Abstention and Threshold Identification for Uncertainty Management in Clinical Decision Tools: A Case Study using Human-In-The-Loop Pediatric Autism Classifiers

Aiden Ko, Aaron Kline, Kaitlyn Dunlap, SaiMourya Surabhi, Parnian Azizian
Department of Pediatrics (Clinical Informatics), Stanford University, Stanford, CA 94305, USA
Emails: aidensko@stanford.edu, akline@stanford.edu, Kaiti.dunlap@stanford.edu, mourya@stanford.edu, azizian@stanford.edu
Peter Y. Washington
Division of Clinical Informatics and Digital Transformation, Department of Medicine, University of California, San Francisco (UCSF), San Francisco, CA 94143, USA
Email: peter.washington@ucsf.edu
Dennis P. Wall
Departments of Pediatrics (Clinical Informatics), Biomedical Data Science, and Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305, USA
Email: dpwall@stanford.edu

Uncertainty quantification remains an underdeveloped aspect of AI-based clinical decision tools. As AI systems become increasingly prevalent in healthcare, it is essential not only to measure uncertainty but also to manage it in ways that support clinical decision-making. In this study, we investigate abstention as a practical mechanism for managing uncertainty in diagnostic classifiers. To stress-test this approach, we deliberately evaluate abstention performance on a purposefully noisy dataset of pediatric autism video assessments comprising heterogeneous video sources and a diverse range of human raters. We apply abstention strategies to existing autism classifiers trained on diagnostic assessment data, comparing baseline performance to a range of thresholding configurations that trade off retained sample coverage against key clinical metrics. We compare performance gains from prioritizing sensitivity or specificity to targeting a balanced increase in Youden's J to demonstrate a wide variety of use cases that abstention can enable. This work demonstrates a concrete use case of introducing abstention into the output range of clinical decision models, enabling both uncertainty quantification and management in diagnostic classifiers.

Keywords: Uncertainty Management; Human-In-The-Loop AI; Autism; Pediatrics;AI health; AI medicine; AI uncertainty

1. Introduction

Autism is a developmental disorder that is diagnosed by clinician observation and behavioral assessment. Current diagnostic approaches for autism rely on specialists who are outnumbered,[1] use myriad diagnostic aids, e.g. Autism Diagnostic Observation Schedule,[2] Autism Diagnostic Interview,[3] Social Responsiveness Scale,[4] with over 20 clinically recommended tools[5] that are time intensive (up to 8 hours of test time), subjective,[6, 7] disagree on outcome,[7] and that do not produce a numerical score that has consistent performance and reliability across the 1 in 31[8] who will be diagnosed with autism. The wait to access diagnostic assessments exceeds 12 months.[9] This delayed process has kept the average age of first diagnosis close to 5 years[10] (and 8 if from an underrepresented group[11]). Evidence supports the importance of early intervention. Like diagnosis, wait periods to access treatments often exceed 12 months,[12, 13] and are not quality monitored or modified with objective endpoints.

The difficulties navigating the healthcare system, variability in diagnostic practices, and consequent delays in diagnosis strongly underscores the need for an alternative system for diagnosis that is objective, quantified, reproducible, and consistent across geographies and demographics. Previous work has shown that machine learning models can effectively classify autism from neurotypical development when using an optimal feature set from clinical instrument data; our team previously trained models using data from 11,298 children with autism and 2,540 controls (including attention-deficit/hyperactivity disorder (ADHD), speech and language delays, and typical development) that resulted in 8 AI classifiers that use between 5 and 10 social-behavioral features each.[14-18] Together, the models use 30 different social-behavioral features. We have since tested the ability to run these models using human-labeled feature ratings on videos submitted by parents. Experiments show that video feature tagging by video analyst “crowdworkers”[19-28] requires a median of only 4 minutes and can be done with clinical accuracy >=92% in independent validation.[29-31,32] While the models were trained to produce binary outcomes, they produce point estimates that indicate the severity of the phenotype and range from severe forms of autism to typical development with no signs of developmental delay.

We seek to build upon this body of work by exploring the concept of uncertainty quantification and management in AI-powered diagnostic tools, with a specific emphasis on abstention-based thresholding as a way to make AI predictions more clinically interpretable and trustworthy. In clinical contexts like autism diagnosis, where behavioral variation is high, the incoming video quality can vary, symptom presentations may be complex, and even high-performing models may not perform at a level needed for clinical decision support. Forcing a model to make a binary decision in such cases can lead to overdiagnosis or missed opportunities for early intervention. This highlights the need for tools that can identify when they are uncertain and explicitly defer judgment, a concept known as abstention. Abstention allows models to say “I don’t know” when presented with low-confidence inputs, enabling a safety-aware approach that aligns with clinical practice, where ambiguous cases are often escalated for further review.

This approach is especially relevant in behavioral health domains where ground truth is often fuzzy and diagnostic boundaries overlap. When implemented via thresholding on a model's probabilistic outputs, abstention can be tuned to meet the needs of a particular use case, for example by maximizing sensitivity for screening or maximizing Positive Predictive Value (PPV) for referral decisions. Despite its importance, abstention remains underutilized in diagnostic AI tools, and its integration with video-based autism assessment has not been systematically studied.

One prominent exception is CanvasDx, a machine learning-based medical device approved by the U.S. Food and Drug Administration (FDA) for autism diagnosis. This software device uses video submitted by a caregiver and human-provided feature ratings to produce a diagnostic classification using machine learning.[33, 34] As AI-informed clinical diagnostics are developed, it is critical to improve the quantification and management of uncertainty for these diagnostic tools.[35, 36] One key mechanism for improving these is adding an abstention mechanism with clinical performance-informed threshold values. CanvasDx improved its management of uncertainty using thresholding scores where if this score is above or below two prespecified thresholds, a positive autism or negative autism output is produced.[37, 38]

The focus of the current paper is to examine how we can apply such thresholding methods to a naturalistic but noisy set of video and ratings data that utilizes videos sourced from multiple settings (clinical, mobile game, and home, sourced from the U.S. and abroad) and ratings from raters of different backgrounds (cultural, geographic, professional). By testing performance on this larger dataset containing wildtype videos and nonexpert human ratings (as reflected in real-world scenarios), we perform a comprehensive evaluation of abstention as an uncertainty management mechanism and describe the tradeoffs between diagnostic coverage and performance on clinically informative performance metrics.

2. Methods

2.1 *Noisy Dataset Creation*

We sought to collect, aggregate, and curate a purposefully noisy and heterogeneous dataset as a representative but inherently challenging collection in which abstention for certain samples may be warranted. This combined dataset was collected through a series of studies conducted following approval by the Stanford University Institutional Review Board (IRB), protocol #39562. The data include videos submitted by parents/guardians of children with and without a diagnosis of autism via several different studies conducted by our team (N=665 videos) as well as videos sourced via Youtube.com (N=366 videos). Most videos in the dataset originated from smartphone videos taken by a parent of their child playing alone or with others. For these videos, parents were instructed to take short 1.5–5-minute videos of their children playing. Smaller subsets of videos were taken by clinical coordinators or behavioral therapists, following parent consent, of children waiting for therapy in a medical research center (N=217). Some of these clinically sourced videos are from Bangladesh.[31] Another subset of videos were collected via a mobile game app[39-41] where parents are asked to play a charades style game with their child, while the phones front camera records the child acting out a prompt presented with a corresponding image (i.e., elephant, astronaut, playing basketball) (N=380). Our video dataset contained children who have been clinically confirmed to have a diagnosis of autism, another condition, such as speech language condition (SLC), or neurotypical development. For the purposes of this study, we classified videos as depicting a child with autism or a child without autism. Demographics including age and gender provided by the parent and/or clinician, and for Youtube videos, video meta data and descriptions were used to assign age and gender.

To safeguard against the potential for self-reporting bias in parent submitted video, we required the caregiver to confirm that their child's autism diagnosis came from a formal medical assessment[42] and asked parents to provide answers to a set of clinical instruments that included the Vineland Adaptive Behavior Scales-Parent Report,[43] Social Responsiveness Scale,[4] Mobile Autism Risk Assessment,[44] Social Communication Questionnaire,[45] and Autism Symptom Dimensions Questionnaire[46] to establish the autism diagnosis or lack thereof. Additionally, videos from YouTube and Caregivers were independently examined in a blinded fashion by a clinician with expertise in autism diagnosis to provide a clinical diagnostic impression.[22, 47] We required the parent reported diagnosis to match the clinician's observational diagnosis to include the video for analysis. This rigorous two-point consensus was used to ensure diagnostic ground truth while minimizing parental burden in the participation.

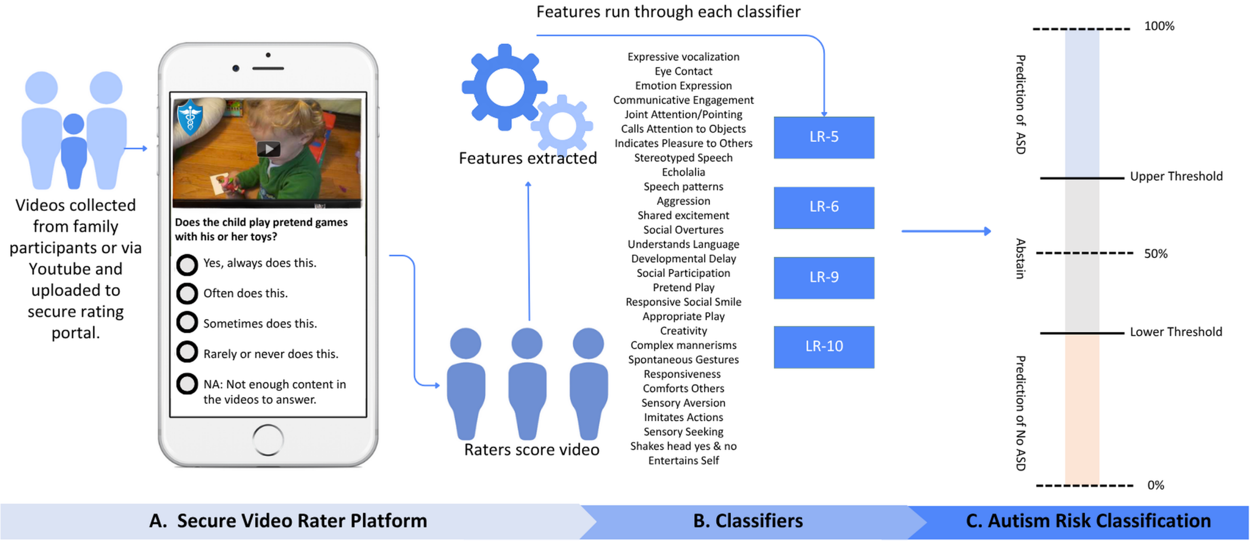

Figure 1. (A) Parents consented to participate in a study where they submitted short home videos of their child to be rated by human labelers in a secure rating platform. Video raters accessed a secure rating platform where 30 behavioral features were rated on a 4-point scale. (B) These features were then used in the following analysis by our previously developed machine learning classifiers to determine an autism risk classification. (C) The current study examines how we can determine and utilize thresholding values to tell the machine learning models to abstain when a certain level of uncertainty is reached.

Feature ratings were collected via a process we previously designed and evaluated (**Figure 1**) where raters recruited by our research lab under the same IRB protocol were asked to view a set of videos in a secure portal and rate a set of 30 features (i.e., answers to multiple choice questions that are clinically relevant to diagnosing autism).[26, 30] Video raters composed of clinicians (N=56 raters and N=3,025 sets of 30-feature ratings) and non-clinician students and crowdworkers (N=50 raters and N=5119 sets of 30 feature ratings) provided numerically coded responses to each of the 30 questions necessary to create the feature vectors for input into the models.

2.2 *Thresholding Experiments*

We leveraged an established and previously-validated set of binary logistic regression (LR) autism classifiers that have previously demonstrated high performance in more homogenous video and rater settings.[26, 30, 31] These models were trained on medical records generated through the clinical administration of popular autism instruments, with training feature sets ranging in size from 30 to over 100. Each identified a specific subset of behavioral features particularly salient to their performance. Given this, we use a short form notation to indicate the number of features for each of the models used in our evaluations, namely LR5, LR6, LR9 and LR10.

We then supplied the ratings from the dataset described above as inputs for inference to the models, taking the median of all ratings available for each video, due to the crowdsourced nature of the ratings and the potential for outliers. The results serve as baseline performance for the thresholding experiments, representing the initial metrics based on a default single 0.5 threshold.

We then introduced an abstention window by replacing that single threshold with dual thresholds t_L and t_U. For each prediction p_i that the model makes, if $p_i < t_L$, we consider it a negative diagnosis, whereas if $p_i > t_U$, we consider it a positive diagnosis. If p_i instead falls on or between the two thresholds, we consider it an abstention. In turn, we can reevaluate each model with different threshold pairs to demonstrate the modified performance alongside an abstention rate (= abstained cases/total cases).

We first measure baseline performance on our dataset using the logistic regression models. Then, to better understand how abstention might improve clinical performance, we conducted a comprehensive grid search over possible upper and lower decision thresholds for each model. Rather than fixing a single decision boundary (e.g., 0.5), we examined the full landscape of trade-offs between diagnostically-relevant performance metrics (sensitivity, specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV)) and the rate at which the model abstains from making a prediction. In the grid search, we set the minimum value as the lowest predicted probability (p_{min}) and the maximum value as the highest (p_{max}). We utilized 2000 evenly spaced steps between p_{min} and p_{max} to iterate through all pairs of t_L and t_U along those steps, generating millions of threshold combinations. For each valid combination (where $t_L < t_U$), we calculated the same performance metrics as before, as well as the abstention rate.

This enables us to ask, for example: if we want a classifier that achieves 80% sensitivity and 80% specificity, what is the minimum abstention rate required to meet that target? Or alternatively: if we can tolerate an abstention rate of up to 25%, what is the best performance we can achieve? By evaluating these scenarios across all valid threshold pairs, we create a flexible framework for tailoring model behavior to different clinical priorities.

3. Results

3.1 *Dataset Compilation*

In the dataset, we identified a total of 8144 sets of ratings across 1031 unique children after preprocessing for videos deemed ‘unscorable’ by raters and duplicate entries. Of those, 588 children had a confirmed positive ASD diagnosis while 443 did not, including those who were neurotypical or had other diagnoses that were not ASD (**Table 1**). Several children in our dataset had video ratings from multiple independent raters (**Figure 2**).

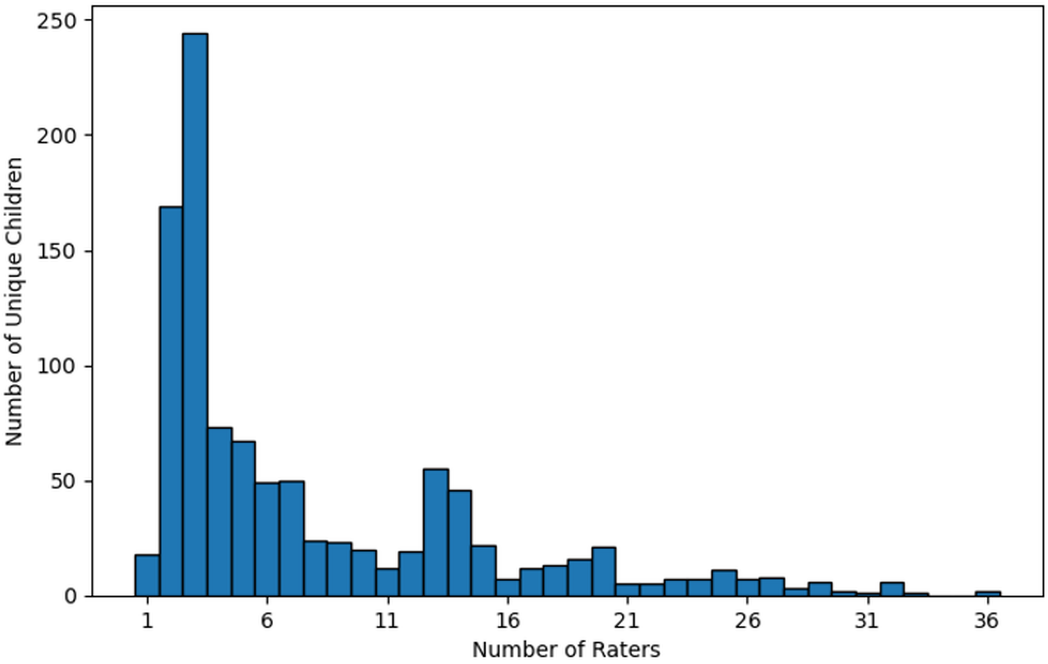

Figure 2. Frequency distribution of the number of raters per child, where the majority of children were given at least three sets of ratings.

Table 1. Demographic characteristics of the children in the video ratings utilized in the analysis.

	Number of Children	**Average Age (Months)**	**Gender Distribution**
Autism	588	65.07 (SD=29.77)	M: 456; F: 132
Non-Autism	443 (38% with other DD)	60.04 (SD=123.14)	M: 257; F:186
Total	1031	62.91 (SD=83.8)	M: 713; F: 318

3.2 *Baseline Classification*

As seen in **Table 2,** accuracies for our baseline classification ranged from 62% to 72%, with LR9 achieving the highest accuracy (71.87%) and sensitivity (82.31%) but lower specificity (58.01%). The other models demonstrated more balanced sensitivity and specificity, though at slightly lower overall accuracy. These results suggest that while our baseline classifiers perform reasonably well given the noisiness (by design) of the video data and human ratings, the addition of abstention may boost performance into more clinically acceptable levels.

Table 2. Baseline performance of logistic regression binary classifiers on the curated collection of 1031 naturalistic video samples.

Model	Accuracy	Sensitivity	Specificity	PPV	NPV
LR5	62.08%	56.12%	69.98%	71.27%	54.58%
LR6	68.57%	66.50%	71.33%	75.48%	61.60%
LR9	71.87%	82.31%	58.01%	72.24%	71.19%
LR10	66.92%	64.63%	69.98%	74.07%	59.85%

3.3 *Thresholding Experiments*

First, we ran a preliminary empirical scan of the data in which we moved the thresholds incrementally further apart for all 4 models. This showed that the general behavior remained consistent across models and further showed for all 4 models a near linear relationship between the threshold width ($t_U – t_L$) and the abstention rate, indicating that widening the distance, and consequently decreasing coverage, expectedly improves performance metrics. We then chose LR6 for the additional in-depth evaluation steps described below due to the fact that it showed the greatest equivalence between sensitivity and specificity, likely offering more room and opportunity to study the impact of uncertainty management on performance (while LR9 showed the highest baseline accuracy (Table 2), its performance favors positives (given the high sensitivity and comparably low specificity), making it potentially less balanced overall).

We evaluated LR6 over a grid of performance targets. For each metric pair (Sensitivity/Specificity and NPV/PPV), we varied the target values from 0.5 to 1.0 in increments of 0.0025, calculating the minimum abstention rate required to simultaneously satisfy both target constraints at each point. Subsequently, we calculated a maximum performance we can achieve under a clinically viable abstention rate. We choose Youden's J (*sensitivity* + *specificity* - 1) as a balanced metric for performance for each set of thresholds and iterate through a set of abstention ceilings, finding the set of thresholds that achieved the maximum Youden's J. Additionally, we set a sensitivity requirement and measure the effect of the ceiling on the best possible specificity and vice versa.

Table 3. LR6 model performance with optimized thresholds (t_L,t_U) under varying abstention ceilings, searching for the highest Youden's J.

Abstain ceiling	Accuracy (%)	Sensitivity (%)	Specificity (%)	Youden's J	t_L	t_U
Baseline	68.57	66.50	71.33	.3783	N/A	N/A
0.1	71.77	71.37	72.28	.4365	0.4027	0.5480
0.2	73.58	69.89	78.33	.4822	0.3685	0.6923
0.3	76.31	74.45	78.78	.5323	0.2966	0.7530
0.4	79.64	85.87	70.52	.5639	0.1944	0.7060
0.5	81.59	86.38	74.88	.6126	0.1645	0.8459

Table 4. LR6 model performance with optimized thresholds (t_L,t_U) under varying abstention ceilings and a baseline condition for sensitivity/specificity while searching to maximize the other.

Abstain ceiling	Additional Condition	Accuracy (%)	Sensitivity (%)	Specificity (%)	t_L	t_U
Baseline	N/A	68.57	66.50	71.33	N/A	N/A
0.1	Sensitivity>=71.33	71.58	71.70	71.43	0.3944	0.5382
0.2	Sensitivity>=71.33	73.70	75.43	71.47	0.3235	0.5867
0.3	Sensitivity>=71.33	75.90	79.15	71.33	0.2384	0.6532
0.4	Sensitivity>=71.33	78.04	82.51	71.43	0.2032	0.7060
0.5	Sensitivity>=71.33	81.48	88.05	71.43	0.1640	0.7593
0.1	Specificity>=66.50	70.72	66.73	75.79	0.4619	0.5969
0.2	Specificity>=66.50	72.92	66.60	81.04	0.4155	0.7354
0.3	Specificity>=66.50	74.52	67.26	83.08	0.3524	0.8400
0.4	Specificity>=66.50	75.08	66.56	84.54	0.3206	0.8841

Abstain ceiling	Additional Condition	Accuracy (%)	Sensitivity (%)	Specificity (%)	t_L	t_U
Baseline	N/A	68.57	66.50	71.33	N/A	N/A
0.1	Sensitivity>=71.33	71.58	71.70	71.43	0.3944	0.5382
0.2	Sensitivity>=71.33	73.70	75.43	71.47	0.3235	0.5867
0.3	Sensitivity>=71.33	75.90	79.15	71.33	0.2384	0.6532
0.4	Sensitivity>=71.33	78.04	82.51	71.43	0.2032	0.7060
0.5	Specificity>=66.50	77.76	66.67	89.58	0.2643	0.9320

As displayed in **Table 3**, increasing the maximum tolerated abstention rate allows the model to achieve a significantly higher optimal performance. The maximum Youden's J consistently improves with higher ceilings, and each ceiling shows a significant increase in performance over the baseline in terms of sensitivity and specificity. We can also see in **Table 4** how setting a baseline for sensitivity or specificity and maximizing the other displays monotonic growth as the abstention ceiling is raised. Moreover, a particularly compelling point can be found at the 30% ceiling in the Youden's J optimization, as this level of data coverage would make it viable in clinical contexts while also displaying a clear performance benefit: both sensitivity and specificity are enhanced, rising to 74.5% and 78.8% respectively, compared to the baseline's 66.5% and 71.3%.

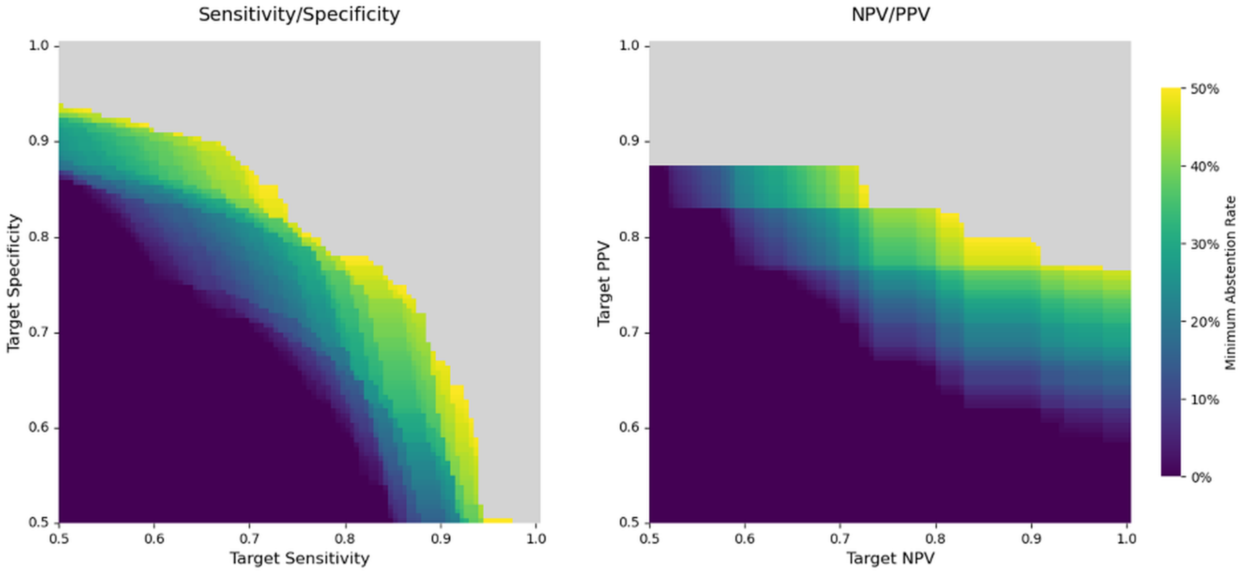

Figure 3. Performance trade-off landscapes for the LR6 model, showing achievable targets at an abstention rate of 50% or less. The grey area represents target combinations that are either impossible to achieve or require an abstention rate greater than 50%.

We also plot the results of our model evaluation over ranges of performance targets in **Figure 3.** This illustrates a promising result, displaying the minimum abstention rate required to simultaneously achieve target performance levels for two key pairs of metrics: sensitivity/specificity (left) and NPV/PPV (right). While some targets are not feasible with any threshold pair at all, we also intentionally drop any data points where the minimum required abstention is greater than 50% as a reasonable ceiling for a scenario where a model would be clinically viable.

Figure 3 demonstrates that the model's performance can be significantly improved over the baseline performance in **Table 2** by allowing it to abstain from making predictions on low-confidence samples. For instance, in the left panel, achieving an approximate balanced 80% sensitivity and 80% specificity requires an abstention rate of approximately 20-30%. Pushing for higher targets, such as 90% in both metrics, drastically pushes the abstention rate past 40%, or it may not be possible within our parameters.

A particularly promising result is seen in the NPV/PPV plot. The model can achieve a high target NPV (e.g., 0.95) while maintaining a moderate but still clinically useful target PPV (e.g., 0.75) with a minimal abstention rate of less than 20%, demonstrating its potential promise as a triaging tool, confidently identifying negative cases while flagging a smaller, manageable subset of potential positive cases for further review. This is particularly valuable in the context of autism, where it could help reduce the backlog of patients waiting for full evaluation by prioritizing those that pass this initial screening.

4. Discussion

We demonstrate that abstention-based thresholding can serve as a flexible mechanism for managing uncertainty in AI-based clinical diagnostic tools, and the possibilities that this dual-threshold approach can provide in the context of pediatric autism classification. By applying abstention strategies to video-based autism classifiers trained on traditional instruments, we were able to explore trade-offs between PPV, NPV, and the proportion of samples retained for classification. Our results show that carefully selecting upper and lower thresholds can meaningfully shift clinical performance metrics without retraining the underlying model, offering a post-hoc yet clinically interpretable mechanism to tailor model behavior to the needs of diverse deployment contexts. These findings build on prior work in medical AI that emphasizes the importance of communicating uncertainty[36, 37] by offering a concrete pathway to operationalize it within the diagnostic process.

Additionally, our results demonstrate that dynamic thresholding can permit us to utilize the existing models in many different contexts. For instance, as described above, it could be employed as a triage tool, with a fixed high NPV and trading off PPV to fit abstention requirements. Another use case would be as a “first pass” model, requiring a high PPV and NPV but also accepting a large abstention rate. This would permit one to filter for “easier” cases (i.e. cases where a model has high confidence) cheaply then transfer the remaining cases to a more resource intensive next step. Overall, working with abstention allows us to explore a plethora of new applications for a single model, all without requiring retraining.

There are limitations worth highlighting that will be the target of future work. First, our dataset, though diverse and significantly larger than in prior work, is still constrained by the availability of labeled videos with verified diagnoses. Publicly available videos, especially those sourced from platforms like YouTube, often lack complete demographic or clinical metadata, limiting the ability to perform stratified or subgroup analyses. Second, the crowdsourced ratings used to construct the input features, although aggregated, may contain suboptimal variations in rater quality. While we employed clinicians and trained raters in many instances, we did not systematically quantify or control for inter-rater reliability across the full dataset. This introduces uncertainty that may impact classifier threshold optimization. Given that the goal of this study was to demonstrate the general tradeoffs of sweeping across a wide span of threshold points and abstention windows, as opposed to the identification of specific points for clinical decision making based on thresholds aligned with real-world distribution, we used pre-existing classifiers without additional calibration. However, we acknowledge that model calibration is a critical step for any future work that would target such real-world clinical goals. Furthermore, we note that future work exploring feature importance contributing to abstain decisions would provide additional insight into specific factors that drive models to identify samples as indeterminate, resulting in increased model explainability, a key facet of uncertainty management. Finally we acknowledge that selecting thresholds and evaluating performance on the same inference outputs could introduce optimistic bias. Although inference was performed using frozen classifiers that were previously trained on an entirely different dataset, and our thresholds were utilized to examine the tradeoffs of abstain windows and performance, future work will benefit from use of a dedicated evaluation dataset and/or nested-cross validation.

The paradigm of abstention in clinical informatics leads to several opportunities for future work. A natural follow up of this work is to study patient and clinician-facing consequences of abstention from a human-centered design perspective (e.g., user trust, follow-up burden, and missed diagnosis costs). Future work should incorporate decision-analytic frameworks or prospective studies to align abstention behavior and threshold selection with clinical utility and patient outcomes. In addition, adaptive thresholding strategies that respond to contextual features, such as demographic variables, symptom severity, or prior history, may enable more personalized uncertainty management decision making. Exploring abstention in multi-class or multi-label diagnostic settings, including differential diagnosis among overlapping developmental conditions, would enable abstention management in more challenging diagnostic scenarios (e.g., distinguishing between autism vs. ADHD vs. both vs. none).[48, 49] Finally, we plan to explore the concept of abstention in less noisy contexts, such as by using a single source of videos and a highly filtered subset of human raters.

5. Conclusion

Abstention is a powerful mechanism for enhancing the quantification and management of uncertainty in AI-based clinical diagnostic tools. We demonstrated how adjusting classification thresholds on models trained with traditional autism measurement tools can meaningfully trade off between performance metrics, such as PPV, NPV, and coverage, without requiring retraining or architectural changes. These thresholding strategies provide more "degrees of freedom" to align model outputs with diverse clinical success criteria, depending on the setting, population, and tolerance for false positives or negatives.

As AI tools increasingly move from research settings into regulated clinical workflows, it is essential to develop and formalize strategies for managing uncertainty in ways that are interpretable, clinically relevant, and adaptable. Abstention offers one such strategy, enabling systems to defer low-confidence predictions and support safer, more trustable decision-making. Future work should build upon this paradigm by integrating abstention into clinician-facing interfaces, evaluating its real-world impact, and developing guidelines that support its responsible use across a range of clinical domains.

6. Acknowledgements

This work was funded by National Institutes of Health, under awards R01LM013364 and R01LM014342, and Stanford Center for Digital Health. PW acknowledges funding from the National Institutes of Health (NIH) via the NIH Director's New Innovator Award (award DP2-EB035858). We acknowledge the use of ChatGPT in polishing sections of the manuscript for readability; all emitted outputs were thoroughly reviewed and edited by the coauthors.

References

1. Thomas KC, Ellis AR, Konrad TR, Holzer CE, Morrissey JP. County-level estimates of mental health professional shortage in the United States. Psychiatr Serv. 2009;60(10):1323-8. Epub 2009/10/03. doi: 10.1176/ps.2009.60.10.1323. PubMed PMID: 19797371.
2. Gotham K, Risi S, Pickles A, Lord C. The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity. J Autism Dev Disord. 2007;37(4):613-27. Epub 2006/12/21. doi: 10.1007/s10803-006-0280-1. PubMed PMID: 17180459.
3. Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism Dev Disord. 1994;24(5):659-85. Epub 1994/10/01. PubMed PMID: 7814313.
4. Constantino JN, Gruber CP. Social responsiveness scale, second edition (SRS-2): Western Psychological Services Los Angeles, CA; 2012.
5. Falkmer T, Anderson K, Falkmer M, Horlin C. Diagnostic procedures in autism spectrum disorders: a systematic literature review. European Child & Adolescent Psychiatry. 2013;22(6):329-40. doi: 10.1007/s00787-013-0375-0.
6. Lebersfeld JB, Swanson M, Clesi CD, O'Kelley SE. Systematic Review and Meta-Analysis of the Clinical Utility of the ADOS-2 and the ADI-R in Diagnosing Autism Spectrum Disorders in Children. Journal of Autism and Developmental Disorders. 2021;51(11):4101-14. doi: 10.1007/s10803-020-04839-z.
7. Kamp-Becker I, Albertowski K, Becker J, Ghahreman M, Langmann A, Mingebach T, Poustka L, Weber L, Schmidt H, Smidt J, Stehr T, Roessner V, Kucharczyk K, Wolff N, Stroth S. Diagnostic accuracy of the ADOS and ADOS-2 in clinical practice. European Child & Adolescent Psychiatry. 2018;27(9):1193-207. doi: 10.1007/s00787-018-1143-y.
8. Shaw KA. Prevalence and Early Identification of Autism Spectrum Disorder Among Children Aged 4 and 8 Years—Autism and Developmental Disabilities Monitoring Network, 16 Sites, United States, 2022. MMWR Surveillance Summaries. 2025;74.
9. Shannon J, Salomon M, Kraft C, Chettiath T, Taraman S, Badesch S, editors. Wait Times and Processes for Autism Diagnostic Evaluations: A First Report Survey of Autism Centers in the US. JOURNAL OF DEVELOPMENTAL AND BEHAVIORAL PEDIATRICS; 2024: LIPPINCOTT WILLIAMS & WILKINS TWO COMMERCE SQ, 2001 MARKET ST, PHILADELPHIA ….
10. National Autism Data Center AJ. How early does diagnosis happen? Autism by the Numbers supported by Autism Speaks. In: Institute DA, editor.
11. Mandell DS, Listerud J Fau - Levy SE, Levy Se Fau - Pinto-Martin JA, Pinto-Martin JA. Race differences in the age at diagnosis among medicaid-eligible children with autism(0890-8567 (Print)).
12. Gordon-Lipkin E, Foster J, Peacock G. Whittling Down the Wait Time: Exploring Models to Minimize the Delay from Initial Concern to Diagnosis and Treatment of Autism Spectrum Disorder. Pediatr Clin North Am. 2016;63(5):851-9. Epub 2016/08/28. doi: 10.1016/j.pcl.2016.06.007. PubMed PMID: 27565363; PMCID: PMC5583718.
13. Vu M, Duhig AM, Tibrewal A, Campbell CM, Gaur A, Salomon C, Gupta A, Kruse M, Taraman S. Increased delay from initial concern to diagnosis of autism spectrum disorder and associated health care resource utilization and cost among children aged younger than 6 years in the United States(2376-1032 (Electronic)).

14. Duda M, Ma R, Haber N, Wall DP. Use of machine learning for behavioral distinction of autism and ADHD. Transl Psychiat. 2016;6(2):e732. doi: ARTN e732
10.1038/tp.2015.221. PubMed PMID: WOS:000373892800003.
15. Kosmicki JA, Sochat V, Duda M, Wall DP. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Transl Psychiat. 2015;5(2):e514. doi: ARTN e514
10.1038/tp.2015.7. PubMed PMID: WOS:000367652200002.
16. Levy S, Duda M, Haber N, Wall DP. Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism. Mol Autism. 2017;8:1-17.
17. Wall DP, Dally R, Luyster R, Jung JY, Deluca TF. Use of artificial intelligence to shorten the behavioral diagnosis of autism. Plos One. 2012;7(8):e43855. Epub 2012/09/07. doi: 10.1371/journal.pone.0043855. PubMed PMID: 22952789; PMCID: PMC3428277.
18. Wall DP, Kosmicki J, DeLuca TF, Harstad E, Fusaro VA. Use of machine learning to shorten observation-based screening and diagnosis of autism. Transl Psychiat. 2012;2(4):e100. doi: ARTN e100
10.1038/tp.2012.10. PubMed PMID: WOS:000306218400003.
19. Washington P. A Perspective on Crowdsourcing and Human-in-the-Loop Workflows in Precision Health. J Med Internet Res. 2024;26:e51138. Epub 11.4.2024. doi: 10.2196/51138. PubMed PMID: 38602750.
20. Washington P, Chrisman B, Leblanc E, Dunlap K, Kline A, Mutlu C, Stockham N, Paskov K, Wall DP. Crowd annotations can approximate clinical autism impressions from short home videos with privacy protections. Intelligence-Based Medicine. 2022;6:100056. doi: https://doi.org/10.1016/j.ibmed.2022.100056.
21. Washington P, Kalantarian H, Tariq Q, Schwartz J, Dunlap K, Chrisman B, Varma M, Ning M, Kline A, Stockham N, Paskov K, Voss C, Haber N, Wall DP. Validity of Online Screening for Autism: Crowdsourcing Study Comparing Paid and Unpaid Diagnostic Tasks. J Med Internet Res. 2019;21(5):e13668. Epub 2019/05/28. doi: 10.2196/13668. PubMed PMID: 31124463.
22. Washington P, Leblanc E, Dunlap K, Kline A, Mutlu C, Chrisman B, Stockham N, Paskov K, Wall DP. Crowd Annotations Can Approximate Clinical Autism Impressions from Short Home Videos with Privacy Protections. medRxiv. 2021:2021.07.01.21259683. doi: 10.1101/2021.07.01.21259683.
23. Washington P, Leblanc E, Dunlap K, Penev Y, Kline A, Paskov K, Sun MW, Chrisman B, Stockham N, Varma M. Precision telemedicine through crowdsourced machine learning: testing variability of crowd workers for video-based autism feature recognition. Journal of personalized medicine. 2020;10(3):86.
24. Washington P, Leblanc E, Dunlap K, Penev Y, Varma M, Jung J-Y, Chrisman B, Sun MW, Stockham N, Paskov KM, Kalantarian H, Voss C, Haber N, Wall DP. Selection of trustworthy crowd workers for telemedical diagnosis of pediatric autism spectrum disorder. Biocomputing 2021: WORLD SCIENTIFIC; 2020. p. 14-25.
25. Washington P, Tariq Q, Leblanc E, Chrisman B, Dunlap K, Kline A, Kalantarian H, Penev Y, Paskov K, Voss C, Stockham N, Varma M, Husic A, Kent J, Haber N, Winograd T, Wall DP. Crowdsourced feature tagging for scalable and privacy-preserved autism diagnosis. medRxiv. 2020:2020.12.15.20248283. doi: 10.1101/2020.12.15.20248283.
26. Washington P, Tariq Q, Leblanc E, Chrisman B, Dunlap K, Kline A, Kalantarian H, Penev Y, Paskov K, Voss C, Stockham N, Varma M, Husic A, Kent J, Haber N, Winograd T, Wall DP.

Crowdsourced privacy-preserved feature tagging of short home videos for machine learning ASD detection. Scientific Reports. 2021;11(1):7620. doi: 10.1038/s41598-021-87059-4.

27. al. PWe. Evaluating Privacy-Preserving Mechanisms for Crowdsourcing Pediatric Autism Diagnoses from Home Video. Under review. 2020.

28. Wall DP, Kosmicki J, Deluca TF, Harstad E, Fusaro VA. Use of machine learning to shorten observation-based screening and diagnosis of autism. Transl Psychiatry. 2012;2:e100. Epub 2012/07/27. doi: 10.1038/tp.2012.10. PubMed PMID: 22832900; PMCID: 3337074.

29. Fusaro VA, Daniels J, Duda M, DeLuca TF, D'Angelo O, Tamburello J, Maniscalco J, Wall DP. The Potential of Accelerating Early Detection of Autism through Content Analysis of YouTube Videos. Plos One. 2014;9(4):e93533. doi: ARTN e93533
10.1371/journal.pone.0093533. PubMed PMID: WOS:000336863900022.

30. Tariq Q, Daniels J, Schwartz JN, Washington P, Kalantarian H, Wall DP. Mobile detection of autism through machine learning on home video: A development and prospective validation study. Plos Med. 2018;15(11):e1002705. doi: 10.1371/journal.pmed.1002705; PMCID: PMC6258501.

31. Tariq Q, Fleming SL, Schwartz JN, Dunlap K, Corbin C, Washington P, Kalantarian H, Khan NZ, Darmstadt GL, Wall DP. Detecting Developmental Delay and Autism Through Machine Learning Models Using Home Videos of Bangladeshi Children: Development and Validation Study. J Med Internet Res. 2019;21(4):e13822. Epub 24.04.2019. doi: 10.2196/13822. PubMed PMID: 31017583; PMCID: PMC6505375.

32. Tariq Q, Daniels J, Schwartz JN, Washington P, Kalantarian H, Wall DP. Mobile detection of autism through machine learning on home video: A development and prospective validation study. PLoS Med. 2018;15(11):e1002705. Epub 2018/11/28. doi: 10.1371/journal.pmed.1002705. PubMed PMID: 30481180; PMCID: PMC6258501 following competing interests: DPW is the scientific founder of Cognoa, a company focused on digital pediatric healthcare; the approach and findings presented in this paper are independent from/not related to Cognoa. All other authors have declared no competing interests exist.

33. Megerian JT, Dey S, Melmed RD, Coury DL, Lerner M, Nicholls C, Sohl K, Rouhbakhsh R, Narasimhan A, Romain J, Golla S, Shareef S, Ostrovsky A, Shannon J, Kraft C, Liu-Mayo S, Abbas H, Gal-Szabo DE, Wall DP, Taraman S. Performance of Canvas Dx, a Novel Software-based Autism Spectrum Disorder Diagnosis Aid for Use in a Primary Care Setting (P13-5.001). Neurology. 2022;98(18_supplement):1025. doi: doi:10.1212/WNL.98.18_supplement.1025.

34. Shannon J, Taraman S, Liu-Mayo S, Salomon M, Seal M, Kraft C, Wall D, editors. Exploring the real-world performance of an artificial intelligence-based diagnostic device for autism: an aggregate analysis of early Canvas Dx prescription and output data. ANNALS OF NEUROLOGY; 2023: WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA.

35. Abdar M, Khosravi A, Islam SMS, Acharya UR, Vasilakos AV. The need for quantification of uncertainty in artificial intelligence for clinical data analysis: increasing the level of trust in the decision-making process. IEEE Systems, Man, and Cybernetics Magazine. 2022;8(3):28-40. doi: 10.1109/MSMC.2022.3150144.

36. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. Npj Digit Med. 2021;4(1):4. doi: 10.1038/s41746-020-00367-3.

37. Wall DP, Liu-Mayo S, Salomon C, Shannon J, Taraman S. Optimizing a de novo artificial intelligence-based medical device under a predetermined change control plan: Improved ability to

detect or rule out pediatric autism. Intelligence-Based Medicine. 2023;8:100102. doi: https://doi.org/10.1016/j.ibmed.2023.100102.

38. Salomon C, Heinz K, Aronson-Ramos J, Wall DP. An analysis of the real world performance of an artificial intelligence based autism diagnostic. Scientific Reports. 2025;15(1):29503. doi: 10.1038/s41598-025-15575-8.

39. Kalantarian H, Jedoui K, Washington P, Tariq Q, Dunlap K, Schwartz J, Wall DP. Labeling images with facial emotion and the potential for pediatric healthcare. Artificial Intelligence in Medicine. 2019;98:77-86. doi: https://doi.org/10.1016/j.artmed.2019.06.004.

40. Kalantarian H, Washington P, Schwartz J, Daniels J, Haber N, Wall DP. Guess What? Journal of Healthcare Informatics Research. 2018. doi: 10.1007/s41666-018-0034-9.

41. Washington P, Kalantarian H, Kent J, Husic A, Kline A, Leblanc E, Hou C, Mutlu OC, Dunlap K, Penev Y, Varma M, Stockham NT, Chrisman B, Paskov K, Sun MW, Jung J-Y, Voss C, Haber N, Wall DP. Improved Digital Therapy for Developmental Pediatrics Using Domain-Specific Artificial Intelligence: Machine Learning Study. JMIR Pediatr Parent. 2022;5(2):e26760. Epub 8.4.2022. doi: 10.2196/26760. PubMed PMID: 35394438.

42. Fombonne E, Coppola L, Mastel S, O'Roak BJ. Validation of Autism Diagnosis and Clinical Data in the SPARK Cohort. Journal of Autism and Developmental Disorders. 2022;52(8):3383-98. doi: 10.1007/s10803-021-05218-y.

43. Sparrow SS, Cicchetti DV, Balla DA, Doll EA. Vineland adaptive behavior scales: Survey forms manual: American Guidance Service; 2005.

44. Duda M, Daniels J, Wall DP. Clinical Evaluation of a Novel and Mobile Autism Risk Assessment. J Autism Dev Disord. 2016;46(6):1953-61. Epub 2016/02/14. doi: 10.1007/s10803-016-2718-4. PubMed PMID: 26873142; PMCID: PMC4860199.

45. Lord C, Rutter M. Social communication questionnaire (SCQ). Torrance, CA: WPS. 2003.

46. Frazier TW, Dimitropoulos A, Abbeduto L, Armstrong-Brine M, Kralovic S, Shih A, Hardan AY, Youngstrom EA, Uljarević M, Team VB. Psychometric evaluation of the Autism Symptom Dimensions Questionnaire. Developmental Medicine & Child Neurology. 2025;67(6):758-69.

47. Canale RR, Larson C, Thomas RP, Barton M, Fein D, Eigsti I-M. Investigating frank autism: clinician initial impressions and autism characteristics. Mol Autism. 2024;15(1):48. doi: 10.1186/s13229-024-00627-z.

48. Jaiswal A, Wall DP, Washington P, editors. Challenges in the Differential Classification of Individual Diagnoses from Co-Occurring Autism and ADHD Using Survey Data. 2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI); 2024: IEEE.

49. Jaiswal A, Kruiper R, Rasool A, Nandkeolyar A, Wall DP, Washington P. Digitally Diagnosing Multiple Developmental Delays Using Crowdsourcing Fused With Machine Learning: Protocolfor a Human-in-the-Loop Machine Learning Study. JMIR Res Protoc. 2024;13:e52205. Epub 8.2.2024. doi: 10.2196/52205. PubMed PMID: 38329783.