# Towards Transparency in Dermatology Image Datasets with Skin Tone Annotations by Experts, Crowds, and an Algorithm

MATTHEW GROH, Massachusetts Institute of Technology, USA
CALEB HARRIS, Massachusetts Institute of Technology, USA
ROXANA DANESHJOU, Stanford, USA
OMAR BADRI, Northeast Dermatology Associates, USA
ARASH KOOCHEK, Banner Health, USA

While artificial intelligence (AI) holds promise for supporting healthcare providers and improving the accuracy of medical diagnoses, a lack of transparency in the composition of datasets exposes AI models to the possibility of unintentional and avoidable mistakes. In particular, public and private image datasets of dermatological conditions rarely include information on skin color. As a start towards increasing transparency, AI researchers have appropriated the use of the Fitzpatrick skin type (FST) from a measure of patient photosensitivity to a measure for estimating skin tone in algorithmic audits of computer vision applications including facial recognition and dermatology diagnosis. In order to understand the variability of estimated FST annotations on images, we compare several FST annotation methods on a diverse set of 460 images of skin conditions from both textbooks and online dermatology atlases. These methods include expert annotation by board-certified dermatologists, algorithmic annotation via the Individual Typology Angle algorithm, which is then converted to estimated FST (ITA-FST), and two crowd-sourced, dynamic consensus protocols for annotating estimated FSTs. We find the inter-rater reliability between three board-certified dermatologists is comparable to the inter-rater reliability between the board-certified dermatologists and either of the crowdsourcing methods. In contrast, we find that the ITA-FST method produces annotations that are significantly less correlated with the experts' annotations than the experts' annotations are correlated with each other. These results demonstrate that algorithms based on ITA-FST are not reliable for annotating large-scale image datasets, but human-centered, crowd-based protocols can reliably add skin type transparency to dermatology datasets. Furthermore, we introduce the concept of dynamic consensus protocols with tunable parameters including expert review that increase the visibility of crowdwork and provide guidance for future crowdsourced annotations of large image datasets.

CCS Concepts: • **Human-centered computing** → **Computer supported cooperative work**; **Empirical studies in collaborative and social computing**; • **Applied computing** → **Health informatics**.

Additional Key Words and Phrases: crowdsourcing, artificial intelligence, healthcare, fairness, accountability, transparency

**ACM Reference Format:**
Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. 2022. Towards Transparency in Dermatology Image Datasets with Skin Tone Annotations by Experts, Crowds, and an Algorithm. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 521 (November 2022), 26 pages. https://doi.org/10.1145/3555634

Authors' addresses: Matthew Groh, groh@mit.edu, Massachusetts Institute of Technology, USA; Caleb Harris, Massachusetts Institute of Technology, USA; Roxana Daneshjou, Stanford, USA; Omar Badri, Northeast Dermatology Associates, USA; Arash Koochek, Banner Health, USA.


2573-0142/2022/11-ART521
https://doi.org/10.1145/3555634

## 1 INTRODUCTION

Artificial intelligence (AI) algorithms hold promise for improving image-based clinical diagnosis tasks ranging from identifying breast cancer in mammograms [74] to classifying skin lesions based on a single image [31] to predicting the diagnosis of hundreds of diverse skin conditions based on a few images and a brief patient history [67]. The combination of algorithmic predictions with physician diagnostic skill has the potential to create large efficiency and welfare gains in healthcare [95]. In particular, AI systems can enhance specialists' diagnostic performance on specific tasks (e.g. identifying pneumonia on chest radiographs [90] and predicting hypoxaemia risk from operating room data [70]) but incorrect predictions from an AI system can mislead specialists and generalists alike [43, 55, 102]. In fact, inaccurate advice regardless of whether it comes from an AI or human tends to decrease physicians' accuracy on diagnostic tasks [36]. Moreover, the algorithm appreciation effect [68] suggests that inaccurate advice from an algorithm is likely to have more negative effects than the same advice given by a human.

Given the consequences of inaccurate advice in healthcare, ethical and responsible algorithm-in-the-loop decision systems should require the systems to be both accurate and also unbiased with regard to sensitive attributes like race and gender. Moreover, these systems should be transparent such that medical experts can reliably assess algorithmic performance [39, 47]. These principles for ethical systems are particularly important because algorithms are prone to make unexpected errors on out-of-distribution data. Due to biases in dataset representation, protected classes are more likely to be out-of-distribution [4, 9, 66]. Moreover, when accurate yet non-equitable algorithmic risk assessments are used as decision support tools they have been shown to alter decision maker's risk aversion and lead to unexpected and sometimes unwanted shifts in human decision-making [40].

Yet the vast majority of AI algorithms for diagnostic tasks in dermatology are trained on datasets that lack transparency with regards to demographic and skin tone attributes [22, 108]. Due to this lack of transparency, it is difficult to assess what data may be out-of-distribution and this leads to the potential for unexpected errors that could have otherwise been addressed. For example, given the under representation of dark skin in educational resources [2, 5, 28, 65, 69] and online dermatology atlases [44], it is unknown the full extent to which dark skin is under-represented in many of the large dermatology image datasets. For the few datasets that do include skin type information, dark skin types are underrepresented [22]. This is particularly problematic because AI algorithms for classifying the skin condition in an image are more accurate on images that match the skin color upon which the algorithm is trained than images that do not match the skin color [44]. An analysis of three AI algorithms (ModelDerm [48], DeepDerm [31], and Ham 1000 [103]) reveal that images of dark skin show a drop in all three models' accuracy rates relative to rates in images of light skin [23].

One approach for increasing transparency in dermatology image datasets and their resulting AI algorithms is to annotate skin tone with Fitzpatrick Skin Type (FST) like the algorithmic audit of accuracy disparities in facial recognition by Buolamwini and Gebru 2018 [13]. FST is a clinical measurement developed and used by dermatologists to assess patients' sun sensitivity for dosing phototherapy or chemophototherapy. Clinical FST has been criticized for subjectivity [45], is not designed for classifying race or skin color [107], and often involves not just assessing skin tone but assessing a patient's hair color and eye color [109]. Despite the imperfections and biases of clinical FST as a proxy for skin tone and an assessment of differential healthcare risks [87], AI researchers have appropriated FST to estimate skin tone labels for algorithmic audits of tasks like classifying skin disease [27, 44, 67, 91] and facial recognition algorithms [13, 50]. In this paper, we distinguish FST as recorded in a clinical patient-provider interaction as "clinical FST" and FST as recorded based on a single image as "estimated FST."

While estimated FST has been frequently used in computer vision tasks, basic questions have not been explored about its use for labeling image datasets: Who is qualified to annotate images with estimated FSTs? More specifically, should estimated FST annotations on large-scale datasets be limited to board-certified dermatologists? How concordant are board-certified dermatologists, particularly on the kinds of datasets used in computer vision? Would the annotations of trained annotators, crowdsourced labor, or algorithms differ significantly from board-certified dermatologists? These are empirical questions, which are connected more broadly to questions about what makes desirable data and how race and gender should be annotated in image datasets [99, 100]. Notably, we limit the focus on estimated FST because it is a method used in algorithmic audits based on clinical medicine and it allows granular analysis which would not be captured by race alone [13]. While most large image datasets with estimated FST annotations are labeled by dermatologists [13, 27, 67, 91], the "Casual Conversations" dataset is annotated by trained annotators [50] and the "Fitzpatrick 17k" annotations are generated by applying a dynamic consensus protocol to crowdsourced annotations [44].

As an alternative to human-annotated estimated FST, researchers have proposed and used the Individual Typology Angle to FST (ITA-FST) algorithm, a computer vision algorithm that converts the RGB values of an image into a single metric for constitutive pigmentation, to estimate apparent skin tone from images [60, 63]. Prior work shows that ITA-FST is strongly correlated with Melanin Index [109], which is sometimes used in assigning clinical FSTs [30]. However, recent research in photodermatology suggests that ITA used for constitutive pigmentation is a poor proxy for clinical FST [89].

Prior work suggests that crowdsourced estimated FST annotations are generally within 1 point of an expert board-certified dermatologist's annotation, but Groh et al (2021) compared crowd annotations with only a single expert, do not include statistical analyses of inter-rater reliability, do not compare ITA with experts' annotations, and do not examine nuances around the compositions of the crowd or edge cases where the crowd is prone to err [44]. We present evidence that the inter-rater reliability between three board-certified dermatologists is comparable to the inter-rater reliability between board-certified dermatologists and crowdsourcing methods but not the ITA-FST algorithm. However, for a subset of images with high disagreement between crowd annotators, we find higher inter-rater reliability between board-certified dermatologists than board-certified dermatologists and the crowd.

In summary, our contributions are the following:

**(1)** We evaluate the inter-rater reliability between three medical experts, a computer vision algorithm, and two crowdsourcing approaches for annotating images of skin conditions with estimated FST, which is useful for increasing transparency into how algorithms perform on images of different skin tones. On a set of 320 images drawn from dermatology textbooks [12, 41, 46, 54, 57, 58, 78, 110], we do not find a statistically significant difference when comparing the Pearson Correlation Coefficients ($\rho$) between three medical experts with the $\rho$ between each medical expert and either of the crowdsourcing methods. In contrast, we do find a statistically significant difference in the annotations produced by the ITA-FST algorithm. These results suggest that crowdsourcing (but not the ITA-FST algorithm) can be a reliable source for generating estimated FST annotations on large-scale datasets of images intended for training and evaluating AI models to classify skin disease. However, we include important caveats. First, our qualitative results show the crowd will sometimes make errors that the medical experts would be unlikely to make. Second, a quantitative follow-up with 140 images drawn from two online dermatology atlases [3, 35] shows the results are robust to 70 images randomly sampled from the 91% of images with relatively low crowd disagreement but on a random sample of 70 images from the 9% of images with relatively high crowd disagreement, we find the crowd annotations can be significantly different from the experts'

annotations. Third, the image-based estimated FST annotations are subject to lighting, image quality, and pose variability that are not an issue for in-person assessments

**(2)** In order to increase visibility into the process of human annotation of large image datasets and guide future work, we introduce and describe a dynamic consensus protocol for aggregating crowdsourced estimated FST annotations using the following transparent, adjustable criteria: (1) **consensus thresholds**, (2) **qualified annotations**, (3) **failure reports** [14], (4) **agreement metrics** and (5) **expert review**. We apply this procedure to the publicly available "Fitzpatrick 17k" dataset of 16,577 images to evaluate inter-rater reliability across crowdsourcing annotation methods, estimate the proportion of images that experts should review, and conduct expert review on 140 images.

## 2 BACKGROUND AND RELATED WORK

### 2.1 Data Documentation for Increasing Transparency and Accountability in Algorithms

Critical frameworks documenting both machine learning datasets and their resulting models promote transparency and accountability by enabling nuanced analyses that can expose unwanted biases. Examples of guiding frameworks for detailing data, its definitions, and its associated models' potential harms include *Data Statements for Natural Language Processing*, *The Dataset Nutrition Label* [51], *Model Cards for Model Reporting* [79], and *Datasheets for Datasets* [37]. The seminal algorithmic audit of accuracy disparities in facial recognition by Buolamwini and Gebru 2018 relied on documenting estimated FST annotations and evaluating algorithmic performance across FSTs and found significant intersectional accuracy disparities [13]. Estimated FST annotations have also been helpful in documenting accuracy in machine learning models for dermatology [23, 44]. With appropriate, inclusive data, algorithms can increase accountability by both serving as a diagnostic tool to detect discrimination and formalizing our definitions around a social problem like inequities in healthcare across gender, race, and skin color [1, 61].

Beyond cataloguing the elements of a dataset, data documentation can also question the existence of categories within the data and inform the question posed by Miceli et al 2022: "Is this information sufficient in itself to explicate unjust outcomes" [75]? For a large number of datasets with images of humans, the definitions of both race and gender in databases lack critical engagements, are overly reductive, and require more than an outside observer looking at a photograph to annotate appropriately [77, 100]. This is particularly problematic because the definition of a category, class, or outcome will impact how disparate treatment and disparate impact arise in the data [9]. For example, Obermeyer et al 2019 report that an algorithm for predicting health risk of millions of people in the United States using cost of care as a proxy for health needs led to the following bias: "At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses" [85]. This racial bias in a healthcare setting is not only a problem of selecting the right outcome measure, but a deeper problem that involves a history in the United States of "segregated hospital facilities, racist medical curricula, and unequal insurance structures, among other factors" [11]. Dataset documentation is an initial step that enables critical researchers to both identify empirical biases, question the definitions of specific data features, and inspect the data generating process. Data documentation is particularly helpful to bridge collaboration between data scientists and subject matter experts to make knowledge and processes explicit such that both groups of people can ask the right question [72]. As bias is uncovered in data, researchers can offer new insights into bias as a starting point for "studying up" [84] with a critical focus on accountability and power dynamics in the underlying data generation process [8].

Extracting categories and clusters from complex data involves value judgments. For example, Scheuerman et al 2021 highlight the tensions in the development of computer vision datasets between efficiency and care, universality and contextuality, impartiality and positionality, and model work and and data work [99]. In crowd sourcing tasks, categorizing can become problematic when crowdworkers have limited attention and expertise [6] and when crowdworkers are overly constrained by power dynamics such that the crowd annotates data based on their expectations of how a client sees the world rather than their own sense of how the world looks [76]. One recent applied example from CSCW shows that accessible interfaces with high degrees of freedom enable crowdworkers to categorize data that can appropriately filter harmful content generated by AI [71]. Another example from recent research in CSCW highlights the potential for failure reports [14] – open-ended descriptions of model errors – to help navigate unexpected systematic failures. We expand on the concept of failure reports in Section 3.3 where crowdworkers can transcend the menu of multiple choice annotations (in our case estimated FST I-VI and "not applicable") to free-text responses where images can be flagged for being incorrectly labeled, inappropriate, or irrelevant.

In dataset documentation, epistemic authority is an important value judgment. How should data be annotated and who or what should do it? Data annotation is often less straightforward and more complex than it seems. Data work is often time-consuming, opaque (unless there's good documentation), and not well rewarded; in field interviews with data workers, one interviewee exclaimed, "Everyone wants to do the model work, not the data work," which is a sentiment shared by many interviewees [98]. Moreover, reasonable people often disagree on color classification [38] and medical experts often disagree on medical diagnoses [93, 94]. In fact, in one study comparing referral and final diagnoses across 280 patients, significant disagreements appear in 21% of cases [105]. Instead of assigning epistemic authority to any particular individual or algorithm for a subjective task, we follow prior work that treats epistemic authority based on inter-annotator agreement [20, 32, 33, 52]. Disagreement between annotators is not necessarily indicative of poor quality annotations or bias, but instead, disagreement can help reveal the subjectivity involved in a particular task and a particular example [7].

In order to answer who or what is epistemically qualified to annotate data with information that can provide transparency and accountability into potential biases, we need to examine the level of agreement produced by different methods. The first step involves measuring the subjectivity of the task by measuring the degree of disagreement among experts. Next, we compare alternative methods (e.g. an algorithm and crowd methods) to the level of disagreement among experts. If an alternative method does not disagree significantly more with experts than experts do with each other, then we can call the alternative method generally comparable to experts. While an alternative method may be generally comparable to experts, edge cases may arise where experts have significantly lower disagreement among themselves than with the crowd. For example, in the case of estimated FST images where a rare skin disease has transformed the color of the skin, non-experts may have higher levels of disagreement than experts. In the framework of Muller et al 2019 *How Data Science Workers Work with Data* [82], this annotation process would be described as "Ground Truth as Created" where human expertise applied to images informs an analysis of the similarity of various methods for annotating estimated FST. While Muller et al 2021 [83] write that "it is widely agreed that SME-labeled data [data labeled by subject matter experts] is the "gold standard" data source for high quality labeled data for specialized tasks," we take a step back from this assumption and empirically evaluate how well crowds and an algorithm compare to estimated FST annotations of images by board-certified dermatologists. The dynamic consensus threshold process described in 3.3 can represent both Muller et al 2021's "Principled design" and "Iterative design" to ground truth annotation because the annotation process is planned and well-defined, but aspects of the dynamic consensus threshold process (e.g. failure reports and expert review) allow

for clarifications, adjustments, and potential re-definitions based on collaboration back and forth between the human annotators examining data at the record level and data scientists examining the data at the dataset level.

### 2.2 Designing Transparency into Clinical Decision Support Systems

Clinical decision support systems (CDSSs) are systems designed to support healthcare providers in medical decision-making. Past work at CSCW has documented the following relevant onboarding criteria for healthcare providers to interact with CDSSs: capabilities and limitations, functionality, design objective, relative strengths and weaknesses of an algorithm, performance of a model on domain specific cases including how the model's idiosyncrasies compare with human idiosyncrasies [16, 17]. Transparency on subgroups within the data is an integral component to onboarding healthcare providers such that they develop an understanding for when they should override the system, which is important when an algorithm makes an erroneous prediction [24]. In practice, healthcare professionals seek to compare algorithmic errors in CDSSs with their own errors [15]. Well-documented data enables healthcare providers (and researchers) to examine subcategories on which algorithms are likely to error, which is important for establishing trust that the algorithm will lead to positive results for vulnerable patients [29, 106] and useful for identifying what kind of data should be collected to reduce accuracy disparities [19].

Recent deployments of deep learning systems for health reveal that algorithms trained on retrospective datasets may not be ecologically valid [10]. In particular, an algorithm applied to data that deviates from the training data is prone to unexpected errors on the out-of-distribution examples. One approach to handling out-of-distribution data is training a classifier to predict whether an image is out-of-distribution and if it is to abstain from generating an algorithmic classification [96]. Decision support systems tend to be less effective on out-of-distribution examples; in evaluations of algorithms that generally outperform humans, the performance gap in accuracy between humans informed by the algorithm and the algorithm is lower in out-of-distribution examples than in-distribution examples [66]. If a particular skin tone is out-of-distribution for a particular disease, this is important for clinicians and model developers to know, so they are aware what kind of data might constitute a context shift [42].

Recently, Jain et al 2021 completed a retrospective study that showed how a deep learning based CDSS may help non-specialists such as primary care physicians and nurse practitioners diagnose skin disease with higher accuracy (defined as agreement with reference conditions) and possibly reduce biopsy and referral rates to dermatologists than the providers would without the system [56]. Decision support systems have the potential to improve the quality of dermatological care, and as such, it is important to evaluate the underlying skin disease classification algorithm on diverse skin tones to address potential accuracy disparities given the context of skin tone and race in the United States healthcare system and computer vision applications. However the algorithm used in Jain et al 2021 was trained on only 46 images of FST VI skin and 510 images of FST V, which was 0.3% and 3.2% of the entire training set, respectively [67]. The lack of images of dark skin types in this dataset means this model may be prone to a higher level of unexpected algorithmic errors on future images of dark skin types [44].

## 3 METHODS FOR FITZPATRICK SKIN TYPE ANNOTATIONS

The Fitzpatrick labeling system is a six-point clinical scale used by dermatologists for classifying skin types based on photo-reactivity of skin and was originally intended to be used for photochemotherapy [34]. See Table 3 in Appendix for a copy of the original description of the Fitzpatrick Scale. We note that the original scale does not include nuanced skin tones or color beyond white, brown, and black. While the Fitzpatrick scale is highly correlated with an individual's melanin index (measured

by narrow-band spectrophotometric devices), the Fitzpatrick scale is a subjective measure [59]. In clinical practice, clinical FSTs are visually assessed by dermatologists based on the colors of a patient's skin, hair, and eyes and their history of sunburns [109]. In a study comparing self-reports to a single dermatologist's clinical FST determination, the dermatologist's assessment was found to be significantly more reliable than individuals' self reports [30]. Recently, researchers have used the estimated FST to annotate images and evaluate algorithmic fairness of AI models across apparent skin tones [13, 22, 44, 56, 67]. While the original Fitzpatrick scale was not designed to categorize skin color, it is often used as such in clinical practice [107] and it serves as a starting point (albeit imperfect given the coarseness of categories for skin color along sepia tones) for algorithmic audits [13].

### 3.1 Expert Labels from Board-Certified Dermatologists

We asked three board-certified dermatologists – experts with deep experience examining skin conditions and assessing patients' clinical FST – to annotate images with the estimated FST. Each expert provided independent estimated FST annotations for 320 images collected from dermatology textbooks and 160 images collected from online dermatology Atlases. We collected 1,380 estimated FST annotations from experts. Given the inherent subjectivity of this task, we present ranges of experts' annotation across these images: 3-5% Type I, 28-31% Type II, 29-30% Type III, 14-15% Type IV, 14-15% Type V, and 4-9% Type VI. Likewise, the distribution across the the 160 images is 4-20% unknown, 0-3% Type I, 17-28% Type II, 26-34% Type III, 20-24% Type IV, 9-13% Type 5, and 1-8% Type 6.

The experts noted that estimated FSTs will not necessarily match in-person assessments because clinical FST relies on not just skin color but eye color, hair type and color, and history of sunburns. Moreover, clinical FST based on an in-person assessment considers an individual's entire body across varying lighting conditions while estimated FSTs based on a single image are restricted to a limited view of the body under a single lighting condition. As such, image-based estimated FST assessments will be have less information and be fundamentally more noisy with less inter-annotator agreement than clinical, in-person assessments. For example, clinical images of dermatological conditions differ in what part of the body is photographed, how the photograph is framed (from the camera's angle and zoom level to the patient's pose), how the lighting illuminates the image, and how the skin disease has transformed the patient's skin. We discuss further limitations of estimated FST annotations in the Limitations section.

### 3.2 Algorithmic Labels from Individual Typology Angle

Following computer vision papers using ITA-FST for algorithmic audits [60, 63], we compute ITA-FST annotations for each image. ITA was designed to classify skin color in a Caucasian population based on healthy skin in an image [18]. While ITA was not designed for all people, research shows ITA-FST correlates with both Melanin Index and clinical FST [25, 109]. However, ITA-FST and clinical FST are designed to measure constitutive pigmentation and sun-reactivity, respectively, and recent research suggests they are poor proxies of one another [89]. In order to calculate ITA-FST more precisely, researchers developed YCbCr masks to mask pixels outside a range of pre-specified colors [62] to reduce the noise of ITA-FST estimates. YCbCr masks are imperfect and often mask healthy skin or fail to mask non-skin parts of an image in the range of skin colors, but without YCbCr masks ITA-FST estimates are even more varied because very light or very dark backgrounds can influence the estimate. For example, YCbCr often fails to mask white underwear of dark skin people leading to the ITA-FST algorithm making errors in estimating skin tone that a reasonable human would not make.

We calculate ITA using the default D65 illuminant over the healthy skin pixels identified by YCbCr masks, and we convert the scores to estimated FSTs that minimize discrepancy between algorithmic labels and the experts' labels on the 320 textbook images and 140 images from dermatology atlases following procedures described by Groh et al 2021 and Krishnapriya et al 2022 [44, 63]. See Algorithm 2 in Appendix for details on transforming ITA scores to FSTs.

### 3.3 Dynamic Consensus Protocol for Crowd Labels

In order to crowdsource estimated FST annotations for images, we collaborated with Scale AI and Centaur Labs, two companies that specialize in labeling large image datasets via dynamic consensus protocols applied to crowdworkers' annotations. In this section, we identify five key components of a dynamic consensus protocol based on the process at Centaur Labs.

Dynamic consensus refers to the process of transforming multiple annotations from independent sources at different times on a single image into a consensus annotation. A dynamic consensus differs from a standard consensus metric like a mean, median, or mode because a dynamic consensus pre-specifies a **consensus threshold**, which must be met before annotations are transformed into accepted responses. For example, the annotations produced by Centaur Labs included a consensus threshold defined as either (a) a single category (across the 6 FSTs and a category for not applicable) has 3 more annotations than any other category or (b) the majority label if a consensus has not been reached after 20 annotations.

**Qualified annotations** are defined as annotations by individuals who have passed a task specific quality control procedure. In contrast, disqualified annotations are annotations by individuals who have failed the task specific quality control. The third category, non-qualified annotations, are annotations by individuals who have not yet been assessed by a task specific quality control procedure. In general, quality control is determined by the proportion of an individual's annotations that correspond to a set of expert annotations. Given the subjectivity of estimated FST, we use both an expert's annotations to compare against 320 annotations collected from both Scale AI and Centaur Labs and the dynamic crowd consensus annotations on the rest of the images as measures on which to evaluate annotation quality. For both Scale AI and Centaur Labs, we seed the dynamic crowd consensus protocol with expert annotations to avoid crowd prejudice equilibria that can arise in cold-start annotation tasks [26]. In the dynamic crowd consensus protocol devised with Centaur Labs, we included a qualified minimum agreement of 40% and qualified minimum and maximum annotations at 25 and 50, which means an individual is qualified only after attaining 40% agreement on 25 images and then an individual only remains qualified as long as her agreement remains above 40% for the 50 most recently annotated images. We selected the 40% minimum agreement threshold with Centaur Labs for two reasons: first, it is significantly above random guessing, which would be 16.7%, and second, we had previously found that 48% of consensus annotations by Scale AI matched expert 1's annotation exactly, so we rounded down to the nearest multiple of ten. The qualified minimum and maximum annotation levels were suggested by Centaur Labs based on past performance of their crowdworkers on other similar datasets. We did not include a dynamic quality control procedure for (dis)qualifying annotations with Scale AI.

For the 320 textbook images, Scale AI provided 156,566 annotations (ranging from 378 to 1094 annotations per image) and Centaur Labs provided 7,999 qualified annotations (ranging from 2 to 93 qualified annotations per image). For an additional 16,577 images from the Fitzpatrick 17k dataset, Scale AI provided 62,710 annotations (with an interquartile range of 4 to 4 annotations per image) and Centaur Labs provided 265,279 qualified annotations (with an interquartile range of 9 to 20 qualified annotations per image) [44]. In total, we collected 492,554 estimated FST annotations from crowd workers.

In addition to estimated FST annotations, we collected **agreement metrics** for measuring the agreement and difficulty of annotating images with estimated FSTs. These agreement metrics are weighted by each individual annotator's agreement with the expert annotations and defined for each image as follows: agreement is the weighted, qualified annotations with the consensus label divided by the weighted, qualified annotations; difficulty is the weighted, qualified annotations without the consensus label divided by the weighted, qualified annotations. In algebraic notation, agreement and difficulty can be written as $A = \frac{Q_c}{Q}$ and difficulty as $D = \frac{Q \notin Q_c}{Q}$ where $Q_c$ is the weighted number of qualified annotations with the consensus label, $Q$ is the weighted number of qualified annotations, and $Q \notin Q_c$ is the weighted number of qualified annotations that are not the consensus label.

Another criteria for assessing and improving the reliability of an images' annotations is the incorporation of **failure reports** [14]. Failure reports are comments on flagged images by annotators indicating that an image is either incorrectly labeled or inappropriate or irrelevant. Failure reports allow crowdsourced workers to transcend the 7 multiple choice labels (the FST scale and the not applicable option) to provide text-based feedback on the image. In the annotations by Centaur Labs, we stop labeling any image which was flagged as inappropriate or irrelevant once or flagged as incorrect twice. Across, the 320 textbook images, we received 20 failure reports on 17 images. We discuss the details of these failure reports in Section 4.2.

The final criteria for crowdsourcing is **expert review**, which is particularly useful for focusing the efforts by experts on the edge cases with high disagreement among crowd annotators. Expert review consists of experts reviewing flagged images without seeing the distribution of labels to adjudicate the annotation. We discuss the results of expert review in Section 4.3.

## 4 RESULTS COMPARING ANNOTATIONS ON 320 TEXTBOOK IMAGES

We asked three board-certified dermatologists to annotate 320 images with FSTs, and we find that the annotations of any two experts match exactly on 50-55% of images and match within one unit on 92-94% of images. Figure 1 presents a confusion matrix comparing the annotations of the first two experts.

In comparison to the two experts' labels, the algorithmically generated annotations for the 320 images are much less similar. The ITA-FST algorithm produces Fitzpatrick labels identical to expert 1, 2, and 3 in 27%, 31%, and 40% of images, respectively, and is off by no more than a single unit (i.e., FST I vs FST II) in 70%, 69%, and 76% of images, respectively. See Figure 5 and Figure 6 in the Appendix for confusion matrices examining annotation discrepancies between experts 1 and 2 and the Scale AI, Centaur Labs, and ITA-FST algorithm.

The inter-rater reliability between the two experts' and the crowds' annotations is much more similar across the 320 images. The labels produced by Scale AI and Centaur Labs match expert 1 exactly in 48% and 40% of images, match expert 2 exactly in 50% and 38% of images, and match expert 3 exactly in 58% and 43% of images, all respectively. Likewise, the annotations produced by Scale AI and Centaur Labs are off by no more than a single unit from expert 1's annotations in 94% and 87% of images, expert 2's annotations in 91% and 79% of images, and expert 3's annotations in 93% and 88% of images.

### 4.1 Quantitative Assessment of Inter-Rater Reliability on 320 images

In light of the subjectivity of estimated FST annotations, we evaluate annotation performance by comparing inter-rater reliability between pairs of experts with the inter-rater reliability between experts and each non-expert annotation method. Specifically, we measure inter-rater reliability

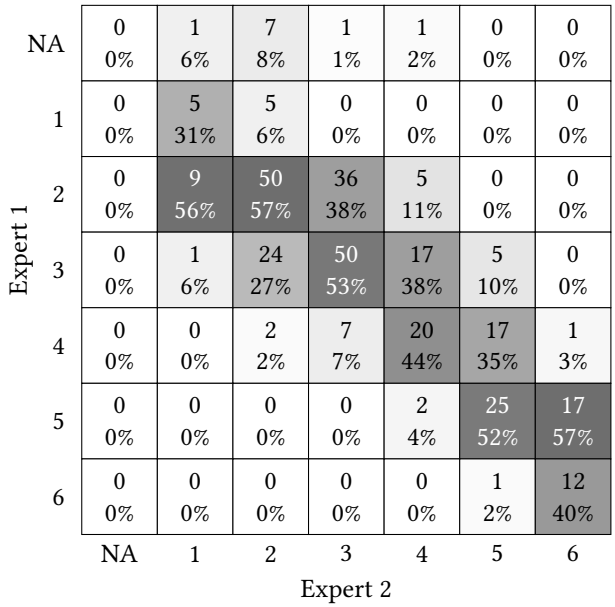


Fig. 1. Confusion matrix comparing two board-certified dermatologists' Fitzpatrick skin type annotations on 320 images from dermatology textbooks.

using the Pearson Correlation Coefficient ($\rho$) between two annotation methods, and we evaluate the statistical significance following the Fisher Z transformation for comparing independent correlations [32]. We describe the pseudocode for comparing $\rho_{E_i,E_j}$ with $\rho_{E_{i,Method}}$ in Algorithm 1 in the Appendix where $E_i$ and $E_j$ refer to one of the three experts and $E_{Method}$ refers to one of the non-expert annotation methods. When calculating the $\rho_{X,Y}$ between two annotation methods X and Y, we drop annotations that either method marks as not applicable.

We find the inter-rater reliability of the ITA-FST algorithm is significantly lower than the inter-rater reliability of experts. The correlation between the first two experts' annotations is $\rho_{E_1,E_2} = .84$ ($E_1$ and $E_2$ refer to expert 1 and 2, respectively) whereas the correlation between the ITA-FST algorithm and any of the experts is $\rho_{ITA,E_1} = .57$, $\rho_{ITA,E_2} = .52$, and $\rho_{ITA,E_3} = .55$. The differences between any pair of experts $\rho_{E_i,E_j}$ and $\rho_{E_1,ITA}$, $\rho_{E_2,ITA}$, and $\rho_{E_3,ITA}$ are statistically significant ($p < 0.00000001$). We present the correlations and the p-value of the comparisons of correlations in Table 1. We also present a heatmap of inter-rater reliability as measured by $\rho$ in Figure 2.

In contrast to the low inter-rater reliability between experts and the algorithm, we find the inter-rater reliability of expert and crowdsourced annotations to be comparable. Notably, the crowdsourced annotations are slightly more correlated with experts' annotations in five of six comparisons – $\rho_{E_1,S} = .88$, $\rho_{E_1,C} = .88$, $\rho_{E_2,S} = .86$, $\rho_{E_2,C} = .83$, $\rho_{E_3,S} = .87$, $\rho_{E_3,C} = .87$ (S and C refer to Scale AI and Centaur Labs, respectively) – than experts' annotations are correlated with each other ($\rho_{E_1,E_2} = .84$, $\rho_{E_2,E_3} = .85$, and $\rho_{E_1,E_3} = .86$). We do not find statistically significant differences between the experts' correlation with each other and either expert's correlation with any of the crowdsourced methods.

In addition to examining the inter-rater reliability across methods, we examine how inter-rater reliability changes depending on the number of non-qualified annotations. Instead of assessing

| Method | $E_1$ ($\rho$) | $E_1$ p-value | $E_2$ ($\rho$) | $E_2$ p-value | $E_3$ ($\rho$) | $E_3$ p-value |
|---|---|---|---|---|---|---|
| $E_1$ | | | 0.84 | 0.73 | 0.86 | 0.66 |
| $E_2$ | 0.84 | 0.44 | | | 0.85 | 0.66 |
| $E_3$ | 0.86 | 0.44 | 0.85 | 0.73 | | |
| ITA-FST | 0.57 | <0.001 | 0.52 | <0.001 | 0.55 | <0.001 |
| Scale AI | 0.88 | 0.08 | 0.86 | 0.43 | 0.88 | 0.08 |
| Centaur Labs | 0.88 | 0.08 | 0.83 | 0.50 | 0.87 | 0.32 |

Table 1. Inter-rater reliability based on Fisher Z transformations of Pearson Correlation Coefficients ($\rho$). The $E_x$ ($\rho$) columns display the correlation between the method in the row and the method in the column. The p-value columns show the minimum p-value based on Algorithm 1 in the Appendix applied to all pairwise correlations of experts; as an example, the cell in the $E_1$ p-value column and ITA-FST row presents the minimum p-value comparing (a) $\rho_{E_1,ITA}$ and $\rho_{E_1,E_2}$, (b) $\rho_{E_1,ITA}$ and $\rho_{E_1,E_3}$, and (c) $\rho_{E_1,ITA}$ and $\rho_{E_2,E_3}$

.

FSTs based on a dynamic consensus procedure, we compare expert 1's annotations with the crowd mean of 25 random draws from the Scale AI annotations (which were non-qualified meaning that crowdworkers were not filtered by a task specific quality control procedure) in samples of the following sizes: 3, 6, 12, 24, 48, and 96 annotations. We find a logarithmic relationship between $\rho_{S,E_1}$ and sample size that plateaus with $\rho_{S,E_1}$ approaching 0.88; see Figure 2 for a visualization of this relationship. For example, an increase from 3 to 12 annotations per image is associated with a 10 percentage point increase in $\rho_{S,E_1}$; the mean $\rho_{S,E_1}$ is 0.74 with a standard deviation of 0.026 when evaluating across 3 annotations per image and 0.84 with a standard deviation of 0.01 when evaluating across 12 annotations per image. A further increase from 12 to 24 annotations per image is associated with another 2 percentage points increase in $\rho_{S,E_1}$. We also find a similar relationship when comparing the varying size of the crowd with expert 2 and 3.

### 4.2 Qualitative Assessment of Inter-Rater Reliability

We examine inter-rater reliability qualitatively by illustrating similarities and differences in annotations across methods and examining images flagged by failure reports. In Figure 3, we present **qualitative confusion matrices** that showcase how different annotation methods lead to different annotations. These qualitative confusion matrices are intended to contextualize and illustrate similarities and discrepancies in subjective annotations and build upon the finding that alternative representation of confusion matrices can improve non-expert understanding of performance [101].

Across the 320 textbook images, annotators flagged 17 images as inappropriate or incorrect. Three of these flagged images were originally marked by expert 1 as "Not Applicable." Unlike most images, all three of these images contain multiple photographs under multiple lighting conditions, expert 2 provided a different annotation than expert 1, and the Centaur Labs and Scale AI crowd labels are discordant. Another two of these flagged images are marked as confusing and neither the expert annotations or the crowd annotations agree with one another. The final 12 of the flagged images contain messages that the annotator is confident that the expert's label is wrong; in 5 of these 12 images, expert 2 and both crowd consensuses agree that expert 1's annotation is one unit off, in 6 of these 12 images, expert 2 agrees with expert 1 while both crowd consensuses disagree with the experts, and in 1 of these 12 images, there is disagreement across experts and crowd consensuses. These results suggest failure reports are generally useful in identifying images that are likely to be problematic and extremely subjective for one reason or another.

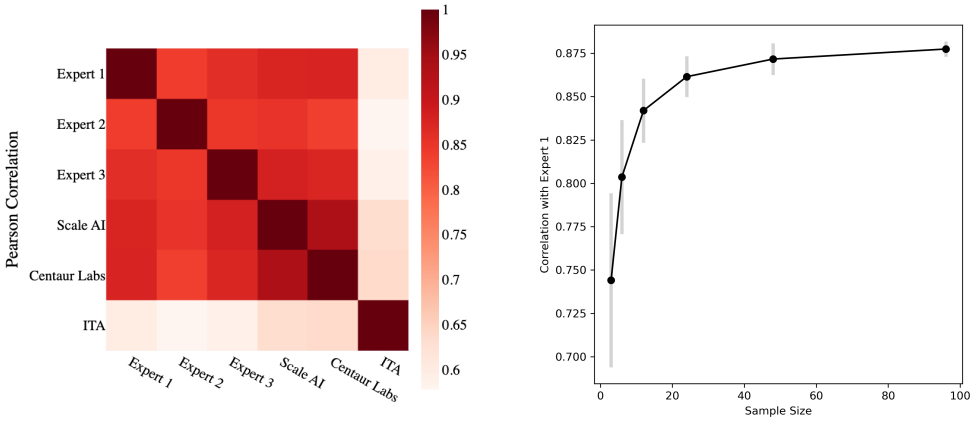


Fig. 2. **Left**: Heatmaps showing inter-rater reliability as measured by Pearson's Correlation Coefficient. These heatmaps include 296 images and exclude the 24 images rated by any expert or crowdsourcing method as "Not Applicable." **Right**: Inter-rater reliability by crowd size based on 25 random bootstrapped samples from the Scale AI annotations. The y-axis presents the correlation between expert 1's annotations and the crowd's mean FST annotation. The x-axis presents the number of annotations per image. The gray bars represent the 95% confidence interval. As the number of annotations increases the confidence interval decreases and the Pearson Correlation Coefficient ($\rho$) approaches 0.88.

### 4.3 Scaling Annotations on the Fitzpatrick 17k

For resource constrained developers of large-scale image datasets, it is orders of magnitude less resource intensive to annotate images with an algorithm or crowdsourcing than with board-certified dermatologists [88, 97, 104]. Given the lower inter-rater reliability of the ITA-FST algorithm, we limit our analysis of scaling annotations on the full Fitzpatrick 17k dataset [44] to crowdsourcing methods. 85% of consensus FST annotations by Scale AI and Centaur Labs are within one unit of each other. In Figure 4, we present a confusion matrix, which reveals that large discrepancies in annotations between sources are rare. An expert review of all applicable annotation discrepancies that are off by more than one unit would involve examining 9% (1,365 of the 13,865 images) of the Fitzpatrick 17k dataset. Error reports by annotators from Centaur Labs indicate that the consensus annotation for 166 images are incorrectly labeled and 21 images are inappropriate or irrelevant for the task.

### 4.4 Expert Review of Scaled Annotations

As a final step in evaluating dynamic consensus protocols, we collect labels from 3 board-certified dermatologists on 140 images randomly selected from the 16,577 images in the Fitzpatrick 17k dataset. We stratified this random selection on two features: (1) Scale AI's estimated FST annotations and (2) a binary variable for discrepancy between Scale AI's and Centaur Labs' annotations of more than 1 estimated FST annotation. As a result, there are 20 images with each Scale AI estimated FST type and 20 images annotated by Scale AI as not applicable. In addition, 70 of these images have been annotated by Scale AI and Centaur Labs within 1 estimated FST of each other and the other 70 images have been annotated with estimated FST that differ by more than 1.

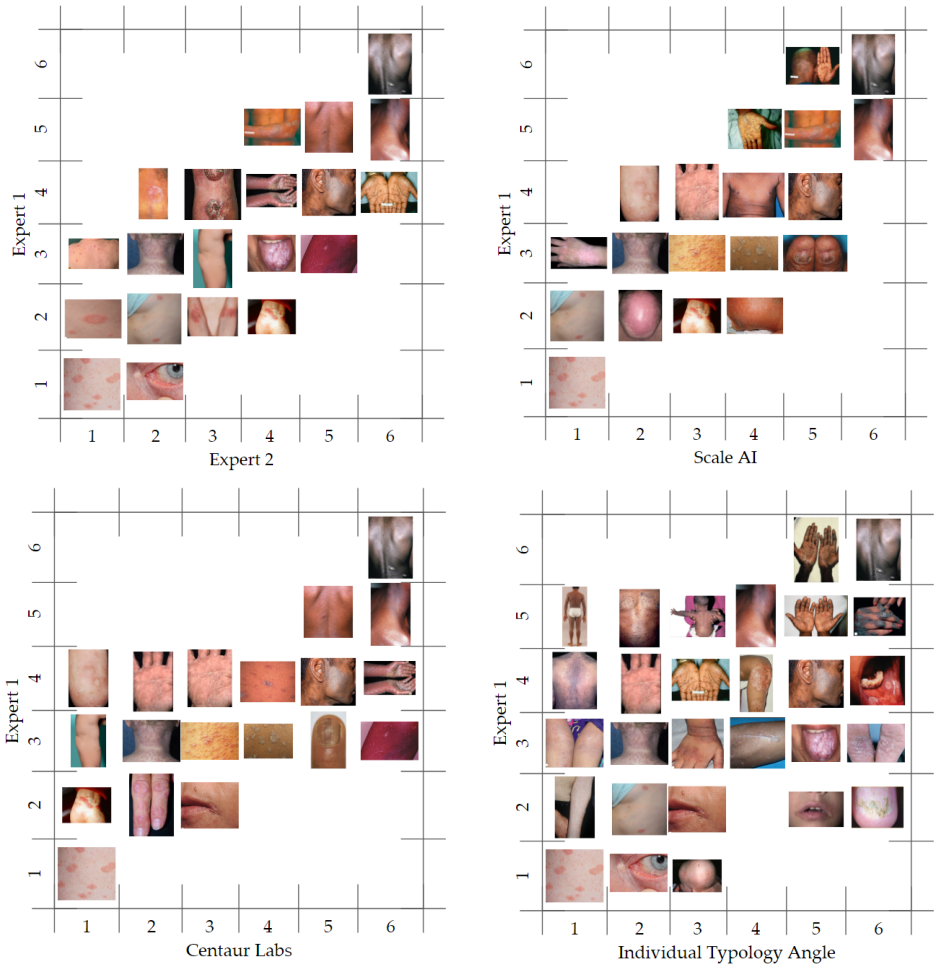


Fig. 3. Textbook images [12, 41, 46, 54, 57, 58, 78, 110] of skin conditions plotted according to Expert 1's annotations (on the Y-axis) and 4 other methods (Expert 2, ITA-FST algorithm Scale AI, and Centaur Labs on the X-axis).

For the 70 images with similar annotations, the correlation between experts ranges from 83% to 87% and the crowds correlation with experts ranges from 86% to 89%. We do not see any statistically significant difference between experts and crowds.

However, for the 70 images with greater than 1 unit discrepancies across the two crowd methods, we do find significant differences between inter-annotator reliability across experts and the crowd. The correlation between experts ranges from 59% to 66% and the crowds' correlation with experts ranges from 32% to 63%. We examine the inter-rater reliability between Scale AI and Centaur Labs and experts by conducting 12 tests of statistical significance to cover all possible comparison permutations. We find that 3 of 6 comparisons of inter-rater reliability between Scale AI and experts show Scale AI's annotations are less correlated and the p-value is less than the standard 5% threshold for statistical significance. Likewise, we find that 1 of 6 comparisons of inter-rater reliability between Centaur Labs and experts show Centaur Labs' annotations are less correlated and

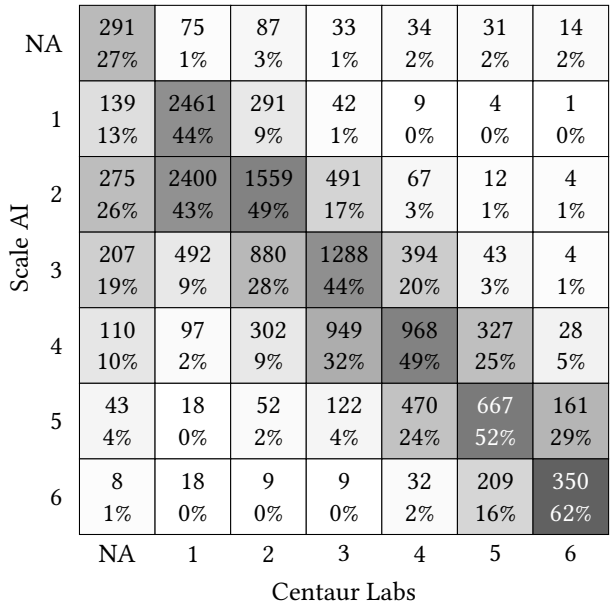


Fig. 4. Confusion matrix comparing two crowdsourcing methods for annotating the 16,577 images in the Fitzpatrick 17k dataset

the p-value is less than the standard 5% threshold for statistical significance. In Table 2, we present the inter-rater reliability Pearson correlation coefficients and lowest p-values for tests of statistical significance. This table also includes an examination of estimated ITA-FST on these 70 images, and we find estimated the correlation between ITA-FST and experts' annotations approaches 0 for this selection of images for expert review.

| Method | $E_1$ ($\rho$) | $E_1$ p-value | $E_2$ ($\rho$) | $E_2$ p-value | $E_3$ ($\rho$) | $E_3$ p-value |
|---|---|---|---|---|---|---|
| $E_1$ | | | 0.59 | 0.29 | 0.66 | 0.29 |
| $E_2$ | 0.59 | 0.29 | | | 0.66 | 0.58 |
| $E_3$ | 0.66 | 0.29 | 0.66 | 0.58 | | |
| ITA-FST | 0.05 | <0.001 | -0.06 | <0.001 | 0.08 | <0.001 |
| Scale AI | 0.50 | 0.04 | 0.32 | <0.001 | 0.57 | 0.19 |
| Centaur Labs | 0.63 | 0.54 | 0.47 | 0.02 | 0.53 | 0.09 |

Table 2. Analysis of subset of 70 images with high disagreement showing the inter-rater reliability based on Fisher Z transformations of Pearson Correlation Coefficients ($\rho$). The $E_x$ ($\rho$) columns display the correlation between the method in the row and the method in the column. The p-value columns show the minimum p-value based on Algorithm 1 in the Appendix applied to all pairwise correlations of experts; as an example, the cell in the $E_1$ p-value column and ITA-FST row presents the minimum p-value comparing (a) $\rho_{E_1,ITA}$ and $\rho_{E_1,E_2}$, (b) $\rho_{E_1,ITA}$ and $\rho_{E_1,E_3}$, and (c) $\rho_{E_1,ITA}$ and $\rho_{E_2,E_3}$

.

The comparison of inter-annotator agreement on the images selected for expert review reveals important nuances that researchers should keep in mind when annotating future datasets. While the inter-rater reliability on estimated FST is just as high between experts as it is between experts and the crowd consensus for most images, the annotations by crowds on images with low agreement may be less reliable than experts' annotations. By incorporating **expert review** into a subset of crowd annotations with low agreement, a dynamic consensus protocol can adjudicate edge cases such that adjudication leads to a higher likelihood of agreement with other experts.

This particular expert review of 140 images highlights edge cases where experts tend to agree with each other more often than they agree with dynamic consensus labels from crowdworkers. However, it is important to note that inter-rater reliability across experts on the 70 images randomly drawn from the 9% of images with two discordant crowd ratings ranges from 59% to 66% whereas inter-rater reliablity on the other 70 images (randomly drawn from the 91% of images with two concordant crowd ratings) ranges from 83% to 87%. On these 70 images with discordant crowd annotations, experts agree with each other significantly less than they do on the overwhelming majority of images. This lower rate of expert agreement and significantly lower rate of crowd worker and expert agreement demonstrates the subjectivity of estimating FST of an individual in an image can vary considerably across images.

## 5 DISCUSSION

How well does the ITA-FST algorithm and various crowdsourcing methods compare to board-certified dermatologists in annotating images with estimated FSTs? Our results reveal that the inter-rater reliability between three board-certified dermatologists (as measured by $\rho_{E_x,E_y}$) is comparable to the inter-rater reliability between each board-certified dermatologist and each of the two crowdsourcing methods (as measured by $\rho_{E,Crowd}$). However, inter-rater reliability of the ITA-FST algorithm (as measured by $\rho_{E,ITA}$) is significantly lower than the inter-rater reliability between any two experts.

Estimated FST annotations on images are highly subjective. We find that three experts agree with each other exactly on estimated FST in only 50-55% of images (although they agree with each other within a one unit difference in 92-94% of images). Rather than treat this subjectivity as a bias, we treat subjectivity on a per annotation basis as a measure of signal and noise. We find the differences between the three experts' annotations are not significantly larger than the differences between the experts and either crowdsourced method. In other words, expert annotations generally have the same amount of signal and noise as crowd annotations. This general finding comes with a caveat: there are identifiable edge cases where experts' annotations demonstrated significantly higher inter-rater reliability than crowdsourced annotations. Nonetheless, our results suggest that crowdsourcing methods (but not the ITA-FST algorithm) can be reliable for annotating large scale dermatology image datasets with skin type annotations especially when expert review is included.

This is particularly important for increasing transparency in machine learning for dermatology because skin type annotations are one of the items on the CLEAR Derm checklist for the evaluation of image-based AI algorithms [21] and an important consideration for evaluating medical AI devices for FDA approvals [111]. Transparency on skin tone information can be useful for evaluating both the distribution (and potential under-representation) of various skin tones in image datasets and how AI algorithms in dermatology perform across different skin tones, which is then useful as evidence for holding the fields of computer vision and dermatology accountable for addressing the unwanted biases.

While crowdsourced annotations are comparable with experts' annotations in aggregate, there are many examples where experts agree with each other yet the crowd differs. One approach for reducing crowdsourcing disagreement with experts is to include more annotations per image, which

we find is effective for reducing errors from crowd sizes of 3 to 12 but less effective for reducing errors from larger crowd sizes. A second approach is to integrate expert review into crowdsourcing. In particular, expert review examines edge cases that are flagged based on failure reports, agreement metrics (e.g. low agreement scores, high difficulty scores), and random samples for review. For the overwhelming majority of images, experts and the crowd have similar inter-rater reliability, but for the edge cases, expert review can offer additional reliability because inter-rater reliability of experts on edge cases can be higher than inter-rater reliability of crowds on edge cases.

The comparison between methods for annotating subjective labels provides a replicable methodology for answering when an algorithm or crowdsourced methodology can reliably be used in lieu of experts for annotating data. The goal of this kind of data annotation is to increase transparency in dataset biases to motivate greater accountability in sociotechnical decision-making systems. However, this kind of transparency comes at a cost. Human labor by experts or crowd annotators requires time and energy and should be compensated appropriately whereas the resources needed to compute ITA-FST scores are neglible. The low agreement between ITA-FST and experts is best to avoid because it may leave analyses prone to data cascade errors [98]. On the other hand, the relatively high agreement between experts and the crowd (and the opportunity to augment crowd labels with expert review) makes crowd annotations of estimated FST on images more attractive than expensive experts. We note that the crowd labels here come from Scale AI and Centaur labs, which represent very different ecosystems than the decentralized requester marketplace of Amazon Mechanical Turk (AMT) [73]. In particular, Scale AI and Centaur Labs work directly with individuals rather than through AMT, and as such, both these services avoid the "root problem... of unfair requesters" [49] in the AMT marketplace and the problem of turkers' uncertainty about the fairness of a particular requester [53]. Moreover, the ability to submit error reports with Centaur Labs creates a tractable opportunity for communication between crowdworkers analyzing the images and data scientists analyzing the data.

## 6 LIMITATIONS

We focused our comparisons on how three experts, one algorithm, and two crowdsourcing methods retrospectively annotate estimated FST across 320 images collected from dermatology textbooks and 140 images collected from online dermatology atlases. The 320 images showcase skin of all six skin types, but the distribution of skin types is not uniform across these images because dark skin types are underrepresented in dermatology textbooks [2, 5, 28]. Based on experts' annotations, only 18-26% of the 320 images show the two darkest skin types.

In our evaluation, we consider the ITA-FST algorithm applied to images with YCbCr masks, and we find it exhibits higher variability than expert and crowd-based annotations. While the ITA-FST algorithm may not be a reliable method for annotating estimated FST, future algorithms applied to images (especially segmented portions of images) may be able to match the inter-rater reliability of experts.

The lighting conditions are heterogeneous across these images, which makes assessing estimated FST more difficult than it would be in images with a single, consistent cross-polarized light source. Guidelines for photographing images of skin conditions on dark skin suggest images use indirect, natural light and a separate light for the hair and should avoid backgrounds with bright colors or patterns [64]. A recent perspective piece in the British Journal of Dermatology presents a series of images where the only difference is lighting source (cross-polarized light vs. white light) that reveals cross-polarized light reduces specular reflections and increases the contrast between healthy and unhealthy skin [86].

The variability of estimated FST annotations in images is much higher than in-person assessments because in-person assessments are not limited by lighting sources and enable a dermatologist to

include an assessment of the patients' skin color, eye color, hair color, and history of sunburns. We leave the comparison of in-person FST annotation to image-based estimated FST annotations to future research.

The Fitzpatrick scale is a starting point but not a perfect method for annotating skin color [13, 107]. The Fitzpatrick classification system was originally designed for classifying skin based on skin's reaction to the sun (burning vs tanning) and not skin color [34]. Moreover, the original Fitzpatrick classification labeled FST I-IV as white, FST V as brown, and FST VI as black, which contrasts with how researchers describe today's usage of FST as pale-white for I, white for II, beige for III, brown for IV, dark brown for V, and black for VI [96]. We re-created the original scale in Table 3 in the Appendix for quick reference. Annotating images with estimated FSTs helps to document the diversity of dermatology datasets and inspect algorithms for discrimination based on the color of one's skin, but estimated FSTs serve as a blunt proxy (biased towards lighter skin colors) that fail to capture the global diversity of skin colors [107]. In order to avoid singularly optimizing future AI algorithms on a biased proxy [81], future research and data collection should consider additional methods and metrics for annotating the diversity and complexity of skin color including factors such as self-reported versus observer reported skin tone [80], in-person or image based assessment, and the number of response categories [92].

## 7 CONCLUSION

By annotating large datasets of dermatology images with FSTs, researchers can increase transparency and enable relatively straight-forward evaluations of algorithmic performance across skin types for AI models trained to classify skin conditions. While image-based FST annotations are subjective, we find the annotations of experts and crowds are highly comparable while the annotations produced by the ITA-FST algorithm are more variable. In light of the higher variability of annotations generated by the ITA-FST algorithm, we recommend that researchers do not augment their datasets of clinical dermatology images algorithmically and instead use a crowdsourcing or expert-based approach.

We find some instances where the experts concur yet the crowd consensus disagrees. We recommend the most efficient and thorough approach to annotating images of skin conditions with FSTs is to combine experts and the crowd. Expert review can adjudicate both images flagged for error reports and images with low agreement or high difficulty scores. While we propose this approach for annotation of FSTs, our recommendation for hybrid dynamic consensus protocols with experts and crowds may extend to other domains in which annotations are similarly subjective for experts and non-experts alike.

## DATA AND CODE AVAILABILITY

The datasets and code generated and analyzed during the current study are available in our public Github repository, https://github.com/mattgroh/fitzpatrick17k.

## ACKNOWLEDGMENTS

We thank Erik Duhaime and Kira Prentice at Centaur Labs and Aerin Kim at Scale AI for providing data annotation services for free, the many crowdworkers for annotating the images, Bruke Wossenseged for valuable research assistance, and Rosalind Picard, Ziv Epstein, and Luis Soenksen for thoughtful feedback.

## REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM,

Barcelona Spain, 252–260. https://doi.org/10.1145/3351095.3372871
[2] Ademide Adelekun, Ginikanwa Onyekaba, and Jules B. Lipoff. 2020. Skin color in dermatology textbooks: An updated evaluation and analysis. *Journal of the American Academy of Dermatology* (April 2020), S0190962220307003. https://doi.org/10.1016/j.jaad.2020.04.084
[3] Jehad Amin AlKattash. [n. d.]. DermaAmin. *https://www.dermaamin.com/site/* ([n. d.]).
[4] Ali Alkhatib and Michael Bernstein. 2019. Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
[5] Savannah M. Alvarado and Hao Feng. 2020. Representation of dark skin images of common dermatologic conditions in educational resources: a cross-sectional analysis. *Journal of the American Academy of Dermatology* (June 2020), S0190962220311385. https://doi.org/10.1016/j.jaad.2020.06.041
[6] Paul André, Aniket Kittur, and Steven P Dow. 2014. Crowd synthesis: Extracting categories and clusters from complex data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 989–998.
[7] Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36, 1 (2015), 15–24.
[8] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 167–176.
[9] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *SSRN Electronic Journal* (2016). https://doi.org/10.2139/ssrn.2477899
[10] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[11] Ruha Benjamin. 2019. Assessing risk, automating racism. *Science* 366, 6464 (2019), 421–422.
[12] Jean L Bolognia, Julie V Schaffer, and Lorenzo Cerroni. 2018. *Dermatología*. Elsevier Health Sciences.
[13] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.
[14] Ángel Alexander Cabrera, Abraham J. Druck, Jason I. Hong, and Adam Perer. 2021. Discovering and Validating AI Errors With Crowdsourced Failure Reports. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–22. https://doi.org/10.1145/3479569
[15] Carrie J. Cai, Martin C. Stumpe, Michael Terry, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, and Greg S. Corrado. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300234
[16] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–24. https://doi.org/10.1145/3359206
[17] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–7. https://doi.org/10.1145/3411763.3443435
[18] Alain Chardon, Isabelle Cretois, and Colette Hourseau. 1991. Skin colour typology and suntanning pathways. *International journal of cosmetic science* 13, 4 (1991), 191–208.
[19] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *Advances in neural information processing systems* 31 (2018).
[20] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
[21] Roxana Daneshjou, Catarina Barata, Brigid Betz-Stablein, M. Emre Celebi, Noel Codella, Marc Combalia, Pascale Guitera, David Gutman, Allan Halpern, Brian Helba, Harald Kittler, Kivanc Kose, Konstantinos Liopyris, Josep Malvehy, Han Seung Seog, H. Peter Soyer, Eric R. Tkaczyk, Philipp Tschandl, and Veronica Rotemberg. 2021. 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 Dermatology* (Dec. 2021). https://doi.org/10.1001/jamadermatol.2021.4915
[22] Roxana Daneshjou, Mary Smith, Mary Sun, Veronica Rotemberg, and James Zou. 2021. Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. (2021), 8.
[23] Roxana Daneshjou, Kailas Vodrahalli, Weixin Liang, Roberto A Novoa, Melissa Jenkins, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. 2022. Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set. *arXiv preprint arXiv:2203.08807* (2022).

[24] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[25] Sandra Del Bino and FJBJoD Bernerd. 2013. Variations in skin colour and the biological consequences of ultraviolet radiation exposure. *British Journal of Dermatology* 169 (2013), 33–40.

[26] Nicolás Della Penna and Mark D. Reid. 2012. Crowd & Prejudice: An Impossibility Theorem for Crowd Labelling without a Gold Standard. *arXiv:1204.3511 [cs]* (April 2012). http://arxiv.org/abs/1204.3511 arXiv: 1204.3511.

[27] Brittany Dulmage, Kyle Tegtmeyer, Michael Z. Zhang, Maria Colavincenzo, and Shuai Xu. 2020. A Point-of-Care, Real-Time Artificial Intelligence System to Support Clinician Diagnosis of a Wide Range of Skin Diseases. *Journal of Investigative Dermatology* (Oct. 2020), S0022202X20321679. https://doi.org/10.1016/j.jid.2020.08.027

[28] Tobechi Ebede and Art Papier. 2006. Disparities in dermatology educational resources. *Journal of the American Academy of Dermatology* 55, 4 (Oct. 2006), 687–690. https://doi.org/10.1016/j.jaad.2005.10.068

[29] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[30] Steven Eilers, Daniel Q Bach, Rikki Gaber, Hanz Blatt, Yanina Guevara, Katie Nitsche, Roopal V Kundu, and June K Robinson. 2013. Accuracy of self-report in assessing Fitzpatrick skin phototypes I through VI. *JAMA dermatology* 149, 11 (2013), 1289–1294.

[31] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (Feb. 2017), 115–118. https://doi.org/10.1038/nature21056

[32] Ronald A Fisher. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10, 4 (1915), 507–521.

[33] Ronald A Fisher. 1921. On the'probable error'of a coefficient of correlation deduced from a small sample. *Metron* 1 (1921), 1–32.

[34] Thomas B Fitzpatrick. 1988. The validity and practicality of sun-reactive skin types I through VI. *Archives of dermatology* 124, 6 (1988), 869–871.

[35] Samuel Freire da Silva. [n. d.]. Atlas Dermatologico. *http://atlasdermatologico.com.br/* ([n. d.]).

[36] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J. Berkowitz, Eva Lermer, Joseph F. Coughlin, John V. Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine* 4, 1 (Dec. 2021), 31. https://doi.org/10.1038/s41746-021-00385-9

[37] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[38] Charles Goodwin. 2000. Practices of color classification. *Mind, culture, and activity* 7, 1-2 (2000), 19–36.

[39] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[40] Ben Green and Yiling Chen. 2021. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.

[41] Christopher Griffiths, Jonathan Barker, Tanya O Bleiker, Robert Chalmers, and Daniel Creamer. 2016. *Rook's textbook of dermatology*. John Wiley & Sons.

[42] Matthew Groh. 2022. Identifying the Context Shift between Test Benchmarks and Production Data. https://doi.org/10.48550/ARXIV.2207.01059

[43] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences* 119, 1 (2022).

[44] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. 2021. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1820–1828.

[45] Vishal Gupta and Vinod Kumar Sharma. 2019. Skin typing: Fitzpatrick grading and others. *Clinics in dermatology* 37, 5 (2019), 430–436.

[46] Thomas Habif. 2010. Clinical dermatology: A color guide to diagnosis and therapy. (2010).

[47] Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S Greene, et al. 2020. Transparency and reproducibility in artificial intelligence. *Nature* 586, 7829 (2020), E14–E16.

[48] Seung Seog Han, Ilwoo Park, Sung Eun Chang, Woohyung Lim, Myoung Shin Kim, Gyeong Hun Park, Je Byeong Chae, Chang Hun Huh, and Jung-Im Na. 2020. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *Journal of Investigative Dermatology* 140, 9 (2020), 1753–1761.

[49] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.

[50] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. [n. d.]. Towards measuring fairness in AI: the Casual Conversations dataset. ([n. d.]), 9.

[51] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677* (2018).

[52] Harold Hotelling. 1953. New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological)* 15, 2 (1953), 193–232.

[53] Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 611–620.

[54] Diane Jackson-Richards and Amit G Pandya. 2014. *Dermatology atlas for skin of color*. Springer.

[55] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry* 11, 1 (2021), 1–9.

[56] Ayush Jain, David Way, Vishakha Gupta, Yi Gao, Guilherme de Oliveira Marinho, Jay Hartford, Rory Sayres, Kimberly Kanada, Clara Eng, Kunal Nagpal, Karen B. DeSalvo, Greg S. Corrado, Lily Peng, Dale R. Webster, R. Carter Dunn, David Coz, Susan J. Huang, Yun Liu, Peggy Bui, and Yuan Liu. 2021. Development and Assessment of an Artificial Intelligence–Based Tool for Skin Condition Diagnosis by Primary Care Physicians and Nurse Practitioners in Teledermatology Practices. *JAMA Network Open* 4, 4 (April 2021), e217249. https://doi.org/10.1001/jamanetworkopen.2021.7249

[57] Ajay Kailas. 2017. Taylor and Kelly's dermatology for skin of color. *Journal of the American Academy of Dermatology* 76, 2 (2017), e75.

[58] Sewon Kang. 2019. *Fitzpatrick's Dermatology, 2-Volume Set (Fitzpatricks*.

[59] Arshad T Khalid, Charity G Moore, Christopher Hall, Flora Olabopo, Nigel L Rozario, Michael F Holick, Susan L Greenspan, and Kumaravel Rajakumar. 2017. Utility of sun-reactive skin typing and melanin index for discerning vitamin D deficiency. *Pediatric research* 82, 3 (2017), 444–451.

[60] Newton M. Kinyanjui, Timothy Odonga, Celia Cintas, Noel C. F. Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R. Varshney. 2019. Estimating Skin Tone and Effects on Classification Performance in Dermatology Datasets. *arXiv:1910.13268 [cs, stat]* (Oct. 2019). http://arxiv.org/abs/1910.13268 arXiv: 1910.13268.

[61] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. 2020. Algorithms as discrimination detectors. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30096–30100.

[62] S. Kolkur, D. Kalbande, P. Shimpi, C. Bapat, and J. Jatakia. 2017. Human Skin Detection Using RGB, HSV and YCbCr Color Models. In *Proceedings of the International Conference on Communication and Signal Processing 2016 (ICCASP 2016)*. Atlantis Press, Lonere, India. https://doi.org/10.2991/iccasp-16.2017.51

[63] KS Krishnapriya, Gabriella Pangelinan, Michael C King, and Kevin W Bowyer. 2022. Analysis of Manual and Automated Skin Tone Assignments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 429–438.

[64] JC Lester, L Clark Jr, E Linos, and R Daneshjou. 2021. Clinical Photography in Skin of Color: Tips and Best Practices. *The British Journal of Dermatology* (2021).

[65] J.C. Lester, J.L. Jia, L. Zhang, G.A. Okoye, and E. Linos. 2020. Absence of images of skin of colour in publications of COVID-19 skin manifestations. *British Journal of Dermatology* 183, 3 (Sept. 2020), 593–595. https://doi.org/10.1111/bjd.19258

[66] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–45. https://doi.org/10.1145/3479552 arXiv: 2101.05303.

[67] Yuan Liu, Ayush Jain, Clara Eng, David H. Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, Vishakha Gupta, Nalini Singh, Vivek Natarajan, Rainer Hofmann-Wellenhof, Greg S. Corrado, Lily H. Peng, Dale R. Webster, Dennis Ai, Susan J. Huang, Yun Liu, R. Carter Dunn, and David Coz. 2020. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine* 26, 6 (June 2020), 900–908. https://doi.org/10.1038/s41591-020-0842-3

[68] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[69] Patricia Louie and Rima Wilkes. 2018. Representations of race and skin tone in medical textbook imagery. *Social Science & Medicine* 202 (April 2018), 38–42. https://doi.org/10.1016/j.socscimed.2018.02.023

[70] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2018. Explainable machine-learning predictions for the

prevention of hypoxaemia during surgery. *Nature biomedical engineering* 2, 10 (2018), 749–760.
[71] Travis Mandel, Jahnu Best, Randall H Tanaka, Hiram Temple, Chansen Haili, Sebastian J Carter, Kayla Schlechtinger, and Roy Szeto. 2020. Using the crowd to prevent harmful AI behavior. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.
[72] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How data scientistswork together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (2019), 1–23.
[73] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 224–235.
[74] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shravya Shetty. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (Jan. 2020), 89–94. https://doi.org/10.1038/s41586-019-1799-6
[75] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–14.
[76] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.
[77] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting computer vision datasets: an invitation to reflexive data practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 161–172.
[78] Robert G Micheletti, William D James, Dirk Elston, and Patrick J McMahon. 2022. *Andrews' Diseases of the Skin Clinical Atlas, E-Book*. Elsevier Health Sciences.
[79] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.
[80] Ellis P Monk Jr. 2015. The cost of color: Skin color, discrimination, and health among African-Americans. *Amer. J. Sociology* 121, 2 (2015), 396–444.
[81] Sendhil Mullainathan and Ziad Obermeyer. 2021. On the Inequity of Predicting A While Hoping for B. In *AEA Papers and Proceedings*, Vol. 111. 37–42.
[82] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.
[83] Michael Muller, Christine T Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, et al. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
[84] Laura Nader. 1972. Up the anthropologist: Perspectives gained from studying up. (1972).
[85] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (Oct. 2019), 447–453. https://doi.org/10.1126/science.aax2342
[86] Yuna Oh, A Markova, SJ Noor, and Veronica Rotemberg. 2021. Standardized clinical photography considerations in patients across skin tones. *British Journal of Dermatology* (2021).
[87] UK Okoji, SC Taylor, and JB Lipoff. 2021. Equity in skin typing: why it's time to replace the Fitzpatrick scale. *The British Journal of Dermatology* (2021).
[88] Silas Ørting, Andrew Doyle, Arno van Hilten, Matthias Hirth, Oana Inel, Christopher R Madan, Panagiotis Mavridis, Helen Spiers, and Veronika Cheplygina. 2019. A survey of crowdsourcing in medical image analysis. *arXiv preprint arXiv:1902.09159* (2019).
[89] Muhammad Osto, Iltefat H Hamzavi, Henry W Lim, and Indermeet Kohli. 2022. Individual Typology Angle and Fitzpatrick Skin Phototypes are Not Equivalent in Photodermatology. *Photochemistry and photobiology* 98, 1 (2022), 127–129.
[90] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. 2019. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine* 2, 1 (2019), 1–10.
[91] Michael Phillips, Helen Marsden, Wayne Jaffe, Rubeta N Matin, Gorav N Wali, Jack Greenhalgh, Emily McGrath, Rob James, Evmorfia Ladoyanni, Anthony Bewley, et al. 2019. Assessment of accuracy of an artificial intelligence

algorithm to detect melanoma in images of skin lesions. *JAMA network open* 2, 10 (2019), e1913436–e1913436.
[92] Carolyn C Preston and Andrew M Colman. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica* 104, 1 (2000), 1–15.
[93] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Robert Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. [n. d.]. Direct Uncertainty Prediction for Medical Second Opinions. ([n. d.]), 10.
[94] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).
[95] Michael Allen Ribers and Hannes Ullrich. 2020. Machine Predictions and Human Decisions with Variation in Payoffs and Skills. *SSRN Electronic Journal* (2020). https://doi.org/10.2139/ssrn.3726018
[96] Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, et al. 2022. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *Medical Image Analysis* 75 (2022), 102274.
[97] Muskaan Sachdeva, Kyla N Price, Jennifer L Hsiao, and Vivian Y Shi. 2020. Gender and rank salary trends among academic dermatologists. *International Journal of Women's Dermatology* 6, 4 (2020), 324.
[98] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
[99] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–37. https://doi.org/10.1145/3476058
[100] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–35. https://doi.org/10.1145/3392866
[101] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–22.
[102] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26, 8 (Aug. 2020), 1229–1234. https://doi.org/10.1038/s41591-020-0942-0
[103] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* 5, 1 (2018), 1–9.
[104] Joseph D Tucker, Stephen W Pan, Allison Mathews, Gabriella Stein, Barry Bayus, and Stuart Rennie. 2018. Ethical concerns of and risk mitigation strategies for crowdsourcing contests and innovation challenges: scoping review. *Journal of medical Internet research* 20, 3 (2018), e75.
[105] Monica Van Such, Robert Lohr, Thomas Beckman, and James M Naessens. 2017. Extent of diagnostic agreement among medical referrals. *Journal of evaluation in clinical practice* 23, 4 (2017), 870–874.
[106] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–39. https://doi.org/10.1145/3476068
[107] Olivia R Ware, Jessica E Dawson, Michi M Shinohara, and Susan C Taylor. 2020. Racial limitations of fitzpatrick skin type. *Cutis* 105, 2 (2020), 77–80.
[108] Davis Wen, Saad Khan, Antonio Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, Carlos de Blas Perez, Alastair Denniston, Xiaoxuan Liu, and Rubeta Matin. 2021. Characteristics of publicly available skin cancer image datasets: a systematic review. *Lancet Digital Health* (2021).
[109] Marcus Wilkes, Caradee Y Wright, Johan L du Plessis, and Anthony Reeder. 2015. Fitzpatrick skin type, individual typology angle, and melanin index in an African population: steps toward universally applicable skin photosensitivity assessments. *JAMA dermatology* 151, 8 (2015), 902–903.
[110] K Wolff, L Goldsmith, S Katz, B Gilchrest, A Paller, and D Lafell. 2008. Fitzpatricks Textbook of Dermatology in General Medicine.
[111] Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E Ho, and James Zou. 2021. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nature Medicine* 27, 4 (2021), 582–584.

## APPENDIX

**Algorithm 1** Fisher Z transformation for comparing independent correlations

**for** Each Expert **do**
  **for** Each Method **do**
    $z_{E_A,E_B} \leftarrow \frac{1}{2} \ln \frac{1+\rho_{E_A,E_B}}{1-\rho_{E_A,E_B}}$
    $z_{Expert,Method} \leftarrow \frac{1}{2} \ln \frac{1+\rho_{Expert,Method}}{1-\rho_{Expert,Method}}$
    $Z \leftarrow |\frac{z_{E_A,E_B} - z_{Expert,Method}}{\sqrt{2/(n-3)}}|$ where n= # of images
    Convert Z score to p-value
  **end for**
**end for**

**Algorithm 2** Individual typology angle threshold adjustment

$T_{12} \leftarrow Mean(ITA_1.Quantile(1), ITA_2.Quantile(3))$
$T_{23} \leftarrow Mean(ITA_2.Quantile(1), ITA_3.Quantile(3))$
$T_{34} \leftarrow Mean(ITA_3.Quantile(1), ITA_4.Quantile(3))$
$T_{45} \leftarrow Mean(ITA_4.Quantile(1), ITA_5.Quantile(3))$
$T_{56} \leftarrow Mean(ITA_5.Quantile(1), ITA_6.Quantile(3))$
$T \leftarrow \{T_{12}, T_{23}, T_{34}, T_{45}, T_{56}\}$
**for all** $t \in T$ **do**
  $Max\_Concordant \leftarrow Sum(Annotation\_E_1 = Annotation\_E_2 = ITA(t))$
  $I \leftarrow \{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$
  **for all** $i \in I$ **do**
    $t_i \leftarrow t + i$
    $Concordant \leftarrow Sum(Annotation\_E_1 = Annotation\_E_2 = ITA(t_i))$
    **if** $Concordant > Max\_Concordant$ **then**
      $Max\_Concordant \leftarrow Concordant$
      $t \leftarrow t_i$
    **end if**
  **end for**
**end for**

Received January 2022; revised April 2022; accepted August 2022

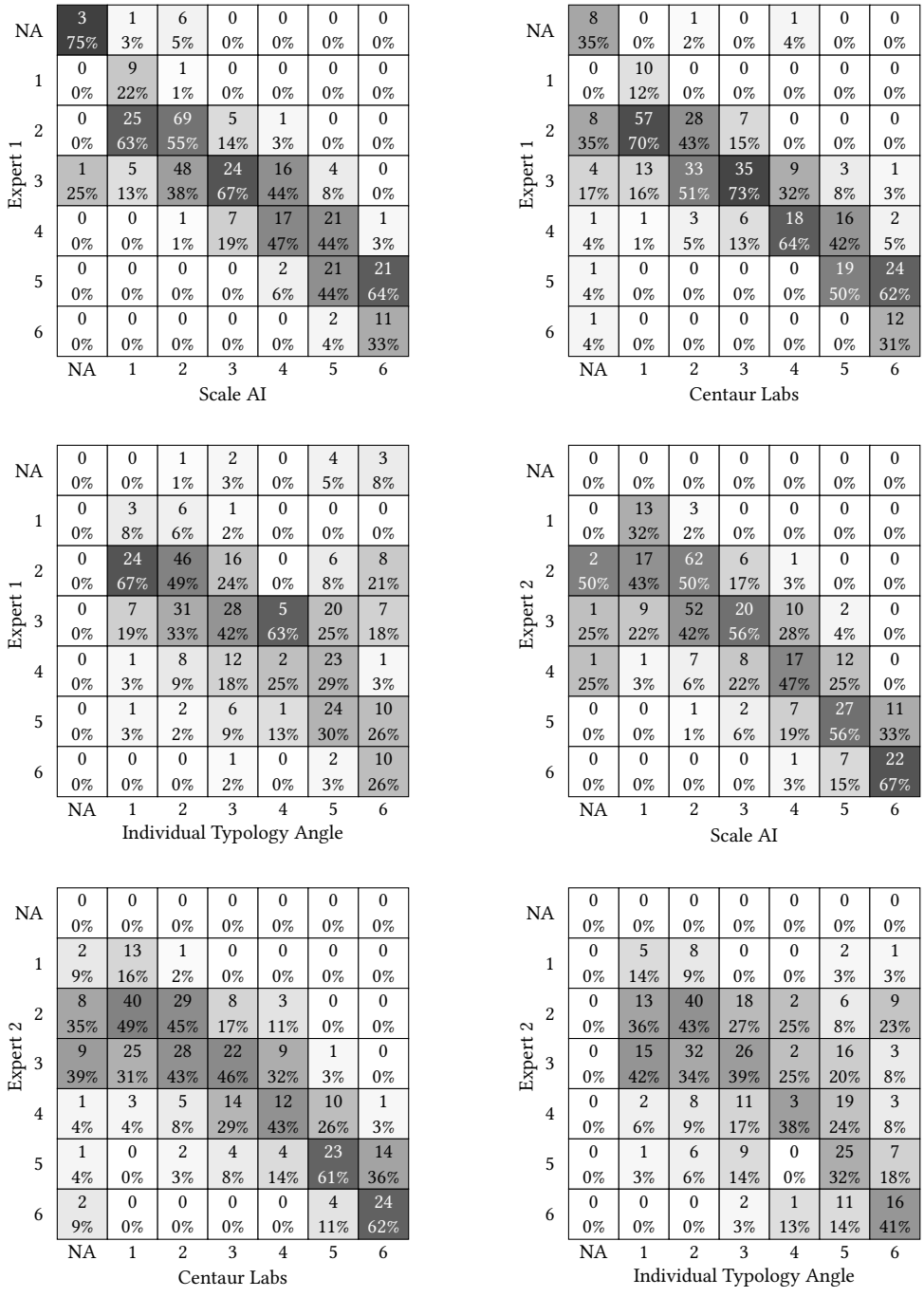


Fig. 5. Confusion matrices comparing experts 1 and 2 to the ITA algorithm and crowdsourcing methods.

| Skin Type | Skin Color | Sunburn | Tan |
|---|---|---|---|
| I | White | Yes | No |
| II | White | Yes | Minimal |
| III | White | Yes | Yes |
| IV | White | No | Yes |
| V | Brown | No | Yes |
| VI | Black | No | Yes |

Table 3. The Fitzpatrick skin type scale from Fitzpatrick et al 1988 [34]. The scale is intended for classifying sun-reactive skin types. Notably, the original scale does not include nuanced skin tones or color beyond white, brown, and black. In dermatology practice, the Fitzpatrick scale is commonly used to describe constitutive skin color [107]. Recent research published in the *Medical Image Analysis* describes the Fitzpatrick skin types as pale-white, white, beige, brown, dark brown, and black [96]. We informed crowd annotators by presenting example images of each FST.

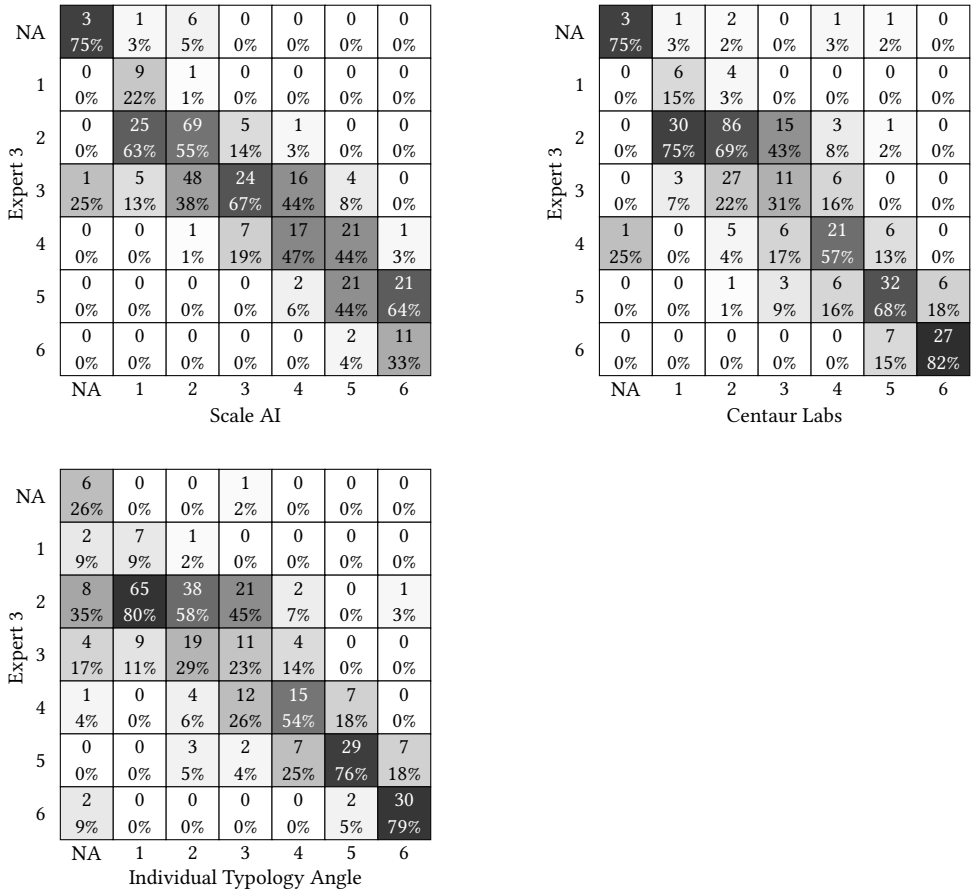


Fig. 6. Confusion matrices comparing expert 3 to the ITA algorithm and crowdsourcing methods.