# Artificial intelligence and dermatology: opportunities, challenges, and future directions

Daniel I Schlessinger, BA[1]; Guillaume Chhor, MS[2,3]; Olivier Gevaert, PhD[3,4];
Susan M Swetter, MD[5,6]; Justin Ko, MD, MBA[5]; and Roberto A Novoa, MD[5]

## Abstract

The application of artificial intelligence (AI) to medicine has considerable potential within dermatology, where the majority of diagnoses are based on visual pattern recognition.  Opportunities for AI in dermatology include the potential to automate repetitive tasks; optimize time-consuming tasks; extend limited medical resources; improve interobserver reliability issues; and expand the diagnostic toolbox of dermatologists. To achieve the full potential of AI, however, developers must aim to create algorithms representing diverse patient populations; ensure algorithm output is ultimately interpretable; validate algorithm performance prospectively; preserve human-patient interaction when necessary; and demonstrate validity in the eyes of regulatory bodies.

Semin Cutan Med Surg 38:E31-E37 © 2019 Frontline Medical Communications

In 2016, artificial intelligence (AI) pioneer Andrew Ng wrote, "If a typical person can do a mental task with less than one second of thought, we can probably automate it using AI either now or in the near future."[1] Although the potential for AI in health care is controversial,[2-6] it has already achieved notable results in various fields of medicine, ranging from dermatology[7-11] to pathology[12,13] and critical care medicine.[14] In 2017, a team of Stanford researchers published a landmark paper in *Nature*[7] describing the results of an AI algorithm capable of classifying pigmented skin lesions and keratinocytic carcinomas as accurately as 21 dermatologists. While subsequent discussions of AI in dermatology have generally focused on pigmented skin lesions, newer AI algorithms have since been developed to help diagnose onychomycosis[10] and nonpigmented skin cancer.[11] Additionally, novel applications of AI in other fields of medicine may shed light on the future of AI for dermatology. This article aims to review several potential opportunities and important pitfalls of AI in dermatology.

## What is AI?

AI is a broad term referring to the use of computers to mimic human intelligence, and machine learning refers to the computational method employed to develop such algorithms. Although AI was described in 1956,[15] it has recently experienced a resurgence after advances in the field of computer vision—specifically, the 2012 development of the AlexNet algorithm by Krizhevsky, Sutskever, and Hinton—allowed for the development of highly accurate AI algorithms for image classification.[16] The "learning" phenomenon results from the nature of the artificial neural network, a computational structure that is the backbone of most modern-day AI algorithms. A type of neural network called a convolutional neural network (CNN) is particularly suited for most modern-day image recognition tasks. Neural networks are arranged in a minimum of 3 layers: one layer that accepts an input (eg, an image of face), one layer of outputs (eg, the probabilities of the image representing different conditions), and at least one "hidden" layer in between. A "deep" neural network (DNN) refers to a neural network with 3 or more hidden layers. The "thinking" happens in these hidden layers, each of which detects some feature within the input. For example, one layer may focus on the image's colors, another may detect edges, and another may detect specific shapes. With each successive case, the precise weights assigned to each layer are adjusted via an optimization formula until the DNN achieves high enough accuracy. Remarkably, the ideal number of layers and their connections in the DNN is not determined by humans, but by the data itself; hence, the DNN is said to be able to "learn."

Depending on the type of task desired, DNNs are trained using either "supervised learning" or "unsupervised learning." Supervised learning involves training the DNN with a set of cases (eg, images of pigmented skin lesions) and their respective labels (eg, melanoma, seborrheic keratosis, etc.). With each successive case, the DNN "learns" to identify patterns in the cases and associate these patterns with the labels. Because the DNN is given the ground-truth labels for each case, supervised learning is best suited for tasks that involve classification (eg, diagnosis of pigmented skin lesions)[7] or regression (eg, prediction of in-hospital mortality from electronic medical record data).[17] Unsupervised learning involves training a DNN using unlabeled data and is therefore best suited for tasks in which the target is not previously known, such as data clustering (ie, identification of previously unknown cancer subtypes from genomic data)[18] and outlier identification (eg, identification of possible medication order entry errors).[19]

[1]Department of Dermatology, Northwestern University, Evanston, Illinois.
[2]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, California.
[3]Department of Biomedical Data Science, Stanford University, Stanford, California.
[4]Department of Medicine, Stanford University, Stanford, California.
[5]Department of Dermatology, Stanford University, Stanford, California.
[6]Dermatology Service, Veterans Affairs Palo Alto Health Care System, Palo Alto, California.

*Disclosure:* Dr. Ko reports personal fees from LEO Pharma, personal fees from Enspectra, personal fees from Hims,  outside the submitted work. Dr. Novoa reports  disclosures unrelated to the work:  Enspectra- consultant fees; Novartis Argentina-Speaker honorarium; HealthCert-Speaker honorarium; Dermpedia- Speaker honorarium. All other authors have nothing to disclose.

*Correspondence:* Roberto A. Novoa, rnovoa@stanford.edu

1085-5629/13$-see front matter © 2019 Frontline Medical Communications
DOI: 10.12788/j.sder.2019.007

## Five potential opportunities for AI in dermatology

### Repetitive tasks

Whereas human clinicians tire when repeatedly performing the same task, a computer algorithm actually gains diagnostic ability with each successive case. Consider a patient with a prior history of melanoma who presents for skin cancer screening and, on examination, is found to have hundreds of clinically atypical nevi. Although a dermatologist can use a combination of dermoscopy, total-body photography, and meticulous documentation to decide which lesions are the most concerning, DNNs are resistant to exhaustion and diagnostically improve with each successive case exposure, as more data are gathered.

### Time-consuming tasks

Although training a DNN may take days or weeks, most finished algorithms generally run in less than a few seconds, making them useful in several areas of dermatology and medicine. In the field of pathology, an AI-powered pathologist support tool developed by Google AI was able to outline potential areas of micrometastases in lymph node sections taken during mastectomies, making pathologists' jobs subjectively "easier" (when pathologists were surveyed) and halving average slide review time from 2 minutes to 1 minute per slide.[12,13] Similar algorithms could be developed for dermatopathologists, identifying features on slides for pathologists to aid in otherwise laborious histopathologic interpretation—for example, assessment of dermal mitotic rate in invasive cutaneous melanomas.

Within clinical dermatology, reflectance confocal microscopy (RCM) technology is promising for noninvasive diagnosis of pigmented skin lesions,[20] inflammatory skin diseases,[21] and hair disorders,[22] but it is limited by the time and training required to both image and analyze lesions of interest. Diagnosis of a single lesion can take up to an hour, requiring 20 to 30 minutes to first image a lesion and as much time to subsequently interpret it.[20] Research on machine learning analysis of RCM images was done as early as 2008,[23] and newer tools are now able to automatically delineate the dermal–epidermal junction, calculate stratum corneum thickness *in vivo*, and meaningfully quantify RCM interpretation.[24-27]

### Resource-limited tasks

Teledermatology was born from a need for dermatologists in resource-limited areas, but it still requires a dermatologist on the other end of the line; therefore, it merely redistributes dermatologist availability from one place to another. Because it is remote, it is limited by a lack of triage when virtual appointment requests are forwarded by patients or nondermatologist referrers. AI has the potential to substantially improve the teledermatology process by allowing patients to be digitally triaged (or even diagnosed). When a DNN was adapted to images obtained via a Japanese teledermatology system, it was reportedly able to diagnose a wide variety of skin conditions with 99.5% accuracy.[28] In light of the rise of virtual whole slide imaging in recent years,[29] similar algorithms could also be applied to pathology, a field with a 2030 projected net deficit of more than 5,700 full-time equivalent pathologists.[30]

### Tasks with poor interobserver reliability

Diagnostic tasks with poor interobserver reliability pose a double-edged sword for machine learning algorithms: while AI has the potential to substantially improve diagnostic consistency, an AI algorithm is only as reliable as the data labels from which it "learns." Consider the problem of inconsistency of histopathologic diagnosis of dysplastic nevi. A 2017 study of melanocytic lesion diagnosis by 187 board-certified pathologists found a 75% disagreement rate when differentiating between moderately and severely dysplastic nevi and early melanoma.[86] Furthermore, many pathologists did not even demonstrate good intraobserver reliability: when presented with the same case on 2 separate occasions, diagnostic concordance was only 67% overall and was as low as 34% for mildly dysplastic nevi. The potential for a well-trained, highly accurate AI algorithm that could standardize histopathologic assessment of melanocytic lesions cannot be understated—but this presumes that the data labels the algorithm learns from are also accurate. Ideally, any given algorithm is only trained using cases that have been verified by several blinded reviewers using forced agreement to settle on a diagnosis.

AI has already made strides in the standardization of other histopathologic measurement, such as mitotic rate of melanoma, the measurement of which is compromised by both poor interobserver and intraobserver reliability.[31] Recent AI algorithms such as the iDermatoPath system (Munich, Germany) are able to overcome these issues by automatically detecting tumor regions in the whole slide image, identifying mitotic cells with high sensitivity, and presenting the mitosis candidates (ordered by relevance) to the pathologist for approval.[32] This system exemplifies AI at its best: a clinical decision support tool that meaningfully aids, but does not replace, the physician.

Interobserver reliability issues also abound within dermatologic outcome measurement, especially in clinical trials. Researchers have now developed a proprietary mobile application using computer vision to quantify alopecia areata Severity of Alopecia Tool (SALT) scores,[33] a notable achievement that overcomes the time-intensiveness of human-rated Severity of Alopecia Tool scoring (which can necessitate up to 15 minutes per scalp) and the technological challenge of judging color and textural differences between patches of true hair and barren scalp. Additional AI-powered rating tools are on the horizon, including a patient-facing mobile application using AI for the grading of vitiligo severity,[34] although this has not yet been implemented in a clinical research setting.

### Creative diagnostics via AI

Although most uses of AI in dermatology to date have utilized algorithms trained via supervised learning, the field of dermatology could also benefit from algorithms trained via unsupervised learning. As unsupervised learning solely presents cases, but not data labels, to the incipient algorithm, the possibilities are not limited to our scope of knowledge and therefore can be more "out of the box" or creative than supervised learning algorithms. These techniques lend themselves well to clustering the "big data" generated via genomics research. For example, next-generation (high-throughput) sequencing generates massive amounts of data from a person's genome, but traditional statistical methods of analysis to identify similar clusters of data are inefficient. Unsupervised learning algorithms, however, are able to more efficiently identify and cluster data to find novel associations, such as previously unknown genotypic subtypes of cancer.[18,35] Expansion of this research into dermatology may help subtype melanoma from large-scale genomic data and eventually guide the development of more targeted precision therapies for advanced cases.

Even supervised learning algorithms have the potential to identify novel associations. For example, a DNN trained to prognosticate cardiovascular risk from retinal images was incidentally found to be extremely accurate at using the retinal images to identify age, biological sex, smoking status, and systolic blood pressure—tasks that were never previously considered by ophthalmologists to be possible.[36] Similarly novel associations could potentially be discovered in the skin, especially when DNNs are applied to noninvasive imaging modalities, such as dermoscopy, confocal microscopy, and multispectral imaging.

The aforementioned algorithms are discriminative algorithms: they are able to take an image as input and return probable labels as output. Another type of neural network is called a generative network: it takes a data label as input and generates a real-appearing (but "fake") image as output. How does this work? Consider a discriminative neural network that is trained to recognize different types of animals. A generative neural network would work in the reverse, attempting to synthesize entirely new images of animals. One can even link the 2 algorithms together in a kind of symbiotic feedback loop called a generative adversarial network (GAN), in which the generative network repeatedly synthesizes images, and those images are then then tested for authenticity by the discriminative network. With the help of the discriminative network's feedback, the generative network gradually improves at generating authentic-looking images, and the increasing number of new cases also improves the discriminative network's accuracy. GANs have already been used for a wide range of medical applications,[37] from the generation of novel biologic drug chemical structures to the synthesis of computed tomography images[38] from magnetic resonance imaging images, and they have several potential uses in dermatology. One interesting application is data augmentation, which is the use of a GAN to synthesize real-appearing images of otherwise rare medical conditions in order to train better AI algorithms. This has already been prototyped for synthetic images[39,40] ranging from skin lesions to mammograms[41,42] and echocardiograms.[43] Consider the recently developed DNN for the diagnosis of nonpigmented skin cancer.[11] Although it achieved a high overall accuracy, it was poor at diagnosing rare nonpigmented skin lesions such as clear cell acanthomas because these lesions were underrepresented in the training set. If one used a generative network to augment the dataset with synthesized images of rare nonpigmented skin lesions, the discriminative network would see more examples of those cancers, potentially improving its accuracy.

## Five potential pitfalls of AI in dermatology

### Improper training data

An AI algorithm is only as strong as the data from which it learned. If one wanted to create a new DNN for image classification of animals, it would be important to train the algorithm with a wide variety of animals. Similarly, a dermatologic lesion classifier should be trained using a diversity of skin conditions, skin types, anatomic body sites, and patient ages. A pigmented skin lesion classifier that was training on a predominantly Caucasian skin dataset was able to achieve high accuracy when tested on lesions in Caucasian skin but not when tested on lesions in Asian skin—and vice versa when the authors trained a DNN on a predominantly Asian skin dataset.[9] However, when the algorithm was trained on the combined set of images from each dataset, the algorithm was able to achieve high accuracy for both ethnicities. The current International Skin Imaging Collaboration: Melanoma Project, a widely used, open-source, publicly available archive of pigmented skin lesions, is culled from patients in United States, Europe, and Australia, most of whom are lighter skin types, and data are not able to be filtered by skin type.[44] Building on these findings, Adamson and Smith[45] have called for training data diversity so that future AI algorithms in dermatology are able to recognize skin disease accurately in all skin types. Future algorithms should also ensure diversity in anatomic body site and patient ages.

The granularity with which the training images are labeled will also guide the range of outputs of any given algorithm. This poses a problem in dermatology and dermatopathology, in which a great number of diagnoses could either be "lumped" or "split" (eg, the difference between "dysplastic nevus" versus "mildly dysplastic nevus" versus "moderately dysplastic nevus" versus "severely dysplastic nevus"). If one tends to "lump" the data labels into major classes, the resultant algorithm will have more cases of each class to learn from, but its outputs will also be less useful to the end user. On the converse, if one tends to "split" the data labels into subclasses, the algorithm will be able to associate specific features with each subclass, but it might not have enough training cases from each subclass to achieve high accuracy in its outputs. Therefore, improper granularity in data labeling (too much or too little) is a potential pitfall of AI applications in dermatology.

### Uninterpretable output

If AI systems are to be adapted as clinical support tools, clinicians will want a way of verifying the reasoning that goes into an algorithm's decision—akin to a radiologist's report that supports the diagnosis line. Although DNNs may well achieve human expert-level accuracy at diagnosis, they are naturally opaque, leading many to label DNNs as "black boxes."[5,46-49] While humans' "gut instinct" is another example of a black box, most physicians could still identify which reasons swayed their intuition. With AI, though, there is often a tradeoff between predictive accuracy and explainability: the methods that are the most accurate are often the least explainable (eg, DNNs)—and vice versa.[50] Just as physicians are expected to provide rationales for treatment decisions (to prove to themselves, insurers, and other physicians that their decisions are reasonable), AI algorithms should be too. Furthermore, because a single algorithm's outputs might be potentially employed by numerous health care providers, the potential implications of a single erroneous and opaque AI algorithm (ie, one that incorrectly and systematically diagnoses a type of melanoma as benign) are far greater than those of a single flawed physician.

Several techniques exist for improving DNN transparency. One well-known technique is the use of heatmaps, overlaid on top of representative images, allowing the DNN to emphasize which parts of an image most influenced its decision. The results are sometimes surprising, highlighting the DNN's unique ability to inappropriately "piggyback" on context clues. For example, Esteva et al.[7] found that their pigmented lesion classifier "learned" to assign a greater probability of malignancy if the image contained a ruler, likely reflecting increased clinician concern for malignancy when imaging these lesions.[51] Similarly, a DNN trained to diagnose cardiomegaly from inpatient chest x-rays learned to assign a greater probability of heart failure if it noticed that the film was portable, likely reflecting patients who were too sick to get out of bed for an erect posterior-anterior film.[52] Furthermore, while heatmaps can aid in pointing out clear algorithmic flaws for image classification models, they cannot be used for regression models. Finally, what should a heatmap highlight in an image that the DNN decides is disease free? It is more intuitive to highlight problematic areas of the image than to highlight reassuring areas.

Another method for improving DNN transparency involves image segmentation to supplement the subsequent diagnosis. For example, De Fauw et al.[53] designed an "interpretable" DNN to evaluate three-dimensional optical coherence tomography (OCT) eye scans and make subsequent referral recommendations (eg, urgent, semi-urgent, routine, observation only). Importantly, their DNN was designed to produce 2 outputs for each case: an OCT segmentation (in which the DNN color-coded the OCT images) and a diagnosis/referral (in which the DNN used the color-coded OCT segmentation to make a diagnosis and referral suggestion). Such an algorithm improved transparency for the treating physician: instead of simply seeing an opaque diagnosis and referral suggestion, the physician could see what factors contributed to the decision. Just as a dermatologist is able to refer to a dermatopathologist's histopathologic findings if the diagnosis line is in question, future algorithms should similarly attempt to disambiguate results with "explainable" outputs.

### Improper comparison to humans

Just as 2 rating devices in a controlled trial should be compared in identical treatment settings, human raters and AI algorithms are best compared in similar settings. For example, the large majority of trials evaluating DNNs for image classification (eg, pigmented skin lesion classifiers) have utilized DNNs *in silico*, which means that the DNN was not actually utilized in a real-world clinical setting.[54,55] Instead, the images fed to these DNNs were sometimes preprocessed and selected from datasets, which may misrepresent the true diversity of cases. Accordingly, if an AI pigmented skin lesion classifier is only trained and/or tested on lesions that had a histopathologic diagnosis, it would be less accurate for otherwise common skin conditions that are diagnosed clinically (eg, seborrheic keratoses, cherry angiomas, etc.). Just as the clinical validation of new pharmaceuticals necessarily involves hypothesis generation, prototyping, feasibility testing, safety and efficacy validation, and—finally—deployment to real-world practicing clinicians, the best way to validate a DNN involves prospective clinical validation after initial *in silico* feasibility testing.[56] It should be noted that these prospective clinical studies will most likely produce results showing inferior performance of AI algorithms when compared to the *in silico* results.[54]

Not only would DNNs likely perform worse when studied *in vivo*, but humans would also likely perform better. For example, in a 2-part study of detection of breast cancer nodal metastases in lymph node dissection samples,[57] several AI algorithms were compared to a panel of 11 pathologists. In the first part, each pathologist was given 2 hours to review all 129 test slides (less than 1 minute per slide). As Golden points out in an accompanying editorial,[58] not only is this an unrealistically short time to review such a large number of consecutive slides, but pathologists are also likely to request additional special stains in real practice when the diagnosis is in question. In the second part, one pathologist was given unlimited time to review the same slide set. The pathologist took 30 hours but outscored the time-limited group, demonstrating that performance was dramatically underestimated in the unrealistic testing environment imposed on the first cohort. Similar problems have been noted in comparisons of AI algorithms to radiologists,[59] who have been shown to perform the best when using a high-resolution computer screen in a dimly lit reading room.[60]

The same principles of appropriate clinical comparison apply to dermatologists, who perform best when given clinical context regarding the patient. For example, Haenssle et al.[8] compared 58 human raters to a DNN pigmented skin lesion classifier, both with and without the provision of additional clinical information about the patient in question (age, sex, body site, and close-up images). Dermatologists performed better with the information than without it. Although they were still outperformed by the DNN, the 58 humans included in the study were actually composed of 34 dermatologists, 21 dermatology residents, and 3 "anonymous" participants, and the underperformance relative to the DNN was likely due to the latter 2 groups.

### Improper task

The nature of the clinician–patient interaction is fundamentally different from that of the AI–patient interaction. Before asking whether AI might fill a role within medicine, it is instructive to consider the nature of the task with regards to the patient. Generally speaking, the more that a human has to interact with the patient, the less well-suited AI is to that task. For example, while an automated system might be able to accurately process a return for an online order, most people would prefer to speak with an actual human on the other end of the line. Within medicine, AI performs at superhuman level on tasks that do not require any clinician–patient interaction, such as checking drug–drug interactions.[61] AI performs at the expert level for tasks that may require clinician–patient interaction but that can largely be done with a sole image, such as dermoscopic diagnosis of melanoma[7] or radiographic diagnosis of wrist fractures.[62] Deep learning has also been considered as a tool for the classification of diagnosis billing codes.[63-65] AI appears unsuited, however, for tasks that are completely dependent upon the clinician–patient interaction, such as counseling and emotional support—although rudimentary AI-powered tools have even begun to be considered for such tasks, as well.[66-68]

Within dermatology, the problem is more complicated. Dermatologic diagnosis is not often straightforward. Pigmented skin lesions, which the majority of dermatology AI algorithms have addressed, are generally discrete and easily captured by a single, high-quality photograph. Rashes, on the other hand, are highly multifarious, elicit numerous differential diagnoses, and merit a thorough physical exam and history of present illness.

Consider the diagnosis of acne, a "bread-and-butter" dermatologic condition: it does not simply involve binary classification (eg, acne or not), but it, importantly, involves scoring (eg, mild, moderate, severe) as well as a consideration of subtypes (eg, vulgaris, excoriée, steroid-induced, hormonal). For example, an algorithm evaluating a teenager who has acne excoriée might correctly diagnose "acne" but misclassify it as simply "severe acne" upon seeing the erythematous background. If tasked to supply treatment recommendations, the AI might then suggest treatment with antibiotics or isotretinoin, when in reality, the patient needs to address skin-picking tendencies. Even if the AI was able to secure the proper diagnosis of acne excoriée, it is a condition the management of which requires a great degree of clinician trust, ample counseling, and the gradual formation of a strong therapeutic alliance. This is especially true for dermatologic disease because of the chronic nature of many dermatologic conditions (eg, psoriasis, atopic dermatitis, skin cancer). Just because AI algorithms can complete tasks on par with human experts does not mean that they should or that they will be welcomed as such by patients.

Although AI could theoretically play a role in autonomous surgery,[69,70] this field is still in its infancy, and it is unclear exactly how useful such a device would be for dermatologists. Dermatologic surgery often involves cosmetically sensitive areas and necessitates delicacy, creativity, and an artistic eye to achieve cosmesis. As opposed to all other surgical fields, dermatologic surgery involves awake patients who may feel more anxious if the stress of surgery was exacerbated by a robotic, rather than human, surgeon. Finally, AI would be less useful in dermatologic surgery than in other fields, as dermatologic surgeries (eg, biopsies, electrodessication, and curettage) are generally quite fast and involve few sutures.

Finally, it is important to bear in mind the nature of the task from a statistical perspective. Some diseases (eg, Merkel cell carcinoma) are sufficiently rare that, even with a highly accurate system, the positive predictive value will suffer. This has already been noted in a real-world evaluation of an AI-based screening tool for diabetic retinopathy (DR) in a primary care setting, which was effective for ruling out disease but produced a high rate of false positives (15/17) for DR, translating to a specificity of 92% and a positive predictive value of just 12%. Until algorithms are able to achieve high overall accuracy while maintaining a low rate of false positives, AI-powered triage or screening may be more suited to development than true diagnostic tools.[71]

### Legal challenges

In the end, the use of AI-powered clinical decision support tools will be dictated by regulatory approval. To date, the Food and Drug Administration (FDA) has approved 4 AI-powered devices.[62,72-74] Consider the obstacles encountered by IDx-DR (Coralville, Iowa), the first FDA-approved, AI-powered DR diagnostic system for the primary care setting.[72] First, as these tools are considered medical devices, the FDA cannot approve an algorithm that continues to dynamically learn and change during and after the approval process. Why? Although the device would most likely only continue to improve its accuracy with a greater number of cases, there is also the possibility that it would worsen. Some have even postulated that nefarious hackers might take advantage of such an AI-powered device by intentionally feeding it "adversarial" cases, decreasing the algorithm's accuracy.[75] Therefore, the deep learning aspect of the IDx-DR system had to be "locked" prior to the clinical trial—meaning that, although the system initially used AI to learn dynamically and fine-tune its accuracy, it no longer is an autodidactic algorithm.[76] Such rules will likely apply to future AI-powered medical devices. Second, IDx-DR was approved under an alternate device approval process called the De Novo premarket review pathway, a regulatory pathway for some low- to moderate-risk devices that are novel and for which there is no prior legally marketed device.[72] Finally, IDx-DR was approved for a very specific indication: "for use by health care providers to automatically detect more than mild diabetic retinopathy in adults (22 years of age or older) diagnosed with diabetes who have not been previously diagnosed with diabetic retinopathy." Taking these obstacles into account, any future AI-powered tools for the field of dermatology will likely be required to have very specific predefined conditions (ie, usage to detect toenail onychomycosis in adults who are otherwise healthy), and autodidactic functionality will be locked prior to the approval process.

Noting a need to adapt to the digital age, the FDA has developed several pathways for the approval of novel digital devices, including the Digital Health Pre-Certification (Pre-Cert) Program.[77] Pre-Cert aims to streamline the approval process for devices labeled as Software as a Medical Device by building trust with a group of selected companies to free them of the traditional burdens imposed on companies during the drug- and device-approval processes.[78,79] In a speech, FDA Commissioner Scott Gottlieb said, "Employing the Pre-Cert approach to AI may allow a firm to make certain minor changes to its devices without having to make submissions each time."[80] On the other hand, "streamlined" Software as a Medical Device approval via such a program may have unintended consequences, such as methodologic inadequacy and lack of transparency.[81,82]

## The future of AI in dermatology

AI will likely have far-reaching effects within all fields of medicine, including dermatology. Dermatologic AI has been tested in specific use cases *in silico*, but no studies have yet prospectively validated these tools in a clinical setting. Not only do dermatologic tasks generally involve a great deal of counseling, emotional support, history taking, and patient examination, but patients may not welcome an AI clinician. Rather, the best and most likely use of AI within dermatology will be as an adjunct clinical decision support tool, both for dermatologists and nonspecialists. Some advocate the creation of augmented intelligence tools, which integrate AI with human capabilities, complementing—rather than replacing—clinicians.[83] For example, AI-powered clinical decision support tools may assist primary care providers in more appropriately referring patients with dermatologic concerns by more frequently referring patients who need to be managed by a dermatologist (eg, patients with skin cancer) and less frequently referring patients who can be managed without a dermatologist (eg, patients with tinea).

Therefore, the net effect on referrals may very well be zero. The dermatologist, meanwhile, will still be the patient's primary diagnostician and guide counseling and management. However, tasks that are time-consuming or repetitive (eg, the evaluation of total-body photographs in patients with numerous atypical moles) or that have poor intraobserver reliability (eg, various outcome measures in clinical trials) may be automated. Just like with traditional medical devices, dermatologists will review the AI tools' outputs prior to clinical decision-making and discuss the benefits and risks of a further management (eg, a biopsy of a concerning mole) based off of the shared diagnostic acumen of the clinician and AI tool. Although AI has been suggested as an existential threat to dermatology,[84,85] the nature of the dermatologist–patient relationship is fundamentally irreplaceable. However, doctors are increasingly burdened by repetitive tasks and administrative responsibilities that few imagined when starting medical school—the kind of tasks AI is equipped to address. We imagine a bright future in which AI allows dermatologists to spend less time sifting through reams of electronic health record data and coding diagnoses, and more time "doctoring"—examining, counseling, and treating patients.

## References

1. Ng A. What artificial intelligence can and can't do right now. *Harvard Business Review.* https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now. Published November 9, 2016. Accessed January 22, 2019.
2. Hinton G. Deep learning-a technology with the potential to transform health care. *JAMA.* 2018;320(11):1101-1102. doi:10.1001/jama.2018.11100.
3. Mar VJ, Soyer HP. Artificial intelligence for melanoma diagnosis: How can we deliver on the promise? [published online May 22, 2018]. *Ann Oncol*. doi:10.1093/annonc/mdy191.
4. Naylor CD. On the prospects for a (deep) learning health care system. *JAMA.* 2018;320(11):1099-1100. doi:10.1001/jama.2018.11103.
5. Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. *JAMA.* 2018;320(11):1107-1108. doi:10.1001/jama.2018.11029.
6. Zakhem GA, Motosko CC, Ho RS. How should artificial intelligence screen for skin cancer and deliver diagnostic predictions to patients? *JAMA Dermatol.* 2018;154(12):1383-1384. doi:10.1001/jamadermatol.2018.2714.
7. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115-118. doi:10.1038/nature21056.
8. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol.* 2018;29(8):1836-1842. doi:10.1093/annonc/mdy166.
9. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. *J Invest Dermatol.* 2018;138(7):1529-1538. doi:10.1016/j.jid.2018.01.028.
10. Han SS, Park GH, Lim W, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network. *PloS One.* 2018;13(1):e0191493. doi:10.1371/journal.pone.0191493.
11. Tschandl P, Rosendahl C, Akay B, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks [published online November 28, 2018]. *JAMA Dermatol*. doi:10.1001/jamadermatol.2018.4378.
12. Liu Y, Kohlberger T, Norouzi M, et al. Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection [published online October 8, 2018]. *Arch Pathol Lab Med.* doi:10.5858/arpa.2018-0147-OA.
13. Steiner DF, MacDonald R, Liu Y, et al. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *Am J Surg Pathol.* 2018;42(12):1636-1646. doi:10.1097/PAS.0000000000001151.
14. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med.* 2018;24(11):1716-1720. doi:10.1038/s41591-018-0213-5.
15. Russell SJ, Norvig P, Davis E. *Artificial Intelligence: A Modern Approach.* 3rd ed. Upper Saddle River, New Jersey: Prentice Hall; 2010.
16. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*. 2017;60(6):84-90. doi:10.1145/3065386.
17. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine.* 2018;1(1):18.
18. Young JD, Cai C, Lu X. Unsupervised deep learning reveals prognostically relevant subtypes of glioblastoma. *BMC Bioinformatics.* 2017;18(Suppl 11):381. doi:10.1186/s12859-017-1798-2.
19. Dos Santos HDP, Ulbrich AHDPS, Woloszyn V, Vieira R. DDC-Outlier: Preventing medication errors using unsupervised learning [published online April 17, 2018]. *IEEE J Biomed Health Inform.* doi:10.1109/JBHI.2018.2828028.
20. Rajadhyaksha M, Marghoob A, Rossi A, Halpern AC, Nehal KS. Reflectance confocal microscopy of skin in vivo: From bench to bedside. *Lasers Surg Med.* 2017;49(1):7-19. doi:10.1002/lsm.22600.
21. Agozzino M, Gonzalez S, Ardigo M. Reflectance Confocal Microscopy for Inflammatory Skin Diseases. *Actas Dermosifiliogr.* 2016;107(8):631-639. doi:10.1016/j.ad.2016.01.010.
22. Ardigo M, Agozzino M, Franceschini C, Lacarrubba F. Reflectance Confocal Microscopy Algorithms for Inflammatory and Hair Diseases. *Dermatol Clin.* 2016;34(4):487-496. doi:10.1016/j.det.2016.05.011.
23. Wiltgen M, Gerger A, Wagner C, Smolle J. Automatic identification of diagnostic significant regions in confocal laser scanning microscopy of melanocytic skin tumors. *Methods Inf Med.* 2008;47(1):14-25.
24. Gareau D, Hennessy R, Wan E, Pellacani G, Jacques SL. Automated detection of malignant features in confocal microscopy on superficial spreading melanoma versus nevi. *J Biomed Opt.* 2010;15(6):061713-061713. doi:10.1117/1.3524301.
25. Koller S, Wiltgen M, Ahlgrimm-Siess V, et al. In vivo reflectance confocal microscopy: automated diagnostic image analysis of melanocytic skin tumours. *J Eur Acad Dermatol Venereol.* 2011;25(5):554-558. doi:10.1111/j.1468-3083.2010.03834.x.
26. Kurugol S, Kose K, Park B, Dy JG, Brooks DH, Rajadhyaksha M. Automated delineation of dermal-epidermal junction in reflectance confocal microscopy image stacks of human skin. *J Invest Dermatol.* 2015;135(3):710-717. doi:10.1038/jid.2014.379.
27. Kurugol S, Rajadhyaksha M, Dy JG, Brooks DH. Validation Study of Automated Dermal/Epidermal Junction Localization Algorithm in Reflectance Confocal Microscopy Images of Skin. *Proc SPIE Int Soc Opt Eng.* 2012;8207. doi:10.1117/1112.909227.
28. Imaizumi H, Watanabe A, Hirano H, Takemura M, Kashiwagi H, Monobe S. Hippocra: Doctor-to-Doctor TeleDermatology consultation service towards future AI-based Diagnosis System in Japan. Paper presented at: 2017 IEEE International Conference on Consumer Electronics; June 12–14, 2017; Taipei, Taiwan.
29. Nakayama I, Matsumura T, Kamataki A, et al. Development of a teledermatopathology consultation system using virtual slides. *Diagn Pathol.* 2012;7:177. doi:10.1186/1746-1596-7-177.
30. Robboy SJ, Weintraub S, Horvath AE, et al. Pathologist workforce in the United States: I. Development of a predictive model to examine factors influencing supply. *Arch Pathol Lab Med.* 2013;137(12):1723-1732. doi:10.5858/arpa.2013-0200-OA.
31. Garbe C, Eigentler TK, Bauer J, et al. Mitotic rate in primary melanoma: interobserver and intraobserver reliability, analyzed using H&E sections and immunohistochemistry. *J Dtsch Dermatol Ges*. 2016;14(9):910-915. doi:10.1111/ddg.12797.
32. Andres C, Andres-Belloni B, Hein R, et al. iDermatoPath - a novel software tool for mitosis detection in H&E-stained tissue sections of malignant melanoma. *J Eur Acad Dermatol Venereol.* 2017;31(7):1137-1147. doi:10.1111/jdv.14126.
33. Bernardis E, Castelo-Soccio L. Quantifying Alopecia Areata via Texture Analysis to Automate the SALT Score Computation. *J Investig Dermatol Symp Proc.* 2018;19(1):S34-S40. doi:10.1016/j.jisp.2017.10.010.
34. Nugraha GA, Nurhudatiana A, Bahana R. Vi-da: vitiligo diagnostic assistance mobile application. *Journal of Physics: Conference Series.* 2018;978(1):012003.
35. Sweeney TE, Chen AC, Gevaert O. Combined Mapping of Multiple clUsteriNg ALgorithms (COMMUNAL): A Robust Method for Selection of Cluster Number, K. Vol 52015. *Sci Rep*. 2015;5:16971. doi:10.1038/srep16971.
36. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering.* 2018;2(3):158-164.
37. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. *arXiv (Cornell University).* Submitted September 19, 2018. Accessed January 22, 2019.
38. Nie D, Trullo R, Lian J, et al. Medical image synthesis with context-aware generative adversarial networks. Paper presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention; September 11–13, 2017; Quebec City, Quebec.
39. Baur C, Albarqouni S, Navab N. Generating Highly Realistic Images of Skin Lesions with GANs. In: *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer; 2018:260–267.

40. Baur C, Albarqouni S, Navab N. MelanoGANs: High Resolution Skin Lesion Synthesis with GANs. *arXiv (Cornell University).* Submitted April 12, 2018. Accessed January 22, 2019.

41. Korkinof D, Rijken T, O'Neill M, Yearsley J, Harvey H, Glocker B. High-resolution mammogram synthesis using progressive generative adversarial networks. *arXiv (Cornell University).* Submitted July 9, 2018. Accessed January 22, 2019.

42. Kim G, Shim H, Baek J. Feasibility study of deep convolutional generative adversarial networks to generate mammography images. Paper presented at: conference of the Society of Photo-Optical Instrumentation Engineers (SPIE), Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment; March 7, 2018.

43. Madani A, Ong JR, Tibrewal A, Mofrad MR. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *npj Digital Medicine.* 2018;1(1):59.

44. Codella NC, Gutman D, Celebi ME, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). Paper presented at: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); April 4–7, 2018.

45. Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatol.* 2018;154(11):1247-1248. doi:10.1001/jamadermatol.2018.2348.

46. Brouillette M. Deep Learning Is a Black Box, but Health Care Won't Mind. MIT Technology Review. https://www.technologyreview.com/s/604271/deep-learning-is-a-black-box-but-health-care-wont-mind/. Published April 27, 2017. Accessed January 22, 2019.

47. Adler P, Falk C, Friedler SA, et al. Auditing Black-box Models for Indirect Influence. *arXiv (Cornell University).* Submitted February 23, 2016. Updated November 30, 2016. Accessed January 22, 2019.

48. Cabitza F, Rasoini R, Gensini GF. Benefits and Risks of Machine Learning Decision Support Systems-Reply. *JAMA.* 2017;318(23):2356-2357. doi:10.1001/jama.2017.16635.

49. Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in Medicine. *JAMA.* 2017;318(6):517-518. doi:10.1001/jama.2017.7797.

50. Bologna G, Hayashi Y. A Rule Extraction Study from SVM on Sentiment Analysis. *Big Data and Cognitive Computing.* 2018;2(1):6. doi:10.3390/bdcc2010006.

51. Patel N. Why Doctors Aren't Afraid of Better, More Efficient AI Diagnosing Cancer. The Daily Beast. https://www.thedailybeast.com/why-doctors-arent-afraid-of-better-more-efficient-ai-diagnosing-cancer. December 11, 2017. Accessed January 22, 2019.

52. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv (Cornell University).* Submitted July 2, 2018. Updated July 13, 2018. Accessed January 22, 2019.

53. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med.* 2018;24(9):1342-1350.

54. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7.

55. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *npj Digital Medicine.* 2018;1(1):40.

56. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA.* 2018;320(21):2199-2200. doi:10.1001/jama.2018.17163.

57. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA.* 2017;318(22):2199-2210. doi:10.1001/jama.2017.14585.

58. Golden JA. Deep learning algorithms for detection of lymph node metastases from breast cancer: Helping artificial intelligence be seen. *JAMA.* 2017;318(22):2184-2186. doi:10.1001/jama.2017.14580.

59. Oakden-Rayner L. CheXNet: an in-depth review [blog post]. Word Press. https://lukeoakdenrayner.wordpress.com/2018/01/24/chexnet-an-in-depth-review/. Published January 24, 2018. Accessed January 22, 2019.

60. Pollard BJ, Samei E, Chawla AS, et al. The effects of ambient lighting in chest radiology reading rooms. *J Digit Imaging.* 2012;25(4):520-526. doi:10.1007/s10278-012-9459-5.

61. Kuperman GJ, Bobb A, Payne TH, et al. Medication-related clinical decision support in computerized provider order entry systems: a review. *J Am Med Inform Assoc.* 2007;14(1):29-40. doi:10.1197/jamia.M2170.

62. FDA permits marketing of artificial intelligence algorithm for aiding providers in detecting wrist fractures [press release]. United States Food and Drug Administration. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm608833.htm. May 24, 2018.

63. Huang J, Osorio C, Sy LW. An Empirical Evaluation of Deep Learning for ICD-9 Code Assignment using MIMIC-III Clinical Notes. *arXiv (Cornell University).* Submitted February 7, 2018. Accessed January 22, 2019.

64. Medori J, Fairon C. Machine learning and features selection for semi-automatic ICD-9-CM encoding. Paper presented at: Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents; June 2010; Los Angeles, California.

65. Suominen H, Ginter F, Pyysalo S, et al. Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: a method description. Paper presented at: Proceedings of the ICML/UAI/COLT Workshop on Machine Learning for Health-Care Applications; July 9, 2008; Helsinki, Finland.

66. Meghdari A, Shariati A, Alemi M, et al. Arash: A social robot buddy to support children with cancer in a hospital environment. *Proc Inst Mech Eng H.* 2018;232(6):605-618. doi:10.1177/0954411918777520.

67. De Silva D, Ranasinghe W, Bandaragoda T, et al. Machine learning to support social media empowered patients in cancer care and cancer treatment decisions. *PloS One.* 2018;13(10):e0205855. doi:10.1371/journal.pone.0205855.

68. Inkster B, Sarda S, Subramanian V. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR Mhealth Uhealth.* 2018;6(11):e12106. doi:10.2196/12106.

69. Moustris GP, Hiridis SC, Deliparaschos KM, Konstantinidis KM. Evolution of autonomous and semi-autonomous robotic surgical systems: a review of the literature. *Int J Med Robot.* 2011;7(4):375-392. doi:10.1002/rcs.408.

70. Shademan A, Decker RS, Opfermann JD, Leonard S, Krieger A, Kim PC. Supervised autonomous robotic soft tissue surgery. *Sci Transl Med.* 2016;8(337):337ra364. doi:10.1126/scitranslmed.aad9398.

71. Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney M, Mehrotra A. Evaluation of artificial intelligence–based grading of diabetic retinopathy in primary care. *JAMA Netw Open.* 2018;1(5):e182665. doi:10.1001/jamanetworkopen.2018.2665.

72. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems [press release]. United States Food and Drug Administration. https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm. April 11, 2018.

73. FDA permits marketing of clinical decision support software for alerting providers of a potential stroke in patients [press release]. United States Food and Drug Administration. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm596575.htm. February 13, 2018.

74. Arterys Receives First FDA Clearance for Broad Oncology Imaging Suite with Deep Learning [press release]. PR Newswire. https://www.prnewswire.com/news-releases/arterys-receives-first-fda-clearance-for-broad-oncology-imaging-suite-with-deep-learning-300599275.html. February 15, 2018.

75. Finlayson SG, Bowers JD, Ito J, et al. Adversarial attacks on medical machine learning. Science. 2019;363(6433):1287.

76. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine.* 2018;1(1):39. doi:10.1038/s41746-018-0040-6.

77. Digital Health Software Precertification (Pre-Cert) Program [computer program]. Silver Spring, MD: United States Food and Drug Administration; 2019.

78. Statement from FDA Commissioner Scott Gottlieb, M.D., on the agency's new actions under the Pre-Cert Pilot Program to promote a more efficient framework for the review of safe and effective digital health innovations [press release]. US Food and Drug Administration. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm629306.htm. January 7, 2019.

79. Ustun B, Tracà S, Rudin C. Supersparse Linear Integer Models for Interpretable Classification. *arXiv (Cornell University).* Submitted June 27, 2013. Updated April 11, 2014. Accessed January 22, 2019.

80. Gottlieb S. Transforming FDA's Approach to Digital Health [speech to Academy Health's 2018 Health Datapalooza]. US Food and Drug Administration. https://www.fda.gov/NewsEvents/Speeches/ucm605697.htm. April 26, 2018.

81. AI diagnostics need attention. *Nature.* 2018;556(7699):285. doi:10.1038/d41586-018-03067-x.

82. Park SH. Regulatory approval versus clinical validation of artificial intelligence diagnostic tools. *Radiology.* 2018;288(3):910-911. doi:10.1148/radiol.2018181310.

83. Ismail, N. Augmented intelligence: why the human element can't be forgotten. Information Age. http://www.information-age.com/augmented-intelligence-human-element-cant-forgotten-123466894/. Published June 21, 2017. Accessed January 22, 2019.

84. Molteni M. If You Look at X-Rays or Moles for a Living, AI Is Coming for Your Job. Wired. https://www.wired.com/2017/01/look-x-rays-moles-living-ai-coming-job/. Published January 25, 2017. Accessed January 22, 2019.

85. Scutti S. 'Automated dermatologist' detects skin cancer with expert accuracy. CNN. https://www.cnn.com/2017/01/26/health/ai-system-detects-skin-cancer-study/index.html. Published 2017. Accessed January 22, 2019.

86. Elmore JG, Barnhill RL, Elder DE, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *BMJ (Clinical research ed).* 2017;357:j2813-j2813.