

Defining the Clinical Phenotypic Ground Truth: Expert Interrater Agreement in Hyperkinetic Movement Disorders

Madelein van der Stouwe, MD, PhD,^{1,2,3} Olof Vermeulen, BSc,^{1,2} Wieke Eggink, MD, PhD,^{1,2} Tove Henriksen, MD, PhD,⁴ Eavan McGovern, MD, PhD,^{5,6} Marenka Smit, MD, PhD,⁷ Selma Aybek, MD, PhD,^{8,9} Agnita Boon, MD, PhD,¹⁰ Rick Brandsma, MD, PhD,¹¹ Andres Deik, MD,¹² Roberto Erro, MD,¹³ Alberto Espay, MD, PhD,¹⁴ Cristian Falup-Pecurariu, MD, PhD,^{15,16} Stephen Frucht, MD,¹⁷ Christos Ganos, MD,^{18,19} Emilia Gatto, MD,^{20,21} Rick Helmich, MD, PhD,^{22,23} Christine Klein, MD,²⁴ Hans Koelman, MD, PhD,²⁵ Vincent Odekerken, MD, PhD,²⁵ Lisette Koens, MD, PhD,²⁶ Abhimanyu Mahajan, MD, MHS,¹⁴ Francesca Morgante, MD, PhD,²⁷ Alexander Münchau, MD, PhD,²⁸ Pattamon Panyakaew, MD, PhD,²⁹ Kathryn Peall, MBChB, PhD,^{30,31} Bélen Pérez-Dueñas, MD, PhD,^{32,33} Emanuel Roze, MD, PhD,³⁴ Anna Sadnicka, PhD, MBBS,^{35,36,37} Susanne Schneider, MD, PhD,³⁸ Petra Schwingenschuh, MD, PhD,³⁹ Tereza Serranova, MD, PhD,⁴⁰ Matej Skorvanek, MD, PhD,^{41,42} Maria Stamelou, MD, PhD,⁴³ Gertrúd Tamas, MD,⁴⁴ Michele Tinazzi, MD, PhD,⁴⁵ Anna Valadas, MD, PhD,^{46,47} Wim Vandenberghe, MD, PhD,^{48,49} Jeroen de Vries, MD, PhD,^{1,2} Rodi Zutt, MD, PhD,⁵⁰ Sterre van der Veen, MD, PhD,^{1,2} Rogier Luiken, MD,⁵¹ Inge Tuitert, MSc, PhD,^{1,2} and Marina Tijssen, MD, PhD, CA^{1,2,*}

Abstract: Background: Reliable clinical phenotyping is crucial for diagnosis, treatment, and research in hyperkinetic movement disorders, yet its interrater reliability remains uncertain.

¹Expertise Center Movement Disorders Groningen, University Medical Center Groningen, Groningen, the Netherlands; ²Department of Neurology, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands; ³Department of Neurology, Ommelanden Ziekenhuis Groningen, Scheemda, the Netherlands; ⁴Movement Disorder Clinic, University Hospital of Bispebjerg, Copenhagen, Denmark; ⁵Department of Neurology, Beaumont Hospital, Dublin, Ireland; ⁶School of Medicine, Royal College of Surgeons in Ireland, Dublin, Ireland; ⁷Department of Neurology, Medical Clinic Velsen, Velsen-Noord, the Netherlands; ⁸Department of Neurology, Faculty of Science and Medicine, University of Fribourg, Fribourg, Switzerland; ⁹Psychosomatic Medicine Unit, Department of Neurology, Inselspital Bern University Hospital, University of Bern, Bern, Switzerland; ¹⁰Department of Neurology, Erasmus MC Rotterdam, Rotterdam, the Netherlands; ¹¹Department of Pediatric Neurology, University Medical Center Utrecht, Utrecht, the Netherlands; ¹²Parkinson's Disease and Movement Disorders Center, University of Pennsylvania, Philadelphia, Pennsylvania, USA; ¹³Department of Medicine, Surgery and Dentistry "Saola Medica Salernitana," University of Salerno, Fisciano, Italy; ¹⁴James J. and Joan A. Gardner Family Center For Parkinson's Disease and Movement Disorders, University of Cincinnati, Cincinnati, Ohio, USA; ¹⁵Faculty of Medicine, Transilvania University, Brasov, Romania; ¹⁶Department of Neurology, County Clinic Hospital, Brasov, Romania; ¹⁷NYU Grossman School of Medicine, New York, New York, USA; ¹⁸Division of Neurology, Department of Medicine, Movement Disorders Clinic, Toronto Western Hospital, UHN, University of Toronto, Toronto, Ontario, Canada; ¹⁹Krembil Brain Institute, Toronto, Ontario, Canada; ²⁰Department of Clinical Medicine, Sanatorio de la Trinidad Mitre, Affiliated to the Faculty of Medicine of the University of Buenos Aires, Buenos Aires, Argentina; ²¹Parkinson's and Movement Disorders Section of the Institute of Neurosciences Buenos Aires, Buenos Aires, Argentina; ²²Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, the Netherlands; ²³Department of Neurology, Center of Expertise for Parkinson and Movement Disorders, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Centre, Nijmegen, the Netherlands; ²⁴Institute of Neurogenetics, University of Lübeck, Lübeck, Germany; ²⁵Department of Neurology, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, the Netherlands; ²⁶Department of Neurology and Clinical Neurophysiology, Martini Ziekenhuis, Groningen, the Netherlands; ²⁷Neurosciences and Cell Biology Institute, Neuromodulation and Motor Control Section, City St George's University of London, London, UK; ²⁸Institute of Systems Motor Science, CBBM, University of Lübeck, Lübeck, Germany; ²⁹Chulalongkorn Centre of Excellence for Parkinson's Disease and Related Disorders, Division of Neurology, Department of Medicine, Faculty of Medicine, King Chulalongkorn Memorial Hospital and Chulalongkorn University, Thai Red Cross Society, Bangkok, Thailand; ³⁰Laboratories of Neuroscience Research, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, UK; ³¹Neuroscience and Mental Health Innovation Institute, Cardiff University, Cardiff, UK; ³²Innovative Therapies in Pediatric Neurology Research Group, Vall d'Hebron Research Institute, Barcelona, Spain; ³³Department of Pediatrics, Universitat Autònoma de Barcelona, Barcelona, Spain; ³⁴Sorbonne University, INSERM, CNRS, Paris Brain Institute, AP-HP, Institute of Neurology, Paris, France; ³⁵Gatsby Computational Neuroscience Unit, University College London, London, UK; ³⁶Department of Clinical and Movement Neurosciences, University College London, London, UK; ³⁷National Hospital for Neurology and Neurosurgery, London, UK; ³⁸INSPIRE-PNRM+, Neuroimaging Center (NIC), University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany; ³⁹Department of Neurology, Medical University of Graz, Graz, Austria; ⁴⁰Department of Neurology, 1st Faculty of Medicine and General University Hospital in Prague, Charles University, Prague, Czech Republic; ⁴¹Department of Neurology, P.J. Safarik University, Kosice, Slovak Republic; ⁴²Department of Neurology, University Hospital of L. Pasteur, Kosice, Slovak Republic; ⁴³Parkinson's Disease and Movement Disorders Department, HYGELA, Athens, Greece; ⁴⁴Department of Neurology, Semmelweis University, Budapest, Hungary; ⁴⁵Biomedicine and Movement Sciences, Neurology Unit, Department of Neurosciences, University of Verona, Verona, Italy; ⁴⁶Department of Neurosciences and Mental Health, Neurology Hospital Santa Maria, CHLUN, Lisbon, Portugal; ⁴⁷Faculty of Medicine, Instituto de Medicina Molecular João Lobo Antunes, University of Lisbon, Lisbon, Portugal; ⁴⁸Department of Neurology, University Hospitals Leuven, Leuven, Belgium; ⁴⁹Department of Neurosciences, KU Leuven, Leuven, Belgium; ⁵⁰Department of Neurology, HagaZiekenhuis, The Hague, the Netherlands; ⁵¹Department of Rehabilitation Medicine, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands

***Correspondence to:** Prof. Dr. Marina Tijssen, Expertise Center Movement Disorders Groningen, Department of Neurology, University Medical Center Groningen, Hanzplein 1, 9700 RB, Groningen, the Netherlands; E-mail: m.a.j.de.koning-tijssen@umcg.nl

Keywords: tremor, myoclonus, dystonia, functional movement disorders, interrater agreement.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Received 28 July 2026; revised 28 July 2026; accepted 4 August 2026.

Published online 00 Month 2026 in Wiley Online Library ([wileyonlinelibrary.com](https://www.wileyonlinelibrary.com)). DOI: 10.1002/mdc3.70808

Objectives: To quantify interrater agreement in clinical phenotyping of hyperkinetic movement disorders among experts in a large, phenotypically diverse cohort.

Methods: Within the NEMO project, 161 patients with hyperkinetic movement disorders underwent standardized video assessments. Each patient was independently classified by three international movement disorders experts in a structured, stepwise procedure. Experts first assessed videos alone, then received clinical history and examination, and finally diagnostic test results. At each step, they assigned primary and secondary phenotypes and rated their diagnostic confidence. Interrater agreement was assessed using Fleiss' and Cohen's kappa.

Results: Across all phenotypes, interrater agreement on the primary phenotype increased with added information ($\kappa = 0.457$ to $\kappa = 0.607$). Agreement with the NEMO reference classification similarly improved ($\kappa = 0.478$ to $\kappa = 0.656$). Among patients with a single dominant phenotype, agreement reached $\kappa = 0.78$. In 93% of cases, at least two out of three experts reached consensus. Agreement was substantially lower in patients with mixed or complex presentations, such as myoclonus-dystonia and functional movement disorders. Diagnostic certainty rose throughout, though secondary phenotype agreement remained poor. Experts rarely revised initial impressions, suggesting robust pattern recognition but also potential confirmation bias.

Conclusions: Movement disorders specialists can reliably identify hyperkinetic phenotypes, even without in-person assessment. While complex, mixed and functional presentations remain diagnostically challenging, our findings reinforce the value of expert phenotyping in clinical care, training, and research, and provide a benchmark for future studies and computational approaches.

Movement disorders such as tremor, myoclonus, dystonia, chorea, and tics are all characterized by excessive involuntary movements. These movements hinder patients in their daily activities and negatively impact their quality of life. To the untrained eye, these hyperkinetic movement disorders may all appear similar. But to the trained eye of a neurologist—and especially to movement disorders experts—careful observation and classification of the phenotype are the cornerstone of patient care. Phenotyping shapes every step of the diagnostic trajectory: it informs decisions to conduct further laboratory testing, clinical neurophysiological examination, brain imaging, or genetic analysis.^{1–3} As pharmacological and surgical treatments in movement disorders are typically symptomatic and phenotype-specific, accurate classification is essential not only to the diagnosis, but to clinical outcomes. Deep brain stimulation, for instance, relies fundamentally on the movement disorder phenotype to select the optimal anatomical target.⁴

As Sadnicka and Edwards have recently pointed out, thousands of disease processes can lead to just a handful of movement disorder phenotypes.⁵ These core phenotypes (tremor, myoclonus, dystonia, tics, chorea, bradykinesia) each seem to represent a stable pattern of disordered movement with distinct clinical features. They propose that this convergence, in which many different conditions give rise to a limited number of phenotypes, suggests that the sensorimotor system may only “break” in a limited number of ways: an insight that underscores the value of phenotype classification in both clinical care and mechanistic research. That said, in certain clinical scenarios, accurate phenotypic classification itself serves as a shortcut to etiological diagnosis: recognizing alcohol-responsive myoclonus-dystonia as a distinct syndrome, for instance, immediately directs the clinician toward targeted genetic investigation rather than prolonged debate over the precise phenomenological boundaries between

myoclonus and dystonia. In most clinical situations, careful phenotyping remains the essential first step in guiding the diagnostic workup.

Phenotyping also holds a central place in the education of neurologists. Yet it remains a skill that resists easy quantification: shaped not by rule-following but by pattern recognition developed through experience, feedback and close mentorship with movement disorders specialists. In research, the need for robust and reproducible phenotyping is becoming increasingly urgent. Advances in pathophysiology, genetics, and computational neurology depend on reliable clinical labels.⁶ Whether in imaging studies, biomarker discovery, or training machine learning models, the quality of any analysis is fundamentally constrained by the accuracy of the phenotypic input.

Despite its importance, the reliability of expert-based phenotyping is not well quantified. What level of agreement can be expected among movement disorder experts? Most clinicians will have experienced a spirited debate between two specialists over whether a case is best described as “tremor” or “myoclonus”—a reminder that even expert opinion is subject to variation.

Several prior studies have investigated interrater agreement on movement disorder phenotyping, yet often in limited populations. Three studies focused on functional movement disorders. Van der Salm et al. assessed agreement among experts evaluating 60 patients with myoclonus, tics, and functional jerks, reporting moderate agreement ($\kappa = 0.56$; κ ranges from -1 to 1 , with values near 1 indicating strong agreement and values near 0 reflecting chance-level agreement⁷).⁸ Morgante et al found only fair agreement among both movement disorder and general neurologists classifying 28 patients with either functional or non-functional movement disorders, with improvement after access to clinical information.⁹ In the context of functional versus

primary tics, Rigas et al. reported fair agreement based on video assessment of 24 patients ($\kappa = 0.21$), which improved to moderate ($\kappa = 0.51$) with additional clinical context.¹⁰ Two other studies have examined pediatric populations. Eggink et al. found modest agreement ($\kappa = 0.35$) in children with spasticity, dystonia and ataxia,¹¹ while Yilmaz et al demonstrated significant discordance among pediatric neurologists, particularly when multiple phenotypes coexisted.¹²

Other studies have illustrated the diagnostic gray zones movement disorders specialists regularly encounter: Becktepe et al and Du et al both highlighted the particular difficulty of dystonia-adjacent cases: agreement on tremor syndromes varied by subspecialty, and subtle dystonic features in essential tremor patients were prone to divergent interpretation.^{13,14} Fearon et al further highlighted these challenges in the assessment of soft signs, showing that even among movement disorder specialists, agreement on the presence of subtle neurological abnormalities was only fair to moderate ($\kappa = 0.31$ – 0.42).¹⁵

If phenotyping is to support therapeutic development and mechanistic research its reliability must be demonstrable. Therefore, the current study builds on earlier work by examining interrater agreement in a large and diverse cohort of hyperkinetic movement disorders. Leveraging the data infrastructure of the Next Move in Movement Disorders (NEMO) project,¹⁶ we use a stepwise expert assessment design that includes a structured video protocol, clinical history, and diagnostic test results.¹⁷ This approach allows us to evaluate phenotyping reliability across both clearly recognizable and diagnostically challenging cases, with the aim of setting a benchmark for phenotyping reliability and to support efforts toward standardization in both clinical and research contexts.

Methods

This study was conducted at the Expertise Centre for Movement Disorders at the University Medical Center Groningen (UMCG), within the NEMO project.^{17,18} The broader aim is to develop a machine learning-assisted tool to support clinical phenotyping in movement disorders. The present substudy focuses on the reliability of expert-based visual classification using standardized clinical video assessments combined with additional clinical information. The study was approved by the Medical Ethical Committee of the UMCG (METc 2018-444) and conducted in accordance with the Declaration of Helsinki.

Patients

Patients were eligible for inclusion if they were aged ≥ 16 , and had a clinically confirmed hyperkinetic movement disorder with involvement of the upper limbs. Recruitment was carried out through the expertise centers outpatient clinic and patient database, and was extended to affiliated hospitals and patient associations where necessary. All patients provided written informed consent prior to enrolment.

Recruitment was structured around two broad phenomenological tiers. The first comprised patients with a single dominant

movement disorder phenotype: tremor, myoclonus, dystonia, ataxia, chorea, tics, spasticity or functional movement disorders—mirroring the classification options present to expert raters. Within this tier, specific clinical entities were selected as representative examples: essential tremor, enhanced physiological tremor, cortical myoclonus, benign idiopathic myoclonus, isolated dystonia and jerky/tremulous functional movement disorders. The second tier comprised patients with combined or mixed phenotype, for whom myoclonus-dystonia (with and without a confirmed *SGCE* mutation) served as the primary example. We aimed to include >20 patients in the groups of tremor, myoclonus and dystonia in the first tier, with a smaller number of patients with chorea, tics, ataxia, spasticity, dystonic tremor, parkinsonian tremor, and enhanced physiological tremor to reflect the diversity encountered in clinical practice. The group sizes were determined in part by the requirements of the broader NEMO project, which aimed to generate sufficient data for machine learning model development. Spasticity was included given the clinical overlap with dystonia that can make differentiation challenging; dystonic tremor and parkinsonian tremor were included to reflect the full spectrum of tremor presentations encountered in a tertiary movement disorders clinic, rather than restricting the cohort to phenotypically unambiguous cases.

Assessment Protocol

All patients underwent the full, standardized NEMO assessment protocol including clinical data collection, and video recordings.¹⁷ Video data were acquired using the Intel® RealSense™ D435 depth-sensing camera, which captures synchronized RGB and depth video (3D).

Each patient was recorded while performing a fixed sequence of 24 movement tasks selected to evoke a broad range of hyperkinetic movement disorders.¹⁹ These tasks included rest, maintenance of various postures, goal-directed actions, repetitive movements such as finger tapping, and performance under cognitive distraction. All recordings were conducted by trained instructors following a standardized protocol to ensure consistency.

Clinical and demographic data were collected using REDCap electronic data capture tools hosted at the UMCG.²⁰ Collected variables included age, sex, age at symptom onset, symptom duration, medication use, affected body regions, contributing factors and comorbidities. Additional data relevant to the diagnostic work-up—including findings from clinical neurophysiology, neuroimaging, and genetic testing—were extracted from the patients' clinical records and also entered into REDCap. These data were later on used to supplement the video-based ratings in a stepwise fashion.

Expert Ratings

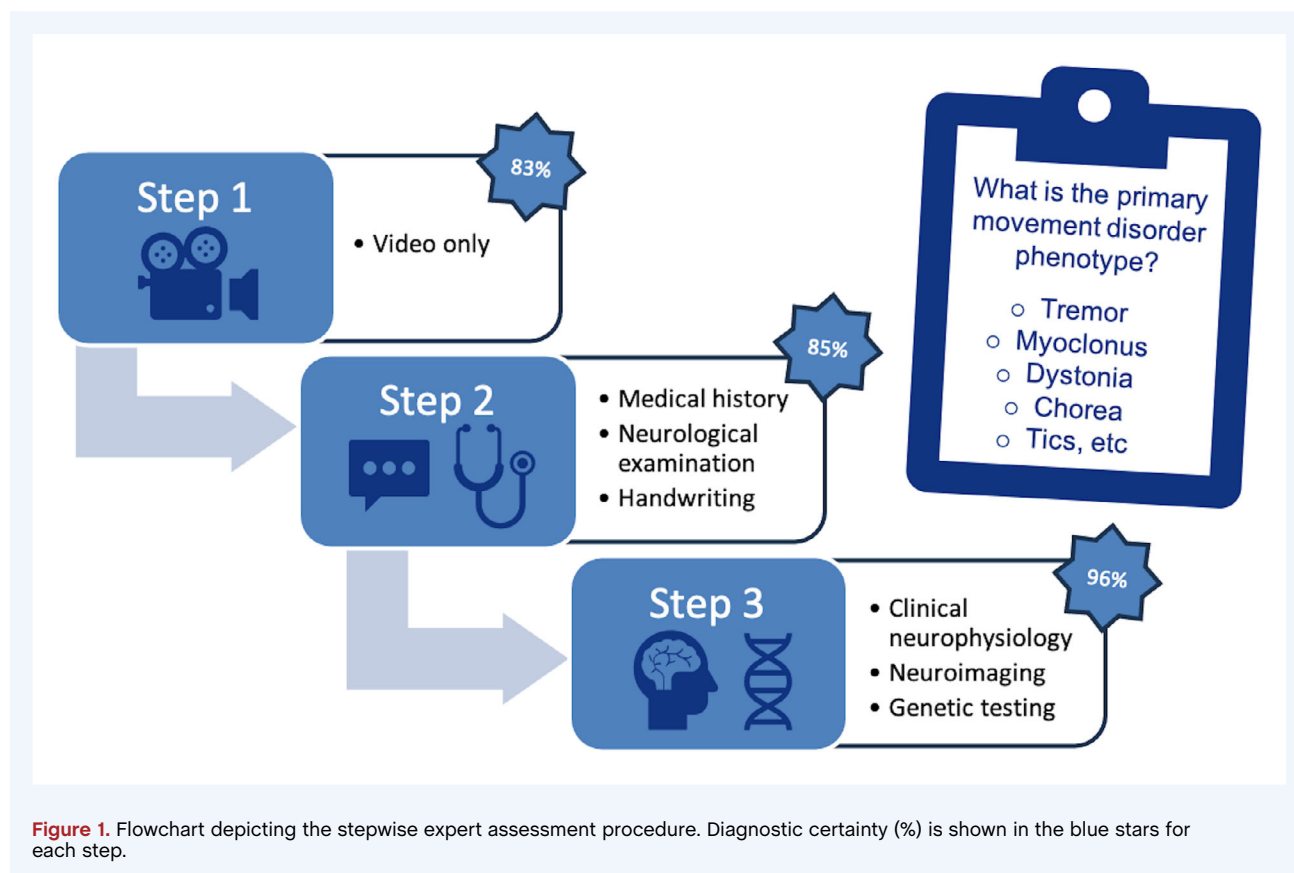
Expert assessment was carried out by an international panel of movement disorders specialists (authors HE-RZ). Experts were invited to participate based on predefined criteria: all are experienced movement disorders specialists with recognized clinical

and academic expertise in hyperkinetic movement disorders, and membership of the MDS. The final panel was selected following discussion between the senior investigator (AMMS and MdK), with the aim of ensuring international representation and broad subspecialty coverage. Experts were asked to assess between 10 and 15 patients; some participated twice as the study progressed. Every patient was independently rated by three different experts. To avoid bias, experts were blinded to the a priori classification assigned by the NEMO research team, which was made based on all available information.

Assessments were conducted in a structured, stepwise fashion (Fig. 1). Step 1 involved rating based on video recordings alone. In step 2, clinical information was added, simulating what a neurologist would typically have after a first consultation. This included a brief medical history, clinical information such as age and mode of onset, findings from general neurological examination, and a sample of handwriting and spiral drawing. In the final step 3, additional diagnostic data were provided, including results from clinical neurophysiology, neuroimaging, and genetic testing if available. At each of these three steps, experts were asked to determine (a) the primary movement disorder phenotype (choices: tremor, myoclonus, dystonia, ataxia, chorea, tics, spasticity, functional movement disorder); (b) whether a secondary phenotype was present (same options plus “none”); and (c) how confident they were in their phenotyping decision on a scale from 0% to 100%.

Statistics

All statistical analyses were performed using IBM SPSS Statistics, version 28 (IBM Corp., Armonk, NY, USA). To evaluate interrater agreement on the primary movement disorder phenotype among experts, Fleiss' kappa was calculated for each of the three diagnostic steps. Interpretation of kappa values followed standard thresholds: values $\kappa < 0.20$ were considered slight agreement, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and >0.80 almost perfect agreement.²¹ We analyzed the interrater agreement on the primary movement disorder phenotype among experts across the full cohort. In addition, observed agreement on the primary phenotype was also reported as the percentage of cases in which 2/3 or all three experts assigned the same phenotype. Agreement between individual expert classifications and the reference phenotype assigned by the NEMO team was assessed using Cohen's kappa for each step, using the same thresholds for interpretation as for Fleiss' kappa. Diagnostic certainty was summarized per step and tested for change over the course of the rating process using Friedman's analysis of variance, and, in case of significant differences, subsequent pairwise comparisons using Bonferroni corrected significance values. Similarly, we investigated whether diagnostic certainty differed in cases dependent on the observed agreement among experts using an independent-samples Kruskal-Wallis test with pairwise comparisons. Agreement on the secondary movement disorder



phenotype was evaluated using Fleiss' κ at each diagnostic step. Finally, we recorded the number of diagnostic changes made by each expert between steps to further explore the effect of added clinical information on classification stability. $P < 0.05$ were considered statistically significant. It should be noted that kappa values are sensitive to the prevalence distribution of categories in the sample; in unbalanced cohorts, kappa may underestimate observed agreement. Observed agreement percentages are therefore reported alongside kappa values to also provide a more intuitive measure. The reference phenotype assigned by the NEMO team represents a clinical consensus classification based on all available information, rather than an independent biological gold standard. As no such gold standard exists for the majority of hyperkinetic movement disorders, Cohen's kappa in this context reflects agreement with the best available clinical reference rather than diagnostic accuracy in an absolute sense.

As a post hoc analysis, we also examined interrater agreement on the primary movement disorder phenotype among experts in three clinically relevant subgroups: patients with single phenotypes only, patients with myoclonus dystonia (as an example of a mixed phenotype), and patients with functional jerky or tremulous movement disorders (as a potentially diagnostically challenging group). These subgroup classifications were assigned by the NEMO research team based on clinical consensus among the team prior to expert assessment. The post hoc analyses were performed to explore whether agreement patterns differed in phenotypically more or less complex cases, and to support interpretation of the findings across distinct clinical scenarios.

Results

A total of 161 patients were included in this study, covering a broad range of hyperkinetic movement disorders. Patients included those exhibiting tremor ($n = 49$; clinical diagnoses included essential tremor ($n = 29$), enhanced physiological tremor ($n = 10$), parkinsonian tremor ($n = 8$), Holmes tremor ($n = 2$)), myoclonus ($n = 26$: cortical myoclonus ($n = 22$), benign idiopathic myoclonus ($n = 4$)), dystonia ($n = 22$), tics ($n = 2$), chorea ($n = 2$), ataxia ($n = 2$), and spasticity ($n = 2$). Among the patients with myoclonus-dystonia ($n = 30$), 17 had a confirmed *SGCE*-mutation and 13 did not. Four patients were included with dystonic tremor. Finally, a group of patients with jerky/tremulous functional movement disorders ($n = 22$) was included. Demographic and clinical characteristics of the included patients are presented in Table 1. A total of 38 international movement disorders experts participated, each assessing a median of 11 patients (IQR 10–13, range 7–30), resulting in 483 assessments.

Interrater Agreement among Experts

Expert interrater agreement on the primary phenotype ranged from a Fleiss' κ of 0.457 based on videos only, to 0.525 after

receiving additional clinical information, and 0.607 after disclosing all diagnostic test results (Fig. 2).

Observed Agreement among Experts

Full agreement on the primary phenotype among 3/3 experts ranged from 24% of patients based on videos only, 25% after receiving additional clinical information, to 27% upon receiving all diagnostic test results. Partial agreement among 2/3 experts ranged from 88% based on videos only, 89% with additional clinical information, and 93% after receiving all diagnostic test results (Fig. 2).

Interrater Agreement between Experts and the NEMO Team

Assessment of the interrater agreement between expert classifications and the reference primary phenotype assigned by the NEMO team resulted in a Cohen's κ of 0.478 based on videos only, 0.566 after additional clinical information, and 0.656 after receiving all diagnostic test results.

Diagnostic Certainty

When asked to rate the perceived certainty of their primary phenotype classification, the experts' certainty increased from a median 83% (Q1: 65%–Q3: 95%) in step 1, to 85% (70%–96%, $p < 0.001$) in step 2, to 92% (77%–100%, $p < 0.001$) in step 3 (Fig. 3). Diagnostic certainty increased with expert agreement: 78% in cases with no agreement, 87% in cases with 2/3 agreement and 96% in cases with 3/3 agreement ($p < 0.001$).

Secondary Phenotypes

Experts identified a secondary phenotype in 39% of patients in step 1, 41% in step 2, and 42% in step 3. Interrater agreement among experts on the secondary phenotype was slight, as indicated by $\kappa = 0.085$ in step 1, 0.101 in step 2, and 0.091 in step 3.

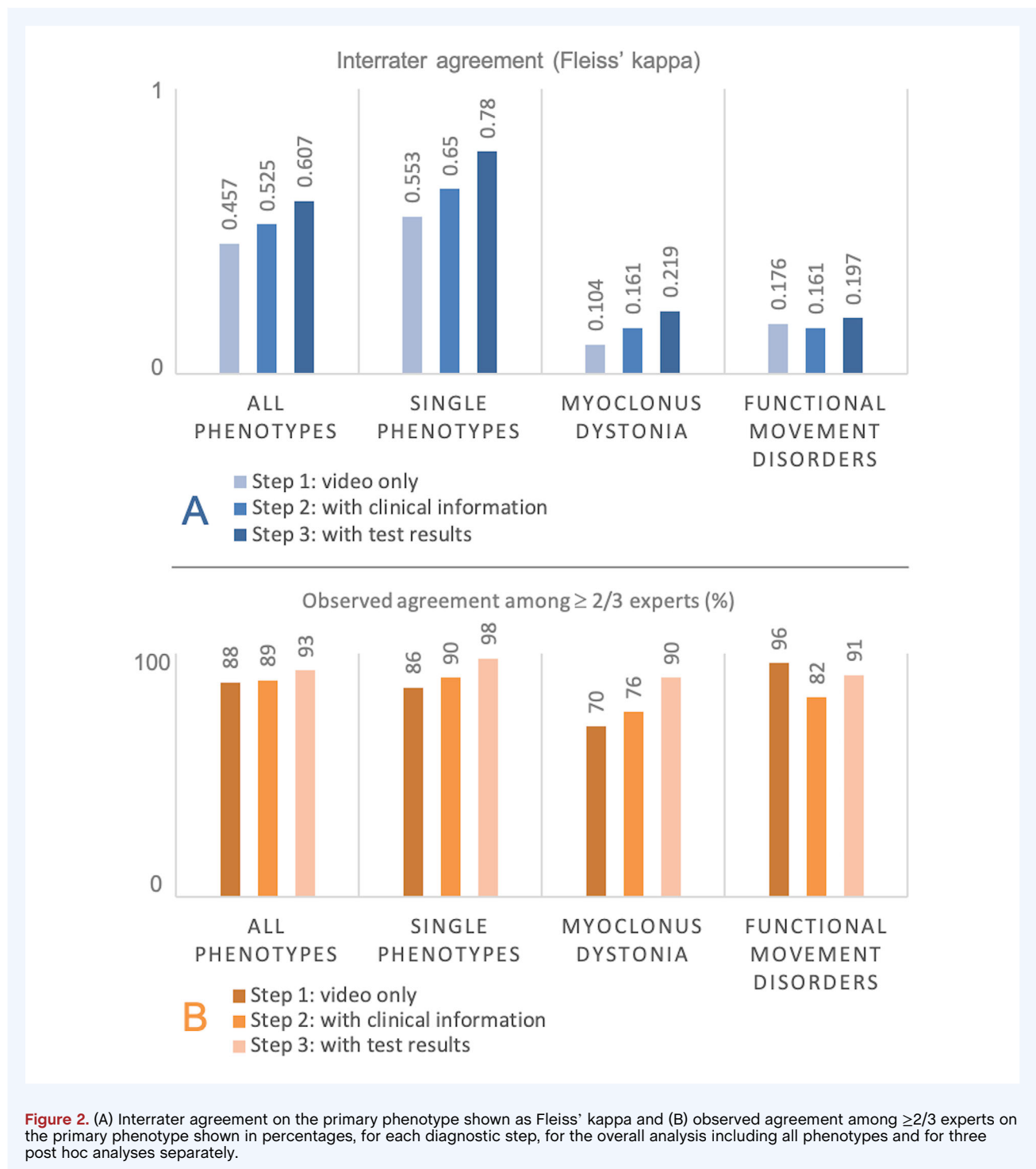
Changes in Classification

Experts could change their phenotype classification twice: 12% of classifications were changed after receiving additional clinical information, whereas 11% of classifications were changed upon receiving all diagnostic test results. In total, this resulted in a total of 113 changes in classification, with most changes in primary phenotype between tremor and myoclonus (23%), from chorea, dystonia and tics to myoclonus (25%), and from functional movement disorders to dystonia, tremor, myoclonus and spasticity (19%) (Fig. 3).

TABLE 1 Clinical characteristics and test results

	Tremor (n = 49)	Myoclonus (n = 26)	Dystonia (n = 22)	Myoclonus dystonia (n = 30)	Tics (n = 2)	Chorea (n = 2)	Ataxia (n = 2)	Spasticity (n = 2)	FMD (n = 22)
Age	65 (25) range 19–81	24.5 (20) range 16–77	46 (36) range 16–76	27 (40) range 17–75	31 range 31–31	55.5 range 37–74	63.5 range 62–65	46 range 30–62	63 (23) range 25–75
Female sex	31%	69%	50%	34%	50%	50%	50%	50%	41%
Age at onset	39 (44)	12 (15)	14 (35)	6 (8)	5	30	52	21.5	45.5 (29)
Disease duration	13 (30.5)	13 (9.5)	16 (16)	23 (36.5)	26	25.5	11.5	24.5	12.5 (21.5)
Family history of movement disorders	59%	31%	18%	83%	100%	50%	100%	0%	52%
Polymyography	17	20	10	8	0	0	0	0	13
Fits with phenotype									
Not performed	32	6	12	22	2	2	2	2	11
Cortical origin	0	12	0	2	0	0	0	0	0
No cortical origin	0	8	0	5	0	0	0	0	1
Not performed	49	6	22	23	2	2	2	2	21
Brain MRI	2	2	1	0	0	0	1	2	0
Relevant lesions	12	12	12	3	0	1	0	0	6
No relevant lesions	35	12	9	27	2	1	1	0	16
FDOPA-PET	6	0	0	0	0	0	0	0	0
Dopaminergic deficit	5	0	1	0	0	0	0	0	0
Normal	38	26	21	30	2	2	2	2	22
Not performed	1	8	4	18	0	1	2	0	0
Relevant variant	4	11	7	6	0	1	0	0	0
No relevant variants	44	7	11	6	2	0	0	2	22
Not performed									
Genes	PARK-GBA1	ATX/MYC- NUS1, n = 2 ATX-MT-ATP6 MARCH6 SCN1A, n = 2 STARD7 16p11.2	DYT-TOR1A DYT-KMT2B DYT-THAP1 DYT-TUBB4A	MYC/DYT- SGCE, n = 17 Silver Rusel syndrome	-	CHOR-NKX2-1	ATX-ATXN3, n = 2	-	-

Clinical characteristics are reported as median (interquartile range) or percentage, test results as counted frequencies. FMD: functional movement disorder. Genes are listed in line with the International Parkinson and Movement Disorders Society genetic nomenclature where possible.



Post Hoc Subgroup Analyses

In the subgroup of patients with a single dominant movement disorder phenotype ($n = 95$), interrater agreement on the primary phenotype increased across steps from $\kappa = 0.553$, to 0.650, to 0.780. This subgroup comprised patients with a dominant phenotype of tremor (essential tremor [$n = 29$], enhanced physiological tremor [$n = 10$]), myoclonus (cortical myoclonus

[$n = 22$], benign idiopathic myoclonus [$n = 4$]), dystonia (isolated dystonia [$n = 22$], tics [$n = 2$], chorea [$n = 2$], ataxia [$n = 2$], spasticity [$n = 2$]). Each clinical diagnosis was thus grouped under the corresponding phenomenological category used in the expert classification framework; experts were asked to classify at this phenomenological level, and distinguishing between specific entities within a category, (eg, essential tremor vs. enhanced physiological tremor), was not part of the rating

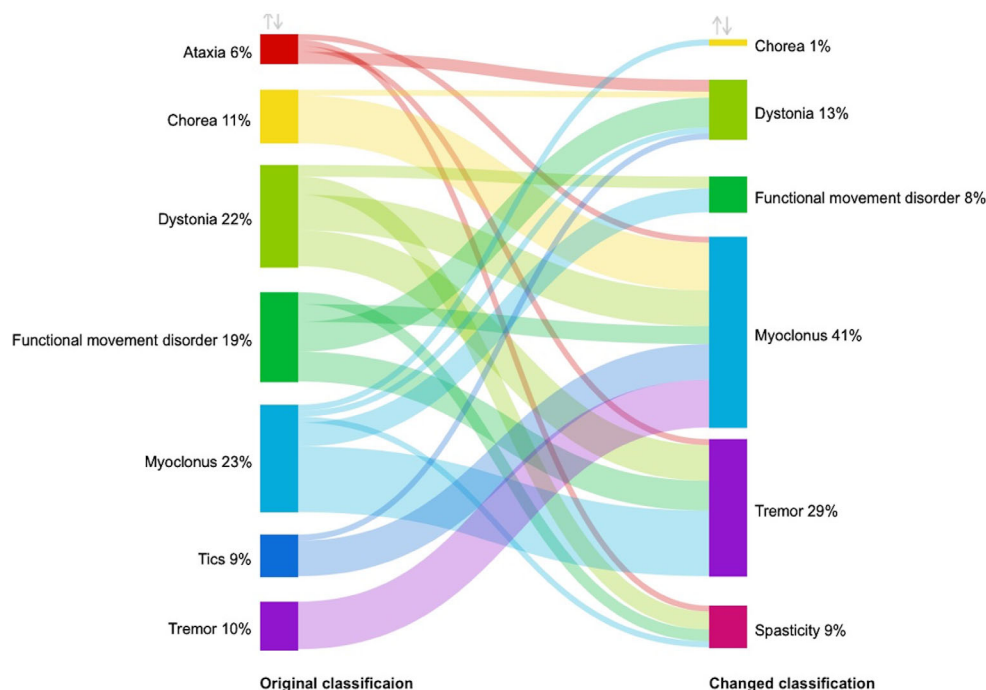


Figure 3. Sankey plot depicting changes in classification ($n = 113$) of the primary movement disorder phenotype across the entire assessment process, with the experts' original classification on the left and the changed classification on the right.

task. For patients with myoclonus dystonia, interrater agreement was lower, ranging from $\kappa = 0.104$, to 0.161 – 0.219 . In this group, after step 3, 63% of experts selected myoclonus as the primary phenotype, of whom 46% reported no secondary movement disorder and 37% identified dystonia as secondary. Conversely, 18% selected dystonia as the primary phenotype in step 3, of whom 44% chose myoclonus as secondary and 38% no secondary phenotype (see Supplementary Materials for an overview of myoclonus dystonia classifications). In the functional jerky/tremulous movement disorder group, interrater agreement remained low across all steps ($\kappa = 0.176$, 0.161 , and 0.197). After full information disclosure in step 3, 55% of experts identified a functional movement disorder as the primary phenotype, while others chose tremor (21%), myoclonus (8%), dystonia (8%), chorea (6%), or ataxia (3%). Functional movement disorder was selected as a secondary phenotype in only 3% of assessments.

Observed agreement followed a similar pattern. For single phenotypes, full agreement (3/3 experts) increased from 25% to 30% to 32%, whereas partial agreement (at least 2/3 experts) ranged from 86% to 90% to 89%. In the myoclonus dystonia group, full agreement was reached in 7%, 3% and 3% of patients across steps, while partial agreement improved from 70% to 76% to 90%. For functional movement disorders, full agreement remained stable at 18% across all steps, while partial agreement was 96% in step 1, 82% in step 2, and 91% in step 3.

Diagnostic certainty mirrored these trends. In single phenotype patients, median certainty increased from 82% (IQR

69–95%) in step 1, to 87% (70–98%) in step 2 ($p < 0.001$), to 94% (80–100%) in step 3 ($p < 0.001$). For myoclonus dystonia, median certainty was 80% (65–93%) in step 1, 80% (70–95%) in step 2 ($p = 0.656$), and increased to 95% (75–100%) in step 3 ($p = 0.008$). In functional movement disorders, diagnostic certainty remained stable across steps at 80% (61–90%), 81% (64–90%), and 81% (62–94%) ($p = 0.259$).

Discussion

This study provides a detailed evaluation of interrater agreement on movement disorder phenotyping in the largest and most clinically diverse cohort of patients with hyperkinetic movement disorders to date. Across this spectrum, agreement among movement disorders specialists on the main phenotype after receiving all clinical information was substantial. In 93% of cases, at least two out of three experts independently reached consensus, underscoring the reliability of expert judgment even in the context of video-based assessments.

These findings extend prior work on interrater agreement on myoclonus, tics and functional movement disorders,^{8–10} and on classification of movement disorders in children.^{11,12} Previous studies highlighted the challenges of clinical classification, with kappa values never extending beyond 0.56. In comparison, our data suggest that higher agreement may be more readily observed

in broader clinical populations: our cohort included a wide range of hyperkinetic movement disorders, beyond classifications of functional versus non-functional or complex pediatric presentations, both of which may pose particular diagnostic challenges. With $\kappa = 0.61$ across the cohort, and $\kappa = 0.78$ in patients with a single movement disorder phenotype, agreement was substantial, and partial consensus in a least 2/3 experts can nearly always be expected. These findings suggest that our shared language, training, and clinical frameworks support reliable phenotyping of core movement disorder phenotypes. However, this encouraging picture is tempered by the low rate of full agreement across the cohort as a whole, and by the notably poor agreement in mixed and functional presentations, where only slight agreement was reached.

Important differences were found between agreement levels in patients with single phenotypes and those with mixed features. As expected, agreement was highest in cases with a clear single dominant phenotype, such as essential tremor, cortical myoclonus or isolated dystonia, reflected in a Fleiss' kappa of 0.78, and lower in cases with mixed presentations such as myoclonus-dystonia and functional movement disorders where only slight agreement was reached ($\kappa = 0.22$ and 0.20 , respectively). These differences highlight the added complexity and subjectivity involved in classifying patients with overlapping symptoms, where even experienced specialists may disagree on their interpretation. This complexity was further reflected in the poor interrater agreement on the secondary movement disorder phenotype, suggesting that experts differed in how they weighed additional features. Together, these results may in part reflect the cognitive limits of visual diagnostic reasoning: experts must integrate multiple simultaneously occurring movement features, of varying prominence and clarity, while drawing on pattern recognition gained from experience. Studies in medical decision-making suggest that while clinicians are highly skilled at recognizing prototypical presentations, ambiguity or competing cues can reduce consistency.^{22–24} In the context of mixed movement disorders, these factors could contribute to differences in interpretation.

The particularly low agreement in functional movement disorders warrants separate consideration. Functional movement disorders occupy an ambiguous position in phenotypic classification: while it is in many respects an etiological diagnosis, it is also associated with specific clinical signs (such as distractibility, variability, and entrainment) that can be considered phenotypic markers in their own right. Recognition of these signs, however, depends on targeted video tasks; while our protocol included relevant tasks such as contralateral tapping and cognitive distraction, we cannot ascertain whether all experts systematically reviewed these specific clips, and furthermore, a standardized fixed-duration protocol cannot replicate the intuition-guided flexibility of a clinical examination, where an experienced clinician can extend or adapt distraction tasks until a clear picture emerges. This may have contributed to the low agreement in this group, and is consistent with lower agreement on functional movement disorders reported in prior literature.^{8–10} The very low rate of functional movement disorder as a secondary phenotype (identified in only 3% of assessments) further reflects this: in this study, experts were unlikely to identify functional features as an accompaniment to another primary phenotype.

Experts reported high clinical certainty across the assessment process, often already after video-only evaluation. Certainty increased mildly across the stepwise procedure: from 82% to 93% certainty, except for functional movement disorders where certainty remained at 81% throughout. The combination of stable but modest certainty and persistently low interrater agreement in the functional movement disorders group suggests that individual experts were confident in their own classifications while reaching different conclusions: a pattern consistent with high subjective confidence in the face of genuine diagnostic ambiguity, and one that merits attention in future classification frameworks. Similarly, most experts did not revise their initial classification when clinical information was added: 88–89% of all phenotype classifications remained unchanged when experts were presented with the opportunity to change their mind. This is in line with previous findings by van der Salm et al.,²⁵ who similarly observed that movement disorders specialists relied predominantly on initial clinical impressions and neurological examination to guide decision-making. This consistency likely reflects the robustness of pattern recognition among experienced clinicians. At the same time, it raises the possibility that early diagnostic impressions may shape subsequent interpretation, leading to confirmation bias, with clinicians primarily noticing features that confirm their initial classification, at the expense of critical re-evaluation.²⁶ Observer-level cognitive biases, including anchoring on an initial impression and framing effects where the context in which information is presented influences its interpretation, may contribute to the variability seen across experts, independent of phenotypic complexity. These biases are well recognized in clinical neurology and are relevant to any diagnostic process that relies on sequential information gathering.²⁷ Regarding changes in classification, myoclonus emerged as the most reclassified disorder, which highlights how difficult it can be to distinguish myoclonus from other movement disorders, particularly tremor, and emphasizes the importance of neurophysiology in these patients.²⁸

It is important to acknowledge that video-based assessment differs in key ways from in-person evaluation. Several experts commented that they felt constrained by the inability to interact with patients, repeat tasks, or hear the patient's history in their own words. In-person assessments also allow for undirected observation, which may reveal signs not captured on video. Additionally, our stepwise information flow (video first, then clinical data) differs from the usual clinical reasoning process. While this format was chosen to foreground phenomenology, even higher agreement might be possible in more naturalistic settings. Still, the substantial agreement between experts and the *a priori* NEMO classification is encouraging. It suggests that structured video recordings, combined with stepwise clinical information, can provide a sufficiently rich dataset to support reliable phenotyping, reinforcing the validity of video-based assessments in research settings.

Our study has other limitations. Although the cohort is large and phenotypically diverse, it is not evenly distributed across phenotypes. Disorders such as tremor, myoclonus, dystonia, myoclonus-dystonia, and functional movement disorders are well represented, while tics, chorea, ataxia, and bradykinesia are underrepresented. This imbalance reflects both the focus areas of

our expertise center and the group size requirements of the broader NEMO project, in which adequate representation of the primary phenotype groups tremor, myoclonus and dystonia was prioritized to support machine learning model development. In addition, technical issues affected data quality in some cases. For example, a microphone failure led to missing speech data for part of the cohort, and the use of a 3D camera optimized for depth recording came at the cost of lower-quality RGB video. Moreover, the quality of video streaming within the VRW depended on users' internet connection, which may have impacted the rating experience for some experts.

Nonetheless, our study has several strengths as well. It is based on a large and well-characterized cohort, with structured, extended video and clinical information provided to experts in a controlled, stepwise manner. The use of secure data platforms such as REDCap and the VRW ensured high standards of data integrity and protection. Importantly, our expert panel consisted of experienced raters from different national backgrounds, strengthening the generalizability of our results.

The results provide a reassuring foundation for both clinical practice, training and research. In line with Sadnicka and Edwards,⁵ who emphasize that recognizable “substates” of disordered movement arise across diverse etiologies, our field relies on clinicians learning to identify these stable patterns. Our results support this principle: pattern recognition in movement disorders phenotyping appears to be a teachable and reproducible skill. In cases where diagnostic uncertainty remains, video case conferences—already quite common in clinical practice—may complement individual judgment to strengthen phenotyping reliability. In research contexts, our data support the use of expert consensus as a valid method for patient inclusion. In the NEMO study, we defined reliable phenotypes based on agreement between at least two of three experts and a Fleiss' $\kappa > 0.8$ across a group.¹⁷ The results presented here lend empirical support to that approach and may guide similar efforts in future machine learning research.

Conclusion

This study shows that movement disorders specialists can reliably identify hyperkinetic phenotypes, even in the absence of in-person assessment. While complex, mixed and functional presentations remain diagnostically challenging, our findings reinforce the value of expert phenotyping in clinical, training and research settings, and provide a benchmark for future studies and computational approaches.

Author Roles

(1) Research Project: A. Conception, B. Organization, C. Execution; (2) Statistical Analysis: A. Design, B. Execution, C. Review and Critique; (3) Manuscript Preparation: A. Writing of the First Draft, B. Review and Critique.

MS: 1A, 1B, 2A, 2B, 3A.

OV: 1B, 3B.

WE: 1C, 3B.
 TH: 1C, 3B.
 EMG: 1C, 3B.
 M.Smit: 1C, 3B.
 SA: 1C, 3B.
 AB: 1C, 3B.
 RB: 1C, 3B.
 AD: 1C, 3B.
 RE: 1C, 3B.
 AE: 1C, 3B.
 CFP: 1C, 3B.
 SJF: 1C, 3B.
 CG: 1C, 3B.
 EG: 1C, 3B.
 RH: 1C, 3B.
 CK: 1C, 3B.
 HK: 1C, 3B.
 VO: 1C, 3B.
 LK: 1C, 3B.
 A.Mahajan: 1C, 3B.
 FM: 1C, 3B.
 A.Munchau: 1C, 3B.
 PP: 1C, 3B.
 KP: 1C, 3B.
 BP: 1C, 3B.
 ER: 1C, 3B.
 AS: 1C, 3B.
 SS: 1C, 3B.
 PS: 1C, 3B.
 TS: 1C, 3B.
 M.Skorvanek: 1C, 3B.
 M.Stamelou: 1C, 3B.
 GT: 1C, 3B.
 MT: 1C, 3B.
 AV: 1C, 3B.
 WV: 1C, 3B.
 JV: 1C, 3B.
 RZ: 1C, 3B.
 SV: 1B, 3B.
 RL: 1B, 3B.
 IT: 1A, 1B, 2C, 3B.
 M.Tijssen: 1A, 2A, 2C, 3B.

Acknowledgments

We thank Sebastian Sippach and Paula Swaving for their assistance in organizing the expert assessments.

Disclosures

Ethical Compliance Statement: The study was approved by the Medical Ethical Committee of the UMCG (METc 2018-444) and conducted in accordance with the Declaration of

Helsinki. All patients provided written informed consent prior to enrolment. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this work is consistent with those guidelines.

Financial Disclosures and Conflicts of Interest: This work was supported by the European Fund for Regional Development of the European Union in collaboration with the province of Fryslan (grant number 01492947), and a ZonMW TOP Grant (grant number 91218013). The authors declare no competing interests relevant to the submitted manuscript.

AMMS reports grant support from ZonMw-Off Road and ZonMw-Klinische fellowship. AE has received grant support from the NIH and the Michael J. Fox Foundation; personal compensation as a consultant/scientific advisory board member for Mitsubishi Tanabe Pharma America (formerly, Neuroderm), Amneal, Acorda, Abbvie, Bial, Supernus (formerly, USWorldMeds), NeuroDiagnostics, Inc. (SYNAPS Dx), Intrance Medical Systems, Inc., Merz, Praxis Precision Medicines, Citrus Health, and Herantis Pharma; Data Safety Monitoring Board (chair) of AskBio; and publishing royalties from Lippincott Williams & Wilkins, Cambridge University Press, and Springer. He is co-inventor of the patent "Compositions and methods for treatment and/or prophylaxis of proteinopathies." He cofounded REGAIN Therapeutics to fund preclinical studies but relinquished the right to any personal income from future treatments. He serves on the editorial boards of the Journal of Parkinson's Disease, Journal of Alzheimer's Disease, European Journal of Neurology, Movement Disorders Clinical Practice, and JAMA Neurology. FM received research support from NIHR, Innovate UK. Speaking honoraria from Abbvie, Medtronic, Boston Scientific, Bial, Merz; Travel grants from the International Parkinson's disease and Movement Disorder Society; Advisory board fees from Merz and Boston Scientific; Consultancies fees from Boston Scientific, Merz and Bial; Research support from Boston Scientific, Merz; Royalties for the book "Disorders of Movement" from Springer; member of the editorial board of Movement Disorders, Movement Disorders Clinical Practice, European Journal of Neurology, European Journal of Neurology. AM has received funding from the Parkinson's foundation, Sunflower Parkinson disease Foundation and the Dystonia Medical Research Foundation, outside of the submitted work. He served on the scientific advisory board of adaptive biosciences, through the Parkinson's Study Group. He served as a consultant for Vima therapeutics Inc. He receives a stipend for serving as an associate editor for the journal, *Movement Disorders Clinical Practice*. ER received honorarium for speech from Orkyn, Aguetant, Elivie, Merz-Pharma, Janssen, Teva and for participating in advisory boards from Merz-Pharma, Elivie, Teva, and BIAL, and he received research support from Merz-Pharma, Orkyn, Elivie, Everpharma, Aguetant. GT was partly supported by grant No: TKP2021-EGA-25. MAJT reports grants from the European Fund for Regional Development from the European Union (01492947) and the province of Friesland, ZonMw-TOP (91218013), from Stichting Wetenschapsfonds, and unrestricted grants from Merz.

Author disclosures are available in the Supporting Information.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request. ■

References

1. Zutt R, van Egmond ME, Elting JW, et al. A novel diagnostic approach to patients with myoclonus. *Nat Rev Neurol* 2015;11(12):687–697. <https://doi.org/10.1038/nrneuro.2015.198>.
2. Van Egmond ME, Kuiper A, Eggink H, et al. Dystonia in children and adolescents: a systematic review and a new diagnostic algorithm. *J Neurol Neurosurg Psychiatry* 2015;86(7):774–781. <https://doi.org/10.1136/jnnp-2014-309106>.
3. van de Wardt J, van der Stouwe AMM, Dlrkx M, et al. Systematic clinical approach for diagnosing upper limb tremor. *J Neurol Neurosurg Psychiatry* 2020;91(8):822–830. <https://doi.org/10.1136/jnnp-2019-322676>.
4. Krack P, Volkmann J, Tinkhauser G, Deuschl G. Deep brain stimulation in movement disorders: from experimental surgery to evidence-based therapy. *Mov Disord* 2019;34(12):1795–1810. <https://doi.org/10.1002/MDS.27860>.
5. Sadnicka A, Edwards MJ. Between nothing and everything: phenomenology in movement disorders. *Mov Disord* 2023;38(10):1767–1773. <https://doi.org/10.1002/MDS.29584>.
6. De A, Bhatia KP, Volkmann J, Peach R, Schreglmann SR. Machine learning in tremor analysis: critique and directions. *Mov Disord* 2023; 38(5):717–731. <https://doi.org/10.1002/MDS.29376>.
7. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012;22(3):276–282. <https://doi.org/10.11613/bm.2012.031>.
8. van der Salm SMA, de Haan RJ, Cath DC, van Rootselaar AF, Tijssen MAJ. The eye of the beholder: inter-rater agreement among experts on psychogenic jerky movement disorders. *J Neurol Neurosurg Psychiatry* 2013;84(7):742–747. <https://doi.org/10.1136/jnnp-2012-304113>.
9. Morgante F, Edwards MJ, Espay AJ, Fasano A, Mir P, Martino D. Diagnostic agreement in patients with psychogenic movement disorders. *Mov Disord* 2012;27(4):548–552. <https://doi.org/10.1002/MDS.24903>.
10. Rigas A, Mainka T, Pringsheim T, et al. Distinguishing functional from primary tics: a study of expert video assessments. *J Neurol Neurosurg Psychiatry* 2023;94(9):751–756. <https://doi.org/10.1136/jnnp-2022-330822>.
11. Eggink H, Kremer D, Brouwer OF, et al. Spasticity, dyskinesia and ataxia in cerebral palsy: are we sure we can differentiate them? *Eur J Paediatr Neurol* 2017;21(5):703–706. <https://doi.org/10.1016/j.ejpn.2017.04.1333>.
12. Yilmaz S, Vermilion J, Dean S, Pourdeyhimi R, Mink JW. Inter-rater agreement for movement disorder classification in children with hyperkinetic movement disorders. *Mov Disord Clin Pract* 2024;11(12):1598–1603. <https://doi.org/10.1002/MDC3.14252>.
13. Becktepe J, Gövert F, Balint B, Schlenstedt C, Bhatia K, Elble R, Deuschl G. Exploring interrater disagreement on essential tremor using a standardized tremor elements assessment. *Mov Disord Clin Pract* 2021;8(3): 371–376. <https://doi.org/10.1002/MDC3.13150>.
14. Du W, Bain PG, Defazio G, et al. The conundrum of dystonia in essential tremor patients: how does one classify these cases? *Tremor Other Hyperkinet Mov (N Y)* 2022;12(1):15. <https://doi.org/10.5334/TOHM.690>.
15. Fearon C, Espay AJ, Lang AE, et al. Soft signs in movement disorders: friends or foes? *J Neurol Neurosurg Psychiatry* 2019;90(8):961–962. <https://doi.org/10.1136/jnnp-2018-318455>.
16. van den Brandhof EL, Tuitert I, Madelein van der Stouwe AM, et al. Explainable machine learning for movement disorders – classification of tremor and myoclonus. *Comput Biol Med* 2025;192(Pt B):110180. <https://doi.org/10.1016/j.COMPBIOMED.2025.110180>.
17. van der Stouwe AMM, Tuitert I, Giotis I, et al. The next move in movement disorders (NEMO): developing a computer aided classification tool for hyperkinetic movement disorders. *BMJ Open* 2021;11(10): e055068. <https://doi.org/10.1136/bmjopen-2021-055068>.
18. Dalenberg JR, Peretti DE, Marapin LR, van der Stouwe AMM, Renken RJ, Tijssen MAJ. Next move in movement disorders: neuroimaging protocols for hyperkinetic movement disorders. *Front Hum Neurosci* 2024;18:18. <https://doi.org/10.3389/FNHUM.2024.1406786>.
19. Sciacca G, van der Stouwe AMM, Centen LM, et al. Next move in movement disorders (NEMO): the best clinical tasks for the visibility of

- essential tremor, dystonia, cortical myoclonus and myoclonus-dystonia. *Parkinsonism Relat Disord* 2025;138:107963. <https://doi.org/10.1016/J.PARKRELDIS.2025.107963>.
20. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. *J Biomed Inform* 2019;95:103208. <https://doi.org/10.1016/j.jbi.2019.103208>.
 21. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;76(5):378–382. <https://doi.org/10.1037/h0031619>.
 22. Norman GR, Eva KW. Diagnostic error and clinical reasoning. *Med Educ* 2010;44(1):94–100. <https://doi.org/10.1111/J.1365-2923.2009.03507.X>.
 23. Eva KW. What every teacher needs to know about clinical reasoning. *Med Educ* 2005;39(1):98–106. <https://doi.org/10.1111/J.1365-2929.2004.01972.X>.
 24. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Arch Intern Med* 2005;165(13):1493–1499. <https://doi.org/10.1001/ARCHINTE.165.13.1493>.
 25. van der Salm SMA, van Rootselaar AF, Cath DC, de Haan RJ, Koelman JHTM, Tijssen MAJ. Clinical decision-making in functional and hyperkinetic movement disorders. *Neurology* 2017;88(2):118–123. <https://doi.org/10.1212/WNL.0000000000003479>.
 26. Sweller J. Cognitive load during problem solving: effects on learning. *Cogn Sci* 1988;12(2):257–285. https://doi.org/10.1207/S15516709COG1202_4.
 27. Vickrey BG, Samuels MA, Ropper AH. How neurologists think: a cognitive psychology perspective on missed diagnoses. *Ann Neurol* 2010; 67(4):425–433. <https://doi.org/10.1002/ana.21907>.
 28. Latorre A, van der Veen S, Pena A, et al. IAPRD new consensus classification of myoclonus. *Parkinsonism Relat Disord* 2025;132:107216. <https://doi.org/10.1016/J.PARKRELDIS.2024.107216>.

Supporting Information

Supporting information may be found in the online version of this article.

Data S1. Supporting Information.

Data S2. Table on myoclonus dystonia classifications.

Each row represents the classifications given by an individual rater: 3 raters for each patient. Myoclonus dystonia, n = 30 patients, 90 assessments.