The Wisdom of the Confident Crowd in Medical Image Decision-making

Eeshan Hasan[1,2] Jennifer Trueblood [1,2]

[1]Department of Psychological and Brain Science, Indiana University Bloomington [2]Cognitive Science Program, Indiana University Bloomington

Abstract

We explore methods for improving the accuracy of medical image decisions by aggregating repeated decisions. Novices (undergraduates) and experts (medical professionals) made classification decisions (cancer vs. not cancer) and confidence judgments on cell images, viewing and classifying each image twice. We show that the maximum confidence slating algorithm, which uses the most confident response for an image as the final response, improves performance for novices and experts at the individual level. We then show that aggregation algorithms based on confidence weighting scale to larger groups of participants, with the performance of groups of novices reaching that of individual experts.

Experimental Design

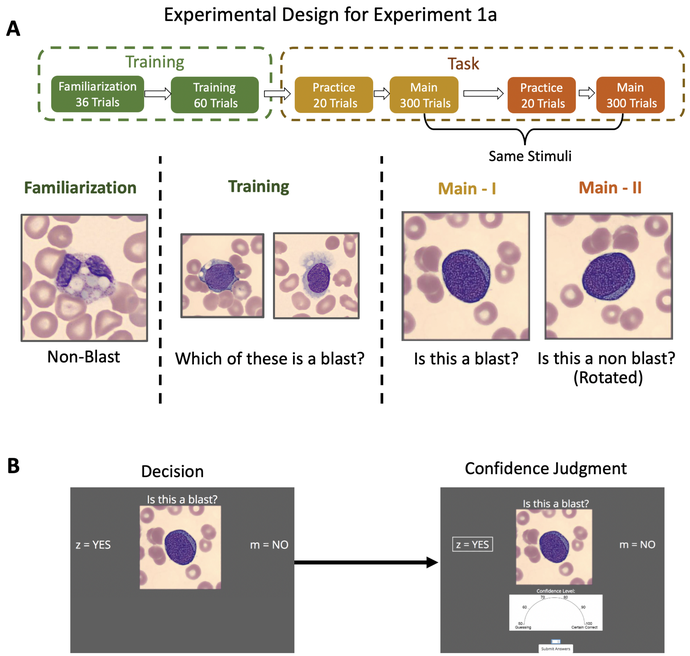

Figure 1. Panel A illustrates the structure of Experiment 1a. Participants first completed a brief training phase before the main experiment. In the main task, two responses were collected for each participant on each of the 300 images. In the first pass (shown in yellow), participants were presented with the image of a cell and were asked the question 'Is this a blast?' (i.e. Is this cancerous?). In the second pass (shown in orange), the same image was rotated and participants were asked the question 'Is this a non-blast?' (i.e. Is this non-cancerous?). Each set of main trials was preceded by practice trials. Experiment 1b was similar to Experiment 1a except that the same question was asked in both of the main blocks (i.e., 'Is this a blast?') and images were not rotated. Experiment 2 was a shorter version of Experiment 1a and was designed for expert participants. Panel B shows the two parts of every trial in the main task. In the first part, participants decided whether a cell was a blast or not. In the second part, participants indicated their confidence on a scale of 50-100, where 50 was 'guessing' and 100 was 'certain correct'.

Results: Combining two responses

Table 1. The mean performance of each algorithm for Experiments 1a, 1b, and 2.

Algorithm	Exp. 1a Novice (Reframing)	Exp. 1b Novice (No Reframing)	Exp. 2 Expert (Reframing)
Average Response - Within	66.1%	66.5%	71.6%
Max. Conf. Slating - Within	67.4%*	67.4%*	73.7%*
Average Response - Between	66.7%	66.5%	72.4%
Max. Conf. Slating - Between	70.0%*	70.0%*	78.5%*

Note. *Significant improvement as compared to average response with p<0.0001

Results: Combining Multiple Responses

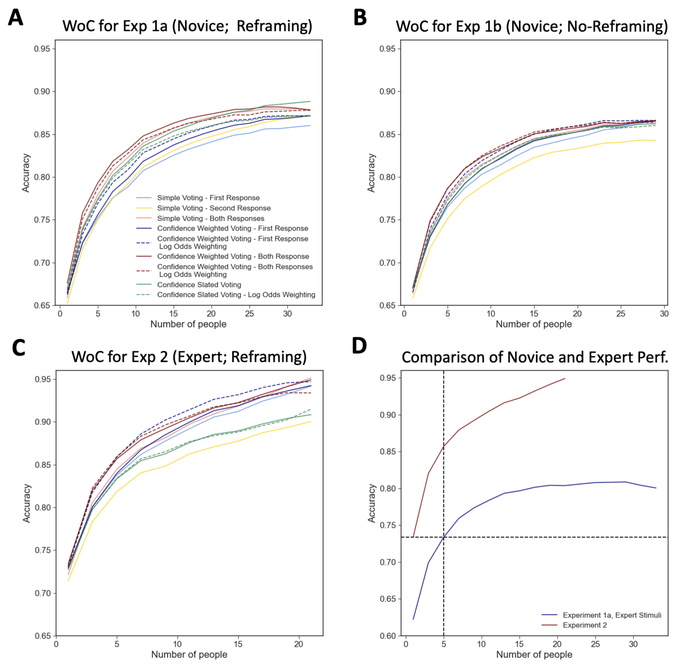

Figure 2. Panels A, B and C plot the accuracy obtained by applying all of our Wisdom of the Crowd algorithms to data from Experiments 1a, 1b, and 2 respectively. The legend for Panel B, C and D is identical to the one is Panel A. We observe that as the responses are pooled from more participants, the accuracy increases. Panel D compares the best performing algorithm, Confidence Weighted Voting with Both Responses, for novice participants from Experiment 1a and expert participants from Experiment 2. We restrict the responses from Experiment 1a to the stimuli that were used in the main trials of Experiment 2 and apply the algorithm. We observe that the performance of the expert participants is greater than that of the novice participants. We also observe that 5 novice participants can match the performance of 1 expert.

Results: Comparison of the Algorithms

Table 2. Results of the different Wisdom of the Crowd (WoC) algorithms when groups of 7 decision makers are considered.

Algorithm	Exp. 1a Novice Reframing	Exp. 1b Novice No Reframing	Exp. 2 Expert Reframing
First Response (1 Person) [a]	66.5%	67.0%	73.0%
Simple Voting - First Resp.	77.6%	78.7%	86.2%
Simple Voting - Second Resp.	77.5%	77.5%	84.0%
Simple Voting - Both Resp.	80.7%	80.0%	86.9%
Conf. Weight Voting - First Resp.	78.3%	79.2%	86.7%
Conf. Weight Voting - First Resp. - Log Odds Weight	79.4%	80.4%	**88.6%**
Conf. Weight Voting - Both Resp.	**81.8%**	**81.0%**	**87.9%**
Conf. Weight Voting - Both Resp. - Log Odds Weight	**81.3%**	**81.0%**	**88.3%**
Conf. Slated Voting	80.2%	79.3%	85.5%
Conf. Slated Voting - Log Odds Weight	79.9%	79.8%	85.7%

Note. The best performing algorithm for each experiment is in bold.
[a] The average accuracy using only the first response for a group size of one is provided as a baseline comparison.

Conclusions

- Maximum confidence slating - within improves performance for Novices and Experts.
- Maximum confidence slating - between is more effective than maximum confidence slating within. Using responses from two people is more useful than asking the same person the same question again.
- Aggregating decisions from a large number of individuals dramatically improves performance.
- Repeated decision making and confidence weighting improves performance at the group level. This is especially larger when the frame of the question is changed.
- It is better to average confidence judgments than slating to the more confident decision.
- Aggregating decisions from a small number of novices (i.e. 5 novices) matches the performance of experts.

Acknowledgments

We would like to thank my co-authors Quentin Eichbaum, Adam Seegmiller and Charles Stratton for their contribution to the project and their medical expertise. I would also like to thank Payton O'Daniels for his excellent research assistance.
This work was supported by a Clinical and Translational Research Enhancement Award from the Department of Pathology, Microbiology, and Immunology, Vanderbilt University Medical Center. This work was also supported by NSF grant 1846764.

References

[1] Eeshan Hasan, Quentin Eichbaum, Adam C Seegmiller, Charles Stratton, and Jennifer Trueblood. Harnessing the wisdom of the confident crowd in medical image decision-making. 2021.

[2] Eeshan Hasan, Quentin Eichbaum, Adam C Seegmiller, Charles Stratton, and Jennifer S Trueblood. Harnessing the wisdom of the confident crowd in medical image decision-making. 2021. URL `osf.io/ckvxz`.

[3] Asher Koriat. When are two heads better than one and why? *Science*, 336(6079):360–362, 2012.

[4] Aleksandra Litvinova, Ralf HJM Kurvers, Ralph Hertwig, and Stefan M Herzog. How experts' own inconsistency relates to their confidence and between-expert disagreement. *Scientific Reports*, 12(1):1–12, 2022.