crowdEEG: A Platform for Structured Consensus Formation in Medical Time Series Analysis

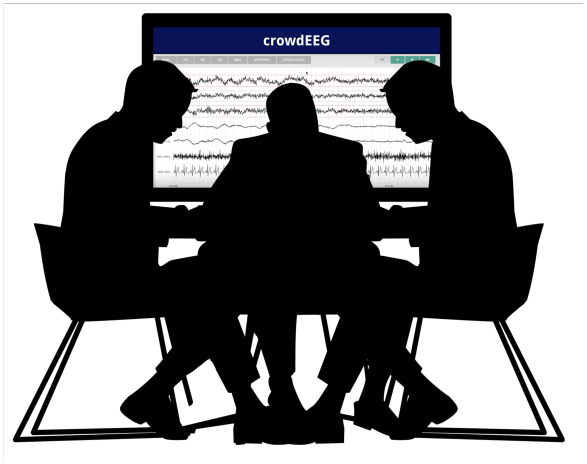

Figure 1: In-person adjudication among a panel of three experts. Disagreement epochs in the EEG recording are discussed until a consensus is reached.

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

ABSTRACT

Disagreement among domain experts in medical image interpretation is a wide-spread, yet poorly managed phenomenon. With the exception of only a few medical disciplines like radiology, the practice of second reads and adjudication of divergent expert assessments is rarely implemented in clinical

WISH at CHI'19, May 04–05, 2019, Glasgow, UK

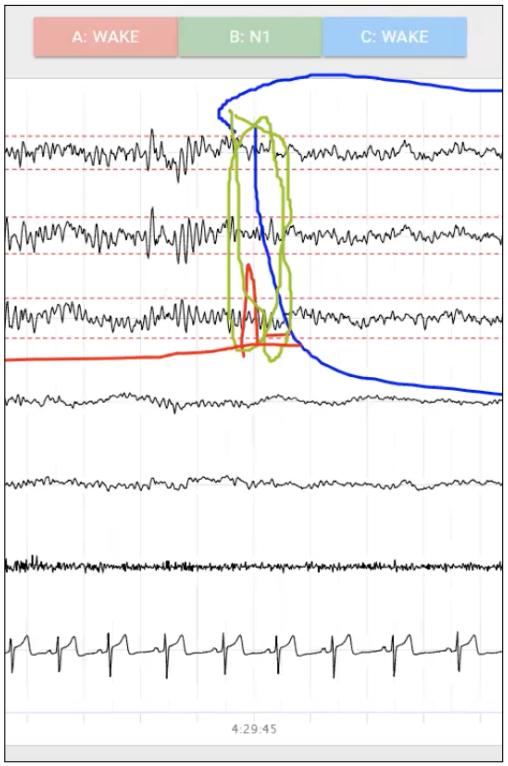

Figure 2: Remote adjudication via video conference. Three panel experts annotate the same recording on a shared screen while discussing divergent interpretations of signal patterns.

[1] http://crowdeeg.ca

workflows. We posit that sparse adoption of adjudication procedures in medicine is in part due to the lack of effective tools supporting consensus formation. Addressing this gap, we conducted an iterative design exploration with the goal to develop a web-based adjudication platform for structured consensus formation among panels of medical experts. In this work, we report our findings from this design journey within the application domain of medical time series 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*.

KEYWORDS

Platform; Adjudication; Disagreement; Medical time series

ACM Reference Format:
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 Workgroup on Interactive Systems in Healthcare (WISH) at CHI'19: ACM Conference on Human Factors in Computing Systems, May 04–05, 2019, Glasgow, UK*. ACM, New York, NY, USA, 4 pages.

INTRODUCTION

High inter-rater variability is pervasive across various domains of medical image interpretation [1, 3], and prior work has shown that panel-based adjudication can improve the quality of group decisions [2, 4]. Effective tools that enable collaborative diagnostic consensus formation, however, are sparse. In this work, we contribute findings from an iterative design exploration through the development of **crowdEEG**,[1] a web-based platform for collaborative annotation and adjudication of medical time series. Our design exploration 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 steps 1 and 2. The crowdEEG platform was used as a signal viewer for all three steps in the process, but only in step 3, adjudication of disagreements was conducted directly within the platform. Our design study was embedded in the application domain of sleep stage classification, the task of mapping a sequence of 30-second epochs of multimodal medical time series (*polysomnogram*) to a sequence of discrete sleep stages (*hypnogram*). Each epoch is classified into one of five stages of sleep—Wake, NREM1, NREM2, NREM3 or REM sleep. Expert agreement rates in sleep stage classification average around 82.6% [3].

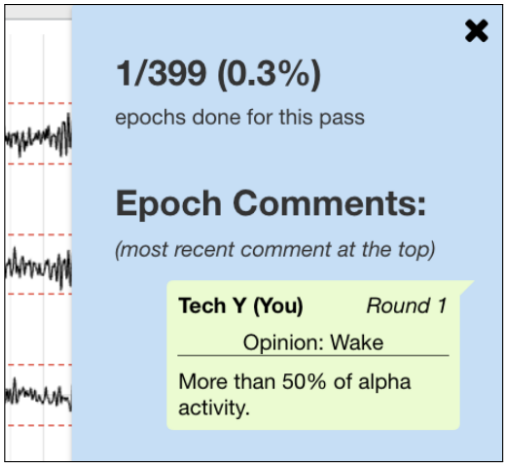

Figure 3: Web-based adjudication using free-form comments to explain rationales.

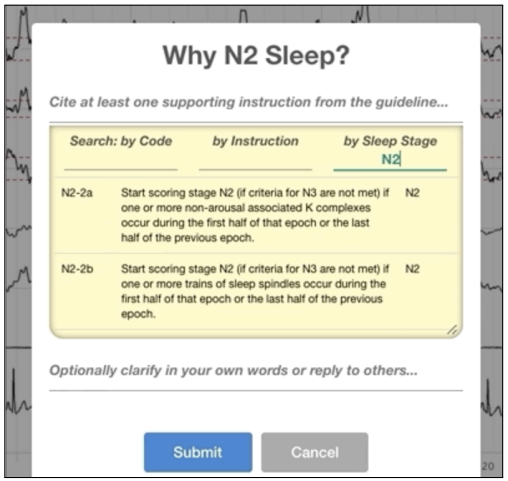

Figure 4: Web-based adjudication with integrated, citable scoring guidelines.

ITERATIVE DESIGN STUDY

Our iterative design exploration consisted of a 3-step process with the goal of identifying relevant design considerations for tools that support consensus formation in medical time series analysis.

In-Person Adjudication

An initial formative session of in-person adjudication was conducted with three board-certified sleep technologists. We describe the details of the procedure and the data selection criteria in [5]. After an initial round of independent scoring, researchers organized an in-person meeting in the hospital to discuss select disagreement epochs. All members of the expert panel convened at a set time and place to discuss disagreement cases in front of a shared screen (Figure 1). During the in-person session, it became apparent that certain scoring guidelines, as well as individual patterns or features in the signal (e.g., sleep spindles and arousals) played important roles both as sources of disagreement and as evidence to support consensus. In-person discussions were influenced by inter-personal factors such as perceived grader experience and the effectiveness of individual communication and argumentation skills—observations serving as design considerations for next steps.

Remote Adjudication via Video Conference

In a second step, we conducted an exploratory adjudication session with the same three sleep technologists, enabling remote discussion via video conference. 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 in the signal (Figure 2) in real time during discussion. A key insight from remote adjudication via video conference was that the localization of ambiguous features and the identification of feature boundaries were helpful mechanisms to pinpoint sources of disagreement during discussion, and that the discussion around individual features was consistently rooted in the context of specific grading guidelines.

Web-based Adjudication Platform

In order to control for the effects of inter-personal dynamics observed during both in-person adjudication and adjudication via video conference, we implemented a web-based adjudication platform through which a pool of anonymous graders participated in round-robin reviews of a recording in panels of three. After independent annotation, each grader in a panel was asked to review all disagreement cases across the entire recording, one at a time, for a total of three rounds. This experimental design was based on our prior observation that the quality of individual scoring decisions can depend on the number of passes a grader has made over a recording. Adjudication took place in two forms.

Free-form Discussions. Readers deliberated over disagreement cases and entered rationales for their scoring decisions through a free-form input field (cf. Figure 3). The goal here was to collect a diverse range of arguments without constraining graders. Our previous observation that expert discussions tended to be rooted in explicit grading guidelines was confirmed in this part of the study.

Integration of Grading Guidelines. As a result, a comprehensive set of grading guidelines was incorporated into the web-interface (Figure 4) enabling readers to cite explicit rules for their scoring decisions. Graders could still provide free-form comments, but were required to cite at least one guideline instruction in support of each scoring decision. This implementation allowed for highly structured data to be collected during the adjudication process. Our preliminary analysis revealed that the vast majority of scoring decisions required only one guideline rule for justification, while a small set of cases required citation of two or three guideline rules.

CONCLUSION

In this work, we reported findings from an iterative design exploration with the goal of developing crowdEEG, a web-based platform for structured consensus formation among medical experts. Our study illuminated various dynamics of group deliberation in the medical domain through three different lenses: in-person adjudication, adjudication via video conference and web-based adjudication using both free-form comments and structured rationales to justify individual assessments.

ACKNOWLEDGMENTS

We thank Rui DeSousa for his invaluable help in recruiting participants for this study. This work was funded by NSERC CHRP (CHRP 478468-15) and CIHR CHRP (CPG-140200).

REFERENCES

[1] 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

[2] 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

[3] 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

[4] 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

[5] 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.