HHS Public Access
Author manuscript
Pac Symp Biocomput. Author manuscript; available in PMC 2021 March 15.

Published in final edited form as:
Pac Symp Biocomput. 2021 ; 26: 14–25.

Selection of trustworthy crowd workers for telemedical diagnosis of pediatric autism spectrum disorder

Peter Washington[1], **Emilie Leblanc**[2,3], **Kaitlyn Dunlap**[2,3], **Yordan Penev**[2,3], **Maya Varma**[4], **Jae-Yoon Jung**[2,3], **Brianna Chrisman**[1], **Min Woo Sun**[3], **Nathaniel Stockham**[5], **Kelley Marie Paskov**[3], **Haik Kalantarian**[2,3], **Catalin Voss**[4], **Nick Haber**[6], **Dennis P. Wall**[2,3]

[1]Department of Bioengineering, Stanford University, Palo Alto, CA, 94305, USA

[2]Department of Pediatrics (Systems Medicine), Stanford University, Palo Alto, CA, 94305, USA

[3]Department of Biomedical Data Science, Stanford University, Palo Alto, CA, 94305, USA

[4]Department of Computer Science, Stanford University, Palo Alto, CA, 94305, USA

[5]Department of Neuroscience, Stanford University, Palo Alto, CA, 94305, USA

[6]School of Education, Stanford University, Palo Alto, CA, 94305, USA

Abstract

Crowd-powered telemedicine has the potential to revolutionize healthcare, especially during times that require remote access to care. However, sharing private health data with strangers from around the world is not compatible with data privacy standards, requiring a stringent filtration process to recruit reliable and trustworthy workers who can go through the proper training and security steps. The key challenge, then, is to identify capable, trustworthy, and reliable workers through high-fidelity evaluation tasks without exposing any sensitive patient data during the evaluation process. We contribute a set of experimentally validated metrics for assessing the trustworthiness and reliability of crowd workers tasked with providing behavioral feature tags to unstructured videos of children with autism and matched neurotypical controls. The workers are blinded to diagnosis and blinded to the goal of using the features to diagnose autism. These behavioral labels are fed as input to a previously validated binary logistic regression classifier for detecting autism cases using categorical feature vectors. While the metrics do not incorporate any ground truth labels of child diagnosis, linear regression using the 3 correlative metrics as input can predict the mean probability of the correct class of each worker with a mean average error of 7.51% for performance on the same set of videos and 10.93% for performance on a distinct balanced video set with different children. These results indicate that crowd workers can be recruited for performance based largely on behavioral metrics on a crowdsourced task, enabling an affordable way to filter crowd workforces into a trustworthy and reliable diagnostic workforce.

dpwall@stanford.edu.

Keywords

Crowdsourcing; Machine Learning; Diagnostics; Trust; Privacy; Autism

1. Introduction

Autism spectrum disorder (ASD, or autism) is a pediatric developmental condition affecting 1 in 40 children in the United States [1], with prevalence continuing to rise [2]. While access to care relies on a formal diagnosis from a clinician, an uneven distribution of diagnostic resources across the United States contributes to increasingly long waitlists. Some evidence suggests that 80% of counties lack sufficient diagnostic resources [3], with underserved communities disproportionately affected by this shortage [4]. Telemedicine has the potential to minimize this gap by capitalizing on the increasing pervasiveness and affordability of digital devices. Such diagnostic solutions are especially pertinent during times of pandemic, most notably the coronavirus, which further hinders access to diagnosis and care.

Mobile digital autism interventions administered on smartphones [5-12] and on ubiquitous devices [13-27] passively collect structured home videos of children with neuropsychiatric conditions for use in subsequent diagnostic data analysis [27-28]. In order for the video data collected from digital therapies to become widely used, trustworthy data sharing methodologies must be incorporated into the diagnostic pipeline [29]. One possible approach, which we realize in the present study, is to carefully recruit a trustworthy set of workers to transform the video streams into a secure, quantitative, and structured format. While modern computer vision algorithms could handle this task in several domains, extracting complex behavioral features from video is currently beyond the scope of state-of-the-art machine learning methods and therefore requires human labor. However, the collected videos naturally contain highly sensitive data, requiring careful selection of trustworthy and reliable labelers who are allowed access to protected health information (PHI) after completion of Health Insurance Portability and Accountability Act (HIPAA) training, Collaborative Institutional Training Initiative (CITI) human subjects training, and whole disk encryption.

In the present study, we examine strategies for quantitatively determining the credibility and reliability of crowd workers whose labels can be trusted by researchers. It is important that the metrics for evaluating workers are speedy and simple, as formally credentialing recruited crowd workers through institutional channels is laborious and slow. We crowdsource the task of providing categorical feature labels to videos of children with autism and matched controls. For each crowdsourced worker, we evaluate correlations of their mean classifier probability of the correct class (PCC) using their answers as input with (1) the mean L1 distance between their responses to the same video spaced one month apart, (2) the mean L1 distance between their answer vector to each video and all other videos they rated, (3) the mean time spent rating videos, and (4) the mean time and L1 distance of answers when the worker is explicitly warned about not spending enough time rating a video and provided with a chance to revise their response. We then feed the metrics which are correlated with PCC into a linear regression model predicting the PCC.

2. Methods

2.1. Clinically representative videos

We used a set of 24 publicly available videos from YouTube of children with autism and matched neurotypical controls (6 females with autism, 6 neurotypical females, 6 males with autism, and 6 neurotypical males). Criteria for video selection and inclusion were that (1) the child's hand and face must be visible, (2) opportunities for social engagement must be present, and (3) an opportunity for using an object such as a toy or utensil must be present. Child diagnosis was determined through the video title and description. The videos were short, with a mean duration of 47.75 seconds (SD = 30.71 seconds). The mean age of children in the video was 3.65 years (SD = 1.82 years).

2.2. Crowdsourcing task for Microworkers

Prior work has validated the capability of subsets of the crowd recruited from the Amazon Mechanical Turk crowdsourcing platform [30] to provide feature tags of children with autism comparable to clinical coordinators working with children with autism on a daily basis [31-32]. We instead recruited workers from Microworkers.com, as Microworkers consists of a diverse representation of worker nationalities [33] compared to Mechanical Turk, which contains workers mostly from the United States and India [34]. Furthermore, Microworkers provides built in functionality for allowing workers to revise their answers if a requester is unsatisfied but believes the worker can redeem their response. This functionality was crucial for our trustworthiness metric.

The task consisted of a series of 13 multiple choice questions identified, in prior work which employed feature selection algorithms on electronic health records [35-44], as salient categorical ordinal features for autism prediction. Workers were asked to watch a short video and answer the multiple-choice questions using the interface depicted in Fig. 1. Microworkers automatically records the time spent on each task.

Through a pilot study of internal lab raters providing 9,374 video ratings for which we logged labeling times, we observed that the mean time per video was 557.7 seconds (9 minutes 18 seconds), with a standard deviation of 929.7 seconds (15 minutes 30 seconds). The pilot task consisted of answering 31 multiple choice questions, while the Microworkers task only contained 13 questions; the proportional mean time is 233.9 seconds (3 minutes 54 seconds). We therefore required workers to spend at least 2 minutes per video, a time threshold significantly below the 233.9 second mean proportional time. If any crowd worker spent less than 2 minutes rating a video, we leveraged the built-in functionality on Microworkers to prompt these users to revise their answers and sent them a warning message disclosing that we know the "*Impossibly short time spent on task*." We measured the additional time spent by the worker, if any, as well as the changes in the answer vector (L1 distance) after receiving this message.

We posted all tasks for all 24 videos exactly 30 days after the original task, allowing workers who completed the first task to complete the task again while minimizing the chance that they could use the memory of their prior responses to bias the test. Previous studies which evaluate test-retest reliability consider 2 weeks to be sufficient time to prevent memorization of prior administrations of the questionnaire [45-48], and we increased this time frame to 30 days to minimize the likelihood that any memory of the workers' previous answers remained. The same video of the child was provided for both administrations of the task. Workers were not provided with their original answers for reference. The difference between the worker's original answers and their revised answers on the same video served as quantitative information about the *reliability* of the worker.

2.3. Classifier to evaluate performance

For a gold standard, we use a previously published and validated [49-54] logistic regression classifier (Fig. 2), trained on electronic health record databases of autism diagnostic scoresheets filled out by expert clinicians, which emits a probability score of autism using the crowd workers' multiple-choice responses as categorical ordinal feature vectors. Because logistic regression classifiers produce a probability, we treat the probability as a confidence score of the crowdsourced workers' responses. We analyze the probability of the correct class (referred to as PCC), which is p when the true class is autism and $1-p$ when the true class is neurotypical. When assessing classifier predictions, we use a threshold of 0.5. We use a worker's average PCC for videos the worker has rated as a metric of the worker's video tagging capability, with a higher mean PCC corresponding to greater mean performance by the worker.

2.4. Metrics evaluated

We strive to develop metrics which only take input parameters that do not depend on *a priori* knowledge about the correct classification score of the videos. We test the following metrics for correlation with the PCC, where N is the number of videos rated by a worker, M is the number of questions per video rating task (inputs to the diagnostic classifier), and $A_{i,j,k}$ is the answer for video i and question j for the k^{th} time.

Mean same-child L1 distance (MSCL$_1$): We asked crowd workers to rate the same child at least one month apart. Workers did not have access to their originally recorded answers and were unaware that they would be asked to rate the same video a second time when providing the first set of ratings. We observe the mean deviation for all videos between a worker's original ratings for the video and their subsequent ratings one month later. We call this metric the *mean same-child L1 distance (MSCL$_1$)*, which we consider as a metric of the worker's *test-retest reliability*. Higher values for the MSCL$_1$ correspond to greater variation in worker responses when re-rating the same video one month apart. Formally, MSCL$_1$ is calculated as:

$$MSCL_1 = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} |A_{i,j,2} - A_{i,j,1}|}{N}$$

Mean pairwise internal L1 distance (MPIL$_1$): To analyze the reliability of the worker's answers across videos, we look at the mean L1 distance between a worker's answer to each video and all other videos they rated. We call this metric the *mean pairwise internal L1 distance (MPIL$_1$)*. MPIL$_1$ is high when workers provide a wide variety of answer patterns across videos. If the worker answers all questions the same way per video, the MPIL_1 will be 0. Formally, MPIL_1 is calculated as:

$$MPIL_1 = \frac{\sum_{i1=1}^{N} \sum_{i2=1}^{N} \sum_{j=1}^{M} |A_{i2,j} - A_{i1,j}|}{0.5\,N\,(N-1)}, i1 < i2$$

Penalized time (PT): We aimed to build a metric that prioritizes rewarding workers who spent sufficient time rating the first time while rewarding, to a lesser extent, workers who spend sufficient time rating after receiving a warning. We also aimed to penalize workers who either do not spend more time rating after receiving a warning or who do not sufficiently update their answers. We create a metric of worker *trustworthiness* taking both of these factors into account which we call the *penalized time (PT)*. If workers spend longer than a time threshold T rating, then they are not asked to revise their answers and receive a baseline score M. If they do not spend a sufficient time (T) rating, then they are asked to spend more time and to revise their answers. In this case, the metric consists of two terms, balanced by a weighting constant c. The first term is the “revision” mean same-child L1 distance ($RMSCL_1$) between initial and revised answers only for videos that the worker was explicitly asked to revise. The second term is the mean of the total time spent rating, which is the time spent initially (t_1) and the time spent revising the answers (t_2). Formally, PT is calculated as:

$$PT = \begin{cases} M, & t_1 \geq T \\ \frac{t_1 + t_2}{N} + c\,RMSCL_1, & t_1 < T \end{cases}$$

Time spent: Finally, we record the mean amount of time spent rating per video, in seconds. We hypothesized that workers who spend more time on the rating task will tend towards achieving higher performance.

We hypothesized that all four metrics are correlated with PCC. We only calculate metrics for workers who rated at least 10 videos. Because 13 questions were asked, an MSCL_1 or MPIL_1 of 13 means that, on average, the worker’s answer differed by 1 categorical ordinal answer choice per question (e.g., the difference between “*Mixed: some regular echoing of words and phrases, but also some language*” and “*Mostly echoed speech*” in Fig. 1).

2.5. Prediction of crowd worker performance from metrics

We train and test a linear regression model to predict the mean PCC of the workers using 5-fold cross validation. We evaluate all non-empty subsets of the correlative metrics described in section 2.4 as inputs to the model. Since not all workers reopened the task after receiving a warning and not all workers conducted the second task in the series, we evaluated our model both using all available workers with complete data for all metrics as well as using the subset of 55 workers with data for all metrics.

3. Results

3.1. Correlation between metrics and probability of the correct class

Correlations of each of the worker metrics with their mean PCC are displayed in Fig. 4. Mean values per worker are only plotted and analyzed if at least 5 data points are available for the worker. $MSCL_1$, $MPIL_1$, and mean time spent were all significantly correlated with PCC (r=0.31, p=0.0212 for $MSCL_1$; r=0.57, p<0.0001 for $MPIL_1$; r=0.16, p=0.0284 for time), supporting the predictive power of these metrics. Intuitively, this means that higher variability in worker answers for the same video and across videos correlates with increased worker performance. We note that only $MPIL_1$ passes Bonferroni correction. Penalized time was not significantly correlated with PCC (r=0.17, r=0.1413 for penalized time).

Interestingly, Fig. 4 reveals that the presence of enough data to calculate certain metrics is in itself predictive of worker performance. Fig. 4C shows that there are several workers who had a mean PCC below 50%. However, none of these workers appear in the plot for $MSCL_1$ (Fig. 4A), $MPIL_1$ (Fig. 4B), or penalized time (Fig. 4D), indicating that workers with low average performance did not rate videos again after one month and did not revise their answers when prompted.

We evaluate all values of the weighting constant *c* for the penalized time metric in the interval [0.05, 10.0] using a step size of 0.05. No value resulted in a metric that positively correlates with PCC. To investigate, we review the correlation between both terms of penalized time: (1) the mean total time spent rating post-warning and (2) the mean L1 distance between the answer vector before and after the warning (Fig. 5). Neither of these metrics are correlated with PCC (r=−0.10, p=0.3414 for revision L1 distance; r=0.11, p=0.2908 for total time), explaining the inability of the penalized time metric to predict PCC regardless of the parameters chosen.

3.2. Regression prediction of the mean probability of the correct class

Table 1 contains the mean average error (MAE) of a linear regression model predicting the probability of the correct class for each worker using metrics on the same set of videos. There were 55 workers with data for all 3 metrics used in the regression model. For these workers, all metrics predicted the PCC with less than 10% MAE.

The MAE when using all 3 features performs nearly identically, to two decimal places, compared to using only $MSCL_1$ and $MPIL_1$. Mean time does not contribute much predictive power given the other metrics. Interestingly, the most predictive input configuration when using the same 55 workers is $MPIL_1$ together with mean time (6.97% MAE), followed by $MPIL_1$ alone as a close second (6.98% MAE). This is a testament to the success of the $MPIL_1$ metric.

Table 2 contains the mean average error of a linear regression model predicting the probability of the correct class for each worker using metrics from one set of children and mean probability of the correct class calculations for a distinct set of children. The most predictive input feature configuration ($MSCL_1$ and $MPIL_1$) results in a MAE of 10.41%, only 3.44% higher than the best MAE when training and testing on the same set of videos and workers using cross-validation (Table 1). $MPIL_1$ is involved in all of the top-4 input metric configurations resulting in the lowest MAE, again verifying the success of the $MPIL_1$ metric.

4. Discussion and Future Work

We identify three metrics which are individually highly correlated with the mean probability of the worker's categorical behavioral feature tags predicting the correct class. In particular, one of our two reliability metrics - the mean pairwise internal L1 distance, which is the mean L1 distance between a worker's answer to each video and all other videos they rated - stood out as the most predictive metric. Mean pairwise internal L1 distance alone can predict a worker's PCC within 7% MAE when trained on the same set of workers as in the test set but with different videos, and it can predict PCC within 11% MAE when trained on one group of workers and tested on an entirely district set of workers and videos. This metric alone therefore provides a powerful behavioral predictor of worker performance and is therefore likely to be useful for rapidly filtering workers. The positive correlation shown in Fig. 4B suggests that unreliable workers will provide the same or similar patterns of answer sequences for each task. We see that an increasing diversity of answers between tasks results in a higher PCC for the entire spectrum of possible L1 distances. Intuitively, this may be a result of the diverse set of features exhibited by the heterogeneous behavioral characteristics of the children in our dataset.

Interestingly, the raw time metric is not particularly correlative with PCC, indicating that analyzing the answer domain is more informative than the time domain. For workers who received a warning for low time spent, neither the time spent revising post-warning nor the L1 distance between the original and revised set of answers was predictive of the workers' final performance. It is possible that once workers are aware that their time is tracked, they idly keep the rating interface open, accumulating time without accumulating thoughtful work. This hypothesis is speculative, and more fine-grained timing information must be recorded to evaluate such hypotheses.

Future work should evaluate workers on a larger scale, which will validate the preliminary findings of the present study. It is possible that predictive time-based trustworthiness metrics exist. Evaluation on a larger scale in conjunction with more fine-tuned worker metrics will lead to more precise predictions.

5. Conclusion

We demonstrate that behavioral metrics about crowd workers can predict, with a high degree of accuracy, the performance of crowd workers on behavioral feature extraction tasks for the binary diagnosis of autism. Metrics like these can be used for quickly and efficiently identifying crowd workers who are trustworthy and reliable enough for exposure to highly sensitive PHI based on a quantification of their reliability.

Acknowledgments

This work was supported by awards to DPW by the National Institutes of Health (R01EB025025, R01LM013083, and R21HD091500). Additionally, we acknowledge the support of grants to DPW from The Hartwell Foundation, the David and Lucile Packard Foundation Special Projects Grant, Beckman Center for Molecular and Genetic Medicine, Coulter Endowment Translational Research Grant, Berry Fellowship, Spectrum Pilot Program, Stanford's Precision Health and Integrated Diagnostics Center (PHIND), Wu Tsai Neurosciences Institute Neuroscience: Translate Program, and Stanford's Institute of Human Centered Artificial Intelligence as well as philanthropic support from Mr. Peter Sullivan. PW would like to acknowledge support from the Schroeder Family Goldman Sachs Stanford Interdisciplinary Graduate Fellowship (SIGF).

References

1. Kogan Michael D., Vladutiu Catherine J., Schieve Laura A., Ghandour Reem M., Blumberg Stephen J., Zablotsky Benjamin, Perrin James M. et al. "The prevalence of parent-reported autism spectrum disorder among US children." Pediatrics 142, no. 6 (2018).
2. Fombonne Eric. "The rising prevalence of autism." Journal of Child Psychology and Psychiatry 59, no. 7 (2018): 717–720. [PubMed: 29924395]
3. Ning Michael, Daniels Jena, Schwartz Jessey, Dunlap Kaitlyn, Washington Peter, Kalantarian Haik, Du Michael, and Wall Dennis P.. "Identification and quantification of gaps in access to autism resources in the United States: an infodemiological study." Journal of Medical Internet Research 21, no. 7 (2019): e13094. [PubMed: 31293243]
4. Howlin Patricia, and Moore Anna. "Diagnosis in autism: A survey of over 1200 patients in the UK." autism 1, no. 2 (1997): 135–162.
5. Escobedo Lizbeth, Nguyen David H., Boyd LouAnne, Hirano Sen, Rangel Alejandro, Garcia-Rosas Daniel, Tentori Monica, and Hayes Gillian. "MOSOCO: a mobile assistive tool to support children with autism practicing social skills in real-life situations." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2589–2598. 2012.
6. Hashemi Jordan, Campbell Kathleen, Carpenter Kimberly, Harris Adrianne, Qiu Qiang, Tepper Mariano, Espinosa Steven et al. "A scalable app for measuring autism risk behaviors in young children: a technical validity and feasibility study." In Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare, pp. 23–27. 2015.
7. Kalantarian Haik, Jedoui Khaled, Dunlap Kaitlyn, Schwartz Jessey, Washington Peter, Husic Arman, Tariq Qandeel, Ning Michael, Kline Aaron, and Wall Dennis Paul. "The Performance of Emotion Classifiers for Children With Parent-Reported Autism: Quantitative Feasibility Study." JMIR Mental Health 7, no. 4 (2020): e13174. [PubMed: 32234701]
8. Kalantarian Haik, Jedoui Khaled, Washington Peter, and Wall Dennis P.. "A mobile game for automatic emotion-labeling of images." IEEE Transactions on Games (2018).
9. Kalantarian Haik, Jedoui Khaled, Washington Peter, Tariq Qandeel, Dunlap Kaiti, Schwartz Jessey, and Wall Dennis P.. "Labeling images with facial emotion and the potential for pediatric healthcare." Artificial intelligence in medicine 98 (2019): 77–86. [PubMed: 31521254]
10. Kalantarian Haik, Washington Peter, Schwartz Jessey, Daniels Jena, Haber Nick, and Wall Dennis P.. "Guess what?." Journal of Healthcare Informatics Research 3, no. 1 (2019): 43–66. [PubMed: 33313475]
11. Kalantarian Haik, Washington Peter, Schwartz Jessey, Daniels Jena, Haber Nick, and Wall Dennis. "A gamified mobile system for crowdsourcing video for autism research." In 2018 IEEE international conference on healthcare informatics (ICHI), pp. 350–352. IEEE, 2018.
12. Li Wei, Abtahi Farnaz, Tsangouri Christina, and Zhu Zhigang. "Towards an "in-the-wild" emotion dataset using a game-based framework." In 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1526–1534. IEEE, 2016.
13. Boyd LouAnne E., Rangel Alejandro, Tomimbang Helen, Conejo-Toledo Andrea, Patel Kanika, Tentori Monica, and Hayes Gillian R.. "SayWAT: Augmenting face-to-face conversations for adults with autism." In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 4872–4883. 2016.

14. Daniels Jena, Haber Nick, Voss Catalin, Schwartz Jessey, Tamura Serena, Fazel Azar, Kline Aaron et al. "Feasibility testing of a wearable behavioral aid for social learning in children with autism." Applied clinical informatics 9, no. 1 (2018): 129. [PubMed: 29466819]
15. Daniels Jena, Schwartz Jessey N., Voss Catalin, Haber Nick, Fazel Azar, Kline Aaron, Washington Peter, Feinstein Carl, Winograd Terry, and Wall Dennis P.. "Exploratory study examining the at-home feasibility of a wearable tool for social-affective learning in children with autism." NPJ digital medicine 1, no. 1 (2018): 1–10. [PubMed: 31304287]
16. Kaliouby El, Rana, and Robinson Peter. "The emotional hearing aid: an assistive tool for children with Asperger syndrome." Universal Access in the Information Society 4, no. 2 (2005): 121–134.
17. Haber Nick, Voss Catalin, and Wall Dennis. "Making emotions transparent: Google Glass helps autistic kids understand facial expressions through augmented-reaiity therapy." IEEE Spectrum 57, no. 4 (2020): 46–52.
18. Kline Aaron, Voss Catalin, Washington Peter, Haber Nick, Schwartz Hessey, Tariq Qandeel, Winograd Terry, Feinstein Carl, and Wall Dennis P.. "Superpower glass." GetMobile: Mobile Computing and Communications 23, no. 2 (2019): 35–38.
19. Madsen Miriam, Kaliouby Rana El, Goodwin Matthew, and Picard Rosalind. "Technology for just-in-time in-situ learning of facial affect for persons diagnosed with an autism spectrum disorder." In Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility, pp. 19–26. 2008.
20. Nyström Pär, Thorup Emilia, Bölte Sven, and Falck-Ytter Terje. "Joint attention in infancy and the emergence of autism." Biological psychiatry 86, no. 8 (2019): 631–638. [PubMed: 31262432]
21. Ravindran Vijay, Osgood Monica, Sazawal Vibha, Solorzano Rita, and Turnacioglu Sinan. "Virtual reality support for joint attention using the Floreo Joint Attention Module: Usability and feasibility pilot study." JMIR pediatrics and parenting 2, no. 2 (2019): e14429. [PubMed: 31573921]
22. Strobl Maximilian AR, Lipsmeier Florian, Demenescu Liliana R., Gossens Christian, Lindemann Michael, and De Vos Maarten. "Look me in the eye: evaluating the accuracy of smartphone-based eye tracking for potential application in autism spectrum disorder research." Biomedical engineering online 18, no. 1 (2019): 1–12. [PubMed: 30602383]
23. Voss Catalin, Haber Nick, and Wall Dennis P.. "The Potential for Machine Learning–Based Wearables to Improve Socialization in Teenagers and Adults With Autism Spectrum Disorder—Reply." Jama Pediatrics 173, no. 11 (2019): 1106–1106.
24. Voss Catalin, Schwartz Jessey, Daniels Jena, Kline Aaron, Haber Nick, Washington Peter, Tariq Qandeel et al. "Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial." JAMA pediatrics 173, no. 5 (2019): 446–454. [PubMed: 30907929]
25. Voss Catalin, Washington Peter, Haber Nick, Kline Aaron, Daniels Jena, Fazel Azar, De Titas et al. "Superpower glass: delivering unobtrusive real-time social cues in wearable systems." In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, pp. 1218–1226. 2016.
26. Washington Peter, Voss Catalin, Haber Nick, Tanaka Serena, Daniels Jena, Feinstein Carl, Winograd Terry, and Wall Dennis. "A wearable social interaction aid for children with autism." In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 2348–2354. 2016.
27. Washington Peter, Voss Catalin, Kline Aaron, Haber Nick, Daniels Jena, Fazel Azar, De Titas, Feinstein Carl, Winograd Terry, and Wall Dennis. "SuperpowerGlass: a wearable aid for the at-home therapy of children with autism." Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies 1, no. 3 (2017): 1–22.
28. Nag Anish, Haber Nick, Voss Catalin, Tamura Serena, Daniels Jena, Ma Jeffrey, Chiang Bryan et al. "Toward Continuous Social Phenotyping: Analyzing Gaze Patterns in an Emotion Recognition Task for Children With Autism Through Wearable Smart Glasses." Journal of Medical Internet Research 22, no. 4 (2020): e13810. [PubMed: 32319961]
29. Washington Peter, Park Natalie, Srivastava Parishkrita, Voss Catalin, Kline Aaron, Varma Maya, Tariq Qandeel et al. "Data-driven diagnostics and the potential of mobile artificial intelligence for digital therapeutic phenotyping in computational psychiatry." Biological Psychiatry: Cognitive Neuroscience and Neuroimaging (2019).

30. Paolacci Gabriele, Chandler Jesse, and Ipeirotis Panagiotis G.. "Running experiments on amazon mechanical turk." Judgment and Decision making 5, no. 5 (2010): 411–419.

31. Washington Peter, Leblanc Emilie, Dunlap Kaitlyn, Penev Yordan, Kline Aaron, Paskov Kelley, Sun Min Woo et al. "Precision Telemedicine through Crowdsourced Machine Learning: Testing Variability of Crowd Workers for Video-Based Autism Feature Recognition." Journal of personalized medicine 10, no. 3 (2020): 86.

32. Washington Peter, Kalantarian Haik, Tariq Qandeel, Schwartz Jessey, Dunlap Kaitlyn, Chrisman Brianna, Varma Maya et al. "Validity of online screening for autism: crowdsourcing study comparing paid and unpaid diagnostic tasks." Journal of medical Internet research 21, no. 5 (2019): e13668. [PubMed: 31124463]

33. Hirth Matthias, Hoßfeld Tobias, and Tran-Gia Phuoc. "Anatomy of a crowdsourcing platform-using the example of microworkers. com." In 2011 Fifth international conference on innovative mobile and internet services in ubiquitous computing, pp. 322–329. IEEE, 2011.

34. Ipeirotis Panagiotis G. "Analyzing the amazon mechanical turk marketplace." XRDS: Crossroads, The ACM Magazine for Students 17, no. 2 (2010): 16–21.

35. Abbas Halim, Garberson Ford, Liu-Mayo Stuart, Glover Eric, and Wall Dennis P.. "Multi-modular Ai Approach to Streamline Autism Diagnosis in Young children." Scientific reports 10, no. 1 (2020): 1–8. [PubMed: 31913322]

36. Abbas Halim, Garberson Ford, Glover Eric, and Wall Dennis P.. "Machine learning approach for early detection of autism by combining questionnaire and home video screening." Journal of the American Medical Informatics Association 25, no. 8 (2018): 1000–1007. [PubMed: 29741630]

37. Duda Marlena, Daniels Jena, and Wall Dennis P.. "Clinical evaluation of a novel and mobile autism risk assessment." Journal of autism and developmental disorders 46, no. 6 (2016): 1953–19 [PubMed: 26873142]

38. Duda M, Haber N, Daniels J, and Wall DP. "Crowdsourced validation of a machine-learning classification system for autism and ADHD." Translational psychiatry 7, no. 5 (2017): e1133–e1133. [PubMed: 28509905]

39. Duda M, Kosmicki JA, and Wall DP. "Testing the accuracy of an observation-based classifier for rapid detection of autism risk." Translational psychiatry 4, no. 8 (2014): e424–e424. [PubMed: 25116834]

40. Duda M, Ma R, Haber N, and Wall DP. "Use of machine learning for behavioral distinction of autism and ADHD." Translational psychiatry 6, no. 2 (2016): e732–e732. [PubMed: 26859815]

41. Paskov Kelley M., and Wall Dennis P.. "A low rank model for phenotype imputation in autism spectrum disorder." AMIA Summits on Translational Science Proceedings 2018 (2018): 178.

42. Wall Dennis P., Dally Rebecca, Luyster Rhiannon, Jung Jae-Yoon, and DeLuca Todd F.. "Use of artificial intelligence to shorten the behavioral diagnosis of autism." PloS one 7, no. 8 (2012): e43855. [PubMed: 22952789]

43. Wall Dennis Paul, Kosmicki J, Deluca TF, Harstad E, and Fusaro Vincent Alfred. "Use of machine learning to shorten observation-based screening and diagnosis of autism." Translational psychiatry 2, no. 4 (2012): e100–e100. [PubMed: 22832900]

44. Washington Peter, Paskov Kelley Marie, Kalantarian Haik, Stockham Nathaniel, Voss Catalin, Kline Aaron, Patnaik Ritik et al. "Feature selection and dimension reduction of social autism data." In Pac Symp Biocomput, vol. 25, pp. 707–718. 2020. [PubMed: 31797640]

45. Deyo Richard A., Diehr Paula, and Patrick Donald L.. "Reproducibility and responsiveness of health status measures statistics and strategies for evaluation." Controlled clinical trials 12, no. 4 (1991): S142–S158.

46. Paiva Carlos Eduardo, Barroso Eliane Marçon, Carneseca Estela Cristina, de Pádua Souza Cristiano, Dos Santos Felipe Thomé, López Rossana Verónica Mendoza, and Paiva Sakamoto Bianca Ribeiro. "A critical analysis of test-retest reliability in instrument validation studies of cancer patients under palliative care: a systematic review." BMC medical research methodology 14, no. 1 (2014): 8. [PubMed: 24447633]

47. Polit Denise F. "Getting serious about test–retest reliability: a critique of retest research and some recommendations." Quality of Life Research 23, no. 6 (2014): 1713–1720. [PubMed: 24504622]

48. Vilagut Gemma. "Test-retest reliability." Encyclopedia of quality of life and well-being research (2014): 6622–6625.
49. Bone Daniel, Goodwin Matthew S., Black Matthew P., Lee Chi-Chun, Audhkhasi Kartik, and Narayanan Shrikanth. "Applying machine learning to facilitate autism diagnostics: pitfalls and promises." Journal of autism and developmental disorders 45, no. 5 (2015): 1121–1136. [PubMed: 25294649]
50. Fusaro Vincent A., Daniels Jena, Duda Marlena, DeLuca Todd F., D'Angelo Olivia, Tamburello Jenna, Maniscalco James, and Wall Dennis P.. "The potential of accelerating early detection of autism through content analysis of YouTube videos." PLOS one 9, no. 4 (2014): e93533. [PubMed: 24740236]
51. Kosmicki JA, Sochat V, Duda M, and Wall DP. "Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning." Translational psychiatry 5, no. 2 (2015): e514–e514. [PubMed: 25710120]
52. Levy Sebastien, Duda Marlena, Haber Nick, and Wall Dennis P.. "Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism." Molecular autism 8, no. 1 (2017): 65. [PubMed: 29270283]
53. Tariq Qandeel, Daniels Jena, Schwartz Jessey Nicole, Washington Peter, Kalantarian Haik, and Wall Dennis Paul. "Mobile detection of autism through machine learning on home video: A development and prospective validation study." PLoS medicine 15, no. 11 (2018): e1002705. [PubMed: 30481180]
54. Tariq Qandeel, Fleming Scott Lanyon, Schwartz Jessey Nicole, Dunlap Kaitlyn, Corbin Conor, Washington Peter, Kalantarian Haik, Khan Naila Z., Darmstadt Gary L., and Wall Dennis Paul. "Detecting developmental delay and autism through machine learning models using home videos of Bangladeshi children: Development and validation study." Journal of medical Internet research 21, no. 4 (2019): e13822. [PubMed: 31017583]

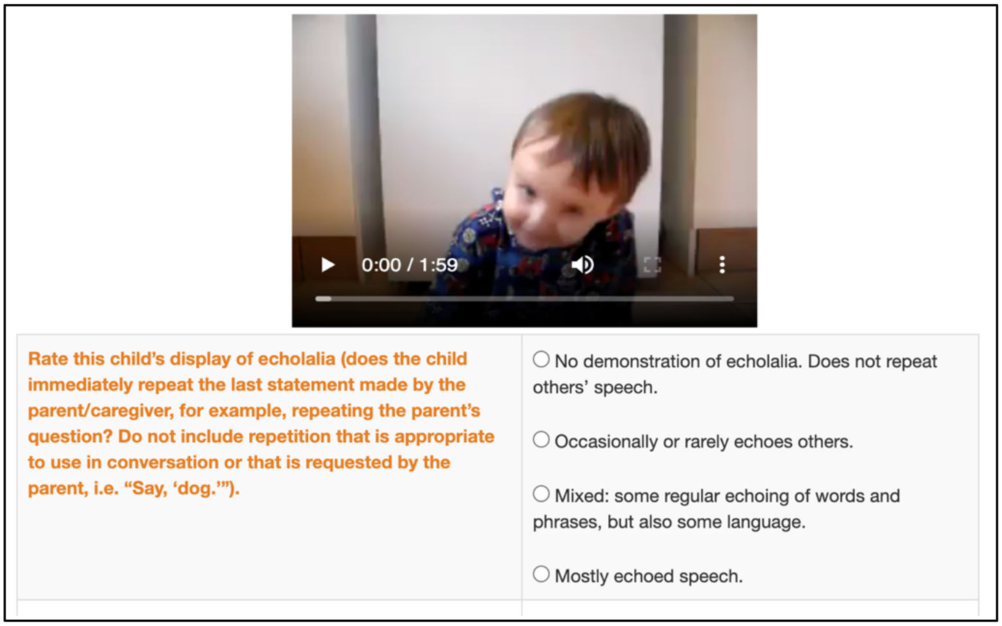

Fig. 1.

Crowd worker feature tagging user interface deployed on Microworkers.com. Each worker answered a series of multiple-choice questions corresponding to each input feature of a gold standard classifier.

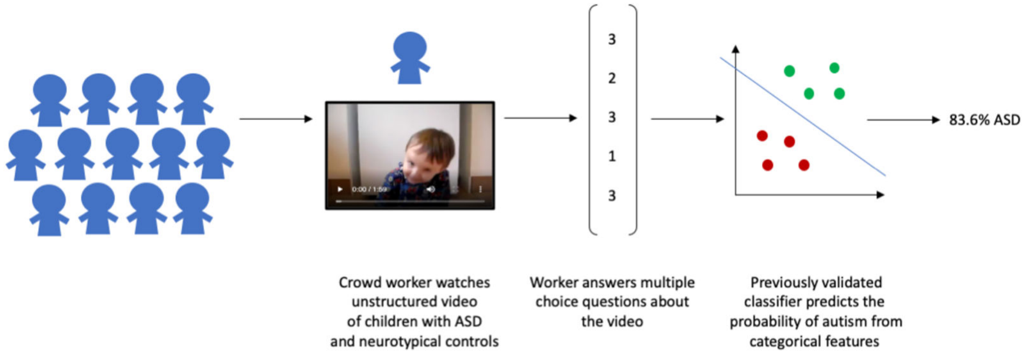

Fig. 2.

Process for collecting the data needed to evaluate trust and reliability metrics for crowd workers. Each crowd worker watches unstructured videos of children with autism and neurotypical controls, answering multiple choice questions about each video. These multiple-choice answers serve as categorical ordinal feature vectors for a previously validated logistic regression classifier, trained on clinician-filled electronic health records, that predicts the probability that a child has autism.

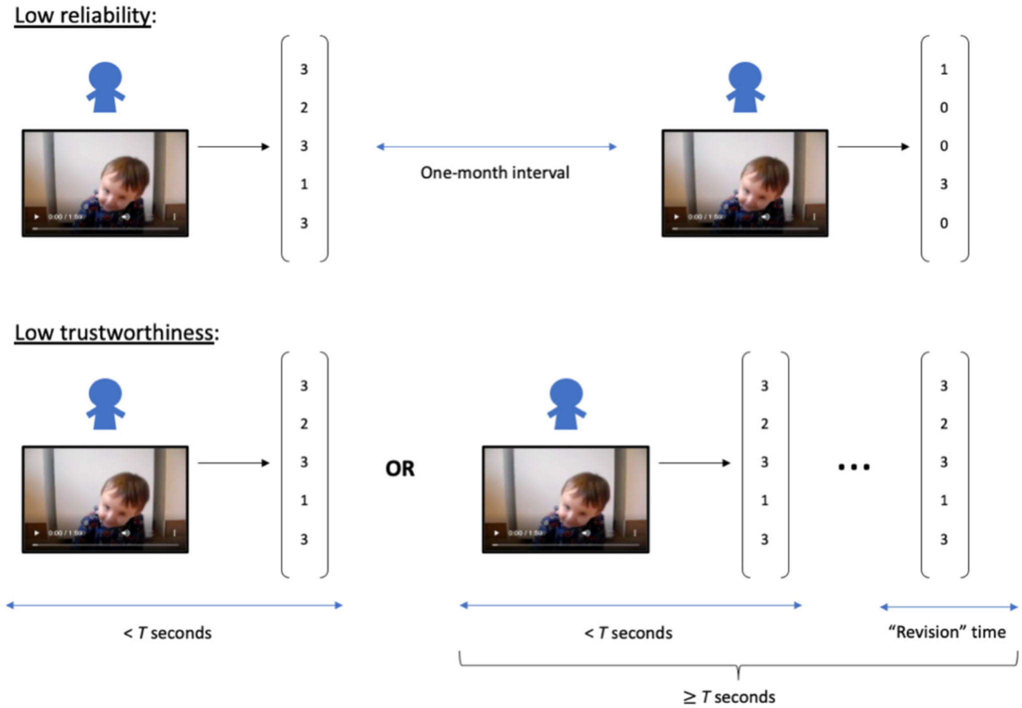

Fig. 3.

Process for calculating trust and reliability metrics for crowd workers. The reliability of workers is determined by how different their answers are when rating the same video one month apart. The trustworthiness of workers is determined by whether they spend the minimal amount of time needed to properly answer the questions, whether they spend sufficient time when receiving a warning, and whether their original answers change after receiving the warning.

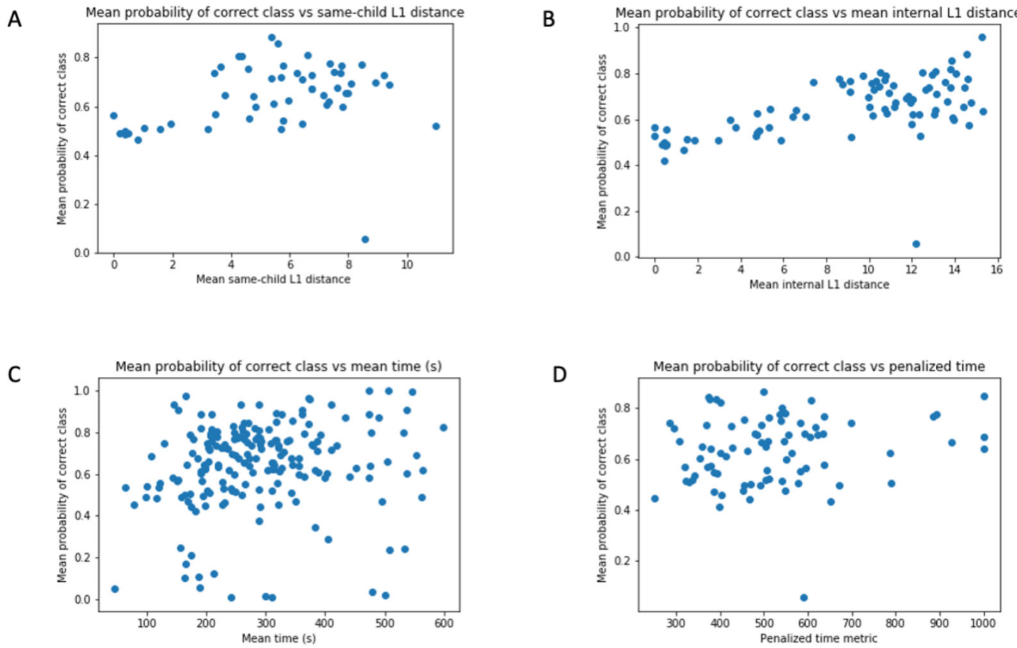

Fig. 4.

Correlations between metrics and probability of the correct class (PCC). (A) Correlation between mean same-child L1 distance and PCC. (B) Correlation between mean pairwise internal L1 distance and PCC. (C) Correlation between time spent (s) and PCC. (D) Lack of correlation between penalized time and PCC.

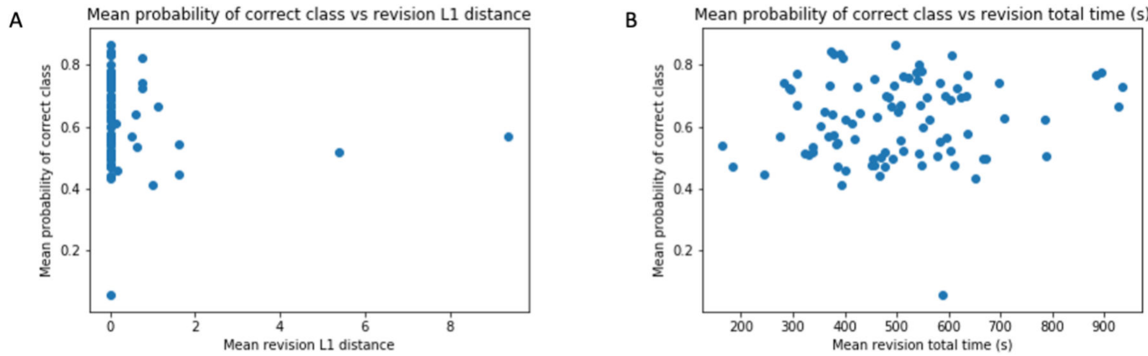

Fig. 5.

Lack of correlation between PCC and (A) the total time spent rating post-warning and (B) the L1 distance between the answer before and after the warning.

Table 1.

5-fold cross validated mean average error (MAE) of a linear regression model predicting the probability of the correct class for each worker using metrics on the same set of videos.

Input Features	5-fold MAE (%) All data points	5-fold MAE (%) 55 workers with all metric data	N
$MSCL_1$, $MPIL_1$, mean time	7.51	7.51	55
$MSCL_1$, mean time	8.89	8.89	55
$MPIL_1$, mean time	7.43	6.97	81
$MSCL_1$, $MPIL_1$	7.51	7.51	55
$MSCL_1$	9.24	9.24	55
$MPIL_1$	7.39	6.98	81
Mean time	15.56	9.83	193

Table 2.

Mean average error (MAE) of the linear regression model predicting the probability of the correct class for each worker using the same metric data and resulting classifier weights *for the workers and videos used in* Table 1 and mean probability of the correct class calculations for a *distinct set of videos* for a *distinct set of workers*.

Input Features	MAE (%) All data points
$MSCL_1$, $MPIL_1$, mean time	10.93
$MSCL_1$, mean time	13.03
$MPIL_1$, mean time	11.50
$MSCL_1$, $MPIL_1$	10.41
$MSCL_1$	11.87
$MPIL_1$	10.91
Mean time *	12.10

* Mean time as the only feature is the only configuration of input features that requires a different set of data points: N=102 instead of a subset of size N=62 for all other configurations.