Iterative Quality Control Strategies for Expert Medical Image Labeling

Beverly Freeman

IEEE webinar, April 2022

Google Health

IEEE
Advancing Technology for Humanity

Agenda

About me

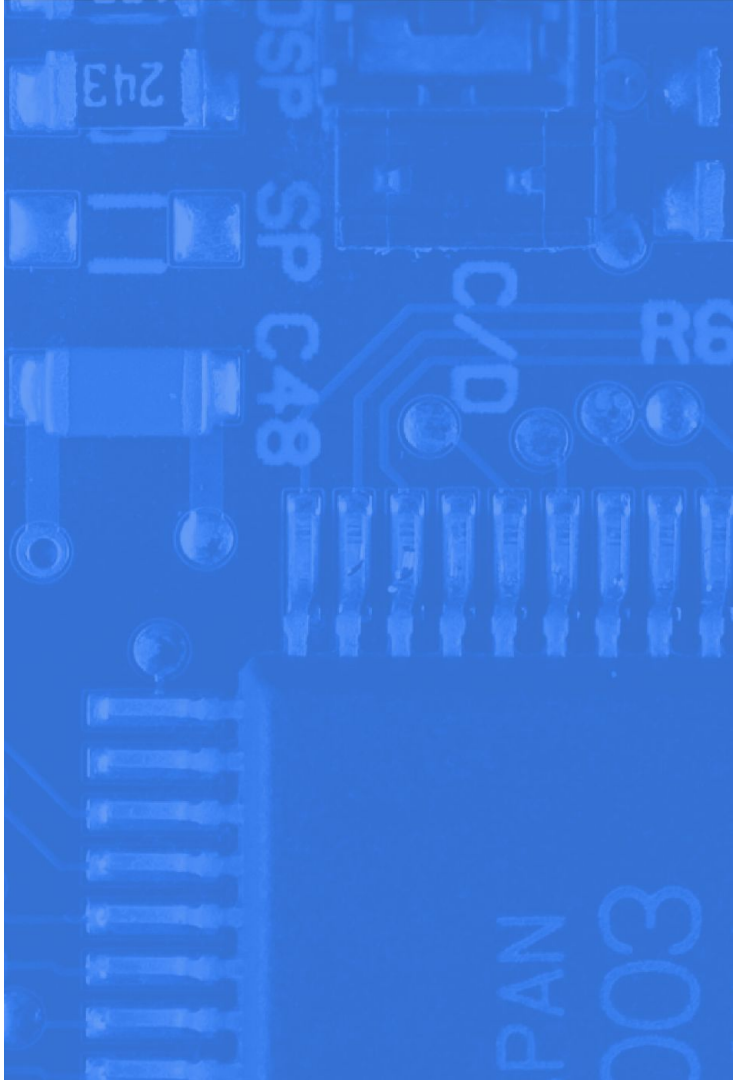
Why is **data quality** important?

What have we **learned**?

- Challenges of expert labeling
- Strategies for expert labeling
- 2 key perspective shifts

1

About me



About me

Sr. UX Researcher @ Google

Previously: eBay, Paypal, Intuit, Oracle

MS Human Factors, SJSU

Strengths: Asking Qs, synthesizing data

Our work

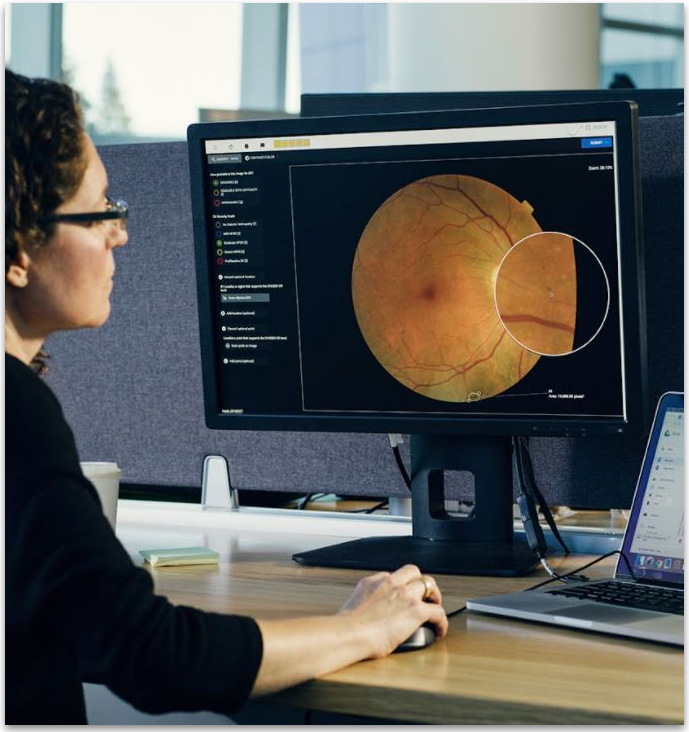
Imaging & Diagnostics | Genomics | Publications

AI-enabled imaging and diagnostics previously thought impossible

In partnership with healthcare organizations globally, we're researching robust new AI-enabled tools focused on diagnostics to assist clinicians. Drawing from diverse datasets, high-quality labels, and state-of-the-art deep learning techniques, we are making models that we hope will eventually support medical specialists in diagnosing disease. We're excited to further develop this research towards new frontiers—and to demonstrate that AI has the ability to enable novel, transformative diagnostics.

https://health.google/

Our work



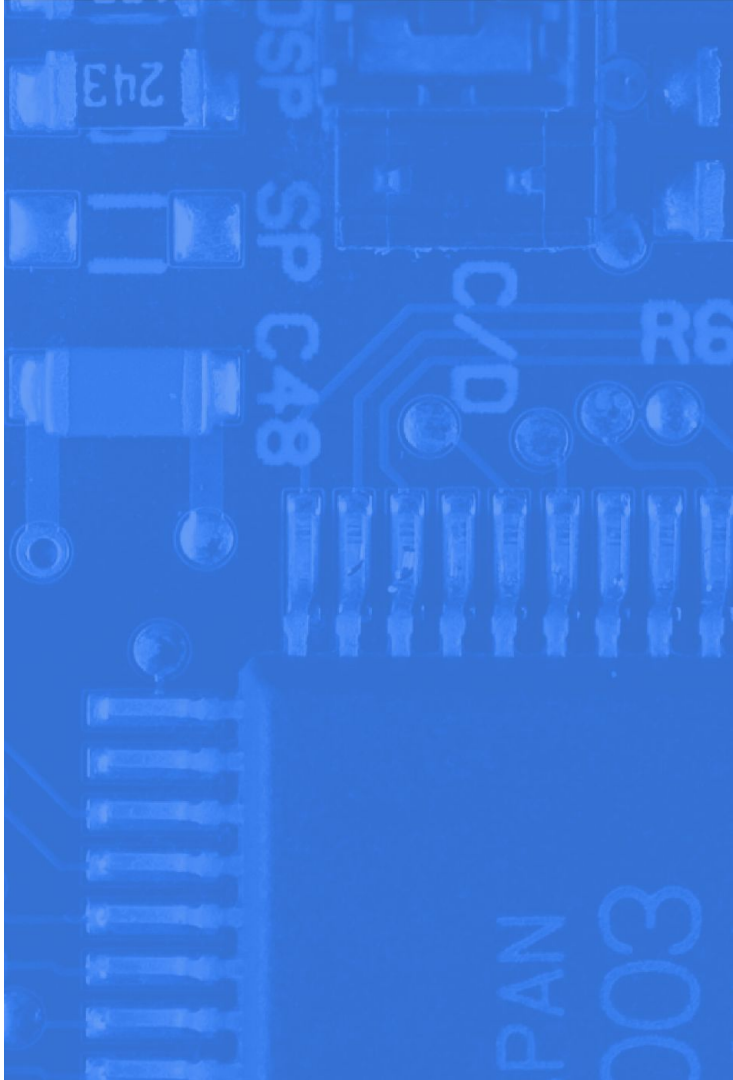
Medical experts label health data

Labels train AI to detect disease

- Dermatology
- Ophthalmology
- Pathology
- Radiology

2

Why is data quality important?



“Garbage in, garbage out”

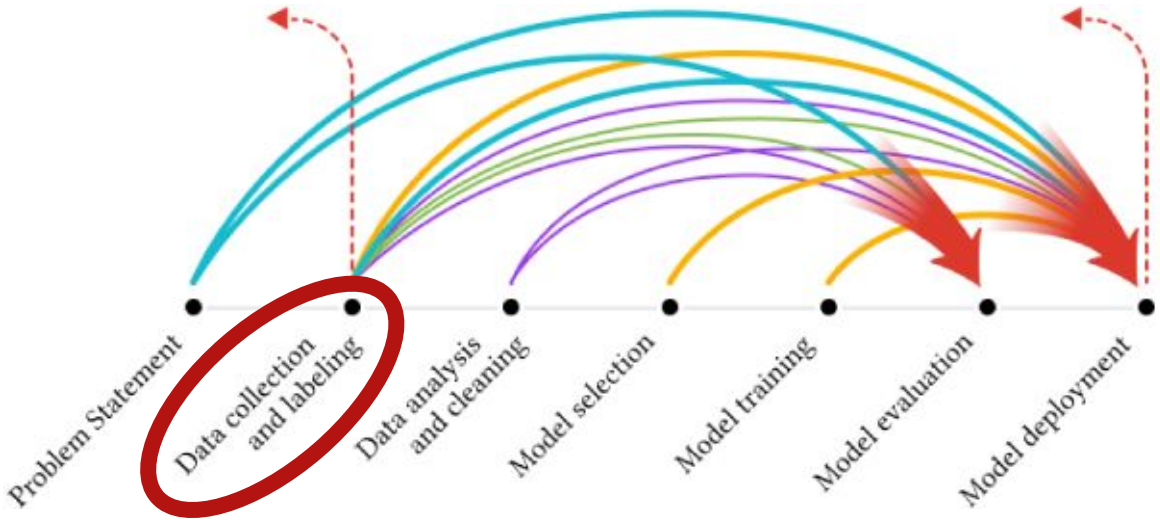
Low-quality **labels** → Lower-performing **models**

Bad data in medical imaging: Impact

High $$$ of medical experts

Patient harm (e.g. false negatives / positives)

Bad data & “data cascades”

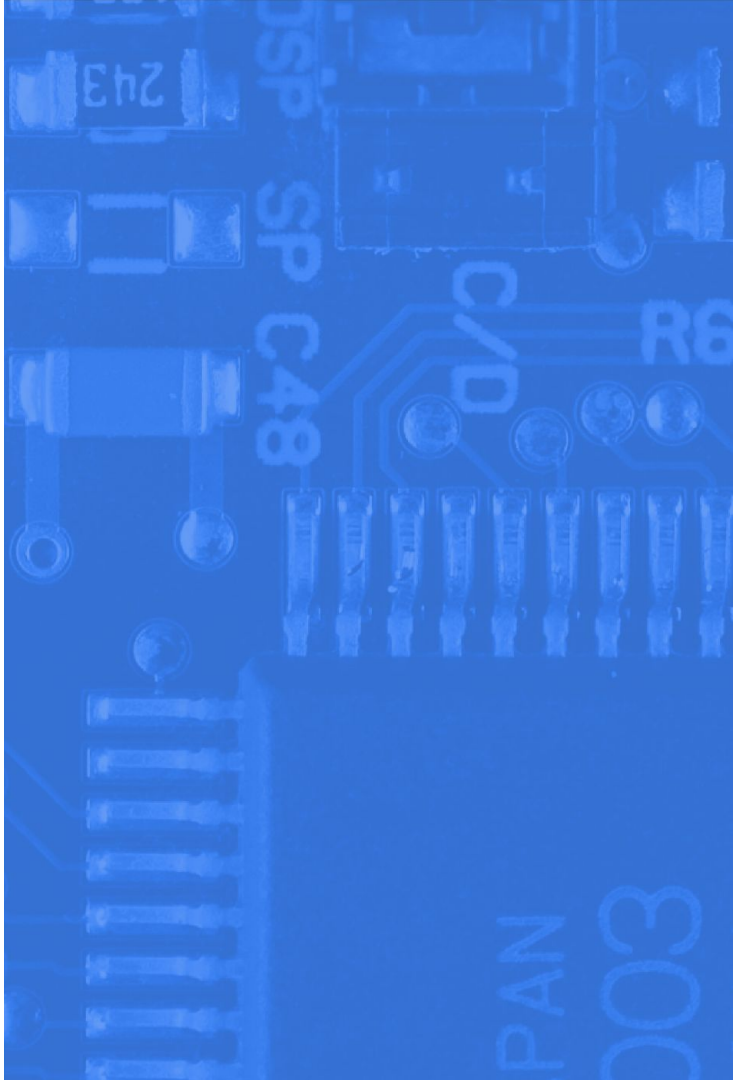


“Compounding events causing negative, downstream effects from data issues, resulting in technical debt over time.”

Sambasivan et al. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. CHI 2021. https://doi.org/10.1145/3411764.3445518

3

What have we learned?



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

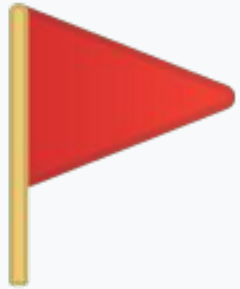
Freeman et al. "Iterative Quality Control Strategies for Expert Medical Image Labeling." AAAI HCOMP 2021.
https://ojs.aaai.org/index.php/HCOMP/article/view/18940

Learning from teams' stories

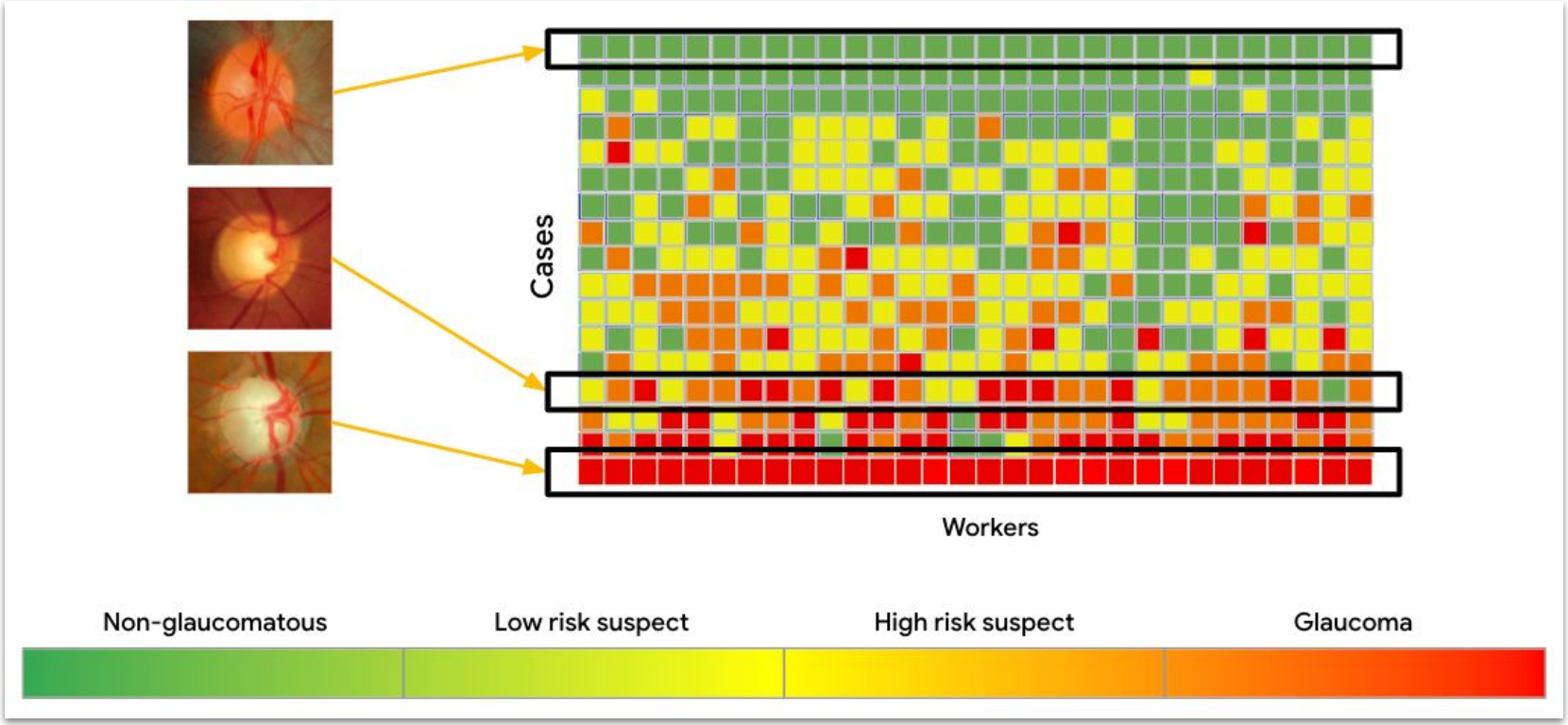
"Describe a time when you had concerns about label quality."

How did it come to your **attention**?	What did you **do**?	How has your approach **changed**?
↓	↓	↓
🔎 Signals	🛠️ Interventions	💡 Strategies

Expert medical image labeling: Key challenges



Inter-worker variability



“It turned out that half of the (workers) interpreted a question **one way**, and half of them **another way**.

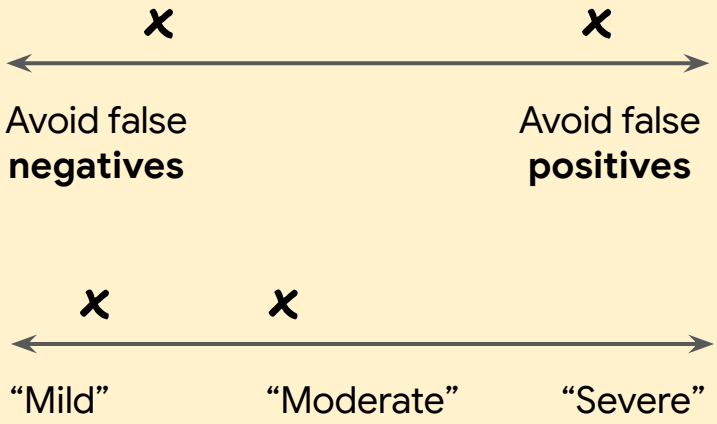
At least half of our data was **useless**.”

Reason #1: Differences among workers

Experience / training

“You end up doing what you learned during training.”

Individual tendencies



Reason #2: Gap between contexts

Clinical 🩺

Data abundance
Medical history, multiple images, etc.

Expert Gestalt intuition
“Glaucoma: I know it when I see it.”

Focus on most salient pathology
Treat the most severe issue

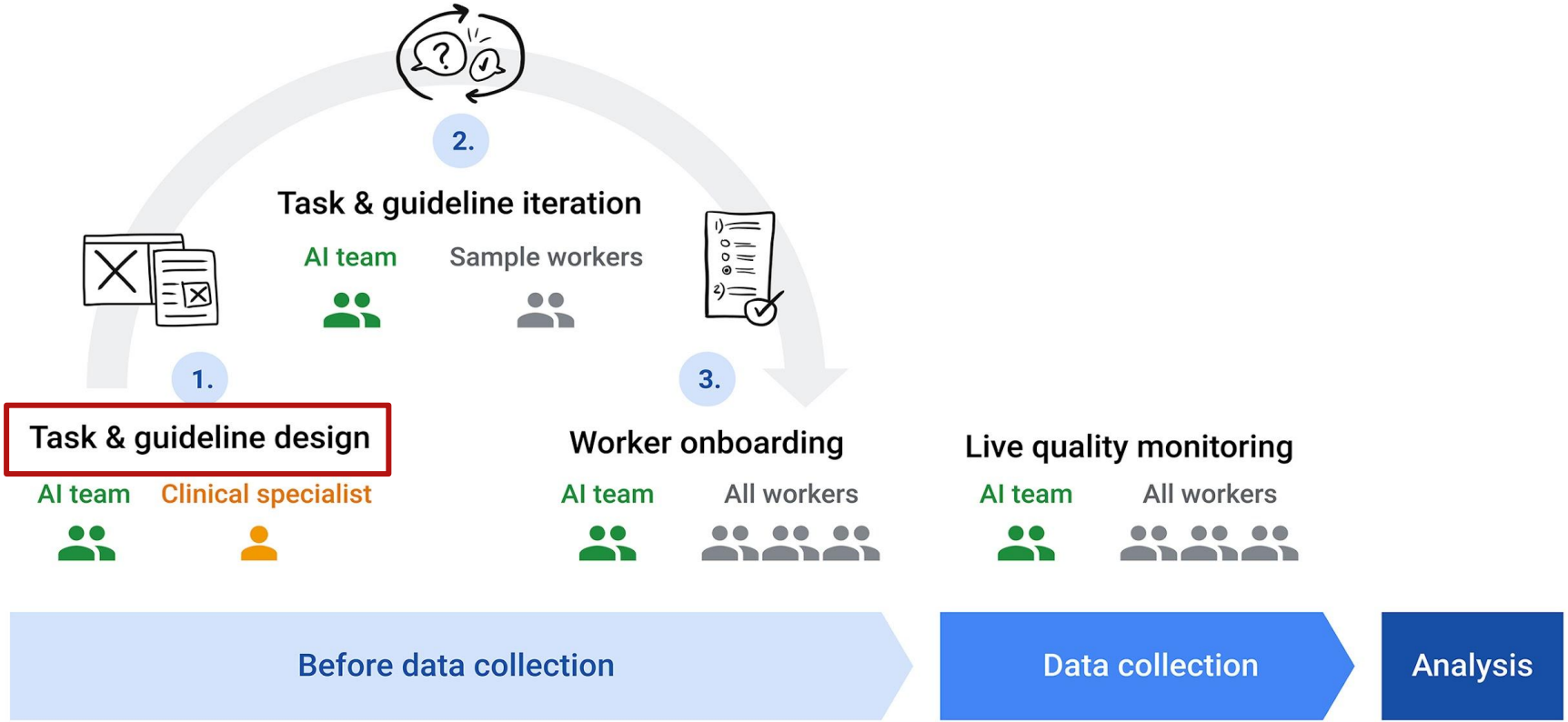
Labeling for AI 🤖

Data scarcity
Single image

Need for structured data
For consistency, explainability

Focus on all pathologies
Screen for early mild pathology

Expert medical image labeling: Strategies


2.
Task & guideline iteration
AI team
Sample workers
1.
Task & guideline design
AI team
Clinical specialist
3.
Worker onboarding
AI team
All workers
Live quality monitoring
AI team
All workers
Before data collection
Data collection
Analysis

1: Task & guideline design

Task design 📝

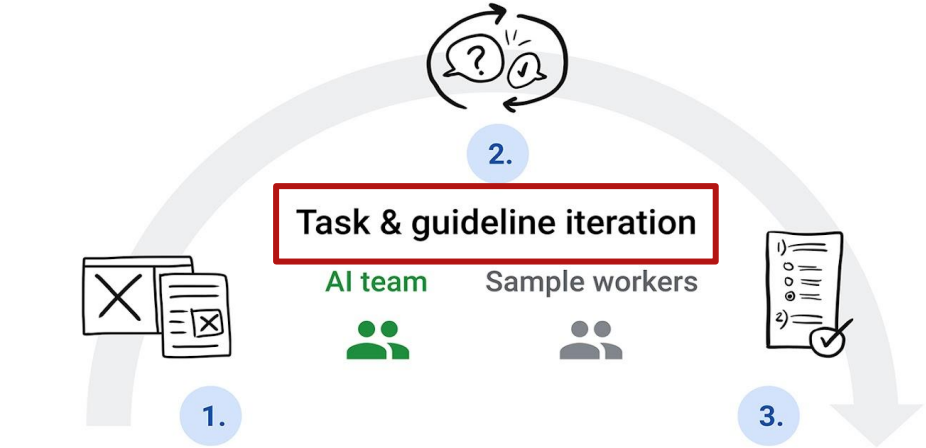
“Global” + “Local” questions
e.g. Checklist + overall risk

Careful wording
“Can [condition] be ruled out?” vs. *“Is [condition] present?”*

Guideline design 📋

Provide context (prime workers)
e.g. “Imagine these are from a screening clinic where [disease] is relatively common.”

Reduce ambiguity with examples
e.g. images of risk indicators



2.

Task & guideline iteration

AI team Sample workers

1.

3.

Task & guideline design

AI team Clinical specialist

Worker onboarding

AI team All workers

Live quality monitoring

AI team All workers

Before data collection

Data collection

Analysis

2: Task & guideline iteration

Qualitative methods 💬

Expert reviews

“To find glaringly obvious mistakes before we start a full production run of label collection.”

Worker feedback sessions

“We had focus groups, asking if there was anything we should change.”

Dry run with sample tasks

“We find holes in the guidelines. They’re exercising the whole thing and giving feedback right away.”

2: Task & guideline iteration

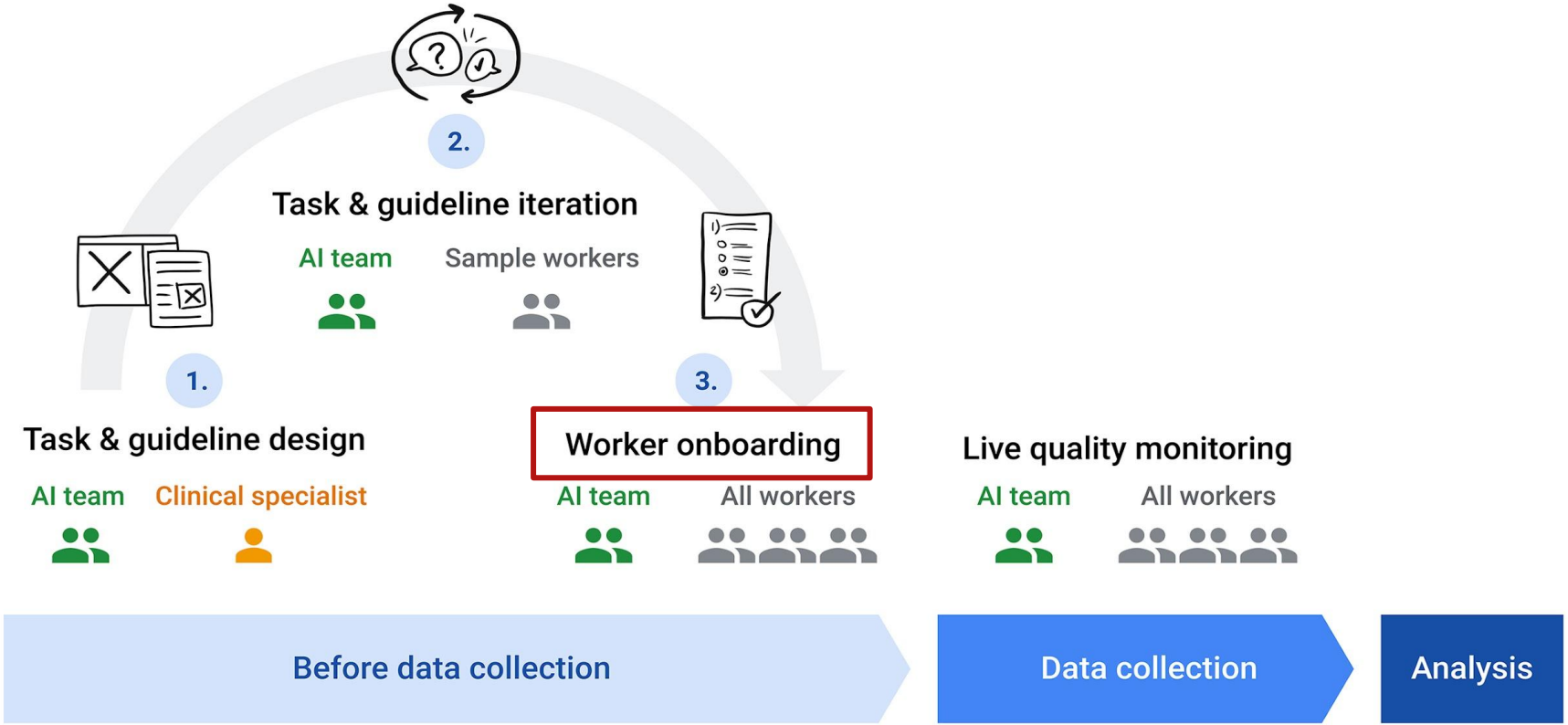
Quantitative methods 📊

Inter-rater agreement

*"We look at cases with high disagreement. We **talk to the workers** and try to figure out: What did they misinterpret?*

*We **modify the guidelines**, and then deploy Round 2. We **repeat as needed** until we hit (a preset) threshold."*

For improving tasks & **guidelines**... NOT for assessing **workers**!


2.
Task & guideline iteration
AI team
Sample workers
1.
Task & guideline design
AI team
Clinical specialist
3.
Worker onboarding
AI team
All workers
Live quality monitoring
AI team
All workers
Before data collection
Data collection
Analysis

3: Worker onboarding via tests

Guideline *comprehension*

Answer questions about the guidelines

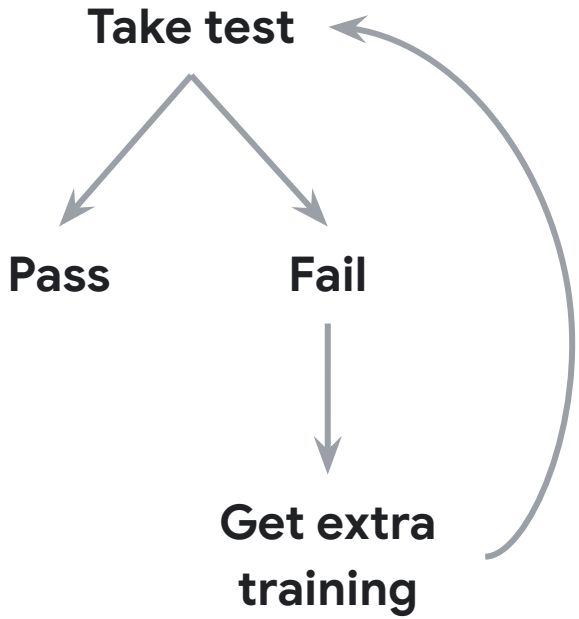
(written knowledge test)

Guideline *application*

Perform sample tasks

(behind-the-wheel driving test)

3: Worker onboarding via tests



For helping workers **succeed** in applying their expertise to labeling... NOT for **filtering out** “bad” workers!

Expert medical image labeling: 2 key perspective shifts

Domain experts
as data **providers**

Domain experts
as **partners**

Measuring quality once large-scale data collection **launches**

Proactively identifying and addressing issues **upstream**

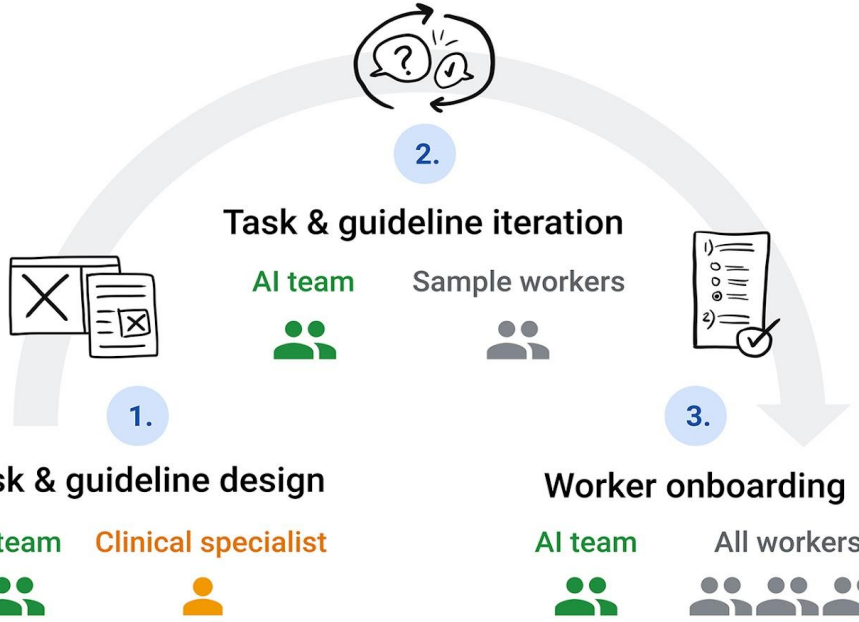
Thank you!

Expert medical labelers

Study participants

Co-authors

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



Thank You!

Beverly Freeman

beverlyf@google.com

Before data collection | Data collection | Analysis