Title: **Assessing infant risk of cerebral palsy with video-based motion tracking**

Authors: Melanie Segado, PhD[1,2,5], Laura Prosser, PT, PhD[*4,5], Andrea F. Duncan, MD, MS[4,6], Michelle J. Johnson, PhD[1,7,8,9], Konrad P. Kording, PhD[1,2]

1. Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, United States
2. Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, United State
3. Department of Physical Therapy, The Children's Hospital of Philadelphia, Philadelphia, PA, USA
4. Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
5. Division of Rehabilitation Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA, USA
6. Division of Neonatology and Department of Pediatrics, Children's Hospital of Philadelphia
7. Department of Physical Medicine and Rehabilitation, University of Pennsylvania, Philadelphia, PA, USA
8. Department of Mechanical Engineering and Applied Mechanics, University of Pennsylvania, Philadelphia, PA, USA
9. Rehabilitation Robotics Lab, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

*corresponding author

E-mail: prosserl@chop.edu

Phone: 215-590-2495

Location:

Children's Seashore House
Room 208A
3405 Civic Center Blvd
Philadelphia, PA 19104

Abstract:

Cerebral Palsy (CP) is a common (about 1 in 500 children) health condition caused by abnormal brain development that affects the ability to control movement. Early risk assessment happens through the General Movements Assessment (GMA), a test administered by trained clinicians at 3-4 months of age that has high predictive value for CP. With recent improvements in video-based motion tracking, automated risk assessment for CP based on the GMA is being explored. However, studies generally have used small datasets or were limited in terms of methodological rigor. Here we acquired a large dataset (1060 infants) of videos from a clinical population with elevated CP risk. In a preregistered pipeline using a lock-box set that was not used before algorithm submission we find that our machine learning predictions are highly predictive of the clinician-assessed GMA (AUC=0.79). Given its low cost, our video-based approach may be useful for clinical screening applications, particularly in low-resource settings.

Introduction

Cerebral Palsy (CP) is the most common cause of motor impairment leading to physical disability in children, affecting an estimated 2-3 out of 1000 infants globally[1]. In the USA alone, this results in approximately 1 million people living with impaired mobility due to CP at any given time, many of which have lifelong disability. Early detection and rehabilitation before two years of age is critical, as beginning rehabilitation within this sensitive period for neural plasticity and motor development is associated with functional outcomes [2,3]. Atypical movement patterns that indicate a high risk of developing CP are reliably detectable through visual observation of movements by a trained physician at or before 10 weeks of age, but many infants are not evaluated by a physician until after severe, overt motor impairments have already developed. In practice, this means that CP is typically diagnosed between 6 to 24 months of age, which is near the end of the optimal window for intervention. There is, therefore, a need to develop automated early pre-screening tools that can detect atypical patterns of motor development before they progress to more severe impairment, allowing for more efficient use of costly medical resources and improved outcomes particularly in low resources settings.

CP risk is routinely assessed by clinicians based on visual observation of movements. One such assessment is the General Movement Assessments (GMA)[4], which is predictive of CP as early as 3 months of age based on the expert classification of spontaneous infant movements. It distinguishes between "typical" and "atypical" general movements (GMs), including the identification of "fidgety movements" (FMs) at 3-4 months, which are a precursor to coordinated, volitional movement. The absence of FMs at this age is 95% predictive of CP when combined with abnormal findings on brain MRI[2]. The GMA assessment is typically scored from video and considers characteristics of movement quality, variability, and complexity. If these relevant movement features can be reliably computed from videos, then algorithmic approaches for predicting infant risk from movement features should work robustly.

Many parallel efforts by various research groups are underway to automate GMA assessment using video-derived skeletal tracking[5–17]. However, the potential for these approaches to scale beyond the dataset on which they were trained is currently limited. Existing models rely on hand-annotated, or custom fine-tuned models, which are specific to each research group's dataset. The advent of pre-trained vision transformers has enabled better feature extraction and multi-scale information fusion, leading to improved performance on data with occlusions and poses and joint or limb segment occlusions, both of which are common in spontaneous infant movement and cause significant issues for infant pose estimation algorithms[18–23]. The combined advancements in deep learning, open datasets, and open-source tools have significantly improved the reliability and accuracy of pose estimation and tracking outcomes[24,25]. Pre-trained vision transformers should be sufficiently good to capture the movement features that are relevant for clinical assessment, without the need for custom models that risk overfitting.

Existing video-based automated risk assessment models often perform well, but are limited either in terms of sample size, generalizability, or methodological rigor. For instance, Gao et al. (2023) trained a transformer model on clips of hand-labeled movements and counted the proportion of video clips in their sample labeled as FMs after training[6]. This approach was highly effective at detecting FMs, however, as they noted this approach cannot be extended without

the need for retraining on other hand-labeled segments. Others such as Ihlen et al. (2019) found high levels of sensitivity and specificity, comparable to clinician GMA assessments, but the model relied on a backward prediction of over 900 features, raising the concern that the precise featurization may overfit to the specific dataset[26,27]. Moreover, none of these models with promising results have been made fully openly available, limiting the extent to which generalization to a new dataset can be tested, and none employed a "lock-box" set (ie. held-out data points that were not used at any point during the hyperparameter optimization process), raising the possibility that results are overly optimistic[28]. There is, therefore, still a need to test whether the effects observed in the literature replicate in a large sample, with explainable features, and a pre-registered analysis pipeline.

We addressed these limitations using a large dataset of clinician-labeled videos from our institutions' United States CP Early Detection and Intervention Network site data. To assess risk based on video data we developed a classification pipeline. To compute accurate movement features, we started by selecting an open-source pose estimation algorithm that had high precision on our infant dataset. Out of the 2D pose estimates we, based on clinician feedback, computed 38 features that described posture, velocity, acceleration, left-right symmetry, and complexity of movements. We show that movement features can predict GMA scores in the largest infant dataset used to date, using an automated machine learning approach that limits bias in hyperparameter optimization, and a fully pre-registered pipeline.

Methods

Ethics approval for this study was provided by the University of Pennsylvania (Penn) Institutional Review Board (IRB Protocol Number: 833180), acting as the single IRB or record and a subsequent reliance agreement between Penn and the Children's Hospital of Philadelphia (CHOP) Institutional Review Board (IRB Protocol Number: 19-016641).

Collection of a large clinical dataset

Data were collected as part of standard clinical care by team members of the CHOP site of the US CP Early Detection and Intervention Network in a REDcap database between May 2019 and December 2023. This included the secure uploading of iPad- or iPhone-recorded videos, GMA scores and demographic information. Access to this clinical database was restricted to hospital staff. The GMA was administered in accordance with CHOP's participation in the Cerebral Palsy Foundation's Early Detection and Intervention network, who follow the international diagnostic guidelines [2]. For all infants who were between 10-20 weeks post-term age (corrected for preterm birth, if applicable) at the time of a clinic visit, and whose parents or legal guardians agreed to video recording for clinical care, clinicians captured a 1-2 minute video of the infant lying in supine. This is the usual age for an infant's first visit with the Neonatal Follow-up Program high-risk infant follow-up clinic. Infants were observed in minimal attire for unobstructed visibility of the trunk, shoulders, and extremities to facilitate the observation of natural movements (typically wearing a diaper only. The use of pacifiers, toys, or engagement in communication with the infant during the assessment was prohibited and other distractions that could potentially influence the outcome were minimized. If patients missed their clinic visit

during this time period, parents were instructed how to capture the video and provided a link to upload the video into REDcap.

The evaluation process was characterized by the involvement of over 20 clinicians, including physical and occupational therapists, nurse practitioners and physicians, who had completed training in GMA assessment by the General Movements Assessment Trust, several with additional advanced training. The GMA score (FMs present, absent or abnormal) was determined after adjudication by two independent clinician reviewers. In instances where disparities in assessment arose, a third evaluator was consulted. Videos eliciting uncertainty were further examined in weekly meetings convened by the site's early detection team.

For children still hospitalized at the time of the fidgety-aged GMA, Early Detection Team members captured the videos in the hospital as part of standard care. Exclusions were applied to intubated patients, those under the influence of sedation medications, within a week post-operative, on ECMO support, or diagnosed with myelomeningocele. The full dataset comprised 1060 infants. 129 were then excluded for meeting one or more exclusion criteria listed above. Six infants with a GMA score of 3 ("atypical fidgety") were also excluded, since there were not enough infants in this group for model training/testing, and movement patterns differ from those of "absent fidgety" infants. The remaining 931 videos were split into an analysis set (744) and a lock box holdout set (187) (Figure 1.). The analysis set was further split into train/val/test sets (558, 93, 93), each of which had a 12% representation of the "absent fidgety" movement type. The splits were stratified to preserve the ratios of boy/girl infants, as well as age, and race/ethnicity. There was a total recording duration of 60-120 s per infant.

Developing pipeline for robust skeletal tracking

To estimate pose from monocular hand-held video, we implemented a top-down 2D pose estimation pipeline using tools from open-source library OpenMMLab, MMDetection was used for infant detection[25], and MMPose was used for 2D pose estimation[24]. Infant detection was carried out using an RTMDet[29] model pre-trained for person detection trained on the Common Objects in Context (COCO) dataset[30]. 2D frame-wise pose estimation was carried out using ViTPose[18], a 10B parameter standard vision transformer that has been shown to be robust and domain-adaptable compared with previously state-of-the-art approaches like OpenPose[31]. We found that it had robust out of the box performance on the infant video data compared to several other models we tested[31–35], obviating the need for fine-tuning (Figure 2). However, to ensure best performance, only the first (highest-confidence) detection was considered from each frame, and only frames where all 17 key points (OpenPose coordinate system) for the detection were above a confidence threshold of 0.8 were considered. As in Chambers et al. (2020)[5] missing frames were linearly interpolated, outliers were removed using a rolling-median filter (1 second window) and smoothed using a rolling-mean filter (1 second window). Kinematic features were then computed from the resulting smooth 2D pose estimates using python code[36]. Pose estimation was conducted using high performance computing servers at CHOP, ensuring that all (identifiable) video data remained within the CHOP hospital system and accessible only to CHOP staff in accordance with ethics guidelines.

Computing Movement Feature Vector

Following pose estimation, infant IDs were divided into training, validation, and lockbox test sets using a stratified split to preserve a 12% representation of the absent fidgety movement type, and preserve the ratio of Male/Female infants, and Race/Ethnicity described in the dataset description. The video IDs corresponding to each split were pre-registered prior to carrying out any additional analyses, along a set of 38 clinician-selected kinematic features, which were used in previously published work[5,37,38]. Following pre-registration, the kinematic features were computed from the smoothed keypoint timeseries using the open-source python code adapted from Chambers et al. (2020)[5,36]. The kinematic features captured the displacement, speed, velocity, acceleration, and entropy of the extremities (wrists/ankles) and joint angles (elbows/knees) (Figure 3), which were described by clinicians as capturing the relevant information needed for visual scoring of the GMA[4]. No features specific to the GMA (FMs) were computed to minimize the risk of overfitting to the clinical dataset.

Model Training

A binary classifier was trained to predict which infants had a GMA score of "absent fidgety", which indicates higher risk of developing CP[2–4,39,40]. Because decisions made by researchers during feature selection, model selection and hyperparameter optimization may lead to overly optimistic results on trained classifiers even when no data leakage has occurred[27], feature selection was done in consultation with clinicians prior to any data analysis and pre-registered in 2018[38]. Model selection and hyperparameter optimization were done using automated machine learning package AutoSklearn 2.0[41,42] with "Vanilla AutoSklearn" settings, which restricted the ensemble size to one such that the current best model would always be chosen according to its performance on the validation set. Balanced Accuracy was selected as the optimization metric due to the imbalance in class sizes (roughly 10:1)[43]. Meta-feature free Portfolios were used for efficient meta-learning, and training/validation splits were automatically selected using successive halving [41]. 5-folds cross validation (standard way of validating ML results) was carried out to ensure that model performance was generalizable across different data splits within the training/validation set, and the resulting trained model was pre-registered on May 22, 2024, prior to testing on the held-out lock box data[37].

Results

Patient characteristics

To ask how well we could predict GMA score from clinician-selected movement features in a large sample, we recruited 1060 participants from the Children's Hospital of Pennsylvania. The sample of 1060 infants was sex-balanced, with 55% girls, 45% boys, and 1% Unknown/Unspecified. It also comprised a wide range of race/ethnicities, including White (38%), Black/African American (35%), Other (10%), and all other responses (Multi-Racial, Asian, Indian, American

Indian/Alaskan, Native Hawaiian, and Not Reported/Unknown/Other) making up 16% of responses. Of reported ethnicities, 8% were Hispanic/Latino.

Each of the video recordings used for analysis was determined evaluable by the clinical reviewers. In cases where infants were distracted during the recording session, a second video was obtained. Only the final videos used for GMA scoring were considered in this dataset. Average infant age was mean corrected age 14.6 weeks (+/- 2.1 weeks). Of the 931 infants that remained after applying exclusion criteria, 820 were scored as having FMs (normal) and 105 infants were scored with having absent FMs. The remaining six infants were scored as having abnormal movements and were excluded from further analyses.

Pre-trained vision transformers led to robust skeletal tracking

To develop a classification pipeline, we first had to get reliable, accurate skeletal tracking from video. Videos of infants are challenging for pose-estimation algorithms due to the high frequency of irregular body poses, the tendency for there to be many face/body-like objects in frame (toys, dolls, cartoons), and the high level of occlusion (primarily self-occlusion during rolls, variable limb movements or crunching). Algorithms like OpenPose often fail to identify body parts when they are occluded, or mis-attribute their location if another body-like object is in frame (Figure 2, left), leading to unreliable estimates of pose and movement[21]. To some extent, these limitations can be overcome with fine tuning on each infant dataset; however, we found that a large pre-trained vision transformer, ViTPose-H, performed better than any fine-tuned algorithm we tested, fully obviating the need for fine-tuning on the infants, and generalized much better to new, unseen datasets (Figure 2, right). This was validated through iterative inspection of 2D pose estimates by clinicians trained in the GMA and tested for generalizability on 2 fully out-of-sample infant datasets from other experiments (total of 280 infants 0 - 4 months). We thus have a fast, scalable, and precise way of converting videos into skeletal tracking.

A simple movement feature vector was sufficient for clinical score prediction

The recognition algorithm needs as inputs a description of an infant's movement, a so-called feature vector, which is computed from the outputs of the skeletal tracking pipeline. This feature vector was based on extensive interviewing of expert clinicians, pre-registered in 2018 [5,37,38] and published as part of a CP risk-assessment pipeline based on Bayesian Surprise[5,38]. All of the features were selected and pre-registered prior to collecting the infant videos in order to eliminate any risk of overfitting. All 38 movement features were computed for each infant using python code which has been made available on github[36]. We found a high degree of overlap across the features for infants with and without FMs (Fig. 3), with no feature clearly distinguishing between GMA scores. In other words, every individual feature fails to reveal if a given infant is at high risk of developing CP.

Decisions made by researchers during feature selection, model selection and hyperparameter optimization may lead to overly optimistic results on trained classifiers even when no data leakage has occurred[27]. To minimize these procedural overfitting risks, we used a pre-registered set of clinician-selected features, and a very well-established standard machine learning process (Auto-sklearn 2.0[41], also pre-registered). For the machine learning classifier we used "vanilla" auto-sklearn with default settings and a duration parameter of 1h[42]. Using this approach, the model was able to achieve a AUC-ROC of 0.72 on the test set. To verify that this result was generalizable within the training dataset, we ran 5-folds cross-validation runs using 6 different random seeds, which gave an average AUC-ROC of 0.73 ± 0.05. This is the standard way in which medical machine learning scientists report results and represents the rate of True Positives relative to False Positives. In other words, the feature vector contains sufficient information to predict GMA scores.

Many machine learning models have shown similar results, however these approaches still suffer from the problem that machine learning experts often try multiple algorithms, selecting only the one that works best on their specific dataset. This can make even the best-performing models useless on new data. We took several steps to avoid this. Prior to any model training, we randomly selected another test set, a so-called 'lock box' set of 187 infants (22 with absent FMs and therefore higher risk for CP) which the team did not use until all features, pre-processing steps, and algorithms were pre-registered. This yielded an AUC-ROC of 0.79 (Figure 4), a value very similar to the AUC-ROC of the machine learning pipeline we obtained in cross validation. In other words, there is no sign that we have any degree of overfitting to the test-set. We thus obtained good performance in a setting where a lack of overfitting can be guaranteed.

Discussion

Here we have developed an ML algorithm to predict GMA score (a strong indicator of CP risk) from video-based pose estimates using rigorous methods. We used an exceptionally big sample (training set: 558, overall >1000 infants), a simple and explainable movement-based feature vector, and we pre-registered each step of the process prior to testing on a randomly-selected "lock-box" set of 187 infant videos. We found that our algorithm performs well (AUC-ROC 0.79). We have utilized an AutoML approach to minimize the risk of overfitting. We have further minimized the risk of false positive data using a lock-box set and pre-registered our analysis before running it. We have made data and algorithms publicly available on the OSF pre-registration site, Github, and Figshare. Based on our rigorous pre-registered approach with a lock-box set we can be confident that we did not do any overfitting and that it will generalize well to other datasets.

While the GMA has been shown to have a high level of sensitivity and specificity in clinical settings, we did not predict the main important target future outcome – diagnosis of CP – as long-term outcomes were not available at the time of model training. Instead we predicted GMA, a clinician powered risk measure. This is common throughout the automated CP risk prediction literature, with multiple research groups focusing on predicting GMA score, or detecting Fidgety Movements directly, as opposed to predicting CP diagnosis. This approach is not ideal, as it introduces an additional source of noise from potential human error during assessment, in addition to the noise inherent in the GMA assessment itself. FMs, while highly indicative, are still not a perfect biomarker for CP and multiple items are necessary

for CP diagnosis (biomarkers, clinical history, functional motor assessment and neurological assessment). Over-reliance on FMs risks missing other, perhaps more indicative features or combinations of features that are not readily apparent. Moreover, the extremely low prevalence of Abnormal FMs makes training a model that captures this movement type infeasible, meaning that some infants at high risk are not accounted for in models trained only to detect FMs (or their absence). Future efforts should focus directly on predicting CP outcomes.

In order to detect subtle movement differences, it is likely that our features are suboptimal. The clinician-selected movement features offer only a coarse description of movement, whereas we know from the clinical literature that the difference between infants whose movements are typically developing and those that are not is often subtle. This is especially true if we push towards early-prediction, before the 3-4 months where the difference between movements is captured by the GMA, or push towards models that work for widespread pre-screening in the general population. Many efforts have been made to identify a precise featurization using machine learning [14,15], however all of these efforts risk overfitting since the size of the datasets is very small. By contrast, deep learning models trained on very large datasets of infant movements, as well as important context like clinical history and other assessments) promise to give more precise feature vectors that capture these subtle differences, boosting performance of the GMA prediction model and enabling even earlier prediction of CP.

The wide range of ages at which CP is typically diagnosed reflects the fact that less severe movement deficits are often not evident to untrained observers until later in an infant's development when they start missing major milestones, whereas indicators of more severe impairment may be evident to clinicians (and caregivers) much earlier. The infants included in the model all spent time after birth in the Neonatal Intensive Care Unit (NICU), meaning that they were already at an elevated risk of CP. This limitation is prevalent throughout the automated CP detection literature [8,10,11,16], since collecting videos of infants for the purposes of training a ML prediction model is most feasible in a hospital setting. As such the movements that distinguish the two groups in our sample may not be representative of infants from the general infant population. Future work can address this by including infants from the general population as well as those from the higher risk NICU cohort. However, other people have shown that movement features can be used to predict GMA scores in at-home videos of infants that are not at high risk[15], so the approach should generalize if trained on the bigger sample. This should be imminently feasible now that we have released a pose estimation and preprocessing pipeline that is open, easy to share, and does not require fine-tuning across videos of infants in different settings and at different ages.

We have shown that a simple movement-based automated prediction approach works in an exceptionally big sample (>1000 infants). This is significant since models are often trained on very small datasets (< 50) which risk overfitting and limit generalizability. Given that the clinician-selected movement features can predict GMA score in such a large sample, we have a strong indication that models that include even more data from a wider sample of infants, and more precise features, should perform even better.

To further minimize the risk of overfitting, we pre-registered all of our analysis pipeline, and our model, prior to testing on a "lock-box" set. Both of these steps minimize the risk of selecting features that are unique to a specific dataset, or selecting features that are biased by the pre-processing steps taken prior to model training. Such bias results in many machine learning models being overfit, and failing when applied to an out-of-sample dataset. Pre-registration and lock-box testing greatly reduce this risk, and should be standard for developing machine learning pipelines in clinical contexts.

All of the methods used are ethologically doable on a phone camera. Data collection was done using a hand-held iPad camera, and the pose estimation pipeline was tested both on videos from this dataset, as well as ~300 infant videos from other datasets to ensure generalizability across various contexts. While video-based pose estimation for infant movement estimation is common in the automated detection literature, each site typically uses their own custom fine-tuned algorithm with a post-processing pipeline that is tailored to their specific dataset. The pre-trained vision transformer we used was not fine-tuned on any of the infant videos, and as such should work equally well at other clinical sites and on at-home videos. Training on datasets across multiple sites and various contexts should now be possible, and drive towards a globally available at-home pre-screening tool.

Overall we have shown that advances in pose estimation now make it entirely realistic to get precise movements from infant videos without the need for any specialized camera setup or fine-tuning. We have shown that movement features derived from these pose estimates are sufficient for predicting GMA scores in a very large sample, and that our model generalizes well to unseen data. This holds tremendous potential for the creation of a global prescreening tool, especially if we boost performance using deep-learned feature vectors and train on more data, including from many infants across various contexts, and includes CP outcomes as opposed to just clinical scores.

References:

1. Sarah McIntyre, Goldsmith S, Webb A, et al. Global prevalence of cerebral palsy: A systematic analysis. *Dev Med Child Neurol*. 2022;64(12):1494-1506. doi:10.1111/dmcn.15346
2. Novak I, Morgan C, Adde L, et al. Early, Accurate Diagnosis and Early Intervention in Cerebral Palsy: Advances in Diagnosis and Treatment. *JAMA Pediatr*. 2017;171(9):897-907. doi:10.1001/jamapediatrics.2017.1689
3. Herskind A, Greisen G, Nielsen JB. Early identification and intervention in cerebral palsy. *Dev Med Child Neurol*. 2015;57(1):29-36. doi:10.1111/dmcn.12531
4. Einspieler C, Prechtl HFR. Prechtl's assessment of general movements: A diagnostic tool for the functional assessment of the young nervous system. *Ment Retard Dev Disabil Res Rev*. 2005;11(1):61-67. doi:10.1002/mrdd.20051
5. Chambers C, Seethapathi N, Saluja R, et al. Computer Vision to Automatically Assess Infant Neuromotor Risk. *IEEE Trans Neural Syst Rehabil Eng*. 2020;28(11):2431-2442. doi:10.1109/TNSRE.2020.3029121
6. Gao Q, Yao S, Tian Y, et al. Automating General Movements Assessment with quantitative deep learning to facilitate early screening of cerebral palsy. *Nat Commun*. 2023;14(1):8294. doi:10.1038/s41467-023-44141-x
7. Adde L, Brown A, Broeck C van den, et al. In-Motion-App for remote General Movement Assessment: a multi-site observational study. *BMJ Open*. 2021;11(3):e042147. doi:10.1136/bmjopen-2020-042147
8. Hashimoto Y, Furui A, Shimatani K, et al. Automated Classification of General Movements in Infants Using a Two-stream Spatiotemporal Fusion Network. Published online July 4, 2022. Accessed September 27, 2023. http://arxiv.org/abs/2207.03344
9. Groos D, Adde L, Aubert S, et al. Development and Validation of a Deep Learning Method to Predict Cerebral Palsy From Spontaneous Movements in Infants at High Risk. *JAMA Netw Open*. 2022;5(7):e2221325. doi:10.1001/jamanetworkopen.2022.21325
10. Irshad MT, Nisar MA, Gouverneur P, Rapp M, Grzegorzek M. AI Approaches towards Prechtl's Assessment of General Movements: A Systematic Literature Review. *Sensors*. 2020;20(18):5321. doi:10.3390/s20185321
11. Kwong AKL, Doyle LW, Olsen JE, et al. Parent-recorded videos of infant spontaneous movement: Comparisons at 3–4 months and relationships with 2-year developmental outcomes in extremely preterm, extremely low birthweight and term-born infants. *Paediatr Perinat Epidemiol*. 2022;36(5):673-682. doi:10.1111/ppe.12867
12. Morais R, Le V, Morgan C, et al. Robust and Interpretable General Movement Assessment Using Fidgety Movement Detection. *IEEE J Biomed Health Inform*. Published online 2023:1-12. doi:10.1109/JBHI.2023.3299236
13. Nguyen-Thai B, Le V, Morgan C, Badawi N, Tran T, Venkatesh S. A Spatio-temporal Attention-based Model for Infant Movement Assessment from Videos. *IEEE J Biomed Health Inform*. 2021;25(10):3911-3920. doi:10.1109/JBHI.2021.3077957
14. Passmore E, Kwong AL, Greenstein S, et al. Automated identification of abnormal infant movements from smart phone videos. *PLOS Digit Health*. 2024;3(2):e0000432. doi:10.1371/journal.pdig.0000432
15. Redd CB, Karunanithi M, Boyd RN, Barber LA. Technology-assisted quantification of movement to predict infants at high risk of motor disability: A systematic review. *Res Dev Disabil*. 2021;118:104071. doi:10.1016/j.ridd.2021.104071
16. Silva N, Zhang D, Kulvicius T, et al. The future of General Movement Assessment: The role of computer vision and machine learning – A scoping review. *Res Dev Disabil*. 2021;110:103854. doi:10.1016/j.ridd.2021.103854
17. Spittle AJ, Olsen J, Kwong A, et al. The Baby Moves prospective cohort study protocol: using a smartphone application with the General Movements Assessment to predict neurodevelopmental outcomes at age 2 years for extremely preterm or extremely low birthweight infants. *BMJ Open*. 2016;6(10):e013446. doi:10.1136/bmjopen-

2016-013446
18. Xu Y, Zhang J, Zhang Q, Tao D. ViTPose++: Vision Transformer for Generic Body Pose Estimation. Published online December 14, 2023. doi:10.48550/arXiv.2212.04246
19. Liu W, Bao Q, Sun Y, Mei T. Recent Advances of Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective. *ACM Comput Surv*. 2022;55(4):80:1-80:41. doi:10.1145/3524497
20. Wei K, Kording KP. Behavioral tracking gets real. *Nat Neurosci*. 2018;21(9):1146-1147. doi:10.1038/s41593-018-0215-0
21. Seethapathi N, Wang S, Saluja R, Blohm G, Kording KP. Movement science needs different pose tracking algorithms. Published online July 23, 2019. doi:10.48550/arXiv.1907.10226
22. Hesse N, Bodensteiner C, Arens M, Hofmann UG, Weinberger R, Schroeder AS. Computer Vision for Medical Infant Motion Analysis: State of the Art and RGB-D Data Set. In: *Computer Vision - ECCV 2018 Workshops*. Springer International Publishing; 2018.
23. Ostadabbas S. Fine-tuned Domain-adapted Infant Pose (FiDIP). Published online August 17, 2023. Accessed September 6, 2023. https://github.com/ostadabbas/Infant-Pose-Estimation
24. MMPose Contributors. OpenMMLab Pose Estimation Toolbox and Benchmark. Published online August 2020. https://github.com/open-mmlab/mmpose
25. MMDetection Contributors. OpenMMLab Detection Toolbox and Benchmark. Published online August 2018. https://github.com/open-mmlab/mmdetection
26. Ihlen EAF, Støen R, Boswell L, et al. Machine Learning of Infant Spontaneous Movements for the Early Prediction of Cerebral Palsy: A Multi-Site Cohort Study. *J Clin Med*. 2020;9(1):5. doi:10.3390/jcm9010005
27. Powell M, Hosseini M, Collins J, et al. I Tried a Bunch of Things: The Dangers of Unexpected Overfitting in Classification. Published online February 14, 2020:078816. doi:10.1101/078816
28. Hosseini M, Powell M, Collins J, et al. I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neurosci Biobehav Rev*. 2020;119:456-467. doi:10.1016/j.neubiorev.2020.09.036
29. Lyu C, Zhang W, Huang H, et al. RTMDet: An Empirical Study of Designing Real-Time Object Detectors. Published online 2022. https://arxiv.org/abs/2212.07784
30. Lin TY, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context. Published online February 20, 2015. doi:10.48550/arXiv.1405.0312
31. Cao Z, Hidalgo G, Simon T, Wei SE, Sheikh Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans Pattern Anal Mach Intell*. 2021;43(1):172-186. doi:10.1109/TPAMI.2019.2929257
32. Wang W, Xie E, Li X, et al. PVT v2: Improved baselines with Pyramid Vision Transformer. *Comput Vis Media*. 2022;8(3):415-424. doi:10.1007/s41095-022-0274-8
33. Mathis A, Mamidanna P, Cury KM, et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat Neurosci*. 2018;21(9):1281-1289. doi:10.1038/s41593-018-0209-y
34. Pereira TD, Tabris N, Matsliah A, et al. SLEAP: A deep learning system for multi-animal pose tracking. *Nat Methods*. 2022;19(4):486-495.
35. Toshev A, Szegedy C. DeepPose: Human Pose Estimation via Deep Neural Networks. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. ; 2014:1653-1660. doi:10.1109/CVPR.2014.214
36. Melanie Segado MJKK Claire Chambers, Nidhi Seethapathi, Rachit Saluja, Laura Prosser. Infant Movement Assessment: Update. Published online 2024. https://github.com/quietscientist/Infant_movement_assessment
37. Segado M. Update: Predicting clinical assessments of infants' risk of neuromotor disease from 2-dimensional videos. Published online September 2023. doi:10.17605/OSF.IO/SD6FA
38. Chambers C. Predicting clinical assessments of infants' risk of neuromotor disease from 2-dimensional videos.

Published online November 2018. doi:10.17605/OSF.IO/HV7TM
39. Ferrari F, Cioni G, Einspieler C, et al. Cramped Synchronized General Movements in Preterm Infants as an Early Marker for Cerebral Palsy. *Arch Pediatr Adolesc Med*. 2002;156(5):460-467. doi:10.1001/archpedi.156.5.460
40. Einspieler C, Yang H, Bartl-Pokorny KD, et al. Are sporadic fidgety movements as clinically relevant as is their absence? *Early Hum Dev*. 2015;91(4):247-252. doi:10.1016/j.earlhumdev.2015.02.003
41. Feurer M, Eggensperger K, Falkner S, Lindauer M, Hutter F. Auto-Sklearn 2.0: Hands-free AutoML via Meta-Learning. Published online October 4, 2022. doi:10.48550/arXiv.2007.04074
42. Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and Robust Automated Machine Learning. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, eds. *Advances in Neural Information Processing Systems*. Vol 28. Curran Associates, Inc.; 2015. https://proceedings.neurips.cc/paper_files/paper/2015/file/11d0e6287202fced83f79975ec59a3a6-Paper.pdf
43. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.

Figures

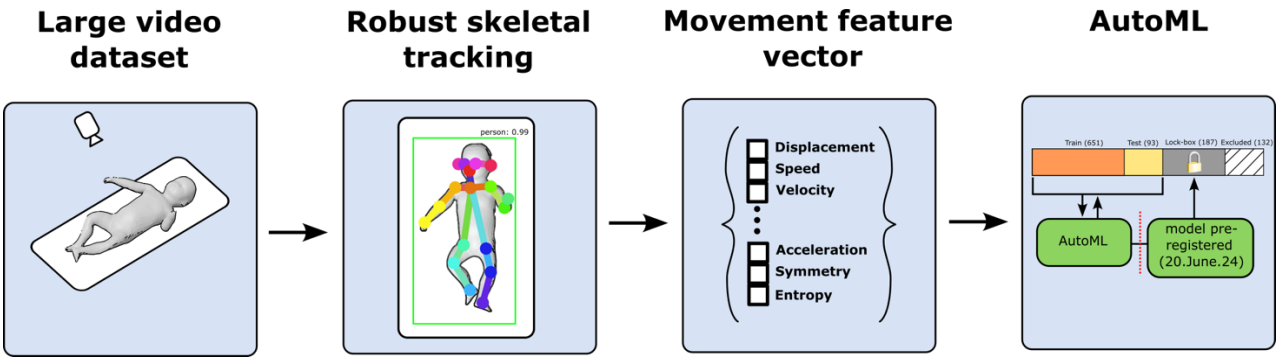

Figure 1 Process for rigorous evaluation of automated clinical score prediction. *Each step of the model development process was pre-registered. Lage video dataset had 1060 infants, 132 were excluded from the dataset for meeting one or more medical exclusion criteria prior to any analysis. Training(651)/Test(93) infant videos and "Lock-box" videos (187) were pre-registered prior to pose-tracking algorithm selection and movement feature computation. Features were pre-registered prior to model training. Model was pre-registered prior to testing on lock-box*

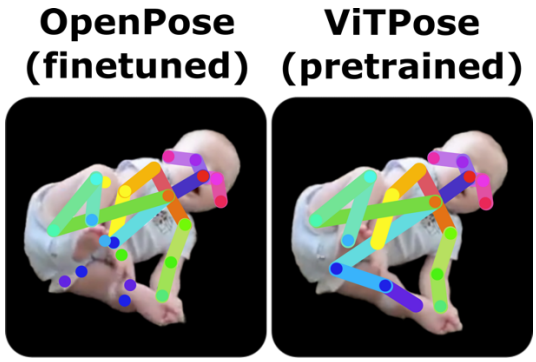

Figure 2 Improvements in skeletal tracking with pre-trained vision transformers. *Fine-tuned OpenPose algorithm (Left) still struggles with issues frequently encountered in infant videos such as occlusions and complex poses. Modern approaches, notably those leveraging pretrained vision transformers such as ViTPose-H (Right), are more robust. Occluded keypoints are not included in pose estimate using previous algorithms (Left). Transformer-based approach (Right) learns skeletal structure from adult human data and can infer occluded keypoints even on infants.*

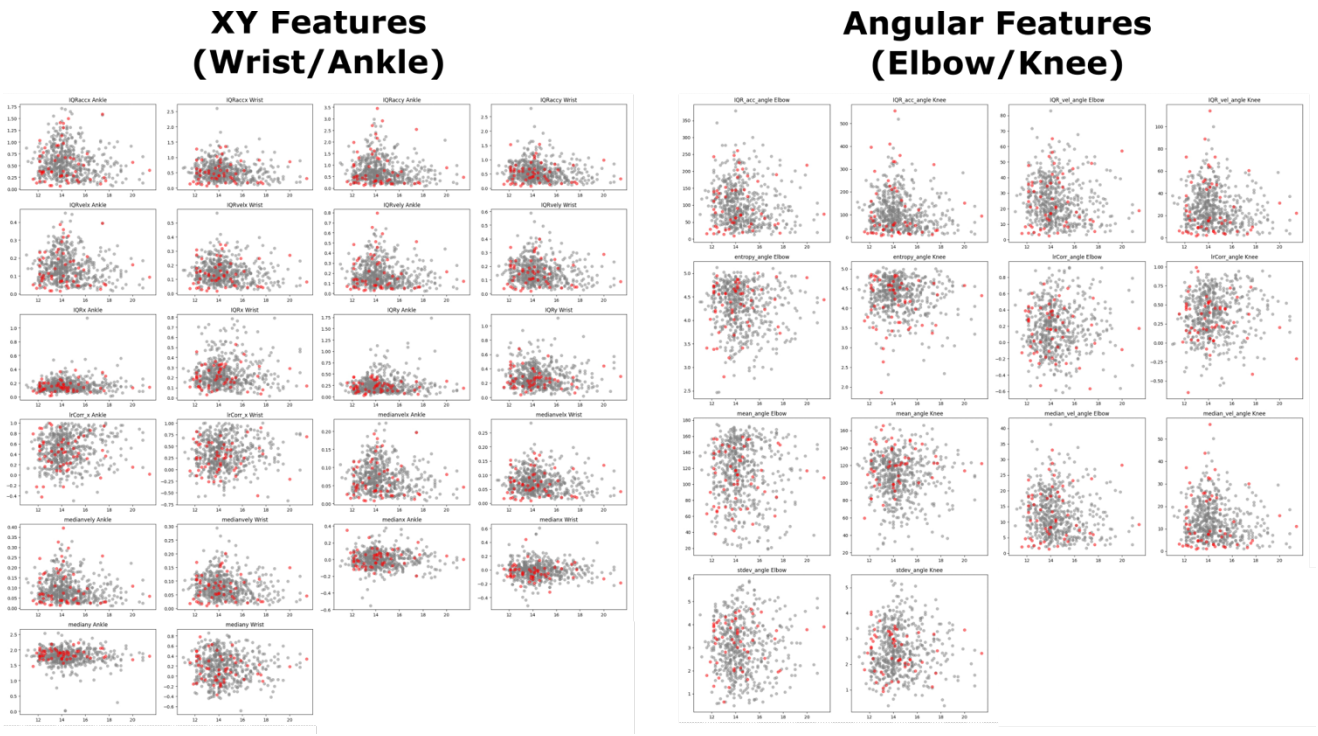

Figure 3 Individual features are highly correlated. *Clinician-selected features, including XY features of the wrists/ankles (Left) and angular features of the elbows/knees (Right), which are typically used for human assessment of risk are highly overlapping for Fidgety (Grey) and Absent Fidgety (Red) movement types, with no individual feature clearly predicting GMA score.*

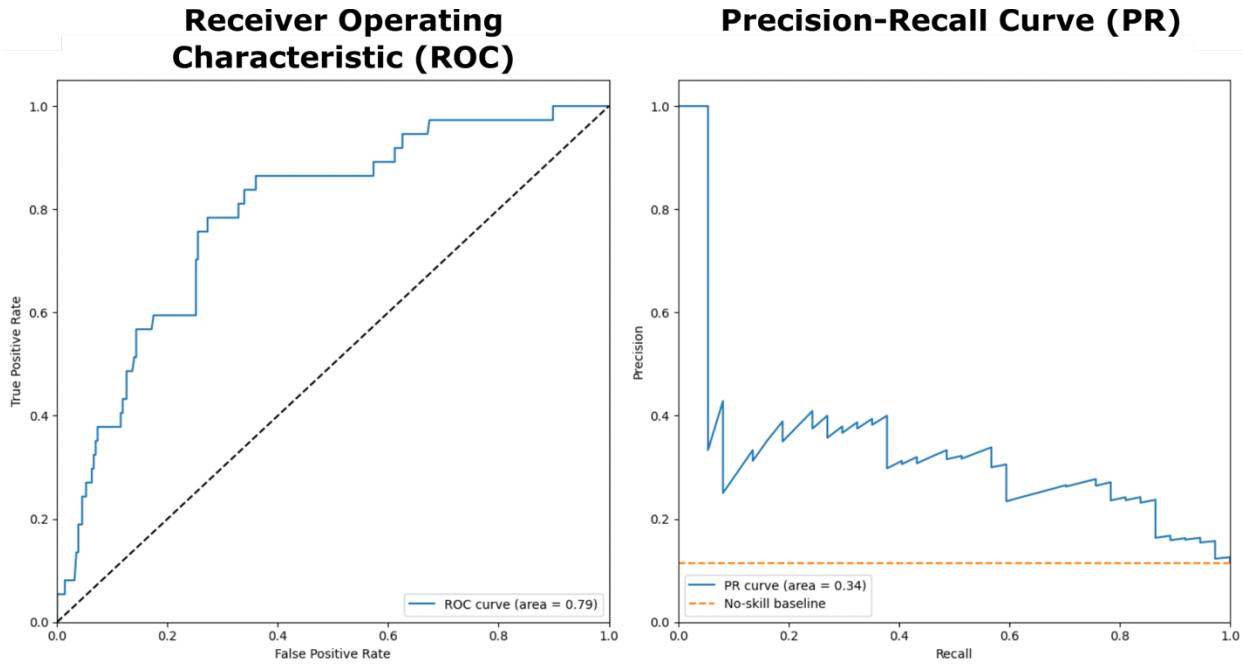

Figure 4 Model generalizes to lock-box set. *Classifier trained on clinician-selected features using vanilla auto-sklearn shows a high AUC-ROC of 0.79 (Left) and Precision-Recall of 0.34 (Right) on lock-box set of 187 infants, having 12% representation of absent fidgety movement type. True positive rate is equal to the Sensitivity of the classifier, False positive rate is equal to 1-Specificity.*