Understanding Expert Disagreement in Medical Data Analysis through Structured Adjudication

MIKE SCHAEKERMANN, University of Waterloo, Canada
GRAEME BEATON, University of Waterloo, Canada
MINAHZ HABIB, University of Toronto, Canada
ANDREW LIM, University of Toronto, Canada
KATE LARSON, University of Waterloo, Canada
EDITH LAW, University of Waterloo, Canada

Expert disagreement is pervasive in clinical decision making and collective adjudication is a useful approach for resolving divergent assessments. 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 work, 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.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**; **Human computer interaction (HCI)**; *Collaborative interaction*; *Empirical studies in collaborative and social computing*.

Additional Key Words and Phrases: Ambiguity; Disagreement; Adjudication; Medical time series

ACM Reference Format:
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, 1, Article 1 (November 2019), 23 pages. https://doi.org/01.0001/0000001

1 INTRODUCTION

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

Authors' addresses: Mike Schaekermann, University of Waterloo, Waterloo, Canada, mschaeke@uwaterloo.ca; Graeme Beaton, University of Waterloo, Waterloo, Canada, graeme.beaton@edu.uwaterloo.ca; Minahz Habib, University of Toronto, Toronto, Canada, minahz.habib@mail.utoronto.ca; Andrew Lim, University of Toronto, Toronto, Canada, andrew.lim@utoronto.ca; Kate Larson, University of Waterloo, Waterloo, Canada, kate.larson@uwaterloo.ca; Edith Law, University of Waterloo, Waterloo, Canada, edith.law@uwaterloo.ca.

2573-0142/2019/11-ART1 $15.00
https://doi.org/01.0001/0000001

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 [19, 49], and that resolving disagreements is not always possible or desirable, even if human graders are given the opportunity to deliberate on a case [47]. In the clinical domain, collaborative, team-based decision making has long been deemed superior to individual diagnosis by the National Academy of Medicine [5]. 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 cover related work on expert disagreement, group decision support, and adjudication in medical data analysis, 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.

2 RELATED WORK

2.1 Ambiguity and Sources of Inter-rater Disagreement

Ambiguity, or an openness to multiple interpretations, 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., [20, 33, 51]) and medical (e.g., [3, 34, 40]) literature, but also within the human-computer interaction (HCI, e.g., [10, 12, 24]), human computation (e.g., [2, 18, 32]), and computer-supported cooperative work (CSCW, e.g., [1, 28, 47]) communities. The problem of inter-rater variability is particularly important within medicine where clinical decisions rely on the interpretation of patient data. Indeed, Raghu et al. [40] concluded that label disagreement among medical experts presents a *"full-fledged clinical problem in the healthcare domain."* In the same vein, Paletz et al. [34] stressed the importance of detecting uncertainty among diagnosticians to triage incorrect diagnoses.

In the context of non-expert crowd work, Aroyo and Welty [2] 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 expert

disagreement in medical data analysis, 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 [33] 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 [20] 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 [24] found that disagreement among crowd workers in visual question answering tasks can arise from differing levels of annotator expertise. Kairam and Heer [28] showed that inter-rater agreement in a crowdsourced entity annotation task is affected by how conservatively or liberally workers follow task instructions. In this work, we assume medical experts to be high-quality graders with varying backgrounds, i.e., differences in professional credentials, geographic location, and work experience.

Data characteristics. In addition to grader-specific factors, disagreement may be an indicator of ambiguity, vagueness, or complexity inherent in the given data [1–3, 39, 40]. Prior works have demonstrated that inter-rater disagreement can be associated with characteristics of individual data objects, including text documents for sentiment classification [39], photographs for visual question answering [25], medical images for eye disease assessment [40], and biomedical time series data for epilepsy diagnosis [3]. In this work, we take into account different measures for characterizing complexity in biomedical time series data to study the interplay between case-specific data characteristics and other sources of expert disagreement, including grader differences, data presentation, and guideline clarity.

Task instructions. Finally, inter-rater disagreement has been attributed to ambiguous category definitions [2, 10, 24, 32] relevant to a given task. Gurari and Grauman [24] identified subjective questions and vocabulary mismatch between crowd workers as sources of disagreement. Chang et al. [10] 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 [32] developed workflows for identifying and refining unclear instructions for crowdsourcing tasks. Our work demonstrates how complex classification guidelines can be integrated directly into data adjudication workflows to collect expert rationales with respect to individual guideline rules and rule-specific evidence.

Data presentation. Our work builds on these prior contributions by studying various factors contributing to expert disagreement in medical data analysis. Our quantitative analysis takes into account differences between graders (in terms of professional credentials, geographic location, and work experience), data characteristics of individual disagreement cases (in terms of pathology and signal complexity), and the role of classification guidelines (in terms of individual guideline rules and rule-specific evidence collected during adjudication) to understand expert disagreement. Furthermore, our work contributes a novel perspective on the expert disagreement problem by incorporating *data presentation* as another potential factor contributing to disagreement. In particular, we incorporate in our analysis 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.

2.2 Systems for Group Deliberation

Beyond the problem of understanding how inter-rater disagreement arises is the question of how to deal with the resulting uncertainty. There is a growing body of work in support of group deliberation as a useful and productive method of consensus formation, where members of a group who disagree about a case gather to exchange arguments for their individual classification decisions and collectively weigh evidence in order to reach a shared decision. Group deliberation is favoured in the literature over alternative techniques like majority vote, which tend to discount argument and dissenting insight in order to promote artificial consensus [50].

A seminal procedure for structured group deliberation is the Delphi method, by which a panel of experts exchange arguments and evidence through multiple rounds, moderated by a facilitator who summarizes individual positions and discards what is assumed to be irrelevant comments after each round [16]. Since its introduction, the Delphi method has been employed for several aims, including to facilitate remote group decision making [26]. Several systems have been built to facilitate online deliberation, though not all of them directly resemble the structure of the Delphi method. Drapeau et al. [18] introduced MicroTalk, a crowdsourcing workflow with functionality for asynchronous argumentation between workers in the context of a semantic relation-extraction task. Drapeau et al. [18] evaluated their workflow with respect to accuracy improvements against a golden reference standard, demonstrating that online, asynchronous worker argumentation can improve answer accuracy over computational aggregation methods. Building on this work, Chen et al. [12] showed that a synchronous workflow enabling crowd workers to engage in real-time, multi-turn discussions, can lead to additional improvements in answer accuracy. Schaekermann et al. [47] also used a synchronous workflow for worker deliberation to analyze how disagreement arises and under what circumstances it can be resolved for text classification tasks with varying levels of subjectivity.

Chang et al. [10] 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. [22] 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. [11] proposed a multi-step crowdsourcing workflow for semantic frame annotation, allowing workers to express disagreement with expert-labelled golden data presented as feedback during labelling.

Our system leverages 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 a web-based, structured adjudication system. 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 annotation guidelines. 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 streamlining the efficiency or accuracy of data labelling workflows.

2.3 Adjudication in Medical Data Analysis

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 [23, 29, 41].

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. [29] 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. [23] 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 [37] 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. [7] 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.

2.4 Computational Models of Argumentation

Argumentation, an approach to reasoning centered around the logic of inferring conclusions from given data, has an extensive history in the field of computer science. Prior research has focused on patterns common to human argumentation in decision making [14, 52], and there is a wealth of literature on mapping natural human language used in argumentative discourse to machine-readable representations [8, 9, 13, 14, 30, 31, 43]. Studies have shown that human argument represented in a machine-readable format can be used to generate new conclusions in the context of novel inquiries [35, 42]. Building on basic concepts of propositional logic, the structure we employ to collect expert rationales during adjudication in the present work is compatible with existing methods (e.g., [35]) and interoperable with logic-based approaches in the field of computer science broadly.

3 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 agreement among two independent expert averages as low as 82.6% [44]. 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.

4 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.

4.1 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. For all three steps, the same signal viewer software[1] 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 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

[1]Part of the crowdEEG research project (http://crowdeeg.ca, [45, 46]).

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.

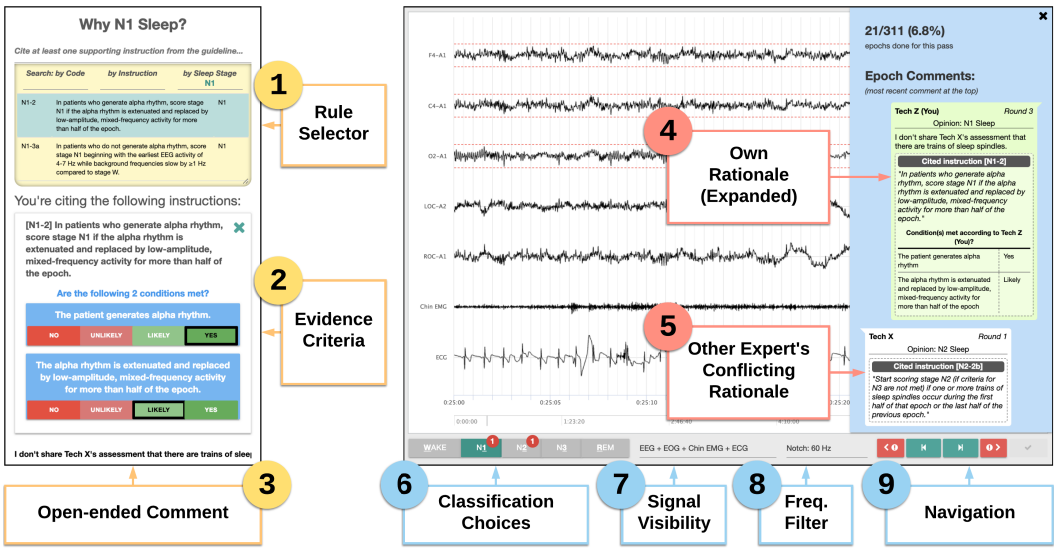
Open-ended discussion comments collected during the pilot were generally free of inter-personal comments, concise (ranging from a few words to one or two sentences), and focused primarily on specific rules from the classification guidelines including low-level features referenced therein. While the majority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

4.2 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,



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

Fig. 1. Interface for structured adjudication of classification decisions in medical time series analysis.

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 [27], 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.

4.3 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 [48] 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.

4.4 User Interface

We designed a user interface (UI) to implement our workflows for classification and adjudication within a web-based platform (Figure 1). 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 1b, components 6 to 9).

Adjudication. The adjudication UI (Figure 1) 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 1b, 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 1b, 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 1b). 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 1b, component 4) from those of other experts in the group (Figure 1b, 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 1a). 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 1a, 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 1a, component 2); and the option to provide an additional open-ended comment (Figure 1a, 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.

5 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 [5], 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.

6 METHODS

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

6.1 Participant Recruitment

We recruited 36 expert participants via domain-specific online platforms. Based on the pre-study questionnaire, 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.

6.2 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.

6.3 Procedure

Before the study, experts first completed a pre-study questionnaire 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, 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.

6.4 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 1.

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

Table 1. 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 [6]
	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

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.

7 RESULTS

Structured adjudication resulted in a 20-30% increase in inter-rater agreement over the course of three rounds (cf. Figure 2). 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); 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.

7.1 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 2, 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**.

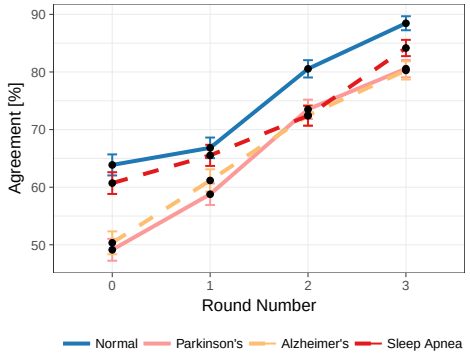


Fig. 2. Agreement rate by adjudication round number and patient's health condition.

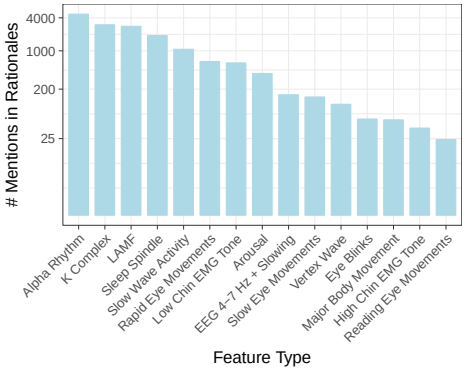


Fig. 3. Number of times each feature type was mentioned in a rationale (log scale).

- 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 2).

7.2 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 2, 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 chances of leaving a case unresolved. Differences in amplitude scaling,

Table 2. 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	***

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 2 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.

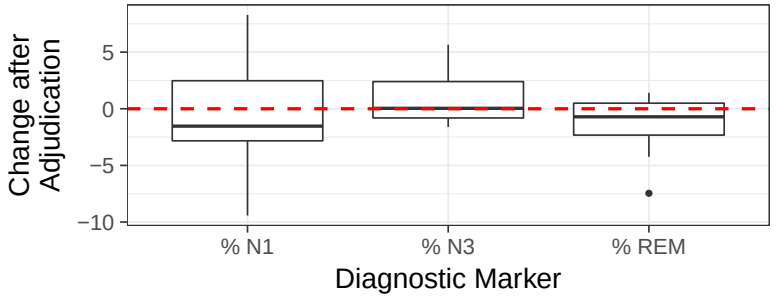


Fig. 4. Change in diagnostic markers from before to after adjudication.

7.3 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 4). 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 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.

8 DISCUSSION

Our core contribution in this present work 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.

8.1 Generalizability and Applications

Our case study was limited to a task specific to biomedical time series analysis, so caution is warranted in generalizing the results to outside domains. 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 [4], and the organization of guidelines into easily identifiable grading recommendations is encouraged within the medical community [21]. 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 [36]. 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 [38]. 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 [15].

8.2 Design Considerations for Expert Adjudication

Davies and Chandler [17] 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 work, 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 [17]. Whether inter-rater agreement rate increases with a synchronous version of our system is a question for future research.

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 [17]. 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 [22]. 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.

8.3 Limitations and Future Work

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 read 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.

9 CONCLUSION

In this work, 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-labelled data in the context of other applications like computer-aided diagnosis.

ACKNOWLEDGMENTS

We thank Rui de Sousa for his invaluable help in recruiting participants for this study. This work was funded by NSERC CHRP (CHRP 478468-15), CIHR CHRP (CPG-140200), and the Google PhD Fellowship Program.

REFERENCES

[1] Paul André, Aniket Kittur, and Steven P Dow. 2014. Crowd synthesis: Extracting categories and clusters from complex data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 989–998.

[2] Lora Aroyo and Chris Welty. 2014. The three sides of CrowdTruth. *Journal of Human Computation* 1 (2014), 31–34.

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

[4] A. Baker, K. Young, J. Potter, and I. Madan. 2010. A review of grading systems for evidence-based guidelines produced by medical specialties. *Clinical Medicine* 10, 4 (8 2010), 358–363. https://doi.org/10.7861/clinmedicine.10-4-358

[5] Erin P. Balogh, Bryan T. Miller, and John R. Ball (Eds.). 2015. *Improving Diagnosis in Health Care*. National Academies Press, Washington, D.C. https://doi.org/10.17226/21794

[6] Forrest S Bao, Xin Liu, and Christina Zhang. 2011. PyEEG: An Open Source Python Module for EEG/MEG Feature Extraction. *Computational Intelligence and Neuroscience* 2011 (2011), 1–7. https://doi.org/10.1155/2011/406391

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

[8] Floris Bex, Henry Prakken, Chris Reed, and Douglas Walton. 2003. Towards a Formal Account of Reasoning about Evidence: Argumentation Schemes and Generalisations. *Artificial Intelligence and Law* 11, 2/3 (2003), 125–165. https://doi.org/10.1023/B:ARTI.0000046007.11806.9a

[9] Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards Argument Mining from Dialogue. In *Computational Models of Argument - Proceedings of {COMMA} 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014*. 185–196. https://doi.org/10.3233/978-1-61499-436-7-185

[10] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM, ACM Press, New York, New York, USA, 2334–2346. https://doi.org/10.1145/3025453.3026044

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

[12] Quanze Chen, Jonathan Bragg, Lydia B. Chilton, and Daniel S. Weld. 2019. Cicero. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1–14. https://doi.org/10.1145/3290605.3300761

[13] Carlos Chesñevar, Jarred McGinnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, and Steven Willmott. 2006. Towards an argument interchange format. *The Knowledge Engineering Review* 21, 04 (12 2006), 293. https://doi.org/10.1017/S0269888906001044

[14] Robin Cohen. 1987. Analyzing the Structure of Argumentative Discourse. *Comput. Linguist.* 13, 1-2 (1 1987), 11–24. http://dl.acm.org/citation.cfm?id=26386.26388

[15] Robin Cohen, Mike Schaekermann, Sihao Liu, and Michael Cormier. 2019. Trusted AI and the Contribution of Trust Modeling in Multiagent Systems. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1644–1648. http://dl.acm.org/citation.cfm?id=3306127.3331890

[16] Norman Dalkey and Olaf Helmer. 1963. An Experimental Application of the DELPHI Method to the Use of Experts. *Management Science* 9, 3 (4 1963), 458–467. https://doi.org/10.1287/mnsc.9.3.458

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

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

[19] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing Ground Truth for Medical Relation Extraction. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (7 2018), 1–20. https://doi.org/10.1145/3152889

[20] Luciana Garbayo. 2014. Epistemic Considerations on Expert Disagreement, Normative Justification, and Inconsistency Regarding Multi-criteria Decision Making. *Constraint Programming and Decision Making* 539 (2014), 35–45. http://link.springer.com/10.1007/978-3-319-04280-0%5C_5

[21] Gowri Gopalakrishna, Miranda W Langendam, Rob JPM Scholten, Patrick MM Bossuyt, and Mariska MG Leeflang. 2013. Guidelines for guideline developers: a systematic review of grading systems for medical tests. *Implementation Science* 8, 1 (12 2013), 78. https://doi.org/10.1186/1748-5908-8-78

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

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

[24] Danna Gurari and Kristen Grauman. 2017. 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*. ACM, ACM Press, New York, New York, USA, 3511–3522. https://doi.org/10.1145/3025453.3025781

[25] Danna Gurari, Kun He, Bo Xiong, Jianming Zhang, Mehrnoosh Sameki, Suyog Dutt Jain, Stan Sclaroff, Margrit Betke, and Kristen Grauman. 2017. Predicting Foreground Object Ambiguity and Efficiently Crowdsourcing the Segmentation(s). (4 2017). http://arxiv.org/abs/1705.00366

[26] Francis T. Hartman and Andrew Baldwin. 1995. Using Technology to Improve Delphi Method. *Journal of Computing in Civil Engineering* 9, 4 (10 1995), 244–249. https://doi.org/10.1061/(ASCE)0887-3801(1995)9:4(244)

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

[28] Sanjay Kairam and Jeffrey Heer. 2016. 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*. ACM Press, New York, New York, USA, 1635–1646. https://doi.org/10.1145/2818048.2820016

[29] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2018. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology* (3 2018). https://doi.org/10.1016/j.ophtha.2018.01.034

[30] John Lawrence and Chris Reed. 2015. Combining Argument Mining Techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining at ACL 2015*. 127–136. https://doi.org/10.3115/v1/W15-0516

[31] John Lawrence and Chris Reed. 2016. Argument Mining using Argumentation Scheme Structures. *Proceedings of the 6th International Conference on Computational Models of Argument (COMMA 2016)* 0 (2016), 379 – 390. https://doi.org/10.3233/978-1-61499-686-6-379

[32] V K Chaithanya Manam and Alexander J Quinn. 2018. WingIt: Efficient Refinement of Unclear Task Instructions. In *The Sixth AAAI Conference on Human Computation and Crowdsourcing*. 108–116. https://www.aaai.org/ocs/index.php/HCOMP/HCOMP18/paper/view/17931

[33] Jeryl L. Mumpower and Thomas R. Stewart. 1996. Expert Judgement and Expert Disagreement. *Thinking & Reasoning* 2, 2-3 (7 1996), 191–212. https://doi.org/10.1080/135467896394500

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

[35] Simon Parsons, Elizabeth Sklar, Jordan Salvit, Holly Wall, and Zimi Li. 2013. ArgTrust: Decision Making with Information from Sources of Varying Trustworthiness. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems (AAMAS '13)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1395–1396. http://dl.acm.org/citation.cfm?id=2484920.2485242

[36] 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. 2017. Sleep architecture and the risk of incident dementia in the community. *Neurology* 89, 12 (2017), 1244–1250.

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

[38] 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. 2019. Risk and predictors of dementia and parkinsonism in idiopathic REM sleep behaviour disorder: a multicentre study. *Brain* 142, 3 (2019), 744–759.

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

IEEE, 179–188.
[40] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Robert Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. 2018. Direct Uncertainty Prediction for Medical Second Opinions. (7 2018). http://arxiv.org/abs/1807.01771
[41] Pranav Rajpurkar, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y. Ng. 2017. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. (7 2017). http://arxiv.org/abs/1707.01836
[42] Chris Reed and Timothy Norman. 2004. *Argumentation Machines*. Argumentation Library, Vol. 9. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-94-017-0431-1
[43] Chris Reed and Doug Walton. 2005. Towards a Formal and Implemented Model of Argumentation Schemes in Agent Communication. 19–30. https://doi.org/10.1007/978-3-540-32261-0{_}2
[44] Richard S. Rosenberg and Steven van Hout. 2013. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *Journal of Clinical Sleep Medicine* (1 2013). https://doi.org/10.5664/jcsm.2350
[45] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019. Capturing Expert Arguments from Medical Adjudication Discussions in a Machine-readable Format. In *Companion Proceedings of The 2019 World Wide Web Conference on - WWW '19*, Vol. 2. ACM Press, New York, New York, USA, 1131–1137. https://doi.org/10.1145/3308560.3317085
[46] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019. crowdEEG: A Platform for Structured Consensus Formation in Medical Time Series Analysis. In *8th Workshop on Interactive Systems in Healthcare (WISH) at CHI 2019*. Glasgow, UK.
[47] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. In *Proceedings of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'18)*. New York City, NY. https://doi.org/10.1145/3274423
[48] Mike Schaekermann, Edith Law, Kate Larson, and Andrew Lim. 2018. Expert Disagreement in Sequential Labeling: A Case Study on Adjudication in Medical Time Series Analysis. In *1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing at HCOMP 2018*. Zurich, Switzerland.
[49] Mike Schaekermann, Edith Law, Alex C Williams, and William Callaghan. 2016. 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.
[50] Miriam Solomon. 2006. Groupthink versus The Wisdom of Crowds : The Social Epistemology of Deliberation and Dissent. *The Southern Journal of Philosophy* 44, S1 (3 2006), 28–42. https://doi.org/10.1111/j.2041-6962.2006.tb00028.x
[51] Miriam Solomon. 2007. The social epistemology of NIH consensus conferences. In *Establishing medical reality*. Springer, 167–177.
[52] D Walton, C Reed, and F Macagno. 2008. *Argumentation Schemes*. Cambridge University Press. https://books.google.ca/books?id=qc3LCgAAQBAJ

Received April 2019; revised June 2019; accepted July 2019