10-1-2019

# Machine Learning Groups Patients by Early Functional Improvement Likelihood Based on Wearable Sensor Instrumented Preoperative Timed-Up-and-Go Tests

Riley A. Bloomfield
*Western University*

Harley A. Williams
*Robarts Research Institute*

Jordan S. Broberg
*Robarts Research Institute*

Brent A. Lanting
*London Health Sciences Centre*

Kenneth A. McIsaac
*Western University*

*See next page for additional authors*

Citation of this paper:
Bloomfield, Riley A.; Williams, Harley A.; Broberg, Jordan S.; Lanting, Brent A.; McIsaac, Kenneth A.; and Teeter, Matthew G., "Machine Learning Groups Patients by Early Functional Improvement Likelihood Based on Wearable Sensor Instrumented Preoperative Timed-Up-and-Go Tests" (2019). *Bone and Joint Institute*. 797.
https://ir.lib.uwo.ca/boneandjointpub/797

## Authors

Riley A. Bloomfield, Harley A. Williams, Jordan S. Broberg, Brent A. Lanting, Kenneth A. McIsaac, and Matthew G. Teeter

The Journal of Arthroplasty 34 (2019) 2267–2271

Contents lists available at ScienceDirect

The Journal of Arthroplasty

journal homepage: www.arthroplastyjournal.org

Artificial Intelligence and Machine Learning

# Machine Learning Groups Patients by Early Functional Improvement Likelihood Based on Wearable Sensor Instrumented Preoperative Timed-Up-and-Go Tests

Riley A. Bloomfield, BESc [a, b, *], Harley A. Williams, MSc [b, c], Jordan S. Broberg, BMSc [b, c], Brent A. Lanting, MD, MSc, FRCSC [d], Kenneth A. McIsaac, PhD [a], Matthew G. Teeter, PhD [b, c, d, e]

[a] *Department of Electrical & Computer Engineering, Western University, London, Ontario, Canada*
[b] *Imaging Research Laboratories, Robarts Research Institute, London, Ontario, Canada*
[c] *Department of Medical Biophysics, Schulich School of Medicine & Dentistry, Western University, London, Ontario, Canada*
[d] *Division of Orthopaedic Surgery, Department of Surgery, Schulich School of Medicine & Dentistry, Western University and London Health Sciences Centre, London, Ontario, Canada*
[e] *Surgical Innovation Program, Lawson Health Research Institute, London, Ontario, Canada*

ARTICLE INFO

*Article history:*
Received 21 February 2019
Received in revised form
13 May 2019
Accepted 29 May 2019
Available online 5 June 2019

*Keywords:*
total knee arthroplasty
wearable sensors
machine learning
functional testing
early recovery

## ABSTRACT

*Background:* Wearable sensors permit efficient data collection and unobtrusive systems can be used for instrumenting knee patients for objective assessment. Machine learning can be leveraged to parse the abundant information these systems provide and segment patients into relevant groups without specifying group membership criteria. The objective of this study is to examine functional parameters influencing favorable recovery outcomes by separating patients into functional groups and tracking them through clinical follow-ups.

*Methods:* Patients undergoing primary unilateral total knee arthroplasty ($n = 68$) completed instrumented timed-up-and-go tests preoperatively and at their 2-, 6-, and 12-week follow-up appointments. A custom wearable system extracted 55 metrics for analysis and a K-means algorithm separated patients into functionally distinguished groups based on the derived features. These groups were analyzed to determine which metrics differentiated most and how each cluster improved during early recovery.

*Results:* Patients separated into 2 clusters ($n = 46$ and $n = 22$) with significantly different test completion times (12.6 s vs 21.6 s, $P < .001$). Tracking the recovery of both groups to their 12-week follow-ups revealed 64% of one group improved their function while 63% of the other maintained preoperative function. The higher improvement group shortened their test times by 4.94 s, ($P = .005$) showing faster recovery while the other group did not improve above a minimally important clinical difference (0.87 s, $P = .07$). Features with the largest effect size between groups were distinguished as important functional parameters.

*Conclusion:* This work supports using wearable sensors to instrument functional tests during clinical visits and using machine learning to parse complex patterns to reveal clinically relevant parameters.

While total knee arthroplasty (TKA) is a successful end-stage treatment for osteoarthritis that commonly improves joint function and reduces pain, further investigation for treatment refinement is needed to decrease the recovery time and economic burden of this procedure. Preoperative function can be a predictor of postoperative functional improvement [1,2]. Rapid functional recovery in the first 3 months following TKA can additionally affect hospital funding, predict long-term recovery, and be used to adjust patient expectations realistically [3]. Objective measurement is

---

One or more of the authors of this paper have disclosed potential or pertinent conflicts of interest, which may include receipt of payment, either direct or indirect, institutional support, or association with an entity in the biomedical field which may be perceived to have potential conflict of interest with this work. For full disclosure statements refer to https://doi.org/10.1016/j.arth.2019.05.061.

[*] Reprint requests: Riley A. Bloomfield, BESc, Department of Electrical & Computer Engineering, Western University, 1151 Richmond Street, London, ON N6A 3K7, Canada.

https://doi.org/10.1016/j.arth.2019.05.061

important for functional evaluation, but subjective patient-reported outcome measures (PROMs) remain a more commonly used tool. These self-reported surveys are convenient to deploy but completion of multiple measures by patients requires significant time and effort. Patients often do not complete the provided questionnaires entirely, rendering the collected data incomplete or inconsistent, and these measures have been shown to differ from objectively assessed functional performance [4–6].

The emergence of wearable sensor systems has permitted more efficient data collection providing objective measures that can be compared across patient populations or multiple clinical time-points of the same patient for recovery analysis. Functional tests such as the timed-up-and-go (TUG) or 3-minute-walk test can be completed by patients quickly in clinic and benefit from portable instrumentation for objective measurement [7]. The TUG test is of interest because it combines knee-relevant activities of standing from a chair, walking to a 3-m goal, turning about the goal, and returning to the seated start position, encompassing a variety of weight-bearing joint stressors. This test has been previously used in assessing TKA patients by timing the completion of the test, and differences of more than 2.27 s represent a clinically meaningful change in function [8]. Further analysis and joint measurement can be leveraged with the use of wearable sensors. As patient functional data are abundantly available with numerous measured parameters, patterns throughout individual patient recovery or across patient populations become difficult to identify by an observer. Unsupervised machine learning offers the ability to identify complex multivariate patterns in data and can group similar patient populations together without predefining group labels or membership criteria. Similarities and differences between patient groups throughout their recovery can reveal functional parameters prevalent in positive or negative outcomes.

The purpose of this study is to measure functional performance of TKA patients before and during their short-term recovery period and apply unsupervised machine learning to separate patients into different functional groups based on derived performance to highlight parameters influencing short-term functional improvement.

## Methods

IRB approval was obtained for studies investigating patients scheduled to undergo primary unilateral TKA as a treatment for osteoarthritis. Patients were prescreened to ensure they did not have inflammatory arthritis or alcoholism before being recruited at their preadmission appointment and each patient provided informed written consent before participation. Individuals who had language and/or cognitive barriers, neuromuscular disorders, and operative or nonoperative leg amputations or who required a wheelchair for mobility were excluded. Demographic information was obtained from an orthopedic database, and Short Form 12, Western Ontario and McMaster Universities Osteoarthritis Index, Knee Society Score (KSS), and University of California Los Angeles (UCLA) Activity Score questionnaires were obtained at the preoperative, 6, and 12-week clinical appointments. Surgeons also completed the Office Knee Evaluation at each of these timepoints.

At preoperative, 2, 6, and 12-week appointments, patients were instrumented with a validated wearable sensor system consisting of an iPod Touch and 4 lightweight inertial sensors (one mounted above and below each knee) while completing 3 trials of the TUG test [9]. A custom software application was used to extract the recorded functional tests from the wearable sensor system and segment the test into 5 phases: sit to stand, walking to the goal, turning about the goal, walking back to the start, and sitting in the starting chair. Observer bias was minimized using autonomous test start, finish, and segment separation and a step detection algorithm was used to extract steps for operative and nonoperative legs. A total of 55 spatiotemporal and functional metrics were extracted from raw test data including lower leg motions, upper leg motions, joint movement in 3 dimensions, flexion/extension velocities and accelerations, step range of motion (ROM), number of steps, turn directions, and segment completion times.

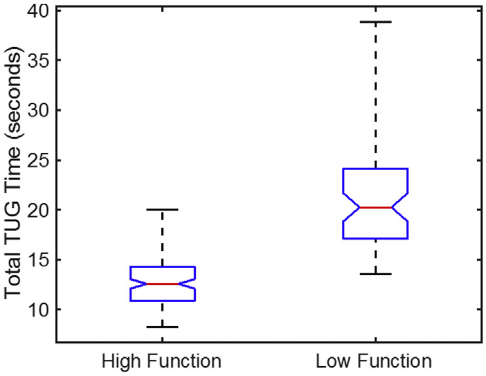

**Fig. 1.** Box and whisker plot of total TUG completion time of both groups preoperatively showing a large overlap in times between groups. TUG, timed-up-and-go.

Extracted functional and spatiotemporal metrics from preoperative tests were combined into feature rows and stacked. Each feature column was standardized to have zero mean and unit standard deviation, so all features varied on a similar scale. All preoperative samples were fed into an unsupervised K-means clustering algorithm which separated patients into 2 groups. This algorithm is effective for finding groupings in unlabeled data that may be otherwise difficult to identify. It iteratively separates data points into $k$ number of groups by comparing the similarity of data samples across all features and rearranges the group membership for maximum separation. Binary groups were not decided initially but this ideal number of groups was validated by observing a minimal Calinski-Harabasz index with $k = 2$, tested on a range of 2-6 [10].

A 1-way multivariate analysis of variance (ANOVA) was used to find PROM differences between the 2 separated clusters and Anderson-Darlington tests were used to confirm the normality of all preoperative PROM distributions ($P < .05$). One-way ANOVA was also used to find significant differences in the derived spatiotemporal and functional metrics between groups. Additionally, Cohen's D was computed for each feature to determine which of the derived metrics had the largest effect size and contributed the most to differentiating the 2 clusters.

The preoperative functional clusters were tracked forward to their 12-week appointments and a mixed-effects ANOVA with repeated measures test was performed on TUG total completion times of both groups at the preoperative, 2-, 6-, and 12-week timepoints. Derived metrics that remained significantly different between groups at both preoperative and 12-week timepoints with the largest effect sizes were observed as the most important persistent functional differences and were also analyzed using a mixed-effects ANOVA with repeated measures to compare improvement between preoperative and 12-week trials.

## Results

A total of 94 patients were eligible to participate in this study but 18 were excluded because they missed/rescheduled their preoperative or 12-week appointments. An additional 8 patients were excluded because they underwent alternate interventions from the

**Table 1**
Mean ± SD Values of Patient Characteristics and Questionnaire Outcomes of High and Low Function Preoperative Clusters.

| Mean ± SD | High Function | Low Function | $P$ Value |
|---|---|---|---|
| Age (y) | **65.6 ± 9.1** | **71.5 ± 10.3** | **.018** |
| BMI (kg/m$^2$) | 32.8 ± 5.8 | 34.9 ± 6.3 | .180 |
| UCLA Activity Score[b] | **4.9 ± 2.1** | **3.8 ± 1.6** | **.001** |
| SF-12 Mental | 53.3 ± 11.0 | 51.7 ± 9.8 | .493 |
| SF-12 Physical | 33.1 ± 8.3 | 29.7 ± 8.9 | .079 |
| WOMAC Pain | 46.6 ± 17.3 | 45.5 ± 17.9 | .682 |
| WOMAC Stiffness | 41.2 ± 19.0 | 46.2 ± 24.3 | .123 |
| WOMAC Function | 49.3 ± 18.7 | 49.8 ± 11.0 | .838 |
| WOMAC Total | 46.1 ± 15.8 | 47.2 ± 13.8 | .620 |
| KSS Symptoms | 16.0 ± 4.1 | 15.3 ± 3.8 | .246 |
| KSS Satisfaction | 13.7 ± 7.6 | 14.5 ± 7.1 | .506 |
| KSS Expectations | **14.0 ± 1.3** | **13.0 ± 2.0** | **<.001** |
| KSS Functional Activities | **37.9 ± 16.7** | **32.3 ± 12.5** | **.029** |
| KSS Knee Objective Indicators[a] | 34.4 ± 18.5 | 33.8 ± 17.8 | .813 |
| Knee Evaluation Function[a] | **50.7 ± 13.6** | **41.8 ± 23.1** | **.007** |
| Knee Evaluation Total Knee[a] | 43.0 ± 15.3 | 39.3 ± 16.1 | .201 |
| Knee Evaluation Total[a] | **93.7 ± 25.4** | **78.1 ± 33.0** | **.003** |

Bolded items indicates the significant differences between groups ($P < .05$).
"a" Indicates surgeon-reported measures and "b" indicates distinctions above minimal clinically important differences.
BMI, body mass index; KSS, Knee Society Score; SD, standard deviation; SF-12, Short Form 12; UCLA, University of California Los Angeles; WOMAC, Western Ontario and McMaster Universities Osteoarthritis Index.

study protocol. The remaining 68 patients (male:female [M:F] = 34:34) were included in the study aged 67.5 ± 9.8 years with a body mass index of 33.5 ± 6.0 kg/m$^2$. Clustering feature rows for all preoperative TUG trials (3 per patient) partitioned patients into 2 groups with sizes 46 and 22 (M:F = 20:26 and M:F = 14:8). Seven patients had at least 1 TUG trial sorted into each group but were separated into the group with majority membership. Comparison of spatiotemporal metrics revealed the mean total TUG completion time (the only traditionally recorded TUG metric) of the larger group's trials was significantly faster than the smaller group (12.6 s vs 21.6 s, $P < .001$) but there was a large overlap in trial times between groups (Fig. 1). Despite the large group containing slow trials and the small group containing fast trials, these 2 groups were labeled "high function" and "low function," respectively, due to their mean total TUG time. The high function group was younger, scored higher in UCLA, KSS Functional Activities, Office Knee Evaluation Function (surgeon reported), and Office Knee Evaluation Total (surgeon reported) questionnaires and had slightly higher (KSS) expectations (Table 1). Most of the novel derived measures (48/55) were also significantly different between groups ($P < .05$) including all spatiotemporal metrics (Table 2) which is expected because the K-means algorithm separates clusters to achieve a maximum separation across all features equally. The 10 derived metrics with the largest effect size can be seen in Table 3.

The mean total TUG time for sorted patients was not only different between groups preoperatively, but also at the 6-week (5.4 s, $P = .01$) and 12-week (4.2 s, $P = .02$) follow-ups. Total time was not significantly different at 2 weeks ($P = .55$) where the mean patient time increased for both high and low function groups (by 9.2 s and 7.6 s, respectively). Mean total time of the high function group improved 0.87 s ($P = .07$) from preoperative to 12 weeks (Table 4) while the low function group improved 4.94 s ($P = .005$). Of the patients in the high function group, 26% (12) had meaningfully improved function (completion time decreased by >2.27 s from preoperation to 12 weeks postoperation), 63% (29) maintained function, and 11% (5) had worsened function (completion time increased by >2.27 s). Of the patients in the low function group, 64% (14) had improved function, 27% (6) maintained function, and 9% (2) had worsened function. A comparison of individual TUG segment times for all trials of both groups at each timepoint can be seen in Figure 2. The most distinguishable features between the 2 groups at 12 weeks and their effect sizes can be seen in Table 5. The top 3 most distinguishable features at the 12-week timepoint (top of Table 5) have been plotted alongside the same dimensions of the 2 cluster centroids found using all 55 metrics to visualize their strong influence on the preoperative group separation (Fig. 3). Six test samples from the low function group with exceptionally low operative step motion are shown in Figure 3 due to their operative leg function being too impaired to correctly detect any distinct steps. A failure to detect these steps has resulted in zero motion during steps but it is expected that this remains a meaningful result. Early improvement of key metrics for both groups can be seen in Table 6.

**Table 2**
Mean ± SD of Spatiotemporal Metric Differences Between Preoperative Patient Clusters.

| Mean ± SD | High Function | Low Function | $P$ Value |
|---|---|---|---|
| Total time | 12.7 ± 2.4 | 21.6 ± 5.7 | <.0001 |
| Sit to stand | 1.1 ± 0.5 | 2.2 ± 1.3 | <.0001 |
| Walking to goal | 3.8 ± 0.8 | 6.6 ± 2.0 | <.0001 |
| Turning at goal | 0.6 ± 0.3 | 1.1 ± 0.5 | <.0001 |
| Walking to chair | 4.8 ± 1.0 | 8.1 ± 2.0 | <.0001 |
| Stand to sit | 1.8 ± 0.5 | 2.8 ± 1.5 | <.0001 |

SD, standard deviation.

**Table 3**
Mean ± Standard Deviation of Top Distinguishable Functional and Spatiotemporal Metrics Between Groups at Preoperation and Their Effect Size (D).

| Metric Description | High Function | Low Function | D |
|---|---|---|---|
| Time taken walking back to test start after turning (s) | 4.8 ± 1.0 | 8.1 ± 2.0 | 1.604 |
| Mean additive operative lower leg motion during steps (°) | 124.5 ± 14.4 | 92.1 ± 12.0 | 1.601 |
| Total test time (s) | 12.6 ± 2.4 | 21.6 ± 5.7 | 1.591 |
| Mean additive nonoperative lower leg motion during steps (°) | 128.7 ± 15.4 | 96.0 ± 11.2 | 1.571 |
| Time taken walking from initial stand to begin of turn (s) | 3.8 ± 0.8 | 6.6 ± 2.0 | 1.510 |
| Mean additive operative upper leg motion during steps (°) | 90.0 ± 14.7 | 63.4 ± 9.8 | 1.470 |
| Mean additive nonoperative upper leg motion during steps (°) | 91.1 ± 14.9 | 65.1 ± 7.6 | 1.468 |
| Mean nonoperative step peak flexion velocity (°/s) | 289.9 ± 57.2 | 196.2 ± 22.0 | 1.428 |
| Mean nonoperative step peak extension velocity (°/s) | 282.3 ± 59.5 | 197.8 ± 37.6 | 1.271 |
| Mean nonoperative step peak flexion acceleration (°/s$^2$) | 5412.1 ± 1692.7 | 3129.9 ± 895.6 | 1.247 |

All features are significantly different between groups ($P < .001$).

## Discussion

The unsupervised clustering performed in this study successfully separated wearable sensor instrumented performance tests

**Table 4**
Mean Functional Group TUG Total Completion Time Changes and CI Between Preoperative Performance and Each Recovery Point (Negative Values Indicate a Total Time Improvement).

| Timepoint Comparison | High Function | 95% CI | $P$ Value | Low Function | 95% CI | $P$ Value |
|---|---|---|---|---|---|---|
| Preoperative to 2 wk | **+9.2 s** | **+13.3 to +5.1** | **<.001** | +7.6 s | +21.4 to +6.1 | .400 |
| Preoperative to 6 wk | +0.4 s | +1.8 to −1.0 | .472 | −2.9 s | +2.1 to −7.9 | .373 |
| Preoperative to 12 wk | −0.9 s | +0.1 to −1.8 | .076 | **−4.9 s** | **− 1.3 to −8.5** | **.005** |

Bolded items indicates the significant differences between groups ($P < .05$).
CI, confidence interval; TUG, timed-up-and-go.

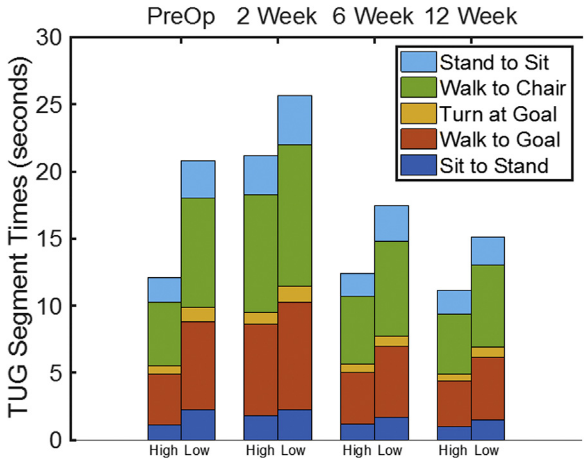

**Fig. 2.** Mean TUG segment times for each functional group over their early recovery period.

based on derived functional metrics into clinically relevant groups. Literature has previously linked preoperative functional performance to postoperative functional improvement; however, this study has highlighted functional parameters that differentiate between patients who are likely to gain function [1,2]. Other work predicting gait recovery following TKA has employed a different method to first label samples as responders or nonresponders of parameter improvement and train a classifier to make this prediction. The authors suggest the unsupervised grouping and cluster analysis performed in the present work is more effective for distinguishing important functional parameters and real-world generalization because group membership is not defined, but organically created based on the unbiased structure of the data [11].

The objective measures extracted from the short and easy-to-implement TUG test have been able to distinguish function preoperatively with many significant differences between groups whereas there are few differences in PROMs and overlapping TUG completion times. It can be seen in Table 1 that despite several PROMs being significantly different between groups, only the UCLA Activity Score varied more than its minimal clinically important difference of 0.92 [12,13]. This presents evidence that there are functional performance differences that cannot be distinguished using subjective self-reported measures when comparing patients with different expectations.

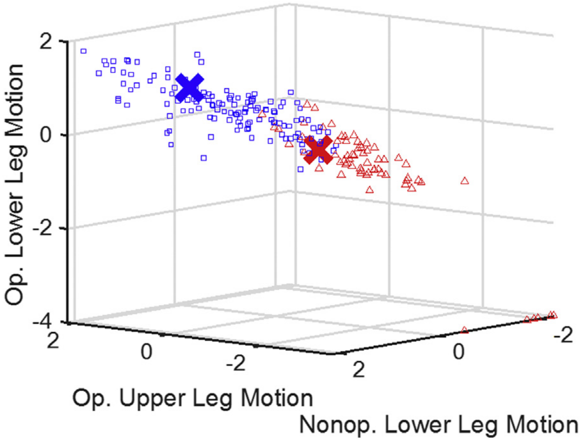

**Fig. 3.** Z-scores of the top 3 most distinguished metrics persisting to the 12-wk follow-up appointments for each trial and their influence on initial preoperative clustering. Solid crosses (x) represent group centroids found using all 55 derived metrics. Blue squares and red triangles indicate high and low function group trials, respectively. Nonop., nonoperative; Op., operative.

**Table 5**
Mean ± Standard Deviation of Top Distinguishable Functional and Spatiotemporal Metrics Between Groups at 12 wk Postoperation and Their Effect Size (D).

| Metric Description | High Function | Low Function | D |
|---|---|---|---|
| Mean additive operative upper leg motion during steps (°) | **96.1 ± 15.0** | **76.8 ± 18.9** | **1.043** |
| Mean additive nonoperative lower leg motion during steps (°) | **135.0 ± 15.4** | **114.8 ± 22.2** | **1.004** |
| Mean additive operative lower leg motion during steps (°) | **131.2 ± 16.3** | **109.5 ± 27.1** | **0.961** |
| Time taken walking from initial stand to begin of turn (s) | **3.4 ± 0.7** | **4.7 ± 2.0** | **0.933** |
| Mean additive nonoperative upper leg motion during steps (°) | **97.5 ± 15.1** | **81.6 ± 16.8** | **0.919** |
| Total test time (s) | **11.8 ± 2.7** | **16.0 ± 6.5** | **0.893** |
| Time taken walking back to test start after turning (s) | **4.5 ± 1.2** | **6.1 ± 2.5** | **0.868** |
| Time taken to stand from the seated position (s) | 1.7 ± 0.7 | 2.1 ± 0.9 | 0.800 |
| Mean additive operative flexion during steps (°) | 87.0 ± 16.1 | 73.9 ± 15.2 | 0.774 |
| Mean operative flexion range during steps (°) | 42.4 ± 7.9 | 36.5 ± 7.5 | 0.714 |

All features were significantly different between groups ($P < .001$).
Bold indicates features that were also distinguished in the preoperative timepoint analysis.

Tracking the 2 functionally separated groups through to 3 months revealed that one group was much more likely to improve relative to the other. Test results for new patients can be compared to the 2 groups clustered in this study to determine the most similar expected recovery path, and this information can be used to provide realistic recovery expectations for patients. The analysis performed has relied on total TUG time as a relatable overall performance metric and tool for labeling groups but it should be highlighted that other derived metrics had larger effect sizes between groups and were determined to be more relevant parameters influencing the resulting group memberships (Table 3 and 5). Identified features distinguishing the functional groups at 12 weeks included both operative and nonoperative leg metrics which suggests test completion and movement compensation strategies were captured. The low function group had larger improvements in all relevant additive motion metrics (Table 6) but these values never exceeded those of the high function group. Increased ROM has been previously linked to better outcomes but the results of this study indicate that the total additive amount of motion expended during activities by both operative and nonoperative legs is also functionally important [14,15]. Both groups improved these metrics significantly during recovery which supports that the metrics are influencing shorter TUG times, which did decrease for both groups but not by a meaningful difference for the high function group.

Although it may be thought that higher functioning preoperative patients have less possible function to regain and will likely show less functional improvement, there is still a benefit of including the instrumented TUG test at preoperative visits because this functional differentiation was possible with the derived metrics and the current work has shown that PROMs alone cannot reliably report function to this granularity. Additionally, it is important to note in the preoperative groups that 7 patients in the "low" function group had mean total TUG times faster than the worst "high" function time. Similarly, there were 7 patients in the "high" function group with times slower than the best "low" function group time, indicating that total TUG time alone would not be enough to sort the groups in this way. Of the 7 patients in the "low" function group with favorable times that have been labeled likely to improve, 5 of them have improved their total time above the TUG minimal clinically important difference of 2.27 s [8]. Of the

**Table 6**
Operative and Nonoperative Improvement of Top Persisting Motion Metrics Between Preoperative and 12-wk Follow-Ups for Each Group.

| Metric Description | High Function | 95% CI | *P* Values | Low Function | 95% CI | *P* Values |
|---|---|---|---|---|---|---|
| Mean additive operative upper leg motion during steps | 6.1 | 10.2-2.0 | .002 | 13.3 | 19.5-7.1 | <.0001 |
| Mean additive nonoperative lower leg motion during steps | 6.3 | 10.7-1.9 | .003 | 18.8 | 25.3-12.4 | <.0001 |
| Mean additive operative lower leg motion during steps | 6.6 | 11.4-1.8 | .004 | 17.4 | 24.6-10.2 | <.0001 |
| Mean additive nonoperative upper leg motion during steps | 6.4 | 10.3-2.4 | .001 | 16.5 | 22.3-10.7 | <.0001 |

CI, confidence interval.

7 patients in the "high" function group with less favorable times labeled not likely to improve, only 3 of them have improved above the same threshold. These patients who were sorted into the low function group who had more favorable times have still shown improvement despite a smaller possible improvement range while patients sorted into the high function group who had less favorable times have not improved despite a larger possible improvement range. These special cases suggest the cluster separation better indicates likelihood of improvement than overall functional level.

Some limitations were noted with the study performed. During data collection, 7 patients repeated one of their trials sufficiently different from the others that the samples were split between both groups. Although a single patient cannot belong to multiple functional groups, the separation of tests may be a valid result because repetitions of the same test can be slightly different due to fatigue, test familiarity, confusion, or perhaps a stumble or stiffness. During clustering, it was decided to keep each trial as a separate sample to best separate functional groups but when future data are compared to find the most relatable path, it may be more practical to take an average across multiple trials for more generalized predictability.

The follow-up time analyzed was limited to the early recovery period of 12 weeks. Despite recovery following TKA usually lasting 1 to 2 years, the authors believe valuable information can be obtained during the early recovery phase, and this time period can be important for health economics [16]. As alternative joint replacement payment models such as Medicare's Bundled Payments for Care Improvement Program are introduced, early outcome prediction becomes valuable for allocating care costs. Under model 2 of the Bundled Payments for Care Improvement Program, hospitals will be reimbursed for costs saved in the first 90 days following surgery and patients with early improvement will likely require less frequent early care [17]. Fast functional improvement and early ambulation reduces hospital length of stay which also reduces the likelihood of costly readmissions due to infection [18–20]. A longer follow-up will be necessary to determine how function changes in each group until patients are fully healed.

The current work has shown that preoperative functional assessment can benefit from the use of wearable sensor instrumentation and machine learning techniques can identify multivariate patterns that would be otherwise difficult to see by an observer. Groups of patients following similar short-term recovery paths have been identified and future test data can be compared to similar path prediction to better influence patient expectations. There was little evidence that the PROMs collected in this study related to the results found using the derived functional and spatiotemporal metrics. Obtaining PROMs proved much more time-consuming for patients and the process was more cumbersome when it became time to store and digitize the measures, further motivating the use of an automated sensor system that can record performance tests in only a few minutes and provide instantaneous analysis.

## Acknowledgments

This study was funded by The Arthritis Society and an Ontario Early Researcher Award. MGT is supported by a CIHR New Investigator Award.

## References

[1] Fortin PR, Clarke AE, Joseph L, Liang MH, Tanzer M, Ferland D, et al. Outcomes of total hip and knee replacement: preoperative functional status predicts outcomes at six months after surgery. Arthritis Rheum 1999;42:1722–8. https://doi.org/10.1002/1529-0131.

[2] Kennedy DM, Hanna SE, Stratford PW, Wessel J, Gollish JD. Preoperative function and gender predict pattern of functional recovery after hip and knee arthroplasty. J Arthroplasty 2006;21:559–66. https://doi.org/10.1016/j.arth.2005.07.010.

[3] Lavernia C, D'Apuzzo M, Rossi MD, Lee D. Is postoperative function after hip or knee arthroplasty influenced by preoperative functional levels? J Arthroplasty 2009;24:1033–43. https://doi.org/10.1016/j.arth.2008.09.010.

[4] Luna IE, Kehlet H, Peterson B, Wede HR, Hoevsgaard SJ, Aasvang EK. Early patient-reported outcomes versus objective function after total hip and knee arthroplasty. Bone Joint J 2017;99-B:1167–75. https://doi.org/10.1302/0301-620X.99B9.BJJ-2016-1343.R1.

[5] Tully MA, Panter J, Ogilvie D. Individual characteristics associated with mismatches between self-reported and accelerometer-measured physical activity. PLoS One 2014;9:1–9. https://doi.org/10.1371/journal.pone.0099636.

[6] Mizner RL, Petterson SC, Clements KE, Zeni JA, Irrgang JJ, Snyder-Mackler L. Measuring functional improvement after total knee arthroplasty requires both performance-based and patient-report assessments: a longitudinal analysis of outcomes. J Arthroplasty 2011;26:728–37. https://doi.org/10.1016/j.arth.2010.06.004.

[7] Sprint G, Cook DJ, Weeks DL. Toward automating clinical assessments: a survey of the timed up and go. IEEE Rev Biomed Eng 2015;8:64–77. https://doi.org/10.1109/RBME.2015.2390646.

[8] Yuksel E, Kalkan S, Cekmece S, Unver B, Karatosun V. Assessing minimal detectable changes and test-retest reliability of the timed up and go test and the 2-minute walk test in patients with total knee arthroplasty. J Arthroplasty 2017;32:426–30. https://doi.org/10.1016/j.arth.2016.07.031.

[9] Bloomfield RA, Fennema MC, McIsaac KA, Teeter MG. Proposal and validation of a knee measurement system for patients with osteoarthritis. IEEE Trans Biomed Eng 2019;66:319–26. https://doi.org/10.1109/TBME.2018.2837620.

[10] Calinski T, Harabasz J. A dendrite method for cluster analysis. Commun Stat 1974;3:1–27. https://doi.org/10.1080/03610927408827101.

[11] Kluge F, Hannink J, Pasluosta C, Klucken J, Gassner H, Gelse K, et al. Pre-operative sensor-based gait parameters predict functional outcome after total knee arthroplasty. Gait Posture 2018;66:194–200. https://doi.org/10.1016/j.gaitpost.2018.08.026.

[12] SooHoo NF, Li Z, Chenok KE, Bozic KJ. Responsiveness of patient reported outcome measures in total joint arthroplasty patients. J Arthroplasty 2015;30:176–91. https://doi.org/10.1016/j.arth.2014.09.026.

[13] Lee WC, Kwan YH, Chong HC, Yeo SJ. The minimal clinically important difference for Knee Society Clinical Rating System after total knee arthroplasty for primary osteoarthritis. Knee Surg Sports Traumatol Arthrosc 2017;25:3354–9. https://doi.org/10.1007/s00167-016-4208-9.

[14] Ha C-W, Park Y-B, Song Y-S, Kim J-H, Park Y-G. Increased range of motion is important for functional outcome and satisfaction after total knee arthroplasty in Asian patients. J Arthroplasty 2016;31:1199–203. https://doi.org/10.1016/j.arth.2015.12.018.

[15] Miner AL, Lingard EA, Wright EA, Sledge CB, Katz JN. Knee range of motion after total knee arthroplasty: how important is this as an outcome measure? J Arthroplasty 2003;18:286–94. https://doi.org/10.1054/arth.2003.50046.

[16] Snyder DJ, Bienstock D, Keswani A, Tishelman J, Ahn A, Molloy IB, et al. Preoperative patient reported outcomes and clinical characteristics as predictors of 90-day cost/utilization and complications. J Arthroplasty 2019;34:839–45. https://doi.org/10.1016/j.arth.2019.01.059.

[17] Siddiqi A, White PB, Mistry JB, Gwam CU, Nace J, Mont MA, et al. Effect of bundled payments and health care reform as alternative payment models in total joint arthroplasty: a clinical review. J Arthroplasty 2017;32:2590–7. https://doi.org/10.1016/j.arth.2017.03.027.

[18] Urish KL, Qin Y, Li BY, Borza T, Sessine M, Kirk P, et al. Predictors and cost of readmission in total knee arthroplasty. J Arthroplasty 2018;33:2759–63. https://doi.org/10.1016/j.arth.2018.04.008.

[19] Saucedo JM, Marecek GS, Wanke TR, Lee J, Stulberg SD, Puri L. Understanding readmission after primary total hip and knee arthroplasty: who's at risk? J Arthroplasty 2014;29:256–60. https://doi.org/10.1016/j.arth.2013.06.003.

[20] Pua Y-H, Ong P-H. Association of early ambulation with length of stay and costs in total knee arthroplasty: retrospective cohort study. Am J Phys Med Rehabil 2014;93:962–70.