Iterative Quality Control Strategies for Expert Medical Image Labeling

Beverly Freeman, Naama Hammel, Sonia Phene, Abigail Huang, Rebecca Ackermann, Olga Kanzheleva, Miles Hutson, Caitlin Taggart, Quang Duong, Rory Sayres

Google Health

{beverlyf, nhammel, sphene, abigailhuang, rebackermann, okanzheleva, hutson, ctaggart, qduong, sayres}@google.com

Abstract

Data quality is a key concern for artificial intelligence (AI) efforts that rely on crowdsourced data collection. In the domain of medicine in particular, labeled data must meet high quality standards, or the resulting AI may perpetuate biases or lead to patient harm. What are the challenges involved in expert medical labeling? How do AI practitioners address such challenges? In this study, we interviewed members of teams developing AI for medical imaging in four subdomains (ophthalmology, radiology, pathology, and dermatology) about their quality-related practices. We describe one instance of low-quality labeling being caught by automated monitoring. The more proactive strategy, however, is to partner with experts in a collaborative, iterative process prior to the start of high-volume data collection. Best practices including 1) co-designing labeling tasks and instructional guidelines with experts, 2) piloting and revising the tasks and guidelines, and 3) onboarding workers enable teams to identify and address issues before they proliferate.

Introduction

As artificial intelligence (AI) applications become more widespread, there is a growing need for high-quality labeled data. Many AI applications require large labeled data sets, on the order of tens of thousands of examples or more (Ting et al. 2017; Phene et al. 2019; Liu et al. 2020) to train and validate a sufficiently high-performing model. Often, such labels can only be collected via a large-scale labeling process (Gulshan et al. 2016; Nagpal et al. 2020).

Label quality has emerged as a key challenge (Daniel et al. 2018). Recent work has demonstrated that training with low-quality labels, identified by methods such as cross-validation, results in poorer-performing models than when such labels are excluded (Hsu et al. 2020). Low label quality can pose many risks, including 1) models that are inaccurate, or that generalize poorly outside of the training sets, 2) significant time and resource costs, and 3) models that amplify worker bias (Jiang and Nachum 2020). Quality issues may not be apparent until after a model is trained and tested against a held-out set.

This challenge is further exacerbated in the application of AI to higher-risk domains, such as medical imaging. AI models have demonstrated performance equal to or greater than that of experts on diagnostic tasks such as identifying eye disease (Ting et al. 2017; Gulshan et al. 2019) or cancer (Esteva et al. 2017; McKinney et al. 2020). But if deployed in real-world clinical settings, poorly-performing or poorly-generalizing models may lead to patient harm (Zou and Schiebinger 2018; Challen et al. 2019).

Moreover, medical-imaging models often require labels from experts. This can be costly, due to the limited pool and availability of workers with sufficient medical training. Training a model based on a large data set before assessing label quality is thus particularly risky in this domain.

As a result, there is strong motivation to establish practices to ensure label quality for medical imaging. What quality-related practices do teams developing medical imaging AI employ? What are the unique challenges and opportunities of expert labeling as they relate to label quality?

This paper addresses these questions by reporting on interviews with teams under real-world constraints of developing AI for clinical deployment. We observe the practical application of principles described in the commodity crowdsourcing literature, and illustrate a novel set of challenges relating to experts' prior heuristics. In contrast to commodity crowdsourcing, workers on medical labeling tasks bring considerable prior experience, much of which reflects Gestalt rather than explicit knowledge. This may result in a mismatch between clinical practice and the needs of labels for AI development.

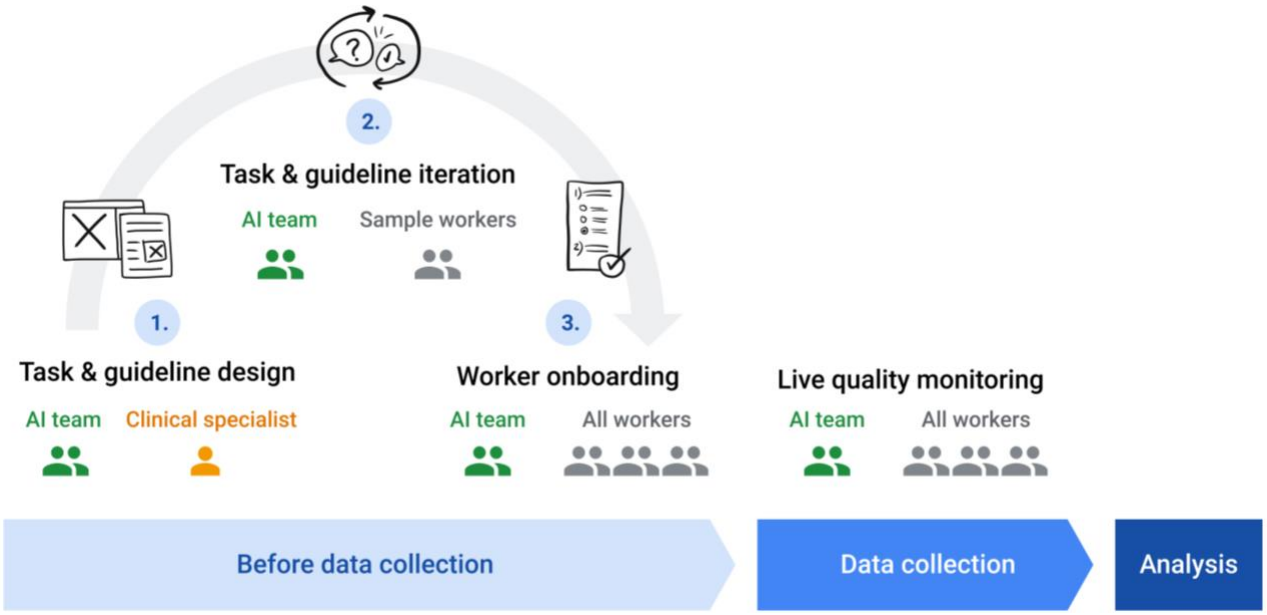


Figure 1: Quality-control mechanisms used by teams developing medical imaging AI. Upstream efforts involved co-designing and iterating labeling tasks and instructions with experts. Downstream efforts included automated label-quality monitoring.

We describe a process designed specifically to address this mismatch, in which AI practitioners partner with experts to 1) co-design labeling tasks and detailed instructional guidelines, 2) iterate the tasks and guidelines via small-scale pilots, and 3) onboard workers via tests that train and ensure guideline compliance. See Figure 1 for an overview of these processes.

A key insight from our interviews is that these practices focus primarily on *partnering with* and *instructing* expert workers, rather than *filtering out* low-performing workers. Iterative guideline development identifies points of misalignment between clinicians' approach to a task and the requirements for labels to train AI systems. Onboarding tests train experts to use explicit guidelines rather than rely solely on their own pre-existing clinical heuristics.

Related Work

Extensive prior literature documents label-quality considerations, much of it focused on commodity crowdsourcing platforms. Daniel et al. (2018) provide an extensive survey and synthesis of prior commodity crowdsourcing literature. They derive a quality model, which formally specifies the entities, dimensions, and attributes relevant to label quality. They review a range of interventions and methods to ensure quality. They further assess how 14 crowdsourcing platforms support different assessment methods for workers. Platforms provide some support for identifying workers with particular skill sets (such as qualification tests for particular tasks), but these tend to focus on relatively simple tasks. For example, Heer and Bostock (2010) show that a qualification task for graphical perception tasks on simple visualizations may be effective.

A common quality-control pattern in commodity tasks is to monitor performance by assigning questions with known ground truth, referred to as "gold standard" data sets (Le et al. 2010). Other approaches focus on measuring consistency among workers, sometimes using algorithms to estimate overall accuracy per worker and item, such as expectation-maximization approaches (Ipeirotis et al. 2010; Huang and Fu 2013). While measures of worker agreement may reflect quality, some conceptual frameworks indicate that agreement only reflects common knowledge of workers. Such common knowledge may not always converge on correct answers (Waggoner and Chen 2014). Yet other approaches involve identifying low-performing workers with adversarial intent, such as workers who are paid per task and are motivated to complete tasks quickly to maximize income, without regard for quality (Checco et al. 2020).

By contrast, other research has focused on improving worker performance by improving the labeling experience itself. Gaikwad et al. (2017) criticize the assumption that "low-quality work is the fault of workers." They propose "prototype tasks," a process in which requesters launch tasks to a small number of workers, solicit feedback, and iterate on the tasks based on the feedback. Similarly, Bragg et al. (2018) describe a system in which workers surface points of confusion and suggest alternative task phrasing or structure. Manam et al. (2019) show that quality issues may

reflect shortcomings in the design of the labeling task questions and/or instructions given to workers, rather than shortcomings in worker skill or conscientiousness.

Relatively few studies have focused on tasks requiring workers with domain expertise. Barrett and Sherman (2019) examine quality assessment of labels from expert workers in a legal task (tagging legal rulings text). They show that inter-worker agreement metrics on sequential batches of tasks can reflect data quality.

Within the medical domain, Ørting et al. (2020) review 57 papers that discuss the use of non-expert crowdsourced workers to label medical images. They identify a range of image domains (including brain, eye, lung, breast, and heart), a range of tasks (most commonly image classification and/or segmentation), and a range of image-based comparisons. Many of the cited papers focus on the use of non-expert crowd workers to label examples to train an expert task, although the authors note that several tasks may not be well-suited to crowd workers. Another analysis of medical image labeling indicates that worker time on task can be a useful signal of low quality, but that time alone is not sufficiently robust to be clearly diagnostic (Hutson et al. 2019).

Outside of the crowdsourcing domain, teams working in medical imaging have developed approaches to involving clinical experts in AI efforts to design effective product onboarding (Cai et al. 2019) and increase the accuracy of medical generalists (Schaekermann, Cai, et al. 2020).

While interest in large-scale label collection for medical imaging has increased, there remains a gap in understanding how the above approaches may or may not apply to this expert domain. The focus of the present research was to understand what practices teams partnering with medical experts to collect medical imaging labels have developed to ensure data quality.

Methods

We conducted 12 1-hour interviews with members of teams developing medical imaging AI in 4 subdomains: ophthalmology, radiology, pathology, and dermatology. A summary of participants is provided in Table 1.

To understand overall labeling processes, we interviewed six program managers (who managed labeling operations, resource allocation, and performance monitoring). To understand the perspectives of other functions involved in labeling, we interviewed four clinical specialists (medical domain experts who apply clinical expertise to labeling efforts), one engineer with AI specialization, and one user experience researcher who consulted on task and guideline design. The 12 participants were all the team members from the organization who were available for us to interview.

Participant	Domain	Role
P1	Ophthalmology	Program manager
P2	Radiology	Program manager
P3	Pathology	Program manager
P4	Pathology	Program manager
P5	Dermatology	Program manager
P6	Dermatology	Program manager
P7	Ophthalmology	Clinical specialist
P8	Ophthalmology	Clinical specialist
P9	Radiology	Clinical specialist
P10	Pathology	Clinical specialist
P11	Ophthalmology	Software engineer
P12	Radiology	User experience researcher

Table 1: Summary of interview participants.

Workers for each project included clinicians with training relevant to the given medical imaging domain. Workers' degree of prior experience varied depending on the project. Some projects included trainees, while others required doctors who had completed their training; still others required board certification for a specific specialization (such as board-certified radiologists or ophthalmologists who had completed retina or glaucoma fellowships). Worker managers, regardless of job title, were responsible for assigning tasks to workers, communicating the task and guidelines to them, and monitoring progress on tasks. Program managers more specifically were responsible for assigning workers to projects, and investigating and resolving worker performance issues, if needed.

Participants worked on a range of AI projects within the medical imaging domain. Typically, each project involved developing supervised learning models on one or more more images in a given domain, such as ophthalmology or digital slide pathology. Tasks primarily focused around classification tasks, such as determining disease severity of a case, or whether a case does or does not contain a specific pathology. Some teams worked on localization tasks in which models are trained to specify a region of interest, such as a suspicious area in an image or volume. One team also developed models that classified disease severity across a large digital slide. During interviews, we asked about label quality within the context of the specific tasks for that domain.

Data labeling generally occurred using image viewers operating within an internal, HIPAA-compliant platform for medical image labeling, using image viewers that were customized for different imaging modalities. Labeling

occurred with strict data-protection policies, such that only workers and team members with explicit permission were able to access images.

Due to the open-ended nature of the research question of how teams ensure quality, we used a semi-structured interview protocol. The focus was on collecting specific stories of quality-related incidents, and understanding the implications of these incidents. The core prompt for each interview was: *"Describe a time when you had concerns about a worker's label quality."* We considered asking specifically about "the last time" but decided to allow interviewees to start with the most salient incident, and then probe on more recent incidents. If an experienced team had developed effective practices for ensuring quality, we wanted to learn about their formative experiences and how they influenced subsequent practices. We probed on several key aspects of each label-quality incident:

- **"How did it come to your attention?"** We sought to understand what *signals* (quantitative or qualitative) could surface potential issues, especially prior to the conclusion of data collection.
- **"What did you do?"** Follow-up questions probed what *interventions* were used to diagnose and rectify quality issues at various points in the process.
- **"How have your processes changed over time?"** We sought to understand what *practices* the teams instituted as a result of these incidents, in order to curtail future issues.

Each interview was conducted by the first author and observed by at least one fellow author. All interviews were video recorded. Using a Reflexive Thematic Analysis approach (Braun and Clarke 2006, 2019), we transcribed the interviews, created initial codes via an inductive (bottom-up) approach, and generated initial themes. Authors participated in multiple rounds of discussion to reflect on interviewee stories, identify commonalities across teams and projects, and iterate on the themes presented below.

Results

Inter-Worker Variability Was a Key Challenge in Medical Image Labeling

All teams interviewed described challenges with label variability across workers. An example is illustrated in Figure 2. These data were shared by a team developing a model to assess glaucoma risk from retina images using a four-point risk scale. The team found that many cases, when labeled by multiple workers, had high variability across labels from different workers. Other teams also reported variability as a key challenge to be managed.

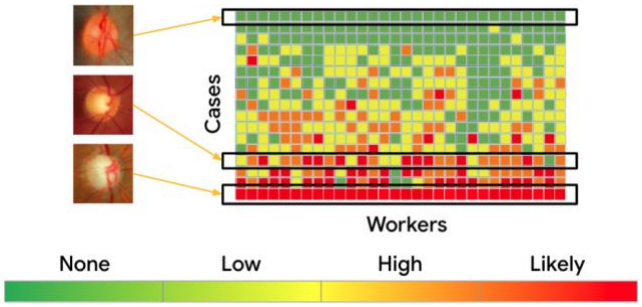


Figure 2: Labeling variability across expert workers. Highlighted rows indicate (from top to bottom): a case that all workers agreed is non-glaucomatous, a case that elicited all 4 possible risk levels, and a case that all workers agreed is at high risk for glaucoma.

In the medical domain, several factors were identified as contributing to variability across workers, including:

- **Differences in training or experience**: Experts may come from different training backgrounds. As P7 stated, "You end up doing what you learned during training.". Those with more experience may benefit from being able to draw from a wider range of real-life data points.
- **Differences in individual tendencies**: Teams have observed differences in labeling approaches based on different inclinations: "(Workers) have a tendency to fall across the spectrum regarding specificity versus sensitivity, which will be more likely to surface if explicit instructions are not provided as to where on the sensitivity/specificity side we prefer them to be" (P9).
- **Differences in understanding of worker guidelines**: Teams realized that workers sometimes approached the same task inconsistently due to different guideline interpretations: "After looking at the data, it turned out that half of the (workers) interpreted a question one way, and half of them another way. At least half of our data set was useless. We learned that we needed to test run the guidelines before opening up the entire job for labeling" (P3).

These factors can impact consistency and quality of labels and resulting AI models.

Task Design Mapped from Clinical to AI Needs

A key mitigation to address differences across workers was careful design of the labeling task and guidelines. *Task design* refers to the detailed structure of questions asked around a set of medical images. "*Guidelines"* refers to documents outside the task that orient workers on how to answer questions.

The guidelines served several purposes. First, guidelines instructed workers on how to *apply pre-existing clinical workflows and heuristics*. For instance, a team developing models for glaucoma risk detection based initial guidelines off recommendations from the American Academy of

Ophthalmology's Primary Open-Angle Glaucoma Suspect Preferred Practice Patterns (Prum et al. 2016).

Second, guidelines *oriented workers from a clinical setting to an AI setting*, taking them from familiar tasks to tasks suited for AI labels. For instance, workers in the glaucoma risk project had to assess risk based on only an image of a patient's retina. By contrast, the clinical risk assessment for which workers were trained involved many inputs in addition to the retinal image, including patient metadata (e.g. age and family history), data from prior visits, other measurements (e.g. intraocular pressure and visual field testing), and potentially images from a different modality (such as optical coherence tomography). Because workers weren't trained in assessing images alone, guidelines needed to be more explicit about the definition of visual features workers were expected to identify.

Third, guidelines *reduced ambiguity by including examples of cases representing different answer choices*. For instance, during glaucoma risk assessment, guidelines showed examples of retina images with and without specific risk indicators. The practice of including examples, especially ambiguous examples, has been shown to improve labeling accuracy (Pradhan and Lease 2018).

Fourth, guidelines *primed workers for the task at hand*. Interviewed teams described the importance of setting context for workers, especially when the labeling project used enriched data sets (in order to obtain more positive examples of rare conditions). For example, one radiology guideline document stated, "Imagine these images are from patients at a [condition] screening clinic in a region of [location] where [condition] is relatively common." Without sufficient context about the data, workers might have biased their answers (consciously or not) with their underlying assumptions about disease prevalence. For example, if during a labeling task, a worker felt they had identified pathologies at a higher rate than they would expect in a clinic, they might have tended to under-diagnose that pathology in the remaining portion of the labeling task.

Given the many functions of task and guideline design, the design process tended to involve several steps, which we describe below.

AI Teams and Clinical Experts Partnered in an Iterative Process Prior to the Launch of High-Volume Data Collection

Team members described an iterative process to identify and address issues related to task design, guideline design, and worker training. These issues were typically identified well before the tasks were launched to all workers for high-volume data collection.

The pre-launch process included the following stages:

- **Initial task and guideline design**: Initial adaptation of a clinical task to a labeling task
- **Task and guideline iteration**: A multi-step process to surface and address ambiguities in the tasks and instructions, using a small group of trusted workers
- **Worker onboarding**: Training a larger pool of workers on the updated guidelines, and validating that they are able to perform the task well

In contrast to these pre-launch practices, post-launch quality-related checks were more variable. One of the four teams adopted automated monitoring of performance during the task, comparable to practices in commodity crowdsourcing (Le et al. 2010; Checco et al. 2020). The other teams relied on spot checks to examine answers to individual tasks.

Initial Task and Guideline Design

Medical domain experts and AI practitioners co-led the designing of labeling tasks and instructional guidelines. This process typically involved an engineer and a clinical specialist. The specialist usually had a background similar to those who would be involved in the labeling, while the engineer understood the requirements of the labels for AI development.

This approach was developed to address two key challenges in creating labeling tasks for medical experts:

- **The gap between clinical and labeling contexts**: In a clinical setting, an expert can directly examine the patient, as well as access their full medical record. When labeling, a worker may have access to only a single photo or scan. It would be "clinically unacceptable" (P7) to make a diagnosis based on such limited data, although doing so may be valuable for other contexts, such as screening.
- **The gap between experts' intuition and the structured data required for AI development**: Clinical specialists described how experts can develop an initial "gut feeling" from years of experience. According to P7, one expert described diagnosing glaucoma as "I know it when I see it."[1]

Clinical specialists helped translate clinical practices into labeling tasks by participating in the drafting of tasks and instructional guidelines.

To bridge the gap between clinical and labeling contexts, teams noted the importance of carefully wording task prompts. As one clinical specialist (P9) described, "The question can predispose your worker to be highly sensitive or highly specific. Asking, 'How confident are you that [condition] may be present?' could add to unpredictable noise in the data. Rephrasing it to, 'Can [condition] be ruled out?' would be a more optimal way to phrase it (if optimizing for sensitivity)."

[1] A phrase used during the 1964 U.S. Supreme Court case *Jacobellis v. Ohio* to describe the act of defining obscenity

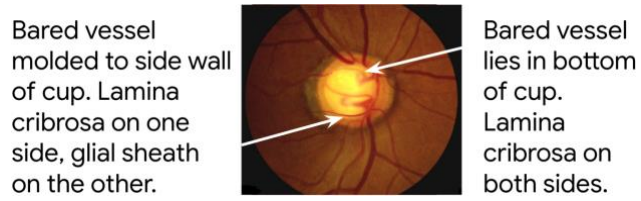


Figure 3: Visual callouts in a guideline document for glaucoma-risk labeling. (Retina image provided via RetinaGallery: Mayo Clinic Jacksonville. Some modifications made for illustrative purposes.)

Teams also described using a combination of "global" and "local" questions in order to bridge the gap between experts' intuition and the need for structured data in AI efforts. In a clinical setting, clinicians do not necessarily perform a detailed assessment of items in a checklist of individual ("local") features. Rather, they often form a clinical assessment based on the overall Gestalt ("global") picture. However, in model development, a checklist is important to ensure consistency among workers and provide data that can be used to explain model output.

For example, a task asked questions about 11 separate "local" optic nerve head features prior to asking the "global" question of overall glaucoma risk. This ensured that workers would methodically examine and consider specific features rather than rely on their initial impression in their assessment. It also allowed the AI team to analyze which features (or combination thereof) correlated with higher glaucoma risk. As P7 described, "The algorithm is doing more than merely approximating physician skills, but finding novel relationships not readily apparent to human beings."

In addition to designing the labeling tasks, clinical specialists also drafted guideline documents for labeling tasks. In commodity crowdsourcing, tasks and instructions are typically simple enough to be presented together in the crowdsourcing platform interface. The medical labeling tasks managed by the interviewed teams, on the other hand, were too complex to fully explain within the labeling interface. Therefore, each team created a guidelines document to accompany each labeling project.

The guidelines created by interviewed teams ranged from two to 87 pages, depending on the number and complexity of subtasks. Figure 3 shows a portion of a glaucoma risk guidelines document.

Task and Guideline Iteration

Teams noted that they could not anticipate all possible ways in which tasks and guidelines could be interpreted before workers started hands-on labeling. Therefore, teams developed a process of testing and revising tasks and guidelines prior to high-volume data collection. This process involved a small set of workers who collaborated closely with the AI and clinical domain specialists who developed the initial tasks and guidelines.

Teams used a variety of mechanisms to gather feedback on tasks and guidelines. One practice was to use clinical experts to review the tasks and guidelines, similar to a heuristic evaluation: "We have three (expert) pathologists who take the guidelines, do some sample tasks, and provide feedback. This is in order to find glaringly obvious mistakes before we start a full production run of label collection" (P3). A second practice was to gather initial feedback on the tasks and guidelines from representative workers, prior to labeling: "We had focus groups to solicit information, asking them what they thought about the prompts, and if there was anything they wished we could change. When you have thousands of cases, you want to make the experience better (for workers)" (P6). A third practice was to pilot the labeling task. Workers labeled a set of sample data, and then shared feedback: "We find holes and gaps in the guidelines. Some (workers) give suggestions on wording and options. They're exercising the whole thing and giving feedback right away" (P1).

This collaborative and iterative feedback process enabled teams to improve the clarity of task instructions. For example, teams described how workers not involved in drafting the guidelines pointed out ambiguous or edge cases not accounted for in the initial draft. As a result, teams documented these ambiguous and edge cases in the guidelines, and clarified how workers should handle them.

Teams also used feedback to improve task prompts. For example, workers sometimes pointed out when answer options were incomplete. For one team, the initial options for a question about the presence of a pathology were "[condition] present" and "[condition] absent." P7 observed that workers struggled with the binary nature of the options: "Sometimes it's impossible to make a yes/no call for any disease, especially based on a single image, and forcing (workers) to do that made them very uncomfortable." As a result, the options were changed from a boolean Yes/No to more nuanced options (e.g. "none," "low," "high," and "likely").

In addition to gathering qualitative feedback from workers, teams also used quantitative analyses to identify potential issues. In particular, several teams analyzed inter-rater agreement (the degree of answer consistency among workers) to assess the effectiveness of guideline iterations. P6 described tracking an inter-rater agreement metric (Krippendorff's alpha) during guideline iteration: "Version 1 of the guidelines goes out with a small pilot number of questions. Then we run an inter-rater agreement analysis. We try to hit a (certain) K-alpha score. If we don't hit it, then we look at the cases that have very high disagreement. We talk to the workers and try to figure out, Where did they misinterpret things? We modify the guidelines, and then we deploy Round 2. Same number of questions, slightly different guidelines. We measure the agreement again. We repeat as needed until we hit that threshold. Once we do, the guidelines are finalized."

Whereas inter-rater agreement is often used in other crowdsourcing contexts to reflect *worker quality* (Daniel et al. 2018) or *adversarial behavior* (Jagabathula et al. 2014), it was used by interviewed teams to implicitly measure *guideline clarity and completeness*. Notably, the interventions used by teams that observed low initial agreement scores were guideline-centric rather than worker-centric. Teams 1) asked workers how they interpreted the questions, and then 2) updated the guidelines accordingly to reduce ambiguity.

Worker Onboarding

After task and guideline iteration, teams conducted onboarding exercises to train a wider set of workers on the task. Training processes consistently involved having workers demonstrate a certain level of proficiency with a pilot task prior to working on high-volume labeling tasks.

Teams used two distinct types of onboarding exercises:

- **Guideline *comprehension* tests**: Tests administered to assess worker attentiveness and understanding of the task instructions.
- **Guideline *application* tests**: Tests administered to assess worker accuracy during labeling tasks.

Some teams reported using guideline comprehension tests outside the context of the labeling platform (e.g. in a Google Form quiz). The intent was to verify that workers had read and understood the task instructions. Teams noted that while attention to detail is necessary to ensure consistency among workers, clinical expertise did not guarantee attentiveness. As P7 stated, "Some highly-trained specialists might think, 'Oh, I know what this is; I don't need to read the guidelines.' Our best-performing (worker) wasn't a glaucoma specialist, but an optometrist. When you are not an expert in the field, you tend to stick to the (provided) guidelines. When you are the expert, you rely on your gut and tend to disregard them." For that reason, the team used guideline comprehension tests to specifically assess attentiveness.

All teams used guideline application tests to onboard workers. Each worker labeled a small "gold" data set, with correct answers determined by established clinical processes. Their answers were evaluated against an answer key. The size and selection of the onboarding tests varied, with the primary consideration being the availability and size of gold data. Workers who didn't achieve the predefined accuracy level for that test often received additional training, and were re-tested on a different data set before being allowed to label larger data sets. Workers who were repeatedly unable to reach the required performance threshold were generally not used in high-volume labeling tasks.

In contrast to screening tests often used in commodity labeling as a *filtering* tool to remove low-performing potential workers, the tests were used as *training* tools for teams to help medical experts successfully apply their expertise in a labeling context. Upon submitting their answers, workers were immediately shown their incorrect answers, the correct answers, and explanations. Teams had workers retake tests until they scored 100%. As P1 described, "The [initial guideline test] is just a warmup." In this way, workers had the chance to improve their understanding of how to apply their clinical expertise to labeling tasks.

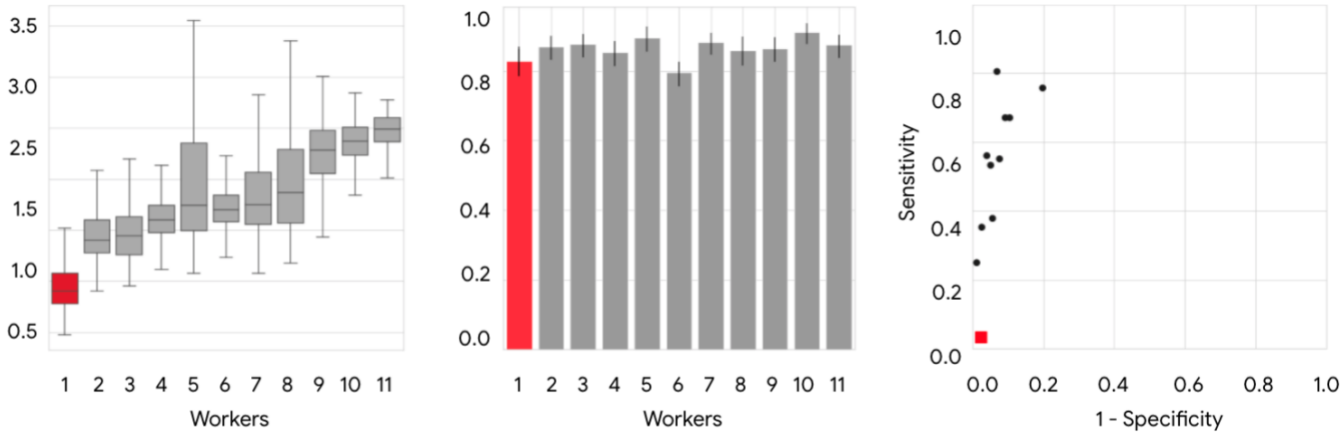
Automated Quality Monitoring Can Surface Distinct Issues during High-Volume Data Collection

In addition to upstream efforts to detect and address quality issues, one interviewed team described a downstream mechanism for assessing the quality of labels after the launch of high-volume data collection. The team achieved this by interspersing a gold data set with known ground truth among other data to be labeled. The task appeared the same to workers regardless of whether they were labeling a case from the gold set or not. However, unlike the onboarding tasks, workers were not aware that their labeling performance was being assessed.

This case of live monitoring identified one worker whose performance was notably faster than that of others (Figure 4a). This worker's overall accuracy on the gold set was not dramatically lower than that of other workers (Figure 4b). However, an analysis of sensitivity and specificity (accuracy among actually-positive and actually-negative cases, respectively) revealed dramatically low sensitivity, almost at zero -- effectively identifying no cases of disease (Figure 4c).

Subsequent investigation indicated that this worker was almost always indicating cases as having no pathology. Because positive cases were relatively infrequent in the test set (which is common for medical image sets), this strategy resulted in high specificity (high accuracy among no-pathology cases). This offset the low sensitivity when measured throughout the data set. Detailed investigation indicated this worker was likely not attending to the task, attempting to maximize the number of cases completed in a short time. The worker was removed from the task, and the labels produced by the worker were removed.

This team observed that the types of quality-related issues they could identify before the launch of high-volume data collection were different from those they could identify after the launch. Before launch, their piloting process with small data sets helped surface issues stemming from unclear or incomplete tasks or guidelines. It was not until after launch, however, that they had the opportunity to employ large data sets to measure worker speed and sensitivity.



(a) Log time on task (b) Accuracy (c) Performance

Figure 4: Automated quality-control monitoring in a medical imaging task. (a) Plot of $\log_{10}$ time spent in a labeling task. Worker 1 (in red) had a dramatically lower median task time, and a very tight distribution, indicating uniformly fast labeling times. (b) Mean accuracy on a gold data set. The accuracy of Worker 1 (in red) was not dramatically lower than that of other workers. Error bars indicate 95% binomial confidence intervals. (c) Performance for each worker plotted as sensitivity over false positive rate (1 - specificity). This display mirrors that used for receiver operating curves used to evaluate categorization performance. The red data point near (0, 0) represents Worker 1.

Discussion

Our study examined how teams developing AI models for medical images ensure high label quality. These interviews indicated that quality-control efforts benefit from a collaborative, iterative process with experts that starts long before the launch of high-volume data collection. Indeed, they represent "methodologies for collecting data from experts" that are missing from many high-stakes AI efforts (Sambasivan et al. 2021). Many quality-control steps occur early in the process, to mitigate the risks and costs of identifying quality issues late in the process. As P3 stated, "Once we get bad data into our system, it's really hard to find it and excise it."

Expert Medical Labeling Faces Challenges Distinct from Other Crowdsourcing Efforts

Our work revealed that teams developing medical imaging AI used a range of practices that, in many respects, mirror iterative practices documented for non-expert crowd work (Le et al. 2010; Gaikwad et al. 2017; Bragg et al. 2018; Manam et al. 2019). However, this work also highlights key themes that, to our knowledge, are not documented in existing literature. In particular, expert labeling practices must manage two key tradeoffs: 1) "Gestalt" expert abilities vs. the need for systematic labeling, and 2) Clinical workflows vs. AI label requirements.

Expert labeling tasks differ from commodity crowd tasks in their reliance on domain expertise. Prior expertise poses a source of variability that must be actively managed in task design. Workers may 1) have strong expectations for how to approach a task that conflicts with the goals of the AI project, 2) be calibrated to different thresholds for detecting signs of disease, or 3) rely on heuristics that are not open to introspection. Our interviews showed that the iterative practices used here explicitly focused on making workers aware of these differences.

Within the domain of medical imaging, AI applications often focus on providing benefit to patients by augmenting or improving existing clinical workflows (Gulshan et al. 2019; Wang et al. 2019; McKinney et al. 2020). As described above, the pre-existing workflows may be ambiguous with respect to how an AI model should handle certain cases, or may specify different diagnostic policies (e.g. favor sensitivity over specificity for screening tasks). Therefore, task design in this expert domain requires explicit reference to existing clinical practice, and extra steps to orient workers to AI-specific labeling requirements. These considerations do not apply when labels are used to train models with new capabilities, without respect to existing workflows.

In addition to highlighting these themes, our study also revealed an emphasis on task and guideline design and worker training over worker monitoring as means to ensure label quality. While both approaches are deployed in commodity crowdsourcing (see Daniel et al. (2018) for an overview, (Gaikwad et al. 2017; Bragg et al. 2018; Manam et al. 2019) for task design, and (Le et al. 2010; Checco et al. 2020) for monitoring), our work suggests further development of task-design methods will have outsized impact on label quality for expert tasks.

Below, we further elaborate on these themes, and consider implications for other expert labeling applications.

Misalignment to Guidelines Is a Primary Error Mode

Many practices in the commodity crowdsourcing literature focus on identifying workers with low ability to perform a task, low conscientiousness, or misaligned incentives (Li 2015; Jagabathula et al. 2014; Checco et al. 2020). However, the most common type of quality issue described in our study was due to guideline misalignment. Interviewed teams developed practices to prevent or correct this misalignment, such as 1) administering tests that enforced careful reading of the guidelines themselves, 2) providing detailed, example-oriented guidelines that clarified how to handle difficult corner cases, and 3) using onboarding exercises that trained workers rather than provide a single pass/fail attempt. These practices both emphasized the importance of the guidelines, and also provided multiple education points for workers (Chi 2006).

Labels used to train medical imaging models must be produced by expert workers with prior experience in a clinical domain. However, variability in clinical labels is well documented (Kalpathy-Cramer et al. 2016; Krause et al. 2018; Schaekermann, Cai, et al. 2020). Therefore, iterative task development must also work to understand and manage variability in workers' experiences and approaches to tasks. For example, clinicians may favor different cutoffs between severity levels in multi-class clinical scales (Chiang 2007; Kalpathy-Cramer et al. 2016; Schaekermann, Cai, et al. 2020). A key role of task guidelines in our study was to provide explicit descriptions to workers of the recommended cutoff points in classification tasks, in order to ensure a consistent approach among workers.

One challenge faced by the clinical specialists we interviewed was contending with ingrained habits of highly experienced and trained clinicians. Specialist clinicians, with many years of experience assessing medical images, may often perform very well at overall diagnosis, but rely on Gestalt mental processes that are not open to introspection ("I know it when I see it"). In order to align workers' labeling approaches, guidelines tended to be explicit, requiring a more deliberative approach (Chi 2006; Kahneman 2011).

Guideline Iteration Aligns Clinical Workflows and AI Labeling Needs

Clinical and AI guidelines may differ in a range of ways. For instance, AI models developed by the teams in our study often assess risks directly from a medical image, whereas related clinical practice may involve other inputs, such as patient metadata or other modalities. As a result, guidelines for the labeling task must be explicitly centered on interpreting the image itself, whereas other clinical guidelines may recommend cross-referencing with other measurements. Further, labels used to train AI models often need comprehensive evaluation of all features in an image. In this way, the model learns to identify all potentially relevant pathologies. In contrast, when there were multiple pathologies, some workers tended to focus on only the most salient pathology. This reflects considerations for in-person clinical practice, where not all pathology is salient to treatment decisions. (A patient presenting with one minor pathology requiring monitoring and one severe pathology requiring immediate treatment would mostly receive recommendations based on the severe pathology.) When models are trained on cases labeled in this way, they may have lower power at detecting mild pathology. This in turn reduces effectiveness in contexts such as screening. Thus, guidelines needed to be refined to explicitly prompt workers to assess lower-severity pathology consistently in all cases.

Another important difference between guidelines for clinical tasks versus labeling tasks for AI development is the need for AI training labels to reflect consistent assessments of the degree of suspicion in images. Doctors vary in their tendency to diagnose cases, particularly in cases with borderline evidence for a condition (Kalpathy-Cramer et al. 2016; Krause et al. 2018). This variation may be valuable in different clinical contexts. For instance, clinicians assessing disease for screening may be more likely to err on the side of identifying disease (high sensitivity), to avoid missing disease that might otherwise go untreated. By contrast, specialists who treat advanced disease may tend to avoid false positives (high specificity), to avoid unnecessary interventions. For training AI models, however, such variability tends to reduce label quality (Guan et al. 2018; Krause et al. 2018).

Iterative guideline development addresses this variability by 1) identifying conditions under which clinicians might give different assessments, 2) clarifying reasons for disagreement, and 3) providing explicit guidance to workers as to how they should reorient themselves for these conditions. The use of explicit callouts of specific image features, as illustrated in Figure 3, is a result of this iterative back-and-forth with workers in understanding sources of disagreement. The partnership between medical experts and AI practitioners in this context is a form of participatory co-design of the labeling experience.

Implications for Expert Labeling Best Practices

The quality-control practices described here may apply broadly to expert labeling domains in general. In particular, two themes stand out: 1) co-designing and refining tasks and guidelines with experts, and 2) training experts to adapt their expertise to labeling. These approaches may ensure labels from highly-trained workers result in high-performing AI models. This may also protect against underutilizing costly and otherwise-capable experts due to miscalibration to guidelines.

Given that worker variability appears to be a prominent source of label variability, methods explicitly focused on calibrating workers against one another may be valuable tools for label quality. Work by Schaekermann, Beaton, et

al. (2020) and Schaekermann, Cai, et al. (2020) indicate that feedback from specialists on sources of disagreement can help improve performance of non-specialist workers on medical imaging tasks. This suggests that as more effective training materials are developed, some projects may be able to use slightly lower-expertise workers. Other methods of task structure, such as comparison-based methods (Kalpathy-Cramer et al. 2016) may further work to calibrate expert workers against each other.

The co-design approach to labeling tasks we describe here addresses the instructional needs of experts who are *contributors* to AI efforts. The approach has some commonality with the development of onboarding materials to address the information needs of clinicians who are *consumers* of an AI-based assistant (Cai et al. 2019). These two studies highlight the importance of close collaboration with experts in a wide range of medical imaging AI efforts.

Our findings also suggest that the commonly-used practice of monitoring quality with the use of "gold" or "silver" datasets may be less central to ensuring quality in many medical imaging tasks. Only one interviewed team reported using this practice, though they did benefit from identifying and removing a low-performing worker from the task. The primary reasons cited by other teams for not engaging in this practice were the difficulty and cost of obtaining a suitable gold set. Since the one team engaging in this practice found a low-performing worker, it is possible these other teams may have had the issue, but not detected it.

Considerations for the effective monitoring of high-volume labeling tasks in medical imaging include the relatively high cost of suitable ground truth for some data sets, and the lower frequency of pathology in many data sets. The low-performing worker described here had very low sensitivity, marking most cases as being without pathology even when pathology was present. In data sets where pathology is rare, this strategy may give the false appearance of relatively high performance. Explicitly measuring sensitivity and specificity (and possibly the related measure of precision) in these gold sets may be needed to quickly identify issues with worker performance.

Implications for Mitigating Label Bias

AI systems in medicine may be subject to many potential forms of bias (Rajkomar et al. 2018), including bias that originates from labels used to train supervised models. As such, practices to alleviate label bias in workers are of high importance.

The iterative processes described here are a natural point for teams to consider potential forms of label bias. They involve domain experts and careful consideration of the AI task in the context of clinical workflows. They surface specific details of the clinical task and can highlight issues such as variability in presentation across patient populations. Iterative guideline development requires measuring variability in grading across workers. Worker variability may be analyzed from the context of different worker cutoffs (discussed above), as well as different clinical approaches. Guidelines that require workers to carefully refer to explicit rules over previously-learned heuristics may mitigate pre-existing worker bias.

With respect to label bias, one drawback to the emphasis on early, iterative processes is that the samples used for guideline development were reported to be small, often on the scale of dozens of examples. Real-world model performance is often expected to show generalization across many different dimensions, such as different patient populations (D'Amour et al. 2020). These small samples are not likely to reflect wider patient populations. Future work should build on these practices through the lens of surfacing different forms of worker bias, and measuring label quality across different dimensions.

Conclusion

In this study, we explored how teams developing AI in medical domains discovered, diagnosed, and learned from apparent data-quality issues throughout the labeling process. We highlighted the unique challenges of bridging the gap between clinical expertise and AI labeling needs. We also articulated a process for bridging this gap via task and guideline design and iteration, as well as worker onboarding. This process emphasizes leveraging experts as not just participants, but also co-creators of the labeling experience.

Quality monitoring during high-volume data collection can surface performance issues that may not arise during onboarding. However, the practices described here occur upstream of high-volume data collection and involve a smaller number of people making low-cost modifications. As such, these practices may deliver an outsized impact on data quality relative to their cost. Further research to quantify label quality may help assess the relative impacts of both iterative guideline development and monitoring on the quality of the resulting labeled data.

Our hope is that AI practitioners working with experts in a range of domains will apply and build on these strategies, to ensure a proactive approach to data quality in a variety of consequential domains.

Acknowledgements

Thank you to our study participants for sharing their stories, artifacts, and insights. Thank you to Richard Gossweiler and Cameron Chen for their valuable comments on the manuscript, and to Michael Sherman for background on quality control mechanisms for expert labeling. We are especially grateful to our expert medical labelers for their ongoing partnership in co-creating labeling practices for medical AI.

References

Barrett, L., and Sherman, M.W. 2019. Improving ML Training Data with Gold-Standard Quality Metrics. In *KDD '19: 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Bragg, J.; Mausam; and Weld, D.S. 2018. Sprout: Crowd-Powered Task Design for Crowdsourcing. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. doi.org/10.1145/3242587.3242598.

Braun, V., and Clarke, V. 2006. Using Thematic Analysis in Psychology. *Qualitative Research In Psychology* 3(2): 77–101.

Braun, V., and Clarke, V. 2019. Reflecting On Reflexive Thematic Analysis. *Qualitative Research in Sport, Exercise and Health* 11(4): 589–597.

Cai, C.J.; Winter, S.; Steiner, D.; Wilcox, L.; and Terry, M. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. Proceedings of the ACM on Human-Computer Interaction CSCW, pp. 1–24, 2019. doi.org/10.1145/3359206.

Challen, R.; Denny, J.; Pitt, M.; Gompels, L.; Edwards, T.; and Tsaneva-Atanasova, K. 2019. Artificial Intelligence, Bias And Clinical Safety. *BMJ Quality & Safety* 28(3): 231–237.

Checco, A.; Bates, J.; and Demartini, G. 2020. Adversarial Attacks on Crowdsourcing Quality Control. *Journal of Artificial Intelligence Research* 67: 375–408. doi.org/10.1613/jair.1.11332.

Chi, M.T.H. 2006. Two Approaches to the Study Of Experts' Characteristics. In *The Cambridge Handbook of Expertise and Expert Performance,* 21–30.

Chiang, M.F. 2007. Interexpert Agreement of Plus Disease Diagnosis in Retinopathy of Prematurity. *Archives of Ophthalmology* 7: 875. doi.org/10.1001/archopht.125.7.875.

D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M.D.; et al. 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning. ArXiv preprint arXiv:2011.03395 [cs.LG].

Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; and Allahbakhsh, M. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Computing Surveys (CSUR)* 51(1): 1–40.

Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; and Thrun, S. 2017. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature 7639*, 115–118. 2017. doi.org/10.1038/nature21056.

Gaikwad, S.; Chhibber, N.; Sehgal, V.; Ballav, A.; Mullings, C.; Nasser, A.; Richmond-Fuller, A.; Gilbee, A.; Gamage, D.; Whiting, M.; et al. 2017. Prototype Tasks: Improving Crowdsourcing Results through Rapid, Iterative Task Design. ArXiv preprint arXiv:1707.05645 [cs.HC].

Guan, M.; Gulshan, V.; Dai, A.; and Hinton, G. 2018. Who Said What: Modeling Individual Labelers Improves Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; Kim, R.; Raman, R.; Nelson, P.C.; Mega, J.L.; and Webster, D.R. 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA: the journal of the American Medical Association* 316(22): 2402–2410.

Gulshan, V.; Rajan, R.P.; Widner, K.; Wu, D.; Wubbels, P.; Rhodes, T.; Whitehouse, K.; Coram, M.; Corrado, G.; Ramasamy, K.; Raman, R.; Peng, L.; and Webster, D.R. 2019. Performance of a Deep-Learning Algorithm vs Manual Grading for Detecting Diabetic Retinopathy in India. *JAMA Ophthalmology*. doi.org/10.1001/jamaophthalmol.2019.2004.

Heer, J., and Bostock, M. 2010. Crowdsourcing Graphical Perception. In *Proceedings of the 28th International Conference on Human Factors In Computing Systems - CHI '10*. doi.org/10.1145/1753326.1753357.

Hsu, J.; Phene, S.; Mitani, A.; Luo, J.; Hammel, N.; Krause, J.; and Sayres, R. 2020. Improving Medical Annotation Quality to Decrease Labeling Burden Using Stratified Noisy Cross-Validation. ArXiv preprint arXiv:2009.10858.

Huang, S.-W., and Fu, W.-T. 2013. Enhancing Reliability Using Peer Consistency Evaluation. In *Human Computation. Proceedings of the 2013 Conference On Computer Supported Cooperative Work - CSCW '13*. doi.org/10.1145/2441776.2441847.

Hutson, M.; Kanzheleva, O.; Taggart, C.; Campana, B.; and Duong, Q. 2019. Quality Control Challenges in Crowdsourcing Medical Labeling. In *Proceedings of Data Collection, Curation, and Labeling for Mining and Learning Workshop at KDD '19*.

Ipeirotis, P.G.; Provost, F.; and Wang, J. 2010. Quality Management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*. doi.org/10.1145/1837885.1837906.

Jagabathula, S.; Subramanian, L.; and Venkataraman, A. 2014. Reputation-Based Worker Filtering in Crowdsourcing. In *Advances in Neural Information Processing Systems*, 2492–2500.

Jiang, H., and Nachum, O. 2020. Identifying and Correcting Label Bias in Machine Learning. In *International Conference on Artificial Intelligence and Statistics*, 702–712.

Kahneman, D. 2011. *Thinking, Fast and Slow*. New York: Macmillan.

Kalpathy-Cramer, J.; Campbell, J.P.; Erdogmus, D.; Tian, P.; Kedarisetti, D.; Moleta, C.; Reynolds, J.D.; Hutcheson, K.; Shapiro, M.J.; Repka, M.X.; Ferrone, P.; Drenser, K.; Horowitz, J.; Sonmez, K.; Swan, R.; Ostmo, S.; Jonas, K.E.; Chan, R.V.P.; Chiang, M.F.; and Imaging and Informatics in Retinopathy of Prematurity Research Consortium. 2016. Plus Disease in Retinopathy of Prematurity: Improving Diagnosis by Ranking Disease Severity and Using Quantitative Image Analysis. *Ophthalmology* 123(11): 2345–2351.

Krause, J.; Gulshan, V.; Rahimy, E.; Karth, P.; Widner, K.; Corrado, G.S.; Peng, L.; and Webster, D.R. 2018. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology* 125(8): 1264–1272.

Le, J.; Edmonds, A.; Hester, V.; and Biewald, L. 2010. Ensuring Quality in Crowdsourced Search Relevance Evaluation: The Effects of Training Question Distribution.

In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, 22–32.

Liu, Y.; Jain, A.; Eng, C.; Way, D.H.; Lee, K.; Bui, P.; Kanada, K.; de Oliveira Marinho, G.; Gallegos, J.; Gabriele, S.; Gupta, V.; Singh, N.; Natarajan, V.; Hofmann-Wellenhof, R.; Corrado, G.S.; Peng, L.H.; Webster, D.R.; Ai, D.; Huang, S.J.; Liu, Y.; Dunn, R.C.; and Coz, D. 2020. A Deep Learning System for Differential Diagnosis of Skin Diseases. *Nature Medicine* 26(6): 900–908.

Li, Y. 2015. Crowdsourcing with Worker Quality Control. Master's thesis, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong.

Manam, V.K.C.; Chaithanya Manam, V.K.; Jampani, D.; Zaim, M.; Wu, M.-H.; and Quinn, A.J. 2019. TaskMate: A Mechanism to Improve the Quality of Instructions in Crowdsourcing. In *Companion Proceedings of The 2019 World Wide Web Conference*, 1121-1130. doi.org/10.1145/3308560.3317081.

McKinney, S.M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; Back, T.; Chesus, M.; Corrado, G.C.; Darzi, A.; Etemadi, M.; Garcia-Vicente, F.; Gilbert, F.J.; Halling-Brown, M.; Hassabis, D.; Jansen, S.; Karthikesalingam, A.; Kelly, C.J.; King, D.; Ledsam, J.R.; Melnick, D.; Mostofi, H.; Peng, L.; Reicher, J.J.; Romera-Paredes, B.; Sidebottom, R.; Suleyman, M.; Tse, D.; Young, K.C.; De Fauw, J.; and Shetty, S. 2020. International Evaluation of an AI system for Breast Cancer Screening. *Nature* 577(7788): 89–94.

Nagpal, K.; Foote, D.; Tan, F.; Liu, Y.; Chen, P.-H.C.; Steiner, D.F.; Manoj, N.; Olson, N.; Smith, J.L.; Mohtashamian, A.; Peterson, B.; Amin, M.B.; Evans, A.J.; Sweet, J.W.; Cheung, C.; van der Kwast, T.; Sangoi, A.R.; Zhou, M.; Allan, R.; Humphrey, P.A.; Hipp, J.D.; Gadepalli, K.; Corrado, G.S.; Peng, L.H.; Stumpe, M.C.; and Mermel, C.H. 2020. Development and Validation of a Deep Learning Algorithm for Gleason Grading of Prostate Cancer from Biopsy Specimens. JAMA Oncology. doi.org/10.1001/jamaoncol.2020.2485.

Ørting, S.; Doyle, A.; van Hilten, A.; Hirth, M.; Inel, O.; Madan, C.R.; Mavridis, P.; Spiers, H.; and Cheplygina, V. 2020. A Survey of Crowdsourcing in Medical Image Analysis. *Human Computation* 7(1):, 1-26. doi.org/10.15346/hc.v7i1.1

Phene, S.; Dunn, R.C.; Hammel, N.; Liu, Y.; Krause, J.; Kitade, N.; Schaekermann, M.; Sayres, R.; Wu, D.J.; Bora, A.; Semturs, C.; Misra, A.; Huang, A.E.; Spitze, A.; Medeiros, F.A.; Maa, A.Y.; Gandhi, M.; Corrado, G.S.; Peng, L.; and Webster, D.R. 2019. Deep Learning and Glaucoma Specialists: The Relative Importance of Optic Disc Features to Predict Glaucoma Referral in Fundus Photographs. *Ophthalmology* 126(12): 1627–1639.

Pradhan, V.K., and Lease, M. 2018. In Search of Ambiguity: A Three-Stage Workflow Design to Clarify Annotation Guidelines for Crowd Workers. Technical Report TR-18-03. Austin, TX: University of Texas at Austin, Department of Computer Science.

Prum, B.E., Jr.; Lim, M.C.; Mansberger, S.L.; Stein, J.D.; Moroi, S.E.; Gedde, S.J.; Herndon, L.W., Jr.; Rosenberg, L.F.; and Williams, R.D. 2016. Primary Open-Angle Glaucoma Suspect Preferred Practice Pattern®. *Ophthalmology* 123(1): 112–151.

Rajkomar, A.; Hardt, M.; Howell, M.D.; Corrado, G.; and Chin, M.H. 2018. Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine* 169(12):866–872. doi:10.7326/M18-1990.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D; Paritosh, P.; and Aroyo, L. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *CHI Conference on Human Factors in Computing Systems*. doi.org/10.1145/3411764.3445518.

Schaekermann, M.; Beaton, G.; Sanoubari, E.; Lim, A.; Larson, K.; and Law, E. 2020. Ambiguity-Aware AI Assistants for Medical Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. doi.org/10.1145/3313831.3376506.

Schaekermann, M.; Cai, C.J.; Huang, A.E.; and Sayres, R. 2020. Expert Discussions Improve Comprehension of Difficult Cases in Medical Image Assessment. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. doi.org/10.1145/3313831.3376290.

Ting, D.S.W.; Cheung, C.Y.-L.; Lim, G.; Tan, G.S.W.; Quang, N.D.; Gan, A.; Hamzah, H.; Garcia-Franco, R.; San Yeo, I.Y.; Lee, S.Y.; Wong, E.Y.M.; Sabanayagam, C.; Baskaran, M.; Ibrahim, F.; Tan, N.C.; Finkelstein, E.A.; Lamoureux, E.L.; Wong, I.Y.; Bressler, N.M.; Sivaprasad, S.; Varma, R.; Jonas, J.B.; He, M.G.; Cheng, C.-Y.; Cheung, G.C.M.; Aung, T.; Hsu, W.; Lee, M.L;. and Wong, T.Y. 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: The Journal of the American Medical Association* 318(22): 2211–2223.

Waggoner, B., and Chen, Y. 2014. Output Agreement Mechanisms and Common Knowledge. In *Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing*.

Wang, P.; Berzin, T.M.; Glissen Brown, J.R.; Bharadwaj, S.; Becq, A.; Xiao, X.; Liu, P.; Li, L.; Song, Y.; Zhang, D.; Li, Y.; Xu, G.; Tu, M.; and Liu, X. 2019. Real-Time Automatic Detection System Increases Colonoscopic Polyp and Adenoma Detection Rates: A Prospective Randomised Controlled Study. *Gut* 68(10): 1813–1819.

Zou, J., and Schiebinger, L. 2018. AI Can Be Sexist and Racist — It's Time to Make It Fair. *Nature* 7714, 324–326. doi.org/10.1038/d41586-018-05707-8.