Cluster analysis for knee arthroplasty

Daan Toben

Astrid de Wind

Carlien Straat

Eva van der Meij

Judith Huirne

Han Anema

2025-10-20

Abstract

## Introduction

Knee arthroplasty (KA) is an effective procedure to mitigate the debilitating impact of osteoarthritis (OA) of the knee [@Li2019]. By 2050, the global prevalence of knee OA is projected to grow by 74.9% [@Steinmetz2023]. Working age patients drive this trend [@Culliford2015; @Nemes2015; @Romanini2019] and their expectations differ from the traditional KA demographic, focusing on the resumption of work and sports activities [@Witjes2017]. A better understanding of recovery following KA and how this impacts these patients’ return to work (RTW) is needed to help patients meet their goals.

Trajectory analysis can be a valuable tool in this regard, mapping the heterogeneity of patients’ recovery trajectories into distinct, clinically meaningful patterns [@Nagin2010]. Prior research using these or similar methods has established a basis for refuting the one-size-fits-all model of recovery. Van Egmond et al. [@Egmond2021] found three patterns, adding a middle-of-the-road responders group to the dichotomous picture presented by others [@Dumenci2019; @Harmelink2021; @Lange2020]. Potentially actionable factors like preoperative pain or anxiety were found to be linked to a poor longitudinal outcome profile. Whether these findings generalize to a working age population remains unclear, and the impact of these patterns on the resumption of work has also not yet been investigated.

In this study, we therefore aim to investigate the heterogeneity of functional recovery following KA surgery in working age patients, and to link the resulting patterns to RTW. We look to contribute to the existing knowledge base by including more data points in our operationalization of recovery. Furthermore, we aim to integrate model selection and substantive interpretation of our clusters in a systematic and reproducible manner. Our research question is: Which patterns of recovery can be identified for working age patients recovering from knee arthroplasty? Our secondary questions are 1) How do patients differ between these patterns in terms of their patient characteristics?; and 2) How do these patterns impact patients’ return to work?

## Methods

### Study design and data

We used data collected through the ACTIVE trial, a multicenter, single-blind, randomized, placebo-controlled trial [@Straat2023]. The trial was registered in the Netherlands National Trial Register (NL8525). It aimed to investigate the (cost)-effectiveness of the mHealth intervention *ikHerstel* in speeding up patients’ recovery following surgery. Recruitment ran from *Oct 2020* through *Feb* *2023*, with a one-year follow-up period, and aimed to include a sample that was generalizable to the Dutch population of working age KA patients. Patients were recruited using waiting lists of surgical departments across eleven hospitals and clinics in the Netherlands. Data were collected using web-based surveys disseminated at standard time points relative to patients’ date of surgery. Measurement times included a preoperative baseline, 4, 6, and 8 weeks as well as 3, 4, 5, 6, 9, and 12 months following surgery. We used the STROBE statement and the GRoLTS-Checklist to guide our report of the findings [@Schoot2017].

#### Variables

Our main outcome measure was physical functioning, operationalized through the Dutch-Flemish PROMIS-PF item v. 1.2 [@Terwee2014]. Specifically, we used a short form tailored to the needs of patients undergoing knee arthroplasty, which condenses the 121 item bank into 29 activities which patients use to select eight activities that they felt were most important to their recovery. Construct validity and responsiveness for using the short form in this way was investigated previously; the short form t-scores were shown to correlate to the mobility subscale of the WHO Disability Assessment Schedule 2.0 as well as the Short Form-Physical Functioning at T0, T2 and T3; and responsiveness was shown for six out of seven hypotheses [@Meij2018].

Return to work was operationalized as the number of days between surgery and full time resumption of work activities. This was measured using  digital surveys where patients reported if they had resumed work, and if so at what date.

Patient characteristics were used as descriptive variables and as predictors for cluster membership. These were measured at baseline and include: age (years), sex (male/female), educational attainment (low/intermediate/high)(CBS, 2025), body mass index (kg/length2), type of surgery (total- or unicompartmental knee arthroplasty), terms of employment (employee/self-employed/both), average work hours per week, recovery expectations prior to surgery (faster than average/slower than average), perceived health prior to surgery (EQ-5D VAS scale 0-100)[@EuroQol1990], experienced fatigue measured through the Multidimensional Fatigue Inventory (MFI) [@Smets1995], Achieving RTW within the study’s data collection period (yes/no), and complications (yes/no).

### Statistical analysis

Statistical analyses were performed using R version 4.3.2. Preliminary analyses involved an assessment of the distribution of the data, data harmonization, and exploration of missing data. Density plots were used for continuous variables and bar charts for categorical variables. Data harmonization was performed in accordance with the secondary character of the data, to safeguard data compatibility and content equivalence [@Adhikari2021]. We used descrptive statistics to compare our dataset to the original, whereby two researchers (DT, AdW) inspected the results and any inconsistencies with the original publication were discussed with its principal author (CS) [INSERT REF CARLIEN].

Missing data analysis was exploratory, consisting of an assessment of the mechanism for missingness through data visualization and Little’s MCAR test. Multiple Imputation, although a valid and reliable approach to missing data in analyses focused on a population parameter, was not performed as its application in cluster analysis is a difficult proposition [@Basagaña2013]. Complete case analysis was used instead.

#### Cluster analysis

A Latent Class Growth Model (LCGM) and a Growth Mixture Model (GMM) were specified in order to cluster each trajectory of physical function following KA into distinct patterns of recovery. Time was modeled as a continuous variable using the measurement times specified by the original trial. Preliminary analysis revealed nonlinear trajectories and these were modeled using a natural spline with 3 degrees of freedom. Our procedure involved four steps [@Jung2008]: First, an unconditional single class growth model was constructed. Second, *k* successive unconditional models were construed, where within-class variance was kept to 0. Third, free estimation of the within-class variance around the intercept was performed. We chose not to explore free estimation of the growth factors to prevent overfitting. At this point, having construed *k* models with targeted free estimation of the random intercepts, the fourth step involved selecting the number of classes.

At each of these steps, we used three model fit indicators to assess model improvement. We used the Bayesian Information Criterion (BIC) which takes the likelihood of the model as well as its number of parameters in consideration. Lower values of BIC indicate better model fit and a decrease of 10 points is frequently cited as sufficient improvement [@Kass1995]. We also used the Bootstrapped Likelihood Ratio Test (BLRT) to test improvement, and we used entropy as a measure of class separation.

Lastly, we interviewed clinicians themselves in order to determine the number of classes, assess the model’s clinical relevance, and to substantively interpret each of the clusters’ trajectories. In determining We conducted structured interviews with five clinical experts representing the entirety of the patient journey for recovery following KA. During the interviews, our preliminary findings were shared with the clinicians using an interactive dashboard. Screenshots of the dashboards are presented in appendix X. The interview guide was focused on free interpretation of the course of each cluster according to the clinician’s expertise, and it closes with their selecting the *k* model they feel best reflects the clinical practice of recovery following KA. It is included in appendix Y.

Descriptive statistics were used to describe and compare clusters on the basis of patient characteristics. We report the mean and standard deviation for normally distributed variables, the median (interquartile range [IQR]) for skewed variables and frequencies (%) for categorical variables. As a test of inter-cluster differences, we used one-way Analysis Of Variance (ANOVA) for normally distributed variables, Kruskal-Wallis ANOVA for skewed variables, and Fisher’s exact test for categorical variables. We applied a statistical significance threshold of α = 0.05.

### Regression analyses

Cox regression analysis was used to investigate the relationship between the patterns and postoperative RTW. Multinomial regression analysis was used to assess the relationship between preoperative characteristics and recovery patterns. We manually implemented the R3step procedure from MPlus to R to correct for the uncertainty inherent in pattern membership.

## Results

### Patient characteristics

Baseline PROMIS-PF data and sufficient follow-up data were available for 256 patients. Table 1 shows the patient characteristics for the entire sample, as well as per cluster. Data harmonization showed no irregularities between our data and the original publication [INSERT REF CARLIEN]. Missing data analysis indicated data weren’t entirely MCAR, in spite of Little’s MCAR test resulting in a p-value of 0.43.

| **Characteristic** | **Overall** N = 256*1* | **Moderate recovery** N = 143*1* | **Good recovery** N = 77*1* | **Poor recovery** N = 36*1* | **p-value***2* |
| --- | --- | --- | --- | --- | --- |
| **Age** | 58.4 (4.6) | 58.1 (4.5) | 59.0 (4.5) | 58.3 (5.2) | 0.41 |
| **Sex** |  |  |  |  | <0.001 |
| Female | 136 (53%) | 85 (59%) | 26 (34%) | 25 (69%) |  |
| Male | 120 (47%) | 58 (41%) | 51 (66%) | 11 (31%) |  |
| **BMI** | 29.4 (26.6-32.1) | 30.4 (27.1-33.0) | 27.8 (26.0-30.0) | 29.9 (26.6-32.2) | 0.005 |
| **Surgery type** |  |  |  |  | <0.001 |
| TKP | 174 (68%) | 106 (74%) | 39 (51%) | 29 (81%) |  |
| UKP | 82 (32%) | 37 (26%) | 38 (49%) | 7 (19%) |  |
| **Prior knee arthroplasty surgery** |  |  |  |  | 0.40 |
| 1e primaire knieprothese | 230 (90%) | 131 (92%) | 66 (86%) | 33 (92%) |  |
| 2e primaire knieprothese | 26 (10%) | 12 (8.4%) | 11 (14%) | 3 (8.3%) |  |
| **Educational attainment** |  |  |  |  | 0.76 |
| low | 24 (9.4%) | 15 (10%) | 5 (6.5%) | 4 (11%) |  |
| medium | 134 (52%) | 77 (54%) | 39 (51%) | 18 (50%) |  |
| high | 98 (38%) | 51 (36%) | 33 (43%) | 14 (39%) |  |
| **MVI total score** | 11.2 (8.4-13.0) | 11.6 (8.8-13.2) | 9.4 (7.4-11.4) | 12.9 (11.4-13.8) | <0.001 |
| **Average pain past week** | 6.60 (1.87) | 6.68 (1.73) | 6.06 (2.12) | 7.47 (1.42) | 0.002 |
| **EQ-5D VAS health score** | 66 (17) | 65 (16) | 72 (17) | 55 (18) | <0.001 |
| **Recovery expectations** |  |  |  |  | 0.007 |
| Quicker than average | 152 (59%) | 84 (59%) | 54 (70%) | 14 (39%) |  |
| Slower than average | 104 (41%) | 59 (41%) | 23 (30%) | 22 (61%) |  |
| **terms of employment** |  |  |  |  | 0.52 |
| Salaried | 220 (86%) | 125 (87%) | 62 (81%) | 33 (92%) |  |
| Self-employed | 27 (11%) | 14 (9.8%) | 11 (14%) | 2 (5.6%) |  |
| Both salaried and self-employed | 9 (3.5%) | 4 (2.8%) | 4 (5.2%) | 1 (2.8%) |  |
| **work hours/week** | 32 (24-40) | 32 (24-40) | 35 (28-40) | 31 (20-40) | 0.023 |
| **Return to work; days** | 112 (68-183) | 119 (80-202) | 85 (35-126) | 180 (127-365) | <0.001 |
| **Censored from RTW** | 222 (87%) | 126 (88%) | 75 (97%) | 21 (58%) | <0.001 |
| *1*Mean (SD); n (%); Median (Q1-Q3) | | | | | |
| *2*Kruskal-Wallis rank sum test; Pearson's Chi-squared test; Fisher's exact test | | | | | |

### Cluster analysis

The one-class cluster analysis solution showed a large degree of heterogeneity in the progression of physical function following surgery. BIC values, presented in Table 2, continued to improve over successive iterations of models without a clear injunction point.

| Models | LL | BIC | AIC | BLRT | Entropy | Npm |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | -8222.282 | 16472.29 | 16454.56 | 0.0000000 | 0.000 | 5 |
| 2 | -7666.339 | 15388.13 | 15352.68 | 0.0000000 | 0.920 | 10 |
| 3 | -7427.090 | 14937.36 | 14884.18 | 0.0000000 | 0.939 | 15 |
| 4 | -7332.374 | 14775.65 | 14704.75 | 0.0020040 | 0.919 | 20 |
| 5 | -7281.120 | 14700.87 | 14612.24 | 0.0060060 | 0.867 | 25 |
| 6 | -7247.719 | 14661.79 | 14555.44 | 0.3086172 | 0.883 | 30 |
| 7 | -7216.124 | 14626.33 | 14502.25 | 0.0490982 | 0.880 | 35 |

Based on testing, the BLRT computed for the 3-class versus the 2-class model resulted in p-value below the significance threshold (p < .001). The BLRT continued to indicate significant improvement through 5 classes but stopped at the extraction of a sixth (p = 0.308). The 3-class solution was chosen as the final model (Figure 3). The entropy value for this model was 0.93 and the average posterior probability of all classes was above 0.97, indicating excellent class separation [@Weller2020].

The three recovery patterns described by this model are of a similar shape, and have been labelled accordingly. The largest, cluster B, consisted of 0% of our sample (n = 0) and was labelled *Moderate recovery* due to the middle-of-the-road character of the recovery curve. A slight over-representation of female patients was observed in this cluster (NaN%), as well as the highest average BMI score (NaN) and a UKP vs TKP distribution (NaN% vs NaN%) similar to the overall sample. The proportion of patients with high expectations (r RCT %>% filter(lcga3 == "B") %>% summarise(``ound(mean(expectations\_rtw == "Quicker than average") \* 100, 2``), 2)) %>% pull()) was likewise similar, and perceived health was over an average level (NaN/100). Few of these patients were self-employed (NaN%.

The next in order of size, cluster A (n = 0), was labelled *Good recovery* due to the sharp and uninterrupted growth trend of the recovery curve. Male patients are overrepresented (NaN%), average BMI is relatively low (NaN), and the ratio KP/TKP is close to a half. Prior to surgery, patients` experienced fatigue score was low (NaN), and their perceived health relatively high (NaN/100). The largest proportion of self-employed patients were observed in this cluster (r RCT %>% filter(lcga3 == "A") %>% summarise(round(mean(``employment == "self-employed") \* 100, 2``)) %>% pull() %).

The smallest group, cluster C (n = 0), was labelled as *Poor recovery* due to the absence of growth across it. This group was characterized by an over-representation of female patients (NaN%) with a BMI close to the population average (NaN) undergoing mostly TKP (NaN%). Prior to surgery, their experienced fatigue was high (NaN) and their perceived health low (NaN/100). Only NaN% of this group was self-employed.

| **Characteristic** | **Moderate recovery** N = 143*1* | **Good recovery** N = 77*1* | **Poor recovery** N = 36*1* |
| --- | --- | --- | --- |
| **Baseline** | 36.1 (33.5-37.9) | 39.6 (37.0-43.3) | 32.5 (30.0-34.4) |
| **4 weeks follow-up** | 34.9 (32.2-37.5) | 40.1 (38.3-44.0) | 28.0 (26.8-31.5) |
| **6 weeks follow-up** | 37.2 (35.2-39.5) | 43.7 (40.8-47.5) | 32.9 (31.2-35.2) |
| **8 weeks follow-up** | 40.4 (38.0-42.2) | 47.1 (42.8-49.2) | 34.1 (31.8-36.3) |
| **3 months follow-up** | 42 (40-45) | 50 (47-54) | 34 (32-37) |
| **4 months follow-up** | 43 (41-46) | 53 (50-57) | 36 (33-38) |
| **5 months follow-up** | 44 (42-46) | 54 (51-57) | 37 (34-39) |
| **6 months follow-up** | 45 (43-48) | 55 (52-60) | 36 (33-39) |
| **9 months follow-up** | 47 (44-49) | 56 (53-61) | 37 (34-40) |
| **12 months follow-up** | 47 (44-50) | 57 (53-61) | 37 (35-41) |
| *1*Median (Q1-Q3) | | | |

### Regression analysis

For each regression analysis, the largest group (*Moderate recovery)* was chosen as the reference group. Table 4 shows the results of the cox regression analysis investigating the relationship between our clusters and Return to Work (RTW). There was a significant difference in RTW between members of the *Good recovery* group compared to *Moderate recovery* (HR = 2.05, 95% CI = 1.51 – 2.77, p = <0.001), indicating a greater chance of RTW for these patients. Conversely, patients in the *Poor recovery* group were half as likely to return to work (HR = 0.46, 95% CI = 0.28 – 0.75, p = 0.002) as were patients in the *Moderate recovery* group. The Schoenfeld residual test shows that the residuals were independent of time (p = 0.09). Time-based interaction terms (RTW cut-off set at 70) showed no interaction (p = 0.86). Kaplan Meier curves are presented in appendix Z.

| **Characteristic** | **log(HR)** | **95% CI** | **p-value** |
| --- | --- | --- | --- |
| **Cluster** |  |  |  |
| *Moderate recovery* | — | — |  |
| *Good recovery* | 0.72 | 0.41, 1.0 | <0.001 |
| *Poor recovery* | -0.77 | -1.3, -0.29 | 0.002 |
| Abbreviations: CI = Confidence Interval, HR = Hazard Ratio | | | |

# weights: 30 (18 variable)  
initial value 263.993978   
iter 10 value 199.463185  
iter 20 value 183.582369  
iter 30 value 183.430691  
final value 183.430657   
converged

For the multinomial logistic regression analysis, all covariates (OR; 95% CI) were entered simultaneously. For the *Good recovery* group, male sex (2.26; 0.15-1.48), Unicompartmental surgery (3.26; 0.48-1.88), and VAS health score (1.03; 0.01-0.05), were significantly associated. For the *poor recovery* group*,* predictors included VAS pain (1.37; 0.02-0.62), VAS health score (0.98; -0.05-0), and low expectations for RTW (2.55; 0.05-1.83) (Table 5).

| **Characteristic** | **log(OR)** | **95% CI** | **p-value** |
| --- | --- | --- | --- |
| Good recovery | | | |
| (Intercept) | -4.5 | -11, 1.7 | 0.2 |
| Age | 0.05 | -0.02, 0.13 | 0.2 |
| Sex |  |  |  |
| Female | — | — |  |
| Male | 0.81 | 0.15, 1.5 | 0.017 |
| BMI | -0.09 | -0.18, 0.00 | 0.061 |
| Surgery type |  |  |  |
| TKP | — | — |  |
| UKP | 1.2 | 0.48, 1.9 | <0.001 |
| Average pain past week | -0.14 | -0.32, 0.04 | 0.12 |
| EQ-5D VAS health score | 0.03 | 0.01, 0.05 | 0.016 |
| Recovery expectations |  |  |  |
| Quicker than average | — | — |  |
| Slower than average | -0.42 | -1.1, 0.28 | 0.2 |
| Censored from RTW | 1.6 | -0.14, 3.3 | 0.072 |
| Poor recovery | | | |
| (Intercept) | -0.95 | -8.3, 6.4 | 0.8 |
| Age | 0.00 | -0.09, 0.09 | >0.9 |
| Sex |  |  |  |
| Female | — | — |  |
| Male | -0.14 | -1.1, 0.82 | 0.8 |
| BMI | -0.02 | -0.10, 0.07 | 0.7 |
| Surgery type |  |  |  |
| TKP | — | — |  |
| UKP | -0.54 | -1.6, 0.49 | 0.3 |
| Average pain past week | 0.32 | 0.02, 0.62 | 0.039 |
| EQ-5D VAS health score | -0.02 | -0.05, 0.00 | 0.035 |
| Recovery expectations |  |  |  |
| Quicker than average | — | — |  |
| Slower than average | 0.94 | 0.05, 1.8 | 0.039 |
| Censored from RTW | -1.4 | -2.4, -0.48 | 0.003 |
| Abbreviations: CI = Confidence Interval, OR = Odds Ratio | | | |

### Interviews

We interviewed five clinicians: two physical therapists (PT), a physician assistant (PA), an orthopedic surgeon (OP), and an occupational physician (CP), using the interactive dashboard to showcase each model. They were not unanimous in their decision for the optimal solution, the physical therapists and the orthopedic surgeon preferred the 3 cluster model, whilst the physician assistant and the occupational physician preferred 4 clusters. The difference between the two groups depended on a balance between parsimoniousness versus validity. For example, one physical therapist noted: *“For practice… use the three groups. That would be easy to understand for everyone, not too complicated. It would be a nice model to bring to the clinic, I think”*.

Clinicians quickly labelled clusters A, B, and C as the good, the average and the poor responders to surgery, respectively. For PA and PO, the addition of cluster D represented an above average group of patients that separated the good (D) from the excellent (A), which they found to be lacking in the 3-class solution. The CP noted, however, that they hadn’t seen many KA surgeries in the past year.

Interestingly, both PTs nuanced their initial negative interpretation of the *Poor recovery* group, criticizing their own clinical perspective as potentially reductionist: *“So even if the green line  has a nice trajectory, clinically speaking, and a nice endpoint given its start, that doesn’t mean that these people are satisfied. Perhaps they much higher expectations or are driven by higher demands put on them by their work.”*  All clinicians except CP felt an interest in early identification of the poor responder group, for both selection and intervention purposes: *“It would be interesting, if this is the prediction, to manipulate it. Could you go about intervening on those poor recovery? Can we improve their predicted course?”*.

## Discussion

We set out to identify and classify heterogeneity in patients’ recovery of physical function following TKA surgery. Our findings reject a one-sized fits all approach to postoperative recovery following KA. Three distinct patterns - which we labelled *Moderate recovery, Good recovery,* and *Poor recovery* - were found. Cox regression analysis revealed membership of the *Good recovery* group was associated with an increased likelihood of RTW, and membership of the *Poor recovery* group with a decreased likelihood of RTW, compared to *Moderate recovery*. Multinomial logistic regression analysis showed female sex, unicompartmental surgery, and higher preoperative perceived health were associated with membership to the *Good recovery* group. At the same time, higher preoperative pain, lower perceived health, and low RTW expectations prior to surgery were associated with membership to the *Poor recovery* group.

Our results were consistent with other studies [@Egmond2021; @Dowsey2015; @Dumenci2019] in finding a *Poor recovery* group. The lack of a clinically relevant increase in physical function between surgery and one year follow-up, begs the question if surgery was beneficial for these patients. At the same time, the strong associations with potentially actionable factors likelow perceived health, low expectations, and increased preoperative fatigue and pain indicate a potential for preoperative intervention or at least selection. This corresponds with previous research [@Vissers2012; @Singh2010] indicating that physical and mental well-being prior to surgery affect functional recovery afterwards. Working age patients seem no different from traditional KA patients in this regard, and neither does their early identification seem to be any less of a priority looking at the group’s decreased chance of RTW.

When it comes to non-actionable factors, the link between *poor* or *moderate* group membership and female sex is surprising, as studies show no difference in functional outcomes based on sex for these surgeries between men and women [@Ayers2023; @Clement2020; @Lim2015]. This may be explained by the higher burden of OA carried by women in terms of prevalence, pain, and functional limitation [@Segal2024] predisposing them to poorer outcomes. This does not necessarily indicate a confounding relationship, as biological factors like menopause may underlie women’s greater burden of OA given the age range of our sample [@Segal2024]. The high proportion of total KA surgery in both poor and moderate groups was expected given the relative severity of the surgery. The fact that no difference was found between the two groups supports the hypothesis that the course of recovery is more than just a function of the type of surgery employed.

Other studies found a *Moderate recovery* pattern [@Egmond2021; @Golineli2023; @Dowsey2015], although its inclusion clashes with some two-class solutions [@Lange2020]. Compared to the *Poor recovery* group, this group’s improved recovery in the presence of a similar proportion of TKP and improved actionable factors like preoperative pain, perceived health, and preoperative expectations is suggestive of the utility of preoperative intervention to improve postoperative outcomes. The psychosocial character aligns with other studies [@Vonk2014; @Levett2019]. Future research might shed light on this relationship by investigating whether the higher perceptions of health reported by these patients are accurate reflections of their health or whether they’re reflections of a more positive outlook, compared to *Poor recovery* patients.

### Strengths and limitations

Several strengths underlie these reflections. The high number of measurements across the 12 month study period resulted in a detailed picture of physical function following KA surgery. It connects patients’ experiences in the acute period after surgery with the mid to long term. This higher temporal resolution has only nuanced the character of recovery for our clusters, however; it did not result in a different solution when comparing to van Egmond et al. (2020), who used fewer measurements. The use of clinical expertise in substantive interpretation and selection of the cluster models is another strength. Their critical reflection on our findings affirmed some of our preconceptions, like our hypothesis that psychosocial factors underlie differences in functional outcomes, whilst also uncovering flaws. Lastly, our selection of working age patients fills a gap in a field which has so far mostly been focused on older patients.

Several limitations must be addressed. The first is the secondary character of our data, which has meant that we could not analyze some variables nevertheless relevant to KA, like the degree of comorbidity. The inclusion of a Patient Reported Experience Measure (PREM) like a patient’s satisfaction regarding the outcome of their surgery, could have provided more insight into the different patient groups represented by each cluster. As highlighted by some clinicians, it would be interesting to challenge our clinical perspective on these clusters, to verify whether patients in the G*ood recovery* group are actually more satisfied with their procedure compared to patients exhibiting *Poor recovery*.

Another limitation lies in the low number of patients included in our study, increasing the risk of type 2 errors. Cluster analysis is more susceptible to this error as it divides a group into an indeterminate amount of subgroups of unequal size. This would warrant extra care in disregarding associations where small groups are concerned, like categorical variables. No such associations were found in this study, however.

The limited size of the sample was compounded by the exclusion of patients who had fewer than 4 PROMIS scores, missing baseline scores, missing the first three follow-up scores, or missing 4 consecutive follow-up scores, leading to the loss of 13% of patients. This represents another limitation, as filling out questionnaires may very well be related to one’s course of recovery or other factors related to it. Appendix Z shows that this seems to be the case, as a greater proportion of self-employed patients are present in this group, with a much faster RTW a greater proportion of censored patients. Possibly, these patients represent a very successful group whose lack of experienced difficulties caused them to renege on filling out the questionnaire.

### Implications for practice and research

Our findings may benefit clinicians and patients in their preoperative appointments. Patients may find it useful to have a visualization of the diversity of experiences while doctors may find it helps them communicate that same message, managing patients’ expectations. At the same time, the factors associated to each recovery pattern may help surgeons select patients who would benefit most from surgery, and filter out those who wouldn’t. Future research may build on this utility by solidifying these results in a prediction model. A different operationalization of recovery would be preferrable over cluster labels, however, as the exploratory character of the analysis undercuts the generalizability of any algorithm predicting cluster membership.

## Conclusions

Recovery from knee- and hip arthroplasty is not only heterogenous, distinct recovery patterns are apparent and the patients exhibiting these patterns can be identified by their characteristics prior to surgery. Some of these, like sex, are static and limited to serve as early detectors of potentially suboptimal recovery prior to surgery. In this role, they may nevertheless benefit recovery by enhancing shared decision making if they are leveraged to manage patients’ expectations. Others, however, seem actionable and might be leveraged to alter a patient’s course from poor recovery to a moderate or even good response.

## Appendix X: Screenshots of the dashboard

Screenshot 1: LCGA raw plot for 1-class

[Dashboard image A](dashboard%20img%20A.PNG)

Screenshot 2: LCGA raw plot for 4-class

[Dashboard image B](dashboard%20img%20B.PNG)

Screenshot 3: LCGA median plot for 4-class

[Dashboard image C](dashboard%20img%20C.PNG)

Screenshot 4: LCGA versus GMM\_int for 5-class

[Dashboard image D](dashboard%20img%20D.PNG)

## Appendix Y: Clinician Interview Guide

### Introduction

1. **Greeting and Purpose**:
   * *“Thank you for taking the time to meet with me. I’d like to hear your thoughts on the patient clusters we’ve identified in our analysis of recovery of physical function following total and unicompartmental knee arthroplasty.”*
   * *“The purpose of this interview is to provide interpretation to our data and hopefully improve their validity and interpretability”*
2. **Dashboard Overview**:
   * Briefly walk them through the interactive dashboard, explaining the key features:
     + Longitudinal outcomes (PROMIS-PF scores)
     + Clustering method

### Main body

**Section 1: General Impressions (5 minutes)**

1. **Initial Thoughts**
   * *“What are your initial impressions on the clusters?*

**Section 2: Cluster-Specific Insights (10 minutes)**

1. **Cluster Interpretation**:
   * *“Looking at* [INSERT SPECIFIC CLUSTER]*, what characteristics or clinical insights come to mind for this group?”*
   * *“Does the recovery pattern for* [INSERT SPECIFIC CLUSTER] *align with any patient groups you encounter?“*
     1. *(If not, does it remind you of any single cases?)*
   * Repeat for other clusters

**Section 3: Clinical relevance (7 minutes)**

1. **Clinical relevance:**
   * *“Based on your experience, do these clusters align with what you see in practice?”*
   * *“Are there any surprising patterns or groups that stand out to you?”*
   * *“Are there patterns that you’re missing?*”
2. **Implications for Practice**:
   * *“Do you see potential for using these clusters in clinical practice?”*
   * *“How would you use this information in clinical practice?”*
   * *“Are there specific interventions that you think might be more effective for one cluster compared to others?”*

**Section 4: Final Thoughts (3 minutes)**

1. **Missing Factors**:
   * *“Is there anything missing that would have made the clusters more clinically useful?”*
2. **Future Directions**:
   * *“Would you be interested in using a similar dashboard/data-driven approach in your own practice?”*
3. **Final thoughts:**
   * *“Do you have any final thoughts or comments?”*

**Conclusion (2 minutes)**

* *“Thank you for your time and valuable input, would you be interested in being kept up-to-date regarding our research’s progress?”*
* *May we approach you for a follow-up meeting?*

## Appendix Z: Excluded versus included patients

| **Characteristic** | **Overall** N = 293*1* | **0** N = 256*1* | **1** N = 37*1* | **p-value***2* |
| --- | --- | --- | --- | --- |
| **Age** | 58.4 (4.6) | 58.4 (4.6) | 58.0 (5.0) | 0.72 |
| **Sex** |  |  |  | 0.74 |
| 1 | 136 (47%) | 120 (47%) | 16 (50%) |  |
| 2 | 152 (53%) | 136 (53%) | 16 (50%) |  |
| **BMI** | 29.4 (26.6-32.2) | 29.4 (26.6-32.1) | 28.3 (26.9-34.0) | 0.89 |
| **Surgery type** |  |  |  | 0.58 |
| 0 | 91 (31%) | 82 (32%) | 9 (27%) |  |
| 1 | 198 (69%) | 174 (68%) | 24 (73%) |  |
| **Prior knee arthroplasty surgery** | 31 (11%) | 26 (10%) | 5 (14%) | 0.56 |
| **Educational attainment** |  |  |  | 0.37 |
| 1 | 1 (0.3%) | 0 (0%) | 1 (3.1%) |  |
| 2 | 27 (9.4%) | 24 (9.4%) | 3 (9.4%) |  |
| 3 | 130 (45%) | 115 (45%) | 15 (47%) |  |
| 4 | 20 (6.9%) | 19 (7.4%) | 1 (3.1%) |  |
| 5 | 97 (34%) | 86 (34%) | 11 (34%) |  |
| 6 | 13 (4.5%) | 12 (4.7%) | 1 (3.1%) |  |
| **MVI total score** | 11.0 (8.2-13.0) | 11.2 (8.4-13.0) | 9.6 (7.6-12.4) | 0.11 |
| **Average pain past week** | 6.56 (1.96) | 6.60 (1.87) | 6.23 (2.63) | 0.76 |
| **EQ-5D VAS health score** | 34 (14) | 34 (14) | 37 (15) | 0.18 |
| **Recovery expectations** |  |  |  | 0.17 |
| 1 | 175 (61%) | 152 (59%) | 23 (72%) |  |
| 2 | 113 (39%) | 104 (41%) | 9 (28%) |  |
| **terms of employment** |  |  |  | 0.026 |
| 1 | 243 (84%) | 220 (86%) | 23 (72%) |  |
| 2 | 36 (13%) | 27 (11%) | 9 (28%) |  |
| 3 | 9 (3.1%) | 9 (3.5%) | 0 (0%) |  |
| **work hours/week** | 32 (24-40) | 32 (24-40) | 31 (20-40) | 0.95 |
| **Return to work; days** | 111 (63-182) | 112 (68-183) | 61 (42-66) | 0.029 |
| **Censored from RTW** |  |  |  | 0.004 |
| = 'Niet' | 39 (15%) | 34 (13%) | 5 (56%) |  |
| = 'Wel' | 226 (85%) | 222 (87%) | 4 (44%) |  |
| *1*Mean (SD); n (%); Median (Q1-Q3) | | | | |
| *2*Wilcoxon rank sum test; Pearson's Chi-squared test; Fisher's exact test | | | | |

## Appendix ZZ: Kaplain Meier curves of Return to Work per cluster

