**Neighbors-based prediction for physical function after total knee arthroplasty**

Chong Kim1

Kathryn L. Colborn, PhD1

Stef van Buuren, PhD2

Timothy Loar, DPT3

Jennifer E. Stevens-Lapsley, PT, PhD3,4

\*Andrew Kittelson, PT, PhD3

1University of Colorado, Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO niversity of Colorado, Physical Therapy Program, Department of Physical Medicine and Rehabilitation, Aurora, CO

2Department of Statistics, University of Utrecht, Utrecht, The Netherlands

3University of Colorado, School of Medicine, Physical Therapy Program, Department of Physical Medicine and Rehabilitation, Aurora, CO

4Veterans Affairs Medical Center, Geriatric Research, Education and Clinical Center, Denver, CO

**\*** Corresponding Author:

Andrew Kittelson, PT, DPT, PhD

School of Medicine, Physical Therapy Program

Department of Physical Medicine and Rehabilitation

Mail Stop C244

13121 East 17th Avenue, Room 3116

Aurora, CO 80045

Andrew.kittelson@ucdenver.edu

**Abstract**

The purpose of this study was to develop and test personalized predictions for functional recovery after Total Knee Arthroplasty (TKA) surgery, using a novel neighbors-based prediction approach. We used data from 397 patients with TKA to develop the prediction methodology and then tested the predictions in a temporally distinct sample of 202 patients. The Timed Up and Go (TUG) Test was used to assess physical function. Neighbors-based predictions were generated by estimating an index patient’s prognosis from the observed recovery data of previous similar patients (a.k.a., the index patient’s “matches”). Matches were determined by an adaptation of predictive mean matching. Matching characteristics included preoperative TUG time, age, sex and Body Mass Index. The optimal number of matches was determined to be m=35, based on low bias (-0.005 standard deviations), accurate coverage (50% of the realized observations within the 50% prediction interval), and acceptable precision (the average width of the 50% prediction interval was 2.33 seconds). Predictions were well-calibrated in out-of-sample testing. These predictions have the potential to inform care decisions both prior to and following TKA surgery.

**Introduction**

Total Knee Arthroplasty (TKA) is the most commonly performed inpatient elective surgery in the United States, at approximately 700,000 procedures per year.1 Although TKA is regarded as effective, the clinical course is highly variable.2 Depending on the patient, recovery of physical function can occur rapidly (within weeks), or it can be an arduous, months-long process.3,4 Moreover, the surgical population is remarkably heterogeneous—some patients engage in sporting activities (e.g., tennis, skiing),5 while others struggle to ambulate at walking speeds sufficient for independence in the community. There is no such thing as the “average” patient.6

To achieve the ideals of person-centered care,7,8 and also because TKA is an elective procedure, clinical decisions should be anchored to the individual patient.9 Yet the determination of an individual patient’s functional prognosis is challenging. Prediction models have been developed in TKA, but these models have several limitations: 1) they perform poorly in out-of-sample testing,10 2) they are based on mathematical functions that are unlikely to be flexible enough to realistically portray the clinical course across all patients,11 or 3) they predict functional outcomes at discrete postoperative time points, which may not overlap with the time frame during which patients are undergoing postoperative care and clinical monitoring.12

Neighbors-based prediction—where an index patient’s prognosis is estimated from the observed recovery data of previous similar patients (a.k.a. the index patient’s matches)—may address these limitations.13 The parameters of the prediction (and the shape of the prognostic trajectory) are allowed to vary substantially across individuals, perhaps better accommodating the heterogeneity in recovery following TKA.14 Additionally, for any individual patient, the prognosis is estimated from a subset of patients with similar characteristics. This may improve the generalizability of the predictions over commonly used mathematical approaches (e.g. multiple regression, mixed effects models), where the parameters of the prediction model are heavily influenced by the characteristics of the sample in which it is developed, potentially limiting its application in samples with different aggregate characteristics (e.g. different sociodemographic profile of patients, etc.).

The purpose of this study was to develop a clinical prediction model for functional recovery after TKA surgery, using a novel neighbors-based prediction approach.15 The outcome of interest for this study was the Timed Up and Go (TUG) test, a clinically feasible test of mobility and a surrogate of lower extremity strength. We utilized a combination of clinical and research data, collected longitudinally over the first six months following surgery. We divided the patients temporally (by date of surgery) into a training set and test set. This split was made to mimic how the approach would be developed and tested in clinical practice. The training set was used to tune the neighbors-based prediction, in particular to choose the optimal number of matches required to achieve precise estimates with low bias and accurate coverage. The training set then also served as the donor dataset for an external validation using patients from the test set.

**Methods**

*Data sources*

This analysis utilized two existing data sources involving patients with primary, unilateral TKA: 1) data collected in routine clinical practice and 2) data from previously published longitudinal research studies. Clinical data were obtained via routine quality improvement procedures at ATI physical therapy (Greenville, SC), with surgery dates between January, 2013 and June, 2017. Research data were obtained from four previously published studies, with surgery dates between June, 2006 and May, 2017. The inclusion/exclusion criteria for these research studies have been reported elsewhere.16-19 Clinical data were not selected based on patient criteria (i.e., all patients with clinical visits were included in the dataset), although only patient records containing a preoperative and postoperative TUG assessment were utilized in this analysis. The combined dataset was divided temporally, based on surgical date, into a training set and a test set (Figure 1). All participants provided informed consent. All records were de-identified prior to use in this study, and all methods were approved by the Colorado Multiple Institutional Review Board (COMIRB) and carried out in accordance with relevant regulations.

*Timed Up and Go (TUG) Test*

The TUG is a brief test of mobility, where a patient rises from a chair, walks a distance of 3 meters and returns to a seated position in the chair. Patients were instructed to perform the test, “as quickly but as safely as possible.” The TUG demonstrates high test-retest reliability and acceptable measurement error.11,20,21 For patients awaiting TKA, the standard error of the measurement for the TUG is 1.07 seconds.11 All testers involved with data collection for this analysis followed the same set of standardized instructions for performing the TUG test.21

*Matching characteristics*

Variables used for selecting matches were patient factors common across all datasets: age (years), sex, Body Mass Index (BMI; kg/m2), and preoperative TUG time (seconds).

*Statistical Analysis*

All analyses were conducted using R version 3.5.1. The steps to generate a neighbors-based prediction by predictive mean matching are summarized in the following sections and also described In Box 1.

*Selection of Matches by Predictive Mean Matching*

Because the source datasets contained TUG assessments at irregular postoperative time-points, we estimated a 90-day postoperative TUG time for all patients using linear mixed effects models via the brokenstick package (R statistical computing etc.).22,23 The 90-day time-point was used as the distal anchor for selecting matches by predictive mean matching.24 Briefly, a brokenstick model was fit to patients in the training data with 4 knots at specific timepoints after surgery (k =  0; 14; 50; 90). Patients in the training data were then matched according to the 90-day predicted TUG time by building a linear model with matching characteristics as predictors and the 90-day brokenstick-estimated TUG time as the outcome variable.

*Flexible Modeling of Observed Data*

For each patient in the training data, the observed postoperative TUG data of the patient’s matches were used to fit a Generalized Additive Model for Location Scale and Shape (GAMLSS)3, 4. The GAMLSS model was chosen for its flexibility in modeling the median (location), variance (scale), skewness and kurtosis (shape) of the TUG as a smooth function of time (i.e. time since TKA). In particular, since TUG times are positively skewed it was preferable to employ a modeling framework that accommodates flexibility in skewness over time. A cubic spline smoother with 3 degrees of freedom (df) for the location parameter and 1 df for the scale and shape parameters was employed.

*Model Tuning via Within-Sample Testing*

The optimal number of matches (m) was chosen by the following procedure: 1) GAMLSS models were fit to the matches’ observed data for each of the 398 patients in the training set with the number of matches ranging from 10 to 397 (i.e., the total number of available patients in the training data), 2) at each increment (i.e. 10 matches; 11; 12; : : : ; 397 matches), the average bias, coverage, and precision of the predictions were calculated, and 3) the optimal number of matches was determined globally by the solution that minimized bias and optimized precision whilst retaining accurate coverage (Box 2).

*Internal and External Validation*

To test the performance of the predictions, we compared predicted vs. observed TUG times via calibration plots. For both the training and test sets, we binned the predicted TUG times by deciles. Within each decile of predicted data, the median and the standard error (95% Confidence Interval) of the observed data were calculated. The median was a better measure of central tendency given the skewness of the TUG data.

**Results**

In the training data set we analyzed information on 397 patients with 1,339 post-operative TUG observations. We used information on 202 patients (604 observations) in the testing data. Patient characteristics from training and testing data are shown in Table 1. Although the sex distribution and BMI were similar across the two data sets, there were statistically significant differences in age and baseline TUG time. Compared to the patients in the training data, patients in the test data were approximately 2 years older on average, with 1 second slower baseline TUG times.

*Selection of Matches and Model Tuning*

Predictive Mean Matching

Age (ß = 0.037; p = 0.001), Sex (ß = 0.92; p <0.001), BMI (ß = 0.037; p = 0.02), and preoperative TUG time (ß = 0.21; p <0.001) demonstrated a statistically significant relationship with brokenstick estimates of the 90-day post-operative TUG time. Preoperative TUG time carried the biggest weight in selecting matches; the standardized coefficient for preoperative TUG time was 4.7 times larger than for BMI.

Examining the Optimal Number Matches

The optimal number of matches was found to be m=35 based on the low bias (0.005 standard deviations) and accurate coverage (proportion of realized observations within the 50% prediction interval: 0.50). Additionally, the average width of the 50% prediction interval with m=35 matches was 2.33 seconds (Figure 2). With m=397 matches (i.e., the full training dataset), the average precision was 3.03 seconds. Thus, the neighbors-based prediction with m=35 matches resulted in a 23% improvement in precision (Figure 3).

*Performance via Internal and External Validation*

Once the number of matches was fixed via tuning procedures in the training dataset, the within-sample and out-of-sample calibration was examined. The training dataset supplied donor data for both of these analyses. This mimics the how the development and testing of the approach would work in practice. Model calibration was good, with close agreement between predicted and observed values of post-operative TUG times (Figure 4).

**Discussion**

We developed and tested a novel, neighbors-based prediction for physical function following TKA. Via predictive mean matching, Body Mass Index (BMI), sex, age, and preoperative TUG time were used to identify the matches for an index patient. In our approach, the observed data from these matches could then be used to generate a prediction for a new patient’s TUG prognosis. One of our primary findings was the number of matches (m=35) required to generate predictions with optimal bias, coverage and precision. This solution demonstrated very low bias and accurate coverage (proportion of realized observations within the 50% prediction interval = 0.50). On average, the 50% prediction interval was 2.33 seconds, which is within the measurement error of the TUG test in this population.

The predictions were well-calibrated in both the training and test datasets. In a temporally distinct test sample (i.e., more recent surgical dates) the predictions performed accurately across all deciles of observed data. This was especially encouraging given the differences in patient characteristics between training and test datasets (Table 1) and considering national-level changes to TKA care, which occurred during the period of data collection.25 To our knowledge, this is the first study to successfully externally validate a prediction model for physical function in TKA.

There are several features of the approach that may have contributed to the accuracy of the predictions. First, estimates were based on flexible models of empirical observations, which may have allowed for more realistic representations of recovery compared to previous approaches. Second, the selection of neighbors (and, hence, the resulting prediction) was performed independently for each patient, thus each patient was the nucleus of his or her prediction. This may have improved external validity since each individual patient’s prediction was generated from similar patients’ observed recovery. Finally, matches were determined by predictive mean matching relative to a distal time-point (i.e., brokenstick estimate of 90-day TUG time). Therefore, the matching characteristics (e.g., age, sex, BMI) were weighted according to the strength of their relation to the outcome of interest, which differs from (more conventional) *k*-nearest neighbors’ approaches that pre-set the measure to express distances between patients without an explicit role for the outcome.

A combination of research and clinical data were used for this analysis. There was substantial variation across the source datasets, both geographically (i.e., sites in South Carolina and Colorado) and in care processes (i.e., multiple surgical and rehabilitation practices). While possible in theory, we did not incorporate these clinical care factors into our analysis, primarily because operationalizing the content and quality of care is an ambitious research challenge in itself.26,27 One could expect that the absence of these factors in our predictions would degrade the quality of out-of-sample predictions (since the variety of clinical practice patterns were not taken into account), but we did not find that in the data. In principle, one could improve the quality of the prediction even further by accounting for by clinical factors such as the quality of the surgery and subsequent rehabilitation. Ideally, such factors should be operationalized as matching characteristics in the construction of the matching metric. Thus, predictions could be conditional not only on patient characteristics, but also on the anticipated plan of care for the patient. Ultimately, one could envision such predictions informing not only prognosis but also potential treatment response (i.e. what is the clinical course given this or that surgical implant or rehabilitation pathway). When done appropriately, such predictions could be valuable in informing clinical decisions that are currently under-informed by evidence.

The temporal validation we performed was a methodologically rigorous form of out-of-sample testing, designed to mimic the development and prospective testing of the neighbors-based predictions in practice. The training set was also used as donor data for external validations, to generate predictions for patients with later surgical dates in the test set.

Our approach is not without limitations. All of the patients included in our analysis underwent a postoperative TUG assessment. Therefore, our dataset may be less likely to include patients who experienced a major postoperative complication, and our predictions may underestimate the likelihood of extremely poor recovery. This challenge (assessing longitudinal outcome in patients with difficulties in recovery) is not unique to our study, and such outcomes are quite rare in TKA, but this potential bias should nevertheless be considered when employing this methodology in future work. Prospective validation of these predictions is still necessary to fully understand the performance.

In conclusion, a novel neighbors-based prediction approach was used to estimate postoperative TUG times following TKA surgery, utilizing patient age, sex, BMI and preoperative TUG time. Predictions performed accurately in estimating observed TUG times at any point during first six months following surgery, according to both within-sample and out-of-sample testing. This approach could be used to inform the understanding of functional prognosis for individual patients for this common elective surgery.

**References**

1 Bernstein, J. & Derman, P. Dramatic increase in total knee replacement utilization rates cannot be fully explained by a disproportionate increase among younger patients. *Orthopedics* **37**, e656-659, doi:10.3928/01477447-20140626-58 (2014).

2 Cheuy, V. A. *et al.* Influence of Diabetes Mellitus on the Recovery Trajectories of Function, Strength, and Self-Report Measures After Total Knee Arthroplasty. *Arthritis Care Res. (Hoboken)*, doi:10.1002/acr.23741 (2018).

3 Bade, M. J., Kohrt, W. M. & Stevens-Lapsley, J. E. Outcomes before and after total knee arthroplasty compared to healthy adults. *J. Orthop. Sports Phys. Ther.* **40**, 559-567, doi:10.2519/jospt.2010.3317 (2010).

4 Judd, D. L., Eckhoff, D. G. & Stevens-Lapsley, J. E. Muscle strength loss in the lower limb after total knee arthroplasty. *Am. J. Phys. Med. Rehabil.* **91**, 220-226; quiz 227-230, doi:10.1097/PHM.0b013e3182411e49 (2012).

5 Weiss, J. M. *et al.* What functional activities are important to patients with knee replacements? *Clin. Orthop. Relat. Res.*, 172-188 (2002).

6 Alemi, F., Erdman, H., Griva, I. & Evans, C. H. Improved Statistical Methods are Needed to Advance Personalized Medicine. *Open Transl. Med. J.* **1**, 16-20, doi:10.2174/1876399500901010016 (2009).

7 Ekman, I. *et al.* Person-centered care--ready for prime time. *Eur. J. Cardiovasc. Nurs.* **10**, 248-251, doi:10.1016/j.ejcnurse.2011.06.008 (2011).

8 Leplege, A. *et al.* Person-centredness: conceptual and historical perspectives. *Disabil. Rehabil.* **29**, 1555-1565, doi:10.1080/09638280701618661 (2007).

9 Dawes, M. *et al.* Sicily statement on evidence-based practice. *BMC Med. Educ.* **5**, 1, doi:10.1186/1472-6920-5-1 (2005).

10 Sanchez-Santos, M. T. *et al.* Development and validation of a clinical prediction model for patient-reported pain and function after primary total knee replacement surgery. *Sci. Rep.* **8**, 3381, doi:10.1038/s41598-018-21714-1 (2018).

11 Kennedy, D. M., Stratford, P. W., Riddle, D. L., Hanna, S. E. & Gollish, J. D. Assessing recovery and establishing prognosis following total knee arthroplasty. *Phys. Ther.* **88**, 22-32, doi:10.2522/ptj.20070051 (2008).

12 Mizner, R. L., Petterson, S. C., Stevens, J. E., Axe, M. J. & Snyder-Mackler, L. Preoperative quadriceps strength predicts functional ability one year after total knee arthroplasty. *J. Rheumatol.* **32**, 1533-1539 (2005).

13 Explaining the Success of Nearest Neighbor Methods in Prediction. *Foundations and Trends® in Machine Learning* **10**, 337-588, doi:10.1561/2200000064 (2018).

14 Page, M. G. *et al.* Distinguishing problematic from nonproblematic postsurgical pain: a pain trajectory analysis after total knee arthroplasty. *Pain* **156**, 460-468, doi:10.1097/01.j.pain.0000460327.10515.2d (2015).

15 van Buuren, S. Curve matching: a data-driven technique to improve individual prediction of childhood growth. *Ann. Nutr. Metab.* **65**, 227-233, doi:10.1159/000365398 (2014).

16 Stevens-Lapsley, J. E., Bade, M. J., Shulman, B. C., Kohrt, W. M. & Dayton, M. R. Minimally Invasive Total Knee Arthroplasty Improves Early Knee Strength But Not Functional Performance: A Randomized Controlled Trial. *J. Arthroplasty*, doi:S0883-5403(12)00126-X [pii]

10.1016/j.arth.2012.02.016 (2012).

17 Stevens-Lapsley, J. E., Balter, J. E., Wolfe, P., Eckhoff, D. G. & Kohrt, W. M. Early neuromuscular electrical stimulation to improve quadriceps muscle strength after total knee arthroplasty: a randomized controlled trial. *Phys. Ther.* **92**, 210-226, doi:ptj.20110124 [pii]

10.2522/ptj.20110124 (2012).

18 Bade, M. J. *et al.* Early High-Intensity Versus Low-Intensity Rehabilitation After Total Knee Arthroplasty: A Randomized Controlled Trial. *Arthritis Care Res. (Hoboken)* **69**, 1360-1368, doi:10.1002/acr.23139 (2017).

19 Loyd, B. J., Kittelson, A. J., Forster, J., Stackhouse, S. & Stevens-Lapsley, J. Development of a reference chart to monitor postoperative swelling following total knee arthroplasty. *Disabil. Rehabil.*, 1-8, doi:10.1080/09638288.2018.1534005 (2019).

20 Naylor, J. M. *et al.* Minimal detectable change for mobility and patient-reported tools in people with osteoarthritis awaiting arthroplasty. *BMC Musculoskelet. Disord.* **15**, 235, doi:10.1186/1471-2474-15-235 (2014).

21 Podsiadlo, D. & Richardson, S. The timed "Up & Go": a test of basic functional mobility for frail elderly persons. *J. Am. Geriatr. Soc.* **39**, 142-148, doi:10.1111/j.1532-5415.1991.tb01616.x (1991).

22 De Kroon, M. L., Renders, C. M., Van Wouwe, J. P., Van Buuren, S. & Hirasing, R. A. The Terneuzen birth cohort: BMI changes between 2 and 6 years correlate strongest with adult overweight. *PLoS One* **5**, e9155, doi:10.1371/journal.pone.0009155 (2010).

23 Anderson, C., Hafen, R., Sofrygin, O. & Ryan, L. Comparing predictive abilities of longitudinal child growth models. *Stat. Med.* **38**, 3555-3570, doi:10.1002/sim.7693 (2019).

24 Buuren, S. v. *Flexible imputation of missing data*. Second edition. edn, (CRC Press, Taylor & Francis Group, 2018).

25 Medicare Program; Comprehensive Care for Joint Replacement Payment Model for Acute Care Hospitals Furnishing Lower Extremity Joint Replacement Services. Final rule. *Fed. Regist.* **80**, 73273-73554 (2015).

26 Zanca, J. M. *et al.* Advancing Rehabilitation Practice Through Improved Specification of Interventions. *Arch. Phys. Med. Rehabil.* **100**, 164-171, doi:10.1016/j.apmr.2018.09.110 (2019).

27 Hart, T. *et al.* A Theory-Driven System for the Specification of Rehabilitation Treatments. *Arch. Phys. Med. Rehabil.* **100**, 172-180, doi:10.1016/j.apmr.2018.09.109 (2019).

**Author Contributions**

All authors contributed to the study conception and design. CK, SvB, KC, and AK contributed to the data analysis and initial drafting of the manuscript. All authors contributed to revisions of the manuscript and approved of the final version.

**Additional Information**

The authors have no conflicts of interest to report

**Figure 1.** CONSORT flow diagram of patient data included in the analysis.

**Figure 2.** Performance metrics for neighbor’s-based predictions across increasing number of matches in the training dataset: a) bias, b) coverage, and c) precision. The optimal number of matches (m=35) is indicated with an red arrow.

**Figure 3.** The 50% prediction interval (PI) for (a) the population level estimate, is wider than the 50% prediction interval for the (b) neighbors-based prediction, for a 55 year old male with BMI of 30 kg/m2 and preoperative TUG time of 8 seconds.

**Figure 4.** Calibration plots for neighbors based predictions in: a) training and b) test datasets. Training and test datasets were were divided into deciles according to the predicted TUG times . For each decile, the median observed TUG time is plotted against the median predicted TUG time. Error bars indicate the standard error of the median.

**Table 1.** Baseline Characteristics of Training and Test Datasets

|  |  |  |  |
| --- | --- | --- | --- |
|  | Train | Test p-value*a* | |
| (n = 397, 1339 observations) | (n = 202, 604 observations) |  |
| Age, years; mean (sd) | 64.04 (8.43) | 65.90 (8.84) | 0.012 |
| Sex distribution, n (% male) | 185 (46.6) | 84 (41.6) | 0.280 |
| BMI, kg/m2; mean (sd) | 31.33 (5.82) | 31.98 (6.20) | 0.208 |
| Preop TUG, seconds; mean (sd) | 9.98 (4.95) | 11.00 (5.04) | 0.018 |

*a*Continuous variables tested with one-way analysis of variance; Categorical variables tested with χ2 test. Abbreviations: Preop TUG: Preoperative Timed Up and Go time; sd: standard deviation; BMI: Body Mass Index.