Title
Multivariate Regression Calibration for Inter-Rater Consistency in Hypertonia Evaluation

Journal
IEEE Sensors Letters, 10(7)

ISSN
2475-1472

Authors
Verrett, Mya
Liu, Jiaxi
Wieand, Alyssa
et al.

Publication Date
2026

DOI
10.1109/lsens.2026.3700488

This article has been accepted for publication in IEEE Sensors Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSENS.2026.3700488

IEEE Sensors Letters

IEEE Sensors Council

Article Subject (Wearable Sensor)

Multivariate Regression Calibration for Inter-Rater Consistency in Hypertonia Evaluation

Mya Verrett,[1] Jiaxi Liu,[2] Alyssa Wieand,[3] Andrew Skalsky,[3,4] and Tse Nga Ng[2]*

[1] *Department of Bioengineering, University of California San Diego, La Jolla, CA 92093, USA*
[2] *Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92093, USA*
[3] *Rady Children's Hospital, San Diego, CA, USA*
[3] *Department of Orthopedic Surgery, University of California San Diego, La Jolla, CA 92093, USA*
** Fellow, IEEE*

Abstract— Clinical evaluations of neuromuscular disorders, such as hypertonia, have been based on perception and can vary from physician to physician. In efforts to provide an objective assessment a multimodal sensing glove was developed, measuring the force and motion trajectories, recording the muscle's resistance to evaluate severity. However, due to grip and evaluation speed differences, variations in the output magnitude between raters require calibration for effective cross comparison. In this work, we demonstrate a multivariate regression model to address inter-rater variability across clinicians. The model uses both the acceleration and force recorded from the glove to calibrate each rater's assessment. Additionally, it effectively increased the kernel density overlapping area between clinicians after calibration, reducing variability and increasing force output consistency. Ultimately, this work will enhance hypertonia monitoring accessibility to support rehabilitation specialists.

Index Terms— Calibration Model, Hypertonia, Wearable Sensor

I. INTRODUCTION

Hypertonia is an abnormal increase in muscle tone resulting from upper motor neuron lesions, commonly arising from conditions such as cerebral palsy, traumatic brain or spinal injuries, and stroke [1]. Because hypertonia limits mobility and daily activities, it is treated with medication; however, finding the right dose is difficult because current practice relies on subjective evaluation[2] Standard methods of hypertonia assessment have been based on perception. A clinician will move the patient's limbs in flexion and extension maneuvers, ranking muscle resistance according to subjective scales such as the Modified Ashworth Scale or Tardieu Scale [3], [4], [5]. The imprecise assessments hinder patient monitoring and medication adjustments.

In efforts to provide an objective evaluation tool [6], [7], [8], [9], [10], [11], [12] we previously developed a portable, multimodal sensing glove to be worn by the evaluator during assessment (Fig. 1A) [13], [14]. The glove measures the applied force by pressure sensors (Tekscan) and captures motion trajectories by inertial motion unit (IMU, MotionNode), to record the patient's muscle resistance to movement (Fig. 1B). The multimodal glove effectively classified patient conditions under medication or placebo in a double-blind study [14], but a key limitation was the reliance on a single rater. Due to grip differences, variations in data magnitude between different raters point to the need for further calibration to enable cross comparisons of results obtained by different raters.

Here a multivariate regression method is demonstrated to

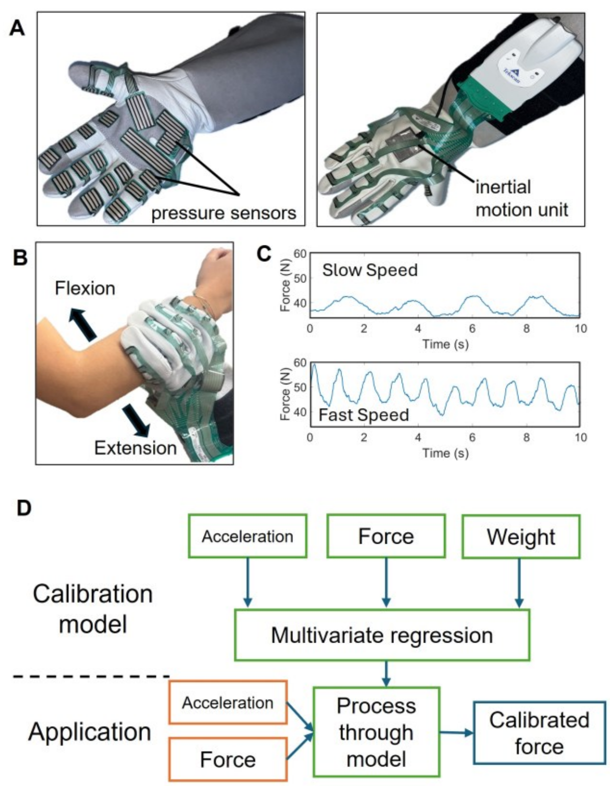

Fig. 1. (A) Photos of the sensor glove, with pressure sensors on the palm and inertial motion unit on the back. (B) A rater performing maneuver cycles to evaluate the patient's elbow tricep/bicep muscle resistance. (C) Muscle resistance under different movement speeds. (D) Flow chart of data processing through the calibration model.

address inter-rater variability between clinicians. This calibration process accounts for human variances such as fluctuations in movement speed and grip, which are naturally adjusted to avoid patient discomfort and ensure safety. As seen in Fig. 1C, the speed of the evaluation maneuvers affects the resistance in hypertonic muscles. Fast movements triggered higher reactive force than slower movements performed on the same limb [15]. Hence, this work includes motion acceleration into the multivariate regression model to standardize data across raters (Fig. 1D). In contrast to prior works in which calibration was carried out using only force sensors, this model weighs in acceleration variations, accounting for the motion component that may alter measurement outputs. Applying multivariate regression to establish user-specific calibration allows for different raters to consistently evaluate and monitor muscle tone.

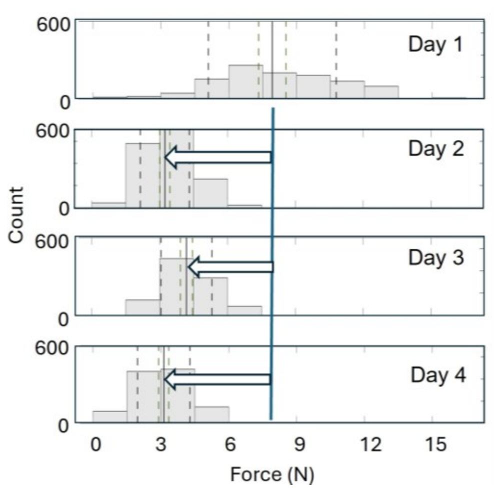

Fig. 2. Histograms of muscle resistive force over 4 days. The median force decreased from day 1, indicating improvement under medication.

The following analysis presents the adjustments to force data before and after passing through the regression model. The force output is the focus, because the histograms of force measurements have been shown to be a clear metric for quantifying changes in muscle tone [14]. For example in Fig. 2, shifts in force histograms correspond to changes in muscle tone in a patient undergoing baclofen treatment. The decreasing force required to maneuver the elbow joint indicates reduced muscle resistance and an improvement in hypertonia severity.

II. METHODS

Calibration Data Collection. Data used for the model training was collected from each rater. The sensor glove hardware descriptions are found in Ref. [13], [14], [16], [17], [18]. Wearing the sensor glove, each rater held a series of dumbbells ranging from 0.5 to 6 lbs while performing the flexion and extension maneuvers as they would on a patient's limb, for at least 10 maneuver cycles. The dumbbells served as known weights to standardize measurements across raters. The recorded force, acceleration, and corresponding dumbbell weights were used in the multivariate regression model fitting.

Development of Multivariate Regression Model. A multivariate linear regression model was created for each rater. The data processing was done in *MATLAB* using the *fitlm* function, forcing the y-intercept to 0, to reflect that there is no hypertonia being measured without movement of the limb. The model parameters were weighted to increase accuracy when compared to unweighted models. Adding weights reduced square error when compared to unweighted, addressing the potential heteroscedasticity of the force and acceleration input. To create the weights, the L2 norms of the calibration angular acceleration and force were taken and squared [19]. These

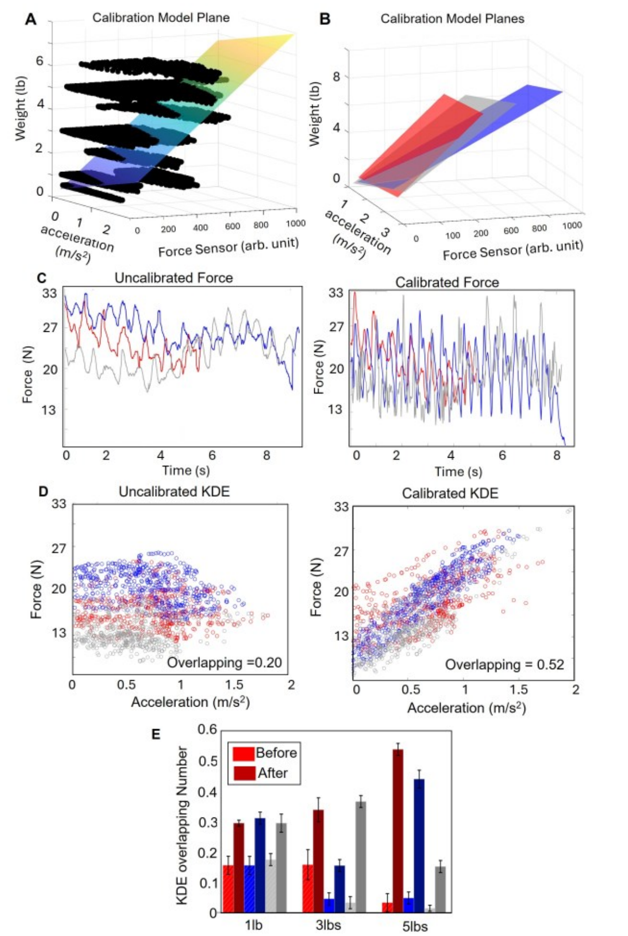

Fig. 3. (A) Calibration data collected by moving known weights from 0.5 to 6 lbs. Force sensor readout were in the 2^{11} bit scale of the readout electronics. (B) Multivariate regression planes for three raters. (C) Force versus time as raters carried out cyclic flexion and extension maneuvers holding a 4 lb dumbbell, before and after passing data through the regression model. Each rater represented by a different color. (D) Force versus acceleration graphs for kernel density estimation (KDE) to evaluate inter-rater correlation using 4 lb dumbbell. (E) Summary of KDE overlap metrics for three raters. Each color represents an individual rater, before (bright) and after (dark) calibration to improve inter-rater consistency.

weights were used to fit the model. The regression aims to minimize the loss function built into the *fitlim* MATLAB function:

$$Loss = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (1)$$

where the equation is the sum of squared errors between the fit and data. The regression models were trained with the calibration data and weights, producing a plane for each rater (Figs. 2A, 2B). The models for each rater with their corresponding parameter weights can be seen below where x is the angular acceleration and y is the force glove output:

$$f_1(x,y) = 0.79x + 0.00062y \qquad (2)$$

$$f_2(x,y) = 0.68x + 0.00096y \qquad (3)$$

$$f_3(x,y) = 0.23x + 0.0012y \qquad (4)$$

The root mean square error (RMSE) was used as confirmation that the models had the least amount of variance, with RMSE below 0.10 for all users. These user-specific models were applied to the angular acceleration and force readings, to standardize each rater's force data to the known dumbbell weights, based on the correlations captured by the multivariate regression models.

III. RESULTS AND DISCUSSION

A. Characterization of inter-rater consistency by kernel density estimation

Comparison of force data before and after (Fig. 3C) adjustment through the regression models shows increased overlap among raters. Inter-rater consistency was quantified using kernel density estimation (KDE). KDE is a nonparametric method that estimates the probability density without assumptions about the underlying data distribution, shape, or structure, relying instead on locations of the sample points over an area (Fig. 3D) [20], [21]. For the overlapping number calculation, the same grid was used for each rater during the kernel density estimation. The bandwidth was selected individually for each measurement using Silverman's rule of thumb [20], minimizing the approximate mean integrated squared error (MISE). Once the kernel density was calculated for each rater over the same grid, the minimum of the 3 densities was integrated over the total range where $\hat{f}_1(x,y)$ is the probability density of a rater respectively:

$$overlap\ area = \iint \min\left(\hat{f}_1(x,y), \hat{f}_2(x,y), \hat{f}_3(x,y),\right) dxdy \ (5)$$

The overlapping area between the two rater's kernel density estimations, normalized to 1, was used as a metric of variability between the two measurements. The greater the overlapping area, the greater probability density the three raters share, thus indicating improvements in consistency and reduced variability. To measure the confidence intervals for the overlapping area, the calculation was bootstrapped 100 times of the KDE estimates, randomizing the order of functions being computed. A bootstrapping limit of 100 times was chosen as it was the limit until no changes to the confidence intervals were evident. The error bars in Fig. 3E shows the lower and upper confidence interval. The KDE overlapping areas all increased after the calibration models were applied, across all dumbbell weights for three raters as seen in Fig 3E based on the calibration data. Next, we proceeded to analyze real patient data.

B. Inter-rater consistency of patient data after calibration

The sensor glove has been approved for use with patients under Institutional Review Board #180115. All measurements were taken with written informed consent of the patients and their legal guardians when applicable. In Fig. 4A, the force measurements on a patient's elbow joint were taken by three raters. In Fig. 4B, the corresponding force histograms before and after passing through the calibration multivariate regression models show the data converging to the same range. The weighted multivariate regression model proved to be the optimal method compared to unweighted versions of the models by reducing error and improving KDE overlap. The RMSE of unweighted models were 0.4 and above. Therefore, using a weighted model provided the least amount of square error, which was also reflected in KDE overlap values.

In addition, another patient was measured by two clinician raters, to evaluate muscle hypertonia of elbow and knee joints. The patient was undergoing baclofen treatment and stayed in the hospital for 4 days, and daily measurements were taken by the two raters to evaluate consistency between their results.

For elbow evaluations, the KDE overlapping areas between datasets increased after applying the calibration method in Fig.

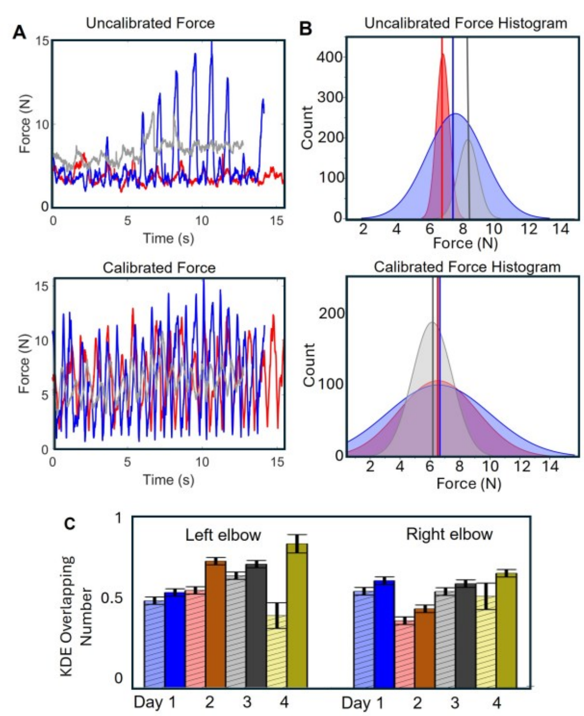

Fig. 4. (A) Exemplar force versus time data measured on actual patient by three raters. (B) Histogram of force pre- and post-calibration for the three raters. The colors in panel A and B indicate different raters. (C) KDE overlapping area pre- and post-calibration, for data measured by two raters over 4 days on the same patient undergoing baclofen treatment. Each color accounts for measurements in each day, before (hashed) and after (solid) calibration.

4C. The results indicate that the multivariate regression models reduced inter-rater variability, and these improvements were longitudinally maintained, allowing the same model to be used across the four measurement days. The calibration process on knee joint datasets was less successful, where the KDE overlapping areas showed both increase and decrease. Given the greater mass of muscles around the knees compared with the elbows, higher calibration weights than the 5 lb range used in the current experiment would be necessary to improve inter-rater consistency and overlap. Overall, it is evident that the calibration models reduce the inter-rater variability for lighter limbs such as the elbows.

C. Limitation of Study

This study has two limitations to be addressed in future research. The first is increasing the calibration data weight range to reduce variability in heavier limbs, such as the knees. Increasing the weight range will provide a larger coverage of the regression model and increase accuracy when more force is needed for the hypertonia assessment. This change may improve the overall accuracy of the calibration models as the limbs get heavier in weight. In this work, the method is successful for lighter limbs, such as elbows, however heavier limbs and knees are less consistent. The second task is to apply this calibration method to an increased number of raters, allowing for further validation and reproducibility of the calibration models.

IV. CONCLUSION

This work demonstrates a calibration method based on multivariate regression models to reduce inter-rater variability and enable consistent cross comparisons. Individual models created for each user adjusted the datasets to converge towards common reference standards. Evaluating the overlapping area from the kernel density estimation provides further confidence in reduced variability after calibration. Applied to a patient's elbow data, the model is applicable across four measurement days, which can save calibration time. This standardization for symptom assessments enhances monitoring accessibility. In the future, this work can be extended and built upon to support at-home evaluations by caregivers and allowing additional support to rehabilitation specialists.

ACKNOWLEDGMENT

This project was supported by the National Science Foundation CBET#2054517.

REFERENCES

[1] S. H. Evans, M. W. Cameron, and J. Michael Burton, "Hypertonia," *Curr. Probl. Pediatr. Adolesc. Health Care*, vol. 47, no. 7, pp. 161–166, Jul. 2017.

[2] A. L. Boster *et al.*, "Best Practices for Intrathecal Baclofen Therapy: Dosing and Long-Term Management," *Neuromodulation*, vol. 19, no. 6, pp. 623–631, Aug. 2016.

[3] A. A. Alhusaini, C. M. Dean, J. Crosbie, R. B. Shepherd, and J. Lewis, "Evaluation of spasticity in children with cerebral palsy using Ashworth and Tardieu Scales compared with laboratory measures.," *J. Child Neurol.*, vol. 25, no. 10, pp. 1242–1247, 2010.

[4] J. F. M. Fleuren *et al.*, "Stop using the Ashworth Scale for the assessment of spasticity," *J. Neurol. Neurosurg. Psychiatry*, vol. 81, p. 46, 2009.

[5] A. A. Puzi, S. N. Sidek, I. M. Khairuddin, H. Md Yusof, and H. Mat Rosly, "Inter-rater and intra-rater reliability of quantitative upper limb spasticity evaluation based on modified ashworth scale tool," in *2018 IEEE EMBS Conference on Biomedical Engineering and Sciences, IECBES 2018 - Proceedings*, 2019, pp. 126–130.

[6] L. Bar-On *et al.*, "A clinical measurement to quantify spasticity in children with cerebral palsy by integration of multidimensional signals," *Gait Posture*, vol. 38, no. 1, pp. 141–147, 2013.

[7] X. Li, H. Shin, S. Li, and P. Zhou, "Assessing muscle spasticity with Myotonometric and passive stretch measurements: validity of the Myotonometer.," *Sci. Rep.*, vol. 7, no. February, p. 44022, 2017.

[8] S. Y. Song, Y. Pei, S. R. Tippett, D. Lamichhane, C. M. Zallek, and E. T. Hsiao-Wecksler, "Validation of a wearable position, velocity, and resistance meter for assessing spasticity and rigidity," *Frontiers in Biomedical Devices - 2018 Design of Medical Devices Conference*, pp. 6–9, Apr. 2018.

[9] M. Amit, L. Chukoskie, A. J. Skalsky, H. Garudadri, and T. N. Ng, "Flexible Pressure Sensors for Objective Assessment of Motor Disorders," *Adv. Funct. Mater.*, p. 1905241, 2019.

[10] M. Amit, R. K. Mishra, Q. Hoang, A. M. Galan, J. Wang, and T. N. Ng, "Point-of-use robotic sensors for simultaneous pressure detection and chemical analysis," *Mater. Horiz.*, vol. 6, pp. 604–611, 2019.

[11] J. Liu *et al.*, "Stretchable Textile Bands for Ambulatory Electrocardiogram and Oximetry," *IEEE Journal on Flexible Electronics*, vol. 1, no. 4, pp. 214–222, Nov. 2022.

[12] L. H. Sloot *et al.*, "Motorized versus manual instrumented spasticity assessment in children with cerebral palsy," *Dev. Med. Child Neurol.*, vol. 59, no. 2, pp. 145–151, 2017.

[13] M. Amit, C. Yalcin, J. Liu, A. J. Skalsky, H. Garudadri, and T. N. Ng, "Multimodal assessment of spasticity using a point-of-care instrumented glove to separate neural and biomechanical contributions," *iScience*, vol. 25, no. 11, p. 105286, Nov. 2022.

[14] J. Liu *et al.*, "Longitudinal monitoring of hypertonia through a multimodal sensing glove," *Biosens. Bioelectron.*, vol. 267, p. 116829, Jan. 2025.

[15] J. Ganguly, D. Kulshreshtha, M. Almotiri, and M. Jog, "Muscle tone physiology and abnormalities," *Toxins (Basel).*, vol. 13, no. 4, p. 282, 2021.

[16] C. Yalcin *et al.*, "Artifacts Mitigation in Sensors for Spasticity Assessment," *Advanced Intelligent Systems*, p. 2000106, 2020.

[17] P. Jonnalagedda *et al.*, "An instrumented glove for improving spasticity assessment," *IEEE Healthcare Innovation Point-of-Care Technologies Conference*, pp. 167–170, 2016.

[18] J. Liu *et al.*, "Instrumented Glove Assessment of Asymmetric Spasticity Progression," in *IFETC 2023 - 5th IEEE International Flexible Electronics Technology Conference, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023.

[19] J. Chukwura and I. Chinenye Jecinta, "A Review of Techniques for Regularization," *International Journal of Research in Engineering and Science (IJRES) ISSN*, vol. 11, pp. 360–367, 2023, [Online]. Available: www.ijres.org

[20] B. W. Silverman, "Density Estimation for Statistics and Data Analysis," in *Monographs on Statistics and Applied Probability*, London: Chapman and Hall, 1986.

[21] C. Chen, J. Krijthe, and R. Karlsson, "An Empirical Study of Adaptive Kernel Density Estimation in Detecting Distributional Overlap," Delft University of Technology, The Netherlands, 2023. [Online]. Available: http://repository.tudelft.nl/.