Detecting Tardive Dyskinesia Using Video-Based Artificial Intelligence

Anthony A. Sterns, PhD; Joel W. Hughes, PhD; Bradley Grimm, MS; Loren Larsen, MS; Fred Ma, MD, PhD; Rakesh Ranjan, MD; Carlene MacMillan, MD; Bretton H. Talbot, PhD; Joseph H. Friedman, MD; and Owen S. Muir, MD

Abstract

Objective: Tardive dyskinesia (TD) is a late-onset adverse effect of dopamine receptor-blocking medications, characterized by involuntary movements primarily affecting the mouth, though other body parts may be involved. Severity of TD varies from mild to debilitating and is usually irreversible. Despite the existence of treatments such as VMAT2 inhibitors, TD remains underdiagnosed, with 40,000 patients treated of an estimated 2.6 million affected US individuals. This study demonstrates a novel, efficient, and reliable method to detect and bring TD to psychiatrists' attention using video-based artificial intelligence.

Methods: Individuals taking antipsychotic medications were enrolled in Study 1 (n = 46) between March and November 2018, in Study 2 (n = 136) between May 2023 and May 2024, and in Study 3 (n = 174) between July 2023 and May 2024. Participants completed video assessments. A vision transformer machine-learning architecture was evaluated by calculating the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity compared with a reference standard of the trained raters' evaluation of TD on the Abnormal Involuntary Movement Scale.

Results: The algorithm reached an AUC of 0.89 in the combined validation cohort across Studies 1, 2, and 3. The model demonstrated strong and reliable levels of agreement, outperforming human raters.

Conclusion: Our algorithm reliably detected suspected TD, reaching higher sensitivity and specificity than trained raters using the standard assessment. The algorithm can be used to monitor patients taking antipsychotic medications, allowing scarce resources to assess identified patients for a conclusive diagnosis by psychiatrists.

J Clin Psychiatry 2025;86(3):25m15792

Author affiliations are listed at the end of this article.

Tardive dyskinesia (TD) is a serious and often disabling neurological syndrome primarily caused by prolonged exposure to antipsychotic medications.[1] It is characterized by involuntary, repetitive movements of the face, trunk, and extremities,[2] such as grimacing, blinking, and abnormal posturing. Incidence rates are estimated at 2%–4% per year for second-generation antipsychotics and higher for first-generation agents.[3–5] TD significantly diminishes patients' quality of life,[6] compounding distress and cognitive impairment and contributing to nonadherence to antipsychotic treatment.

Early detection of TD allows the deployment of effective interventions to mitigate morbidity. Traditional assessment is accomplished with one of several validated scales. The Abnormal Involuntary Movement Scale (AIMS) is the most frequently cited.[7,8] This assessment is more reliable when performed by raters with significant experience with TD. However, it is difficult even for skilled diagnosticians to devote in-person resources, as often as 2–4 times per year, as would be necessary to provide every patient with the recommended standard of care regarding monitoring for TD.

There is a lack of consensus about which conditions can be managed by telepsychiatry alone. Given that antipsychotic prescribing comes with the risk of TD, one understandable concern is the ability to evaluate this potential adverse effect appropriately. This article specifically addresses the reliability of machine-learning approaches to augment the initial screening of TD performed remotely. Importantly, although our method describes a completely remote technology, it does not replace an in-person assessment by a physician for the definitive diagnosis and subsequent management of the condition. The increase in telemedicine-first care, along with increasing demand for psychiatric services, has placed more demands on the health care workforce, frequently at the expense of time-consuming safety monitoring protocols like those for TD.[9]

Clinical Points

- Screening for tardive dyskinesia (TD) is a standard of care. However, it is not routinely implemented and has low sensitivity and specificity.
- For patients taking antipsychotic medication, clinicians might consider if automated screening tools will enhance their ability to screen for TD, which now has breakthrough treatments available.

Here, we report the results of 3 studies that sought to stratify the risk of suspected TD in patients exposed to antipsychotic medications. Beyond our algorithm's ability to identify the presence or absence of suspected TD, it can also stratify the severity of the disorder. Using a visual transformation algorithm,[10] our primary end point was performance at or greater than 90% area under the curve (AUC) in detecting the presence or absence of suspected TD compared to its presence/absence as defined by AIMS ratings on individuals taking an antipsychotic medication as assessed by experienced raters.

METHODS

We collected video data from individuals taking antipsychotic medications across 3 studies. All participants were at risk for TD, and video data were captured via a smartphone app that guided participants through a standardized protocol. Each participant's video answers were assessed by our algorithm and compared with a trained rater assessment of an AIMS as the ground truth of the presence or absence of TD.

For Studies 1 and 2, the participants completed a standard AIMS procedure by sitting across from a device on a stand rather than an assessor. For those elements focused on the face and mouth, a close-up video image captured the individual's face, trunk, and hands. In Study 1, following the AIMS, the individuals responded to 6 open-ended questions (eg, "What kind of pet do you like?") captured on video. The individual elements of the AIMS procedure were later viewed and assessed by 3 trained raters and scored.

Study 2 similarly involved the same AIMS procedure captured by a device on a stand. Study 2 reduced the open-ended questions to the 3 that generated the most detailed responses in Study 1. The AIMS elements were assessed similarly for each participant, with 3 trained raters completing independent assessments. For both studies, a consensus conference was held where AIMS scoring differed. The consensus served as the standard in the subsequent evaluations of the algorithm.

Study 3 utilized a group of participants who were given a standard AIMS assessment by a single trained rater. The video assessment protocol was again simplified for use in home settings. Demographic information, including race/ethnicity and biological sex, was collected using a self-report questionnaire.

The Study 3 video assessment consisted of 4 steps captured using the same smartphone app as Studies 1 and 2 but with the following elements: (1) 15 seconds of tapping a hand on the shoulder, (2) 30 seconds of opening the mouth and sticking out the tongue followed by sitting still for 30 seconds, and (3) answering 2 open-ended questions. The first 2 components are similar to 3 specific elements of the standard AIMS. In the first 2 studies, the video data of participants were collected in a clinical setting, and the assessment was done remotely online. For the third study, the AIMS was conducted in a clinical setting, and the video data were conducted in the participant's homes or other convenient locations. All 3 studies began with obtaining informed consent from the participants.

Participants

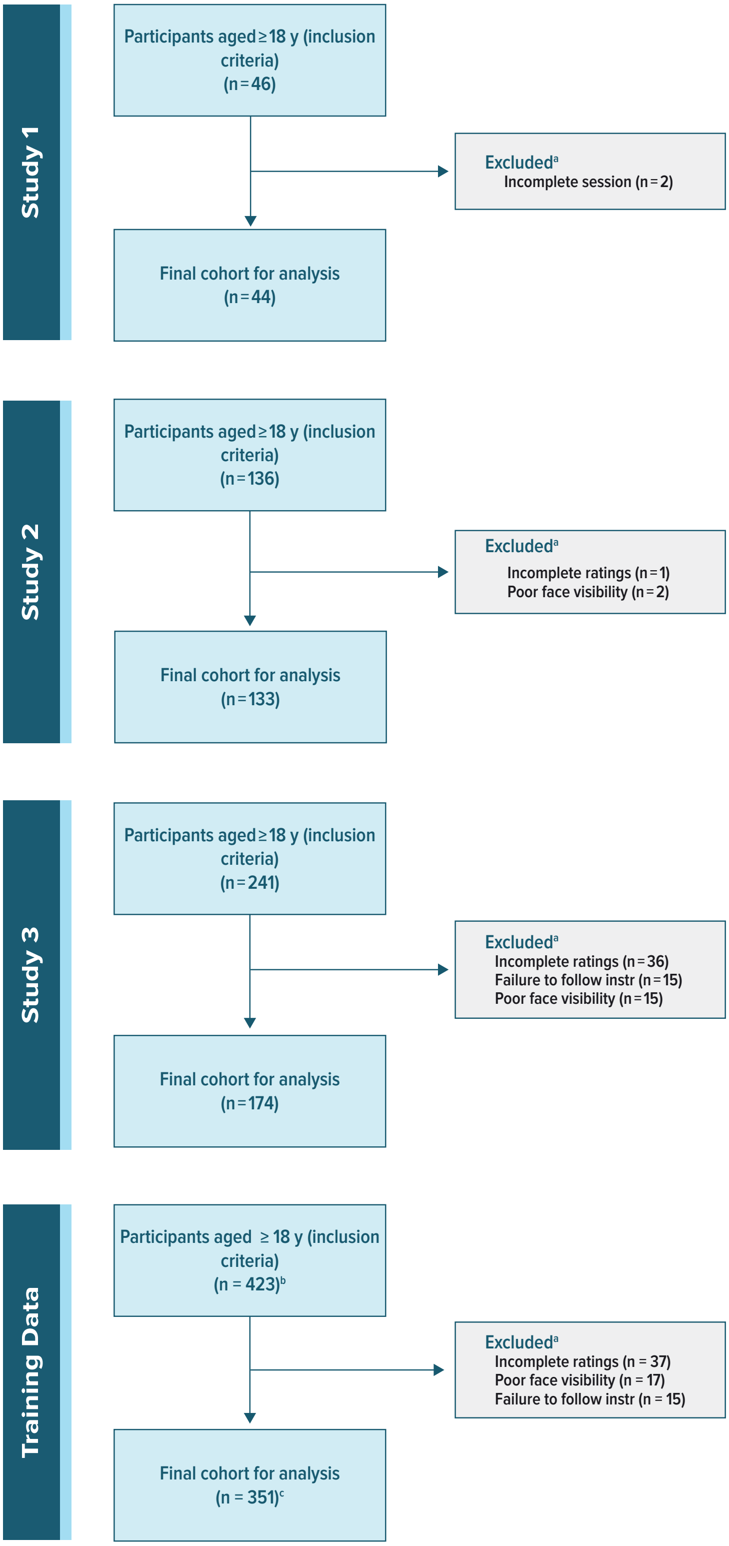
The model's training included data from the pilot (Study 1) and the follow-on study (Study 2). The model was iteratively trained: Study 1 included 46 participants, and Study 2 added 136 participants. Participants were recruited from clinic populations of county social services organizations in Northeast Ohio and behavioral health community clubhouse settings in New York City with institutional review board–approved materials and were provided informed consent. The principal inclusion criteria were that participants had taken an antipsychotic medication for at least 90 days. Exclusion criteria included head injury in the last year, a history of cognitive or developmental disability that would inhibit answering questionnaires, and severe visual impairment.

Participants were enrolled to ensure an even mix of individuals previously diagnosed with TD and those without, as defined by chart review. Our final sample contained 25% with no signs of TD, 60% with minimal to mild, and 15% with moderate-to-severe TD using the global severity rating from the AIMS. At model training time, participants were excluded if (1) participants' faces could not be assessed due to issues of poor lighting or improper positioning, or (2) participants did not follow instructions, such as wearing sunglasses or masks, preventing adequate detection of facial movements. Of the total participants, 2 were removed from Study 1 and 3 from Study 2 (Figure 1). All patients with chewing gum or loose dentures were asked to remove them in the protocol.

A panel of 4 trained evaluators tested to pass criteria using commercially available clinical training tools (see

Figure 1.

Flow Diagram for Sample Size and Data Exclusion


Study 1
Participants aged ≥ 18 y (inclusion criteria) (n = 46)
Excluded[a]
Incomplete session (n = 2)
Final cohort for analysis (n = 44)
Study 2
Participants aged ≥ 18 y (inclusion criteria) (n = 136)
Excluded[a]
Incomplete ratings (n = 1)
Poor face visibility (n = 2)
Final cohort for analysis (n = 133)
Study 3
Participants aged ≥ 18 y (inclusion criteria) (n = 241)
Excluded[a]
Incomplete ratings (n = 36)
Failure to follow instr (n = 15)
Poor face visibility (n = 15)
Final cohort for analysis (n = 174)
Training Data
Participants aged ≥ 18 y (inclusion criteria) (n = 423)[b]
Excluded[a]
Incomplete ratings (n = 37)
Poor face visibility (n = 17)
Failure to follow instr (n = 15)
Final cohort for analysis (n = 351)[c]

[a] Exclusion criteria included incomplete sessions/ratings, poor face visibility, and failure to follow instructions.

[b] Data represented include all participants initially enrolled across 3 studies (N = 423).

[c] The combined final cohort (n = 351) was used for the primary analysis.

Abbreviation: instr = instructions.

https://aapp.org/aims/tips/2022) individually conducted a traditional AIMS evaluation of the videos as a diagnostic comparison.

Study 3 added 174 participants. The final dataset included 351 participants with 3,979 video responses, each containing clips around 1–2 seconds containing normal or abnormal movements. As a result of poor-quality video below the threshold required for analysis, 72 total participants (17%) were excluded from the evaluation dataset. Demographic information across all 3 studies is shown in Table 1.

The video clips were split 75%/25% to create a training and testing set. The AIMS score was the target output to train a neural network to evaluate the videos. The level of agreement between the algorithm's conclusion and the raters' consensus was compared.

Table 1.
Sample Characteristics

Variables	No TD (n = 201)	TD (n = 150)	Full sample (n = 351)
Age, mean ± SE, y	49.1 ± 13.8	55.2 ± 13.9	51.88 ± 14.2
Sex, n (%)			
Female	103 (29.3)	92 (26.2)	195 (55.6)
Male	97 (27.6)	58 (16.5)	155 (44.2)
Other	1 (0.3)	0 (0.0)	1 (0.3)
Race/ethnicity, n (%)			
Non-Hispanic Black	82 (23.4)	23 (6.6)	105 (29.9)
Non-Hispanic White	100 (28.5)	112 (31.9)	212 (60.4)
Hispanic	6 (1.7)	1 (0.3)	7 (2.0)
Other	13 (3.7)	14 (4.0)	27 (7.7)

Abbreviation: TD = tardive dyskinesia.

Procedures

In 2022, Sterns et al[11] presented the results of a pilot study (Study 1) using machine learning to categorize video responses of patients (n = 46) at risk of TD. To test the feasibility and validate the algorithm, 3 trained raters evaluated the videos of individuals with and without diagnosed TD completing all components of the AIMS. The AIMS elements are identical to the standard AIMS, except the evaluation was done on video rather than by the person directly. The AIMS instructions state that tremors should not be rated. While parkinsonism clearly may mask TD or reduce the severity, tremors and rigidity do not enter the evaluation process.

A machine learning algorithm using convolutional neural networks evaluated the open-ended questions focusing on the upper trunk and facial detection only. The machine learning engine demonstrated the ability to detect differences between the individuals with TD and without TD with an AUC of 0.77 (95% CI, 0.679–0.859), comparing on the ground truth of the presence or absence of TD as established by the consensus of the panel of 3 trained raters based on their conclusion using the AIMS.[12,13]

A vision transformer (ViT) is a deep-learning model designed for image recognition tasks. Unlike traditional convolutional neural networks, ViT divides images into fixed-size patches, which are then linearly embedded and processed through a transformer architecture originally designed for natural language processing. This allows the model to capture long-range dependencies within images, making it particularly effective for detecting the subtle, involuntary movements associated with TD. Instead of using convolutional layers typical in convolutional neural networks, a vision transformer divides an image into fixed-size patches, linearly embeds each patch, and feeds the resulting sequence of linear embeddings into a transformer encoder. This approach leverages the transformer's ability to capture long-range dependencies, leading to high performance on image classification tasks when trained on large datasets[10] Here, we compare the vision transformer model with an evaluation of the videos by trained raters using the AIMS. We use the average of 3 raters in Studies 1 and 2 evaluating video-recorded AIMS and 1 rater in Study 3 completing an AIMS in person.

We expected TD to be diagnosed by capturing video data of the face, shoulders, trunk, arms, and hands. We recognize that we will miss potential movement. However, leg and trunk involvement may move the upper body. Similarly, foot movement may rock the upper body slightly. If just toe movement is involved, then it would be missed in our method. Study 1 utilized 943 videos to train the algorithm. We looked to past research to calculate the sample size needed to achieve an AUC above 0.9. Based on Jacques and colleagues'[14] and Rosenberg and Ekman's [15] results, high-performing models can be achieved with 1,000 samples per class.

In Study 1, additional power was achieved with multiple samples per abnormal behavior in each video. Estimating from the sample, we expected each participant would provide an average of 22 segments in each video where abnormal behavior is recognizable. This gave a predicted total of 3,300 samples of abnormal behavior and significantly more normal behavior segments.[16–18] Therefore, to appropriately power Studies 2 and 3, our estimate concluded that 300 additional participants with TD (n = 150) and without TD (n = 150) would be sufficient. Data collection included participants across broad demographics (age, gender, race/ethnicity).

Model Architecture

The model was trained in 2 parts. The goal of the first part was to compress a large amount of video and audio data into a digestible embedding (a vector of

Table 2.
Comparison of Data Source and Review Processes

	Study 1	Study 2	Study 3
Reviewers	3 Reviewers	3 Reviewers + calibration/consensus	1 Reviewer
AIMS	Full[a]	Full[a]	Full[b]
Used in training	Y	Y	Y
Used in validation	Y (cross validated)	Y (cross validated)	N
Open ended questions	6	2	2
AIMS instructions	16	16	2[c]
Total videos/instructions	22	18	4
Input features	All videos	All videos	All videos
Model's target score	AIMS total score[d]	AIMS total score[d]	AIMS total score[e]

[a] AIMS assessment conducted via recorded videos in a controlled environment.
[b] AIMS assessment performed by clinician in face-to-face clinical setting.
[c] Assessment protocol simplified to 2 specific instructions (tongue protrusion and shoulder tapping movements).
[d] Final score calculated as the average of 3 independent reviewer ratings after calibration.
[e] Assessment score calculated by a single trained reviewer.
Abbreviations: AIMS = Abnormal Involuntary Movement Scale, N = no, Y = yes.

numbers with a compressed representation). To do this, we utilized a pretrained InternVideo model[19] combined with a Convolutional Recurrent Neural Network (CRNN) of the Mel-Frequency Cepstral Coefficients (MFCC) to process the video and audio of each response. This combined network was further trained to detect adventitious movements characteristic of TD.

Model Explainability

Our approach incorporated exploratory analyses to identify key features influencing TD outcomes. Specifically, our model labels and predicts each region separately, which allows us to isolate which specific regions are detected as "at risk" for the movements leading to a TD-positive risk score output.

RESULTS

Model Performance

The model's performance was evaluated by iteratively adding data collected across 3 studies. The models were tested using each test set and combined test sets (see Table 2). In all cases, the AUC of the model increased when additional data were included. When the model was trained on all available data, the AUC ranged from 0.85 to 0.98 across the available test sets. Performance improved from an initial AUC of 0.77 (Study 1) to 0.98 on the Study 1 test data using the full model trained on all 3 training sets (Figure 2).

In addition to evaluating the model's performance using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, we also calculated the sensitivity and specificity at the crossover point, where sensitivity equals specificity. This threshold was determined to be 0.81.

Heterogeneity of Treatment Effect

The assessment of nonrandom variation associated with patient characteristics is known as the heterogeneity of treatment effect. The model's performance was assessed across different demographic subgroups, including gender, ethnicity, and age. The results indicate low heterogeneity of treatment effect and uniformly high predictiveness across all demographic splits analyzed (Table 3).

Comparison of CRNN with Human Raters

Measuring the Cohen κ[20] of our reviewers highlights the difficulty of accurately detecting the presence of TD. In Study 2, we measured the average Cohen κ of the reviewers on the binary presence of TD, resulting in 0.37 ± 0.05. The Fleiss κ[21] on the same reviewers was 0.35.

To improve overall consistency, reviewers were allowed to discuss differences, update their ratings based on feedback, and were encouraged to reassess ratings where strong disagreement was found. After this iterative review process, the average Cohen κ was a moderate 0.57 ± 0.03 with the Fleiss κ = 0.58.

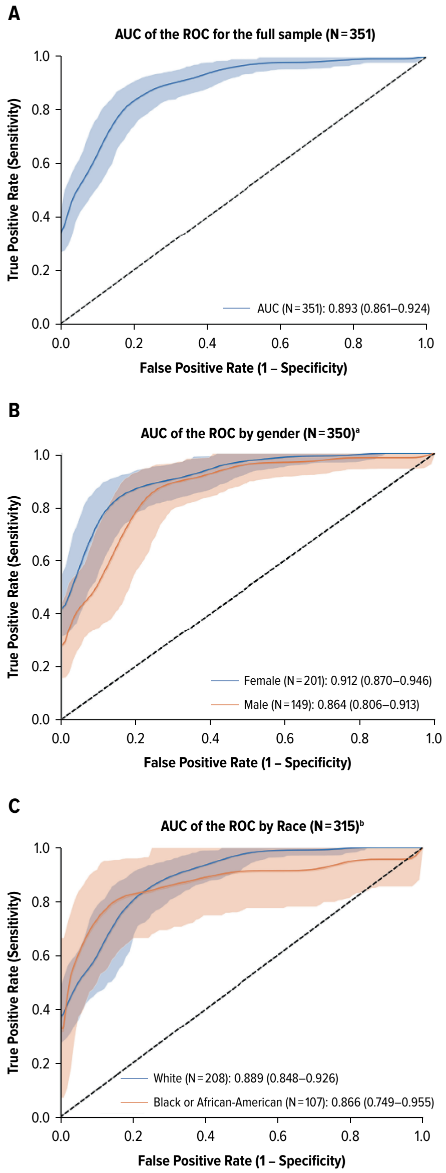
Our machine learning (ML) model was evaluated using the same data as the reviewers. The model achieved a Cohen κ of 0.51, demonstrating greater consistency than the reviewers' initial assessments. Furthermore, when utilizing the full dataset, the model's Cohen κ increased to 0.61, considered a strong and reliable level of agreement, outperforming human raters.

DISCUSSION

The results demonstrate the potential to utilize video-based machine learning algorithms to monitor for the

Figure 2.

The Area Under the Curve of the Receiver Operating Characteristic Curve for the Full Sample (A), by Gender (B), and by Race (C)



[a]ROC curves by gender demonstrating consistent performance across male and female participants.

[b]ROC curves by race showing comparable model performance across racial/ethnic groups.

Abbreviations: AUC = area under the curve, ROC = receiver operating characteristic.

presence or absence of suspected TD. When the algorithm identifies suspected TD, a psychiatrist is justified in evaluating and determining a diagnosis and actions to follow. It represents an opportunity to provide the patient with the regulated required monitoring at the standard pace or with increased frequency. An AIMS can be completed by trained care team professionals with greater urgency when an alert is reported.

Leveraging a Vision Transformer algorithm to detect TD based on AIMS ratings from experienced clinicians, our model achieved an AUC ranging from 0.85 to 0.98 when trained on the full dataset. This level of performance demonstrates the model's ability to reliably identify TD, often outperforming human raters in terms of both sensitivity and specificity. The sensitivity was 0.820, and the specificity was 0.821 when the threshold was set at 5.1. This compares favorably to the historical studies evaluating the performance of the most commonly used instrument, the AIMS. The algorithm demonstrated less bias compared to the interrater reliability of the human raters.

Additionally, the temporal nature of our embeddings used in our visual transformer approach enables us to pinpoint the time points at which risk-positive movements occurred. This dual capability enhances the interpretability of our artificial intelligence (AI) predictions, aligning with the framework suggested by Joyce et al.[22]

An additional benefit of our model is that it produces a continuous risk score, as it predicts the total AIMS score. The model predicts the "total" AIMS score (ie, the sum of the 7 body regions observed) even though the algorithm does not assess legs and feet directly. We picked a threshold on this score grouped into a binary prediction (TD/No-TD). All the metrics reported here are these binary metrics (AUC, sensitivity, specificity), which all refer to that binary prediction. That said, there is a benefit to utilizing the continuous value, as it allows the selection of multiple thresholds with multiple groupings (eg, low/medium/high). The visual transformer model can be calibrated to optimize sensitivity or specificity, which allows the tool to account for the different pretest probabilities of TD in varied populations.

These results suggest that the algorithm effectively captures complex interactions and nonlinear relationships between variables. The high predictiveness of the model across different subgroups is notable in that clinical populations often differ from clinical trial populations, limiting the generalizability of expected benefits from breakthrough treatments.

Diagnosing TD may be challenging for trained raters due to the subtle and variable nature of its symptoms. The AIMS used in this study, like all valid observation scales,[12,13] relies on subjective judgment. Even with rigorous training, human raters reliably exhibit significant assessment variability. The higher Cohen κ for the algorithm indicates that it is more reliable than trained raters. It is expected to improve over time with more data. Factors such as fatigue, cognitive biases, and inconsistencies in detecting and interpreting subtle movements contribute to variability among raters.

Although no curative treatment for TD is approved, 2 drugs to control its symptoms are FDA-approved as breakthrough treatments. These are deutetrabenazine

Table 3.

Model Performance Across Different Studies

Training data	AUC (95% CI)	Validation sample	Sensitivity (95% CI)[a]	Specificity (95% CI)[a]
1	0.770[b]	1	–	–
1 + 2	0.826 (0.72–0.92)	1 + 2	0.762 (0.65–0.90)	0.744 (0.66–0.89)
1 + 2	0.919 (0.82–0.99)	1	1.0 (0.67–1.0)	0.878 (0.73–1.0)
1 + 2	0.823 (0.71–0.92)	2	0.778 (0.61–0.88)	0.783 (0.61–0.88)
1 + 2 + 3	0.893 (0.86–0.92)	1 + 2 + 3	0.82 (0.76–0.86)	0.821 (0.78–0.86)
1 + 2 + 3	0.984 (0.93–1.0)	1	1.0 (1.0–1.0)	0.951 (0.89–1.0)
1 + 2 + 3	0.851 (0.72–0.96)	2	0.778 (0.67–0.93)	0.791 (0.66–0.92)

[a] The sensitivity and specificity were measured at the nearest point to where their values cross when choosing a threshold.

[b] Study 1 data are from the reported paper cited in Study 1.

Abbreviation: AUC = area under the curve.

and valbenazine.[23,24] Given the reported rates of antipsychotic treatment, the expected prevalence in the US of TD is between 800,000 and 2,600,000. Currently, approximately 40,000 patients are prescribed either deutetrabenazine or valbenazine to reduce the symptoms of TD.[25,26] There is a wide gap between those estimated with the condition and those who have received a diagnosis and prescribed treatment.

The absence of early detection hinders the ability to advance treatment for TD. While the relationship between early identification and reversibility remains incompletely characterized in the literature, timely detection allows for prompt intervention strategies such as medication adjustment or symptom management. Current evidence suggests that prolonged exposure to causative agents may increase the likelihood of permanent symptoms; however, long-term longitudinal studies specifically examining outcomes based on detection timing are limited. Early identification enables clinicians to implement risk mitigation strategies before symptoms potentially progress to more severe manifestations that impact patient functioning and quality of life. Whether to spot TD to change the treatment course or treat symptoms with medication, clinicians would benefit from an efficient method to identify and monitor the condition. Early identification is the first step in understanding the irreversibility of the disorder, which is particularly important for second-generation antipsychotics being increasingly used to treat depression, for which there are non-neuroleptic alternatives.

Impairment and disability are often experienced by patients who develop TD. Without efficient screening, many cases of TD go unnoticed, probably increasing the likelihood of irreversibility. However, it is difficult even for skilled diagnosticians to devote the in-person resources necessary to provide every patient prescribed an antipsychotic medication the 4–6 annual assessments to meet the standard of care for TD monitoring. The increase in telemedicine-first care, along with increasing demand for psychiatric services, has placed more demands on the health care workforce, frequently at the expense of time-consuming safety monitoring protocols.[9] Despite best efforts, TD is usually recognized only after the involuntary movements become permanent. A rapid and automatic TD detection method would enable timely diagnosis and avoid morbidity, potentially obviating the need for expensive lifelong treatment.

Limitations

Studies 1 and 2 focused on obtaining video data with both the AIMS and the open-ended questions in a clinical setting, while Study 3 was gathered in a home setting. In Study 3, 17% of participants had to be excluded because of poor-quality video. To increase the likelihood of proper video capture in the future, the tool should provide real-time feedback on whether video quality is sufficient so it can coach the user to correct issues, Future versions of the algorithm will probably, with additional data, be able to reduce the quality threshold to restrict fewer people, In cases where there is an unresolvable issue with network, camera, environment, or user ability to correctly follow instructions, the person can be referred for an in-person or telehealth AIMS with the provider. Even with the current limitations, more monitoring and fewer in-person assessments are accomplished. The variety of training videos gathered is vital to the final model's ability to generalize to future evaluations and significantly increase performance.

While the algorithm significantly enhances the detection of TD, it cannot function independently as a diagnostic tool. A health care professional's evaluation is essential to confirm the diagnosis required for prescribing treatment. A key concern is that the subtle early symptoms of TD often go unrecognized by clinicians, further emphasizing the need for broader training and awareness that can be increased with AI-based tools.

Another limitation of the reported approach is that it excludes foot and toe movements from the assessment protocol. While rare, isolated cases of TD affecting only the feet and toes have been observed. Although foot assessment is relatively straightforward, requiring participants to remove shoes presents practical challenges, particularly for elderly patients or those with mobility issues.

Future Research

Future research should emphasize a longitudinal approach in which patients are monitored monthly or quarterly with medication monitoring to demonstrate the potential of smartphone-based patient monitoring of TD fully. Another avenue for future research is to explore the integration of this AI-based diagnostic tool into routine telemedicine sessions. This could enhance ongoing care for individuals on antipsychotic medications by providing continuous, remote monitoring for early signs of TD, allowing for timely intervention.

Future research should also focus on integrating medication adherence monitoring with TD diagnostics. Understanding patients' adherence patterns will provide critical insights into how treatment strategies can be adjusted to prevent the progression of TD and reduce the risk of irreversible symptoms.

CONCLUSION

The combined studies demonstrate that self-administered, smartphone-recorded video interviews can reliably yield data scored using algorithms produced using highly discriminating machine learning approaches. The underdiagnosis of TD, exacerbated by insufficient professional training and the time constraints of assessments, underscores the value of our AI-based tool. Enabling efficient, accurate, and scalable detection of TD, followed by a psychiatrist's diagnostic assessment completed following the algorithm's detection, this technology has the potential to significantly improve early diagnosis and patient outcomes, especially in remote care settings where resources are the scarcest.

Article Information

Published Online: May 28, 2025. https://doi.org/10.4088/JCP.25m15792
© 2025 Physicians Postgraduate Press, Inc.

Submitted: January 14, 2025; accepted March 18, 2025.

To Cite: Sterns AA, Hughes JW, Grimm B, et al. Detecting tardive dyskinesia using video-based artificial intelligence. *J Clin Psychiatry* 2025;86(3):25m15792.

Author Affiliations: iRxReminder, Cleveland, Ohio (Sterns, Ma, Muir); Scripps Gerontology Center, Miami University, Oxford, Ohio (Sterns); Department of Information Management Systems, Ambassador Crawford College of Business and Entrepreneurship, Kent, Ohio (Sterns); Department of Psychological Sciences, Kent State University, Kent, Ohio (Hughes); Videra Health, Orem, Utah (Grimm, Larsen, Talbot); Charak Health and Wellness, Cleveland, Ohio (Ranjan); Fermata, Brooklyn, New York (MacMillan, Muir); Osmind, Inc, San Francisco, California (MacMillan); Butler Hospital, Providence, Rhode Island (Friedman); Acacia Clinics, Sunnyvale, California (Muir).

Corresponding Author: Anthony A. Sterns, PhD, iRxReminder, 1800 63rd St, Cleveland, OH 44103 (asterns@irxreminder.com).

Relevant Financial Relationships: Dr Sterns is the CEO of iRxReminder, a company that provides telehealth solutions, and he holds equity in the company and serves on its board. **Dr Hughes** is a professor at Kent State University and has received research funding from iRxReminder for this study. **Mr Grimm** is an employee of Videra Health, which provides AI-based healthcare solutions, and holds equity in the company. **Mr Larsen** is the CTO of Videra Health and holds equity in the company. **Dr Ma** is the COO of iRxReminder, a company that provides telehealth solutions, and holds equity in the company and serves on its board. **Dr Ranjan** is the owner of Charak Health and Wellness Clinics and has received research support from iRxReminder. He has received research funding from various pharmaceutical companies and serves on advisory boards, but none directly related to this study. **Dr MacMillan** is the CEO of Fermata and has received consultancy fees from iRxReminder. She is the Chief Medical Officer of Osmind and holds equity in Fermata and Osmind. **Dr Talbot** is the Chief Clinical Officer of Videra Health and holds equity in the company. **Dr Friedman** has received research funding from various pharmaceutical companies and serves on advisory boards, but none directly related to this study. **Dr Muir** is the CMO of iRxReminder, a company that provides telehealth solutions, and CMO of Fermata, a mental health care provider, and holds equity in both companies.

Funding/Support: Funding for this research was provided in part by the National Institute of Mental Health under grant no. 2R44MH114763.

Role of the Sponsor: The funding source had no role in the study's design, conduct, or reporting. Data included and described as Study 3 were obtained through appropriate institutional review board consent as part of the real-world IMPACT-TD Registry sponsored by Teva Branded Pharmaceutical Products, R&D, Inc. Teva had no further contributions (financial or otherwise) to the work described herein, aside from a courtesy medical review of the manuscript prior to submission.

Ethics Approval: The study was approved and overseen by the Creative Action LLC IRB, a National Institutes of Health–registered Institutional Review Board (FWA #10005).

Patient Consent: All patients were given informed consent, and consent was documented and reviewed.

References

1. Shalit AM. 2020. http://www.nlm.nih.gov/medlineplus/ency/article/000685.htm
2. Simpson GM, Lee JH, Zoubok B, et al. A rating scale for tardive dyskinesia. *Psychopharmacology*. 1979;64(2):171–179.
3. Besag FMC, Vasey MJ, Salim I, et al. Tardive dyskinesia with antipsychotic medication in children and adolescents: a systematic literature review. *Drug Saf*. 2024;47(11):1095–1126.
4. Correll CU, Schenk EM. Tardive dyskinesia and new antipsychotics. *Curr Opin Psychiatry*. 2008;21(2):151–156.
5. Correll CU, Leucht S, Kane JM. Lower risk for tardive dyskinesia associated with second-generation antipsychotics: a systematic review of 1-year studies. *Am J Psychiatry*. 2004;161(3):414–425.
6. Vrijens B, Vincze G, Kristanto P, et al. Adherence to prescribed antihypertensive drug treatments: longitudinal study of electronically compiled dosing histories. *BMJ*. 2008;336(7653):1114–1117.
7. Guy W. Abnormal Involuntary Movement Scale (AIMS). *ECDEU Assess Man Psychopharmacol*. 1976;338:534–537.
8. Lane RD, Glazer WM, Hansen TE, et al. Assessment of tardive dyskinesia using the Abnormal Involuntary Movement Scale. *J Nerv Ment Dis*. 1985; 173(6):353–357.
9. Doraiswamy PM, Blease C, Bodner K. Artificial intelligence and the future of psychiatry: insights from a global physician survey. *Artif Intell Med*. 2020;102: 101753.
10. Dosovitskiy, A, Beyer, L, Kolesnikov, A, et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv preprint; 2020.
11. Sterns A, Larsen L, Grimm B, et al. Remote monitoring and AI for detecting TD and improving patient outcomes. *Gerontechnology*. 2022;21. https://doi.org/10.4017/gt.2022.21.s.706.opp3
12. Abnormal Involuntary Movement Scale (AIMS). *Psychopharmacol Bull*. 1988; 24(4):781–783. Accessed August 21, 2024. https://search.ebscohost.com/login.aspx?direct=true&db=mdc&AN=3249784&site=ehost-live
13. Munetz MR, Benjamin S. How to examine patients using the Abnormal Involuntary Movement Scale. *Hosp Community Psychiatry*. 1988;39(11): 1172–1177. Accessed August 21, 2024. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1989-19996-001&site=ehost-live.
14. Jacques JCS Jr, Güçlütürk Y, Pérez M, et al. First impressions: a survey on vision-based apparent personality trait analysis. *IEEE Trans Affect Comput*. 2022;13(1): 75–95. Accessed August 21, 2024. https://arxiv.org/pdf/1804.08046.pdf

15. Rosenberg EL, Ekman P. What the face reveals. In: Rosenberg EL, Ekman P, eds. *Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. 3rd Ed.. Oxford University Press; 2020. Accessed August 21, 2024.
16. Dai Z, Liu H, Le QV, et al. Coatnet: marrying convolution and attention for all data sizes. *Adv Neural Inf Process Syst*. 2021;34:3965–3977. Accessed August 21, 2024. https://arxiv.org/pdf/2106.04803v2.pdf
17. Zhai X, Kolesnikov A, Houlsby N, et al. Scaling vision transformers. In: *InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2022*. arXiv preprint:12104–12113. Accessed August 21, 2024. https://arxiv.org/pdf/2106.04560v1.pdf
18. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115(3):211–252.
19. Wang Y, Li K, Li Y, et al. *Internvideo: General Video Foundation Models Via Generative and Discriminative Learning*. arXiv Preprint; 2022. https://doi.org/10.48550/arXiv.2212.03191
20. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med*. 2012;22(3):276–282. Accessed August 21, 2024. https://search.ebscohost.com/login.aspx?direct=true&db=mdc&AN=23092060&site=ehost-live.
21. Fleiss JL, Levin B, Paik MC. The measurement of interrater agreement. *Stat Methods Rates Proportions*. 1981;2(212–236):22–23.
22. Joyce DW, Kormilitzin A, Smith KA, et al. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *NPJ Digit Med*. 2023;6(1):6.
23. Martino D, Karnik V, Bhidayasiri R, et al. Scales for antipsychotic-associated movement disorders: systematic review, critique, and recommendations. *Mov Disord*. 2023;38(6):1008–1026.
24. Hauser RA, Factor SA, Marder SR, et al. KINECT 3: a phase 3 randomized, double-blind, placebo-controlled trial of valbenazine for tardive dyskinesia. *Am J Psychiatry*. 2017;174(5):476–484.
25. Neurocrine Biosciences. Neurocrine Biosciences reports second quarter 2024 financial results and raises 2024 INGREZZA sales guidance. August 1, 2024. https://www.prnewswire.com/news-releases/neurocrine-biosciences-reports-second-quarter-2024-financial-results-and-raises-2024-ingrezza-sales-guidance-302211905.html
26. Teva Pharmaceuticals. Teva announces strong growth in second quarter revenues mainly driven by generics products in all regions and AUSTEDO; raises 2024 financial guidance. 2024. https://ir.tevapharm.com/news-and-events/press-releases/press-release-details/2024/Teva-Announces-Strong-Growth-in-Second-Quarter-Revenues-mainly-driven-by-Generics-Products-in-All-Regions-and-AUSTEDO-Raises-2024-Financial-Guidance/default.aspx