Evaluating inter-rater reliability of physical behavioral annotations with clinical data to advance PathML

Casey Nash, Storey Balko, Jennifer Perez Aguilar, JT Bernal, Jonah David, Edgardo Hernandez, & Sarah Kozey Keadle
Department of Kinesiology and Public Health, California Polytechnic State University, San Luis Obispo

Introduction

PathML is software developed by Sentimetrix and Cal Poly SLO that uses computer vision to automate the labeling of video data for assessing physical activity, mobility, balance, and muscular endurance (Keadle et al., 2024). Our study evaluates the agreement between student video annotations and clinical scores for two functional tests to ensure the accuracy of data used to train PathML.

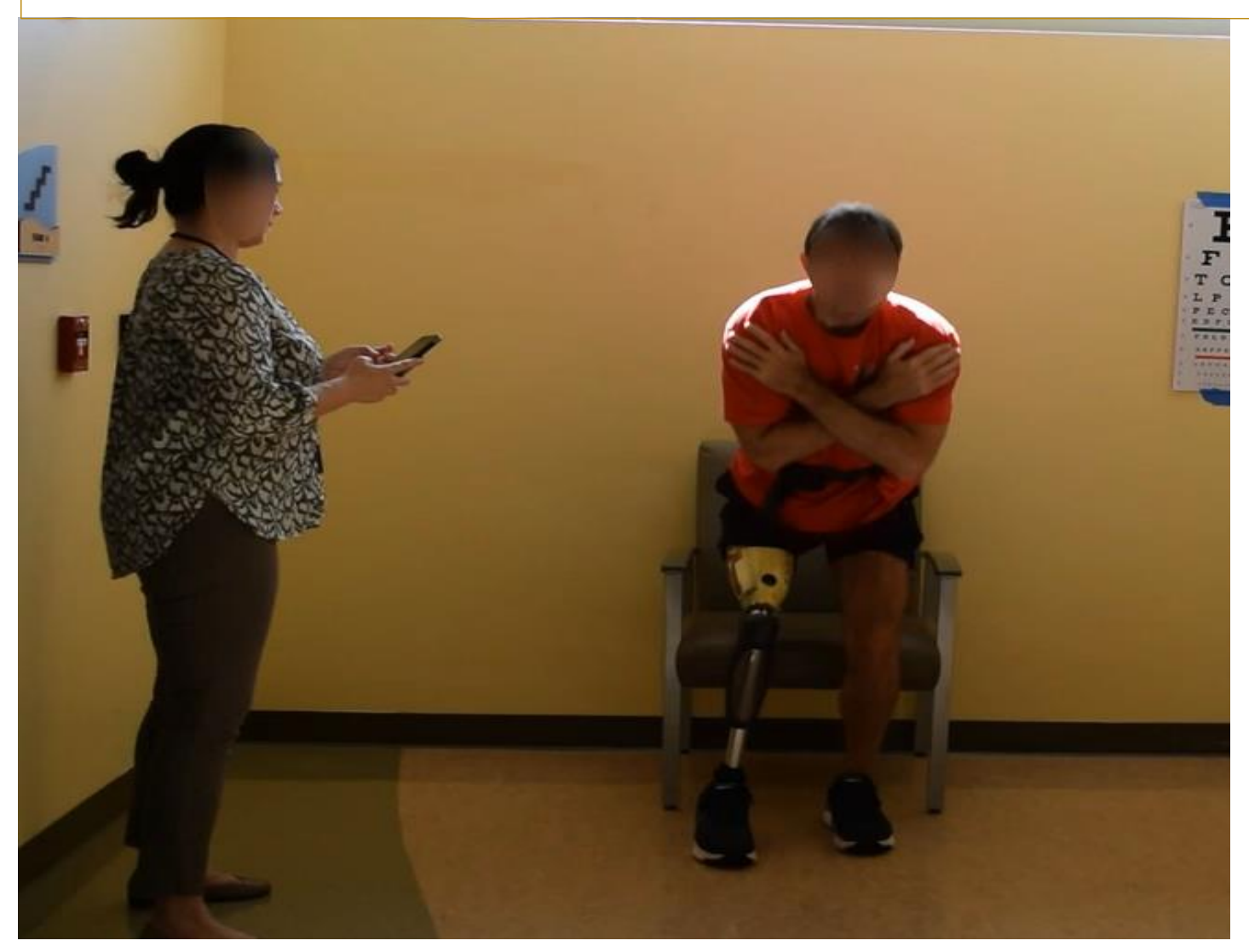


Figure 1: Proctor (left) and participant (right) during the Sit-to-Stand (S2S) functional test. S2S tests leg strength and endurance by measuring the number of times a subject comes to a full standing position within 30 seconds along with support level and failed attempts.

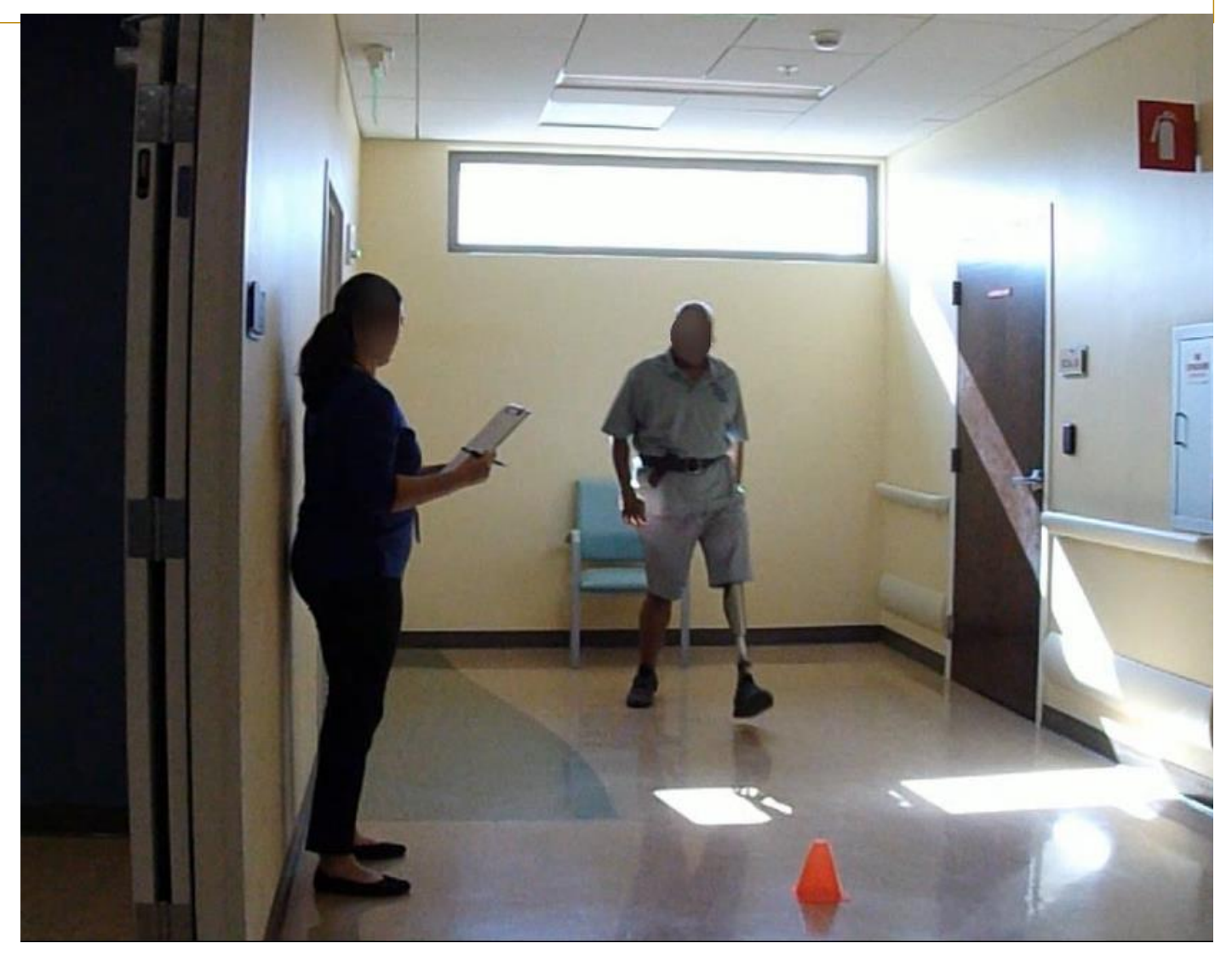


Figure 2: Proctor (left) and participant (right) during the Timed Up and Go (TUG) functional test. TUG assesses mobility, balance, and walking ability by determining the amount of time it takes the subject to stand up, walk three meters, and then return to a seated position.

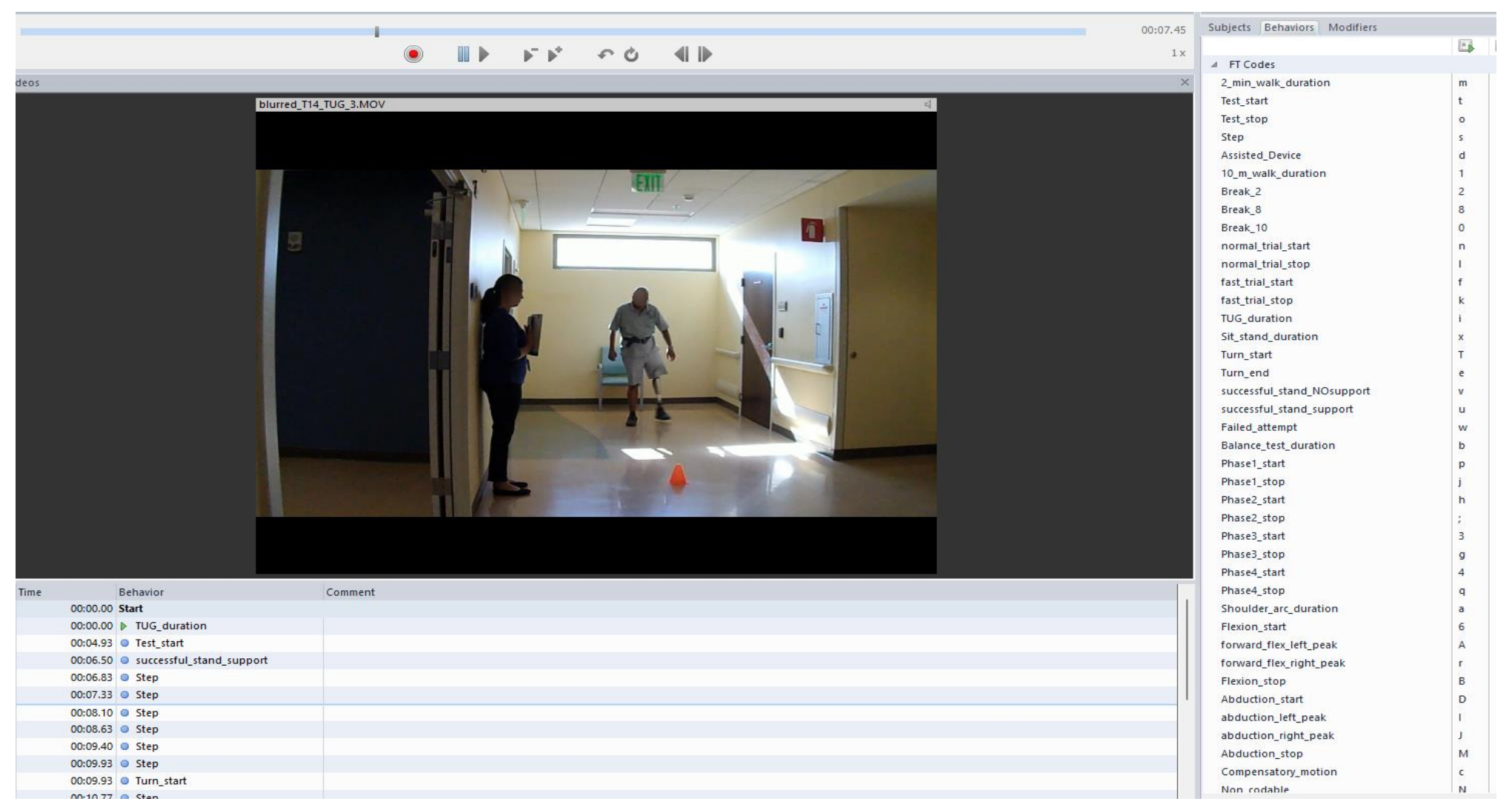
Methods

IN THE LAB:

- 58 participants were recruited at the Tampa VA their scores on Sit-to-Stand (S2S) and Timed Up and Go (TUG) were recorded by trained clinicians at the VA.
- Student researchers at Cal Poly annotated video data using Noldus Observer XT Software (n=46), 20% of videos were double-coded to measure inter-rater reliability (IRR).

ANALYSIS:

- SPSS was used to compare Cal Poly labeled video scores to the clinical values recorded at the VA.
- For continuous outcomes, mean differences were tested using t-tests and intra-class correlation (ICC).
- For categorical outcomes (at-risk/not-at-risk), agreement was measured using Cohen's Kappa statistics.



FT Codes	
2_min_walk_duration	m
Test_start	t
Test_stop	o
Step	s
Assisted_Device	d
10_m_walk_duration	1
Break_2	2
Break_8	8
Break_10	0
normal_trial_start	n
normal_trial_stop	l
fast_trial_start	f
fast_trial_stop	k
TUG_duration	i
Sit_stand_duration	x
Turn_start	T
Turn_end	e
successful_stand_NOsupport	v
successful_stand_support	u
Failed_attempt	w
Balance_test_duration	b
Phase1_start	p
Phase1_stop	j
Phase2_start	h
Phase2_stop	;
Phase3_start	3
Phase3_stop	g
Phase4_start	4
Phase4_stop	q
Shoulder_arc_duration	a
Flexion_start	6
forward_flex_left_peak	A
forward_flex_right_peak	r
Flexion_stop	B
Abduction_start	D
abduction_left_peak	I
abduction_right_peak	J
Abduction_stop	M
Compensatory_motion	c
Non_codable	N

Time	Behavior	Comment
00:00.00	Start	
00:00.00	TUG_duration	
00:04.93	Test_start	
00:06.50	successful_stand_support	
00:06.83	Step	
00:07.33	Step	
00:08.10	Step	
00:08.63	Step	
00:09.40	Step	
00:09.93	Step	
00:09.93	Turn_start	
00:10.77	Step	

Figure 3: Screenshot of the Noldus Observer XT Software being used to label a TUG functional test. Student researchers coded key aspects of physical behaviors including posture changes, steps, start, and stop (see coding scheme located on right side of the image for specific codes).

Results

- There were no significant differences between coder and clinical scores **(TUG: p = 0.101; S2S: p = 0.486)**. Inter-class correlation scores were excellent **(TUG average = 0.993; S2S = 0.993)**, indicating strong agreement between scorers.
- Cohen's Kappa tests were run on categorical scoring (at-risk/not-at-risk) demonstrating high inter-rater reliability **(TUG average = 0.954; S2S = 0.777)**.
- At-risk was defined as ≥12sec for the TUG test. For S2S, at-risk was classified if assistance was required for successful stand (e.g., hands on knees/chair or caregiver aid) or if participants had one or more failed attempts.

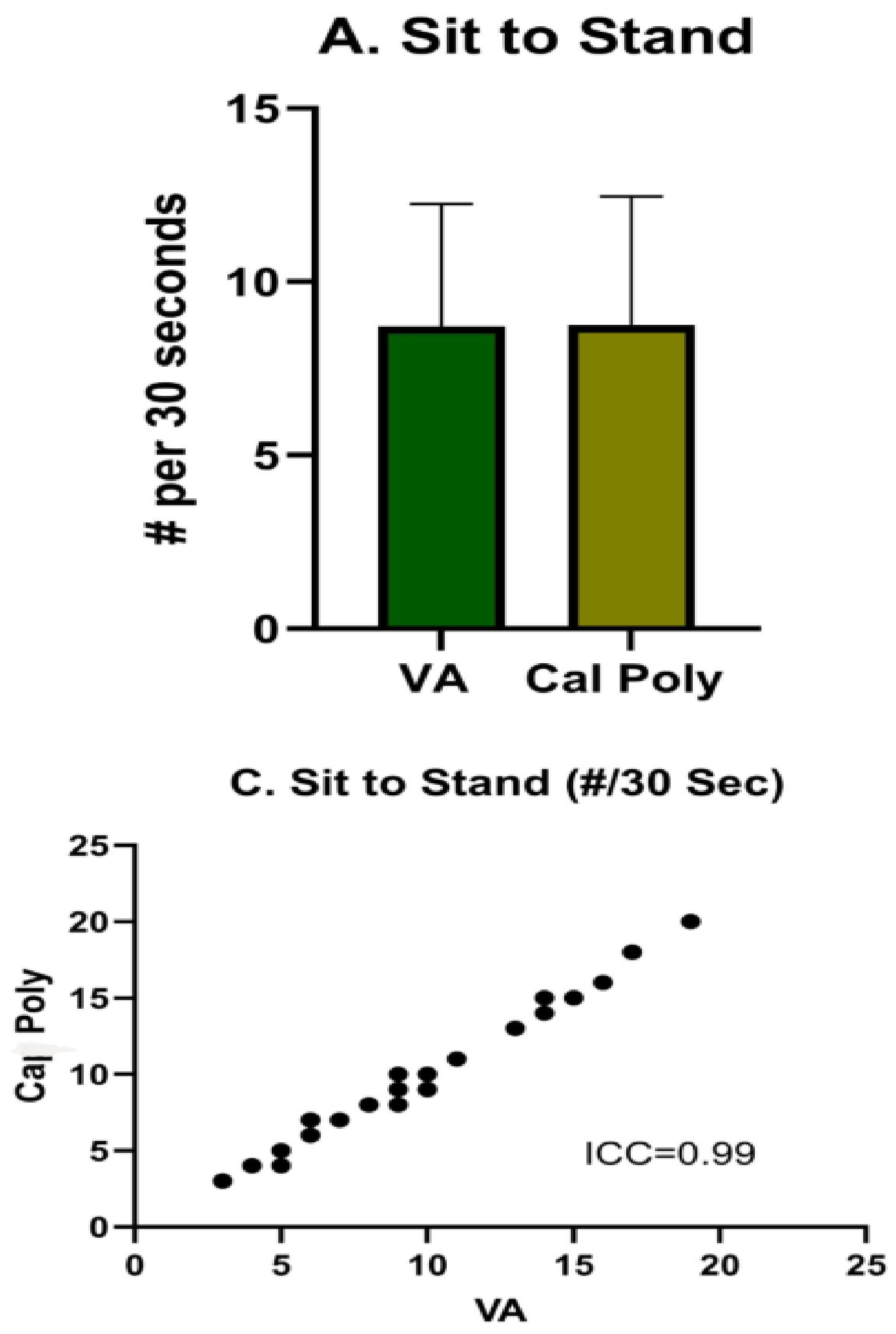


Figure 4a: Comparison of mean coder and clinical S2S scores. *Figure 4c:* ICC between Tampa VA and Cal Poly number of stands in 30 second S2S.

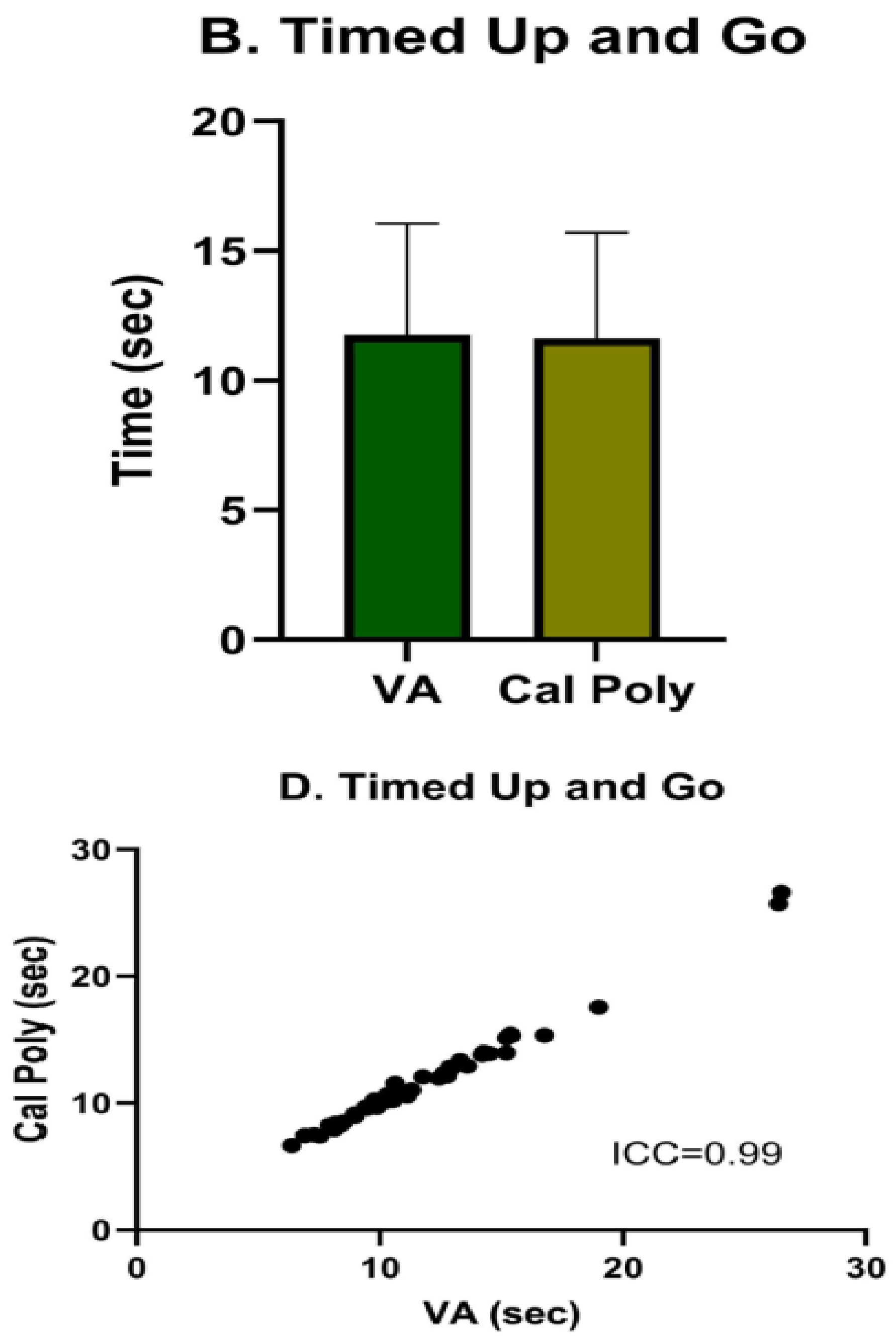


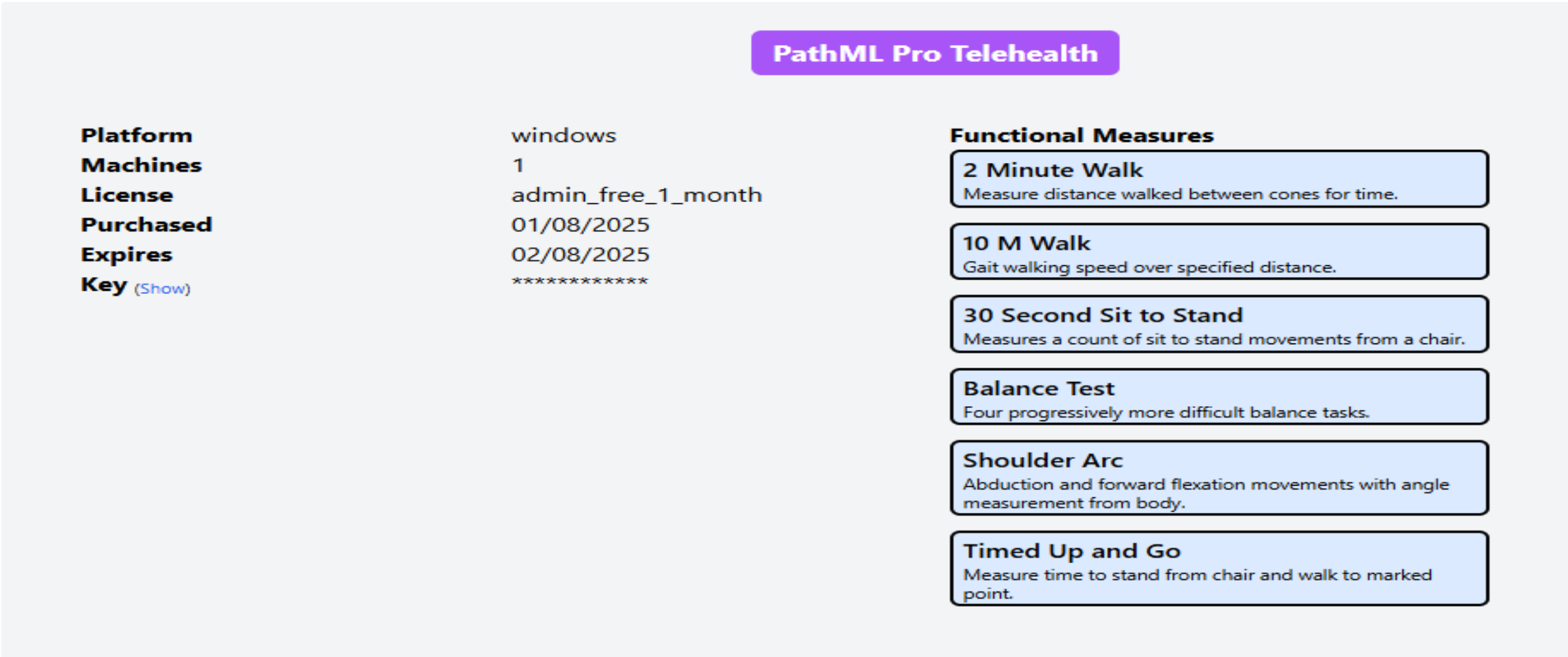
Figure 4b: Comparison of mean coder and clinical TUG scores. *Figure 4d:* ICC between Tampa VA and Cal Poly average TUG durations.

Acknowledgements

Thank you to all who have contributed: Cal Poly Center for Health Research, Dr. Sarah Kozey Keadle, Tampa VA, William and Linda Frost Foundation, NIH/NCI Small Business Innovation Research (75N91023C00031) , & Sentimetrix Inc.

Discussion

- **High Annotation Agreement:** Student coders showed excellent agreement with clinician scores on TUG and S2S tasks using the PathML framework.
- **Reliable Scoring:** Strong inter-rater reliability confirmed that students could match clinician-level accuracy for pass/fail scoring.
- **Validated Training Method:** Supports using student annotations to train PathML for automated functional test labeling from video.
- **Comparable Software Accuracy:** PathML is expected to perform similarly to trained coders, aligning with findings by Keadle et al. (2024) on the reliability of digital assessment tools.
- **Clinical Utility:** PathML may reduce clinician workload and increase access to standardized assessments, especially in telehealth settings.
- **Improved Access and Outcomes:** Could benefit patients in remote or underserved areas by enabling consistent rehab monitoring.
- **Public Health Application:** PathML could be adapted for large-scale physical activity surveillance across populations.
- **Environmental Labeling Potential:** Following Carlson et al. (2018), PathML may also be used to label built environment features alongside physical behaviors.
- **Support for Computer Vision:** Reinforces growing evidence that computer vision methods can reliably assess functional movement (Keadle et al., 2024).
- **Ongoing Testing:** PathML is undergoing user testing for usability and integration; Sentimetrix Inc. continues to refine the tool.



PathML Pro Telehealth

Platform	windows
Machines	1
License	admin_free_1_month
Purchased	01/08/2025
Expires	02/08/2025
Key (Show)	************

Functional Measures

- **2 Minute Walk** Measure distance walked between cones for time.
- **10 M Walk** Gait walking speed over specified distance.
- **30 Second Sit to Stand** Measures a count of sit to stand movements from a chair.
- **Balance Test** Four progressively more difficult balance tasks.
- **Shoulder Arc** Abduction and forward flexation movements with angle measurement from body.
- **Timed Up and Go** Measure time to stand from chair and walk to marked point.

Figure 5: Screenshot of current version of PathML Pro Telehealth taken during one of the first rounds of user testing to assess the accuracy, usability, and consistency of the automatic labeling software.

References

Carlson, Jordan A et al. "Unique Views on Obesity-Related Behaviors and Environments: Research Using Still and Video Images." Journal for the measurement of physical behaviour vol. 1,3 (2018): 143-154. doi:10.1123/jmpb.2018-0021

Keadle, Sarah Kozey et al. "A Framework to Evaluate Devices That Assess Physical Behavior." Exercise and sport sciences reviews vol. 47,4 (2019): 206-214. doi:10.1249/JES.0000000000000206

Keadle, Sarah K et al. "Using Computer Vision to Annotate Video-Recoded Direct Observation of Physical Behavior." Sensors (Basel, Switzerland) vol. 24,7 2359. 8 Apr. 2024, doi:10.3390/s24072359