Human-AI Interaction in the Presence of Ambiguity

From Deliberation-based Labeling to Ambiguity-aware AI

by

Mike Schaekermann

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2020

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Loren Terveen Professor, Department of Computer Science and Engineering, University of Minnesota
Supervisors:	Edith Law Associate Professor, David R. Cheriton School of Computer Science, University of Waterloo
	Kate Larson Professor, David R. Cheriton School of Computer Science, University of Waterloo
Internal Member:	Daniel Vogel Associate Professor, David R. Cheriton School of Computer Science, University of Waterloo
Internal-External Member:	James R. Wallace Associate Professor, School of Public Health and Health Systems, University of Waterloo
Other Member(s):	Oliver Schneider Assistant Professor, Department of Management Sciences, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This dissertation includes first-authored peer-reviewed material that has appeared in conference and journal proceedings published by the Association for Computing Machinery (ACM) and by the Association for Research in Vision and Ophthalmology (ARVO). The ACM's policy on reuse of published materials in a dissertation is as follows[1]:

> *"Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included."*

ARVO's publication license contains the following statement[2]:

> *"The Author(s) shall retain the non-exclusive right to any use of the Work, so long as the Author(s) provide(s) attribution to the place of original publication."*

The following list serves as a declaration of the Versions of Record for works included in this dissertation:

Portions of Chapter 3:

Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 154 (November 2018), 19 pages.

DOI=10.1145/3274423
https://doi.org/10.1145/3274423

Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019. Understanding Expert Disagreement in Medical Data Analysis through Structured Adjudication. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 76 (November 2019), 23 pages.

DOI=10.1145/3359178
https://doi.org/10.1145/3359178

[1] https://authors.acm.org/author-resources/author-rights. Accessed on March 24, 2020.

[2] http://arvojournals.org/DocumentLibrary/ARVOLicensetoPublish.pdf. Accessed on March 24, 2020.

Mike Schaekermann, Naama Hammel, Michael Terry, Tayyeba K. Ali, Yun Liu, Brian Basham, Bilson Campana, William Chen, Ji Xiang, Jonathan Krause, Greg S. Corrado, Lily Peng, Dale R. Webster, Edith Law, and Rory Sayres. Remote Tool-Based Adjudication for Grading Diabetic Retinopathy. *Trans Vis Sci Tech.* 2019; 8(6):40.

DOI=10.1167/tvst.8.6.40
https://doi.org/10.1167/tvst.8.6.40

Portions of Chapter 4:

Mike Schaekermann, Carrie J. Cai, Abigail E. Huang, and Rory Sayres. 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 (CHI '20)*. ACM, New York, NY, USA, Paper 163, 13 pages.

DOI=10.1145/3313831.3376290
https://doi.org/10.1145/3313831.3376290

Portions of Chapter 5:

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, New York, NY, USA, Paper 379, 14 pages.

DOI=10.1145/3313831.3376506
https://doi.org/10.1145/3313831.3376506

Abstract

Ambiguity, the quality of being open to more than one interpretation, permeates our lives. It comes in different forms including linguistic and visual ambiguity, arises for various reasons and gives rise to disagreements among human observers that can be hard or impossible to resolve. As artificial intelligence (AI) is increasingly infused into complex domains of human decision making it is crucial that the underlying AI mechanisms also support a notion of ambiguity. Yet, existing AI approaches typically assume that there is a single correct answer for any given input, lacking mechanisms to incorporate diverse human perspectives in various parts of the AI pipeline, including data labeling, model development and user interface design.

This dissertation aims to shed light on the question of how humans and AI can be effective partners in the presence of ambiguous problems. To address this question, we begin by studying group deliberation as a tool to detect and analyze ambiguous cases in data labeling. We present three case studies that investigate group deliberation in the context of different labeling tasks, data modalities and types of human labeling expertise.

First, we present CrowdDeliberation, an online platform for synchronous group deliberation in novice crowd work, and show how worker deliberation affects resolvability and accuracy in text classification tasks of varying subjectivity. We then translate our findings to the expert domain of medical image classification to demonstrate how imposing additional structure on deliberation arguments can improve the efficiency of the deliberation process without compromising its reliability. Finally, we present CrowdEEG, an online platform for collaborative annotation and deliberation of medical time series data, implementing an asynchronous and highly structured deliberation process. Our findings from an observational study with 36 sleep health professionals help explain how disagreements arise and when they can be resolved through group deliberation.

Beyond investigating group deliberation within data labeling, we also demonstrate how the resulting deliberation data can be used to support both human and artificial intelligence. To this end, we first present results from a controlled experiment with ten medical generalists, suggesting that reading deliberation data from medical specialists significantly improves generalists' comprehension and diagnostic accuracy on difficult patient cases. Second, we leverage deliberation data to simulate and investigate AI assistants that not only highlight ambiguous cases, but also explain the underlying sources of ambiguity to end users in human-interpretable terms. We provide evidence suggesting that this form of ambiguity-aware AI can help end users to triage and trust AI-provided data classifications.

We conclude by outlining the main contributions of this dissertation and directions for future research.

Acknowledgements

This thesis is dedicated to an indispensable bunch of awesome individuals:

- all the amazing folks at the Human-Computer Interaction lab who have made this PhD a fun, crazy and rewarding experience for me — I'm proud to say that some of you have become friends for life;
- my parents Kerstin and Horst, my sister Helen and my brother Raoul, who have always sent their love and support from across the Atlantic Ocean;
- Ellie who has not only been the best and most loving partner in crime I could wish for, but who has also co-authored the paper forming the basis for Chapter 5;
- my advisors Edith Law and Kate Larson, who helped me navigate this unknown terrain with fantastic guidance and a never-ceasing smile on their faces;
- my committee members Loren Terveen, Daniel Vogel, James Wallace and Oliver Schneider, for providing valuable feedback in refining this thesis;
- Rory Sayres and Michael Terry, who have been wonderful mentors to me at Google;
- Andrew Lim and Farrah Mateen, our incredible collaborators who have built the bridges needed to conduct research with medical experts;
- Rui de Sousa for his invaluable help in recruiting expert participants from all over the world;
- all of our study participants and human annotators enlisted through hospitals and crowdsourcing platforms;
- Dr. Rajiv Raman for permission to use the fundus photograph shown in Figure 3.7;
- many other friends, colleagues and collaborators at the University of Waterloo and at Google without whom this research would not have been possible: William Callaghan, Alex Williams, Jessy Ceha, Jay Henderson, Kyle Robinson, Alexandra Vtyurina, Bahareh Sarrafzadeh, Graeme Beaton, Joslin Goh, Robin Cohen, Carrie Cai, Naama Hammel, Sonia Phene, Yun Liu, Abigail Huang, Abi Jones, Kasumi Widner, Cristhian Cruz, Quang Duong, Olga Kanzheleva, Lily Peng, Dale Webster.

This research was supported by several grants including NSERC CHRP (CHRP 478468-15), CIHR CHRP (CPG-140200) and a Google PhD Fellowship. Thank you for making graduate school possible!

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

> Man must not attempt to dispel the ambiguity of his being but, on the contrary, accept the task of realizing it.
>
> — Simone de Beauvoir, *The Ethics of Ambiguity*

Major advances in artificial intelligence (AI) have enabled a new era of decision support, shifting from relatively constrained and well-defined problems to more complex and nuanced domains. Many complex domains essential to our modern society, including journalism, criminal justice, healthcare or science, are full of problems for which it is difficult or impossible to find a single correct answer to a given problem. For example, jurors can arrive at different verdicts in light of the same evidence, or doctors can form conflicting diagnoses looking at the same medical image. At the heart of many of these disagreements lies *ambiguity*, the quality of being open to more than one interpretation.[1] An acceptable approach for decision making in the presence of ambiguous problems often requires elicitation and synthesis of diverse human perspectives to inform more well-defined, complete, balanced or ethical decisions.

However, existing AI approaches typically assume that there is a single correct answer for any given input, lacking mechanisms to incorporate diverse human perspectives. This assumption is prevalent in various steps of the AI pipeline, including data labeling, model development and user interface (UI) design. In data labeling, it has motivated the notion that inter-rater disagreement is mere "noise in the signal" originating from human mistakes, and has thus given rise to post-processing techniques that eliminate disagreement. In model

[1] https://en.oxforddictionaries.com/definition/ambiguity (accessed on 12 May 2020)

development, it has favoured learning techniques and evaluation metrics that support no more than a single correct output for any given input. In UI design for AI applications, it has hampered the exploration of interfaces that communicate and explain ambiguous cases to end users.

This thesis aims to shed light on the question of how humans and AI can be effective partners in the presence of ambiguous problems. We take the stance that inter-rater disagreement carries valuable information that can and should be captured to better understand the structure of ambiguous problems.

One primary contribution of this work is to introduce and study group deliberation in the context of data labeling, as a tool to detect and analyze ambiguous cases, and to either resolve inter-rater disagreements organically through deliberation, or otherwise mark them as irresolvable. In support of this contribution, we present insights from three case studies about collaborative online platforms that all integrate group deliberation into data labeling workflows, in the context of different data modalities, labeling tasks and types of human labeler expertise. We further demonstrate how the resulting deliberation data can be used not only to enable other human labelers to calibrate their own reasoning for better comprehension and accuracy on difficult cases, but also to investigate how AI can effectively communicate ambiguity to end users.

The remainder of this chapter introduces the central thesis of this dissertation and provides an overview of our main contributions as well as the scope of the research conducted to support our thesis.

1.1 Thesis Statement

This dissertation aims to defend the following thesis statement:

> *Ambiguity, the quality of being open to more than one interpretation, permeates our lives. It can take various forms including linguistic and visual ambiguity, arise for various reasons including heterogeneous data or vague definitions, and give rise to inter-rater disagreements that can be hard or impossible to resolve. Human and artificial intelligence can benefit from novel methods that aim to detect and explain instances of ambiguity. The expected advantages of such methods are a better understanding of why disagreement arises and when it can be resolved, as well as better approaches for handling ambiguity in human decision making—both unassisted and when assisted by artificial intelligence.*

To defend this statement, the remainder of this dissertation addresses the following research questions:

1. How can we capture the structure of **ambiguous problems in data labeling**?
   (a) How can **group deliberation** be used to analyze ambiguity in data labeling?
   (b) How can group deliberation be integrated into **non-expert crowdsourcing**?
   (c) How can group deliberation be integrated into **expert labeling tasks**?
2. How can we leverage deliberation data to improve **human decision making**?
   (a) How can we use **deliberation data as training material** for human labelers?
   (b) How do labelers **perceive** deliberation data as a form of training feedback?
   (c) How does reading deliberation data affect **labeling decisions** for future cases?
3. How can we leverage deliberation data to **communicate ambiguity in AI output**?
   (a) How can we simulate an AI that **detects and explains ambiguous data**?
   (b) How do end users **perceive** an ambiguity-aware AI assistant?
   (c) How does an ambiguity-aware AI assistant affect end users' **labeling decisions**?

Next, we summarize the research contributions made through the work presented in this dissertation.

1.2 Research Contributions

In this dissertation, we make the following research contributions:

1. We introduce group deliberation as a tool to detect and explain ambiguity in data labeling. To this end, we implement and investigate deliberation workflows within three different contexts:

   (a) First, we present **Crowd Deliberation**, an online platform for synchronous group deliberation in the context of non-expert crowd work. We study how worker deliberation affects resolvability and accuracy using case studies with both an objective and a subjective task, involving 316 crowd workers.

   (b) Next, we translate our findings to the **expert domain** of medical image classification to study asynchronous group deliberation among medical specialists. We present findings from an experiment with 15 retina specialists showing that structuring deliberation arguments around a set of low-level decision criteria improves the efficiency of the deliberation process.

   (c) Finally, we build on the findings from the first two studies to create **CrowdEEG**, an online platform for collaborative annotation and deliberation of medical time series data. CrowdEEG implements an asynchronous and highly structured deliberation process. We present findings from an observational study with 36 sleep technologists about factors contributing to disagreement and resolvability.

2. We demonstrate how deliberation data can be used as training material to **calibrate human expert labelers**. To this end, we present results from a controlled experiment with ten medical generalists. Our results show that reading deliberation data produced by specialists substantially improves generalists' comprehension and diagnostic accuracy on difficult patient cases.

3. We leverage deliberation data to simulate and **investigate ambiguity-aware AI**, i.e., AI that not only detects, but also explains ambiguous data classifications. We present results from a controlled experiment with twelve sleep technologists suggesting that ambiguity-aware AI can improve the ability of end users to triage and trust AI-provided output, and that the relevance of AI-provided ambiguity explanations is crucial for expert end users to accurately disambiguate difficult cases.

4. We **publish two novel datasets** containing deliberation data in the context of both non-expert and expert classification tasks.

1.3 Research Scope

Here, we define the scope of research conducted as part of this dissertation. We clarify the problem we aim to solve, the tasks and domains in which we seek to provide support for the problem, and various aspects of the proposed solutions we investigated.

1.3.1 Tasks

We aim to address the problem of better supporting ambiguity in human-AI collaborative tasks. Specifically, we focus on supervised learning settings for data classification that rely on human-labeled data to train and evaluate AI algorithms. In this context, we are interested in better understanding and supporting scenarios where the same data instance is labeled by more than one human and where there is disagreement among those humans over the correct classification label. In particular, we are interested in instances of disagreement that arise for reasons more profound than simple human mistakes and which are hard or impossible to resolve even when reviewed and discussed collectively as a group. While there exist labeling tasks that naturally allow multiple labels for a single data instance, e.g., assigning more than one genre to a single movie title, we are interested in classification tasks that aim for a single label per data instance, but that leave room for interpretation as to which the correct label should be. Within this scope, we study ways to support ambiguity in different types of:

- **Classification scales:** We begin by studying *binary* classification tasks and move on to multi-class classification on both *nominal* and *ordinal* scales.
- **Data modalities:** We investigate classification ambiguity in the context of various data modalities including *text* documents, *images* and *times series* data.

1.3.2 Domains

The issue of ambiguity is not constrained to interpretations of people that lack a specific training background for a given task. In fact, ambiguity and inter-rater disagreement are well known phenomena even in seemingly well-defined expert domains. This dissertation investigates the issue of ambiguity in both of these worlds:

- **Novice domain:** We begin by studying inter-rater disagreement and deliberation within the context of *non-expert crowd work* facilitated through crowdsourcing marketplaces like Amazon's Mechanical Turk.

- **Expert domain:** Later, we translate our findings to the expert domain of *medical data interpretation*. In particular, we present case studies from two separate expert tasks: (1) assessment of eye disease by eye care physicians based on retinal images, and (2) sleep stage classifications by trained sleep health professionals based on biosignal recordings from a sleep laboratory.

1.3.3 Deliberation

Finally, we study group deliberation, our primary contribution for detecting and analyzing ambiguity, in different shapes and forms. Specifically, we provide results about deliberation dynamics with varying:

- **Group size:** While our initial study in focuses on groups of *two* or *three* members, the remainder of the dissertation relies on groups of three for the simple reason that a majority vote can be computed for disagreements among three, but not among two.
- **Workflow:** In the context of non-expert crowd work, we study *synchronous* deliberation conducted soon after the initial labeling task is completed because crowd workers quickly move on to other tasks and it can be hard to motivate workers to review disagreements at a later point in time. For our expert tasks, we facilitate an *asynchronous* deliberation workflow to better fit review and discussion activities into the busy schedule of health professionals.
- **Argument structure:** In the course of our three case studies, we evolve the way deliberation arguments are captured from an *open-ended* to a highly *structured* format. The underlying motivation is twofold. On the one hand, our results suggest that imposing an explicit structure on the way arguments are delivered can make the deliberation process more efficient. On the other hand, collecting structured data is useful for our other objective of simulating and investigating ambiguity-aware AI assistants.

1.4 Thesis Overview

Here, we provide an overview of the research included in this dissertation. We begin by outlining the related literature that our work is situated within or builds on, followed by a series of chapters that contribute to our goal of better supporting ambiguity in human-AI collaborative workflows for data classification, by capturing and leveraging information about instances ambiguity.

Chapter 3: We present three case studies of collaborative platforms that integrate group deliberation into data labeling as a tool to detect and analyze instances of ambiguous data. In the course of the chapter, we evolve the proposed deliberation methods in various ways: (a) we begin by studying deliberation within the novice domain of microtask crowd work and later translate our findings to two different expert domains of medical data interpretation; (b) in the context of switching from novice to expert work, we transition from a synchronous to an asynchronous deliberation workflow; (c) in each case study, we incrementally impose additional structure on how our systems collect deliberation arguments to increase efficiency, facilitate quantitative analysis and ultimately make deliberation data accessible for use by AI assistants. We provide insights into why inter-rater disagreements arose, under what circumstances they could be resolved and on how deliberation affected accuracy, efficiency and higher-level decision making.

Chapter 4: We present and study a novel approach based on the idea that deliberation data can be leveraged as training material for human labelers. In particular, we present qualitative and quantitative findings from a controlled experiment suggesting that medical generalists substantially benefit from reading deliberation discussions from medical specialists in the context of a complex diagnostic image assessment task. We demonstrate that reading deliberation discussions not only helps expert labelers calibrate their reasoning towards better comprehension on difficult cases, but also towards improved accuracy on a held-out test dataset. We discuss the implications of our findings for increasing the pool of trained expert labelers and medical diagnostic training.

Chapter 5: We present and study an AI assistant that not only highlights, but also explains ambiguous data to end users in a human-AI collaborative data classification task. We call this AI assistant "ambiguity-aware" and use a Wizard-of-Oz approach to study its effects on the perception and behaviour of expert end users in a medical time series classification task for sleep stage analysis. Our qualitative and quantitative results suggest that ambiguity-aware AI assistants can help end users better triage and trust AI-suggested classification labels, and that the quality of AI-provided ambiguity explanations strongly affects experts' ability to disambiguate difficult cases.

Chapter 6: Finally, we synthesize our findings from the previous three chapters and contextualize the collective impact of the research presented towards the issue of ambiguity in human-AI interaction. We summarize the main contributions and takeaways of our research in this space and conclude by discussing directions of interest for future work.

Appendices:

- **Appendix A**: We provide questionnaires used to understand how deliberation was perceived by human labelers, why they disagreed and under what circumstances disagreements could be resolved in Chapter 3.
- **Appendix B**: We provide questionnaires used to understand how human labelers perceived deliberation data as a form of training feedback in Chapter 4.
- **Appendix C**: We provide questionnaires used to understand how expert end users perceived ambiguity-aware AI assistants in Chapter 5.

1.5 Terminology

Here, we list and define central terms used throughout the dissertation:

1. *Ambiguity* — The quality of being open to more than one interpretation.[2]
2. *Deliberation* — The collaborative process of discussing contested issues by considering various perspectives in order to form opinions and guide judgement.[3]
3. *Adjudication* — Synonym of deliberation used in the context of medical data interpretation.

[2] https://en.oxforddictionaries.com/definition/ambiguity (accessed on 12 May 2020)
[3] https://www.comm.pitt.edu/oral-comm-lab/argument-deliberation/argument-deliberation-introduction (accessed on 12 May 2020)

Chapter 2

Background Literature

In this chapter, we provide an overview on background literature relevant to the content of this dissertation. We begin by outlining the role of ambiguity and how it is currently handled in various steps of the AI pipeline, including data labeling, model development and user interfaces. We then introduce the reader to existing protocols and systems to facilitate group deliberation which will later serve as a tool to detect and analyze ambiguity in data classification. Finally, we summarize related work from the particular domain of medical decision making and the role that ambiguity, deliberation and AI play in that context.

Before we begin, we introduce our high-level perspective on the central terms used in this work and on their relationship to each other, including ambiguity, uncertainty, inter-rater disagreement and group deliberation. In the context of this dissertation, *ambiguity* is defined as the quality of being open to more than one interpretation. As such it is a specific type of *uncertainty* arising from the circumstance that there may exist multiple conflicting, yet equally valid interpretations of the same phenomenon. However, there exist other forms of uncertainty, e.g., uncertainty due to a lack of information about the current state of the world, or uncertainty around the precise consequences of a specific action for the future state of the world.

Ambiguity and *inter-rater disagreement* are strongly linked to each other in that ambiguity typically reveals itself in the form of disagreement among multiple independent human observers. However, both concepts are not identical with each other, nor does one imply the other. For example, given an ambiguous case, a specific group of observers may happen to agree with each other while another combination of observers may have arrived at conflicting interpretations. Conversely, two conflicting interpretations are not necessarily also equally valid, e.g., if one observer made a mistake that ought to be corrected.

This is where the process of *group deliberation* comes into play. It can help us understand whether a given disagreement is either based on mistakes and can therefore be resolved or whether a disagreement is based on some form of underlying ambiguity and is therefore irresolvable. We thus consider classification ambiguity the central problem this work aims to address and group deliberation the central method used to expose if and how an instance of inter-rater disagreement is linked to ambiguity.

2.1 Ambiguity in Human-AI Interaction

2.1.1 Ambiguity in Data Labeling

Ambiguity and the associated issue of inter-rater disagreement in data classification, are both topics that have received ample attention not only in the epistemological (e.g., [55, 97, 136]) and medical (e.g., [10, 103, 114]) literature, but also within the human-computer interaction (HCI, e.g., [25, 29, 64]), human computation (e.g., [9, 42, 93]), and computer-supported cooperative work (CSCW, e.g., [8, 76, 95]) communities.

Aroyo and Welty [9] view inter-rater disagreement as a function of three phenomena: variability among human *annotators*, characteristics of the *data* at hand, and the quality of the *task instructions*. We adopt a similar approach in exploring inter-rater disagreement in data classification, and synthesize prior work within these three categories below. In addition, we take into account *data presentation* as another potential source of inter-rater disagreement, pertaining to differences in the way that human annotators view the data at hand.

Annotator differences. Mumpower and Stewart [97] offer an early theoretical account in which expert disagreement is discussed in three forms: (1) *personality-based* disagreement, beget by expert ideology, venality, or incompetence, (2) *judgement-based* disagreement, due to information gaps, and (3) *structural* disagreement, due to experts holding different organizing principles or problem definitions. Garbayo [55] distinguishes verbal disagreement—disagreement due to differences in terminology or semantics between experts with respect to the problem definitions mentioned previously—from legitimate disagreement, arising despite experts having access to the same evidence. Gurari and Grauman [64] found that disagreement among crowd workers in visual question answering tasks can arise from differing levels of annotator expertise. Kairam and Heer [76] showed that inter-rater agreement in a crowdsourced entity annotation task is affected by how conservatively or liberally workers follow task instructions.

Data characteristics. In addition to grader-specific factors, disagreement may be an indicator of ambiguity, vagueness, or complexity inherent in the given data [8, 9, 10, 111, 114]. Prior works have demonstrated that inter-rater disagreement can be associated with characteristics of individual data objects, including text documents for sentiment classification [111], photographs for visual question answering [64], medical images for eye disease assessment [114], and biomedical time series data for epilepsy diagnosis [10].

Task instructions. Finally, inter-rater disagreement has been attributed to ambiguous category definitions [9, 25, 64, 93] relevant to a given task. Gurari and Grauman [64] identified subjective questions and vocabulary mismatch between crowd workers as sources of disagreement. Chang et al. [25] found that worker disagreement can arise due to ambiguous or incomplete category definitions, and proposed a system to analyze crowd-generated conceptual structures post-hoc. Manam and Quinn [93] developed workflows for identifying and refining unclear instructions for crowdsourcing tasks.

Our work builds on these prior contributions by studying various factors contributing to inter-rater disagreement in various settings of data classification. Our quantitative analyses in Chapter 3 take into account various factors, including differences between graders (in terms of professional credentials, geographic location, and work experience), data characteristics of individual disagreement cases (e.g., in terms of data complexity), and the role of classification guidelines (in terms of individual guideline rules) to understand inter-rater disagreement.

Furthermore, Section 3.3 contributes a novel perspective on the problem of inter-rater disagreement by incorporating **data presentation** as another potential factor contributing to disagreement. In particular, that study incorporates the question to what extent differences in how experts choose to view the data at hand—configuring the viewer interface—may be associated with experts arriving at divergent interpretations of the same data.

2.1.2 Ambiguity in Model Development

Prior systems have generally taken one of three approaches to the problem of ambiguity in AI-based data classification:

Eliminating Ambiguity. Traditional machine learning classification methods eliminate class diversity using automatic procedures like majority vote [75], or expectation maximization [107]. These systems tend to view ambiguity as a proxy for noise to be reduced or eliminated in the data. [23, 148].

Aggregating Multiple Outputs. Other systems retain disagreement labels for the purpose of training multiple models (e.g., one for each human labeler [62]); these systems

typically produce multiple AI predictions which are aggregated into a single label before being presented to the end user.

Label Distribution Learning. A more ambiguity-centric approach to data classification is label distribution learning (LDL) [57], where machines are trained to predict not just one label for a given case, but a distribution of possible classification labels [32, 112]. Standard LDL models will assign uncertainty estimates to their classification outputs, providing degrees of plausibility for each possible label.

2.1.3 Ambiguity in AI Interfaces

The question of how systems should *communicate* or visually represent uncertainty to end users has received ample attention in the human-computer interaction (HCI) community [130]. Approaches include visualizing uncertainty as extrinsic annotation (e.g., confidence intervals), abstract, continuous outcomes (e.g, probability density plots), or hypothetical, discrete outcomes (e.g., natural frequencies or icon arrays) [78]. Kay et. al [78] suggest that communicating uncertainty through discrete outcomes can improve decision making on the part of end users.

Prior work has found that collecting explanations around ambiguous cases during data labeling workflows can be leveraged towards more fine-grained and flexible post-hoc data classification [25]. Chen et al. [27] designed an AI-supported analytics tool for the purpose of qualitative coding in social science shifting the focus to identifying disagreement and ambiguity among groups of human coders. Galdran et al. [53] developed a system for vessel classification from retinal images, with the ability to classify uncertain cases and provide direct uncertainty estimates for its labels while achieving state-of-the-art classification performance.

In Chapter 5, simulate and explore an ambiguity-aware AI assistant for medical data analysis, a system that provides human-interpretable rationales for all plausible classification labels. While our AI assistant is simulated in the sense that it does not predict, but merely displays human-annotated data, our work contributes novel insights about how such an ambiguity-aware system affects expert perception and behaviour.

2.2 Group Deliberation

Instances of data ambiguity typically surface in the form of conflicting assessments among a group of independent human observers. We are interested in leveraging inter-rater disagree-

ment to drive collaborative analysis of the underlying sources of ambiguity. In particular, we explore the role of group deliberation as a potential tool to detect and analyze ambiguity in the context of data classification. In this section, we summarize existing protocols and systems for group deliberation to contextualize our work.

2.2.1 Deliberation Protocols

A seminal protocol on structured decision making is the Delphi method [35], where experts provide, justify and reconsider their estimates through questionnaires in multiple rounds. A facilitator controls the information flow by summarizing estimates and filtering out irrelevant justification content at each round. The Delphi process ends after a fixed number of rounds or when unanimous consensus is reached. The key characteristics of the Delphi method are anonymity of the participants, avoidance of any direct interaction among group members, as well as structured and curated information flow as implemented by the facilitator. The Delphi method is typically used for forecasting, policy making and other types of complex decision making processes. Later versions of the Delphi method, e.g., by Hartman and Baldin [67], make use of computer-supported communication to facilitate remote collaboration, larger groups and asynchronous interaction.

Group deliberation is a typical method for generating high-quality answers in expert domains such as medicine [62, 82, 115]; however, little work has shown under what circumstances group deliberation is resolvable or produces better decisions. Several works explored factors that influence the process and outcomes of group deliberation. Nemeth [99] found that when jurors are required to reach a unanimous decision, there is more conflict, more changes in assessments, and higher confidence in the final verdict reported by members of the group. Solomon [135] sees conflict as an important feature of any effective deliberation system. He argues that dissent is both necessary and useful—as "dissenting positions are associated with particular data or insights that would be otherwise lost in consensus formation"—and criticizes procedures that push deliberators to reach consensus. Instead, he advocates for a structured deliberation procedure that avoids the undesired effects of *groupthink* [74]—the tendency to agree with the group by suppressing dissent and appraisal of alternatives—by actively encouraging dissent, organizing independent subgroups to discuss the same problem, and ensuring diversity of group membership. Kiesler and Sproull [80] found that time limits imposed on deliberation tend to decrease the number of arguments exchanged and to polarize discussions. The authors suggest the use of voting techniques or explicit decision rules to structure the deliberation timeline.

2.2.2 Deliberation Systems

Building on some of these early theoretical results, online deliberation systems have been developed and validated in various domains, including public deliberation [84, 85], on-demand fact checking [83], political debate [50] and knowledge base generation [154]. For example, ConsiderIt [50, 84] is a platform for supporting public deliberation on difficult decisions, such as controversial policy proposals made during U.S. state elections. Kriplean et al. [85] explored ways to promote active listening in web discussions by explicitly encouraging discussion members to summarize the points they heard. Kriplean et al. [83] studied the correctness of statements made in public deliberation and developed an on-demand fact-checking system. Zhang et al. [154] introduced recursive summarization of discussion trees to enable large scale discussions. Liu et al. [91] proposed a visualization technique to augment deliberation for multi-criteria decision making. Their results suggest that highlighting disagreement across multiple decision criteria can cause participants to align their opinions for various reasons, from genuine consensus to appeasement. This finding reinforces the importance of designing deliberation procedures to minimize *groupthink*.

The MicroTalk workflow proposed by Drapeau et al. [42] focuses on argumentation within the microtask crowdsourcing setting. In MicroTalk, workers are prompted to provide justifications for their decisions, and an algorithm selects certain justifications (based on a metric of readability) to present them as counterarguments to other workers, triggering them to reconsider their decision. MicroTalk is based on an asynchronous model with no interactive back and forth discussion between workers. Building on this work, Chen et al. [29] showed that a synchronous workflow enabling crowd workers to engage in real-time, multi-turn discussions, can lead to additional improvements in answer accuracy. In this dissertation, we study both synchronous and asynchronous workflows for group deliberation among human annotators to analyze how disagreement arises and under what circumstances it can be resolved.

Chang et al. [25] addressed the issue of ambiguous category definitions by proposing a system to enable flexible, post-hoc analysis of crowd-generated, conceptual structures, as opposed to refining classification guidelines a priori. Goyal et al. [60] designed a shared sense-making interface allowing dyads of participants to synchronously share hypotheses, evidence and other insights in a simulated criminal investigation task, leading to increased decision making performance compared to a baseline interface. Chang et al. [26] proposed a multi-step crowdsourcing workflow for semantic frame annotation, allowing workers to express disagreement with expert-labeled golden data presented as feedback during labeling.

While we embed our work reported in Section 3.1 in a similar context as Drapeau et

al., our focus is on understanding how unfiltered group deliberation, task types and other characteristics impact resolvability, beyond just answer accuracy. The systems we report on in Sections 3.2 and 3.3 leverage a workflow in which expert-provided classifications can be contested by other experts in a round-based collaborative manner. Our work draws inspiration from the Delphi method and builds on prior work above by deploying web-based, structured deliberation systems. We extend the state of the art by translating existing workflows into the complex expert domain of medical data analysis, and by integrating a procedure to collect arguments from experts in the form of explicit rationales, centered around pre-existing domain-specific decision criteria. Our approach differs from several other works in that the primary objective is to achieve a better understanding of the sources and dynamics of expert disagreement, as opposed to optimizing for accuracy within data labeling workflows. In other words, our approach uniquely combines the two notions that (1) there may exist multiple valid answers to one and the same question and that (2) the reasons *why* multiple answers may be equally valid are important for expert end users to understand and that these reasons can be effectively analyzed using the process of group deliberation.

2.3 Medical Decision Making

Except for Section 3.1, all other sections in Chapter 3 as well as Chapters 4 and 5 are anchored within the specific expert domain of medical decision making. Given the strong focus of this dissertation on the medical domain, this section provides a more in-depth overview on the issues of expert disagreement, group deliberation and AI-based decision support in medical decision making scenarios.

2.3.1 Expert Disagreement in Medicine

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 [10, 88, 118, 128]. The issue of inter-rater disagreement is particularly critical within medicine where unreliable clinical decisions can impact patients' lives adversely. Indeed, Raghu et al. [114] concluded that label disagreement poses a "full-fledged clinical problem in the healthcare domain."

Prior work in medical decision making describes that medical experts are susceptible to biases in their reasoning; for instance, "confirmation bias" can lead a medical expert

to look only for evidence that is in line with their pre-existing hypothesis [19]. As sub-optimal decision-making in medicine can have major consequences, it is crucial to combat any reasoning biases medical experts may have. Our simulated ambiguity-aware AI aims to mitigate this bias by putting forth arguments for conflicting medical assessments, encouraging perspective-taking for alternate lines of reasoning.

Related literature suggests that communicating uncertainty can impact cognition and trust, and potentially influence experts' decision-making behaviours [145]. That said, there is a body of work showing that people have a general aversion towards ambiguity [79, 144]. For example, a study by Redelmeier and Shafir suggested that the uncertainty between two medical assessments led some doctors to avoid making a decision altogether [117]. Work done in psychology acknowledges ambiguity-tolerance as a personality variable [20, 72]. Medical education research advocates that given the inevitable nature of uncertainty in contemporary medicine, medical experts must acquire a certain level of tolerance to it [92].

Several works have suggested ways to make productive use of disagreement information in medical data labels (e.g., [10, 15, 32, 71, 114, 124, 129]). Inel et al. [71] 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 [10], and eye disease diagnosis from retinal fundus photographs [114]. Finally, Barnett et al. [15] 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.

2.3.2 Group Deliberation in Medicine

The issue of low inter-rater reliability in the clinical domain has motivated efforts to find methods of adjudicating ambiguous cases in medical data classification tasks. Group deliberation has also garnered support in the medical domain as a method for generating a trusted reference standard for the evaluation of automated classification methods [62, 82, 115].

In the context of a medical imaging study, in which the aim was to diagnose eye disease based on retinal fundus images, Krause et al. [82] found that group deliberation was more effective than majority vote when it came to recall among experts. This same dataset

was used by Guan et al. [62] to demonstrate that ensembles of multiple grader-specific machine learning models could outperform a single-prediction model trained on majority labels, when benchmarked against an adjudicated gold standard. Penzel, Zhang, and Fietze [106] argue that group deliberation, or "consensus scoring" is the optimal training technique for human scorers in the context of sleep stage classification.

Recent work by Barnett et al. [15] showed that automatic pooling of independent opinions from multiple doctors outperformed individual diagnosis across various diagnostic tasks. However, the authors did not investigate the effects of permitting communication or collaboration among doctors to allow for collective adjudication of their diagnoses.

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 [98] and in medical diagnosis [51].

2.3.3 Medical Diagnosis Training

Given that medical generalists far outnumber specialists in various fields [5], research into effective ways to train medical generalists for difficult cases holds the potential to tap into a large pool of high-quality medical labelers.

A scalable approach for medical diagnosis training is through computer-based tutorials [68, 86]. 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. The work we present in Section 4 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) [40, 137]. 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 Chapter 4, 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.

2.3.4 AI-based Clinical Decision Support

Clinical decision support (CDS) is broadly defined as the provision of intelligent assistance to clinicians, medical staff, and patients [101]. CDS can include low-level functions like computerized alerts and reminders for providers and patients, or high-level functions like patient diagnosis [33]. Norman et al. [100] describe a dual process of diagnostic reasoning, where physicians engage in (1) a non-analytic or unconscious process of hypothesis generation, and/or (2) a conscious, analytic process of hypothesis testing. The latter is an extensive computational process, and has motivated efforts to develop AI-based CDS systems for diagnostic support.

In such support systems, a physician cross-checks the algorithmic output against their internal knowledge, but takes responsibility for the final diagnostic decision. Our work takes a similar approach of augmenting, instead of automating, the job of physicians [61].

Barriers to the Adoption of AI-Based CDS Systems

Explainability. ML-based AI systems are typically opaque with respect to their internal functions [96]. In fields where AI is tasked with important decisions, it is imperative that automated decision making be interpretable, especially if the AI is known to be imperfect [21, 81]. The field of XAI [146] emerged as a response to this problem of transparency beginning in the 1970s and 1980s with the deployment of expert systems with explanation capabilities—most notably for medical decisions [3]. Explanations have been found to promote transparency in machine learning algorithms and make users more aware of how a system works [113]. Recent approaches in XAI, e.g., Ehsan et al. [46], demonstrate that AI systems can learn to generate human-like natural language explanations for their decisions. Mittelstaedt et al. [96] argue that there is a mechanistic link between explanation and justification in human discourse, and that machine explanations should emulate human explanations. The simulated AI assistant we present in Chapter 5 instantiates these design principles, by providing human-interpretable rationales for its outputs.

Trust. A lack of trust is arguably the most significant barrier to adoption of AI-based systems. A CDS system can bias a physician to choose the wrong course of action against their own clinical judgement [43]. Human experts may also fail to trust a reliable system.

It is crucial that an appropriate level of trust in automation be established to balance over-reliance and under-reliance. Cai et al. [21] demonstrated a link between explainability and trust by showing that pathologists trusted a CDS tool for cancer diagnosis more if they could tweak its internal representation of image similarity using domain-specific concepts (e.g., number of fused glands).

Addressing the problem of trust becomes more complicated in the context of uncertainty. Psychological uncertainty is an aversive state [145], and thus information must be communicated effectively to hedge against its negative effects. There appears to be a volatile relationship between uncertainty and trust. On the one hand, trust can be undermined by failing to communicate uncertainty; on the other hand, admitting uncertainty can also hinder trust [145]. Thus, it is crucial that machines strike the right balance between communicating and withholding uncertainty information.

Studies on how communicating uncertainty affects trust are limited and have produced mixed results. While there is some evidence that trust can be fostered through explained uncertainty [78], more research is needed. In a recent review of the matter, van der Bles [145] acknowledged that uncertainty does not always produce negative emotional effects. Indeed, in the healthcare domain, Schneider et al. [130] developed a system for communicating uncertainty in fertility prognosis that increased users' understanding of uncertainty without causing them to have a negative view of the system.

In Chapter 5, we study how the workflows and perception of medical experts is affected by an AI assistant capable of identifying and explaining ambiguous cases.

Chapter 3

Group Deliberation for Data Labeling

One primary contribution of this dissertation is to introduce and study group deliberation in the context of data labeling, as a tool to detect and analyze ambiguous cases, and to either resolve inter-rater disagreements organically through deliberation, or otherwise mark them as irresolvable. In this chapter, we present insights from three case studies about collaborative online platforms that all integrate group deliberation into data labeling workflows, in the context of different data modalities, labeling tasks and types of human labeler expertise. Section 3.1 describes case studies on synchronous group deliberation in the context of novice crowd work and text classification tasks of variable subjectivity. In Sections 3.2 and 3.3, our focus shifts from novice crowd work to the expert domain of medical data interpretation, exploring asynchronous and structured deliberation workflows for image labeling and time series labeling respectively.

3.1 Crowd Deliberation for Text Labeling

Crowdsourced classification of data typically assumes that objects can be unambiguously classified into categories. In practice, many classification tasks are ambiguous due to various forms of disagreement. Prior work shows that exchanging verbal justifications can significantly improve answer accuracy over aggregation techniques. In this section, we study how worker deliberation affects resolvability and accuracy using case studies with both an objective and a subjective task. Results show that case resolvability depends on various factors, including the level and reasons for the initial disagreement, as well as the amount and quality of deliberation activities. Our work reinforces the finding that deliberation can increase answer accuracy and the importance of verbal discussion in this process. We contribute a new public data set on worker deliberation for text classification tasks, and discuss considerations for the design of deliberation workflows for classification.

3.1.1 Motivation

Classification is a prevalent task in crowdsourcing as well as many real-world work practices. A common assumption for many classification tasks is that objects can be *unambiguously* classified into categories, and that the quality of the labeled data can be measured by the extent to which annotators agree with one another. As a result, most post-processing techniques designed to filter or aggregate labeled data interpret inter-rater disagreement as "noise in the signal" originating from human mistakes. In practice, many classification tasks are ambiguous, and disagreement can happen for various reasons including missing context, imprecise questions, contradictory evidence, and multiple interpretations arising from diverse levels or kinds of annotator expertise [64].

The independence of individual assessments has traditionally been considered a prerequisite for leveraging the 'wisdom of the crowd', but recent findings from crowdsourcing [42] and social behavioural research [98] have found that deliberation can help improve answer quality. Drapeau et al. [42] showed that crowdsourcing workflows enabling workers to exchange justifications and to reconsider their assessments based on each other's arguments, can improve accuracy over output aggregation techniques. These prior works, however, have focused mostly on answer *correctness*, with the a priori assumption that each disagreement can be resolved to one correct answer.

In this work, we take the stance that inter-rater disagreement carries valuable information [25, 76, 44] and that deliberation should not always lead to a unanimous consensus.

In particular, we investigate factors that contribute to the *resolvability* of a case. We conducted a study on two text classification tasks—a sarcasm classification task, which has been shown to be inherently ambiguous [48], and a semantic relation extraction task, where objective ground truth is available [42]—to investigate how deliberation affects resolvability. Our key contributions are:

1. We study how the deliberation outcomes differ depending on task subjectivity.
2. We present observations showing that the *resolvability* of an ambiguous case depends on the reasons for and level of initial disagreement, the amount and quality of the deliberation activities, as well as the task and case characteristics.
3. We publish a new *dataset on worker deliberation* in text classification tasks, including all original and revised classifications as well as the deliberation dialogues.

The rest of this section describes the details the deliberation workflow designed and implemented for this study, outlines the experimental procedure and findings, and concludes with a discussion of applications and design considerations.

3.1.2 Deliberation Workflow

We designed a workflow enabling crowdworkers to revisit and potentially resolve disagreements in text classification tasks through group discussions. The input to our workflow is a set of cases (e.g., text documents) to be classified. Each disagreement case either gets resolved through discussion or remains unresolved. The output of our workflow are multiple classification labels for each input case and its deliberation data consisting of structured information (i.e., original and reconsidered classification decisions, confidence levels, sources of disagreement, and evidence) and deliberation dialogues (e.g., arguments, explanations, and examples).

This workflow transforms input to output in four time-limited stages (Figures 3.1). Workflow stages A, B, C and D started at consecutive hours (e.g., 3pm, 4pm, 5pm, and 6pm), to ensure that all workers could complete one stage before collectively starting the next.

Stage A: Independent Classification. Workers independently performed 10 classification microtasks. In each task, workers were asked to read a text document, classify it into one of two categories, provide a confidence level for their decision, and highlight

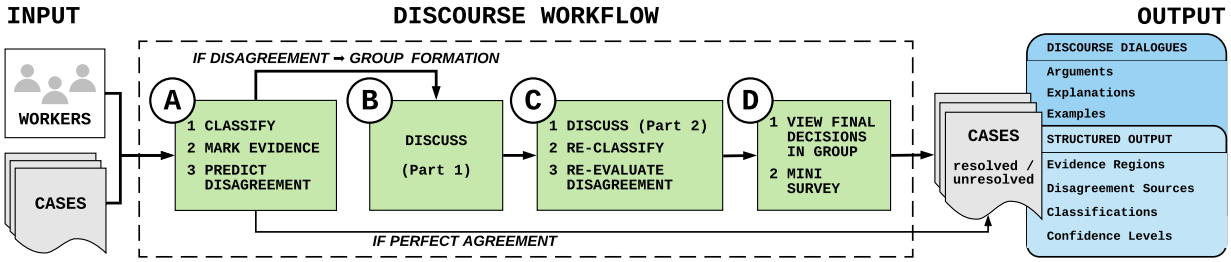


Figure 3.1: Input, output and stages of the deliberation workflow implemented for this study. Each of the green blocks represents one of multiple microtasks in each of the four workflow stages A, B, C and D.

evidence to support their choice. Workers were then asked to predict the level of disagreement for the classification task by indicating whether they expect Substantial Agreement ("I expect most people to agree with me"), Half Agreement ("I expect only about half of the people to agree with me"), or Substantial Disagreement ("I expect most people to disagree with me"). If workers chose one of the latter two options, they were also asked to choose the sources of disagreement they anticipated. They could select multiple options from a list of six preset options as described in Table 3.1—Fuzzy Definition, Missing Context, Contradictory Evidence, Important Details, Expertise Needed, Subjective Case—covering a variety of common sources of disagreement from prior work and our pilot studies. They could provide their own rationale using an optional free-form field. Appendix A.1.2 provides a complete list of questions and answer options for this stage.

Between stage A and B, our system dynamically put crowdworkers into groups of three to deliberate on one or more cases, with the constraint that there was exactly one dissenter for each disagreement case per group. To ensure group heterogeneity, the group formation procedure was randomized, i.e., all workers had the same chance of being assigned to a group, and it was equally likely for them to be grouped together with either two random workers they disagreed with, or one with whom they agreed and one with whom they disagreed.

Stage B: Discussion Round 1. Workers joined their assigned groups to discuss their disagreement cases. For each case, workers were required to leave at least one comment in the group chat explaining why they had chosen their label for the given case. In addition, they could choose to highlight more parts of the text as evidence. The comments and associated evidence were recorded and shown to other workers in real time.

Stage C: Discussion Round 2 and Reconsideration. Workers collectively returned to review each other's comments and participate in a second round of discussion

Table 3.1: Preset choices for sources of disagreement.

Fuzzy Definition	Other people may have different definitions of [sarcasm / relation] in mind.
Missing Context	The text is ambiguous because of missing context (for example, [the identity of the product / some important information about the person or the place] is unknown).
Contrad. Evidence	The text contains some features that indicate [sarcasm / relation] is expressed and other features that indicate [absence of sarcasm / relation is not expressed].
Important Details	The text contains relevant details other people could easily miss.
Expertise Needed	Someone with more experience or expertise may see or understand something about the text that I don't.
Subjective Case	This is a case where a person's answer would depend heavily on their personal preferences and taste.
Other	Requiring free-form answer if selected

on the same disagreement cases. They were asked to further discuss the case by leaving at least one comment in the group chat, and optionally highlight more evidence. After providing at least one additional discussion comment, workers were individually prompted to reconsider their original classification decision and confidence level. To submit their final classification, workers could choose one of the two original class labels or a third option named "irresolvable". Finally, workers were also asked to re-evaluate what they considered the source of disagreement for the given case and group in light of the previous discussion. Workers were again given the six preset options (Table 3.1) and an optional free-form field, as well as the free-form answer they had provided earlier as the anticipated source of disagreement, if any (Appendix A.1.3). By providing the "irresolvable" option, we reduced the bias for consensus, and incentivized workers to change their answer only if they truly believed that their updated classification label was correct. Since we were interested in case resolvability, the reconsidered classification decisions were collected from all discussion members individually instead of enforcing the group to produce one joint decision. Importantly, workers did not see the updated answers of other group members before stage D to reduce opportunities for strategic voting.

Stage D: View Final Decisions. For each disagreement case, workers were presented with the final decision (i.e., either one of the class labels or "irresolvable") from each group member, and they were given a short open-ended survey on why they thought the disagreement was resolved or not resolved, as well as why they had changed or stuck to their original classification decision (Appendix A.1.4).

General Design Considerations

In the workflow design, we made several design decisions regarding the communication medium used for deliberation, ensuring stable pay/work ratios through filler tasks, and

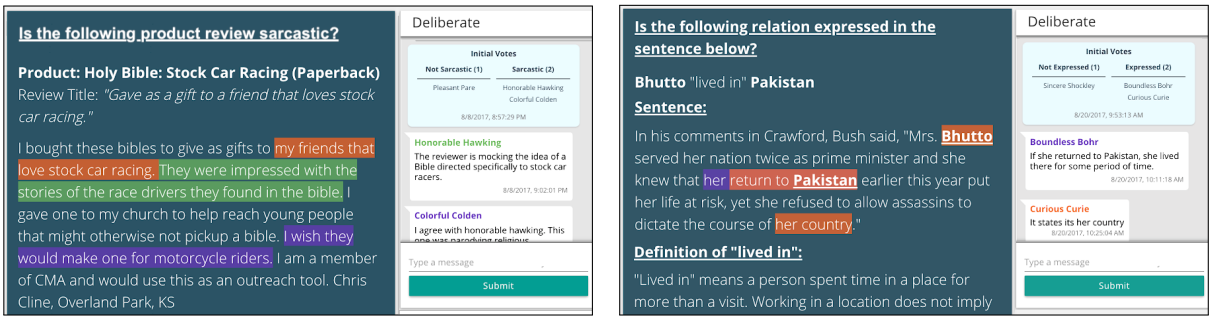


Figure 3.2: Screenshots showing our worker deliberation interface paired with text classification tasks for sarcasm detection (left) and relation extraction (right). The text documents under discussion are shown on the left of each screenshot. Parts of the text are highlighted by different group members as evidence supporting their classification decisions. Written justifications are exchanged through a real-time chat component, with an embedded voting summary, shown on the right of each screenshot.

motivating workers to return for all four consecutive workflow stages despite the intermittent breaks.

Communication Medium. Our decision to use text (versus voice or video) as the communication medium for deliberation was motivated by prior research suggesting that written communication improves outcomes of consensus formation by avoiding bias from the tone of voice or a perceived lack of anonymity [35, 120].

Filler Tasks. If workers had less than 10 disagreement cases to revisit in stages B, C and D, they were asked to perform filler classification tasks (identical to the microtasks in stage A) to fill up the slots. This was to ensure that workers would not be incentivized to provide answers in stage A that were likely to agree with others just to reduce the number of disagreement cases they would have to process in subsequent stages. The data from the filler tasks were not used in our experiment.

Worker Retention. An important design consideration was the mechanism used to encourage workers to collectively return to stages B and C for real-time discussion. Through several pilot studies, we found a combination of monetary incentive and timed notifications to be a successful approach. As a monetary incentive, we paid fixed amounts for participation in each stage and an additional bonus for full participation in all stages. We used the MTurk API to send out reminder emails five minutes before the start of stages B, C and D, with a web link to join the next stage and a notice stating that the next stage could only be joined within three minutes of its start.

Interface

Our deliberation interface is general, i.e., it can be integrated into any task interface for text classification and requires only minor modifications for other data modalities like images. Figure 3.2 shows the interface in action. The interface displays the text document in question and attaches a chat box to the right, through which group members can exchange justifications in real time. Group members are randomly assigned friendly-sounding and easy-to-remember pseudonyms (e.g., 'Joyful Joliot' or 'Enthusiastic Easley') to allow users to address each other without having to reveal their identity. Finally, the interface enables users to highlight parts of the text document as evidence to support their argument. Highlights and pseudonyms are color-coded per deliberator.

3.1.3 Experiment

We conducted an experiment to investigate how deliberation affects the outcomes of crowdsourced text classification tasks. Here, we describe the task types, procedure, participant recruitment and payment associated with our study.

Task Types

We focus on two types of text classification tasks (Figure 3.2) with different degrees of inherent subjectivity. During our pilot studies, we also considered other task types (e.g., image, audio or video classification), but decided to defer those to future work due to the added complexity of synchronously highlighting evidence in such data modalities.

Sarcasm Task. For the first task type, we asked workers to label Amazon product reviews as sarcastic or not sarcastic, using the sarcasm detection data set published by Filatova [48]. We chose this task type and data set because Sarcasm detection is considered by Filatova et al. [48] as an inherently subjective task due to the "absence of a formal definition of sarcasm", an observation further supported empirically by low rates of inter-rater agreement in Filatova's data set. To generate a subset of cases for our experiment, we first identified all reviews for which Filatova [48] reported highest inter-rater disagreement (i.e., 3 sarcastic vs. 2 not sarcastic or vice versa) and then, from this subset, retained the 40 most compact cases based on word count.

Relation Task. For the second task type, workers were asked to indicate whether a certain semantic relation between a person and a place (e.g., "Nicolas Sarkozy *lived in* France", "Pavarotti *died in* Modena") was expressed in a given sentence. In contrast to

the sarcasm detection task, this task is more well-defined and objective, as the ground truth data can be determined from the official label guidelines for the TAC KBP relations *LivedIn* and *DiedIn* as published by the Linguistic Data Consortium. We presented the corresponding relation definition to workers in each individual classification task (see right side of Figure 3.2 for an example) to explicitly make workers aware of the label guidelines. We used all 40 sentences from the data set used by Drapeau et al. [42], of which 25 have ground truth labels.

Procedure

Before the experiment, workers first filled out a pre-study questionnaire eliciting demographic information, including age group, gender, native language and self-rated proficiency in English (see Appendix A.1.1). Participants then collectively stepped through the four consecutive stages of our workflow. They were free to close the browser tab or do other work after completing each stage and before the start of the next one, of which they were notified via email five minutes prior to start.

Participant Recruitment

We recruited 316 participants on Amazon Mechanical Turk, using workers from the US who had completed at least 500 tasks with a 90% acceptance rate. Based on the pre-study questionnaire, almost all of our workers are native English speakers (97%) and have high self-rated proficiency in English (96% and 4% selected the highest and second highest levels on a 5-point Likert scale). The distribution over age groups is: 18-25 (14%), 26-35 (44%), 36-45 (23%), 46-55 (13%) and 56+ (6%). About half of our workers (53%) are female.

Payment

Workers were paid US $1 for each stage that they completed. In addition, we paid a one-time completion bonus of US $2 to workers who completed all four stages. Each stage took workers around 15 minutes to complete, resulting in an approximate payment of US $6 per hour of work for participants who completed all four stages. Note that workers were free to close the browser tab or do other work after completing each stage and before the start of the next one. We incentivized workers to actively engage in discussion by offering an extra bonus of US $0.50 for each disagreement case in which all group members voted for the correct expert answer (which can be one of the two class labels, or "irresolvable") in stage C. As only few cases had an expert answer available, we paid the extra bonus to all groups in the end that reached consensus on one of the two target categories.

3.1.4 Research Questions and Hypotheses

Our study aims to answer three research questions.

Q1: Why do annotators disagree with one another? We expect disagreement to arise due to different reasons in the two task types with different degrees of inherent subjectivity, where the target concepts are more versus less well-defined. Based on this intuition, we hypothesize that:

[H1a] Sources of disagreement differ significantly between the two task types.

[H1b] Annotators can predict levels of disagreement for individual cases better than random.

Q2: Under what circumstances can disagreement be resolved through worker deliberation? A variety of factors can contribute to the resolvability of a given case. First, the characteristics of a task and its associated sources of disagreement can play a role. For example, well-defined target categories may provide better grounding for convincing arguments, enabling groups to more easily come to a consensus. We hypothesize that:

[H2a] Sources of disagreement affect whether a case will be resolved.

[H2b] Task subjectivity affects whether a case will be resolved.

Second, the characteristics of the deliberation activities can influence whether a case is resolved. We hypothesize that:

[H2c] The extent to which members contributed equally affects case resolvability.

Third, we expect the amount of consensus in the label and overlap in highlighted evidence during the independent classification phase (i.e., stage A) to be predictive of a case's resolvability. We hypothesize that:

[H2d] The extent of the disagreement amongst group members affects resolvability.

[H2e] The amount of overlapping evidence between group members affects resolvability.

Q3: What impact does the deliberation workflow have on crowdsourcing outcomes and processes? The deliberation workflow, which encourages workers to consider diverse evidence and arguments, may have a positive effect on crowdsourcing outcomes and processes, such as improving the overall answer correctness and discouraging *groupthink* (i.e., the tendency of group members to blindly follow the majority). We hypothesize that:

[H3a] Worker deliberation improves the quality of the crowdsourced annotations.

[H3b] The probability that a case will be resolved in favour of the initial majority vote is similar to the probability that a case will not be resolved in favour of the initial majority vote.

[H3c] Sources of disagreement and the extent to which members contributed equally affects whether a case will be resolved correctly.

3.1.5 Experimental Conditions

One of the goals in our study is to understand the effect of worker deliberation on the quality of crowdsourced annotations (H3a). To quantify the effects, we tested two additional variants of our workflow *without* the discussion component. Each participant was randomly assigned to one of the following conditions:

Disagree, Discuss and Reconsider (N=316): workers reconsider their position *after discussion* with other group members. This is our main condition testing our full deliberation workflow. If not otherwise noted, all results below are based on data from this condition.

Disagree and Reconsider (N=26): workers reconsider their position after they are shown group disagreement data, but *without* a discussion. This condition was added to isolate potential effects of showing workers information on who agreed vs. disagreed with them.

Reconsider Only (N=24): workers reconsider their position *without* being shown group disagreement data and *without* a discussion. This condition was added to identify any *learning* effects resulting from workers revisiting a case after labeling other cases.

The latter two conditions are used for hypothesis H3a only. In the results section for hypothesis H3a, we use the term **Baseline** for labels submitted during stage A (i.e., before any reconsideration), and otherwise refer to reconsidered labels submitted during workflow stage C.

3.1.6 Data and Analysis

Data. For the analysis, the data include: (a) pre-study questionnaire data about demographics and language proficiency, (b) post-study questionnaire data, probing at workers' thoughts about and experiences with the deliberation process, (c) all the messages exchanged in the deliberation workflow stages, (d) the pre- and post-deliberation classification label and confidence for each worker, (e) the highlighted evidence for each case, (f) the sources of disagreement as anticipated by workers before discussion and re-evaluated by workers after discussion, (g) the anticipated resolvability of each case, and (h) the anticipated level of disagreement for each case.

Method. For each task type, we ran a *breadth* analysis on 40 text documents with up to 3 independent group discussions per document in order to identify the level of ambiguity for each case. This was followed by a *depth* analysis of the 10 most ambiguous cases with up to 16 independent group discussions per case, used to answer Q2 and Q3.

We analyzed the data using both quantitative (e.g., basic descriptive statistics, regression models) and qualitative analysis of worker responses. For the logistic regression models [94], we used the step-wise regression procedure to select the best possible combination of variables that could explain the dependent variable, based on improvements to the Akaike information criterion (AIC). We also test the goodness of fit of each model using the Hosmer-Lemeshow [69], Osius-Rojek [102] and Stukel [139] tests at significance level 0.05. For qualitative data analysis, line-by-line inductive open coding was performed by one of the study authors to identify the emerging themes reported below.

Filtering. Due to worker dropout between the four workflow stages, there were groups that became inactive during the deliberation process or had too few members remaining at the end. Out of all 316 participants, 78.2% completed all four stages, 3.8% and 4.8% dropped out after stages B and C respectively, and 13.2% completed only stage A. Possible explanations for the moderate dropout after stage A are that some workers might not have received or seen their email notifications in time (or not at all) to join stage B, or may simply not have been interested in returning to the same type of task in subsequent sessions. We excluded groups from the analysis if more than one member was inactive (i.e., they did not complete stages B and C), resulting in empty or single person groups, or

where the *minority member* was inactive, leading to groups of two people sharing the same opinion. This resulted in two types of groups that were retained for the final analysis:

- **2 vs. 1** (unbalanced): all group members were active.
- **1 vs. 1** (balanced): one member dropped out, leaving two members with divergent opinions.

From a total number of 418 groups, we excluded 110 (26%) for the aforementioned reasons, resulting in 308 active groups retained for the analysis, of which 206 (67%) were unbalanced (2 vs. 1) and 102 (33%) were balanced (1 vs. 1) groups.

Data Set. The full data set is published at: https://github.com/crowd-deliberation/data.

3.1.7 Results

Analysis based on descriptive statistics shows that there was more disagreement in the Sarcasm task than in the Relation task, confirming our premise that the Sarcasm task is more subjective. Specifically, we computed the level of label disagreement for each of the 40 cases per task type using the data from the *breadth* run. Label disagreement was represented as entropy:

$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$

where p was the proportion of workers who chose the positive category (e.g., Sarcastic). Entropy values range from 0 to 1; higher values mean more disagreement (e.g., 50% choosing Sarcastic and 50% choosing Not Sarcastic) and lower values indicate less disagreement. Cases had mean entropy values of 0.85 ($SD = 0.22$) in the Sarcasm task and 0.61 ($SD = 0.32$) in the Relation task, indicating less disagreement overall in the Relation task. This difference is statistically significant under a two-sided t-test $t(70) = 3.79, p < 0.001$.

Q1: Why do annotators disagree with one another?

[H1a] To discover the sources of disagreement in each task, we analyzed the anticipated and re-evaluated sources of disagreement that workers provided before and after discussion. As workers were allowed to select multiple options, we use the simultaneous Pearson independence test [18], which takes into account correlation between options. Results (in Tables

Table 3.2: Anticipated sources of disagreement (before discussion) where the proportions of workers are significantly different.

	Percentage of Participants	
Task	Relation	Sarcasm
Missing Context	**49%**	4%
Contrad. Evidence	18%	**43%**

Table 3.3: Re-evaluated sources of disagreement (after discussion) where the proportions of workers are significantly different.

	Percentage of Participants	
Task	Relation	Sarcasm
Fuzzy Definition	28%	**43%**
Missing Context	**24%**	4%
Contrad. Evidence	11%	**23%**

3.2 and 3.3) show that certain sources of disagreement reported by the workers depend significantly on the task type, both, as anticipated before discussion ($\chi^2_S = 61.96, p < 0.001$) and as re-evaluated after the discussion ($\chi^2_S = 89.88, p < 0.001$). This result confirms our hypothesis H1a. A higher percentage of workers anticipated Missing Context to be a dominant source of disagreement in the Relation task (49%) than in the Sarcasm task (4%). Note that we included Missing Context to capture the *extent* to which it leads to disagreement, even though, following the standard TAC KBP guidelines for relation extraction, workers were instructed to avoid inferences based on missing information in the Relation task. Conversely, more workers anticipated Contradictory Evidence to be a dominant source of disagreement in the Sarcasm task (43%) than in the Relation task (18%). After discussion, while the same trends are observed, workers also identified Fuzzy Definition as an additional source of disagreement, which is more prominent in the Sarcasm task (43%) than the Relation task (28%).

[H1b] For Q1, we also investigated workers' ability to predict disagreement. Workers were asked to predict the level of disagreement before the discussion by choosing Substantial Disagreement, Half Agreement, or Substantial Agreement. We determined the ground truth "level of disagreement" for each case by running a two-sided proportion test to see if there is a 50/50 split in group opinions about the label. If this null hypothesis could not be rejected at significance level 0.05, we assumed the correct ground truth answer to be Half Agreement. Otherwise, the correct ground truth answer was based on the label chosen by majority vote. Based on the *depth* analysis data, 7 out of 10 cases in the Sarcasm task and 3 out of 10 cases in the Relation task resulted in Half Agreement.

Since very few workers predicted Substantial Disagreement ($< 1\%$ in each task type), this answer option was grouped together with Half Agreement (known as Disagreement from hereon) to allow for further inferential statistical tests. We measured workers' ability to predict the resulting two discrete levels Disagreement and Agreement (previously known as Substantial Agreement). For both task types, we computed workers' rate of being correct

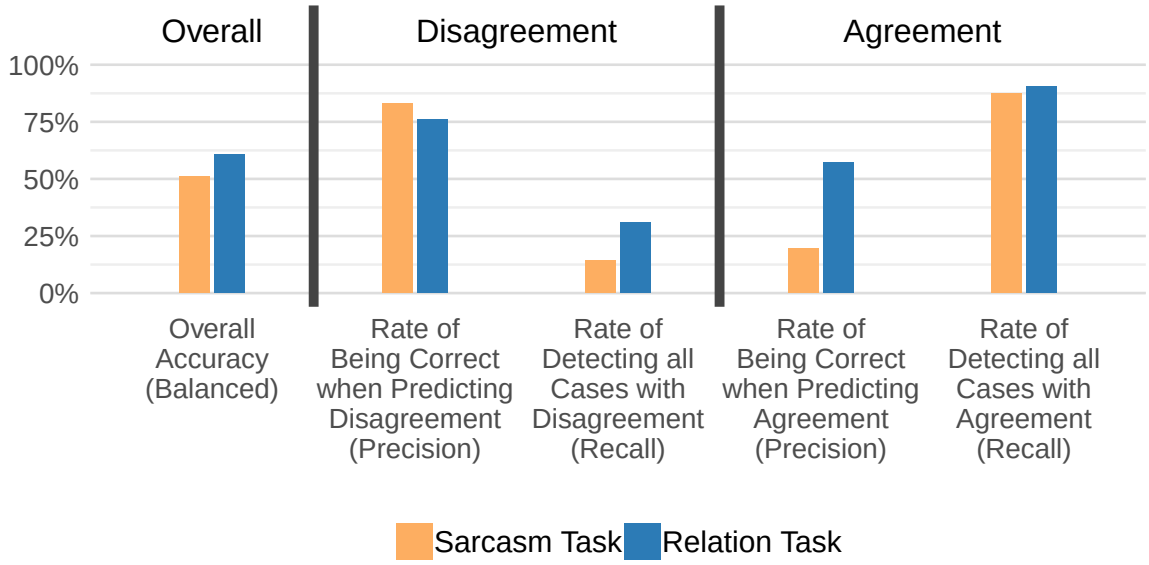


Figure 3.3: Aggregate performance of our worker population at predicting disagreement/agreement by task type.

when predicting Disagreement or Agreement (*precision*), their rate of detecting all cases with Disagreement or Agreement (*recall*), as well as their overall prediction performance, in terms of *balanced accuracy*, a robust measure which accounts for class imbalance, defined as the average of the recall values for Disagreement and Agreement.

Figure 3.3 shows that workers' prediction performance was higher in the Relation task in terms of all the metrics, except for Disagreement precision. In other words, only the rate of being correct when predicting Disagreement was slightly higher in the Sarcasm task; for all other metrics, workers were more successful in the Relation task. In terms of the overall prediction performance, workers' accuracy was significantly better than random in the Relation task at 61% ($\chi^2(1, N = 716) = 158.63, p < 0.001$), whereas this was not the case for the Sarcasm task. This result partially confirms our hypothesis H1b. More interestingly, irrespective of the task type, when workers did predict Disagreement, their rates of being correct were higher than when predicting Agreement; but they were also generally less successful at detecting all cases with Disagreement than those with Agreement.

In summary, for Q1, we confirmed that sources of disagreement differ significantly between the two task types, identified Missing Context as a characteristic source of disagreement for the Relation task, and Contradictory Evidence and Fuzzy Definition for the Sarcasm task (H1a). In addition, we showed that workers can predict levels of disagreement

significantly better than random in the Relation task (H1b).

Q2: Under what circumstances can disagreement be resolved through worker deliberation?

We analyzed the factors contributing to the resolution of disagreement using both quantitative and qualitative analyses of the questionnaire data. Group discussions were considered resolved if and only if all group members converged to one of the two target categories in their final classification.

For the quantitative analysis, a logistic regression model was used to discover factors that affect whether a case is resolved or not. The variables considered include, for each potential source of disagreement, the proportion of workers within a group who selected this source after discussion (H2a), the task type (H2b), various statistics capturing the amount of words contributed by group members (H2c), the level of initial consensus (H2d), and the amount of overlap between the highlighted pieces of evidence among the dissenting parties within a group (H2e), as measured by the Jaccard index, a statistic for the overlap between two sets. The pairwise interactions between the selected factors were excluded due to lack of statistical significance.[1]

Table 3.4: Logistic model for understanding the likelihood of resolving a case.

	Model Parameters			
Variable	$\hat{\beta}$	*Std. Error*	t	*p-value*
Subjective Case	-2.16	0.83	-2.59	**
Fuzzy Definition	-1.45	0.53	-2.73	**
Contrad. Evidence	-1.48	0.65	-2.27	*
# Words Min/Max	-1.49	0.67	-2.24	*
Group 2 vs. 1	-0.67	0.33	-2.04	*
Sarcasm Task	0.61	0.36	1.69	

[H2a] Results (Table 3.4) show that certain sources of disagreement—namely, Subjective Case, Fuzzy Definition, and Contradictory Evidence—decrease the probability of a case being resolved, partially confirming our hypothesis H2a.

[1]Statistically significant results are reported as follows: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*).

In the post-study questionnaire, workers describe some cases as straightforward, requiring only a second glance to reach consensus ("After rereading, I realize the whole review is a joke and the positive points are sarcastic."), while other cases are ambiguous or confusing due to contradictory evidence (e.g., the text contains features of both sarcasm and non-sarcasm) and missing context.

[H2b] We speculated that task subjectivity affects whether a case will be resolved. Results are mixed. Our step-wise regression procedure revealed no detectable differences in case resolvability between the Sarcasm and Relation tasks. However, in the post-study questionnaire, workers in the Relation task mentioned that *well-defined category definitions* helped them resolve disagreement ("We were able to refer to the definition of LivedIn. That made the answer clear."), while the "lack of instruction on what constitutes a sarcastic review" was considered a barrier to resolving cases in the Sarcasm task.

[H2c] We hypothesize that the extent to which members contributed equally predicts whether the case will be resolved. Our model selected # Words Min/Max, i.e., the proportion of words contributed by the least active and the most active contributors within a group, as a significant predictor variable, confirming hypothesis H2c. An increase in this proportion decreases the likelihood of a case being resolved. In other words, if members contributed equally, cases were significantly less likely to be resolved than if some members contributed substantially more than others.

Qualitative analysis also shows that interactions between members of the discussion groups can influence the final outcomes. Workers said that disagreement was resolved by clarifying the *task* ("Members were becoming clearer on interpretation of the task."), providing *examples* ("Using examples, we were able to persuade the group member to change their opinion."), using *evidence* ("We were able to point out things in the sentence that were overlooked by others."), or pointing out *false assumptions* ("Others pointed out things that some of us had assumed in the sentence and changed our opinions.") Many workers identified the *quality of deliberation activity* itself as the main driver for reaching consensus in an argumentative manner, e.g., "People that didn't agree listened to the arguments and were willing to change their mind to sarcastic.") On the other hand, workers also reported some group-related factors that hindered the resolution of disagreement, including divergent, but equally valid *interpretations* ("I think it depends on how people view sarcasm.") and the *lack of communication* ("[My group was] not continuing a conversation. Responding with one word responses does not solve anything.")

[H2d] A consensus level of 2 vs. 1 (i.e., two workers agree, one worker disagrees) also decreases the likelihood of a case being resolved compared to the consensus level of 1 vs. 1, confirming our hypothesis H2d.

[H2e] Finally, the overlap in evidence among group members was not selected as a relevant factor for the optimal model fit, leading us to reject hypothesis H2e.

To summarize the findings for Q2, we found various factors that affect whether a case is resolved or not, including some sources of disagreement (H2a), task type (H2b), the degree to which deliberation activity is balanced within a group (H2c) and the level of initial consensus (H2d). Other factors like overlap in evidence highlighted by dissenting parties (H2e) had no significant effect on resolvability. Finally, our analysis shows that worker characteristics (such as age and personality) can have some influence on the way they deliberate (e.g., their overall tendency to revise their position), which in turn can play a role in resolving a case.

Q3: What impact does the deliberation workflow have on crowdsourcing outcomes and processes?

[H3a] To evaluate whether worker deliberation improves crowdsourcing outcomes, we compared workers' answer correctness between our three experimental conditions **Reconsider Only**, **Disagree and Reconsider**, and **Disagree, Discuss and Reconsider** and our **Baseline** (i.e., original labels submitted in stage A). This part of the analysis is restricted to the 25 cases from the *breadth* run, where we have ground truth to assess correctness.

Correctness was measured by F1-score (i.e., the harmonic mean of precision and recall) of *individual* workers. We did not aggregrate labels across multiple workers as our workflow did not contain a requirement for shared unanimous group decisions, but instead incentivized independent reconsideration from individual workers in all conditions. Assessments that were revised to "irresolvable" were excluded because they were neither right nor wrong. Overall, correctness was significantly higher in **Disagree, Discuss and Reconsider**, with no significant differences between the other three conditions. Figure 3.4 provides a visual comparison. The statistical significance of these differences was confirmed by a one-way analysis of variance ($F(3, 115) = 6.42$, $p < 0.001$), followed by pairwise comparisons with Holm-Bonferroni correction. The pairwise comparisons confirmed that the **Disagree, Discuss and Reconsider** workflow resulted in significantly higher F1-scores than **Baseline** ($t(64) = -3.44, p < 0.01$), **Reconsider Only** ($t(44) = -4.38, p < 0.001$), and **Disagree and Reconsider** ($t(25) = -3.02, p < 0.05$). There were no detectable differences among the other pairings. These results confirm our hypothesis H3a that worker deliberation improves the quality of the crowdsourced annotations. Furthermore, this result suggests that the improvement in correctness is not due to learning effects or knowledge about group disagreement data, but due to the actual discussion process.

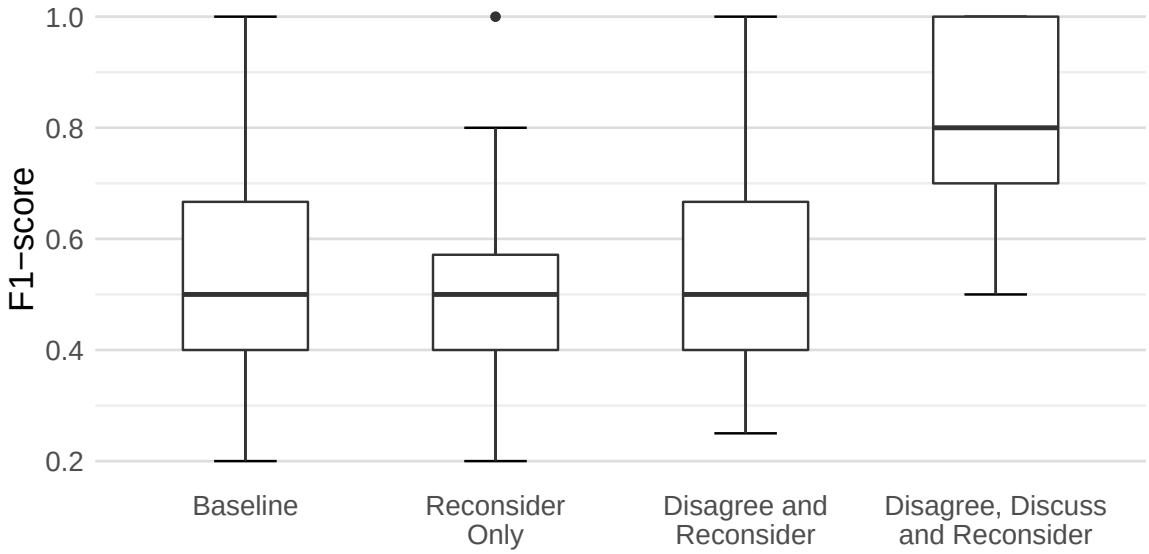


Figure 3.4: Individual workers' answer quality in the Relation task across different reconsideration workflows.

[H3b] For Q3, we were also interested in whether our proposed deliberation workflow discourages *groupthink*, an undesirable effect where discussion members tend to agree with the original majority answer within a group. For this question, we performed a close-up analysis on all groups with a composition of 2 vs. 1, and analyzed whether the proportion of cases resolved in favour of the original majority vote was similar to the proportion of cases not resolved in favour of the original majority vote. We used a two-sided proportion test which confirmed our null hypothesis H3b that both outcomes were not detectably different, $\chi^2(1, N = 136) = 0.60, p = 0.44$. In other words, unbalanced discussion groups were equally likely to converge to the original majority opinion as they were to achieve the opposite outcome, i.e., leave the case unresolved or converge to the original minority opinion. Our quantitative results provide evidence that our proposed deliberation workflow effectively discourages *groupthink*.

[H3c] For Q2, we identified factors that contribute to the resolvability of a case. For hypothesis H3c, we investigate whether some of the factors considered in Q2, e.g., sources of disagrement and the extent to which members contributed equally, also predict whether or not a case will be resolved *correctly*. We performed an analysis on the subset of cases for which ground truth was available and which were resolved in the deliberation process. We used a similar step-wise logistic regression procedure as for Q2, defining the *correctness* of

the final consensus label as the dependent variable and including the same set of predictor variables. Results (Table 3.5) show that certain sources of disagreement made the *correct* resolution of a case more likely (Missing Context) or less likely (Expertise Needed). The extent to which members contributed equally (# Words Min/Max) was a negative predictor for correct resolution. These results partially confirm our hypothesis H3c.

Table 3.5: Logistic model for understanding the likelihood of resolving a case correctly.

	Model Parameters			
Variable	$\hat{\beta}$	*Std. Error*	*t*	*p-value*
Expertise Needed	-3.89	1.95	-1.99	*
Missing Context	2.91	1.42	2.05	*
# Words Min/Max	-3.43	1.73	-1.99	*

To summarize, we provide evidence showing that worker deliberation significantly improves the quality of crowdsourced annotations (H3a), while discouraging undesirable behaviour such as *groupthink* (H3b), and that certain sources of disagreement and the extent to which members contribute equally help predict the correctness of the final resolution (H3c).

3.1.8 Discussion

In this section, we studied how deliberation affects resolvability and answer accuracy, and how deliberation outcomes depend on task subjectivity. Our results demonstrate that legitimate reasons for disagreement can vary by task, and worker deliberation can help resolve some of these cases. Importantly, we identified several factors (such as the level of initial consensus, the amount and quality of deliberation activities, and sources of disagreement) that contribute to case resolvability. We reinforced the finding that deliberation can increase answer accuracy and the importance of verbal discussion in this process. Finally, our deliberation workflow discouraged undesirable behaviour, such as *groupthink*. Our anonymized data set is publicly available for research reuse.

Generalizability and Applications

Our empirical findings are based on experiments using two specific task types, one subjective sarcasm detection task, and one objective (person-to-place) relation extraction task.

Caution is warranted for translating these results to a broader set of classification tasks and data modalities. However, the deliberation workflow we propoosed is general, and can be applied to classification tasks involving other data types (e.g., images, videos, time series data), given minor modifications to the interface to facilitate annotation of evidence in other modalities. In some complex scenarios requiring hierarchical decision processes, our workflow will need to be modified. For example, in image classification, individual image features might be disambiguated first before resolving disagreement on the image level. Our results revolve around atomic classification problems in paid crowdwork, and it is expected that deliberation processes in more complex settings would give rise to more complex dynamics (e.g., circular disagreement in sequential classification).

In general, deliberation workflows like ours can be particularly useful in domains where inter-rater disagreement is rather the norm than exception, e.g., medicine [44, 36, 118]. Beyond enabling workers to discuss and reconsider ambiguous cases, our deliberation workflow collects rich information from human annotators (e.g., confidence levels, arguments, assumptions, examples, inferences, relevant features from the data, and meta information about sources of disagreement) that can be used to teach both humans and machines. For example, prior work has shown that asking annotators to highlight relevant features in input data (i.e., words in a text document or regions in an image) can improve machine inference in sentiment analysis tasks [133, 151, 152, 153] and visual category learning [39]. More generally, we posit that future efforts towards *interpretable* machine learning can use data produced through group discussion to analyze, replicate and mimic human decision making processes. While some of the less structured deliberation output (e.g., verbal arguments) may not yet be fully parsed by automated methods, human learners could significantly benefit from edge-case examples coming with discussions and highlighted features from more experienced annotators. Existing efforts to optimize or harness workers' ability to learn complex tasks [41, 104] could thus leverage data produced through worker deliberation.

Another outcome of deliberation could be the refinement and disambiguation of annotation guidelines or category definitions for expert tasks. For example, scoring manuals in the field of medical imaging undergo regular revisions to increase inter-rater reliability [36, 118], a procedure which could benefit from structured and web-based deliberation workflows.

Design Considerations for Classification Tasks

Redundant Labeling. Our results demonstrate that workers can predict levels of inter-rater disagreement significantly better than random for certain task types. Prior work has shown

that being able to predict answer diversity can reduce cost because fewer annotators are needed when answer agreement is expected [64]. Enlisting human capabilities to predict answer agreement incurs minimal extra cost because it requires one additional human response for each data object.

Task Interfaces. Our results show that certain sources of disagreement are more likely to occur for some task types than for others. In classification tasks (like our Relation task) that require annotators to make decisions solely based on the information provided in the data, the interface could remind annotators who indicate "Missing Context" as the reason for anticipating disagreement to minimize ungrounded assumptions about any latent contextual information.

Deliberation Workflow. We found a variety of factors that affect case resolvability. Deliberation workflows for crowdsourced classification tasks should therefore be equipped with incentive mechanisms to reduce or strengthen the effect of certain factors in a task-specific manner. For example, in task types where objectivity is possible and the goal is to find one correct answer, deliberation procedures should have incentives to reduce the impact of undesirable behaviour (e.g., stubbornness or lack of care) and undesirable group dynamics (like *groupthink* or lack of communication) on the final discussion outcome. Deliberation workflows for more subjective classification tasks where the goal is to uncover multiple divergent, but equally valid interpretations of the task and data (e.g., sentiment analysis, relevance rating, text translation) should incentivize group members to be assertive about their interpretation, and change their assessment only under certain conditions (e.g., when false assumptions or illogical conclusions are pointed out).

Various deliberation systems have been proposed in complex domains like public deliberation [50, 84, 85], on-demand fact checking [83], and knowledge base generation [154]. While our study is primarily embedded in the domain of paid crowd work, our empirical findings and some aspects of our workflow design may be informative for deliberation systems in general. For example, our insight that sources of disagreement, when captured in structured form, can help predict case resolvability (H2a) could be leveraged to categorize debates and streamline consensus building. Our finding that equal contribution among discussion members seemed to negatively affect case resolvability (H2c) *and* final answer correctness (H3c) was surprising as balanced contribution is often considered beneficial for fruitful discussion. This finding could inspire future investigations into the balance between active and passive forms of contribution to deliberation, related to work by Kriplean et al. [85] on active listening in web discussions. Another potentially counter-intuitive finding of our work is the fact that unbalanced (2 vs. 1) groups were *less* likely to resolve a case than balanced (1 vs. 1) groups (H2d). To our knowledge, our study is the first to investigate the effect of initial consensus level on case resolvability by having multiple independent groups

of size two or three discuss the same case in crowdsourced classification tasks. While one may expect that unbalanced groups converge faster, one possible explanation for our observation is the added complexity of communication and coordination among three versus two group members. We showed that a perceived lack of expert knowledge was associated with resolving disagreements incorrectly (H3c). This result has interesting connections to prior work on integrating on-demand fact checking into public deliberation [83], suggesting that crowdsourcing workflows could benefit from similar approaches for on-demand provision of expertise.

Limitations

We studied the effects of deliberation in the context of two binary classification tasks and small groups consisting of only two or three members. In practice, many classification problems have more than two classes and discussion groups can also be larger. Future work can investigate new methods for scaling to more complex classification problems and more complex group structures.

A practical limitation of our workflow is its reliance on multiple rounds of synchronous communication and potential attrition between stages. Dropout was highest after the first stage in our workflow, suggesting promising future research on incentives for workers to return for multiple consecutive sessions, or on systems that initiate discussions immediately as disagreement arises before workers leave the platform.

Our deliberation workflow enables workers to consider alternative views on ambiguous cases through discussion, and produces useful data such as arguments, examples, and evidence. However, deliberation also incurs a cost in terms of time. A promising area for future work is to develop techniques to improve the *efficiency* of the deliberation process [42] and to characterize the cost-benefit of using real-time deliberation in task workflows. Another direction for future work is to explore other deliberation incentives beyond monetary compensation, including peer-based reputation systems for constructive deliberation [52]. In more complicated incentive schemes, moderators could be rewarded for establishing a balanced and constructive deliberation among group members.

3.1.9 Conclusion

This section contributes novel insights into the circumstances and outcomes of worker deliberation for handling inter-rater disagreement in crowdsourced text classification tasks with varying degrees of inherent subjectivity. Based on a custom-designed workflow for

real-time worker deliberation, we investigated the impact of various factors on the probability that a disagreement among small groups of crowdworkers will be resolved through synchronous group discussion. Our results suggest that the reasons for and level of the initial disagreement, the amount and quality of deliberation activities, as well as the task and case characteristics play a role for resolvability. To encourage future work in the field of worker deliberation, we publish our data set including all original classifications, discussion comments, text highlights, and revised positions from crowdworkers. Future work includes developing and validating new deliberation protocols and demonstrating how the information produced by worker deliberation can be used to train both humans and machines.

3.2 Expert Deliberation for Image Labeling

With this second study on group deliberation in data labeling, our focus starts to shift from the domain of novice crowd work to the expert domain of medical data interpretation. Note that in the medical domain, the term *adjudication* is commonly used as a synonym of group deliberation. We therefore use both terms interchangeably throughout the remainder of this dissertation.

This section explores the process of group deliberation among medical doctors in the context of interpreting and classifying medical images for the presence and severity of diabetic retinopathy (DR), a complication of diabetes that can lead to vision loss. In particular, we present and evaluate a remote, tool-based and asynchronous system and structured grading rubric for adjudicating image-based assessments. We compare three different procedures for adjudicating DR severity among panels of retina specialists: (1) in-person adjudication (Baseline); (2) remote, tool-based adjudication for assessing DR severity alone (TA); (3) remote, tool-based adjudication using a feature-based rubric (TA-F). Our results suggest that remote, tool-based adjudication presents a flexible and reliable alternative to in-person adjudication for DR diagnosis, and that feature-based rubrics can help accelerate consensus for tool-based adjudication of DR without compromising label quality.

3.2.1 Motivation

Diabetic retinopathy (DR) is one of the leading causes of vision loss worldwide [141]. The process of grading DR severity involves the examination of the retina and the assessment of several features, such as microaneurysms (MAs), intraretinal hemorrhages, and neovascularization [6]. In a teleophthalmology setting for remote screening, certified graders examine retinal fundus images to determine the presence and severity of disease as it appears in a two-dimensional (2D) photograph [134]. Prior work has shown that this process of human interpretation is subject to individual grader bias, as demonstrated by high intergrader variability, with kappa scores ranging from 0.40 to 0.65 [1, 132, 87, 54, 119, 88].

This moderate-to-poor agreement between graders has led to difficulties in reliable evaluation of both individual graders as well as assistive technologies. Yet, due to limited access to skilled healthcare providers, there continues to be a surge in interest in the development of assistive technologies, such as deep-learning systems, resulting in a sharp increase in the demand for high-quality reference standards of labeled-image data [63, 82, 115, 62, 140, 122, 47, 58, 56, 4]. Prior work has examined different methods for

resolving disagreements among experienced graders when creating a reference standard [82], including majority vote, arbitration of disagreements by a more senior grader, and in-person adjudication among expert panels.

In ophthalmology, a recognized method to obtain a reliable reference standard is expert adjudication of images [82, 142, 49]. Multiple experienced doctors independently grade images and discuss disagreements until resolved. Such "in-person" adjudication has been shown to produce higher-quality labels [82] but can be challenging to schedule: it requires coordination of multiple, highly experienced specialists for in-person sessions, and even small image sets on the order of a few thousand cases can take months to adjudicate due to clinical scheduling conflicts.

In this study, we presented and evaluated a tool-based system for adjudicating images that was suitable for remote grading and removes the need for in-person sessions. Our system allowed doctors to discuss and resolve disagreements on diagnoses remotely, without convening at a set time and place. The practices described in this section aimed to increase the efficiency and flexibility of adjudication, while maintaining the quality of the labels produced. We evaluated our system in the context of DR severity grading based on retinal fundus images.

In addition, we proposed an adjudication system with the ability to impose an explicit structure on the adjudication process by organizing the process of image interpretation around a set of discrete, detailed evaluation criteria. We investigated the effects of such a structure on the efficiency and reliability of adjudication for DR grading. Specifically, we presented a feature-based rubric for adjudication of DR severity grades, in which graders assess individual features (MAs, hemorrhages, neovascularization, etc.) in addition to overall DR severity.

Taken together, these improvements allow high-quality reference standards to be obtained by the community, and have the further benefit of offering flexibility for individual graders to schedule their reviewing activity around their clinical duties.

3.2.2 Methods

Experimental Design

The experiment conducted for this study compared three different adjudication procedures for assessing DR severity based on retinal fundus images as follows: in-person adjudication (Baseline); remote, tool-based adjudication (TA) for assessing DR severity alone; and remote, tool-based adjudication using a feature-based rubric to assess DR severity (TA-F).

Table 3.6: Baseline Characteristics

Characteristic	Value
Number of images	499
Number of images for which an anonymized patient code was available[a]	330
Number of unique individuals out of the images for which a patient code was available	307
DR gradeability distribution according to Baseline adjudication	
Images gradable for DR, *n*/total (%)	472/499 (94.6)
DR severity distribution according to Baseline adjudication, *n* (%)	
No apparent DR	217 (45.9)
Mild NPDR	17 (3.6)
Moderate NPDR	108 (22.9)
Severe NPDR	72 (15.3)
PDR	58 (12.3)

PDR, proliferative diabetic retinopathy.
[a] Patient codes were available for images from two hospitals (Sankara Nethralaya and Narayana Nethralaya) of three.

The experiment implemented a between-subjects design in which independent panels of three retina specialists each graded and adjudicated the same set of images following one of three adjudication procedures (Baseline, TA, TA-F). We describe the image set, each of the three adjudication procedures, and details about the retina specialist graders below. For each design, graders were primarily assessing DR severity, but not diabetic macular edema (DME).

Image Set

We used a subset of 499 images (Table 3.6) from the development dataset used by Krause et al. [82] The image set consisted of central field-of-view images obtained from patients who presented for DR screening at three eye hospitals in India (Aravind Eye Hospital, Sankara Nethralaya, and Narayana Nethralaya). The image set was sampled to include approximately 50% of cases that had some level of DR [82]. Anonymized patient codes

were provided from two of three hospitals, allowing us to verify no patient duplication. For the third hospital, patient codes were not provided; this allows for the possibility that 169 images from this hospital may contain multiple images from the same patient; given that these were sampled from a much larger set of images, duplication is unlikely. Image sizes ranged from 640×480 to 2588×3388 pixels, and were presented to adjudicators at the original resolutions. All images were de-identified according to the Health Insurance Portability and Accountability Act Safe Harbor before transfer to study investigators. Ethics review and institutional review board exemption were obtained through the Quorum Review institutional review board (Seattle, WA).

Adjudication Procedures

Baseline Adjudication. Following the practices described in Krause et al. [82], our Baseline adjudication procedure consisted of the following three stages: (1) an initial independent evaluation; (2) remote review of disagreements; and (3) in-person discussion and final resolution of remaining cases.

For the first stage, three fellowship-trained retina specialists undertook independent grading of the image set. Images in which the independent graders agreed were considered resolved. Next, each of the three retina specialists independently reviewed one-third of the remaining images with any level of disagreement. This independent review procedure was facilitated through the use of online spreadsheets. Cases that remained unresolved after the independent review round were discussed by all three retina specialists in person. During the in-person sessions, all three retina specialists were present at a set time and place, and conflicting grades were reviewed and adjudicated within the panel until all specialists came to an agreement. The time from start of independent grading to full adjudication for the image set was around 3 months. While the total time each grader spent on grading and adjudication activities was not tracked precisely, a substantial portion of the 3-month period was due to difficulties in scheduling the retina specialists to physically convene for in-person discussions.

Tool-Based Adjudication (TA). To ensure the continuity of the adjudication process and to reduce the logistic overhead associated with in-person adjudication, we designed and implemented a tool-based system for remote adjudication that removes the need for in-person sessions (Figure 3.5). Similar to the Baseline procedure, the TA procedure commences with independent grading: each panel member first assesses each image for DR severity. Next, those images with any level of disagreement are reviewed by one panel member at a time in a round-robin fashion until agreement is reached for the given case (Figure 3.6). For each review round, the active grader reviews all grades and comments

Table 3.7: Comparison of adjudication procedures

Property	Adjudication Procedure	
	Baseline	Tool-Based (TA and TA-F)
Image viewer	Web-based image viewer with built-in tools to adjust zoom level and contrast settings; graders submitted their independent assessments using prompts embedded into the image viewer	
Aggregation of grades and identification of disagreements	Exporting results into spreadsheet to manually identify disagreements	Automated process to identify images with disagreement in the grades database
First review round	Remotely in spreadsheet	Remotely, using the web-based image viewer; one grader at a time in a round-robin fashion
Subsequent review rounds	In-person session; all panel members convene at a set time	
Channel for discussion	In-person verbal discussion	Discussion thread integrated into the image viewer; up to one written comment per grader per review round
Scheduling of review rounds	Manual process	No manual scheduling required; grading and review tasks automatically queue up for individual graders in the online platform
Anonymization of graders	Possible only in the first review round, but not during live discussion	Possible throughout the entire procedure
Organization of the disagreement discussion around a set of explicit diagnostic criteria (e.g., lesions)	Challenging to implement during live discussion	Possible using prompt structure integrated into the image viewer

provided in previous rounds, re-grades the given image for DR severity, and provides more detailed comments, or replies to other graders' comments. To handle cases with persistent disagreement, the TA procedure imposes a limit on the number of review rounds for each case. In our studies, each case was limited to a maximum of 15 review rounds (i.e., 5 reviews per grader for a panel of 3 graders). See Table 3.7 for a comparison of the Baseline and TA adjudication procedures.

Tool-Based Adjudication With Feature Rubric (TA-F). Disagreements over DR

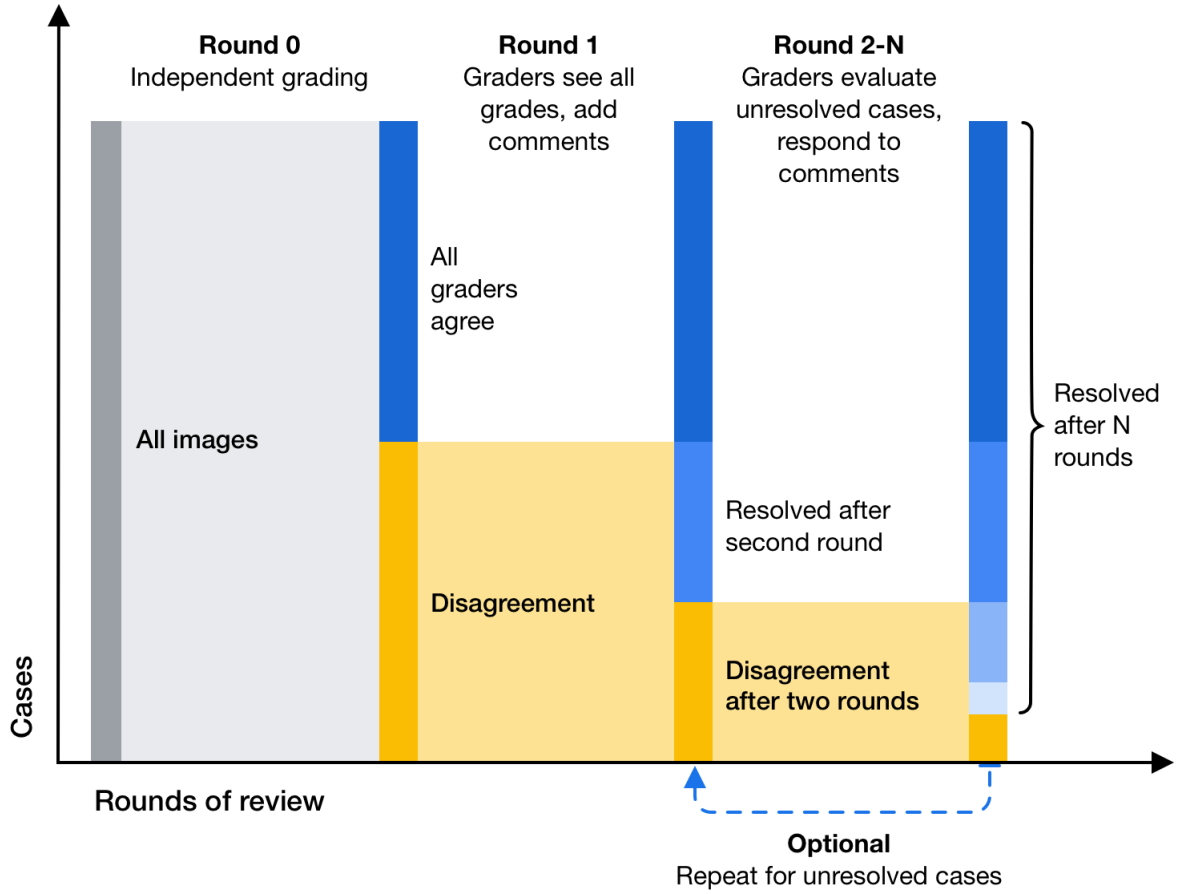


Figure 3.5: Process diagram illustrating remote TA; images are first graded independently by each panel member (round 0); cases with any level of disagreement after independent grading are reviewed by all graders in a round-robin fashion (rounds 1–N); the procedure ends after N review rounds.

severity can arise for various reasons (e.g., due to divergent assessments of the presence and extent of individual features or due to divergent interpretations of whether a retinal pathology is diabetic in nature or not). One benefit of the tool-based adjudication procedure proposed in this section is the ability to impose an explicit structure to the adjudication process by introducing prompts for individual, detailed evaluation criteria. This ability can be leveraged to remind graders of the specific criteria they should apply to assess an image (e.g., from standardized grading guidelines) so that discussions over potential disagreements are grounded in predefined factors relevant to the overall diagnostic

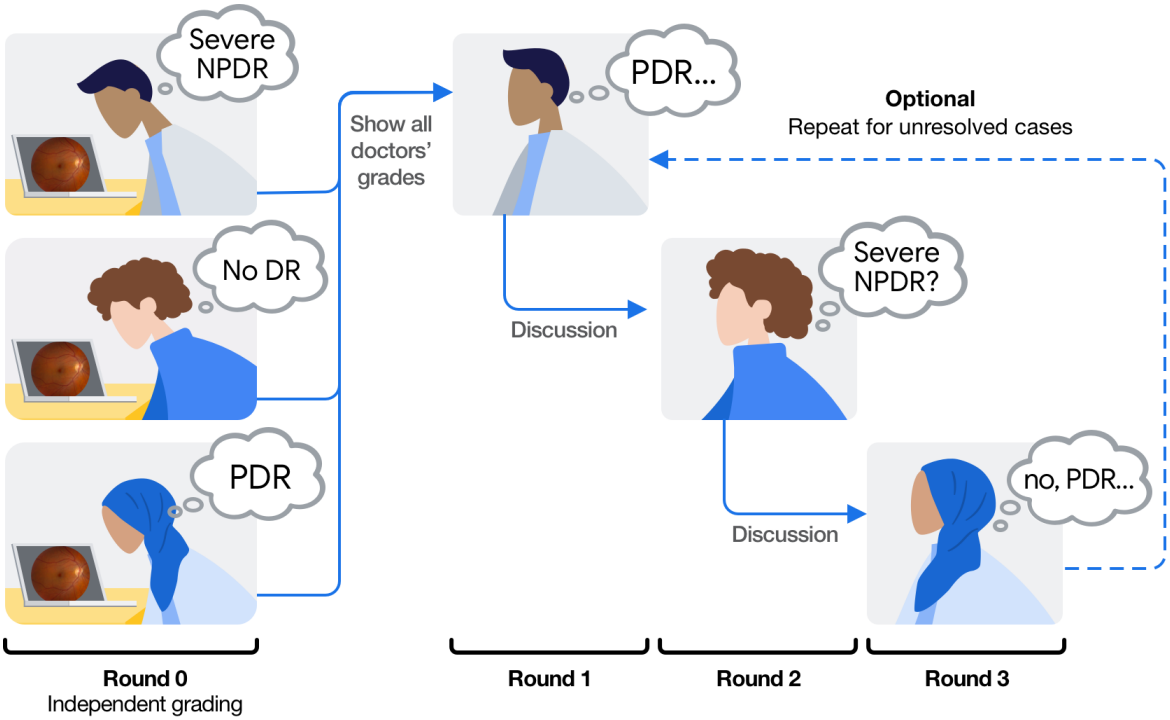


Figure 3.6: Illustration of the round-robin approach for remote TA in the context of DR severity grading.

decision.

In our experiment, we developed a feature-based rubric in which graders were first prompted to assess each image for a set of DR-related features before assessing the image for overall DR severity. Following the International Clinical Diabetic Retinopathy (ICDR) disease severity scale [6], we included the following 10 features in the TA-F procedure: MAs, cotton-wool spots, hard exudates, retinal hemorrhage (heme), venous beading (VB), intraretinal microvascular abnormalities (IRMA), neovascularization or fibrous proliferation, preretinal or vitreous hemorrhage, laser scars from panretinal photocoagulation, and laser scars from focal photocoagulation. Graders assessed whether each feature was present, not present, or ungradable. For heme, graders also assessed whether any retinal hemorrhage was extensive in four quadrants, based on standard photo 2A from the Early Treatment Diabetic Retinopathy Study (ETDRS) [2]. For VB, graders also assessed whether definite venous beading, if present, was observed in two or more quadrants, based on ETDRS standard photo 6A [2]. Similarly, for IRMA, graders assessed whether any IRMA was prominent, based on ETDRS standard photo 8A [2]. Intergrader disagreement may not

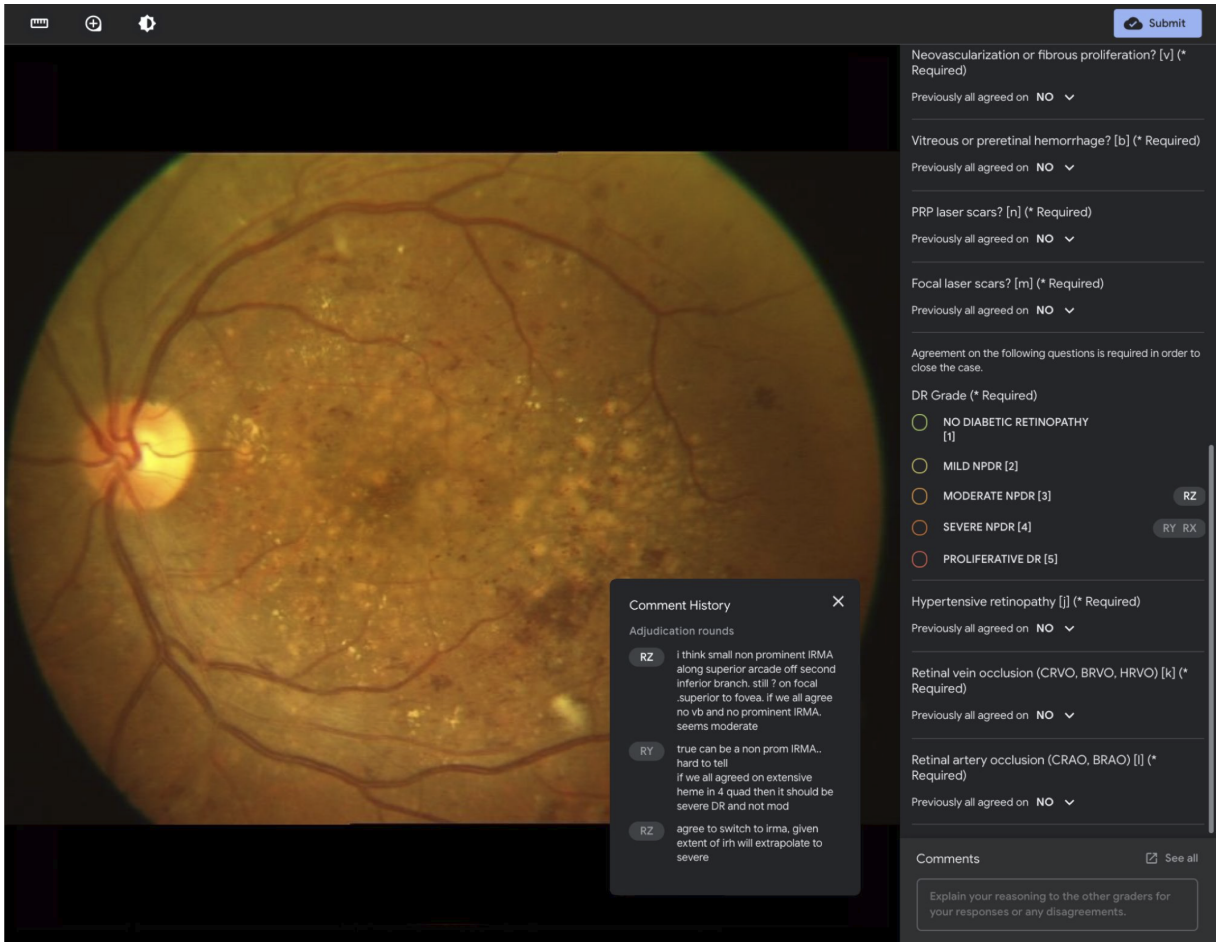


Figure 3.7: Grading interface for remote TA-F for DR severity assessment. Grader pseudonyms (RX, RY, RZ) are used to associate grading decisions and discussion comments from previous rounds with specific (anonymized) grader identities. The current grader's pseudonym is highlighted with *bold white font* (see RZ). The panel on the *right-hand side* lists all prompts included in the TA-F procedure.

only arise over the presence or severity of disease, but also over the specific classification and etiology of an observed pathology. In particular, the appearance of DR may resemble other forms of retinal disease, such as hypertensive retinopathy (HTN), retinal vein occlusion (RVO), and retinal artery occlusion (RAO) [14, 17]. Graders were therefore prompted to assess for the presence of HTN, RVO, and RAO in addition to providing a DR severity assessment. In the adjudication interface (Figure 3.7), disagreements were visualized for

both feature- and diagnosis-level decisions to inform adjudicators about assessments from other panel members. Full agreement within a panel was only required regarding the overall gradeability of an image as well as for the diagnosis decisions (DR, HTN, RVO, RAO) in order to resolve a case; cases could be resolved despite disagreements on individual features.

Graders

We recruited 14 American Board of Ophthalmology–certified fellowship-trained retina specialists to form five adjudication panels, including one panel for the Baseline procedure, two panels for the TA procedure (Panels A and B), and two panels for the TA-F procedure (Panels C and D). Due to the limited availability of retina specialists, one of 14 graders participated in two panels (Baseline and TA Panel B); otherwise, each grader participated in one panel only. Participating retina specialist graders completed their fellowship training between the years 2009 and 2017 and the number of years in practice (post fellowship) at the time of participation in the study ranged from 0.5 to 8.5 years.

Evaluating Tool-Based Adjudication

We evaluated the tool-based adjudication procedures (TA, TA-F) for reliability and efficiency. Reliability was assessed in terms of agreement with the Baseline adjudication procedure, using Cohen's quadratically weighted kappa score [31]. A nonparametric bootstrap procedure [30] with 2000 samples was used to compute confidence intervals (CIs) for the kappa scores. The weighting function for the calculation of kappa scores was the square of the stepwise distance between DR grades on a five-point ordinal scale (e.g., a disagreement between no DR and severe nonproliferative diabetic retinopathy [NPDR], which are three steps apart on the ICDR scale, would receive a weight of $3^2 = 9$ when calculating kappa; larger disagreements would more strongly reduce this metric). Images unanimously deemed ungradable and those with persistent disagreement after 15 review rounds in any of the panels were excluded from kappa score calculations. Exact agreement rates and strikeout rates (i.e., the fraction of images for which grades differed by >2 steps) were calculated as additional measures of agreement for each panel pair.

The efficiency of TA versus TA-F was evaluated using the number of review rounds required to resolve each case in a given panel, and using the cumulative percentage of cases resolved in each round, including independent grading (round 0) and the subsequent review rounds (rounds 1–15). We used the standard permutation test to assess the statistical

significance of these differences [30]. Due to software-related irregularities, in which the full-adjudication discussions were not recorded, 11 images (2%) were excluded from the analysis. Results are based on the remaining 488 cases. For TA-F specifically, the relative efficiency of resolving disagreements on each of the rubric criteria was assessed as the number of review rounds required to reach agreement on a given criterion, or as the round number in which a case was closed despite disagreement on the criterion.

3.2.3 Results

Reliability

	TA		TA-F	
Parameter	Panel A	Panel B	Panel C	Panel D
Baseline	0.948 (0.931–0.964)	0.943 (0.919–0.962)	0.921 (0.886–0.948)	0.963 (0.949–0.975)
TA				
Panel A	/	0.932 (0.911–0.950)	0.917 (0.885–0.944)	0.939 (0.916–0.960)
Panel B	/	/	0.911 (0.873–0.942)	0.936 (0.914–0.953)
TA-F				
Panel C	/	/	/	0.919 (0.882–0.949)

Values are quadratically weighted Cohen's Kappa (95%CI).

Table 3.8: Inter-panel agreement among all pairs of panels as Cohen's kappa.

	TA		TA-F	
Parameter	Panel A	Panel B	Panel C	Panel D
Baseline	0.820	0.828	0.789	0.857
TA				
Panel A	/	0.811	0.811	0.852
Panel B	/	/	0.822	0.816
TA-F				
Panel C	/	/	/	0.820

Table 3.9: Inter-panel agreement among all pairs of panels as exact agreement.

	TA		TA-F	
Parameter	Panel A	Panel B	Panel C	Panel D
Baseline	0.026	0.026	0.027	0.017
TA				
Panel A	/	0.039	0.041	0.038
Panel B	/	/	0.042	0.034
TA-F				
Panel C	/	/	/	0.033

Table 3.10: Inter-panel agreement among all pairs of panels as strikeout rate.

Remote TA grades showed high agreement with the Baseline adjudication procedure (Table 3.8), with Cohen's kappa scores of 0.943 (95%CI, 0.919–0.962) and 0.948 (95%CI,

0.931–0.964) for the two panels assessing DR severity alone without the use of a feature rubric (TA), and 0.921 (95%CI, 0.886–0.948) and 0.963 (95%CI, 0.949–0.975) for the two panels using the feature-based rubric (TA-F). Both TA and TA-F showed high rates of reproducibility, as measured by the Cohen's kappa score between the two independent panels for each procedure. The kappa score for agreement was at 0.932 (95%CI, 0.911–0.950) between the two panels in the TA procedure and at 0.919 (95%CI, 0.882–0.949) for TA-F. Exact agreement rates (Table 3.9) and strikeout rates (Table 3.10) are reported as additional measures of agreement for each pair of panels.

Efficiency

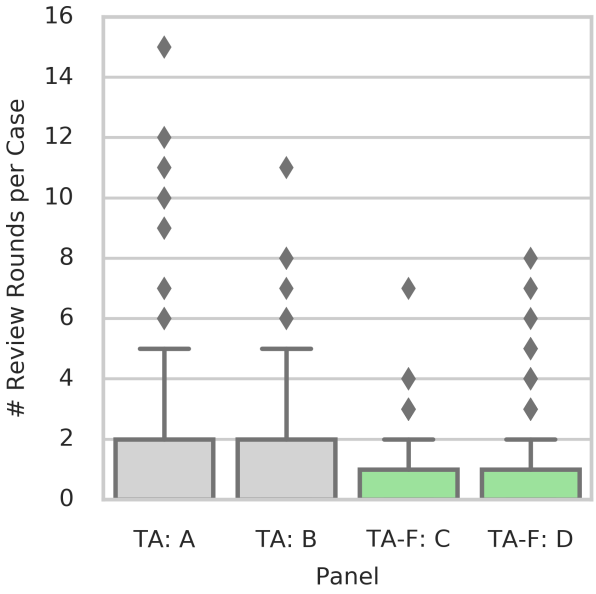


Figure 3.8: Number of review rounds required per case (i.e., number of rounds until agreement or 15 in case of persistent disagreement) for each of the four adjudication panels.

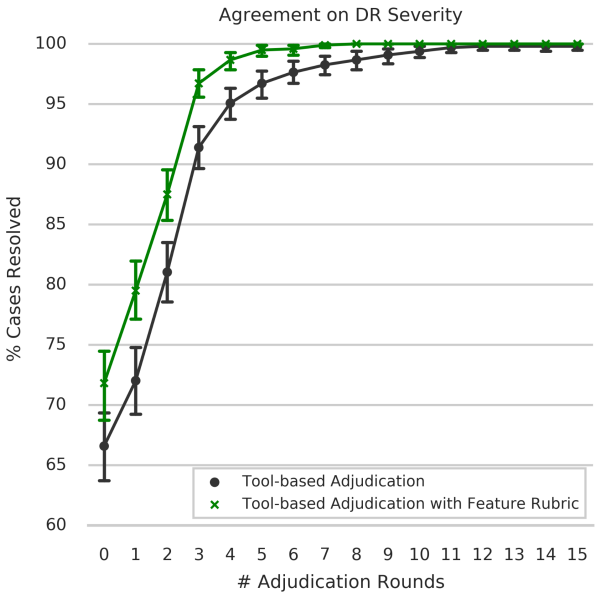


Figure 3.9: Cumulative percentage of cases resolved per adjudication round for TA procedures.

Cases adjudicated using TA-F were resolved in significantly fewer rounds compared with assessing DR severity without the rubric (TA; $p < 0.001$; permutation test, Figure 3.8). During independent grading (round 0), graders were in agreement for 72% of all cases using TA-F, compared with 67% TA, and to 58% in Baseline in-person adjudication. Using

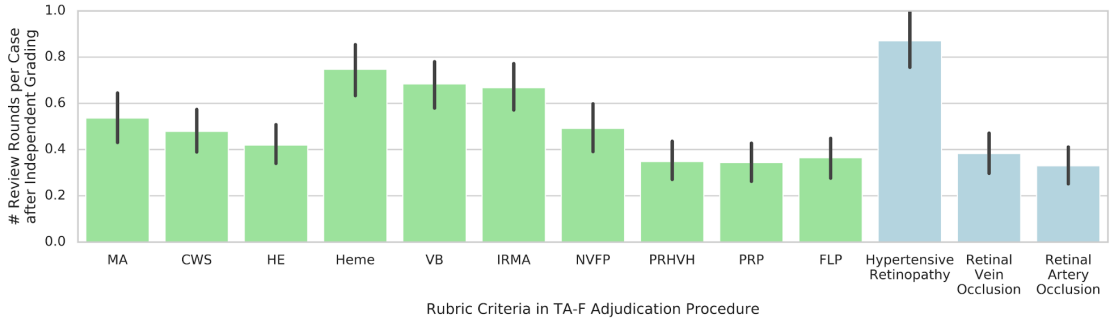


Figure 3.10: Mean number of review rounds required per rubric criterion in remote TA-F. The Y axis indicates the number of rounds after independent grading until either agreement was reached for the given criterion; or the case was closed due to overall agreement on the diagnosis level. Note that the mean number of review rounds may be below 1 because cases not requiring adjudication due to independent agreement were considered to have 0 review rounds. Green bars correspond to feature criteria, blue bars correspond to differential diagnosis criteria. Error bars indicate the 95% confidence intervals. CWS, cotton-wool spot; HE, hard exudate; NVFP, neovascularization or fibrous proliferation; PRHVH, Preretinal or vitreous hemorrhage; PRP, pan-retinal photocoagulation scars; FLP, focal laser photocoagulation scars.

TA-F, only 3% of the cases required more than one full "round-robin" of reviews from the panel (round 3), compared with 9% of the cases in the absence of the feature-based rubric (Figure 3.9). Both differences were statistically significant under a permutation test of two panels for TA versus two panels for TA-F ($p = 0.004$ for round 0, $p < 0.001$ for round 3). The only two cases with persistent disagreement after 15 rounds of review were observed in the TA procedure (Panel A). Overall, cases assessed as mild NPDR or severe NPDR in the Baseline adjudication procedure showed the lowest rates of independent agreement (i.e., before adjudication), with agreement rates of 31.7% and 44.6%, respectively. Mild NPDR and severe NPDR were also the only two categories with any persistent disagreement (1 case each), and with the highest proportion of cases requiring more than two rounds of review for at least one of three graders (3.3% and 1.8%, respectively) in order to reach a consensus.

Among the 10 feature criteria included in the TA-F rubric, assessments of the presence and extent of heme, VB, and IRMA required the greatest number of review rounds on average (Figure 3.10). As for the differential diagnosis section of the TA-F rubric, assessment of HTN required more review rounds on average than assessments of RVO and RAO.

Finally, each of the four panels conducting tool-based adjudication completed all 499

images within 58 days from initial grading to full adjudication, with the fastest panel completing in 19 days. Note that these durations also include intervals of idle time in which the system waited for graders to complete their review passes, and that graders performed other labeling tasks during their own idle intervals. The total amount of time spent on grading and reviewing activities is therefore substantially lower than the corresponding end-to-end durations per panel.

3.2.4 Discussion

As machine-learning methods become more common in ophthalmology, the need to accurately assess diagnostic performance grows. Algorithms that may be used to automate or augment aspects of eye care should be subjected to rigorous evaluation of their performance, against trusted reference standards. This in turn motivates the development of high-quality reference standards, a process that has received relatively little attention in the literature.

Previous studies suggest that adjudication can not only reliably be used to evaluate DR severity, but should be the reference standard used in deep-learning algorithms [82]. While several methods may be used for this process, such as in-person adjudication among expert panels and arbitration of disagreements by a senior grader (Domalpally A, et al. IOVS 2018;59:ARVO E-Abstract 4676) these methods rely on the time and expertise of certain physicians. In the present study, we present a tool-based system for remote expert adjudication of image-based interpretations and evaluate the system in the context of DR severity assessment. In this tool-based method, anonymity allows for an unbiased review of the image, with further clarity added by the rubric feature. Furthermore, the flexibility inherent in the design increases its appeal and ease of use.

Performance of Tool-Based Adjudication

Our study suggests that remote, tool-based adjudication procedures can produce DR grades that are in high agreement with the reference standard of in-person adjudication while offering a range of benefits: an increase in flexibility to accommodate graders' schedules, the possibility to anonymize graders throughout the adjudication process to avoid potential biases grounded in grader identity or seniority, and the option to explicitly structure the adjudication process around detailed evaluation criteria.

Research into efficient and reliable procedures to produce high-quality grading decisions can be applied to manual screening in teleophthalmology settings and to the validation of

automated methods, such as deep-learning systems. In both cases, reliable classification decisions are required to avoid potentially devastating consequences, such as missing cases of advanced disease. Validating the classification performance against a reliable reference standard may be of particular importance for automated methods as, once deployed into a clinical screening setting, these methods can affect large patient populations in a short amount of time.

Beyond the evaluation of our remote TA procedure for adjudicating fundus images for DR severity assessment, we demonstrate how our proposed tool-based procedure can provide structure to the adjudication process itself using explicit prompts for detailed evaluation criteria. The resulting TA-F procedure leads to a significant reduction in the number of rounds needed to resolve disagreements. One possible explanation for the observed efficiency improvement may be that the feature rubric helped graders communicate their rationale and the specific source of disagreement more efficiently than was otherwise achieved through free-form comments (e.g., by focusing communication on the specific guideline criteria that graders are instructed to factor into a diagnosis). Besides efficiency in communication, the guideline-centric rubrics may serve as a lightweight checklist, leading graders to be more consistent in their individual practices. This may reduce variance or allow graders to externalize the diagnostic criteria in a way that reduces their task-related mental workload. As supporting evidence (Figure 3.9), the first-round agreement rates were significantly higher with the use of the rubric, even before further adjudication. Finally, the rubrics lead to the production of structured information (i.e., the specific evaluation criteria applied in each case), facilitating detailed quantitative analyses to examine how and why disagreements arise both across a set of images and for individual cases.

Still, there were specific features of the disease that required more discussion. While the explicit reasons for why heme, VB, and IRMA required the greatest number of review rounds are not clear, it is possible that the overlap between the objective (i.e., simple presence or absence) and subjective (i.e., extent and prominence) features of these particular anatomic abnormalities led to more disagreement. Despite standard reference photographs to help guide whether or not the heme is extensive, VB is definite, or the IRMA is prominent, there is an inherent subjectivity to the process. Ultimately, the physician's gestalt leads her to define disease severity. This same overall impression or pattern recognition may explain why venous and arterial occlusions resolved in fewer rounds, as these diseases have a hallmark appearance. HTN, on the other hand, can overlap with and mimic several other eye diseases, DR being the most common, and giving a definitive diagnosis based on a fundus photograph alone can be challenging. Exploring these feature-based discrepancies may provide more insight on how the model synthesizes the information within the image and also on how to continue to improve it.

Utility in Clinical Practice

We believe the technology we describe here may have several clinical applications. First, our approach for remote adjudication is well suited for integration into existing telemedical workflows, which face the same problem of high intergrader variability as is the case for on-site clinical grading [134]. Here, our proposed system can help resolve ambiguous cases through group decision-making [15] on demand to improve clinical outcomes on a patient-by-patient basis. Apart from adjudication, our tool's functionality of integrating feature-level rubrics into the image interpretation process may facilitate grading by individual graders in difficult cases, by helping list and systematize the image findings.

Second, expanding TA and TA-F use for rare conditions or difficult to diagnose cases, where a patient may otherwise be advised to travel to seek a second or third opinion, could potentially have an important impact on time to diagnosis and treatment, which are likely to impact quality of life and healthcare costs.

Third, and perhaps most importantly, our adjudication tool lends itself naturally for generating highly reliable and trusted reference standards for the validation of automated methods, such as deep-learning models. The process of building and evaluating deep-learning models typically involves at least the three following distinct datasets: a 'development' dataset used to train the model, a 'tuning' dataset used to select high-performing model candidates during the training phase, and a 'validation' dataset used to benchmark the performance of the final model. While development datasets, in many cases, consist of tens or hundreds of thousands of training examples, the datasets used for tuning and validation are typically smaller scale, on the order of several hundred up to a few thousand cases. The methods presented here can facilitate the creation of tuning and validation sets with a substantially reduced overhead, due to lower time and coordination requirements as compared to in-person adjudication. In this study, we demonstrated the feasibility of remote TA for a set of 499 images, positioning it as a useful procedure especially for generating tuning and validation datasets. The availability of a highly trusted validation dataset is of critical importance especially for so-called "black box" systems, where there is limited ability to understand how the model makes its diagnosis. As methods for remote adjudication in clinical decision-making scale, it may become feasible to produce adjudicated datasets large enough to be used for training, which would extend the current state-of-the-art in model development.

Limitations

Our study is not without limitations. First, while we quantify the reliability of each adjudication method using consensus grades from two independent expert panels, the metrics reported in this section remain relative ones given the lack of an absolute, objective gold standard for DR severity assessment in the context of our study. To alleviate this issue, further work may benchmark adjudication decisions from digital fundus images against more rigorous diagnostic procedures (e.g., dilated fundus exam by a retina specialist) [7] or objective outcomes, such as any future development of blindness. Second, graders participating in this study were practicing retina specialists rather than research-grade reading center graders. While reading center gradings may be a more standardized gold standard, the incorporation of insights from clinical practice into the grading may render our results more applicable to real-life scenarios than may otherwise be the case with research-grade readings. Third, we only adjudicated DR severity, but did not adjudicate DME. Agreement levels may be lower overall given difficulties in diagnosing DME on 2D fundus photos. Finally, the grading decisions in this study were based on fundus images without accompanying patient information or clinical records. In practice, DR severity assessment based on digital fundus photography should consider patient history and be complemented by more rigorous diagnostic procedures including dilated fundus examination by a trained eye care professional and optical coherence tomography (OCT) or other imaging techniques when indicated to confirm the diagnosis [7].

Our TA-F procedure included a mechanism to assign pseudonyms to graders to avoid biases grounded in grader identity. Anonymization of graders was not possible during the in-person discussions of the Baseline adjudication procedure, and could not be done for the TA procedure because the functionality for grader anonymization was added at a later stage of our tool's development. Anonymization of members in group-based decision processes generally reduces incentives for groupthink behavior, and thus tends to slow down consensus formation rather than accelerating it [131, 109]. Thus, we reason that our reported benefit of TA-F is an underestimate of the true benefit, relative to comparing TA and TA-F when neither (or both) is anonymized.

Our results show that remote, tool-based adjudication can help organize the consensus formation process especially for those cases that can be resolved in the first few review rounds, but falls short of fully alleviating the problem of small portions of disagreement cases persisting over several review rounds. Future work may explore methods to accelerate resolution for such hard cases, for example, by investigating if aggregation methods like majority vote after the first two review rounds are sufficient proxies for final adjudicated decisions, or by implementing automatic techniques to schedule video conference calls to

discuss small collections of hard cases among panelists without the need to involve a human coordinator.

Other promising avenues for future research revolve around the development of feature rubrics for improved efficiency and reliability of adjudication procedures. Understanding which strategies and practices for rubric development generally result in the biggest improvements across various diagnostic tasks would be helpful for the community so that other researchers can reliably produce effective rubrics for different areas of medical image interpretation.

3.2.5 Conclusion

Remote, tool-based adjudication presents a reliable alternative to in-person adjudication for DR severity assessment. The system allows flexibility so that graders can schedule their reviewing around their clinical duties. Additional benefits include the option of blinding graders from the identity of other panel members and the ability to structure the discussion of controversial cases around a set of discrete evaluation criteria. We found that feature-based rubrics for DR can help accelerate consensus formation for tool-based adjudication without compromising label quality.

3.3 Expert Deliberation for Time Series Labeling

We conclude this chapter with a final case study on group deliberation in data labeling, again within the domain of medical decision making. In this section, we build on our insights from section 3.2 to design and implement CrowdEEG, an online platform enabling groups of medical experts to collaboratively label and adjudicate complex medical time series data.

Prior work shows that expert disagreement can arise due to diverse factors including expert background, the quality and presentation of data, and guideline clarity. In this section, we study how these factors predict initial discrepancies in the context of medical time series analysis, examining why certain disagreements persist after adjudication, and how adjudication impacts clinical decisions. Results from a case study with 36 experts and 4,543 adjudicated cases in a sleep stage classification task show that these factors contribute to both initial disagreement and resolvability, each in their own unique way. We provide evidence suggesting that structured adjudication can lead to significant revisions in treatment-relevant clinical parameters. Our work demonstrates how structured adjudication can support consensus and facilitate a deep understanding of expert disagreement in medical data analysis.

3.3.1 Motivation

Receiving a reliable diagnosis is one of the fundamental steps in health care delivery; it sheds light on the state of a patient's health condition and informs subsequent treatment decisions. The diagnostic process often requires visual analysis of medical data (e.g., x-rays, ultrasounds, electrophysiological signals) and a subsequent classification thereof (e.g., normal vs. abnormal). Expert classification tasks relying on visual analysis, however, tend to give rise to expert disagreement due to their inherently subjective nature—and the medical domain presents no exception in this regard.

In certain non-expert domains (e.g., crowdsourcing), techniques like majority vote and other computational methods (e.g., EM algorithm) are used to aggregate divergent human assessments into what is assumed to be a "correct" answer. By contrast, other approaches acknowledge that disagreement carries valuable information [44, 129], and that resolving disagreements is not always possible or desirable, even if human graders are given the opportunity to deliberate on a case [125]. In the clinical domain, collaborative, team-based decision making has long been deemed superior to individual diagnosis by the National Academy of Medicine [12]. Little is understood, however, about the factors that contribute

to expert disagreement and processes that facilitate resolution of disagreement in medical data analysis from a socio-technical perspective.

Our work addresses this research gap by studying the sources and dynamics of expert disagreement in medical data analysis through structured, collaborative adjudication, i.e., the process of reviewing and potentially resolving divergent assessments collectively as a group. Our findings from an observational case study with 36 experts and 4,543 adjudicated cases in a sleep stage classification task reveal that diverse factors, including expert background, the quality and presentation of data, and classification guidelines, contribute to both initial disagreement and resolvability. Our findings also demonstrate how adjudication can lead to significant revisions in experts' quantification of diagnostic markers, which in turn have the potential to impact patients' lives through changes in treatment outcomes. Our main contributions are:

1. We demonstrate how the sources and dynamics of *expert disagreement* in medical data analysis can be understood through collective *adjudication*.
2. We conducted an *observational study* to analyze expert disagreement, illuminating diverse factors impacting the extent of disagreement, including expert background, the quality and presentation of data, and guideline clarity.
3. We contribute a *structured* adjudication workflow to capture expert rationales in a guideline-centric and interoperable format.

In what follows, we detail the design evolution of our structured adjudication workflow, outline our research questions, methods, and findings, and conclude with a discussion of use cases and design considerations for our approach.

3.3.2 Application Domain

We embed our work in the field of biomedical time-series classification, an expert domain with typically low inter-rater agreement rate, and deploy our adjudication system in the context of sleep stage classification, where the average rate of agreement among two independent experts is approximately 82% with Cohen's kappa of 0.76 [36]. Sleep stage classification lends itself as a task for our case study, as it not only involves lengthy and complex guidelines likely to spur inter-rater disagreement, but sleep data includes a wide range of signal modalities, many of which are integral parts of other diagnostic procedures in medicine. The task of sleep stage classification involves mapping fixed-length segments

of a polysomnogram, i.e., a continuous multimodal medical time series recording, to one of five sleep stages — Wake, Rapid Eye Movement (REM) sleep or one of three non-REM sleep stages (NREM 1, NREM 2, NREM 3). The resulting sequence of sleep stages, called a hypnogram, serves as a relevant artifact in the diagnostic process for various sleep-related disorders and other neurological diseases. The classification of time series segments into sleep stages is based on the presence of distinguishing features of the EEG waveform and other supportive signal modalities like respiratory information.

3.3.3 Structured Adjudication

For the purpose of our study, we designed and implemented a workflow and interface for collective expert adjudication of classification decisions in the context of medical data analysis. Here, we describe our iterative design process and the resulting design considerations that informed our final design and implementation.

Design Evolution

Our design process was structured into three steps: (1) formative sessions of *in-person* adjudication to acquire a better understanding of inter-personal dynamics and expert argumentation patterns used in medical adjudication, (2) adjudication via *video conference* as a testbed for remote adjudication, and (3) *web-based* adjudication informed by insights from the first two steps. In all three steps, the CrowdEEG signal viewer [2] was used for independent classification, but it was only in the final stage where adjudication of disagreements was conducted directly within the web interface.

In-Person Adjudication. An initial formative session of in-person adjudication was conducted with three board-certified sleep technologists. After an initial round of independent classification, researchers organized an in-person meeting in the hospital to host adjudication discussions for select disagreement cases. All members of the expert panel convened at a set time and place to collectively discuss disagreements in front of a shared screen. 106 minutes of discussion content was recorded (using screen capture and audio recording), transcribed, and analyzed. Our findings led to several design considerations both general and specific to our data modality:

- Discussions were primarily centered around the classification guidelines, including the presence of individual patterns or features in the data. This observation primarily

[2] http://crowdeeg.ca

informed our motivation for integrating classification guidelines into the final web-based approach.

- Inter-personal dynamics occasionally distracted from the case at hand (e.g., jokes about the grading style or background of other panel members), or caused bias in favour of certain experts (e.g., the most dominant ones or the ones with highest perceived expertise). Based on this finding, we decided to hide information about expert identity and expert background in our web-based implementation.

- For some disagreement cases, experts requested to review data windows before or after the case in question (specific to sequential data). In addition, resolving certain disagreements triggered consensus on short cascades of subsequent cases in the recording timeline. Based on these two insights, we decided to have experts review all cases in a given recording for our web-based procedure, one expert at a time, to account for any sequential dependencies.

- The configuration of the viewer (e.g., signal visibility and amplitude scaling) played a role in discussing and resolving disagreements. We noticed that, for certain cases, adjusting the viewer settings triggered consensus without further argumentation. Inspired by this observation, we decided to allow experts to configure various aspects of the viewing interface, and to record viewer settings for each classification decision to facilitate quantitative analysis.

Remote Adjudication via Video Conference. In a second step, we conducted a 1-hour experimental session for remote adjudication with the same three experts, this time using video conference as the communication medium. The cases discussed in this step were distinct from the cases previously discussed in person. All three panel members and one moderator (whose role was to ensure adjudication discussions stayed on topic) joined the video conference at the same time. Each expert was assigned one colour (red, green, or blue) that could be used to annotate the location and shape of characteristic features on a shared screen during discussion. Discussions were recorded (via screen capture and audio recording) and analyzed, resulting in additional findings:

- Despite the fact that experts were not co-located in the same room, inter-personal dynamics seemed to influence the discussion based on perceived grader experience and the effectiveness of individual communication or argumentation skills. While part of this behavior may have been influenced by the fact that the same three experts had previously conducted in-person adjudication on separate cases, this observation

reinforced our design consideration to anonymize experts during adjudication and informed our choice of text as a communication medium during web-based adjudication.

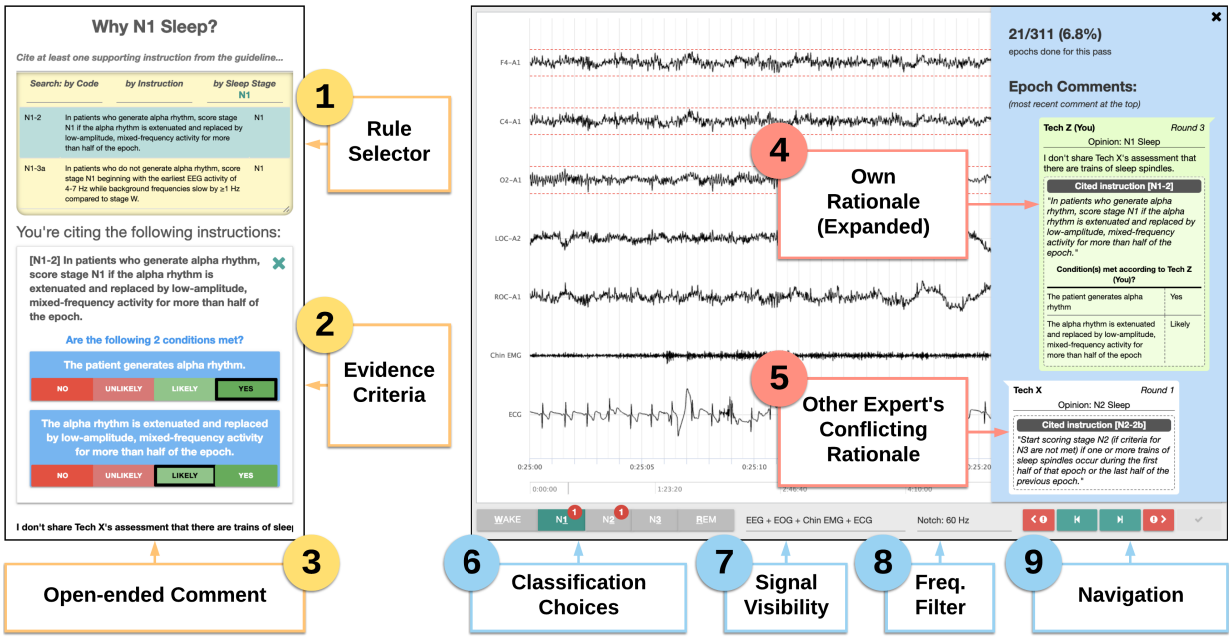
- The logistics of scheduling multiple domain experts to collectively join a meeting at a set time even without the need for a co-located face-to-face setup proved to be prohibitive for a large-scale study. This realization motivated our decision to implement an asynchronous approach for our web-based adjudication workflow in which experts review disagreement cases in a round-robin fashion, one expert at a time.

- The interplay of distinct features within the same classification case, as well as disagreements over the exact transition boundaries between different feature types were topics of contention and became evident through on-screen drawing. Inspired by this observation, we included measures of signal complexity in our data analysis, both with regard to the frequency domain (i.e., how complex is the signal overall?) and from a time-frequency view (i.e., how complex is the signal due to transitions over time?).

Web-based Adjudication. Our design considerations derived from the first two steps informed an early prototype of our web-based adjudication workflow and interface. The primary motivations for moving the adjudication process to a web-based implementation were (1) the ability to orchestrate adjudication at a larger scale involving multiple concurrent expert panels (2), mitigation of certain undesirable factors observed during in-person adjudication and adjudication via video conference, and (3) the introduction of explicit structure to the process of collecting expert rationales for post-hoc quantitative analysis.

Our first iteration of the web-based adjudication workflow addressed the former two motivations by reducing scheduling conflicts among experts through a round-based scheduling approach, by hiding information about grader identity and background, and by using text as a communication medium. We conducted a small-scale pilot using our initial prototype with three independent panels, each with three experts. The objective of the pilot was to validate the overall interface and workflow and to analyze open-ended discussion contents before attempting a more structured approach of collecting expert rationales.

Open-ended discussion comments collected during the pilot were generally free of interpersonal comments, concise (ranging from a few words to one or two sentences), and focused primarily on specific rules from the classification guidelines including low-level features referenced therein. While the majority of comments matched this description, we noticed



(a) Rationale form. (b) Data view for disagreement case with expert discussion.

Figure 3.11: Interface for structured adjudication of classification decisions in medical time series analysis.

that few comments contained arguments not captured by the classification guidelines (e.g., addressing implicit nuances with regard to ambiguous terminology used for individual rules in the guideline or referring to the assumed health condition of the patient). Based on these findings, we decided to proceed with integrating classification guidelines into the workflow in an extensible and structured manner. We also decided to retain the option of providing open-ended comments throughout the process to cover the few cases in which guidelines were insufficient for a comprehensive rationale. The remainder of this section outlines our final design and implementation of web-based, structured adjudication.

Rule-based Representation of Guidelines

In our approach, classification guidelines are represented as a set of inference rules matching a basic template:

IF Evidence A Present **AND** Evidence B Present **THEN** Classify as X

Each rule defines a Boolean proposition or a conjunction (AND connection) of multiple propositions (e.g., rapid eye movements are present AND low-chin EMG tone is present AND low-amplitude, mixed-frequency EEG is present) that need to be true in order to make a certain annotation decision (e.g., classify as REM sleep). We will later refer to the propositions on the left side as *evidence criteria*. More complex rules can be decomposed to match this template. For example, disjunctions (OR connections) can be split into multiple rules relying just on conjunctions.

While our case study demonstrates the utility of our approach using a domain-specific guideline, the overall approach is domain-agnostic, borrowing basic concepts of propositional logic. For our case study, we adapted a domain-specific classification guideline [70], translating it into a set of 36 separate inference rules. These rules referenced a set of 15 unique basic features whose presence were relevant for at least one of the rules. We also included *placeholder* rules, one for each of the possible classification choices, that could be selected in case none of the other rules applied.

Workflow

Our final workflow consists of two stages. First, all panel members independently perform an initial classification pass on the entire data record. Second, each expert reviews and re-classifies all disagreement cases in the record in a round-based fashion, one expert at a time. We describe both workflow stages below.

Classification. The entry point for participants is an automated email notification with a link to sign into our web-based system to proceed with their classification task. During initial classification, each grader classifies all cases in the data record (i.e., 30-second windows within a sleep EEG) into task-specific categories (i.e., one of five sleep stages). To account for sequential dependencies in the classification process [128] as observed in the pilot study, graders can navigate back and forth through all cases and are free to adjust previous classification decisions throughout their pass. Our pilot studies revealed that the specific way in which graders choose to view the data—i.e., which signals they choose to be visible, how they scale individual signal amplitudes, and which frequency filters they apply—can affect individual grading decisions. Our workflow therefore allows for graders to adjust viewer settings throughout their pass and to revise grading decisions accordingly. As graders are free to update prior classifications throughout their pass, our workflow requires that graders explicitly mark a pass as complete, so grading decisions can be locked in for comparison with other panel members.

Adjudication. Our pilot study made clear the fact that scheduling multiple domain

experts to synchronously adjudicate a disagreement at a set time is logistically prohibitive for a large-scale study with geographically remote participants. This insight informed our design consideration to choose an asynchronous, round-based approach for our adjudication workflow. In each round, the active grader reviews all disagreement cases among the data record, remotely and on their own time. Our system notifies individual panel members via email when their adjudication pass becomes available. Upon login, graders are immediately positioned on the first disagreement case in the data record, and can jump to the next or previous disagreement case as they proceed. During each adjudication pass, the active grader reviews each disagreement case at least once before the pass can be marked as complete. The approach to navigation and re-classification is similar to the workflow previously described for independent classification. In addition, graders are required to provide an explicit, structured rationale for each re-classification decision, and must choose at least one domain-specific guideline in support of their classification. For each guideline rule cited, graders are asked to indicate the extent to which they believe the given rule-specific evidence criteria to be met. Finally, graders are given the option to leave an open-ended comment about each decision to account for cases where guidelines are insufficient to explain a comprehensive rationale. All rationales collected from previous adjudication rounds are presented to graders automatically when they navigate to a given disagreement case, encouraging adjudicators to review any prior case-specific discussion within the panel.

A disagreement case is considered resolved when all graders in a panel converge on the same classification. The adjudication process ends when all disagreement cases are resolved, or when a specified number of adjudication rounds have been completed. In our case study, we limit adjudication to three rounds, i.e., one review pass per panel member.

User Interface

We designed a user interface (UI) to implement our workflows for classification and adjudication within a web-based platform (Figure 3.11). Components of the adjudication UI were integrated into the classification UI to ensure contextual vicinity between case-specific expert discussions and signal data. We briefly describe the classification UI below, followed by a more detailed outline of the adjudication UI.

Classification. The primary purpose of the classification UI is to enable experts to view and classify complex data efficiently without violating any existing domain-specific conventions. It is therefore designed to emulate existing viewer software for the domain-specific task at hand (i.e., sleep staging). The largest portion of the screen is devoted to data presentation, with controls streamlined for *efficient* user input. In addition to

on-screen controls, there is hotkey functionality for navigation, classification, and select viewer settings (cf. Figure 3.11b, components 6 to 9).

Adjudication. The adjudication UI (Figure 3.11) is designed for explicit and justified collaborative decision making. Its components are general and can be instantiated in the context of other data classification tasks (e.g., for text documents or images). Our pilot study suggested that a critically important step for experts in understanding disagreements is a compact view of any conflicting classification choices within the group. Therefore, the adjudication UI visualizes group decisions by displaying the number of votes assigned to each classification category using circular indicators attached to classification buttons. Disagreement cases are visually contrasted from agreement cases to guide graders' attention using multiple red-colored vote indicators (Figure 3.11b, component 6).

As disagreement cases can be scattered across a single contiguous data record, our adjudication UI extends the base navigation panel with two additional buttons (and hotkeys) to jump directly to the subsequent and previous disagreement case (Figure 3.11b, component 9).

To facilitate structured communication between members of an expert panel, the adjudication UI includes a discussion component to render case-specific expert rationales. Early prototype testing suggested that some graders re-classified disagreement cases without reviewing prior discussions on the case. We therefore chose to automatically open the discussion component as soon as graders navigate to a disagreement case to encourage active review of prior arguments. Expert rationales and open-ended comments (if any) from all group members are displayed in chronological order (Figure 3.11b). Each guideline rule cited within can be expanded using mouse-over to reveal information about pertinent evidence criteria. As our pilot study showed that inter-personal dynamics can distract from deliberation, we chose to use expert *pseudonyms* allowing group members to distinguish their own rationale (Figure 3.11b, component 4) from those of other experts in the group (Figure 3.11b, component 5) while hiding any information about expert identity or background.

For the purpose of providing justifications for re-classification decisions, the adjudication UI includes a rationale form (Figure 3.11a). The rationale form is triggered when a grader submits a classification choice, and classification choices are saved only after the form has been completed and submitted. The form consists of three parts: a rule selector (Figure 3.11a, component 1) enabling experts to search a catalogue of pre-defined guideline rules and to cite those that best represent their rationale; a component asking experts to specify the extent to which they believe each of the evidence criteria for the selected rule(s) are met (Figure 3.11a, component 2); and the option to provide an additional open-ended

comment (Figure 3.11a, component 3). Graders are required to select at least one guideline rule in support of their classification choice, but can choose to cite additional rules if applicable even if those happen to contradict their classification. The design consideration here was to allow graders to discuss potential nuances or conflicts between multiple guidelines rules by citing several ones and clarifying their reasoning using open-ended comments. The rationale form is domain-agnostic and can be instantiated for a specific application domain by providing a rule-based representation of the pertinent classification guidelines, in the format described above.

3.3.4 Research Questions and Hypotheses

Our study addresses three research questions.

Q1: Why do experts disagree during independent classification?

Diverse factors including training background and preferences in data presentation may cause experts to arrive at divergent classification decisions, beyond characteristics inherent in the data itself. For example, experts with varying credentials or varying levels of work experience may be more likely to disagree. Likewise, our formative design process suggested that experts may disagree solely based on the use of different viewer settings. Based on these intuitions, we hypothesize that:

[H1a] Differences in *expert background* (i.e., credentials, geographic location, and work experience) are associated with higher disagreement.

[H1b] Differences in *viewer settings* (i.e., signal visibility, amplitude scaling, and frequency filters) are associated with higher disagreement.

[H1c] Certain *data characteristics* (i.e., abnormalities in a patient's health condition, and case-specific signal complexity) are associated with higher disagreement.

Q2: Why do certain disagreements persist after collective adjudication?

The same factors that contribute to independent disagreements may similarly contribute to the dynamics of adjudication among panels of experts, and may help explain why certain disagreements get resolved through exchange of arguments while others persist. Beyond this intuition, we take the stance that knowing about the specific criteria over which experts disagree will best explain why certain cases get resolved and others do not. We hypothesize that:

[H2a] Differences in *expert background* affect the likelihood of resolving a case.

[H2b] Differences in *viewer settings* affect the likelihood of resolving a case.

[H2c] *Data characteristics* affect the likelihood of resolving a case.

[H2d] The specific *structure of a disagreement* (i.e., discrepancies over the presence of individual features in the data) carries greater explanatory power for understanding why certain disagreements persist after adjudication, compared to the other factors (i.e., differences in expert background or viewer settings, and data characteristics).

Q3: What impact does adjudication have on clinical decision making?

Collaborative decision making has been championed by national health research institutions [12], which assume that team-based approaches lead to significant improvements in clinical decision making. Adopting the paradigm of collective intelligence in healthcare, we hypothesize that:

[H3a] Experts perceive collective adjudication as useful for arriving at reliable and trustworthy classification decisions.

[H3b] Adjudication can lead to significant revisions in treatment-relevant diagnostic markers.

3.3.5 Methods

Here we describe the details of our observational case study including participant recruitment, data set, procedure and statistical analysis.

Participant Recruitment

We recruited 36 expert participants via domain-specific online platforms. Based on the pre-study questionnaire (Appendix A.2.1), our expert participants were located in the United States (26), Canada (7), the European Union (2), and other unspecified geographic locations (1). The majority of participants (30) were Registered Polysomnographic Technologists (RPSGT); six held lower credentials. More than half of our expert participants (23) reported having at least five years of experience working as sleep technologists. Out of our 36 participants, 31 self-reported as female and five as male. The distribution over

age groups was: 18-25 (1), 26-35 (8), 36-45 (16), 46-55 (8), 56+ (3). Participants were paid US $112.50 for two scoring passes (independent classification and one review pass) via online gift cards, or the equivalent amount in the currency of their specified location, corresponding to an hourly rate of US $37.50 with three hours of estimated total work on average.

Data

For the purpose of our study, we sampled just over 86 hours of sleep recording data from twelve different patients with a mean recording duration of 7.19 hours (SD = 43 mins), reflecting the standard length of a night at a sleep laboratory. Our dataset included patients with four different health conditions (three healthy patients, three with Parkinson's disease, three with Alzheimer's disease and three with sleep apnea). The distribution over patient age groups was: 40-44 (2), 60-64 (1), 65-69 (2), 70-74 (3), 75-79 (4). Six patients were female and six were male. The complete dataset included 10,349 individual classification cases each corresponding to one 30-second window of biosignal data to be classified into one of five different sleep stages. Almost half of all cases (4,543; 44%) resulted in some level of expert disagreement over the correct classification label. Note that agreement rates here refer to exact agreement among three experts whereas rates reported in prior work refer to agreement among just two experts and are therefore expected to be higher. Out of all disagreement cases, about one third (1,667; 37%) remained unresolved after three rounds of collective adjudication.

Procedure

Before the study, experts first completed a pre-study questionnaire (Appendix A.2.1) soliciting demographic information, including age group, gender, geographic location, their highest credential, as well as the number of years of work experience in the sleep health profession. The 36 expert participants were randomly grouped into groups of three and each group was assigned to one of the twelve recordings for collective adjudication. Hence, each expert grader participated in exactly one panel and each recording was scored and adjudicated by the same set of three experts. Experts first performed an initial independent classification pass on their assigned recordings, followed by three rounds of adjudication, one round per grader in the panel. The order in which experts performed the review passes was scheduled based on expert availability in each panel, i.e., for each panel, the three experts were sequenced based on their earliest possible availability for completing a full review pass. Alternative sequencing options such as randomization may be desirable

based on the specific study setup, e.g., if experts are part of multiple distinct adjudication panels. In our case study, where experts are part of exactly one panel and perform one review pass each, individual availability was taken into account as a social requirement to reduce delays between review passes. In each adjudication round, the active grader stepped through each individual disagreement case, re-scored the case, and provided a rationale for their final classification decision. In each adjudication round and for each disagreement case, the active grader was presented with the most recent classifications from all three panel members, as well as the grades and rationales submitted during each of the preceding rounds. Note that our observational case study treats independent classification and adjudication as consecutive workflow stages, rather than distinct experimental conditions. The study concluded with a post-study questionnaire (Appendix A.2.2) allowing participants to provide open-ended feedback about the benefits and drawbacks of the adjudication interface and procedure. We also included two questions to assess the degree to which experts agreed that *'The adjudication process was useful for generating a reliable hypnogram'*, and the degree to which experts agreed that *'The final adjudicated hypnogram can be trusted more than the hypnogram from my first pass'*, both on 5-point Likert scales.

Analysis

For Q1 and Q2, we analyzed how various socio-technical factors like expert background, data characteristics, and viewer settings, were associated with expert disagreement during independent classification (Q1), and with the likelihood of leaving a disagreement unresolved after collective adjudication (Q2). We investigated both research questions using logistic regression models. For Q1, the logistic model was run on all classification cases (N=10,349), the dependent outcome variable indicating whether a case had any level of disagreement (N=4,543) versus perfect agreement among all three experts. For Q2, we ran a sub-analysis on just those cases with any initial disagreement (N=4,543) to understand why some disagreements persisted after three rounds of collective adjudication (N=1,667), whereas other disagreements managed to get resolved. Both analyses shared a base set of independent variables, described in Table 3.11.

For Q2 specifically, we derived additional independent variables from the structured rationales experts submitted during adjudication. The complete set of all 36 guideline rules mentioned 15 unique basic features. We derived one independent variable for each one of these features, which assumed a true value if some, but not all panel members had mentioned the feature in their rationale for a given disagreement case, and false if either all or none had mentioned it. This approach allowed us to condense expert rationales from a complex set of guideline rules into a compact view of basic features to gauge the explanatory

Table 3.11: Factors used as independent variables in Q1 and Q2.

Category	Variable	Description
Grader Differences	**Experience**	true if panel members had different levels of work experience, i.e., if some had 5+ years of work experience, while others did not; false if all panel members had the same level of work experience
	Location	true if panel members were from different geographic locations; false if all were from the same location
	Credentials	true if some, but not all panel members held an RPSGT credential; false if either all or none were RPSGTs
Viewer Differences	**Frequency Filter**	true if some, but not all panel members had activated the frequency filter while making a classification decision; false if either all or none had activated the frequency filter
	Amplitude Scaling	true if some, but not all panel members adjusted the sensitivity of the signals for a given case; false if either all or none had made adjustments to amplitude scaling
	Signal Visibility	true if there were differences among panel members in how many signals were visible when making a classification decision; false if all looked at the same set of signals for a given case
Data Characteristics	**Patient Condition**	one of three disease conditions—Alzheimer's, Parkinson's, or sleep apnea—compared to the healthy baseline
	Signal Complexity	true if the EEG for a given classification case was more complex than the median case with regard to the *frequency domain*; complexity was measured as spectral entropy, which is high if the signal contains multiple dominant frequencies, and low if it only contains one main frequency [13]
	Signal Transitions	true if the EEG for a given classification case was more complex than the median case with regard to the *time-frequency domain*; measured as entropy over the dominant frequencies for each 2-second segment within a 30-second window

power of feature-level expert rationales for understanding why certain disagreements persist after adjudication.

For Q3, we used paired t-tests to compare the value of aggregate diagnostic markers before and after adjudication. A one-sample Wilcoxon signed rank test was used to understand if experts considered the adjudication process useful for making their classification decisions more reliable and trustworthy as per the two questions in the post-study questionnaire.

For qualitative data analysis, line-by-line inductive open coding was performed by one of the study authors to identify the emerging themes reported below.

3.3.6 Results

Structured adjudication resulted in a 20-30% increase in inter-rater agreement over the course of three rounds (cf. Figure 3.12). The machine-readable outputs of our system allowed for several insights to be had regarding the dynamics of our structured adjudication process. We observed vast differences in the role that different features (i.e., distinct evidence criteria mentioned in the classification guidelines) played for adjudication. Not only were certain features mentioned orders of magnitude more often than others (cf. Figure 3.13); different features also contributed to the resolvability of disagreements in diverse ways. Here we present the results of our data analysis with respect to each of our research questions.

Q1: Why do experts disagree?

In determining the causes of disagreement during independent classification, various factors were analyzed across different groups of variables, including differences in grader background and viewer settings, as well as characteristics inherent in the data itself (Table 3.12, left side):

- For **grader background**, differences in work experience, as well as geographic location, were significant determinants in predicting disagreement before adjudication. Differences in grader credentials were not found to be significant in predicting initial disagreement among a panel—results providing partial support for our hypothesis **H1a**.
- Differences in **viewer settings** used by graders during independent classification—frequency filters, amplitude scaling, and signal visibility—all were significant factors for initial agreement rates. However, while differences in frequency filter settings and amplitude scaling were associated with disagreement, differences in signal visibility (i.e., differences in whether graders were viewing all or only a subset of the available signals), was found to be a significant predictor of initial agreement among a panel. Our results partially confirm hypothesis **H1b**.
- With respect to **data characteristics**, and in line with our hypothesis **H1c**, the overall signal complexity for a given case contributed to initial disagreement. Disagreement was significantly higher for patients with Parkinson's and Alzheimer's disease, compared to a baseline of healthy patients. The same insight is also reflected in our observation that these two health conditions exhibited the lowest levels of inter-rater agreement before adjudication (Figure 3.12).

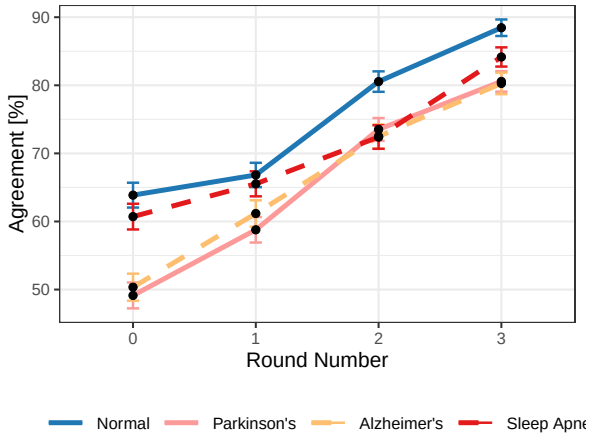


Figure 3.12: Agreement rate by adjudication round number and patient's health condition.

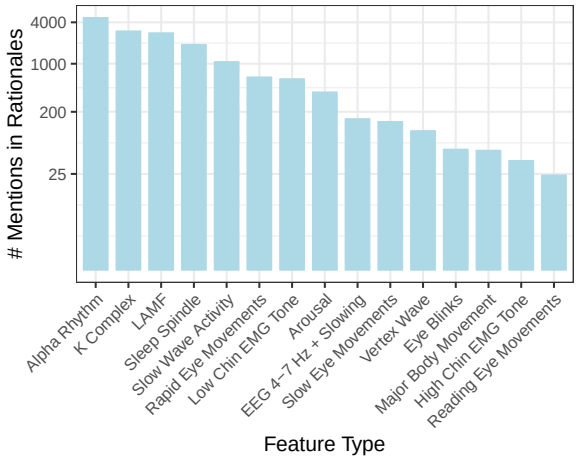


Figure 3.13: Number of times each feature type was mentioned in a rationale (log scale).

Q2: Why do disagreements persist?

In analyzing which factors were associated with persistent disagreement—i.e., cases with initial disagreement that remained unresolved after adjudication vs. those that were resolved through adjudication—similar patterns were observed across variable groups (Table 3.12, right side). Many of the same factors associated with initial disagreement were also significant explanatory variables for the outcome of persistent disagreement after adjudication, offering partial support for hypotheses **H2a**, **H2b**, and **H2c**. There were notable shifts, however, in the way that certain variables were associated with resolving a case compared to how they contributed to initial disagreement. We focus on those variables with differential effects between Q1 and Q2.

Variance in grader credentials, while not found to cause disagreements in Q1, was associated with an increased likelihood of resolving disagreement cases (Q2). Similarly, with respect to data characteristics, sleep apnea patients did not give rise to more disagreement than healthy patients did during independent classification, but disagreement cases could be resolved more readily for sleep apnea patients than for the healthy baseline. On the other hand, Alzheimer's disease did not significantly contribute to the persistence of disagreement, despite the fact that it contributed to initial disagreement. Where the EEG signal itself was concerned, overall signal complexity correlated with greater resolvability, whereas signal complexity in terms of transitions over time was associated with higher

Table 3.12: Logistic models for understanding why experts disagree during independent classification (Q1), and why certain disagreements persist after adjudication (Q2).

	Q1: Why Disagree?				Q2: Why Unresolved?			
Independent Variable	$\hat{\beta}$	*SE*	*t*	*p*	$\hat{\beta}$	*SE*	*t*	*p*
Grader Differences								
Experience	0.69	0.06	12.30	***	0.58	0.13	4.45	***
Location	0.36	0.07	5.06	***	0.50	0.20	2.57	*
Credentials	-0.06	0.07	-0.79		-0.93	0.16	-5.91	***
Viewer Differences								
Frequency Filter	0.51	0.07	7.72	***	0.83	0.14	5.76	***
Amplitude Scaling	0.19	0.05	3.91	***	0.04	0.11	0.32	
Signal Visibility	-0.25	0.07	-3.39	***	-0.44	0.17	-2.62	**
Data Characteristics								
Patient Condition								
Parkinson's	0.73	0.07	10.98	***	0.91	0.16	5.61	***
Alzheimer's	0.27	0.08	3.30	***	-0.18	0.19	-0.97	
Sleep Apnea	0.14	0.09	1.55		-0.59	0.23	-2.58	**
Signal								
Complexity	0.22	0.04	5.05	***	-0.57	0.11	-4.98	***
Transitions	-0.07	0.04	-1.64		0.54	0.10	5.19	***
Feature Disagreements (Q2)								
Slow Wave Activity					5.12	0.21	24.31	***
LAMF					2.16	0.11	19.38	***
Arousal					1.88	0.19	9.65	***
Alpha Rhythm					1.67	0.12	13.64	***
Eye Blinks					1.51	0.39	3.85	***
K Complex					1.33	0.12	10.85	***
Sleep Spindle					1.25	0.13	9.77	***
Reading Eye Movements					0.83	0.54	1.55	
Low Chin EMG Tone					0.71	1.02	0.69	
Vertex Wave					0.61	0.35	1.76	
High Chin EMG Tone					0.57	1.07	0.53	
Rapid Eye Movements					0.48	1.02	0.48	
EEG 4-7 Hz + Slowing					0.05	0.25	0.20	
Slow Eye Movements					-0.82	0.36	-2.32	*
Major Body Movement					-2.88	0.59	-4.88	***

chances of leaving a case unresolved. Differences in amplitude scaling, amenable to viewer settings, were not significant for resolving disagreements, despite causing disagreement during independent classification.

In addition to this base set of variables, Table 3.12 provides a list of 15 EEG features mentioned in at least one of the structured expert rationales from our study. For seven of these, we found that disagreements over feature presence were significantly associated with leaving cases unresolved. We found the opposite to be true for two other features—slow eye movement and major body movement—where discrepancies over feature presence were correlated with consensus formation. Most importantly, however, across all variable groups, it were the feature-level variables that showed the greatest effect sizes for case resolvability overall. This finding confirms our hypothesis **H2d**, the claim that the structure of a disagreement, with respect to feature-level rationales, holds the greatest explanatory power for why disagreements remain unresolved even after adjudication.

Q3: What impact does adjudication have on clinical decision making?

We assessed this question through both qualitative and quantitative measures. Through a post-study survey, expert graders responded that the structured adjudication process was both useful for generating a reliable hypnogram ($p < 0.001$), and that the final adjudicated hypnogram could be trusted more than the original one ($p < 0.001$). These findings support our hypothesis **H3a**. Changes in sleep parameters from before to after adjudication were analyzed as objective measurements for the impact of adjudication (Figure 3.14). We observed a significant decrease ($p < 0.05$) in the percentage of sleep time classified as REM sleep (%REM)—a treatment-relevant diagnostic marker—with shifts ranging between -7.5% and 1.4%. This finding offers support for our hypothesis **H3b**.

Qualitative Feedback. All participants were given the opportunity to provide open-ended feedback regarding both the interface of our adjudication system, as well as the adjudication procedure deployed in our study. Where the design of our platform was concerned, graders commended the clarity of its structured, guideline-centric format, with quick access to a comprehensive list of classification guidelines, and a view through which to appreciate other graders' classifications and rationales.

Grader feedback was unanimously in favour of the collaborative practice of adjudication, especially in a structured format that allowed for the exchange of individual justifications for one's classification decisions. As group discussion was achieved remotely, our graders felt that having a clear-cut time window for each individual grader to complete their pass on the record was beneficial for promoting efficiency of adjudication. However, the sequential

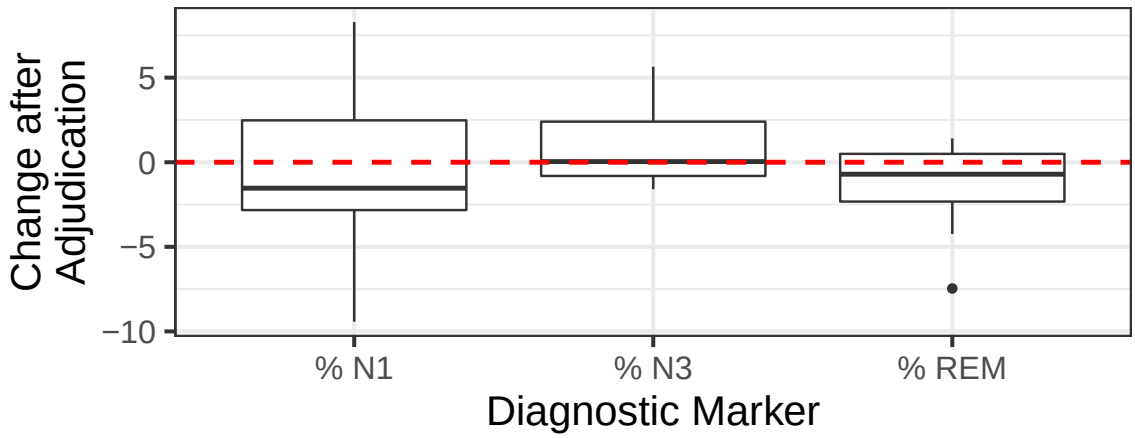


Figure 3.14: Change in diagnostic markers from before to after adjudication.

nature of our procedure, where the first grader must complete their pass before the second grader in the panel can begin theirs, was considered an obstacle by some participants. At the same time, our graders recognized that real-time adjudication may be challenging given the logistical burden of scheduling synchronous sessions among experts, even if facilitated through a remote, web-based system.

Our procedure, requiring graders to construct structured arguments in support of their decisions in terms of rules from the classification guideline, was said to have the potential to improve individual graders' scoring abilities. To borrow a quote from one expert, a guideline-centric adjudication procedure can help both those who have been scoring for years and are thus "stuck in their ways", as well those with minimal scoring experience, who may need a concrete guide. Where the effects of adjudication on consensus are concerned, most participants perceived the exchange of arguments and justifications among the group as a balanced approach for reviewing cases collectively: "Sometimes I still disagreed. Other times I changed my perspective." One expert reinforced our stance that "truly subjective cases" ought to be recognized as such, rather than enforcing artificial consensus.

3.3.7 Discussion

Our core contribution in this section is an observational study of expert disagreement in the domain of medical data analysis. With prior work establishing that group deliberation can be a useful and effective method for resolving disagreement, we built a structured, guideline-centric adjudication system and workflow to facilitate our study.

In addition to offering further support for adjudication as a method of supporting consensus formation, our findings help elucidate the reasons why experts disagree about medical data classification decisions, why some of those disagreements persist, and how adjudication outcomes may translate to clinical outcomes. We discuss the generalizability of our findings and their potential applications, offer design considerations for expert adjudication workflows, and conclude by addressing limitations of our study and directions for future work.

Generalizability and Applications

Our case study was limited to the specific task of interpreting and classifying biomedical time series, so caution is warranted in generalizing the results to outside domains and to task types other than data classification (e.g., policy design, content generation). However, sleep stage classification is a good exemplar for the medical domain, as it shares several characteristics with other diagnostic tasks: (1) expert disagreement is prevalent within; (2) data classification guidelines are lengthy and complex; (3) data analysis includes a wide range of signal modalities (i.e., EEG, ECG, eye movements, muscle activation, etc.), many of which are integral parts also of other diagnostic procedures in medicine; and (4) various low-level features of the data (e.g., alpha rhythm, sleep spindles, K-complexes) provide the basis for higher level assessments (i.e., sleep stage classifications, diagnosis of sleep disorders). We characterize the types of medical data analysis tasks to which our findings may generalize below.

Grader Differences. Findings on the effect of expert background on disagreement dynamics may generalize better to tasks where procedures for expert certification vary between countries or where such procedures may undergo significant changes over time. In those cases, differences in graders' geographic location or professional experience may play a more significant role in contributing to disagreement than for tasks where certification procedures are are globally standardized and remain relatively stable over time.

Viewer Settings. Findings on the effect of viewer settings on disagreement dynamics may generalize better to tasks where complex patient data can be viewed from different perspectives. Perspective adjustment can take various forms, including adjustment of the amount of data viewed (e.g., montage selection in multimodal time series, or region-of-interest adjustment in interpretation of pathology slides), or application of certain filter settings (e.g., frequency filters in time series or audio data, or color filters in image data). Such findings would not directly apply to classification tasks with static data views (e.g., text-based patient records).

Data Characteristics. We included disease condition and signal complexity as variables to understand the effect of data characteristics on expert disagreement. The specific operationalization of these variables may need to be adjusted for other task domains. The idea, however, that certain pathologies or pattern complexity may complicate data interpretation is domain-agnostic and may generalize to other task types.

Guidelines. The proposed guideline-centric adjudication process is general, and applicable to task types where pre-existing guidelines in the expert community can be mapped to a set of classification rules. Evidence-based grading guidelines are widely available across multiple medical subspecialties [11], and the organization of guidelines into easily identifiable grading recommendations is encouraged within the medical community [59]. There are, however, some diagnostic tasks, such as diagnosis of epilepsy or glaucoma, for which comprehensive guidelines yet have to be developed. The approach may generalize better to tasks with only a few classification categories (e.g., sleep staging, diabetic retinopathy grading, prostate cancer grading) than to classification tasks with very large decision spaces (e.g., comprehensive differential diagnosis) or multiple classifications with respect to the same patient record.

In our study, we hypothesized (**H2d**) that disagreements over the presence of specific features in the data would offer the strongest explanatory power for the resolvability of disagreements. Indeed, our results confirm that such feature-level rationales contribute the strongest to explaining why certain disagreements persist after adjudication. This finding suggests that expert disagreement, while influenced by social factors and specifics of data presentation, can be best explained by leveraging feature-level justifications from experts in medical data analysis. Since feature-level justifications were directly derived from guideline rules cited during collaborative adjudication, disagreements on the feature level can be considered a quantitative lens on low-level ambiguities within the guidelines.

Another finding from our study was that structured adjudication can lead to significant revisions in clinical parameters relevant to real-world treatment outcomes (**H3a**). In our sleep stage classification task, adjudication caused a significant decrease in %REM, compared to an independently annotated record. Clinical decision making in many scenarios hinges on the proportion of time spent in REM sleep recorded on an EEG. For instance, REM sleep is decreased in several neurodegenerative disorders, including Parkinson's disease and Alzheimer's disease, and among older adults without Alzheimer's, decreased REM sleep is associated with a higher likelihood of developing the disease in the future [105]. The detection of REM sleep behaviour disorder, which is associated with a high risk of future Parkinson's disease, is critically dependent on the accurate classification of REM sleep [110]. REM sleep is also decreased in sleep apnea, and the restoration of normal amounts of REM sleep can be a marker of therapeutic efficacy in sleep apnea treatment. These

insights position the adjudication process as something more than an academic exercise in consensus formation, but an approach with the potential of altering clinical outcomes as a direct result of changes in diagnostic markers.

Applications. There are several potential applications of our structured adjudication system and procedure. First, a system like our own can be easily implemented in the training of novice readers. In our study, differences in grader experience predicted both discrepancies before adjudication, and persistent disagreement afterwards. Beyond the obvious explanations for this, it is worth repeating that our expert participants highlighted that adjudication may be helpful both for novice graders, and more experienced graders. Our guideline-centric platform allowed for graders to go entirely by-the-book in their approach to classification, but the more seasoned scorers may well have stuck to their tried and true reasons for their classification decisions, perhaps overlooking certain nuances in the data that those following the rules would have better attended to, leading to disagreement cases. Thus, our system may have equal potential for helping more experienced readers reconsider their grading habits.

While structured adjudication was made possible in our study by the fact that standardized, agreed-upon classification guidelines are pre-existent within the expert community, our rationale form retained the option of providing open-ended comments. For domains where standardized classification guidelines do not yet exist (e.g., epilepsy diagnosis), our hybrid approach could offer the potential of mining open-ended arguments to extract explicit inference rules and thus iteratively generate a more structured representation of classification guidelines.

The interoperable output of our structured adjudication system may also lend itself naturally as input to other decision support systems. For example, machine learning models could be trained using structured, ambiguity-aware data sets to not only classify by diagnostic category (e.g., normal vs. abnormal), but also to identify ambiguous cases and to explain those cases in terms of potentially controversial classification guidelines or evidence criteria pertinent to the data at hand [32].

Design Considerations for Expert Adjudication

Davies and Chandler [37] delineate five design categories of an online deliberation system: purpose, population, spatiotemporal distance, communication medium, and deliberative process (e.g., identifiability and structure). In this section, we designed and implemented an adjudication interface for expert users to engage in remote, anonymous, asychronous

adjudication of medical time series data in a web-based environment through a structured, guideline-centric procedure.

Purpose and Population. To facilitate adjudication of medical time series data in the context of sleep stage classification, we designed a system and user interface to emulate existing sleep scoring software, and embedded functionality for adjudication within. This ensured that users could engage in effective group deliberation in a familiar environment. True to the nature of our application domain, our system was aimed at expert users.

By engaging a population of expert users, we discovered that viewer settings play an important role in causing and resolving disagreement. For example, differences in signal visibility increased the likelihood of *resolving* disagreements through adjudication. While the reason for this is debatable, we suggest that experts' preferences and information needs in the context of making clinical decisions may vary with their level and type of professional experience. Designers of expert adjudication systems should therefore take into account the fact that both expert background and preferences for interface settings affect how assessments are made and how divergent assessments are adjudicated. However, differences in certain viewer settings (e.g., gain adjustments) were also associated with initial and persistent disagreement. While providing experts with sufficient amount of flexibility for viewer configurations seems necessary in order to enable exploration of complex medical data, adjudication systems may benefit from ways to share viewer settings between experts. In particular, if differences in viewer settings spur disagreement and make the resolution of discrepancies less likely, a feature allowing experts to view data "through the lens" of another grader and temporarily adopt the other experts' viewer settings may prove helpful for more effective adjudication.

Spatiotemporal Distance and Medium. In order to conduct a large-scale study with numerous expert users, we chose to deploy our system within a web-based environment, and had users participate remotely and asynchronously. While these decisions were largely informed by logistical reasons, we also wanted to design a system to enable effective adjudication in real-world contexts were local, real-time deliberation is infeasible. That said, we acknowledge that synchronous systems may be more effective at fostering agreement between users [37].

Deliberative Process. We enacted an anonymous deliberation process to eliminate user *identifiability* and reduce inter-personal bias during adjudication. Our findings on how grader differences affected disagreement dynamics should therefore be interpreted in the context of how different expert backgrounds may translate into different approaches to reasoning and arguing about corner cases, rather than bias introduced by mere perception of authority.

Adding *structure* to the process of collecting expert rationale allowed for detailed quantitative analyses of adjudication dynamics in our observational study. It is well documented that more structure fosters more deliberative behavior in an online deliberation setting [37]. However, in structuring a system around domain-specific annotation guidelines, structure can limit the efficiency of the workfklow when said guidelines are numerous and complex. Our design may have reduced input efficiency by forcing graders to navigate through a comprehensive set of rules irrespective of their classification decisions. However, unlike more confirmatory UI designs, we argue that this structure encourages participants to consider alternative lines of reasoning during adjudication. We showed how complex classification guidelines can be integrated into adjudication processes in a flexible and interoperable fashion. At the same time, our analysis leveraged a more compact view of expert rationale referencing basic feature types mentioned within the guideline rules. One design consideration by way of promoting input efficiency for structured rationales is to use the presence or absence of distinct, low-level features as an entry point for collecting expert rationales. A hybrid approach may solicit compact, feature-level assessments first, in order to intelligently recommend pertinent classification rules for adjudication in a second step.

Sharing and leveraging the insights of other users in a collaborative workflow has been found to increase task performance [60]. However, in the same study, Goyal and Fussell found that users who collaborated through an interface designed for shared sense making do not report an increased sense of success during the task, and view such an interface as having lower utility than standard setups. These reports suggest that users may need to be informed in real time about the utility of deliberation systems that employ new but important design elements—and may involve extra steps in the workflow—if such systems are to be readily adopted by new users.

Limitations

Despite the demonstrated use cases of adjudication, there are limitations to the process. First and foremost, adjudication is resource-intensive, a factor potentially hindering adoption in real-world contexts. Our study demonstrates how elements of structure can benefit adjudication procedures, and future work may explore how added structure could translate to increased efficiency and reductions in cost. While a quantitative cost-benefit analysis is beyond the scope of our study and will be left for future work, we demonstrate ways to counter the challenges of scheduling synchronous expert meetings through a round-based approach where experts can review cases on their own time. Future work may investigate hybrid methods encouraging turn-based adjudication procedures, while providing the opportunity for real-time communication for times when experts happen to review the same

case concurrently to make efficient use of their resources.

Our pilot study involved the same three experts conducting adjudication both in person and via video conference. While we ensured that experts discussed different cases in both stages, it is possible that certain behaviors observed via video conference may have been influenced by previous face-to-face interactions (e.g., perceived level of experience, word choice, intonation patterns). Our design considerations concerning inter-personal dynamics (e.g., choice of text as a communication medium) were primarily informed by in-person adjudication and subsequently reinforced by the possibility that these may also play a role in adjudication via video conference. Future work may explore the differential effects of communication media in medical adjudication using controlled between-subjects experiments.

Another aspect of the adjudication practice left for future work is the question of when to deploy such a system in a real-world context, clinical or otherwise. In settings where a single expert reader is the norm, what are the costs of introducing collective adjudication, given its demonstrated advantages in medical data analysis? If adjudication is deemed too costly to be routine, what are the indications that may alert clinicians to when group deliberation is necessary? Our findings demonstrate that adjudication outcomes can translate to changes in diagnostic measures, suggesting that the use of adjudication should be prioritized for those *critical* disagreement cases that have the highest potential of impacting patients through revisions in treatment outcomes.

3.3.8 Conclusion

In this section, we introduced a novel perspective on the problem of expert disagreement in medical data analysis using a structured form of collaborative adjudication to study the nature and dynamics of disagreement from a socio-technical perspective. We demonstrated the applicability of our approach in the context of medical time series analysis for sleep stage classification, and showcased how the structured data produced can facilitate a deep understanding of the diverse factors playing a role in generating and resolving disagreements, including expert background, data complexity, viewer settings and classification guidelines. Our proposed workflow for structured adjudication has implications for the design of decision support for clinical group decision making and for the collection of expert-labeled data in the context of other applications like computer-aided diagnosis.

3.4 Conclusion

In this chapter, we have explored the question of how group deliberation can be used as a tool within data labeling workflows to help analyze instances of ambiguity resulting in inter-rater disagreement. We reported three case studies of using both synchronous and asynchronous deliberation workflows not only in novice crowd work, but also in the expert domain of medical data interpretation. Our studies have demonstrated that group deliberation is a versatile approach for addressing ambiguity in various data modalities, including text documents, images and time series data, and that imposing structure on the process does not only improve efficiency, but also helps us understand why disagreement arises and when it may be resolved. The remainder of this dissertation aims to demonstrate that the deliberation data collected in this process can be put into use to not only train less experienced human labelers, but also to simulate and explore AI systems that highlight and explain ambiguity in human-AI collaborative data analysis workflows.

Chapter 4

Deliberation Data for Labeler Training

Workflows for medical data labeling 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.

4.1 Motivation

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 [16]. However, medical specialists sufficiently trained to perform complex diagnoses are exceptionally rare [16].

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 [28]. As a result, contemporary ML-powered algorithms typically rely on medical practitioners to manually label data ground truth [28, 63, 138].

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 [82]. 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 [140]. This is an enticing approach given that medical generalists far outnumber specialists (e.g. optometrists outnumber ophthalmologists 4.9-fold [5]). Yet, generalists are less experienced: difficult patient cases lead to high inter-labeler variability and incorrect diagnoses [82], limiting algorithmic validity and introducing the risk of adverse outcomes for patients' lives.

In this chapter, 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.

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.

4.2 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 [141]. 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 4.1). Digital fundus photos are used both in tele-medical screening [134] and for the development of deep learning models for AI-assisted retinal assessment [63, 108]. 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 [82]. 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 [5]: (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.

4.3 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 inverventions, 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**.

4.4 Methods

4.4.1 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 the remote, round-based protocol for group discussion described in Section 3.2. 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.

4.4.2 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.

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 discus-

sions. 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.

4.4.3 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.

A pre-study questionnaire (Appendix B.1) and a post-study questionnaire (Appendix B.3) were administered before and after the study respectively. The pre-study qustionnaire 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 [122]. Finally, for each case, generalists provided an open-ended explanation of the reasoning behind their rationale. Figure 4.1 shows the task interface including all input prompts for a gradable image.

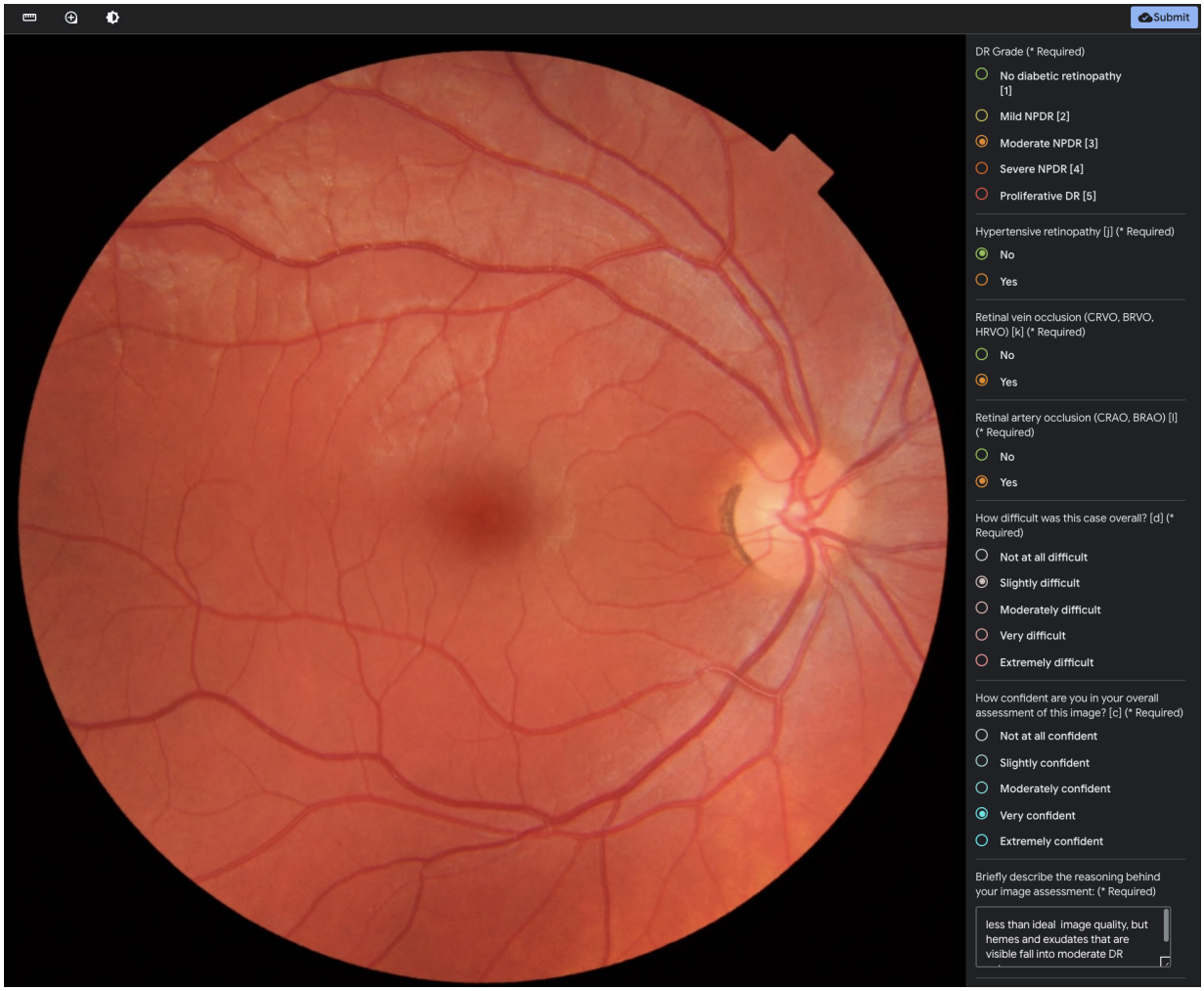


Figure 4.1: Task interface for medical image assessment. The medical image shown is an illustrative example rather than from the real dataset.

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. Generalists were asked to review each case where their answer differed from the answer key.

For each case reviewed, generalists filled out a short questionnaire (Appendix B.2), rating their level of comprehension for the rationale behind the answer key (5-point Likert

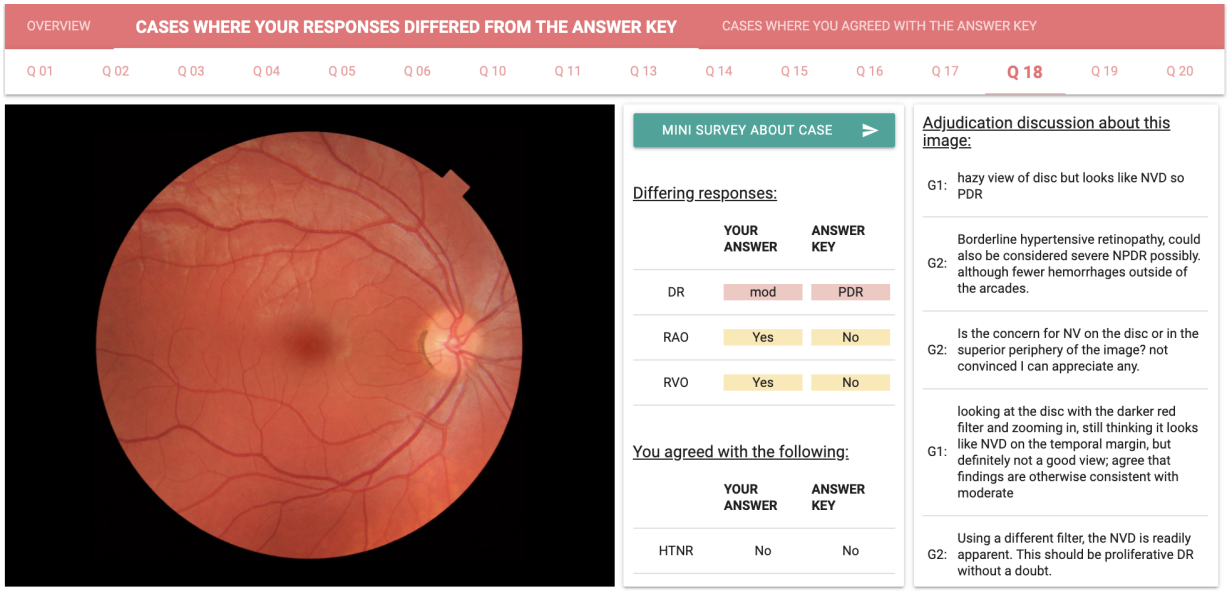


Figure 4.2: Training feedback interface for medical generalists. The medical image shown is an illustrative example rather than from the real dataset.

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 questionnaire 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.

4.4.4 Experimental Conditions

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 4.2, 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 4.2, 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.

4.4.5 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 performed by one of the study authors to identify emerging themes and recurring themes are reported below.

4.5 Results

4.5.1 Quantitative Insights

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 4.3, 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 4.3, middle). Finally, generalists who were provided with adjudication discussions during training

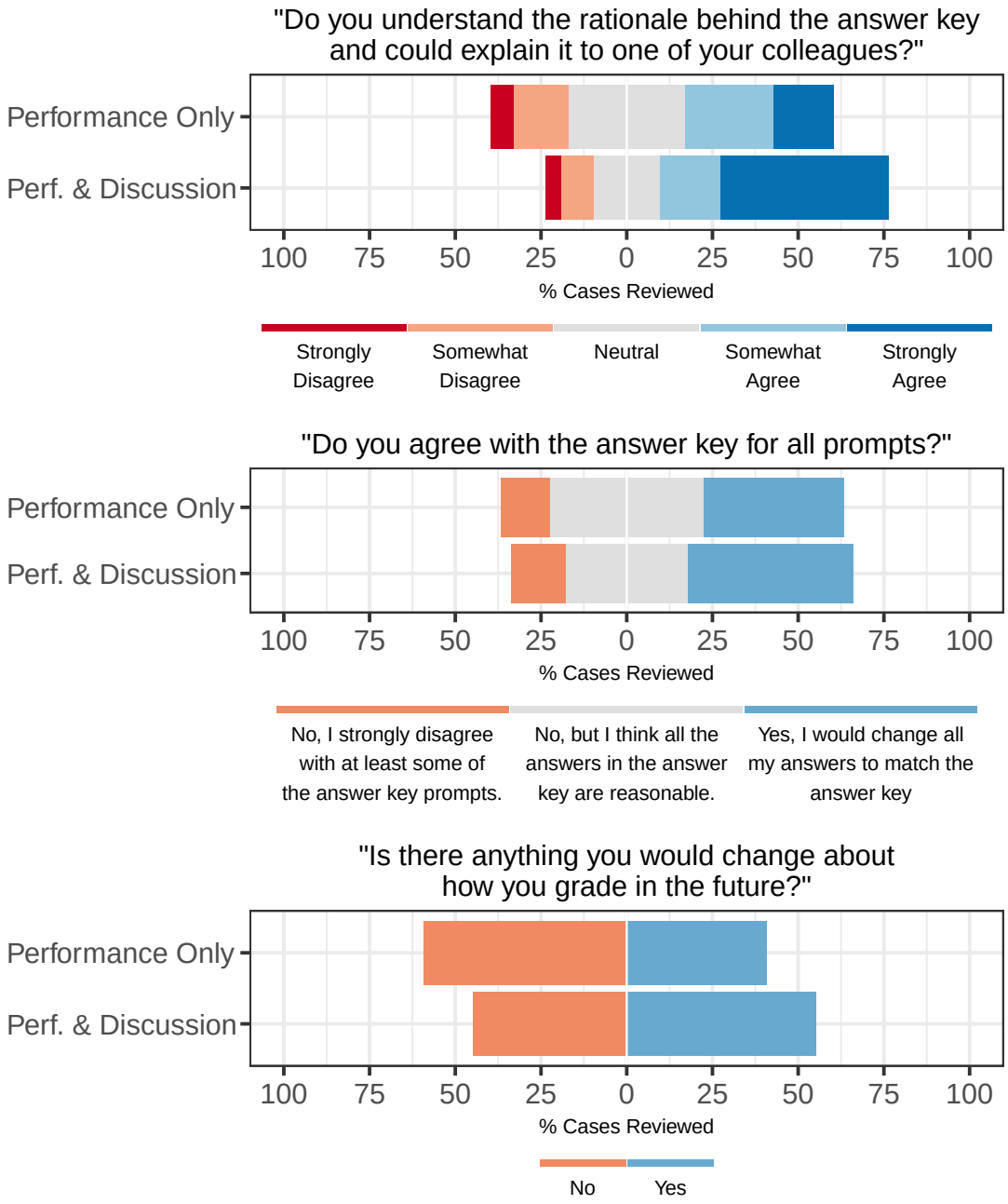


Figure 4.3: Generalists' perception of training feedback.

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 4.3, 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.

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

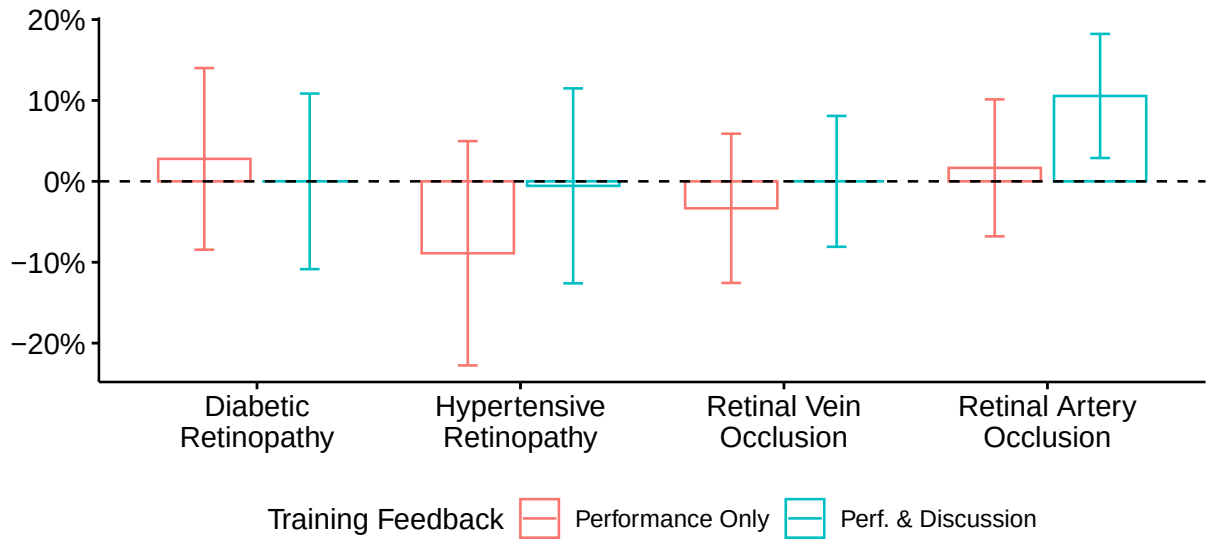


Figure 4.4: Average change in generalists' diagnostic accuracy per case-pair in train set and held-out test set. Error bars indicate 95% confidence intervals.

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 4.4). Generalists exposed to adjudication discussions during training feedback showed significantly greater accuracy improvements for diagnosing RAO (μ = 10.6%, CI [2.9%, 18.2%]) than generalists exposed to performance feedback alone (μ = 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 4.5).

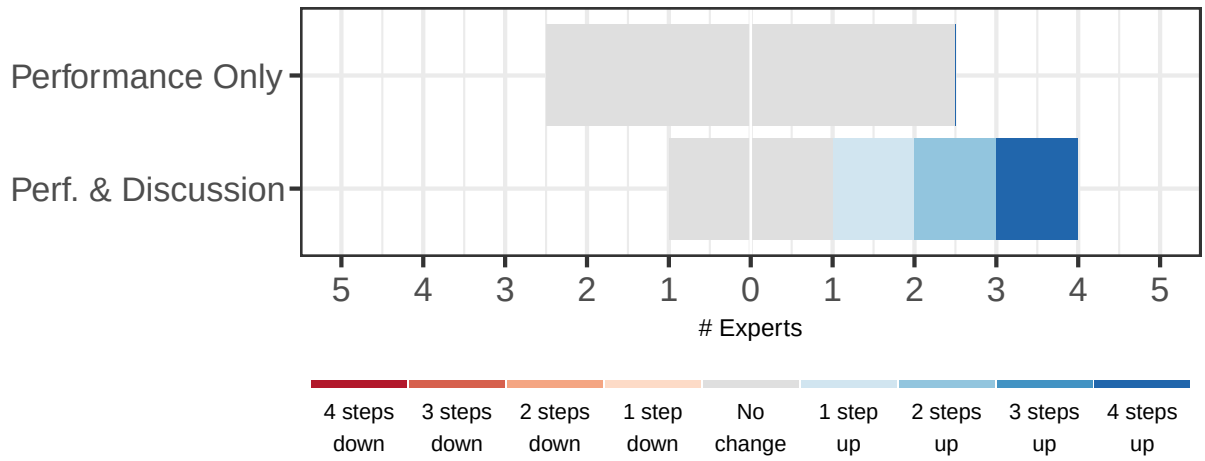


Figure 4.5: Improvement in generalists' self-efficacy score for diagnosis of retinal artery occlusion (RAO) after training feedback.

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 4.6). 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.

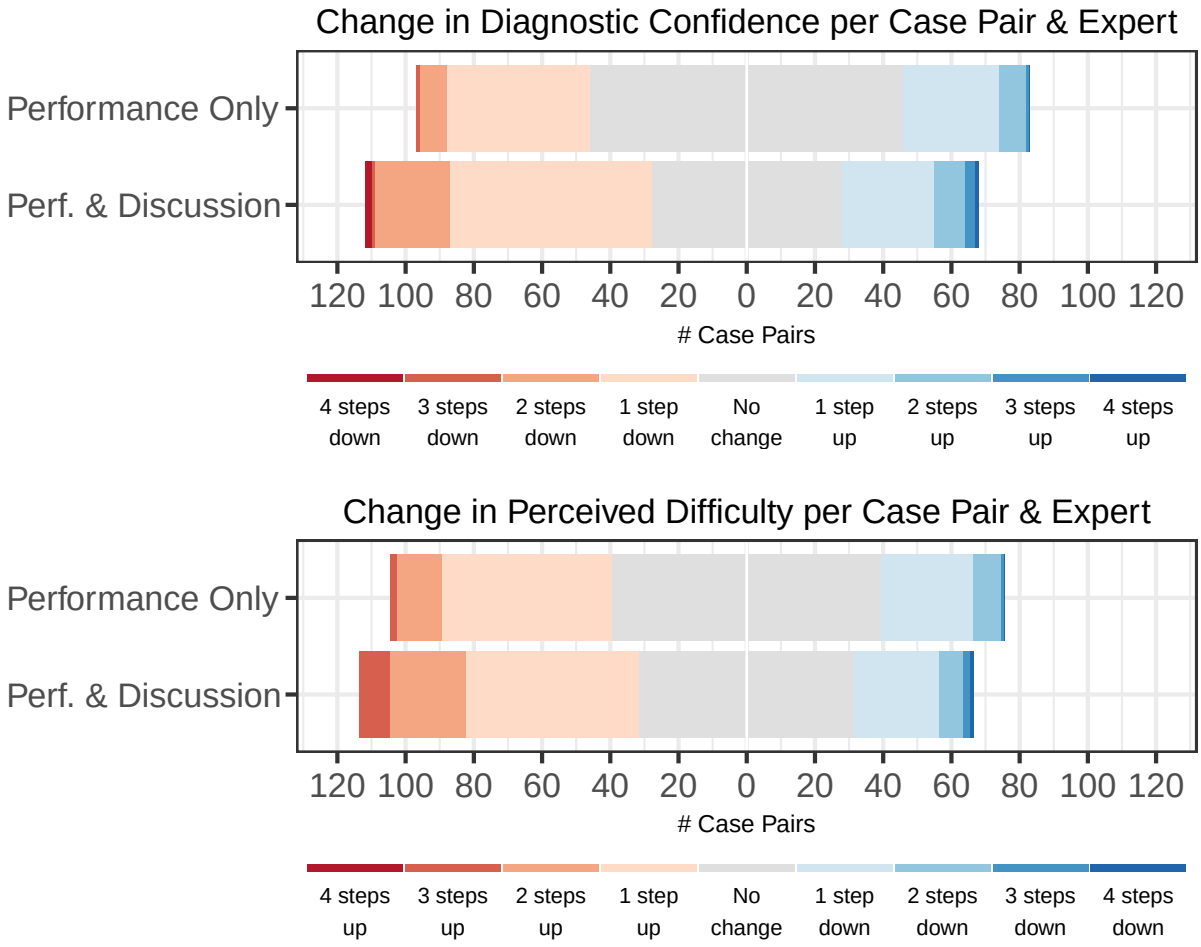


Figure 4.6: Change in generalists' diagnostic confidence and perceived case difficulty per case-pair in train set and held-out test set.

4.5.2 Qualitative Insights

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:

- *"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. [77] 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 [90].

4.6 Discussion

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

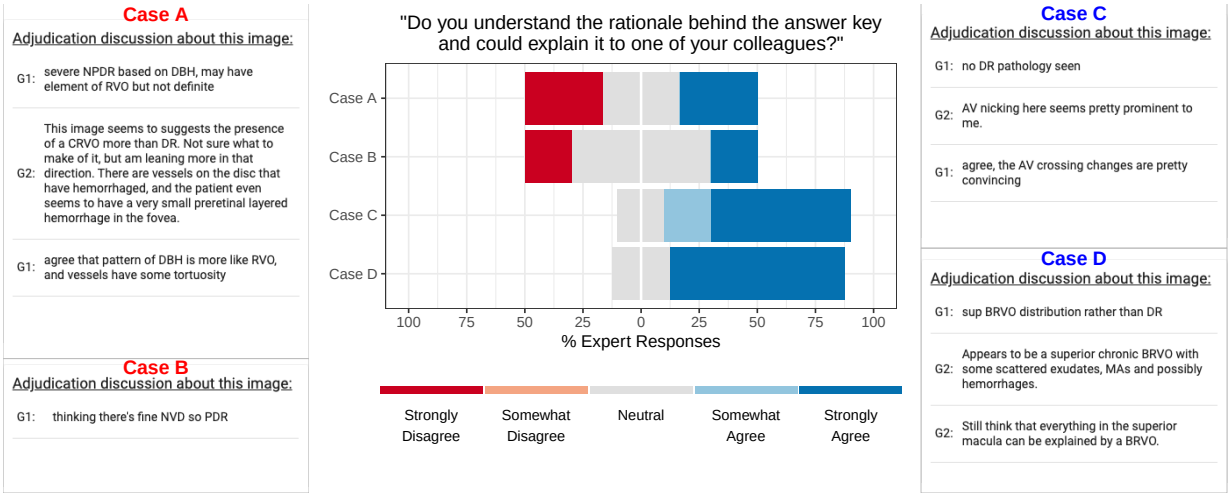


Figure 4.7: Example adjudication discussions with mixed ratings (cases A and B) and consistently high ratings (cases C and D) for answer key comprehension.

4.6.1 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 4.7 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.

4.6.2 Impact on 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 [45] 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.

4.6.3 Learning from Discussions

Our work presented in this chapter 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) [40, 137]. 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 [68] 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.

4.6.4 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 [89], 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 [34, 24]. Likewise, the adjudication discussions we collected may mirror arbitration discussions used in some screening programs [90]. 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.

4.6.5 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 [15] and the effectiveness of discussion-based learning [38] 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 [127].

4.7 Conclusion

In this chapter, 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. The findings presented in this chapter have important implications beyond medical diagnosis training alone, highlighting a practical method applicable to expert labeler training in high-stakes data labeling broadly.

While this chapter has demonstrated how deliberation data can be put into use to improve human comprehension of difficult cases, the next chapter will shed light on the question of how deliberation data can be leveraged to help AI assistants highlight and explain instances of ambiguity to human end users.

Chapter 5

Deliberation Data for Ambiguity-aware AI

AI assistants for clinical decision-making show increasing promise in medicine. However, medical assessments can be contentious, leading to expert disagreement. This raises the question of how AI assistants should be designed to handle the classification of ambiguous cases. Our study compared two AI assistants that provide classification labels for medical time series data along with quantitative uncertainty estimates: conventional vs. *ambiguity-aware*. We simulated our ambiguity-aware AI based on real-world expert discussions to highlight cases likely to lead to expert disagreement, and to present arguments for conflicting classification choices. Our results demonstrate that ambiguity-aware AI can alter expert workflows by significantly increasing the proportion of contentious cases reviewed. We also found that the relevance of AI-provided arguments (selected from guidelines either randomly or by experts) affected experts' accuracy at revising AI-suggested labels. The work we present in this chapter contributes a novel perspective on the design of AI for contentious clinical assessments.

5.1 Motivation

AI systems show increasing promise for numerous clinical applications. Recent advances in deep learning have spawned AI systems with expert-level performance in several domains of medical data classification (e.g., [115, 116, 138]). However, contentious patient cases leading to expert disagreement are prevalent in medicine [82]. Given the gravity of correct

clinical assessments, an important question in the design of AI for medical data analysis is how the system should communicate uncertainty about the classification of ambiguous cases.

State-of-the-art AI systems are capable of providing quantitative uncertainty estimates (e.g., 70% confident that a patient case is abnormal). These estimates are typically derived from posterior probability distributions over the possible classification labels. However, prior work has shown that these estimates do not always reliably predict expert disagreement [114]. Furthermore, numeric representations of uncertainty alone may not be sufficient for human experts to make sense of the underlying reasons behind the AI's uncertainty.

Prior work in explainable AI (XAI) has established the importance of providing reasons for AI-suggested labels to foster model transparency and user trust [3, 113, 147]. Building on this body of work, we argue that explanations for label ambiguity can be leveraged by AI assistants to support medical reasoning. We detail a within-subject study with twelve expert participants who interacted with both a conventional and an *ambiguity-aware* AI assistant, reviewing a total of 4,514 AI-suggested labels, out of which 22% were contentious. Both assistants used quantitative representations to communicate uncertainty, but our ambiguity-aware AI also highlighted contentious cases and explained why they were ambiguous by providing human-interpretable arguments for the conflicting labels. While this feature was simulated using cases and arguments selected from real-world expert discussions, participants were unaware of its simulated nature. Our findings suggest that explaining ambiguity can benefit AI-assisted medical reasoning. Our main contributions are:

1. We present a novel approach for communicating ambiguity in AI-assisted medical reasoning, and provide evidence that ambiguity-aware AI can alter experts' workflows by effectively re-directing their attention and review activity to contentious cases.

2. We demonstrate that while explaining ambiguity can contribute to experts' labeling accuracy, its impact heavily depends on the relevance of the arguments provided (selected from guidelines either randomly or by experts). Specifically, if the arguments are not sufficiently relevant, experts' accuracy can suffer to the point below that of random guessing (i.e., less than 50% accurate).

3. We provide design considerations for communicating uncertainty in AI-assisted medical reasoning, laying a foundation for future implementations of AI systems better capable of conveying information about contentious cases.

In the following sections, we introduce the design of our AI assistants, followed by our research questions, hypotheses and methods. We then detail our quantitative and qualitative findings, and conclude with a discussion of design considerations.

Figure 5.1: Interface for conventional and ambiguity-aware AI assistants in medical data analysis.

5.2 Ambiguity-aware AI Assistance

In this study, we explore how human-AI collaboration is affected by an AI system's ability to not only flag *if* specific cases are on the classification boundary between two or more categories, but also explain *why* a given case may be ambiguous. Specifically, we compare a simulated AI system that provides experts with arguments for conflicting classification choices for a contentious case to a conventional AI assistant that only provides numeric uncertainty estimates. Our ambiguity-aware AI system uses a Wizard of Oz approach. That is, justifications for conflicting classification labels were hand-authored by human experts using the round-based discussion procedure described in Section 3.3. To compare the ambiguity-aware AI assistant to a conventional AI assistant, we led participants to believe that the justifications presented to them were generated by an AI while, in fact, they were manually selected by human experts.

Figure 5.1 illustrates how the two AI assistants—conventional AI vs ambiguity-aware AI—were integrated into an existing expert interface for classification of medical time series

data. Both AIs suggested classification labels based on a state-of-the-art deep learning algorithm for sleep stage classification [138], which has an average accuracy of 87% (when judged against consensus labels from an expert panel). Both AI assistants provided a sequence overview of all suggested labels (hypnogram), in which each label corresponded to a 30-second segment in the timeline of a multi-hour patient recording. Experts could open a case by selecting the corresponding time window in the overview, or by navigating through the recording chronologically.

The key difference between the two AI assistants was in how they communicated uncertainty to expert end users. Typical output from machine learning algorithms includes not only the predicted classification label, but also a likelihood distribution over all possible classification choices. Both of our AI assistants were designed to communicate this type of *quantitative* uncertainty estimate in two ways (Figure 5.1, blue labels 3 and 4): (1) in the timeline overview, quantitative uncertainty was visualized by mapping the confidence level (in percentage) for each possible classification label to a *transparency* value used to display the label option in the timeline—low confidence classification labels were more transparent, and high confidence classification labels were more opaque; (2) in the case detail view, quantitative uncertainty was displayed in a tabular format, listing all possible classification choices ordered from most to least likely along with their *percentage* confidence levels.

While our conventional AI employed this baseline representation of uncertainty, our ambiguity-aware AI also communicated *qualitative* uncertainty based on arguments gathered from real-world expert discussions (Figure 5.1, orange labels 5 and 6). Specifically, the timeline overview was augmented with an additional layer *highlighting contentious cases* that were likely to spur expert disagreement. Note that these suggestions did not dictate the order in which cases were presented to experts for review: experts were still free to decide how to navigate the recording timeline and what cases to review in which order. In addition, the case detail view for contentious cases was extended with an *ambiguity explanation*, listing human-interpretable arguments for conflicting classification choices. These arguments corresponded to discrete scoring rules from the official guidelines for sleep stage classification, and were based on data from real-world expert discussions as described above.

5.3 Research Questions and Hypotheses

Our work addresses two primary research questions about the impact of ambiguity-aware AI on the behaviour (Q1) and perception (Q2) of medical experts.

Q1: How does ambiguity-aware AI affect medical assessments?

Expert time is a limited and expensive resource in clinical settings and should therefore be allocated efficiently. We take the stance that while medical experts should make their clinical assessments with care, AI assistants can help prioritize which cases require their attention the most. Our ambiguity-aware AI is designed to redirect experts' attention towards cases likely to be contentious, and to provide arguments explaining the underlying classification ambiguity.

Our projection is that ambiguity explanations can inform clinical judgement and thus increase experts' classification accuracy without reducing the number of cases reviewed. Specifically, we envision that the relevance of ambiguity explanations is crucial for successfully informing expert judgement. We hypothesize that:

[H1a] The proportion of contentious cases reviewed by experts will be higher with an ambiguity-aware AI.

[H1b] Expert efficiency in terms of the overall number of cases reviewed will not suffer with an ambiguity-aware AI.

[H1c] Expert accuracy in terms of the overall portion of cases reviewed and labeled correctly will be higher with an ambiguity-aware AI.

[H1d] The accuracy of classification labels experts assigned to contentious cases will depend on the relevance of the provided ambiguity explanations.

Q2: How is ambiguity-aware AI perceived by medical experts?

HCI research has established that poor user perception can be a barrier to adoption of technology regardless of performance. It is therefore important to investigate expert perception, beyond the primary outcome of reliability in AI-assisted clinical assessments. We hypothesize that:

[H2a] Experts will have a preference for an ambiguity-aware AI.

[H2b] Experts will consider an ambiguity-aware AI more trustworthy.

[H2c] Highlighting and explaining contentious cases will not increase experts' cognitive load.

[H2d] Experts with higher ambiguity tolerance (as a personality trait) will have a stronger preference for the ambiguity-aware AI.

5.4 Methods

Here we describe the details of our controlled experiment including the task, data set, study procedure, and statistical analysis. In our study, we simulate a scenario in which a medical AI assistant first analyzes a patient case to suggest classification labels of a certain kind. A trained medical expert then reviews and corrects as many AI-suggested classifications as possible within a given time window. This setting represents a future scenario where (imperfect) AI systems are deployed in time-sensitive clinical workflows while requiring oversight from human experts.

5.4.1 Task

We conducted our study in the field of biomedical time-series classification, an expert domain with typically high rates of inter-rater disagreement. In particular, we compared our conventional and ambiguity-aware AIs in the context of assisting trained medical professionals in the task of sleep stage classification (see description in Section 3.3.2). Figure 5.1 shows the expert classification interface used in our study.

5.4.2 Data

We selected two separate patient records (i.e, polysomnographic sleep studies) with similar characteristics (Table 5.1) to examine the two AI assistants under comparable conditions while avoiding learning effects on the side of experts.

Note that patient records where selected such that the AI accuracy measured against just the contentious cases was close to 50% for both patients, meaning that correction by human experts was only required for about half of those cases. In addition to counter-balancing the order in which conventional and ambiguity-aware AI were presented to experts, the assignment of AI assistant to patient record was also fully counter-balanced. A separate third patient record was randomly selected for a practice phase preceding the main task.

Adjudication data. We source the data required to simulate our ambiguity-aware AI (i.e., which cases have high expert disagreement, and what are the arguments for different classification labels) from our previous study reported in Section 3.3. This prior work introduced a round-based procedure to adjudicate clinical classification disagreements among groups of experts using a highly structured argument format. In particular, arguments

Table 5.1: Characteristics of patient records used by the AI assistants.

	Patient A	**Patient B**
Pathology	Dementia	Dementia
Sex	Female	Male
Age Group	70-74 years	75-79 years
Recording Duration	6h 52 min 30 sec	6h 18 min 30 sec
# Cases Total	825	757
% Contentious Cases Total	18%	18%
% Contentious Cases out of all Correct AI Suggestions	12%	11%
% Contentious Cases out of all Incorrect AI Suggestions	48%	51%
AI Accuracy Overall	84%	83%
AI Accuracy on Contentious Cases	55%	51%

were collected in the form of discrete classification rules taken from the official medical guidelines (e.g., *In patients who generate alpha rhythm, score stage N1 if the alpha rhythm is extenuated and replaced by low-amplitude, mixed frequency activity for more than half of the epoch*).

This data set was used to simulate output for our ambiguity-aware AI: cases that had caused expert disagreement and produced conflicting arguments in this data set were highlighted as contentious cases by our ambiguity-aware AI. Arguments put forward during the real-world adjudication process were presented for these cases to explain the ambiguity around conflicting classification labels. For Q1, we sought to examine the impact of argument relevance on clinical decision making for contentious cases (H1d). To this end, we added noise to ambiguity explanations by replacing a random subset (20%) of arguments with scoring rules randomly selected from the same medical guidelines. Otherwise, justifications were displayed as selected by experts during prior discussions, without further manipulation. Our randomization procedure was constrained to ensure that randomly selected arguments were never mentioned in the real-world expert discussion for a given case, and that all arguments presented were still pertinent to their classification choice: for example, an argument for REM sleep could only be replaced with another argument for REM sleep.

Finally, classification accuracy (of either AI or human experts) was measured against the consensus decision of our round-based adjudication procedure involving a panel of three independent experts for each classification decision.

5.4.3 Procedure

We recruited twelve sleep technologists as expert participants for our study. Our experts were recruited with the help of an allied sleep technologist from a local research clinic who posted our recruitment letter to a domain-specific Facebook group with about 4700 sleep technologists from different countries. Each expert was exposed to both AI assistants in a counter-balanced manner.

Consent procedure and pre-study questionnaire. After providing informed consent for participation in the study, experts reported information about their demographics (age, gender, geographic location) and professional background (professional or academic training, number of years of professional experience). We employed the *Intolerance of Ambiguity* scale, a psychometric survey instrument developed by Budner [20], to learn about each expert's general level of tolerance for ambiguity in decision making. We included the phenomenological denial sub-scale consisting of four statements:

- *An expert who doesn't come up with a definite answer probably doesn't know too much.*
- *There is really no such things as a problem that can't be solved.*
- *People who insist upon a yes or no answer just don't know how complicated things really are.*
- *Many of our most important decisions are based on insufficient information.*

Experts rated their level of agreement for each of the four statements on a 7-point Likert scale. Appendix C.1 provides a complete list of questions and answer options from the pre-study questionnaire.

Practice phase. Next, experts familiarized themselves for about 5 minutes with our waveform classification user interface and with the basic interface components common to both AI assistants.

Tasks. Experts performed the same main task twice, once with the ambiguity-aware AI assistant and once with the conventional variant, in a counter-balanced order. In each task, experts were asked to review the waveform of a particular patient record within a limited time window of 15 minutes. The patient record was fully pre-classified by the AI assistant and experts were asked to correct as many of the AI-suggested labels as possible within the given time limit. Experts could revise AI-suggested labels by selecting a different sleep stage label in the classification UI (Figure 5.1, gray labels 1 and 2). After each of the two tasks, experts filled out a brief feedback questionnaire (Appendix C.2) probing for their

perception of each AI assistant. The survey included scales to measure perceived trust towards the AI assistant [73], cognitive load (NASA-TLX; [66]) during the task, perceived diagnostic utility and mental support provided by the AI assistant, and whether experts thought they would use the AI in practice.

Post-study questionnaire. After completing the tasks, experts compared both AI assistants with respect to perceived reliability, trustworthiness, capability and provided an overall preference. Experts rated each of these four items on a 7-point Likert scale ranging from 1 (totally version A), 2 (much more version A than B), 3 (slightly more version A than B), 4 (neutral), etc. to 7 (totally version B). Appendix C.3 provides a complete list of questions and answer options from the post-study questionnaire. After completing the post-study questionnaire, participants received a debrief statement informing them about the simulated nature of the ambiguity-aware AI in this study. Experts were compensated with CA$50 via online gift cards (or the equivalent amount in their preferred currency) for participation in the study, with an average study duration of one hour.

5.4.4 Analysis

For **Q1**, we investigated the impact of our ambiguity-aware AI on experts' behaviour in reviewing AI-suggested classification labels. We used dependent t-tests to compare both AI assistants with respect to the following outcome measures per expert: the proportion of contentious cases out of all reviewed cases (H1a), the number of cases reviewed given a fixed time window (H1b), and the accuracy rate of expert-provided labels (H1c). For our secondary analysis on the relevance of arguments for contentious cases (H1d), we used Pearson's chi-squared test of independence to compare experts' average accuracy at revising AI-suggested labels when presented with either expert-selected arguments only vs. cases with one or more randomly selected arguments.

For **Q2**, we compared experts' perception of both AI assistants. A possible trend in overall preference (H2a) for either of the AI assistants was examined using a one-sample Wilcoxon signed rank test. Self-reported scores for perceived trust (H2b) and cognitive workload (H2c) were compared between both AI assistants using Wilcoxon signed-rank tests. Finally, we used a Pearson's chi-squared test of independence to test whether experts' overall tendency of ambiguity tolerance (ambiguity-tolerant vs. intolerant) was associated with their overall preference for either AI assistant (preference for ambiguity-aware AI vs. conventional AI; H2d).

Finally, line-by-line inductive open coding was performed by one of the study authors to extract emerging themes from open-ended survey responses submitted by experts after

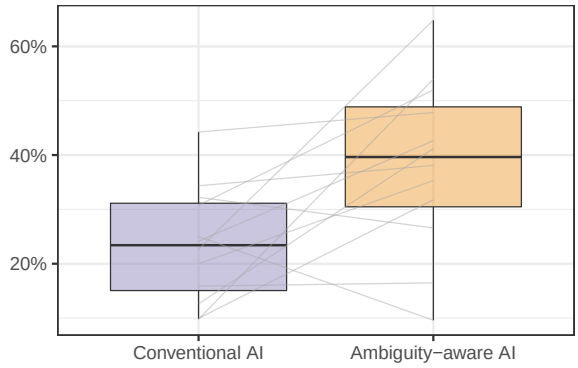


Figure 5.2: Proportion of contentious cases out of all cases reviewed. Ambiguity-aware AI guided experts' attention to contentious cases. Connecting lines correspond to individual experts.

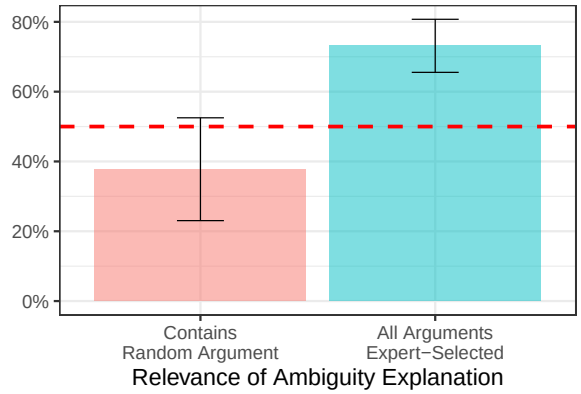


Figure 5.3: Experts' correction rate for cases with ambiguity explanation. The relevance of ambiguity explanations affects clinical assessments of contentious cases. Error bars present 95% confidence intervals.

interacting with each AI. Experts were asked to reflect on how they decided which cases to review and why, what information they used to make these decisions, and how information about the AI's uncertainty affected their decision making. Recurring themes are reported below.

5.5 Results

5.5.1 Expert Participants

Based on the pre-study questionnaire, our expert participants were located in the United States (6), Canada (4), the European Union (1) and one other unspecified location (1). Eleven of our expert participants reported having at least ten years of experience working as sleep technologists, and one participant reported having five to ten years of experience. Out of the twelve experts, five self-reported as female, six as male, and one participant did not specify their gender. The distribution over age groups was: 26-35 (1), 36-45 (7), 46-55 (1), 56+ (2), with one participant who did not specify their age group.

5.5.2 Quantitative Insights

Q1: How does ambiguity-aware AI affect medical assessments?

We hypothesized that the ambiguity-aware AI assistant would alter experts' workflow and increase the number of contentious cases they review in the patient recording **(H1a)**. On average, the proportion of contentious cases out of all cases reviewed was significantly greater with the ambiguity-aware AI (M=.38, SE=.05) than with the conventional AI (M=.23, SE=.03), confirming our hypothesis (Figure 5.2). This difference was significant t(11)=-2.82, $p < .05$, indicating a large effect size r=.48.

We also hypothesized that using the ambiguity-aware AI would not negatively affect the number of cases reviewed by experts **(H1b)**. Our results show that there was no significant difference in the number of cases experts reviewed with the conventional AI (M=197.25, SE=47.60) compared with the ambiguity-aware AI (M=178.92, SE=57.02), t(11)=.50, p=.63. This result provides support for our hypothesis that experts' efficiency at reviewing AI-suggested labels was not negatively affected by being exposed to ambiguity explanations for contentious cases. Our projection that experts would achieve a higher overall labeling accuracy when assisted by the ambiguity-aware AI compared to the conventional one **(H1c)** could not be confirmed, t(11)=1.00, p=.34, r=.53.

Finally, we examined the potential impact of the relevance of ambiguity explanations for contentious cases on the likelihood that an expert would revise an AI-suggested label correctly **(H1d)**. We observed a significant association between the relevance of arguments (whether they contain randomly selected arguments or not) and experts' accuracy rate at revising AI suggestions χ^2=16.83, $p < .001$. In other words, the chance of a label getting revised correctly by an expert was significantly higher if the arguments provided were selected from guidelines via adjudication discussions (i.e., were relevant) than if they were selected from the guidelines randomly (Figure 5.3). The odds of a label getting revised correctly by an expert were 4.48 times higher (odds ratio) if the arguments provided were selected from guidelines by experts than if they were selected from the guidelines randomly.

Q2: How is ambiguity-aware AI perceived by medical experts?

For Q2, we explored experts' perception of both AI assistants. Results for our hypothesis that experts would have an overall preference for the ambiguity-aware AI **(H2a)** were mixed and were not statistically significant (p=.88). Except for two experts who did not have a preference for either AI, preferences were polarized. Out of the ten participants who expressed a preference, half preferred the ambiguity-aware AI assistant and the other half preferred the conventional AI (Figure 5.4).

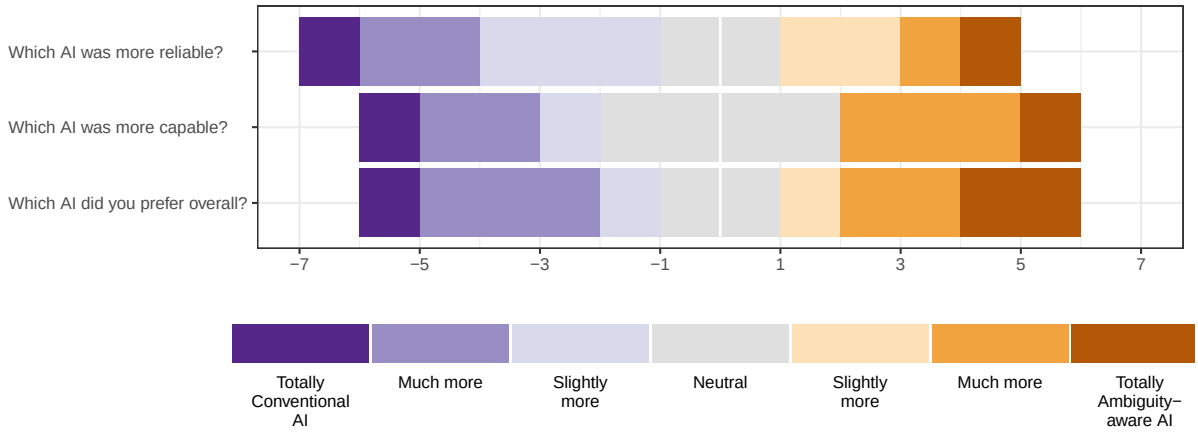


Figure 5.4: Experts' preferences between both AI assistants.

While no significant differences could be detected between both AIs regarding perceived overall trust ($p=.47$), the ambiguity-aware variant was considered to have significantly greater integrity ($p<.05$), and we observed a potential, yet statistically insignificant trend suggesting that experts may have had higher confidence in the ambiguity-aware AI than in the conventional one ($p=.09$; Figure 5.5). These results provide partial support for our hypothesis **H2b**.

Furthermore, there were no detectably significant differences between the cognitive load scores of the two AI assistants on the NASA-TLX scale ($p=.77$), providing support for our hypothesis about their comparable mental demand **(H2c)**.

Finally, while experts varied in their level of ambiguity tolerance ($M=17.25$, $SE=1.17$), ranging from 10 to 26 on a scale from 4 to 28, no significant effect of ambiguity tolerance on expert perception could be detected ($p=.62$), leading us to reject hypothesis **H2d**.

5.5.3 Qualitative Insights

Our qualitative analysis of participant responses to open-ended survey questions yielded insights on how our ambiguity-aware AI assistant can affect experts' workflows and their mental model of AI assistants.

Altering expert workflows. Time constraints play an important role in real-world clinical workflows [143]. Case triaging—determining the priority for which cases receive an

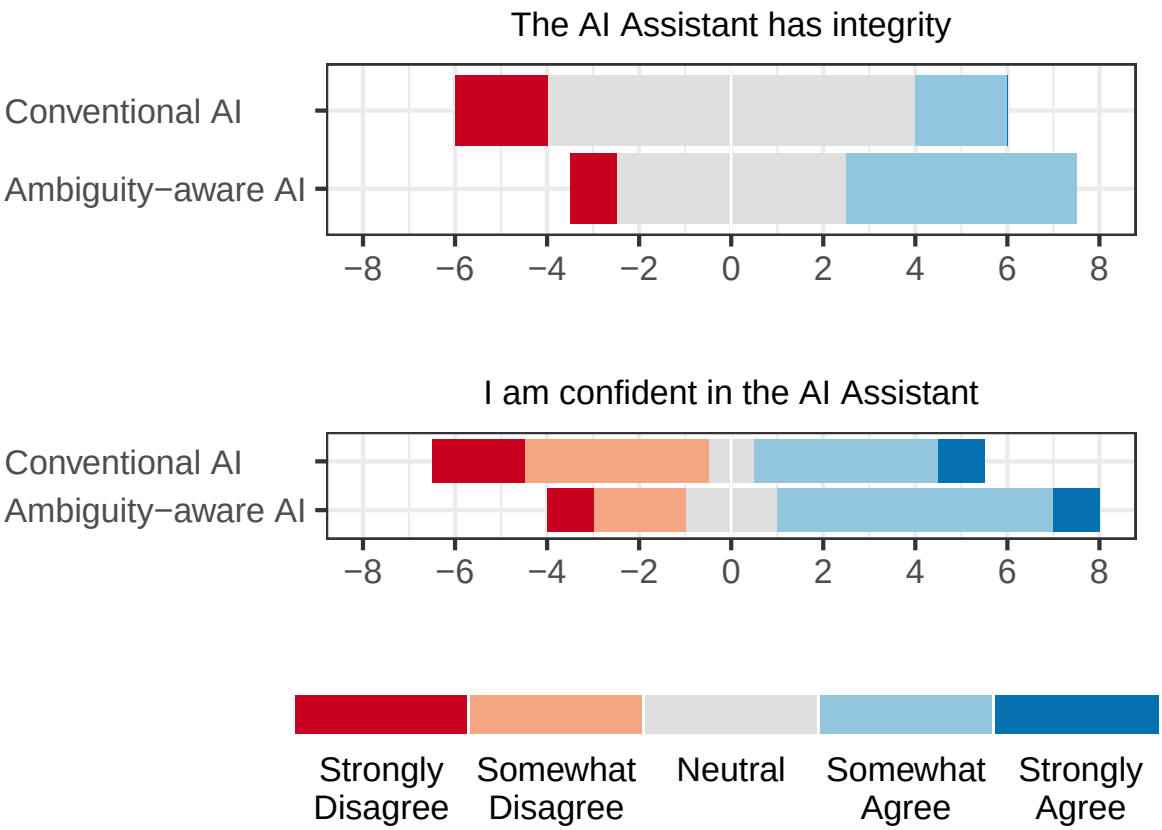


Figure 5.5: Expert ratings for perceived integrity and confidence from trust in automation scale.

expert's attention first—is a common practice in medicine. Similarly, our ambiguity-aware AI assistant triages based on ambiguity by prioritizing contentious patient cases that need more attention from the expert.

Our qualitative findings suggest that some experts found the ambiguity-aware AI system to be more helpful in reducing cognitive load compared to the conventional assistant: *"Assistant B [ambiguity-aware] was more helpful in making me think as it listed the scoring rules that could apply to the epoch."*

Our analysis further highlights the effectiveness of the ambiguity-aware AI assistant in redirecting experts' attention to contentious cases. That is, six out of twelve experts in our study explicitly mentioned that their workflow differed between the two AI assistants, such that they prioritized checking contentious cases using the ambiguity-aware AI: *"I first*

chose the areas that the AI had marked as ambiguous and then tried to check sleep onset, REM onset, and stage 3 as time allowed."

One major criticism to the traditional approach of representing AI uncertainty with numeric confidence values is that it is not sufficient for experts to make sense of the underlying reasons behind the AI's uncertainty. Our qualitative evidence suggests that in choosing between numeric representations of AI uncertainty and human-interpretable ambiguity arguments experts found the latter to be more effective in guiding their attention: *"When I saw that the [conventional] AI had lower than an 80% confidence in the scored stage I tried to double check that epoch... I mostly used the areas marked as ambiguous [by the ambiguity-aware AI] as opposed to the percentage of certainty."*

In our study, we imposed time limits to understand how ambiguity-aware AI would help guide expert attention under the time constraints of real-world workflows. This temporal constraint was received differently by different expert participants. While some experts perceived the timers to be *"very frustrating"*, others found them useful: *"The time limit was great as my first instinct was to review the entire study and see if I was in agreement"*.

Mental models of AI assistants. Experts have preconceived mental models about the level of ambiguity in different cases. For instance, experts may draw from their prior experience of disagreements with other colleagues and have intuitions about what type of medical assessment is the most difficult to agree upon in their specific domain (e.g. certain classifications and stage transitions). It is therefore possible that these intuitions are projected onto the AI assistant to anticipate where the AI would likely make mistakes: *"I had to think where do we, as scoring techs, usually have the strongest disagreement and check those epochs."*

Beyond preconceptions, we also observed that experts developed their own mentals models about the two types of AI systems: *"AI 1 [ambiguity-aware] was rather impressive actually. Although in study 2, the persistent arousals may have interfered with accuracy of AI 2 [conventional]."* Further, their interaction experience with the same AI assistant can also shape their judgement of where they will likely disagree with the system: *"On 'B' [ambiguity-aware], I tried to focus more on the ambiguous epochs indicated by the AI and then on the staging that the AI in 'A' [conventional] did not perform well with."* AI assistants could leverage this insight by grouping contentious cases based on an expert's reviewing and correction behaviour to adjust to their internal representation of specific types of ambiguity.

5.6 Discussion

In this chapter, we studied how highlighting and explaining ambiguity by AI assistants can aid medical experts in their decision making for contentious clinical cases. We conducted a within-subjects study to investigate the use of ambiguity-aware AI assistants by medical experts. Our results show that the ambiguity-aware AI can alter experts' workflows by increasing the proportion of contentious cases reviewed while maintaining overall productivity.

While experts' overall labeling accuracy was not affected by providing ambiguity-awareness, we observed a significant effect of argument relevance on experts' case correction rate. This promising insight motivates future research into the development and validation of ambiguity-aware AI systems capable of providing highly relevant ambiguity explanations for previously unseen cases.

Experts' overall preferences and perceived levels of trust for either AI were polarized. Results suggested higher perceived integrity, and a trend towards higher confidence in the ambiguity-aware AI assistant compared to the conventional variant. These mixed results may indicate the existence of other latent variables (e.g., experts' familiarity with or trust in automation technology) which could shape experts' perception of AI systems generally. Here, we discuss the generalizability and design implications of our findings and conclude with limitations of our study and directions for future work.

5.6.1 Design Implications

Our findings have implications for different stages in the design of AI-based CDS systems, ranging from data collection over model training to the design of user interfaces for AI systems.

Data collection. In our work, we simulate an AI assistant's capability to identify multiple conflicting arguments for why a medical classification decision may be contentious. To this end, we rely on discussion metadata from our previous study on collective adjudication among medical experts reported in Section 3.3. Developing an AI system capable of generating ambiguity explanations for previously *unseen* cases would require that structured information on contentious cases is given in the training data. While several approaches have been suggested to collect unstructured, open-ended arguments for contentious classification cases [29, 42, 125, 129], recent work from the medical domain demonstrates that imposing structure on the discussion process can facilitate a deeper understanding of expert disagreement [123] and accelerate consensus formation [127]. We recommend that data

collection procedures for AI-based CDS systems be equipped with structured discussion procedures to benefit from these findings and facilitate the development of ambiguity-aware classification models.

Model training. Our study suggests that expert workflows and trust can be positively affected by endowing AI-based CDS systems with the ability to not only make classification suggestions, but also to identify which cases may be contentious and why. Implementation of such systems would require that supervised machine learning models are equipped with additional prediction targets beyond classification labels alone. These additional prediction targets could include the likelihood and potential sources of expert disagreement. They could be integrated either into one joint training process or by developing several separate models, one for each target. Cohen et al. [32] describe some additional requirements and challenges in this context.

User interfaces. In this work, we evaluate one specific way of displaying and explaining ambiguity to expert end users by visually emphasizing contentious cases within a collection of cases and by providing text-based arguments for conflicting classification choices. While our results suggest that this representation may be effective, we recommend that future work may explore more complex design considerations such as prioritization of cases based on their disagreement likelihood, and interactive filters to group cases which may be contentious for similar reasons.

5.6.2 Generalizability

Our study sheds light on the use of ambiguity-awareness in the specific domain of sleep stage classification based on biomedical time series data. Therefore, caution is warranted in generalizing the results of this study to outside domains. However, we argue that similar displays of ambiguity explanations can be useful for various types of medical assessments because the issues motivating our study are prevalent across subspecialties.

Despite the abundance of standardized medical guidelines [11], expert disagreement is prevalent across medical disciplines [15, 128], making our approach useful beyond the specific domain of sleep health. For example, differential diagnosis of epilepsy requires that specialized neurologists visually inspect EEG data similar in nature to that used in our study. Ambiguity-aware AI assistants could support the small pool of specialists worldwide in detecting epileptiform abnormalities [10] and thus increase access to healthcare for patients with epilepsy in low- and middle-income countries [149, 150].

The issue of expert disagreement in medical assessments has also been addressed using structured adjudication for other data modalities, e.g., assessment of retinal images for

diabetic retinopathy grading [126, 127] or glaucoma risk assessment [65, 108]. These studies suggest that the recommendations we make for data collection in this work have been considered independently and may be of merit beyond the development of ambiguity-aware AI systems.

5.6.3 Limitations

In this chapter, we conducted a within-subjects study to investigate the use of ambiguity-aware AI assistants by medical experts. Due to the tight working schedule of our experts and the remote nature of our study, it was challenging to control the timing of each step in the experiment precisely. For instance, participants varied in how long they waited after completing the first main task before starting the second one. This lack in experimental control may have impacted the extent to which exposure to the first AI assistant affected how experts interacted with the latter one.

In our Wizard-of-Oz study, the ambiguity-aware AI was *simulated*, in the sense that the assistant presented ambiguity information and arguments generated from real expert discussions. While prior work has demonstrated the potential of predicting the likelihood of expert disagreement directly from raw medical data [114], future work can focus on training machine-learning algorithms based on ambiguity explanation data to provide human-interpretable arguments for previously unseen contentious cases.

Finally, related work shows that medical practitioners seek to understand the specific strengths and weaknesses of an AI *before* interacting with it [22]. Our work offers similar findings by showing that explaining AI uncertainty can be useful also *during* the interaction and help experts allocate cognitive resources and reassess their level of trust appropriately for each specific case. While we did not detect a significant effect of ambiguity tolerance on overall AI preference, we observed a trend that experts with higher ambiguity tolerance exhibited more polarized preferences towards either AI assistant. Future research may explore how different variables such as personality traits [20], domain-specific and culture-specific communication styles [121] may shape these expectations and perceptions on the side of medical experts.

5.7 Conclusion

In this chapter, we provided a novel perspective on the problem of how AI assistants for medical reasoning can explain ambiguous cases to human experts. Our results from a user

study with twelve medical experts comparing a conventional AI assistant to a simulated ambiguity-aware AI assistant suggest that the system's ability to not only flag, but also explain contentious patient cases has merits for end users. In particular, we observed that in comparison to the conventional AI, the ambiguity-aware AI was more effective in guiding experts' attention to contentious medical cases. In addition, our results demonstrate that if explanations contain irrelevant arguments, experts' accuracy at correcting AI-suggested labels can drop below 50%. The work we presented in this chapter has implications for the design of AI-based technology not only in the field of medicine, but more broadly in fields that face similar challenges with classification ambiguity and expert disagreement.

Chapter 6

Conclusion

Through the research presented in this dissertation, we introduced and studied novel approaches for handling ambiguity in human-AI collaborative workflows for data classification problems. We provided methods, open datasets and empirical insights to address ambiguity in various steps of the AI pipeline. Most importantly, we provided a thorough analysis of group deliberation as a tool to understand and capture the structure of ambiguous cases in data labeling. We also showed how the resulting deliberation data can be leveraged to improve outcomes both in the training of human labelers and for communicating classification ambiguity in AI-powered assistive interfaces.

In this chapter, we summarize our main contributions, specify how our findings support the thesis statement, provide design recommendations for handling ambiguity in human-AI collaborative workflows and conclude with several directions for future research in the broader space of human-AI interaction in the context of complex, ill-defined and ambiguous problems.

6.1 Contributions and Impact

We implemented and studied group deliberation as a tool to detect and explain ambiguity in data labeling workflows. Our investigation focused on three different contexts.

First, we presented *Crowd Deliberation*, the first platform enabling synchronous group deliberation in the context of non-expert crowdsourcing. We presented evidence suggesting that group deliberation can not only significantly improve label accuracy, but that

it also produces data that helps to understand why disagreement arises and under what circumstances it can be resolved.

Second, we applied our findings to the *expert domain* of medical image classification, applying an asynchronous deliberation workflow to a complex diagnostic task performed by medical specialist labelers. Our insights from an experiment with 15 retina specialists showed that structuring deliberation arguments around a set of low-level diagnostic criteria significantly improved the efficiency of the deliberation process without compromising its reliability.

Third, we leveraged our findings to design and build *CrowdEEG*, the first online platform enabling expert crowds to collaboratively annotate and deliberate on medical time series data. CrowdEEG implemented an asynchronous workflow combined with an even more structured format for capturing human arguments. Using an observational study with 36 sleep technologists, we analyzed various factors including expert background, data characteristics, labeling guidelines and viewer configurations to better understand how disagreement arises and when it can be resolved.

We also demonstrated how the resulting deliberation data can be put into use to *train human expert labelers*. Our evidence from a controlled experiment with ten medical generalists suggests that reading deliberation data from medical specialists substantially improved generalists' comprehension as well as their diagnostic accuracy on difficult patient cases.

Finally, we leveraged deliberation data for a separate goal, to simulate and *study ambiguity-aware AI*, i.e., AI that not only highlights ambiguous cases, but also explains the underlying sources of ambiguity to end users. Our results from an experiment with twelve sleep technologists demonstrated that this form of ambiguity-aware AI can significantly improve the ability of expert end users to triage and trust AI-provided output. We also provided evidence suggesting not only that expert end users paid attention to AI-provided ambiguity explanations, but also that the relevance of these explanations to the specific case at hand was crucial for human experts to accurately classify difficult ambiguous cases.

To stimulate future research in this space, we made *two novel datasets* publicly available that contain deliberation data from both non-expert and expert classification tasks.

6.2 Support for Thesis Statement

We summarize our hypotheses from Chapters 3, 4 and 5 in Tables 6.1 and 6.2, including the degree to which each hypothesis was supported by our observations. Collectively, our insights provide support for the central claims of our thesis statement.

> **Claim:** *Ambiguity, the quality of being open to more than one interpretation, permeates our lives. It can take various forms including linguistic and visual ambiguity, arise for various reasons including heterogeneous data or vague definitions, and give rise to inter-rater disagreements that can be hard or impossible to resolve.*

We demonstrated that ambiguity in data classification can take various forms. *Linguistic ambiguity* was prevalent not only in the data instances to be classified, e.g., the text documents involved in the sarcasm detection and semantic relation verification tasks in Section 3.1. It could also be found in the classification guidelines used by humans to interpret the data at hand, irrespective of data modality. *Visual ambiguity* was showcased in the context of both image classification (Section 3.2) and time series classification (Section 3.3) tasks.

We provided analyses of the various sources of classification ambiguity: Section 3.1 demonstrated that fuzzy definitions in classification guidelines, discrepancies in labeler expertise, subjectivity, missing context, contradictory evidence or easy-to-miss details in the data can constitute relevant sources of inter-rater disagreements in *novice crowd work*. Section 3.3 provided similar insights in the context of *expert tasks*, assessing potential sources of inter-rater disagreement within four categories: grader differences, viewer differences, data characteristics, classification guidelines. Chapter 3 described three separate case studies that provide support for the claim that certain instances of classification ambiguity give rise to inter-rater disagreement that is hard or impossible to resolve even when addressed through group deliberation.

> **Claim:** *Human and artificial intelligence can benefit from novel methods that aim to detect and explain instances of ambiguity. The expected advantages of such methods are a better understanding of why disagreement arises and when it can be resolved, as well as better approaches for handling ambiguity in human decision making—both unassisted and when assisted by artificial intelligence.*

Chapter 3, specifically Sections 3.1 and 3.3 demonstrate how *group deliberation* is an effective tool to understand why disagreement arises and under which circumstances it can be resolved in the context of data classification.

Chapters 4 and 5 illustrate how the resulting deliberation data can be leveraged to support better approaches for handling ambiguity in human decision making. The case study described in Chapter 4 serves as an example of how deliberation data can be used as training material to improve *unassisted* human decision making on difficult cases in the context of medical image classification. Chapter 5 showcased how deliberation data

can be used to support human decision making *assisted by an AI* capable of not only highlighting ambiguous classifications, but also explaining potential sources of ambiguity in human-interpretable terms.

6.3 Design Recommendations

This dissertation aims to provide guidance for both researchers and practitioners who seek to incorporate the notion of ambiguity into the design and implementation of human-AI collaborative systems. Based on the insights from our studies on group deliberation in data labeling and experiments leveraging deliberation data for labeler training and for ambiguity-aware AI, we make the following recommendations for handling ambiguity within various steps of the AI pipeline.

Data Labeling: Procedures for data labeling should aim to identify and differentiate between instances of ambiguity that are unintended and avoidable, and instances of ambiguity that are either inevitable or an intentional part of the concept being labeled. The recommendations below should be applied with this distinction in mind.

Labeling guidelines are a common source of linguistic ambiguity giving rise to inter-rater disagreement. Unintentional vagueness in the labeling guidelines should be systematically analyzed and removed. However, it is possible that the labeling guidelines are standardized and may not be changed.

Inter-rater disagreement is a useful signal to identify potential instances of ambiguity. However, label disagreements may also be due to other reasons such as human input mistakes. The circumstance that human labelers can guess the likelihood of disagreement for some tasks better than chance may be used to allocate resources for label redundancy more efficiently.

Group deliberation is a useful tool to decide which instances of inter-rater disagreement are hard or impossible to resolve and why. Deliberation therefore can be used to identify truly ambiguous cases as well as the underlying sources of persistent disagreement. Deliberation data may in turn be used to identify and reduce unintentional vagueness in labeling guidelines.

Groupthink dynamics should be mitigated by hiding the true identity and professional credentials (e.g., academic degree) of discussion members, by encouraging balanced contribution among the group and by avoiding explicit incentives for reaching unanimous consensus.

Table 6.1: Summary of hypotheses and their degree of support from Chapter 3.

3.1 Crowd Deliberation for Text Labeling		
	Q1: Why disagree?	
H1a	Sources of disagreement differ by task type.	Supported (***)
H1b	Annotators can predict disagreement levels.	Supported for Relation task (***)
	Q2: Why unresolved?	
H2a	Sources of disagreement affect resolvability.	Supported for Subjective Case (**), Fuzzy Definition (**) and Contradictory Evidence (*)
H2b	Task subjectivity affects resolvability.	Partially supported
H2c	Extent of equal contribution affects resolvability.	Supported (*)
H2d	Level of initial consensus affects resolvability.	Supported (*)
H2e	Amount of overlap in evidence affects resolvability.	Not supported
	Q3: Impact?	
H3a	Worker deliberation improves answer correctness.	Supported (***)
H3b	Groupthink is discouraged by our deliberation incentives.	Supported
H3c	Sources of disagreement and the extent of equal contribution affect whether cases get resolved correctly.	Supported for Expertise Needed (*), Missing Context (*) and the extent of equal contribution (*)
3.2 Expert Deliberation for Image Labeling		
H1	Remote and in-person deliberation produce similar labels.	Supported
H2	Remote deliberation is a reproducible process.	Supported
H3	Imposing argument structure helps resolve disagreements more efficiently in remote deliberation.	Supported (***)
3.3 Expert Deliberation for Time Series Labeling		
	Q1: Why disagree?	
H1a	Differences in expert background predict disagreement.	Supported for Experience (***) and Location (***)
H1b	Differences in viewer settings predict disagreement.	Supported (***)
H1c	Data characteristics predict disagreement.	Supported for Parkinson's Disease (***), Alzheimer's Disease (***) and Signal Complexity (***)
	Q2: Why unresolved?	
H2a	Differences in expert background predict resolvability.	Supported (*)
H2b	Differences in viewer settings predict resolvability.	Supported for Freq. Filter (***) and Signal Visib. (**)
H2c	Data characteristics predict resolvability.	Supported for Parkinson's Disease (***), Sleep Apnea (**), Sig. Complexity (***), Sig. Transitions (***)
H2d	Disagreement sources predict resolvability most strongly.	Partially supported
	Q3: Impact?	
H3a	Experts find deliberation labels reliable and trustworthy.	Supported (***)
H3b	Deliberation can significantly change diagnostic markers.	Supported for %REM (*)

Table 6.2: Summary of hypotheses and their degree of support from Chapters 4 and 5.

4 Deliberation Data for Labeler Training		
	Q1: Perception - Reading specialist discussions ...	
H1a	Improves generalists' comprehension of the correct diagnosis.	Supported (***)
H1b	Increases generalists' agreement with the answer key.	Not supported
H1c	Motivates adaptations in generalists' labeling approach.	Supported (*)
	Q2: Behaviour - Reading specialist discussions ...	
H2a	Improves generalists' diagnostic accuracy.	Supported for retinal artery occlusion (*)
H2b	Increases generalists' diagnostic confidence.	Opposite association found (**)
H2c	Lowers generalists' perceived case difficulty.	Opposite association found (*)
H2d	Improves generalists' diagnostic self-efficacy.	Partially supported
5 Deliberation Data for Ambiguity-aware AI		
	Q1: Behaviour?	
H1a	Experts review more contentious cases with an ambiguity-aware AI.	Supported (*)
H1b	Ambiguity-aware AI does not reduce experts' overall efficiency.	Supported
H1c	Ambiguity-aware AI improves experts' overall label accuracy.	Not supported
H1d	Relevance of ambiguity explanations predicts expert accuracy.	Supported (***)
	Q2: Perception?	
H2a	Experts generally prefer an ambiguity-aware AI.	Not supported
H2b	Experts consider an ambiguity-aware AI more trustworthy.	Supported for Integrity (*)
H2c	Ambiguity-aware AI does not inrease cognitive load.	Supported
H2d	Ambiguity tolerance predicts preference for an ambiguity-aware AI.	Not supported

Asynchronous workflows for group deliberation are suitable for expert tasks as they allow potentially busy labelers to complete their review activities on their own schedule. For fast-paced crowdsourcing marketplaces where human labelers may only be available for a very limited amount of time, synchronous (real-time) deliberation workflows may be more effective.

Structuring arguments during deliberation procedures does not only help resolve disagreements more efficiently, but can also produce useful metadata that can be parsed by machines to understand why disagreements arise and why disagreements persist after group deliberation.

Sources of ambiguity are diverse and can differ by task type. Human labelers may arrive at conflicting interpretations for a variety of reasons, including differences in personal background, differences in viewer settings, characteristics of the data at hand and vagueness

in the labeling guidelines. For factors that can be controlled (e.g., viewer settings) it may be desirable to keep those consistent for all labelers or to store information about the specific configurations labelers make.

Training of human labelers can be enriched with discussion data from group deliberation processes. In particular, less experienced human labelers can improve their comprehension, labeling style and accuracy for difficult-to-classify cases by reading deliberation discussions from more experienced human labelers.

Model Development: While this dissertation does not make explicit contributions to the space of model development, we propose high-level recommendations for future modeling approaches seeking to incorporate the notion of ambiguity.

Uncertainty and ambiguity are not identical. Uncertainty is characterized by a lack of information needed to make accurate judgements about the current or future state of the world. Ambiguity on the other hand is characterized by an inherent openness to multiple interpretations about the same phenomenon. Both uncertainty and ambiguity should be incorporated into model development. Yet, we recommend they be treated as separate concepts.

The *likelihood of ambiguity* may be treated as a prediction target in model development. However, ambiguous examples may be rare compared to non-ambiguous examples in a given training set. We therefore recommend methods to address potential class imbalance when modeling the likelihood of ambiguity.

Sources of ambiguity may be extracted from deliberation data implementing a structured argument format. These may be treated as auxiliary prediction targets for models with strong explainability requirements.

User Interfaces: Communication between human users and AI models is both facilitated and constrained by UI design considerations. Designers should carefully decide whether and how ambiguity-related information is exposed to end users to strike the right balance for an efficient, effective and trusted interaction.

User trust can be promoted by highlighting and explaining instances of ambiguity as predicted by an AI model. However, factors like user-specific or context-specific ambiguity tolerance may affect whether exposing ambiguity information contributes to the establishment or erosion of user trust.

Triaging is the process of prioritizing and allocating human intervention in resource-limited settings. The predicted likelihood of ambiguity for a given case can be a useful

criterion for experts to choose which AI-suggestions to review first and thus save time and cognitive resources on clear-cut cases.

Ambiguity explanations should be human-interpretable, case-specific and accurate. An irrelevant or inaccurate ambiguity explanation may significantly harm a user's ability to disambiguate a case correctly. If accurate ambiguity explanations cannot be reliably produced they should not be exposed to the end user.

The *impact of ambiguity* is a factor that should determine whether users are exposed to ambiguity information. For a resourceful and efficient interaction, we recommend that users should not be exposed to ambiguity information for cases where different interpretations cannot yield different decision outcomes. For example, if different interpretations of a medical image will not affect the final treatment decision for a given patient, doctors should not be informed about the underlying ambiguity to be economical with their cognitive resources.

6.4 Opportunities for Future Work

In the previous three chapters, we discussed design implications, research limitations and directions for future with a specific focus on the research scope for each particular study. In this section, we broaden the scope and address additional directions for future research in the space of human-AI collaborative decision-making for complex, ill-defined and ambiguous problems.

Systems to label data of variable ambiguity: So far, the research presented in Chapters 3 and 4 has produced individual building blocks within the bigger picture of how interactive labeling can be done in the presence of ambiguous data. The future holds the potential to connect these separate components into one integrated system capable of intelligent adaptation when confronted with data of variable ambiguity. I am excited to pursue this research direction centered around the following questions: How can we leverage group deliberation to fuel automatic creation and refinement of labeling guidelines? How to evaluate labeler quality in the presence of ambiguity? How to provide personalized feedback to labelers by triangulating error types, labeling guidelines and group deliberation? Can we synthesize ambiguous cases as training material for human labelers? How can we close the loop by leveraging feedback from end users of an AI system to refine the labeling process?

Building ambiguity-aware AI assistants: The research presented in Chapter 5 has paved the way for understanding the potential benefit of communicating ambiguity in AI

output. Yet, these preliminary contributions relied on a Wizard-of-Oz approach leveraging a *simulated* version of an ambiguity-aware AI assistant. Future research holds the potential to explore the open question of how ambiguity-aware AI can and should be implemented.

One possible pathway towards implementation of ambiguity-aware AI would be to combine techniques from multi-label learning, label distribution learning and model explainability to develop AI assistants capable of not only recognizing, but also explaining ambiguous data. The large corpus of structured deliberation data for over 15,000 cases that we collected within the CrowdEEG platform may lend itself as training input for the purpose of building an initial prototype. However, the availability of ambiguity-related data alone may not be sufficient for building ambiguity-aware AI assistants. The research community may also need to develop novel evaluation metrics that allow us to measure the quality of labels and explanations, produced either by humans or machines, in the presence of inherently ambiguous problems.

Besides the modeling aspect of this broader question, more research is needed to understand when and how AI assistants should communicate ambiguity to end users: Should an AI selectively communicate ambiguous cases, e.g., only if they have the potential to impact meaningful outcomes like medical treatment decisions? Should systems adapt to user characteristics like ambiguity tolerance? What are effective formats to explain ambiguity for different data modalities?

Supporting other forms of problem complexity: The research presented in this dissertation intentionally keeps a narrow focus on ambiguous classification problems. However, there exist many other reasons why humans may face complexity in decision making and interpreting data. Recognizing this circumstance, future research may aim to enhance the people-AI partnership for problems that are complex for reasons other than ambiguity. For example, problems may be complex because they require the ability to make associations that are poorly understood by the current state of science. Arguably, one of the most complex domains in our modern society is the advancement of science. The progress of scientific discovery requires creative exploration and rigorous testing of novel hypotheses. Diversity of perspectives and scientific disagreements are integral to this process.

In recent years, AI systems have acquired the somewhat surprising ability to make complex data associations previously unknown or thought impossible within scientific communities. For example, deep neural networks can predict conditions like anemia or the risk of future heart attacks from just a single photo of a patient's retina. Yet, the immediate benefit of these models to the process of scientific discovery remains limited due the black-box nature of the underlying models. We posit that there is value in exploring methods from Explainable AI (XAI) to foster diversity in scientific reasoning, exploring the following

research directions: How can XAI become part of the modern scientific toolbox? How can XAI help scientists diversify their process of hypothesis generation? How can we support productive scientific discourse through XAI-powered collaborative thinking?

6.5 Summary

In this chapter, we summarized our main contributions, outlined how our findings support the central claims of our thesis statement, provided design recommendations for handling ambiguity in human-AI collaborative workflows and concluded with several directions for future work in the space of human-AI interaction for complex, ill-defined and ambiguous problems. The research presented in this work provides a novel perspective on the question of how humans and AI can be effective partners in the presence of ambiguous problems. We have provided a foundation and proposed directions we believe to be useful to enable future research in this space.

References

[1] Grading diabetic retinopathy from stereoscopic color fundus photographs–an extension of the modified Airlie House classification. ETDRS report number 10. Early Treatment Diabetic Retinopathy Study Research Group. *Ophthalmology*, 98(5 Suppl):786–806, 5 1991.

[2] Diabetic retinopathy PPP 2014: standard photographs 2A, 6A, 8A, 2014.

[3] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–18, New York, New York, USA, 2018. ACM Press.

[4] Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. *Investigative ophthalmology & visual science*, 57(13):5200–5206, 10 2016.

[5] Alaa Al Ali, Stephen Hallingham, and Yvonne M. Buys. Workforce supply of eye care providers in Canada: optometrists, ophthalmologists, and subspecialty ophthalmologists. *Canadian Journal of Ophthalmology*, 50(6):422–428, 12 2015.

[6] American Academy of Ophthalmology. International Clinical Diabetic Retinopathy Disease Severity Scale, Detailed Table, 2010.

[7] American Academy of Ophthalmology. Diabetic retinopathy PPP - updated 2017, 2017.

[8] Paul André, Aniket Kittur, and Steven P Dow. Crowd synthesis: Extracting categories and clusters from complex data. In *Proceedings of the 17th ACM conference*

on Computer supported cooperative work & social computing, pages 989–998. ACM, 2014.

[9] Lora Aroyo and Chris Welty. The three sides of crowdtruth. *Journal of Human Computation*, 1:31–34, 2014.

[10] Elham Bagheri, Justin Dauwels, Brian C. Dean, Chad G. Waters, M. Brandon Westover, and Jonathan J. Halford. Interictal epileptiform discharge characteristics underlying expert interrater agreement. *Clinical Neurophysiology*, 128(10):1994–2005, 10 2017.

[11] A. Baker, K. Young, J. Potter, and I. Madan. A review of grading systems for evidence-based guidelines produced by medical specialties. *Clinical Medicine*, 10(4):358–363, 8 2010.

[12] Erin P. Balogh, Bryan T. Miller, and John R. Ball, editors. *Improving Diagnosis in Health Care*. National Academies Press, Washington, D.C., 12 2015.

[13] Forrest S Bao, Xin Liu, and Christina Zhang. PyEEG: An Open Source Python Module for EEG/MEG Feature Extraction. *Computational Intelligence and Neuroscience*, 2011:1–7, 2011.

[14] Irene A Barbazetto. Diabetic Retinopathy: The Masqueraders. *Retinal Physician*, 7(6), 2010.

[15] Michael L. Barnett, Dhruv Boddupalli, Shantanu Nundy, and David W. Bates. Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians. *JAMA Network Open*, 2(3):e190096, 3 2019.

[16] Andrew Bastawrous and Benjamin D Hennig. The global inverse care law: a distorted map of blindness. *British Journal of Ophthalmology*, 96(10):2–1358, 10 2012.

[17] Abdhish R Bhavsar. Diabetic retinopathy differential diagnoses, 2019.

[18] Christopher R. Bilder and Thomas M. Loughin. Testing for Marginal Independence between Two Categorical Variables with Multiple Responses. *Biometrics*, 60(1):241–248, 3 2004.

[19] Brian H. Bornstein and A. Christine Emler. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *Journal of Evaluation in Clinical Practice*, 7(2):97–107, 5 2001.

[20] Stanley Budner. Intolerance of ambiguity as a personality variable. *Journal of Personality*, 30(1):29–50, 3 1962.

[21] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making. Number 45, 2 2019.

[22] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 11 2019.

[23] Arthur Carvalho and Kate Larson. A Consensual Linear Opinion Pool. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 2518–2524, Beijing, China, 2013. AAAI Press.

[24] Anthony A. Cavallerano and Paul R. Conlin. Teleretinal Imaging to Screen for Diabetic Retinopathy in the Veterans Health Administration. *Journal of Diabetes Science and Technology*, 2(1):33–39, 1 2008.

[25] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, pages 2334–2346, New York, New York, USA, 2017. ACM, ACM Press.

[26] Nancy Chang, Praveen Paritosh, David Huynh, and Collin Baker. Scaling semantic frame annotation. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 1–10, 2015.

[27] Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R Aragon. Using Machine Learning to Support Qualitative Coding in Social Science: Shifting The Focus to Ambiguity. *ACM Transactions on Interactive Intelligent Systems*, (Human-Centered Machine Learning), 2018.

[28] Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. How to develop machine learning models for healthcare. *Nature Materials*, 18(5):410–414, 5 2019.

[29] Quanze Chen, Jonathan Bragg, Lydia B. Chilton, and Daniel S. Weld. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing. In *Proceedings of the*

2019 CHI Conference on Human Factors in Computing Systems - CHI '19, pages 1–14, New York, New York, USA, 10 2019. ACM Press.

[30] L M Chihara and T C Hesterberg. *Mathematical Statistics with Resampling and R*. Wiley, 2018.

[31] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 4 1960.

[32] Robin Cohen, Mike Schaekermann, Sihao Liu, and Michael Cormier. 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, pages 1644–1648, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems.

[33] David A. Cook, Jonathan Sherbino, and Steven J. Durning. Management Reasoning - Beyond the Diagnosis. *JAMA*, 319(22):2267, 6 2018.

[34] Jorge Cuadros and George Bresnick. EyePACS: An Adaptable Telemedicine System for Diabetic Retinopathy Screening. *Journal of Diabetes Science and Technology*, 3(3):509–516, 5 2009.

[35] Norman Dalkey and Olaf Helmer. An Experimental Application of the DELPHI Method to the Use of Experts. *Management Science*, 9(3):458–467, 4 1963.

[36] Heidi Danker-Hopfe, Peter Anderer, Josef Zeitlhofer, Marion Boeck, Hans Dorn, Georg Gruber, Esther Heller, Erna Loretz, Doris Moser, Silvia Parapatics, Bernd Saletu, Andrea Schmidt, and Georg Dorffner. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of Sleep Research*, 18(1):74–84, 3 2009.

[37] Todd Davies and Reid Chandler. Online deliberation design. *Democracy in motion: Evaluation the practice and impact of deliberative civic engagement*, pages 103–131, 2012.

[38] Jasmin Diwan, Chinmay Shah, Saurin Sanghavi, and Amit Shah. Comparison of case-based learning and traditional lectures in physiology among first year undergraduate medical students. *National Journal of Physiology, Pharmacy and Pharmacology*, page 1, 2017.

[39] Jeff Donahue and Kristen Grauman. Annotator rationales for visual recognition. In *2011 International Conference on Computer Vision*, pages 1395–1402. IEEE, 11 2011.

[40] Tim Dornan, Albert Scherpbier, Nigel King, and Henny Boshuizen. Clinical teachers and problem-based learning: a phenomenological study. *Medical Education*, 39(2):163–170, 2 2005.

[41] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. Toward a Learning Science for Complex Crowdsourcing Tasks. In *Proceedings of the 2016 SIGCHI Conference on Human Factors in Computing Systems - CHI '16*, pages 2623–2634, New York, New York, USA, 2016. ACM Press.

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

[43] Stephan Dreiseitl and Michael Binder. Do physicians value decision support? A look at the effect of decision support systems on physician opinion. *Artificial Intelligence in Medicine*, 33(1):25–30, 1 2005.

[44] Anca Dumitrache, Lora Aroyo, and Chris Welty. Crowdsourcing Ground Truth for Medical Relation Extraction. *ACM Transactions on Interactive Intelligent Systems*, 8(2):1–20, 7 2018.

[45] David Dunning. The Dunning–Kruger Effect. pages 247–296. 2011.

[46] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI '19*, pages 263–274, New York, New York, USA, 2019. ACM Press.

[47] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2 2017.

[48] Elena Filatova. Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and*

Evaluation - LREC '12, pages 392–398. European Language Resources Association (ELRA), 2012.

[49] Alan D Fleming, Keith A Goatman, Sam Philip, Gordon J Prescott, Peter F Sharp, and John A Olson. Automated grading for diabetic retinopathy: a large-scale audit using arbitration by clinical experts. *The British journal of ophthalmology*, 94(12):1606–10, 12 2010.

[50] Deen G. Freelon, Travis Kriplean, Jonathan Morgan, W. Lance Bennett, and Alan Borning. Facilitating Diverse Political Engagement with the Living Voters Guide. *Journal of Information Technology & Politics*, 9(3):279–297, 7 2012.

[51] 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. Validation of Consensus Panel Diagnosis in Dementia. *Archives of Neurology*, 67(12), 12 2010.

[52] Snehalkumar (Neil) S. Gaikwad, Mark Whiting, Karolina Ziulkoski, Alipta Ballav, Aaron Gilbee, Senadhipathige S. Niranga, Vibhor Sehgal, Jasmine Lin, Leonardy Kristianto, Angela Richmond-Fuller, Jeff Regino, Durim Morina, Nalin Chhibber, Dinesh Majeti, Sachin Sharma, Kamila Mananova, Dinesh Dhakal, William Dai, Victoria Purynova, Samarth Sandeep, Varshine Chandrakanthan, Tejas Sarma, Adam Ginzberg, Sekandar Matin, Ahmed Nasser, Rohit Nistala, Alexander Stolzoff, Kristy Milland, Vinayak Mathur, Rajan Vaish, Michael S. Bernstein, Catherine Mullings, Shirish Goyal, Dilrukshi Gamage, Christopher Diemert, Mathias Burton, and Sharon Zhou. Boomerang: Rebounding the Consequences of Reputation Feedback on Crowdsourcing Platforms. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST '16*, pages 625–637, New York, New York, USA, 2016. ACM Press.

[53] Adrian Galdran, M. Meyer, P. Costa, MendonCa, and A. Campilho. Uncertainty-Aware Artery/Vein Classification on Retinal Images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 556–560. IEEE, 4 2019.

[54] Sapna Gangaputra, James F Lovato, Larry Hubbard, Matthew D Davis, Barbara A Esser, Walter T Ambrosius, Emily Y Chew, Craig Greven, Letitia H Perdue, Wai T Wong, Audree Condren, Charles P Wilkinson, Elvira Agrón, Sharon Adler, Ronald P

Danis, and ACCORD Eye Research Group. Comparison of standardized clinical classification with fundus photograph grading for the assessment of diabetic retinopathy and diabetic macular edema severity. *Retina (Philadelphia, Pa.)*, 33(7):1393–9, 2013.

[55] Luciana Garbayo. Epistemic Considerations on Expert Disagreement, Normative Justification, and Inconsistency Regarding Multi-criteria Decision Making. *Constraint Programming and Decision Making*, 539:35–45, 2014.

[56] Rishab Gargeya and Theodore Leng. Automated Identification of Diabetic Retinopathy Using Deep Learning. *Ophthalmology*, 124(7):962–969, 2017.

[57] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.

[58] Jeffrey Alan Golden. Deep Learning Algorithms for Detection of Lymph Node Metastases From Breast Cancer: Helping Artificial Intelligence Be Seen. *JAMA*, 318(22):2184–2186, 2017.

[59] Gowri Gopalakrishna, Miranda W Langendam, Rob JPM Scholten, Patrick MM Bossuyt, and Mariska MG Leeflang. Guidelines for guideline developers: a systematic review of grading systems for medical tests. *Implementation Science*, 8(1):78, 12 2013.

[60] Nitesh Goyal and Susan R Fussell. Effects of sensemaking translucence on distributed collaborative analysis. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 288–302. ACM, 2016.

[61] Cosima Gretton. Trust and Transparency in Machine Learning-Based Clinical Decision Support. pages 279–292. 2018.

[62] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who said what: Modeling individual labelers improves classification. In *AAAI Conference on Artificial Intelligence*, 2018.

[63] 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. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *Jama*, 304(6):649–656, 2016.

[64] Danna Gurari and Kristen Grauman. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, pages 3511–3522, New York, New York, USA, 2017. ACM, ACM Press.

[65] Naama Hammel, Mike Schaekermann, Sonia Phene, Carter Dunn, Lily Peng, Dale R Webster, and Rory Sayres. A Study of Feature-based Consensus Formation for Glaucoma Risk Assessment. *Investigative Ophthalmology & Visual Science*, 60(9):164, 2019.

[66] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. pages 139–183. 1988.

[67] Francis T. Hartman and Andrew Baldwin. Using Technology to Improve Delphi Method. *Journal of Computing in Civil Engineering*, 9(4):244–249, 10 1995.

[68] Mark Hartswood, Rob Procter, Paul Taylor, Lilian Blot, Stuart Anderson, Mark Rouncefield, and Roger Slack. 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*, page 909, New York, New York, USA, 2012. ACM Press.

[69] David W. Hosmer and Stanley Lemesbow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, 9(10):1043–1069, 1980.

[70] Conrad Iber, Sonia Ancoli-Israel, Andrew L Cheeson Jr., and Stuart F Quan. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. American Academy of Sleep Medicine, 2007.

[71] Oana Inel, Lora Aroyo, Chris Welty, and Robert-Jan Sips. Domain-Independent Quality Measures for Crowd Truth Disagreement. In *The 12th International Semantic Web Conference (ISWC2013)*, 2013.

[72] Hayley K. Jach and Luke D. Smillie. To fear or fly to the unknown: Tolerance for ambiguity and Big Five personality traits. *Journal of Research in Personality*, 79:67–78, 4 2019.

[73] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71, 3 2000.

[74] Alan M. Jones. Victims of Groupthink: A Psychological Study of Foreign Policy Decisions and Fiascoes. *The ANNALS of the American Academy of Political and Social Science*, 407(1):179–180, 5 1973.

[75] Samed Jukić and Jasmin Kevrić. Majority vote of ensemble machine learning methods for real-time epilepsy prediction applied on eeg pediatric data. *TEM Journal*, 7(2):313, 2018.

[76] Sanjay Kairam and Jeffrey Heer. Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*, pages 1635–1646, New York, New York, USA, 2016. ACM Press.

[77] 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. Plus disease in retinopathy of prematurity: improving diagnosis by ranking disease severity and using quantitative image analysis. *Ophthalmology*, 123(11):2345–2351, 11 2016.

[78] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. When (ish) is My Bus?: User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, pages 5092–5103, New York, New York, USA, 2016. ACM Press.

[79] Gideon Keren and Léonie E.M. Gerritsen. On the robustness and possible accounts of ambiguity aversion. *Acta Psychologica*, 103(1-2):149–172, 11 1999.

[80] Sara Kiesler and Lee Sproull. Group decision making and communication technology. *Organizational Behavior and Human Decision Processes*, 52(1):96–123, 6 1992.

[81] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pages 1–14, New York, New York, USA, 2019. ACM Press.

[82] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S. Corrado, Lily Peng, and Dale R. Webster. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology*, 3 2018.

[83] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. Integrating on-demand fact-checking with public dialogue. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, pages 1188–1199, New York, New York, USA, 2014. ACM Press.

[84] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*, page 265, New York, New York, USA, 2012. ACM Press.

[85] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Andrew Ko. Is this what you meant?: promoting listening on the web with reflect. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, page 1559, New York, New York, USA, 2012. ACM Press.

[86] Joseph D Kronz, Mark A Silberman, William C Allsbrook, and Jonathan I Epstein. A web-based tutorial improves practicing pathologists' Gleason grading of images of prostate carcinoma specimens obtained by needle biopsy. *Cancer*, 89(8):1818–1823, 10 2000.

[87] Helen K Li, Larry D Hubbard, Ronald P Danis, Adol Esquivel, Jose F Florez-Arango, Nicola J Ferrier, and Elizabeth A Krupinski. Digital versus film Fundus photography for research grading of diabetic retinopathy severity. *Investigative ophthalmology & visual science*, 51(11):5846–52, 11 2010.

[88] P R Lichter. Variability of expert observers in evaluating the optic disc. *Transactions of the American Ophthalmological Society*, 74:532–72, 1976.

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

[90] J.C Liston and B.J.G Dall. Can the NHS Breast Screening Programme Afford not to Double Read Screening Mammograms? *Clinical Radiology*, 58(6):474–477, 6 2003.

[91] Weichen Liu, Sijia Xiao, Jacob T Browne, Ming Yang, and Steven P Dow. ConsensUs: Supporting Multi-Criteria Group Decisions by Visualizing Points of Disagreement. *ACM Transactions on Social Computing*, 1(1):4:1–4:26, 1 2018.

[92] Vera P. Luther and Sonia J. Crandall. Commentary: Ambiguity and Uncertainty: Neglected Elements of Medical Education Curricula? *Academic Medicine*, 86(7):799–800, 7 2011.

[93] V K Chaithanya Manam and Alexander J Quinn. WingIt: Efficient Refinement of Unclear Task Instructions. In *The Sixth AAAI Conference on Human Computation and Crowdsourcing*, number HCOMP, pages 108–116, 2018.

[94] Peter McCullagh and John Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, 2 edition, 1989.

[95] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 11 2019.

[96] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pages 279–288, New York, New York, USA, 11 2019. ACM Press.

[97] Jeryl L. Mumpower and Thomas R. Stewart. Expert Judgement and Expert Disagreement. *Thinking & Reasoning*, 2(2-3):191–212, 7 1996.

[98] Joaquin Navajas, Tamara Niella, Gerry Garbulsky, Bahador Bahrami, and Mariano Sigman. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, 1 2018.

[99] Charlan Nemeth. Interactions Between Jurors as a Function of Majority vs. Unanimity Decision Rules. *Journal of Applied Social Psychology*, 7(1):38–56, 3 1977.

[100] Geoffrey R. Norman, Lawrence E. M. Grierson, Jonathan Sherbino, Stanley J. Hamstra, Henk G. Schmidt, and Silvia Mamede. Expertise in Medicine and Surgery. In *The Cambridge Handbook of Expertise and Expert Performance*, pages 331–355. Cambridge University Press, 2018.

[101] J. A. Osheroff, J. M. Teich, B. Middleton, E. B. Steen, A. Wright, and D. E. Detmer. A Roadmap for National Action on Clinical Decision Support. *Journal of the American Medical Informatics Association*, 14(2):141–145, 3 2007.

[102] Gerhard Osius and Dieter Rojek. Normal Goodness-of-Fit Tests for Multinomial Models with Large Degrees of Freedom. *Journal of the American Statistical Association*, 87(420):1145–1152, 12 1992.

[103] Susannah BF Paletz, Joel Chan, and Christian D Schunn. Uncovering uncertainty through disagreement. *Applied Cognitive Psychology*, 30(3):387–400, 2016.

[104] Shengying Pan, Kate Larson, Joshua Bradshaw, and Edith Law. Dynamic Task Allocation Algorithm for Hiring Workers that Learn. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*, pages 3825–3831, New York, 2016.

[105] Matthew P Pase, Jayandra J Himali, Natalie A Grima, Alexa S Beiser, Claudia L Satizabal, Hugo J Aparicio, Robert J Thomas, Daniel J Gottlieb, Sandford H Auerbach, and Sudha Seshadri. Sleep architecture and the risk of incident dementia in the community. *Neurology*, 89(12):1244–1250, 2017.

[106] Thomas Penzel, Xiaozhe Zhang, and Ingo Fietze. Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules. *Journal of Clinical Sleep Medicine*, 9(1):81–87, 2013.

[107] Anh T Pham, Raviv Raich, and Xiaoli Z Fern. Dynamic programming for instance annotation in multi-instance multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2381–2394, 2017.

[108] 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. Deep Learning and Glaucoma Specialists: The Relative Importance of Optic Disc Features to Predict Glaucoma Referral in Fundus Photographs. *Ophthalmology*, 9 2019.

[109] T Postmes and M Lea. Social processes and group decision making: anonymity in group decision support systems. *Ergonomics*, 43(8):1252–74, 8 2000.

[110] Ronald B Postuma, Alex Iranzo, Michele Hu, Birgit Högl, Bradley F Boeve, Raffaele Manni, Wolfgang H Oertel, Isabelle Arnulf, Luigi Ferini-Strambi, Monica Puligheddu,

and others. Risk and predictors of dementia and parkinsonism in idiopathic REM sleep behaviour disorder: a multicentre study. *Brain*, 142(3):744–759, 2019.

[111] Stefan Räbiger, Gizem Gezici, Yücel Saygın, and Myra Spiliopoulou. Predicting worker disagreement for more effective crowd labeling. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 179–188. IEEE, 2018.

[112] Stefan Räbiger, Gizem Gezici, Myra Splilliopoulou, and Yücel Sayg\in. Predicting worker disagreement for more effective crowd labeling. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE*, 2018.

[113] Emilee Rader, Kelley Cotter, and Janghee Cho. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–13, New York, New York, USA, 2018. ACM Press.

[114] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Robert Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct Uncertainty Prediction for Medical Second Opinions. 7 2018.

[115] Pranav Rajpurkar, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y. Ng. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. 7 2017.

[116] 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. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *npj Digital Medicine*, 2(1):25, 12 2019.

[117] D A Redelmeier and E Shafir. Medical decision making in situations that offer multiple alternatives. *JAMA*, 273(4):302–5, 1 1995.

[118] Richard S. Rosenberg and Steven van Hout. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *Journal of Clinical Sleep Medicine*, 1 2013.

[119] Paisan Ruamviboonsuk, Khemawan Teerasuwanajak, Montip Tiensuwan, Kanokwan Yuttitham, and Thai Screening for Diabetic Retinopathy Study Group. Interobserver agreement in the interpretation of single-field digital fundus images for diabetic retinopathy screening. *Ophthalmology*, 113(5):826–32, 5 2006.

[120] Harold Sackman. Delphi assessment: Expert opinion, forecasting, and group process. Technical report, RAND CORP SANTA MONICA CA, 1974.

[121] Elaheh Sanoubari, Stela H. Seo, Diljot Garcha, James E. Young, and Veronica Loureiro-Rodriguez. Good Robot Design or Machiavellian? An In-the-Wild Robot Leveraging Minimal Knowledge of Passersby's Culture. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 382–391. IEEE, 3 2019.

[122] 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. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology*, 126(4):552–564, 4 2019.

[123] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. Capturing Expert Arguments from Medical Adjudication Discussions in a Machine-readable Format. In *Companion Proceedings of The 2019 World Wide Web Conference - WWW '19*, volume 2, pages 1131–1137, New York, New York, USA, 2019. ACM Press.

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

[125] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 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)*, volume 2, pages 1–19, New York City, NY, 11 2018.

[126] Mike Schaekermann, Naama Hammel, Brian Basham, Bilson Campana, Edith Law, Lily Peng, Dale R Webster, and Rory Sayres. Asynchronous Remote Adjudication for Grading Diabetic Retinopathy. *Investigative Ophthalmology & Visual Science*, 60(9):158, 2019.

[127] 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. Remote Tool-Based Adjudication for Grading Diabetic Retinopathy. *Translational Vision Science & Technology*, 8(6):40, 12 2019.

[128] Mike Schaekermann, Edith Law, Kate Larson, and Andrew Lim. 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, 2018.

[129] Mike Schaekermann, Edith Law, Alex C Williams, and William Callaghan. Resolvable vs. Irresolvable Ambiguity: A New Hybrid Framework for Dealing with Uncertain Ground Truth. In *1st Workshop on Human-Centered Machine Learning at SIGCHI 2016*, San Jose, CA, 2016.

[130] Hanna Schneider, Julia Wayrauther, Mariam Hassib, and Andreas Butz. Communicating Uncertainty in Fertility Prognosis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pages 1–11, New York, New York, USA, 2019. ACM Press.

[131] Craig R. Scott. The impact of physical and discursive anonymity on group members' multiple identifications during computer-supported decision making. *Western Journal of Communication*, 63(4):456–487, 12 1999.

[132] Ingrid U Scott, Neil M Bressler, Susan B Bressler, David J Browning, Clement K Chan, Ronald P Danis, Matthew D Davis, Craig Kollman, Haijing Qin, and Diabetic Retinopathy Clinical Research Network Study Group. Agreement between clinician and reading center gradings of diabetic retinopathy severity level at baseline in a phase 2 study of intravitreal bevacizumab for diabetic macular edema. *Retina (Philadelphia, Pa.)*, 28(1):36–40, 1 2008.

[133] Manali Sharma, Di Zhuang, and Mustafa Bilgic. Active Learning with Rationales for Text Classification. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 2015.

[134] Lili Shi, Huiqun Wu, Jiancheng Dong, Kui Jiang, Xiting Lu, and Jian Shi. Telemedicine for detecting diabetic retinopathy: a systematic review and meta-analysis. *British Journal of Ophthalmology*, 99(6):823–831, 6 2015.

[135] Miriam Solomon. Groupthink versus The Wisdom of Crowds : The Social Epistemology of Deliberation and Dissent. *The Southern Journal of Philosophy*, 44(S1):28–42, 3 2006.

[136] Miriam Solomon. The social epistemology of NIH consensus conferences. In *Establishing medical reality*, pages 167–177. Springer, 2007.

[137] Malathi Srinivasan, Michael Wilkes, Frazier Stevenson, Thuan Nguyen, and Stuart Slavin. Comparing Problem-Based Learning with Case-Based Learning: Effects of a Major Curricular Shift at Two Institutions. *Academic Medicine*, 82(1):74–82, 1 2007.

[138] 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. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications*, 9(1):5229, 12 2018.

[139] Thérèse A. Stukel. Generalized Logistic Models. *Journal of the American Statistical Association*, 83(402):426–431, 6 1988.

[140] 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. 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):2211, 12 2017.

[141] Daniel Shu Wei Ting, Gemmy Chui Ming Cheung, and Tien Yin Wong. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public

health challenges: a review. *Clinical & experimental ophthalmology*, 44(4):260–77, 5 2016.

[142] Emanuele Trucco, Alfredo Ruggeri, Thomas Karnowski, Luca Giancardo, Edward Chaum, Jean Pierre Hubschman, Bashir Al-Diri, Carol Y Cheung, Damon Wong, Michael Abràmoff, Gilbert Lim, Dinesh Kumar, Philippe Burlina, Neil M Bressler, Herbert F Jelinek, Fabrice Meriaudeau, Gwénolé Quellec, Tom Macgillivray, and Bal Dhillon. Validating retinal fundus image analysis algorithms: issues and a proposal. *Investigative ophthalmology & visual science*, 54(5):3546–59, 5 2013.

[143] Evangelia Tsiga, Efharis Panagopoulou, Nick Sevdalis, Anthony Montgomery, and Alexios Benos. The influence of time pressure on adherence to guidelines in primary care: an experimental study. *BMJ Open*, 3(4):e002700, 4 2013.

[144] A. Tversky and D. Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, 9 1974.

[145] Anne Marthe van der Bles, Sander van der Linden, Alexandra L. J. Freeman, James Mitchell, Ana B. Galvao, Lisa Zaval, and David J. Spiegelhalter. Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, 6(5):181870, 5 2019.

[146] Michael Van Lent, William Fisher, and Michael Mancuso. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the national conference on artificial intelligence*, pages 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

[147] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pages 1–15, New York, New York, USA, 2019. ACM Press.

[148] Simon C Warby, Sabrina L Wendt, Peter Welinder, Emil G S Munk, Oscar Carrillo, Helge B D Sorensen, Poul Jennum, Paul E Peppard, Pietro Perona, and Emmanuel Mignot. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nature Methods*, 11(4):385–392, 2 2014.

[149] Jennifer Williams, Fodé Abass Cisse, Mike Schaekermann, Foksuna Sakadi, Nana Rahamatou Tassiou, Aissatou Kenda BAH, Abdoul Bachir Djibo Hamani, Andrew Lim, Edward C W Leung, Tadeu A Fantaneau, Tracey Milligan, Vidita Khatri, Daniel

Hoch, Manav Vyas, Alice Lam, Gladia Hotan, Joseph Cohen, Edith Law, and Farrah Mateen. Utilizing a wearable smartphone-based EEG for pediatric epilepsy patients in the resource poor environment of Guinea: A prospective study. *Neurology*, 92(15 Supplement), 2019.

[150] Jennifer A Williams, Fodé Abass Cisse, Mike Schaekermann, Foksouna Sakadi, Nana Rahamatou Tassiou, Gladia C. Hotan, Aissatou Kenda Bah, Abdoul Bachir Djibo Hamani, Andrew Lim, Edward C.W. Leung, Tadeu A. Fantaneanu, Tracey A. Milligan, Vidita Khatri, Daniel B. Hoch, Manav V. Vyas, Alice D. Lam, Joseph M. Cohen, Andre C. Vogel, Edith Law, and Farrah J. Mateen. Smartphone EEG and remote online interpretation for children with epilepsy in the Republic of Guinea: Quality, characteristics, and practice implications. *Seizure*, 71:93–99, 10 2019.

[151] Ainur Yessenalina, Yejin Choi, and Claire Cardie. Automatically Generating Annotator Rationales to Improve Sentiment Classification. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 336–341, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[152] Omar F. Zaidan, Jason Eisner, and Christine D. Piatko. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pages 260–267, 2007.

[153] Omar F. Zaidan, Jason Eisner, and Christine D. Piatko. Machine learning with annotator rationales to reduce annotation cost. In *Proceedings of the NIPS 2008 Workshop on Cost Sensitive Learning*, 2008.

[154] Amy X. Zhang, Lea Verou, and David Karger. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, pages 2082–2096, New York, New York, USA, 2017. ACM Press.

APPENDICES

Appendix A

Group Deliberation for Data Labeling

This appendix includes the complete set of questionnaires used in Chapter 3.

A.1 Crowd Deliberation for Text Labeling

This appendix includes the complete set of questionnaires used in Section 3.1.

A.1.1 Pre-study Questionnaire

1. How old are you? (multiple choice)
    - 18-25
    - 26-35
    - 36-45
    - 46-55
    - 56+
2. What is your gender? (multiple choice)
    - Female
    - Male
    - Other: _____

- Prefer not to say

3. Is English your first language? (multiple choice)
   - Yes
   - No
4. Is English your first language? (5-point Likert scale)
   - 1 - Very Poor
   - 2
   - 3
   - 4
   - 5 - Very Good

A.1.2 Per-case Questionnaire after Independent Classification

1. Do you think other people might choose a different answer than you did? (multiple choice)
   - I expect most people to agree with me. [end of survey]
   - I expect only about half of the people to agree with me.
   - I expect most people to disagree with me.
2. Why do you think other people might choose a different answer? (checkboxes)
   - Other people may have different definitions of [sarcasm / relation] in mind.
   - The text is ambiguous because of missing context (for example, [the identity of the product / some important information about the person or the place] is unknown).
   - The text contains some features that indicate [sarcasm / relation] is expressed and other features that indicate [absence of sarcasm / relation is not expressed].
   - The text contains relevant details other people could easily miss.
   - Someone with more experience or expertise may see or understand something about the text that I don't.

- This is a case where a person's answer would depend heavily on their personal preferences and taste.
- Other: _____

3. Please elaborate on your answer to the previous question, explaining why you think other people might choose a different answer: _____

4. If there were other people who chose a different answer than you did, do you think a group discussion would help to resolve the case? (multiple choice)

   - Yes, a group discussion would help to resolve the case.
   - No, a group discussion would not help to resolve the case.

A.1.3 Per-case Questionnaire after Discussion Round 2

1. Based on your deliberation, why do you think the other people in the group chose a different answer? (checkboxes)

   - [Free-form reason from previous survey if the participant submitted any]
   - Other people may have different definitions of [sarcasm / relation] in mind.
   - The text is ambiguous because of missing context (for example, [the identity of the product / some important information about the person or the place] is unknown).
   - The text contains some features that indicate [sarcasm / relation] is expressed and other features that indicate [absence of sarcasm / relation is not expressed].
   - The text contains relevant details other people could easily miss.
   - Someone with more experience or expertise may see or understand something about the text that I don't.
   - This is a case where a person's answer would depend heavily on their personal preferences and taste.
   - Other: _____

2. Please elaborate on your answer to the previous question (for example, if you changed your mind about the source of disagreement, please explain why): _____

A.1.4 Per-case Questionnaire after Viewing Final Decisions

1. Why do you think this case [could be / could not be fully] resolved? _____
2. Did somebody make you doubt your original answer? Why or why not? (multiple choice)
   - Yes: _____
   - No: _____
3. Did somebody make you change your original answer? Why or why not? (multiple choice)
   - Yes: _____
   - No: _____
4. Did you manage to convince someone to change their answer or confidence level? Why do you think you were able/unable to convince them? (multiple choice)
   - Yes: _____
   - No: _____
5. Describe how you feel about the deliberation process: _____
6. Describe how you feel about the deliberation outcome: _____

A.2 Expert Deliberation for Time Series Labeling

This appendix includes the complete set of questionnaires used in Section 3.3.

A.2.1 Pre-study Questionnaire

1. How old are you? (multiple choice)
   - 18-25
   - 26-35
   - 36-45

- 46-55
- 56+

2. What is your gender? (multiple choice)
    - Female
    - Male
    - Other: _____
    - Prefer not to say
3. Where are you located? (multiple choice)
    - Canada
    - United States
    - European Union
    - Other: _____
4. Do you have any professional or academic training in sleep staging? (multiple choice)
    - Yes
    - No
5. If yes, please specify which kind of training, degree or certificate you hold: _____
6. How long have you worked in sleep? (multiple choice)
    - Not at all
    - < 3 months
    - 3-6 months
    - 6-12 months
    - 1-2 years
    - 2-3 years
    - 3-5 years
    - 5-10 years
    - 10+ years

A.2.2 Post-study Questionnaire

1. What computer did you use to complete our study? Please list the brand/model, and year, if you know this information (e.g., laptop "Macbook" from 2013; or desktop "Dell" from 2015): _____

2. What web browser did you use when running our study? (e.g., Chrome, Firefox, Edge, Opera, etc.): _____

3. How much would you agree with the following statements? (each item answered on a 5-point Likert scale ranging from 1 "Strongly Disagree" to 5 "Strongly Agree")

    - "The adjudication process was useful for generating a reliable hypnogram."
    - "The final adjudicated hypnogram can be trusted more than the hypnogram from my first pass."

4. What could be improved about the adjudication interface? _____

5. What were the good parts about the adjudication interface? _____

6. What could be improved about the adjudication procedure? _____

7. What were the good parts about the adjudication procedure? _____

Appendix B

Deliberation Data for Labeler Training

This appendix includes the complete set of questionnaires used in Chapter 4.

B.1 Pre-study Questionnaire

1. How good do you think you are at grading diabetic retinopathy (DR)? (5-point Likert scale)
    - 1 - Not good at all
    - 2
    - 3
    - 4
    - 5 - Extremely good
2. How good do you think you are at detecting the presence or absence of hypertensive retinopathy (HTNR)? (5-point Likert scale)
    - 1 - Not good at all
    - 2
    - 3
    - 4
    - 5 - Extremely good

3. How good do you think you are at detecting the presence or absence of retinal vein occlusion (RVO)? (5-point Likert scale)
    - 1 - Not good at all
    - 2
    - 3
    - 4
    - 5 - Extremely good
4. How good do you think you are at detecting the presence or absence of retinal artery occlusion (RAO)? (5-point Likert scale)
    - 1 - Not good at all
    - 2
    - 3
    - 4
    - 5 - Extremely good
5. What is your training background? (multiple choice)
    - Optometrist
    - General Ophthalmologist
    - Retina Specialist
    - Other: _____
6. How many years has it been since you finished residency or since you graduated from optometry school? _____
7. Approximately how many DR images have you graded? (multiple choice)
    - less than 100
    - 100 to 500
    - more than 500
8. How many years of experience do you have grading diabetic retinopathy (DR) images? _____

B.2 Per-case Questionnaire after Training Feedback

1. Which of your responses differed from the answer key (if any)? (checkboxes)
    - Diabetic Retinopathy (DR)
    - Hypertensive Retinopathy (HTNR)
    - Retinal Vein Occlusion (RVO)
    - Retinal Artery Occlusion (RAO)
2. After reviewing the answer sheet for this case, do you understand the rationale behind the answer key and could explain it to one of your colleagues? (5-point Likert scale)
    - 1 - Strongly Disagree
    - 2
    - 3
    - 4
    - 5 - Strongly Agree
3. After reviewing the answer sheet for this case, do you agree with the answer key for all prompts? (multiple choice)
    - No, I strongly disagree with at least some of the answer key prompts.
    - No, but I think all the answers in the answer key are reasonable.
    - Yes, I would change all my answers to match the answer key.
4. Please explain in your own words why (or not) you agree with the answer key for all prompts: _____
5. After reviewing the answer sheet for this case, is there anything you would change about how you grade in the future? (multiple choice)
    - No
    - Yes
6. If yes, please explain what you would change and why. Otherwise, please explain why you would not change anything: _____
7. Is there anything else you are still confused about after reviewing the case details? _____

B.3 Post-study Questionnaire

1. How good do you think you are at grading diabetic retinopathy (DR)? (5-point Likert scale)
    - 1 - Not good at all
    - 2
    - 3
    - 4
    - 5 - Extremely good
2. How good do you think you are at detecting the presence or absence of hypertensive retinopathy (HTNR)? (5-point Likert scale)
    - 1 - Not good at all
    - 2
    - 3
    - 4
    - 5 - Extremely good
3. How good do you think you are at detecting the presence or absence of retinal vein occlusion (RVO)? (5-point Likert scale)
    - 1 - Not good at all
    - 2
    - 3
    - 4
    - 5 - Extremely good
4. How good do you think you are at detecting the presence or absence of retinal artery occlusion (RAO)? (5-point Likert scale)
    - 1 - Not good at all
    - 2
    - 3

- 4
- 5 - Extremely good

5. How mentally demanding was the overall task? (5-point Likert scale)
   - 1 - Not at all
   - 2
   - 3
   - 4
   - 5 - Extremely
6. How physically demanding was the overall task? (5-point Likert scale)
   - 1 - Not at all
   - 2
   - 3
   - 4
   - 5 - Extremely
7. How hurried or rushed was the pace of the overall task? (5-point Likert scale)
   - 1 - Not at all
   - 2
   - 3
   - 4
   - 5 - Extremely
8. How successful were you in accomplishing what you were asked to do? (5-point Likert scale)
   - 1 - Perfect
   - 2
   - 3
   - 4
   - 5 - Failure

9. How hard did you have to work to accomplish your level of performance? (5-point Likert scale)
    - 1 - Not at all
    - 2
    - 3
    - 4
    - 5 - Extremely
10. How insecure, discouraged, irritated, stressed, and annoyed were you? (5-point Likert scale)
    - 1 - Not at all
    - 2
    - 3
    - 4
    - 5 - Extremely
11. The personalized feedback provided in this study was useful overall. (5-point Likert scale)
    - 1 - Strongly Disagree
    - 2
    - 3
    - 4
    - 5 - Strongly Agree

Appendix C

Deliberation Data for Ambiguity-aware AI

This appendix includes the complete set of questionnaires used in Chapter 5.

C.1 Pre-study Questionnaire

1. How old are you? (multiple choice)
    - 18-25
    - 26-35
    - 36-45
    - 46-55
    - 56+
2. What is your gender? (multiple choice)
    - Female
    - Male
    - Other: _____
    - Prefer not to say
3. Where are you located? (multiple choice)

- Canada
- United States
- European Union
- Other: _____

4. Do you have any professional or academic training in sleep staging? (multiple choice)
    - Yes
    - No
5. If yes, please specify which kind of training, degree or certificate you hold: _____
6. How long have you worked in sleep? (multiple choice)
    - Not at all
    - < 3 months
    - 3-6 months
    - 6-12 months
    - 1-2 years
    - 2-3 years
    - 3-5 years
    - 5-10 years
    - 10+ years
7. How often do you think two independent sleep experts tend to disagree on the correct sleep stage for a given polysomnography (PSG) epoch? On average, I think, two independent experts will: (multiple choice)
    - disagree on less than 1% of the cases
    - disagree on 2% - 10% of the cases
    - disagree on 11% - 20% of the cases
    - disagree on 21% - 30% of the cases
    - disagree on 31% - 40% of the cases
    - disagree on 41% - 50% of the cases

- disagree on more than half of the cases

8. How good do you think you are at sleep stage classification? (5-point Likert scale)
   - 1 - Not good at all
   - 2
   - 3
   - 4
   - 5 - Extremely good
9. How strongly do you agree with the following statements? (each item answered on a 7-point Likert scale ranging from 1 "Strongly Disagree" to 5 "Strongly Agree")
   - "An expert who doesn't come up with a definite answer probably doesn't know too much."
   - "There is really no such things as a problem that can't be solved."
   - "People who insist upon a yes or no answer just don't know how complicated things really are."
   - "Many of our most important decisions are based on insufficient information."

C.2 Post-condition Questionnaire

1. How strongly do you agree with the following statements? (each item answered on a 5-point Likert scale ranging from 1 "Strongly Disagree" to 5 "Strongly Agree")
   - "The AI Assistant supported my interpretation of the sleep recording."
   - "The AI Assistant helped me think through different options to interpret the sleep recording and organize my thoughts."
   - "I would continue using the AI assistant in practice."
2. How strongly do you agree with the following statements? (each item answered on a 5-point Likert scale ranging from 1 "Strongly Disagree" to 5 "Strongly Agree")
   - "The AI Assistant was deceptive."
   - "The AI Assistant behaved in an underhanded manner."

- "I was suspicious of the AI Assistant's intent, action, or outputs."
- "I was wary of the AI Assistant."
- "The AI Assistant's actions will have a harmful or injurious outcome."
- "I was confident in the AI Assistant."
- "The AI Assistant provided security."
- "The AI Assistant had integrity."
- "The AI Assistant was dependable."
- "The AI Assistant was reliable."
- "I can trust the AI Assistant."
- "I am familiar with the AI Assistant."

3. Please tell us more about how your experience with this task: (each item answered on a 5-point Likert scale ranging from 1 "Not at all" to 5 "Extremely")
   - How mentally demanding was the task?
   - How physically demanding was the task?
   - How hurried or rushed was the pace of the task?
   - How successful were you at accomplishing what you were asked to do?
   - How hard did you have to work to accomplish your level of performance?
   - How insecure, discouraged, irritated, stressed and annoyed were you?
4. How did you decide which epochs to review and why? _____
5. Which information provided by the AI Assistant did you use to decide which epochs to review and why? _____
6. How did information about the AI Assistant's uncertainty affect your decision making? _____

C.3 Post-study Questionnaire

1. Please compare AI Assistant A (the one you interacted with first, right after the playground task) and AI Assistant B (the one you interacted with last): (each item

answered on a 7-point Likert scale ranging from 1 "Totally Assistant A", 2 "Much more Assistant A than B", 3 "Slightly more Assistant A than B", 4 "Neutral", etc. to 7 "Totally Assistant B")

- Which of the two AI Assistants was more reliable?
- Which of the two AI Assistants was more trustworthy?
- Which of the two AI Assistants was more capable?
- Which of the two AI Assistants did you prefer overall?

2. Is there anything else you would like to tell us before completing the study? _____