Expert Discussions Improve Comprehension of Difficult Cases in Medical Image Assessment

Mike Schaekermann[1,2], Carrie J. Cai[1], Abigail E. Huang[1], Rory Sayres[1]
[1]Google, Mountain View, CA, [2]University of Waterloo, Canada
{mikeshake,cjcai,abigailhuang,sayres}@google.com

ABSTRACT

Medical data labeling workflows critically depend on accurate assessments from human experts. Yet human assessments can vary markedly, even among medical experts. Prior research has demonstrated benefits of labeler training on performance. Here we utilized two types of labeler training feedback: highlighting incorrect labels for difficult cases ("individual performance" feedback), and expert discussions from adjudication of these cases. We presented ten generalist eye care professionals with either individual performance alone, or individual performance and expert discussions from specialists. Compared to performance feedback alone, seeing expert discussions significantly improved generalists' understanding of the rationale behind the correct diagnosis while motivating changes in their own labeling approach; and also significantly improved average accuracy on one of four pathologies in a held-out test set. This work suggests that image adjudication may provide benefits beyond developing trusted consensus labels, and that exposure to specialist discussions can be an effective training intervention for medical diagnosis.

Author Keywords

Medical Images; Diagnosis; Adjudication; Labeler Training.

CCS Concepts

•Human-centered computing → Human computer interaction (HCI);

INTRODUCTION

In recent years, major advances in machine learning (ML) have enabled a new era of decision support tools (DST) for critical medical diagnostic tasks. With the increased capabilities of deep learning models, DSTs are being developed to support much more complex diagnostic processes with critical influence on patient outcomes. As these technologies mature, they hold the potential to increase access to healthcare—a demonstrated need for large sections of the developing world [4]. However, medical specialists sufficiently trained to perform complex diagnoses are exceptionally rare [4].

Alongside this growth in deep learning has been a parallel, increased need for large-scale, labeled medical data to power the training of such models. Because medical data is often highly regulated, and patient outcome is typically not available in the original data source, lack of access to ground truth-labeled data has become a key barrier to the development and evaluation of machine learning systems in medical domains [6]. As a result, contemporary ML-powered algorithms typically rely on medical practitioners to manually label data ground truth [6, 17, 42].

The collection of manually-labeled ground truth data from clinicians raises fundamental human challenges. Because many high-stakes medical decisions are highly nuanced and can be subject to personal opinion, even clinician assessments vary markedly. To combat this problem, current labeling approaches enlist a small set of specialized world experts to "adjudicate" the decision via consensus, as a way of producing a more reliable gold standard [21]. However, specialists of this caliber are exceptionally rare and expensive. To make the process more scalable, medical generalists with less training may be recruited to perform labeling at larger scale [43]. This is an enticing approach given that medical generalists far outnumber specialists (e.g. optometrists outnumber ophthalmologists 4.9-fold [1]). Yet, generalists are less experienced: difficult patient cases lead to high inter-labeler variability and incorrect diagnoses [21], limiting algorithmic validity and introducing the risk of adverse outcomes for patients' lives.

In this paper, we address these challenges by introducing and studying *adjudication feedback* for training medical image labelers. Our approach has the potential to help decrease the dependency on specialists by expanding the set of trained labelers to less-specialized workers, like generalists. The central underlying idea is to *reuse* existing metadata from *medical specialists'* adjudication of difficult cases to improve *medical generalists'* comprehension and labeling accuracy. Specifically, we study whether *discussion dialogs*, generated as a side product in the costly process of adjudication, can be repurposed as training material in medical data labeling workflows. We draw inspiration from prior research in crowdsourcing, which has demonstrated benefits of labeler training on the performance of non-experts on the web. Our research applies labeler training to the high-stakes, challenging domain of medicine, advancing our understanding of how to provide feedback to this emerging population of medical labelers.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04.
http://dx.doi.org/10.1145/3313831.3376290

In our controlled experiment, we examined the impact of two different forms of labeler training feedback: individual performance on difficult cases, and specialist discussions from adjudication of these cases. We presented ten certified eye care professionals with either individual performance alone, or individual performance and specialist discussions. Our results suggest that reading specialist discussions has benefits for generalists' comprehension of difficult cases, on their motivation to alter their own labeling approach, and on their diagnostic accuracy on a held-out test set. Our main contributions are:

1. We conducted an empirical study to understand the benefit of presenting adjudication discussions of difficult cases as a form of training feedback in medical data labeling.

2. We present results suggesting that showing adjudication discussions can improve comprehension of the rationale behind the correct diagnosis while motivating changes with respect to medical generalists' labeling approach.

3. We demonstrate that these benefits observed during training also translated into improved diagnostic accuracy in a held-out test set.

Taken together, this research advances our understanding of the emerging field of medical labeling, and provides new implications for how to scale medical data collection on high-stakes tasks with difficult-to-obtain ground truth.

RELATED WORK

The rise of deep learning for AI-assisted medical decision making has created a strong need for large amounts of labeled medical data (e.g., photographs, biomedical time series, x-rays, ultrasounds). In many cases, the ground truth labels required to develop supervised machine learning algorithms (e.g., correct diagnoses) are not given in the raw data. Such settings require the expertise from medical professionals for manual data labeling.

Inter-rater Disagreement in Medical Data Labeling

Like any form of human interpretation, medical data analysis by human experts is a subjective process and can lead to conflicting assessments among independent raters [2, 23, 31, 35]. The issue of inter-rater disagreement is particularly critical within medicine where unreliable clinical decisions can impact patients' lives adversely. Indeed, Raghu et al. [28] concluded that label disagreement poses a "full-fledged clinical problem in the healthcare domain."

Prior work in human computation for medical relation extraction [13] views inter-rater disagreement as a function of three phenomena: differences among human *raters* (e.g., training background, biases), characteristics of the *data* to be labeled (e.g., noisy or heterogeneous), and the quality of the labeling *guidelines* (e.g., subjective or ambiguous classification rules). Exacerbating the problem, human experts often rely on complex viewing technology to inspect medical data. Discrepancies in *viewer settings* (e.g., zoom or filter) [35] and sequential dependencies [39] were found to be additional sources of variablity for assessments in medical time series analysis.

Several works have suggested ways to make productive use of disagreement information in medical data labels (e.g., [2, 3, 8, 19, 28, 33, 36]). Inel et al. [19] introduced domain-independent quality measures for labelers, task instructions and data, based on disagreement information in a medical relation extraction task. Others developed models to predict the likelihood that a given patient case will cause expert disagreement in various medical subspecialties, including epilepsy diagnosis from electrophysiological signals [2], and eye disease diagnosis from retinal fundus photographs [28]. Recent work demonstrates positive effects of such ambiguity-aware models on expert workflows and perceived trust in medical data analysis [36]. Finally, Barnett et al. [3] evaluated different ways of computationally aggregating discordant medical assessments from labelers with varying training background to harness collective intelligence for medical diagnosis. In our work, we leverage the fact that conflicting expert assessments can motivate detailed adjudication discussions about difficult cases, and test whether such discussions can be repurposed to improve training for medical expert labelers at scale.

Resolving Disagreements through Panel Discussions

Reference standards used to evaluate machine learning models for medical applications typically need to meet particularly high quality standards. As a result, reference standards used for model *validation* purposes are often subject to greater scrutiny than labels used in larger-scale datasets for model *training*. A common way to produce these reference standards is to generate expensive gold standard consensus labels from panels of experts, rather than relying on noisy assessments from individual experts alone [16, 29, 30, 42].

A useful approach for resolving disagreements among committees of human labelers is group discussion. The approach of group discussion to improve data classification decisions has been studied both in human-computer interaction [7, 12] and in medicine [21, 34, 35, 38]. Various protocols have been described to facilitate and structure communication among discussion members, including in-person, face-to-face discussions [21] and remote, web-based discussions either in real-time [7, 37] or asynchronously [12, 35, 38].

While there has been long-standing debate about the relative benefits of collective decision making versus the so-called *wisdom of the crowd*, there is evidence suggesting that group discussions can indeed improve accuracy of decisions made both in general intelligence tasks [26] and in medical diagnosis [15].

Medical Diagnosis Training

Given that medical generalists far outnumber specialists in various fields [1], research into effective ways to calibrate medical generalists for difficult cases holds the potential to tap into a large pool of high-quality labelers.

A scalable approach for medical diagnosis training is through computer-based tutorials [18, 22]. Typically, web-based tutorials for medical training present a series of patient cases to the learner, and present case-specific feedback after the learner has submitted their answer. Our approach is similar in that we follow a simple paradigm of presenting feedback after a set of training cases. However, web-based tutorials for medical training typically focus on curation of content for case-specific feedback while selecting mostly clean-cut cases for which clear explanations exist. By contrast, our work emphasizes the use of difficult, contentious cases to test whether pre-existing adjudication discussions not originally intended for training generalists can be re-used for educational purposes.

Our work draws inspiration from the medical education literature about discussion-based learning. The idea that medical students may learn more effectively when engaging in group discussions with their peers has been implemented in the concepts of problem-based learning (PBL) and case-based learning (CBL) [11, 41]. Both approaches aim to improve upon lecture-based learning by fostering collective clinical reasoning through group discussions. CBL is a more structured and guided variant of discussion-based learning in medicine while PBL implements an open-ended approach. In this paper, we examine whether *passive* consumption of specialist discussion about difficult cases can yield similar benefits for diagnostic reasoning as has previously been reported about PBL and CBL, but applied to the context of medical labeling.

APPLICATION DOMAIN

Every year, eye disease causes vision impairments or blindness for millions of people worldwide. In particular, retinal pathologies such as diabetic retinopathy (DR) rank among the leading causes of vision loss in many industrialized countries [44]. To combat the issue, several national governments have established population-wide screening programs for early disease detection.

One of the central diagnostic artifacts in the assessment of retinal disease is fundus photography, i.e., photographs taken of the background of a patient's eye (Figure 1). Digital fundus photos are used both in tele-medical screening [40] and for the development of deep learning models for AI-assisted retinal assessment [17, 27]. Regardless of the setting, expertise from certified medical professionals is required to determine the presence and severity of disease as it appears in the image. While the diagnostic criteria for retinal assessment are governed by official medical guidelines, image interpretation by medical experts remains a subjective process [21]. The resulting inter-rater disagreement may not only arise over the presence of disease, but also over the specific classification of an observed pathology. In particular, the appearance of diabetic retinopathy may resemble other forms of retinal disease such as hypertensive retinopathy (HTNR), retinal vein occlusion (RVO) and retinal artery occlusion (RAO). It is crucial that treatment decisions are formed based on correct differential diagnoses to avoid adverse outcomes for patients.

Eye care professionals with varying levels of specialization are concerned with the assessment of retinal disease [1]: (1) *optometrists* present the largest group of professionals trained for retinal assessment; as generalists, they typically refer difficult-to-assess cases to other experts, such as (2) *general ophthalmologists*, i.e., medical doctors who completed a multi-year residency program in general eye and vision care; at the highest level of specialization, there is a small population of (3) *retina specialists* worldwide—ophthalmologists who completed a two-year fellowship program in retinal assessment after completing their eye care residency.

Our application domain is representative of other medical subspecialties. Not only does it require the subjective process of image interpretation by human experts; it also involves different types of easy-to-confuse pathologies (DR, HTNR, RVO, RAO), that require a deep understanding of symptomatic differences to be reliably differentiated.

RESEARCH QUESTIONS & HYPOTHESES

The case discussions used in our training study are the by-product of an adjudication process designed to analyze and resolve diagnostic disagreements among highly trained medical specialists. As such, the discussion dialogs are expected to reflect types of vocabulary and reasoning grounded in a deep understanding of a certain medical subspecialty. Yet, the case discussions were not collected with an educational purpose in mind. As a result, they may exhibit weaknesses when used for labeler training. For example, the fact that the dialogs are rooted in disagreements and the potential use of specialist jargon may cause confusion among less specialized medical professionals. Our study addresses two primary research questions about how medical generalists *perceive* (Q1) and *act upon* (Q2) the presentation of case-specific adjudication discussions from specialists as a form of medical diagnosis training.

Q1: How do medical generalists perceive reading of specialist discussions as a form of labeler training feedback?

Medical assessments can be contentious and it is possible for one expert to take the perspective of another expert without necessarily agreeing with their final conclusion. Furthermore, even if an expert understands *and* agrees with the diagnostic reasoning for one specific case, it is not guaranteed that this will also motivate a change in their own labeling approach for other cases.

In this study, we examine these three aspects—comprehension, agreement, adaptation—separately, and hypothesize that reading of specialist discussions as a form of training feedback for medical generalists will:

> **[H1a]** Improve **comprehension** of the rationale behind the correct diagnosis.
>
> **[H1b]** Increase **agreement** with the answer key.
>
> **[H1c]** Motivate **adaptations** in generalists' labeling approach.

Q2: How does reading of specialist discussions affect generalists' diagnostic reasoning for future patient cases?

Beyond studying generalists' perception of our training interventions, it is crucial to investigate its effect on future medical assessments. We project that the presentation of case-specific adjudication discussions during labeler training will have benefits for generalists' diagnostic reasoning in a held-out test set.

In particular, we hypothesize that reading of adjudication discussions during training will:

[H2a] Improve diagnostic **accuracy**.

[H2b] Increase case-specific diagnostic **confidence**.

[H2c] Lower perceived case **difficulty**.

[H2d] Improve overall diagnostic **self-efficacy**.

METHODS

Experts

Our study involved two distinct groups of experts with varying levels of specialization who contributed during different stages of our data collection and experimental procedure.

Specialist Adjudicators. Three retina specialists collectively generated the answer key and adjudication discussions for the medical images used in this study. The adjudication process implemented a remote, round-based protocol for group discussion described in prior work [38]. First, each specialist adjudicator labeled each fundus image independently. Images with any level of disagreement were then reviewed in a round-robin fashion, by one specialist at a time.

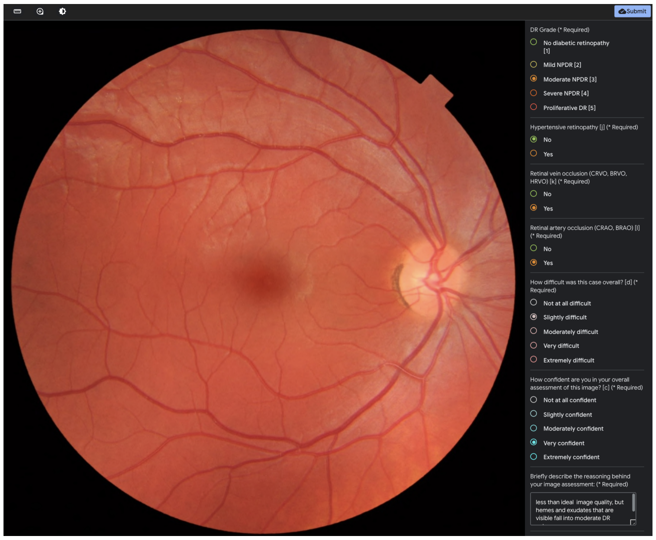
In each review round, the active specialist adjudicator was encouraged to explain the rationale behind their diagnostic reasoning within a text-based discussion thread, and to revise their diagnosis labels if they felt an adjustment was indicated based on insights from the adjudication discussion. The adjudication process ended for a given image when all members of the adjudication committee reached a unanimous consensus on all diagnosis labels for that image (or after a maximum of 15 review rounds, i.e., up to five reviews per adjudicator).

Note that this adjudication procedure was not designed with the purpose of training medical generalists in mind, but to create trusted ground truth labels for the validation of machine learning models. This study explores whether the discussion metadata generated as a side product in the process can be recycled as an effective tool for training medical generalists.

Generalists. Ten certified eye care professionals with varying training backgrounds participated as generalist labelers in the training experiment of our study. These included people at a lower level of specialization and those with substantially fewer years of retina-specific training compared to members of the specialist adjudication committee. We assigned each of the ten generalist labelers to one of the two types of training feedback, ensuring that both groups were relatively balanced with respect to training background and professional experience. There was no overlap between the two groups of specialist adjudicators and generalist labelers in our study.

Image Sets

Our study used two distinct image sets: a **train set** used to elicit each generalist's baseline labeling performance before receiving training feedback, and a held-out **test set** used to measure their labeling performance after training. Our training feedback focused on those image cases in the train set where labels from generalists differed from the answer key. Both image sets consisted of 36 images each.



DR Grade (* Required)

○ No diabetic retinopathy [1]
○ Mild NPDR [2]
● Moderate NPDR [3]
○ Severe NPDR [4]
○ Proliferative DR [5]

Hypertensive retinopathy [j] (* Required)

● No
○ Yes

Retinal vein occlusion (CRVO, BRVO, HRVO) [k] (* Required)

○ No
● Yes

Retinal artery occlusion (CRAO, BRAO) [l] (* Required)

○ No
● Yes

How difficult was this case overall? [d] (* Required)

○ Not at all difficult
● Slightly difficult
○ Moderately difficult
○ Very difficult
○ Extremely difficult

How confident are you in your overall assessment of this image? [c] (* Required)

○ Not at all confident
○ Slightly confident
○ Moderately confident
● Very confident
○ Extremely confident

Briefly describe the reasoning behind your image assessment: (* Required)

less than ideal image quality, but hemes and exudates that are visible fall into moderate DR

Figure 1. Task interface for medical image assessment. The medical image shown is an illustrative example rather than from the real dataset.

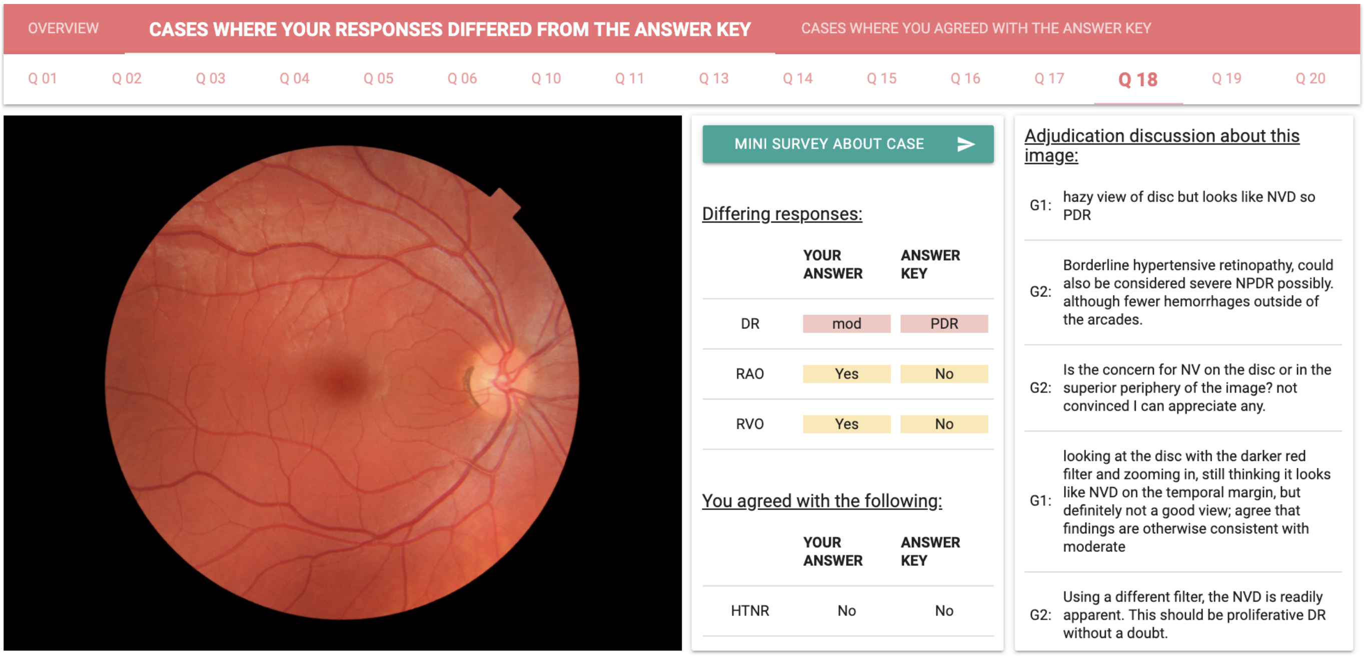
Images were selected from a larger set of 499 cases labeled by our committee of three retina specialists using the adjudication procedure outlined above. Specialists independently agreed on 329 out of the 499 cases, leaving 170 disagreement cases for the round-based review and discussion process. We performed a qualitative content analysis on these 170 disagreement cases based on the dialogs of their corresponding adjudication discussions. The objective of our qualitative analysis was to group difficult cases based on the specific source of disagreement as well as the final adjudicated consensus labels.

Disagreement sources were categorized in a fine-grained and domain-specific manner (e.g., the dark-red filter needs to be activated in order to detect the development of new vessels around the optic disk, evidence suggesting diagnosis of proliferative diabetic eye disease). Based on this fine-grained categorization, we formed pairs of cases sharing the same source of disagreement and final consensus labels. From each pair, we assigned one case to the train set and the other to the test set. Train and test set were thus enriched for difficult cases and each image in the train set matched a separate image in the test set.

In summary, we used 72 distinct cases in our experiment, 36 for training and 36 for testing. These 72 cases were selected from a larger set of 170 disagreement cases following the procedure described above. The remaining 98 cases could not be paired based on their source of disagreement and consensus labels and were therefore not used.

Procedure

Our study was designed to test two different forms of training feedback. The experiment was structured accordingly as a three-step procedure: a training task involving assessment of all images in the train set; a feedback phase providing information about cases from the training task where an generalist's answer differed from the adjudicated answer key; a testing task with all images from the held-out test set.


OVERVIEW
CASES WHERE YOUR RESPONSES DIFFERED FROM THE ANSWER KEY
CASES WHERE YOU AGREED WITH THE ANSWER KEY
Q 01
Q 02
Q 03
Q 04
Q 05
Q 06
Q 10
Q 11
Q 13
Q 14
Q 15
Q 16
Q 17
Q 18
Q 19
Q 20
MINI SURVEY ABOUT CASE

Differing responses:

	YOUR ANSWER	ANSWER KEY
DR	mod	PDR
RAO	Yes	No
RVO	Yes	No

You agreed with the following:

	YOUR ANSWER	ANSWER KEY
HTNR	No	No

Adjudication discussion about this image:

G1:	hazy view of disc but looks like NVD so PDR
G2:	Borderline hypertensive retinopathy, could also be considered severe NPDR possibly. although fewer hemorrhages outside of the arcades.
G2:	Is the concern for NV on the disc or in the superior periphery of the image? not convinced I can appreciate any.
G1:	looking at the disc with the darker red filter and zooming in, still thinking it looks like NVD on the temporal margin, but definitely not a good view; agree that findings are otherwise consistent with moderate
G2:	Using a different filter, the NVD is readily apparent. This should be proliferative DR without a doubt.

Figure 2. Training feedback interface for medical generalists. The medical image shown is an illustrative example rather than from the real dataset.

Pre- and post-study surveys were administered before and after the study. The pre-study survey was used to collect information about generalists' training background and professional experience. We also elicited generalists' self-efficacy at detecting each of the four pathologies both before and after the study. We determined the number of training cases and discussion points to show based on early piloting of the study and taking into account the constraints of the image selection procedure described above.

Training Task. Generalists assessed images for overall gradability and for the presence of four different pathologies: diabetic retinopathy (DR), hypertensive retinopathy (HTNR), retinal vein occlusion (RVO), and retinal artery occlusion (RAO). Generalists also rated their own diagnostic confidence and perceived case difficulty, each on 5-point Likert scales. While there exist alternative ways of measuring confidence, we used a 5-point scale for its granularity, following practices from prior clinical research [32]. Finally, for each case, generalists provided an open-ended explanation of the reasoning behind their rationale. Figure 1 shows the task interface including all input prompts for a gradable image.

Training Feedback. After completing the training task, generalist labelers received an email notification with a link to an automatically generated feedback document. For each case labeled during the training task, the feedback document compared the answer provided by the generalist to the adjudicated answer key. Experts were asked to review each case where their answer differed from the answer key.

For each case reviewed, generalists filled out a short survey, rating their level of comprehension for the rationale behind the answer key (5-point Likert scale), specifying the extent to which they agreed with the answer key (one of three answer options), and indicating whether they would change anything about their future labeling approach (including an open-ended explanation of what they would change). The purpose of the case review surveys was twofold. First, the surveys helped ensure that generalists reviewed the feedback carefully. Second, the surveys were used to collect structured information about generalists' perception of the feedback provided.

Testing Task. After reviewing the feedback for each of the cases where their answer differed from the answer key, generalists were assigned the testing task with images from the held-out test set they had not previously seen. The labeling procedure of the testing task was identical to that of the training task.

Experimental Conditions for Training Feedback

We compared two forms of training feedback for medical generalists to examine the impact of presenting generalists with specialist adjudication discussions for difficult cases:

- **Performance Only**: Our baseline condition identified all cases where any of the diagnosis labels provided by generalists during the training task differed from the adjudicated answer key. For each of these cases, our feedback interface presented the medical image in question along with a list comparing generalist-provided labels with the adjudicated answer key (Figure 2, left and middle).
- **Performance & Discussion**: In addition to providing individual performance feedback about the correctness of labels, our second type of training feedback also presented generalists with case-specific discussions from our specialist adjudication procedure. Specialist discussions were presented in a tabular format listing text-based comments (Figure 2, right). Specialist identities were anonymized to avoid potential biases on the side of generalists.

To support validation of our findings, we make our data, including adjudication discussions, characteristics of the generalist experts, as well as their labeling performance and survey responses, publicly available as auxiliary material.

Analysis

For **Q1**, we analyzed responses to our case review surveys to understand how medical generalists perceive reading of specialist discussions as a form of labeler training feedback. The case review surveys were collected for all cases where one or more generalist-provided labels differed from the answer key. The Mann-Whitney U test was employed to compare Likert type survey responses about perceived level of comprehension of the rationale behind the correct diagnosis **(H1a)**, agreement with the answer key **(H1b)**, and generalists' intention to change their labeling approach in the future **(H1c)**. We also qualitatively analyzed the open-ended explanations for why (or why not) generalists agreed with the answer key and what (if anything) they would change about their future labeling approach and why.

For **Q2**, we leveraged the fact that our two image sets for training and testing were composed of paired case examples. That is, for each case in the train set, there existed a separate case in the test set which had caused disagreement among specialist adjudicators for the same reason as the training example. We refer to these as train example and test example belonging to the same case-pair.

For our hypotheses about improvements in accuracy **(H2a)**, increased diagnostic confidence **(H2b)**, and lowered perceived case difficulty **(H2c)**, we first computed the respective score deltas between the test example and the train example for each case-pair and generalist labeler. Score deltas for correctness were computed separately for each pathology type (1 indicating improvement, i.e., wrong in train and correct in test; 0 indicating no change, i.e., wrong or correct in both train and test; -1 indicating decreased performance, i.e., correct in train, but wrong in test), and averaged across all generalists per group. We then compared the resulting average accuracy improvements per case-pair between both groups using a permutation test (with 9999 bootstrap samples, stratified by case-pair).

Score deltas for confidence and difficulty were computed once for each case-pair and generalist. We tested for differences between both groups using one-sided Mann Whitney U tests.

Finally, we hypothesized that exposing generalists to adjudication discussions from specialists would lead to an improvement in overall diagnostic self-efficacy **(H2d)**. Improvement was measured as the pre-to-post-study difference in diagnostic self-efficacy scores for each generalist and pathology type. Given the limited number of generalists in each group, results for this hypothesis are descriptive and should therefore only be used as an indication.

Open-ended survey responses collected from generalists after reviewing feedback for each training case were analyzed qualitatively. Line-by-line inductive open coding was used to identify emerging themes and recurring themes are reported below.

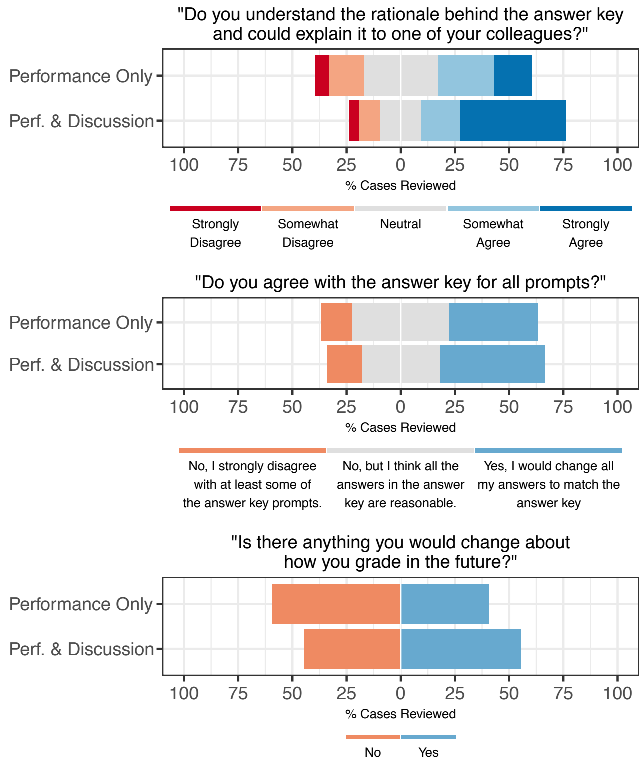


Figure 3. Generalists' perception of training feedback.

RESULTS

Q1: How do medical generalists perceive reading of specialist discussions as a form of labeler training feedback?

Our hypothesis **(H1a)** that exposing generalist labelers to adjudication discussions would facilitate a deeper understanding of the rationale behind the correct diagnosis was confirmed, indicating a very large effect size (U = 4620.50, z = -4.44, $p < 0.001$, r = 0.99). Generalists strongly agreed that they understood the rationale behind the answer key and could explain it to one of their colleagues in about half (49.1%; N = 114) of all cases reviewed along with adjudication discussions, compared to only 17.5% (N = 120) of cases reviewed without adjudication discussions (Figure 3, top). For the question as to whether generalists agreed with the answer key after reviewing the training feedback, no significant difference was detected between the two training feedback conditions, leaving our hypothesis **(H1b)** unconfirmed (U = 6470.50, z = -1.23, n.s., r = 0.28; Figure 3, middle). Finally, generalists who were provided with adjudication discussions during training feedback were significantly more likely to express an intention of changing their labeling approach in the future than generalists who were presented with just performance feedback alone, confirming our hypothesis **(H1c)** (U = 5853.00, z = -2.20, $p < 0.05$, r = 0.49; Figure 3, bottom). Generalists indicated that they would adjust their labeling approach for more than half (55.3%) of the cases reviewed along with adjudication discussions, while generalists in the group with performance feedback alone *denied* any future adjustment to their labeling approach for more than half (59.2%) of the cases reviewed.

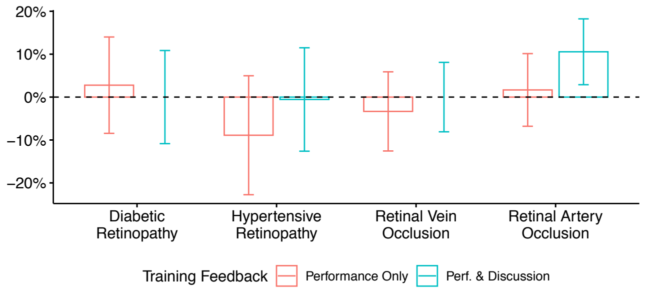


Figure 4. Average change in generalists' diagnostic accuracy per case-pair in train set and held-out test set. Error bars indicate 95% confidence intervals.

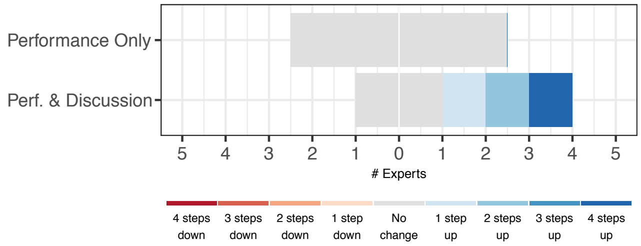


Figure 5. Improvement in generalists' self-efficacy score for diagnosis of retinal artery occlusion (RAO) after training feedback.

Q2: How does reading of specialist discussions affect generalists' diagnostic reasoning for future patient cases?

These benefits of reading adjudication discussions for generalists' perception during training feedback in part also translated to improvements in diagnostic accuracy on the held-out test set **(H2a)** (Figure 4). Generalists exposed to adjudication discussions during training feedback showed significantly greater accuracy improvements for diagnosing RAO ($\mu = 10.6\%$, CI [2.9%, 18.2%]) than generalists exposed to performance feedback alone ($\mu = 1.7\%$, CI [-6.8%, 10.1%]; $p < 0.05$; N = 36 case-pairs). No differences were detected for the other pathology types. Generalists exposed to discussions achieved an absolute test accuracy of 93% for RAO detection, up from 83% in training. Accuracies for DR, HTNR and RVO stayed constant before and after training at 61%, 64% and 83% respectively.

This benefit of showing adjudication discussions for accuracy improvements in RAO diagnosis was accompanied by similar improvements in self-efficacy **(H2d)**: while none of the generalists exposed to performance-only feedback reported any improvements in self-efficacy for RAO diagnosis, the majority of generalists presented with adjudication discussions did (one generalist with one step of improvement, a second generalist with two steps of improvement, and a third generalist with four steps of improvement on the 5-point Likert scale for self-efficacy; Figure 5).

Finally, we hypothesized that the presentation of adjudication discussions would lead to increased case-specific diagnostic confidence **(H2b)** and lowered levels of perceived case difficulty **(H2c)** in the testing task (Figure 6). Both hypotheses were rejected. Indeed, we observed the *opposite* effect: Reading adjudication discussions during training feedback was associated with greater reductions in diagnostic confidence (U = 18555.00, z = -2.74, p < 0.01, r = 0.61) and greater increases in perceived case difficulty (U = 14521.00, z = -2.08, p < 0.05, r = 0.47) compared to training with performance feedback alone.

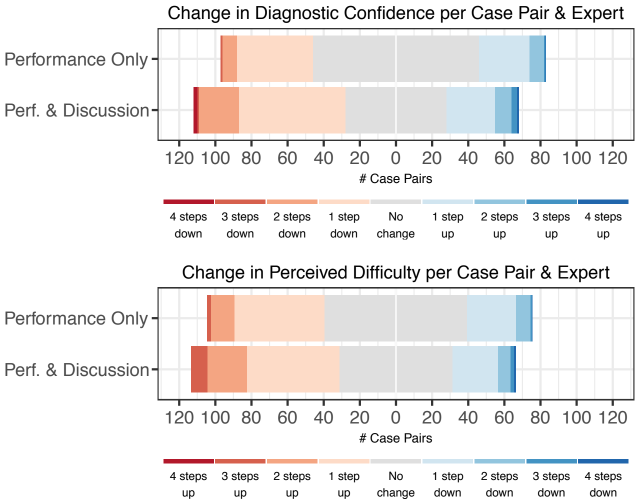


Figure 6. Change in generalists' diagnostic confidence and perceived case difficulty per case-pair in train set and held-out test set.

Qualitative Feedback from Medical Generalists

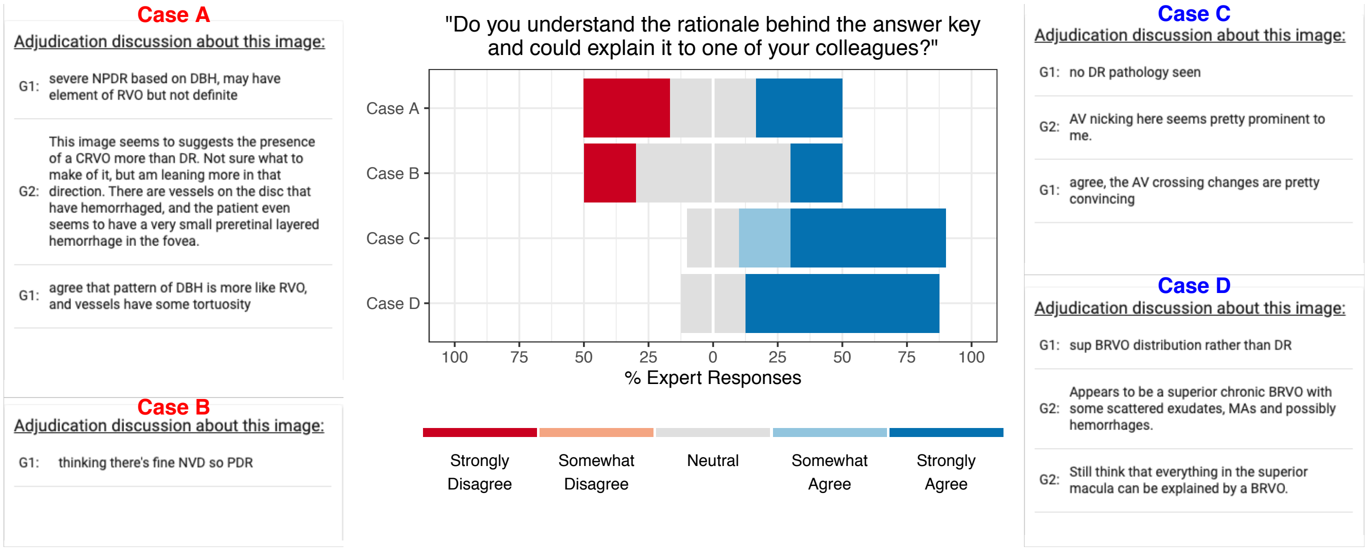
In addition to quantitative measures of diagnosis performance and attitudinal constructs (such as self-reported comprehension), generalists also provided qualitative feedback about their experience reviewing training cases with and without adjudication discussions.

Several themes emerged from this qualitative feedback. We describe some of the more salient themes below, with representative quotes from generalists.

Expressions of confusion and uncertainty when comparing their answers to the answer key: Without specialist discussions, the reasons why generalists were incorrect were often opaque:

- *"I think there could be subtle VB [venous beading]... is that the rationale for severe? I think if the resolution was better, I might be able to clearly see IRMA [intraretinal microvascular abnormalities] temp to the fovea... is that the rationale for severe? I wasn't 100% on these two things, thus the moderate grade."*
- *"would like clarification in the image about the features that make this severe NPDR [non-proliferative diabetic retinopathy] and not a CRVO [central retinal vein occlusion, a potential alternate diagnosis]."*

By contrast, when specialist discussions were present, generalists often cited specific details of their discussions in explaining their understanding of why they were incorrect:



Case A

Adjudication discussion about this image:

G1: severe NPDR based on DBH, may have element of RVO but not definite

G2: This image seems to suggests the presence of a CRVO more than DR. Not sure what to make of it, but am leaning more in that direction. There are vessels on the disc that have hemorrhaged, and the patient even seems to have a very small preretinal layered hemorrhage in the fovea.

G1: agree that pattern of DBH is more like RVO, and vessels have some tortuosity

Case B

Adjudication discussion about this image:

G1: thinking there's fine NVD so PDR

Case C

Adjudication discussion about this image:

G1: no DR pathology seen

G2: AV nicking here seems pretty prominent to me.

G1: agree, the AV crossing changes are pretty convincing

Case D

Adjudication discussion about this image:

G1: sup BRVO distribution rather than DR

G2: Appears to be a superior chronic BRVO with some scattered exudates, MAs and possibly hemorrhages.

G2: Still think that everything in the superior macula can be explained by a BRVO.

Figure 7. Example adjudication discussions with mixed ratings (cases A and B) and high ratings (cases C and D) for answer key comprehension.

- *“blurry view but i can go along with the heme noted by other graders”*
- *“I could see the PDR [proliferative diabetic retinopathy] that they were discussing. THere fore [sic] I could agree with them for DR [diabetic retinopathy].”*

Acknowledgement of missed features: In some cases, generalists originally failed to notice a feature; but when directed to the relevant part of the image by specialist discussion, acknowledged their miss:

- *“Agree with PRP scars, should have been PDR. ... will pay closer attention to laser scars”*
- *“Wasn’t able to detect small/early IRMA ...”* [When asked what they would change about grading behavior, the response was] *“Try to detect IRMA better”*

Calibrating cutoffs to match other clinicians: In many cases, generalists recognized that a pathology was potentially present, but ambiguous. They explicitly called out the potential for disagreement due to subjective differences in the cutoff for a finding being clinically significant. This theme emerged particularly in feedback around hypertensive retinopathy (referred to as HTN in reader comments) and image gradability. In each case, these subjective differences could lead to real changes in clinical outcomes, discussed below.

Much feedback was given around distinguishing hypertensive retinopathy versus diabetic retinopathy. Sample comments:

- *“AV nicking is present and per guidelines that would be considered mild and thus a yes grade for HTN. I do agree that it is mild and I was more generous with grading HTN as G2 mentioned in adjudication ... [I plan on] being more conservative on HTN grading.”*
- *“Overall, tended to undercall HTN in patients with clear DM”*
- *“I don’t think the AV nicking is as prominent as they [adjudicators] say, but I can see why they might have thought that ... I will look out closer for AV nicking”*

In addition to hypertensive retinopathy, generalists in our study also expressed uncertainty around the threshold for considering a low-quality image gradable:

- *“Image is blurry making my grade more of a guess, so I marked ungradable ... [I plan to] mark referable pathology even if image is blurry and there is some doubt”*
- *“As adjuncter G2 noted, image resolution makes the grading of MAs [microaneurysms] more of a guess; I didn’t mark it as ungradable since unlikely moderate or worse DR present.”*

Again, this distinction is subjective, yet has real-world implications: Many screening programs will refer patients to specialists if their image is ungradable. In both of these cases, the subjective differences reflect a common pattern observed in other cases of variability among eye doctors. Prior work by Kalpathy-Cramer et al. [20] demonstrated that disagreements among doctors could be explained by differences in transition points between different severity levels. Doctors tended to order cases by severity in a consistent manner; but varied in the point at which a case was “severe enough”. This suggests that feedback of the sort provided here can substantially improve concordance among doctors, by enabling them to calibrate to the same expert level (i.e., in the case of our study, to a specialist-provided answer key). A similar phenomenon was reported in a national screening program for breast cancer [25].

DISCUSSION

In this work, we introduce a novel perspective on the problem of calibrating medical professionals for accurate assessment of difficult cases in medical image labeling. We demonstrate empirically that specialist discussions from adjudication of difficult cases can be successfully used as training material for generalist labelers.

Impact on Comprehension and Accuracy

Our experimental results suggest that exposure to specialist discussions during training feedback improves generalists' comprehension of the rationale behind the correct diagnosis (H1a), and makes a future adjustment of their labeling approach more likely (H1c). We also demonstrated that these benefits observed during training translate into greater improvements in diagnostic accuracy in a held-out test set (H2a) and diagnostic self-efficacy (H2d) for one of the four retinal pathologies included in the study. While the overall effect on answer key comprehension was strongly positive, some adjudication discussions were perceived as more helpful than others. Figure 7 shows four examples of adjudication discussions: two discussions that received mixed ratings from generalists for answer key comprehension (cases A and B), and two with consistently high ratings among generalists (cases C and D). Case A is an example of a discussion characterized by vague language and phrases of uncertainty on the side of specialists, whereas the discussion for case B consists of a single comment only. While a full semantic analysis of our adjudication discussions is beyond the scope of our study, both language use and overall length of a discussion may have affected generalists' perception its usefulness. Future research may explore ways to motivate specialists a priori (i.e., before or during adjudication) to produce discussion points well suited for training purposes, and evaluate design parameters such as the number of cases and discussion points to show during training.

Impact on Labeling Confidence and Perceived Difficulty

We also hypothesized that presentation of specialist discussions during training feedback would increase generalists' labeling confidence (H2b) and decrease their perceived case difficulty (H2c) in the testing task, compared to showing performance feedback alone. Neither hypothesis could be confirmed. In fact, we observed the opposite effect: generalists who had seen adjudication discussions for difficult cases, scored lower on labeling confidence and higher on perceived difficulty on similar case types in the testing task. One possible explanation for this unexpected observation could be what has been coined the Dunning-Kruger effect [14] or *meta-ignorance*: the phenomenon that performance and confidence are often inversely correlated in intellectual tasks. This effect has been primarily explained with so-called *unknown unknowns* on the side of poor performers, i.e., their relative lack of awareness of deficiencies in their own expertise. Another possible explanation may be that the performance-only training condition did not reveal any information about the difficulty of a case. In other words, it did not transmit any information that would help generalists appreciate how hard the training cases were, whereas the training condition including adjudication discussions made the notion of difficult and contentious cases immediately transparent to generalists.

Learning from Discussions

Our work contributes to the existing body of literature on discussion-based learning. The benefits learners can draw from active participation in group discussions have been established in prior educational and psychological literature. These works have studied differences between learning from online versus face-to-face discussions.

In medical education specifically, the concept of discussion-based learning has been studied under the names of problem-based learning (PBL) and case-based learning (CBL) [11, 41]. Both PBL and CBL differ from lecture-based learning in that they engage medical students in small discussion groups for the purpose of collective clinical reasoning. PBL is a more open-ended form of discussion-based learning while CBL imposes more guidance and structure on the discussion process.

To our knowledge, there has been little prior research in repurposing expert case discussions for training purposes. Previous work [18] examined one potential application for screening mammography using a pre-existing public annotated image set, but cited a range of challenges, including data curation and quality issues. The authors noted that in practice, separating the production and use of data for different purposes was difficult to do cleanly. We believe our work has managed to avoid some of the challenges demonstrated in that work through careful matching of the adjudication and training tasks: generalists were making the same clinical judgments as specialist adjudicators, oriented around the same inputs (only image data, no metadata); both groups had previously been through a certification process for the task; our experimental design included some data curation; and the adjudication format intrinsically elicited more detailed justifications among experts that would not be elicited in a screening context.

Thus, our results extend the existing body of educational literature insofar as they demonstrate the benefits of exposing individuals to consumption of case discussions, as opposed to engaging groups in active discussions. This approach is inherently more scalable and flexible in nature than group-based learning.

Potential Clinical Impact

The contributions described here are framed in the context of training machine learning models: Generalist graders are trained to label images used for training a model, and our interventions aim to bring their performance closer to that of specialists. The adjudication discussions used for training were also collected as part of obtaining a test data set for an ML model. These improvements should enable higher-quality ML models, by improving the quality of training data collected by generalist labelers.

Our training intervention depends on the availability of specialist labels and discussions. Yet, as generalists' labeling accuracy increases through training, the need for label redundancy may decrease [24], enabling more efficient labeling strategies. Our work opens up questions about how best to distribute work between specialists and generalists in the absence of data ground truth. For example, specialists could be recruited to label a small, contained subset of data for training generalists, empowering generalists to take on the rest.

Our results may also translate to clinical practice without relying on ML model development. The labeling workflow we use here is similar to that used in telemedicine enterprises, including screening for eye disease [9, 5]. Likewise, the adjudication discussions we collected may mirror arbitration discussions used in some screening programs [25]. Thus, there may be potential to use discussions generated for screening purposes in training non-experts. Future work in this direction might aim at mapping new cases, with unknown labels, to similar cases in the adjudicated set, allowing clinicians to view discussions around similar cases in the context of a case being screened. In this way, training interventions like the one we demonstrate may expand the reach of screening programs without necessarily requiring ML systems.

Limitations

Our study has several limitations. First, our experiment was conducted with ten medical generalist labelers and three specialist adjudicators. While future work may aim to reproduce our findings with larger participant samples, samples of this size are not uncommon in studies of medical experts like ours where recruitment is a challenge. Given the sample size, we ensured to balance the level of experience of generalists between our two groups, and triangulated our findings with qualitative analysis, to enrich and provide further support for our quantitative findings.

Second, our study is situated in the medical subspeciality of image-based diagnosis in ophthalmology. While the general approach of collecting and presenting adjudication discussions can be easily applied to outside domains (medical or non-medical), caution is warranted in generalizing our findings to other disciplines. That said, prior work has demonstrated the prevalence of expert disagreement [3] and the effectiveness of discussion-based learning [10] across various medical domains, suggesting that our results on passive consumption of specialist discussions may generalize to other subspecialties as well. We encourage future work to validate our approach in other application scenarios.

Third, the remote nature of our study and the tight schedules of our expert participants did not permit precise control over the timing of the individual steps in the procedure. The overall study duration was about one week, but we did not account for potential differences in time experts spent between training and feedback, and between feedback testing phase. Our study also did not include a measure of long-term improvement.

Finally, the cost effectiveness of our proposed technique depends on a pre-existing electronic framework for asynchronous adjudication. While most tele-medical grading centers do not yet use such kind of procedure, there is a growing body of research in HCI on designing and developing methods for online group discussion among crowd workers, that could be leveraged towards a broader applicability of our approach [38].

CONCLUSION

In this work, we provided a novel perspective on the challenge of improving comprehension and diagnostic accuracy in medical data labeling. We demonstrated that existing specialist discussions from adjudication of difficult cases can be reused as training material for generalist labelers—without introducing additional cost to the labeling process. Our results suggest that the presentation of specialist adjudication discussions can improve generalists' comprehension of the rationale behind the correct diagnosis, and make a future adjustment of their labeling approach more likely. Furthermore, we showed that these benefits observed during training also translated into significantly greater improvements in diagnostic accuracy on a held-out test set for one out of four pathologies. Our work has important implications beyond medical diagnosis training alone, highlighting a practical method applicable to expert labeler training in high-stakes data labeling broadly.

ACKNOWLEDGMENTS

We thank Yun Liu for conducting a thorough review of this manuscript and for providing numerous helpful comments that were instrumental to improve this work.

REFERENCES

[1] Alaa Al Ali, Stephen Hallingham, and Yvonne M. Buys. 2015. Workforce supply of eye care providers in Canada: optometrists, ophthalmologists, and subspecialty ophthalmologists. *Canadian Journal of Ophthalmology* 50, 6 (dec 2015), 422–428. DOI: http://dx.doi.org/10.1016/j.jcjo.2015.09.001

[2] Elham Bagheri, Justin Dauwels, Brian C. Dean, Chad G. Waters, M. Brandon Westover, and Jonathan J. Halford. 2017. Interictal epileptiform discharge characteristics underlying expert interrater agreement. *Clinical Neurophysiology* 128, 10 (oct 2017), 1994–2005. DOI: http://dx.doi.org/10.1016/j.clinph.2017.06.252

[3] Michael L. Barnett, Dhruv Boddupalli, Shantanu Nundy, and David W. Bates. 2019. Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians. *JAMA Network Open* 2, 3 (mar 2019), e190096. DOI: http://dx.doi.org/10.1001/jamanetworkopen.2019.0096

[4] Andrew Bastawrous and Benjamin D Hennig. 2012. The global inverse care law: a distorted map of blindness. *British Journal of Ophthalmology* 96, 10 (oct 2012), 1357.2–1358. DOI: http://dx.doi.org/10.1136/bjophthalmol-2012-302088

[5] Anthony A. Cavallerano and Paul R. Conlin. 2008. Teleretinal Imaging to Screen for Diabetic Retinopathy in the Veterans Health Administration. *Journal of Diabetes Science and Technology* 2, 1 (jan 2008), 33–39. DOI:http://dx.doi.org/10.1177/193229680800200106

[6] Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. 2019b. How to develop machine learning models for healthcare. *Nature Materials* 18, 5 (may 2019), 410–414. DOI:http://dx.doi.org/10.1038/s41563-019-0345-0

[7] Quanze Chen, Jonathan Bragg, Lydia B. Chilton, and Daniel S. Weld. 2019a. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1–14. DOI: http://dx.doi.org/10.1145/3290605.3300761

[8] Robin Cohen, Mike Schaekermann, Sihao Liu, and Michael Cormier. 2019. Trusted AI and the Contribution

of Trust Modeling in Multiagent Systems. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1644–1648. http://dl.acm.org/citation.cfm?id=3306127.3331890

[9] Jorge Cuadros and George Bresnick. 2009. EyePACS: An Adaptable Telemedicine System for Diabetic Retinopathy Screening. *Journal of Diabetes Science and Technology* 3, 3 (may 2009), 509–516. DOI: http://dx.doi.org/10.1177/193229680900300315

[10] Jasmin Diwan, Chinmay Shah, Saurin Sanghavi, and Amit Shah. 2017. Comparison of case-based learning and traditional lectures in physiology among first year undergraduate medical students. *National Journal of Physiology, Pharmacy and Pharmacology* (2017), 1. DOI: http://dx.doi.org/10.5455/njppp.2017.7.0204220032017

[11] Tim Dornan, Albert Scherpbier, Nigel King, and Henny Boshuizen. 2005. Clinical teachers and problem-based learning: a phenomenological study. *Medical Education* 39, 2 (feb 2005), 163–170. DOI: http://dx.doi.org/10.1111/j.1365-2929.2004.01914.x

[12] Ryan Drapeau, Lydia B. Chilton, Jonathan Bragg, and Daniel S. Weld. 2016. MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP).*

[13] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing Ground Truth for Medical Relation Extraction. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (jul 2018), 1–20. DOI: http://dx.doi.org/10.1145/3152889

[14] David Dunning. 2011. The Dunning–Kruger Effect. 247–296. DOI: http://dx.doi.org/10.1016/B978-0-12-385522-0.00005-6

[15] Matthew J. Gabel, Norman L. Foster, Judith L. Heidebrink, Roger Higdon, Howard J. Aizenstein, Steven E. Arnold, Nancy R. Barbas, Bradley F. Boeve, James R. Burke, Christopher M. Clark, Steven T. DeKosky, Martin R. Farlow, William J. Jagust, Claudia H. Kawas, Robert A. Koeppe, James B. Leverenz, Anne M. Lipton, Elaine R. Peskind, R. Scott Turner, Kyle B. Womack, and Edward Y. Zamrini. 2010. Validation of Consensus Panel Diagnosis in Dementia. *Archives of Neurology* 67, 12 (dec 2010). DOI: http://dx.doi.org/10.1001/archneurol.2010.301

[16] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. 2018. Who said what: Modeling individual labelers improves classification. In *AAAI Conference on Artificial Intelligence*. https://arxiv.org/pdf/1703.08774.pdf

[17] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *Jama* 304, 6 (2016), 649–656. DOI: http://dx.doi.org/10.1001/jama.2016.17216

[18] Mark Hartswood, Rob Procter, Paul Taylor, Lilian Blot, Stuart Anderson, Mark Rouncefield, and Roger Slack. 2012. Problems of data mobility and reuse in the provision of computer-based training for screening mammography. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*. ACM Press, New York, New York, USA, 909. DOI:http://dx.doi.org/10.1145/2207676.2208533

[19] Oana Inel, Lora Aroyo, Chris Welty, and Robert-Jan Sips. 2013. Domain-Independent Quality Measures for Crowd Truth Disagreement. In *The 12th International Semantic Web Conference (ISWC2013).* http://data.semanticweb.org/workshop/derive/2013/proceedings/paper-01/html

[20] Jayashree Kalpathy-Cramer, J. Peter Campbell, Deniz Erdogmus, Peng Tian, Dharanish Kedarisetti, Chace Moleta, James D. Reynolds, Kelly Hutcheson, Michael J. Shapiro, Michael X. Repka, Philip Ferrone, Kimberly Drenser, Jason Horowitz, Kemal Sonmez, Ryan Swan, Susan Ostmo, Karyn E. Jonas, R.V. Paul Chan, Michael F. Chiang, Michael F. Chiang, Susan Ostmo, Kemal Sonmez, J. Peter Campbell, R.V. Paul Chan, Karyn Jonas, Jason Horowitz, Osode Coki, Cheryl-Ann Eccles, Leora Sarna, Audina Berrocal, Catherin Negron, Kimberly Denser, Kristi Cumming, Tammy Osentoski, Tammy Check, Mary Zajechowski, Thomas Lee, Evan Kruger, Kathryn McGovern, Charles Simmons, Raghu Murthy, Sharon Galvis, Jerome Rotter, Ida Chen, Xiaohui Li, Kent Taylor, Kaye Roll, Jayashree Kalpathy-Cramer, Deniz Erdogmus, Maria Ana Martinez-Castellanos, Samantha Salinas-Longoria, Rafael Romero, Andrea Arriola, Francisco Olguin-Manriquez, Miroslava Meraz-Gutierrez, Carlos M. Dulanto-Reinoso, and Cristina Montero-Mendoza. 2016. Plus disease in retinopathy of prematurity: improving diagnosis by ranking disease severity and using quantitative image analysis. *Ophthalmology* 123, 11 (nov 2016), 2345–2351. DOI: http://dx.doi.org/10.1016/j.ophtha.2016.07.020

[21] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2018. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology* (mar 2018). DOI: http://dx.doi.org/10.1016/j.ophtha.2018.01.034

[22] Joseph D Kronz, Mark A Silberman, William C Allsbrook, and Jonathan I Epstein. 2000. A web-based tutorial improves practicing pathologists' Gleason

grading of images of prostate carcinoma specimens obtained by needle biopsy. *Cancer* 89, 8 (oct 2000), 1818–1823. DOI: http://dx.doi.org/10.1002/1097-0142(20001015)89:8<1818::AID-CNCR23>3.0.CO;2-J

[23] P R Lichter. 1976. Variability of expert observers in evaluating the optic disc. *Transactions of the American Ophthalmological Society* 74 (1976), 532–72. http://www.ncbi.nlm.nih.gov/pubmed/867638http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1311528

[24] Christopher H Lin, Daniel S Weld, and Others. 2014. To re (label), or not to re (label). In *Second AAAI Conference on Human Computation and Crowdsourcing*.

[25] J.C Liston and B.J.G Dall. 2003. Can the NHS Breast Screening Programme Afford not to Double Read Screening Mammograms? *Clinical Radiology* 58, 6 (jun 2003), 474–477. DOI: http://dx.doi.org/10.1016/S0009-9260(03)00063-1

[26] Joaquin Navajas, Tamara Niella, Gerry Garbulsky, Bahador Bahrami, and Mariano Sigman. 2018. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour* (jan 2018). DOI: http://dx.doi.org/10.1038/s41562-017-0273-4

[27] Sonia Phene, R. Carter Dunn, Naama Hammel, Yun Liu, Jonathan Krause, Naho Kitade, Mike Schaekermann, Rory Sayres, Derek J. Wu, Ashish Bora, Christopher Semturs, Anita Misra, Abigail E. Huang, Arielle Spitze, Felipe A. Medeiros, April Y. Maa, Monica Gandhi, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2019. Deep Learning and Glaucoma Specialists: The Relative Importance of Optic Disc Features to Predict Glaucoma Referral in Fundus Photographs. *Ophthalmology* (sep 2019). DOI: http://dx.doi.org/10.1016/j.ophtha.2019.07.024

[28] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Robert Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. 2018. Direct Uncertainty Prediction for Medical Second Opinions. (jul 2018). http://arxiv.org/abs/1807.01771

[29] Pranav Rajpurkar, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y. Ng. 2017. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. (jul 2017). http://arxiv.org/abs/1707.01836

[30] Paisan Raumviboonsuk, Jonathan Krause, Peranut Chotcomwongse, Rory Sayres, Rajiv Raman, Kasumi Widner, Bilson J. L. Campana, Sonia Phene, Kornwipa Hemarat, Mongkol Tadarati, Sukhum Silpa-Archa, Jirawut Limwattanayingyong, Chetan Rao, Oscar Kuruvilla, Jesse Jung, Jeffrey Tan, Surapong Orprayoon, Chawawat Kangwanwongpaisan, Ramase Sukumalpaiboon, Chainarong Luengchaichawang, Jitumporn Fuangkaew, Pipat Kongsap, Lamyong Chualinpha, Sarawuth Saree, Srirut Kawinpanitan, Korntip Mitvongsa, Siriporn Lawanasakol, Chaiyasit Thepchatri, Lalita Wongpichedchai, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2019. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *npj Digital Medicine* 2, 1 (dec 2019), 25. DOI: http://dx.doi.org/10.1038/s41746-019-0099-8

[31] Richard S. Rosenberg and Steven van Hout. 2013. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *Journal of Clinical Sleep Medicine* (jan 2013). DOI: http://dx.doi.org/10.5664/jcsm.2350

[32] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, Shawn Xu, Scott Barb, Anthony Joseph, Michael Shumski, Jesse Smith, Arjun B. Sood, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2019. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology* 126, 4 (apr 2019), 552–564. DOI: http://dx.doi.org/10.1016/j.ophtha.2018.11.016

[33] Mike Schaekermann. 2016. Resolvable vs . Irresolvable Ambiguity : A New Hybrid Framework for Dealing with Uncertain Ground Truth. In *ACM SIGCHI Workshop on Human-Centered Machine Learning*, Marco Gillies and Rebecca Fiebrink (Eds.). ACM, San Jose. http://hcml2016.goldsmithsdigital.com/program/

[34] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019a. Capturing Expert Arguments from Medical Adjudication Discussions in a Machine-readable Format. In *Companion Proceedings of The 2019 World Wide Web Conference - WWW '19*, Vol. 2. ACM Press, New York, New York, USA, 1131–1137. DOI: http://dx.doi.org/10.1145/3308560.3317085

[35] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019b. Understanding Expert Disagreement in Medical Data Analysis through Structured Adjudication. In *Proceedings of the 2019 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2019)*, Vol. 3. Austin, TX, 1–23. DOI: http://dx.doi.org/10.1145/3359178

[36] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. Ambiguity-aware AI Assistants for Medical Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems - CHI '20*. ACM Press, Honolulu, HI, USA. DOI: http://dx.doi.org/10.1145/3313831.3376506

[37] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. In *Proceedings of the 2018 ACM Conference on Computer*

Supported Cooperative Work and Social Computing (CSCW 2018), Vol. 2. New York City, NY, 1–19. DOI: http://dx.doi.org/10.1145/3274423

[38] Mike Schaekermann, Naama Hammel, Michael Terry, Tayyeba K. Ali, Yun Liu, Brian Basham, Bilson Campana, William Chen, Xiang Ji, Jonathan Krause, Greg S. Corrado, Lily Peng, Dale R. Webster, Edith Law, and Rory Sayres. 2019. Remote Tool-Based Adjudication for Grading Diabetic Retinopathy. *Translational Vision Science & Technology* 8, 6 (dec 2019), 40. DOI:http://dx.doi.org/10.1167/tvst.8.6.40

[39] Mike Schaekermann, Edith Law, Kate Larson, and Andrew Lim. 2018. Expert Disagreement in Sequential Labeling: A Case Study on Adjudication in Medical Time Series Analysis. In *1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing at HCOMP 2018*. Zurich, Switzerland.

[40] Lili Shi, Huiqun Wu, Jiancheng Dong, Kui Jiang, Xiting Lu, and Jian Shi. 2015. Telemedicine for detecting diabetic retinopathy: a systematic review and meta-analysis. *British Journal of Ophthalmology* 99, 6 (jun 2015), 823–831. DOI: http://dx.doi.org/10.1136/bjophthalmol-2014-305631

[41] Malathi Srinivasan, Michael Wilkes, Frazier Stevenson, Thuan Nguyen, and Stuart Slavin. 2007. Comparing Problem-Based Learning with Case-Based Learning: Effects of a Major Curricular Shift at Two Institutions. *Academic Medicine* 82, 1 (jan 2007), 74–82. DOI: http://dx.doi.org/10.1097/01.ACM.0000249963.93776.aa

[42] Jens B. Stephansen, Alexander N. Olesen, Mads Olsen, Aditya Ambati, Eileen B. Leary, Hyatt E. Moore, Oscar Carrillo, Ling Lin, Fang Han, Han Yan, Yun L. Sun, Yves Dauvilliers, Sabine Scholz, Lucie Barateau, Birgit Hogl, Ambra Stefani, Seung Chul Hong, Tae Won Kim, Fabio Pizza, Giuseppe Plazzi, Stefano Vandi, Elena Antelmi, Dimitri Perrin, Samuel T. Kuna, Paula K. Schweitzer, Clete Kushida, Paul E. Peppard, Helge B. D. Sorensen, Poul Jennum, and Emmanuel Mignot. 2018. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications* 9, 1 (dec 2018), 5229. DOI: http://dx.doi.org/10.1038/s41467-018-07229-3

[43] Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, Nguyen D. Quang, Alfred Gan, Haslina Hamzah, Renata Garcia-Franco, Ian Yew San Yeo, Shu Yen Lee, Edmund Yick Mun Wong, Charumathi Sabanayagam, Mani Baskaran, Farah Ibrahim, Ngiap Chuan Tan, Eric A. Finkelstein, Ecosse L. Lamoureux, Ian Y. Wong, Neil M. Bressler, Sobha Sivaprasad, Rohit Varma, Jost B. Jonas, Ming Guang He, Ching-Yu Cheng, Gemmy Chui Ming Cheung, Tin Aung, Wynne Hsu, Mong Li Lee, and Tien Yin Wong. 2017. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA* 318, 22 (dec 2017), 2211. DOI: http://dx.doi.org/10.1001/jama.2017.18152

[44] Daniel Shu Wei Ting, Gemmy Chui Ming Cheung, and Tien Yin Wong. 2016. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clinical & experimental ophthalmology* 44, 4 (may 2016), 260–77. DOI:http://dx.doi.org/10.1111/ceo.12696