Published in final edited form as:
Arthritis Care Res (Hoboken). 2022 August ; 74(8): 1359–1368. doi:10.1002/acr.24564.

Application of Heterogeneity of Treatment Effect Methods: Exploratory Analyses of a Trial of Exercise-Based Interventions for Knee OA

Cynthia J. Coffman, PhD[1,2], **Liubov Arbeeva**[3,4], **Todd A. Schwartz, DrPH**[3,5,6], **Leigh F. Callahan, PhD**[3,4,7], **Yvonne M. Golightly, PT, MS, PhD**[3,8,9,14], **Adam P. Goode, PT, DPT, PhD**[10,11,12], **Kim M. Huffman, MD, PhD**[13], **Kelli D. Allen, PhD**[1,3,4]

[1]Center of Innovation to Accelerate Discovery and Practice Transformation, Durham VA Healthcare System, HSRD (152), 508 Fulton Street, Durham NC 27705,

[2]Department of Biostatistics and Bioinformatics, Duke University Medical Center

[3]Thurston Arthritis Research Center, University of North Carolina at Chapel Hill

[4]Department of Medicine, University of North Carolina at Chapel Hill

[5]Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill

[6]School of Nursing, University of North Carolina at Chapel Hill

[7]Departments of Orthopaedics and Social Medicine

[8]Injury Prevention Research Center, University of North Carolina at Chapel Hill

[9]Department of Epidemiology, University of North Carolina at Chapel Hill

[10]Duke Clinical Research Institute, Duke University School of Medicine

[11]Department of Orthopedic Surgery, Duke University Medical Center

[12]Department of Population Health Sciences, Duke University Medical Center

[13]Department of Medicine, Division of Rheumatology, Duke University Medical Center

[14]Division of Physical Therapy, University of North Carolina at Chapel Hill

Abstract

Objective: To evaluate heterogeneity of treatment effects (HTE) in a trial of exercise-based interventions for knee osteoarthritis (OA).

Methods: Participants (n=350) were randomized to standard physical therapy (PT; n=140), Internet-Based Exercise Training (IBET; n=142), or wait list control (WL; n=68). We applied

Corresponding author: Cynthia J. Coffman, PhD, Durham VA Medical Center, HSR&D(152) 508 Fulton St. Durham, NC 27705, Tel: 919-286-6936 Fax: 919-416-5836, Cynthia.Coffman@duke.edu.

Authors Contributions
CJC, LA, and KDA planned the statistical analyses; CJC and LA executed the analyses. KDA, LA, LFC, YMG, APG, KMH and TAS contributed to the trial design and conduct. All authors read and approved the final manuscript.

QUalitative INteraction Trees (QUINT), a sequential partitioning method, and Generalized Unbiased Interaction Detection and Estimation (GUIDE), a regression tree approach, to identify subgroups with greater improvements in Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) score over 4-months. Predictors included 24 demographic, clinical and psychosocial characteristics. We conducted internal validation to estimate optimism (bias) in the range of mean outcome differences among arms.

Results: Both QUINT and GUIDE indicated that for participants with lower body mass index (BMI), IBET was better than PT (improvements of WOMAC ranged from 6.3 to 9.1 points lower) and for those with higher BMI and longer duration of knee OA, PT was better than IBET (WOMAC improvement was 6.3 points). In GUIDE analyses comparing PT or IBET to WL, participants not employed had improvements in WOMAC ranging from 1.8 to 6.8 points lower with PT or IBT vs. WL. From internal validation, there were large corrections to the mean outcome differences among arms; however, after correction some differences remained in the clinically meaningful range.

Conclusion: Results suggest there may be subgroups who experience greater improvement in symptoms from PT or IBET, and this could guide referrals and future trials. However, uncertainty persists for specific treatment effect size estimates and how they apply beyond this study sample.

Trial Registration: NCT02312713

Physical therapy (PT) and exercise-based interventions are core components of knee osteoarthritis (OA) treatment[1,2]. However, overall effects of these interventions tend to be modest, with substantial variability across patients[3–5]. Patients with OA differ substantially from one another in clinical, biomechanical and psychosocial characteristics that can impact the effectiveness of exercise-based interventions[6]. In addition, there are many different types of exercise-based interventions that vary in terms of intensity, duration, delivery mode, amount of supervision, exercise type(s) and physiological target(s)[3]. There is little understanding of which types of exercise-based interventions work best for different patients with OA. This limits our ability to advise patients regarding the exercise-based intervention they may benefit from most, as well as our ability to effectively target interventions in a population-based manner. Consistent with the goals of Precision or Personalized Medicine[7], the OA community needs to develop an understanding of the “right treatment for the right patient at the right time,” in the context of exercise-based interventions, in order to maximize effectiveness.

Exploratory analyses of previous trials provide some evidence that responses to exercise-based interventions for OA may vary according to patient characteristics such as age, gender, pain severity, strength, function, malalignment, radiographic severity and psychological variables[8–12]. However, those analyses have focused on a limited set of potential moderators, since the typical statistical approach of adding interaction terms limits inclusion of a large number of candidate variables. In addition, evaluating potential treatment moderators singly may fail to identify important combinations of variables[13–15]. For example, a given exercise-based intervention may be beneficial for older adults who have low strength levels and low-to-moderate pain severity, or effects of specific interventions may differ based on different OA phenotypes[16]. New data-driven methods allow exploration of multidimensional subgroups that exhibit heterogeneous treatment effects[15,17–20], and these methods can deepen our understanding of patients' responses to exercise-based interventions.

We recently completed the PhysicAl THerapy vs INternet-Based Exercise Training for Patients with Knee Osteoarthritis (PATH-IN) randomized clinical trial that compared PT and an internet-based exercise training (IBET) program, both relative to a wait list control group (WL)[21,22]. We found that effects of PT and IBET were similar to each other and did not differ significantly from WL for the primary outcome of Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) total score in the overall sample. However, a pre-specified aim of the trial was to evaluate heterogeneity of treatment effects (HTE) to understand whether either PT or IBET may have benefits for subgroups of patients compared to each other or to WL. In this manuscript, we applied two different, advanced statistical methods to explore HTE in the PATH-IN study.

Methods

The PATH-IN study randomized individuals with knee OA to standard PT (n=140), IBET (n=142), or WL (n=68) in a 2:2:1 allocation[21,22]. The study was approved by the Institutional Review Board of the University of North Carolina at Chapel Hill.

Overview of HTE Methods

Recursive partitioning methods are the underpinning for many data-driven HTE methods[15,23,24]. The basic idea is to create trees that classify patients into subgroups based on independent variables, where treatment effect sizes are large, in opposite directions or meet some difference threshold. The methods are recursive because each subpopulation may be split again until some stopping criterion is reached.

When selecting from among HTE methods, careful attention must be paid to the specific research question being addressed[15,17–20]. The first question we addressed dealt with two active treatments (PT and IBET), exploring which treatment worked better for whom. This is known as a *qualitative subgroup interaction*, where one treatment may work better for one subgroup, while another treatment may be better for another[25,26]. QUalitative INteraction Trees (QUINT) is a sequential partitioning method that identifies whether or not qualitative subgroup effects are present, and, if so, partitions the sample into three potential subgroups: treatment A is better than B; treatment B is better than A; or, neither treatment is better[25,26]. The second question we addressed dealt with which subgroups showed the greatest improvement, relative to a control group. This is known as a *quantitative subgroup interaction*, which occurs when a treatment produces large improvements in outcomes for some patients but little to no improvement for others[19,27]. We explored which subgroups showed the greatest WOMAC score improvement in IBET compared to PT. We also explored which subgroups showed the greatest WOMAC score improvement compared to WL, including PT, IBET, and WL in one model. Generalized Unbiased Interaction Detection and Estimation (GUIDE) is a regression tree approach that identifies whether or not quantitative subgroups effects are present, and, if so, partitions the sample into subgroups with differential treatment effects[27].

Predictor Variable Selection

The potential predictors were collected at baseline, prior to randomization. We selected the most relevant variables based on our experience and evidence from previous studies[5,8–12]. Following examination of missing data and correlations between predictor variables, 24 candidate variables were selected (Table A1 in Appendix). Due to missing data in included covariates, we excluded n=5 participants; the final sample was: n=138 in PT arm, n=140 in IBET arm, and n=67 in WL control.

HTE Methods

The outcome for our analyses was change from baseline in WOMAC total score (the primary study outcome) at 4 months; a negative change indicates improvement in WOMAC. At 4-months, 45 participants missed follow-up assessment; we used Empirical Best Linear Unbiased Prediction estimates from linear mixed effects models as a single imputation for the 4-month outcome and then calculated the change score[28].

We first applied QUINT (see Appendix for details on algorithm), which is implemented with the package *quint* in R, (R 3.5.1, R Core Team, 2014)[29]. In our analyses we used the “difference in means” option for the outcome and the default options for partitioning criteria: minimum absolute effect size of 0.3 and equal weighting of effect size difference and cardinality component for determining splits[18]. We set the minimum sample size per treatment arm per subgroup to be 15 (i.e., total minimum subgroup n=30), which is close to the default option of 10%. We also ran analysis where we increased the minimum sample size per treatment arm per subgroup to 20 (for a total n=40). Finally, for all analyses, we used the *prune.quint* function to reduce overfitting and to select the optimal tree with the optimal number of subgroups. The number of bootstrap samples was 25, and as a sensitivity analysis we also set the number of bootstraps to 100; results were similar.

We then applied GUIDE (see Appendix for details on algorithm), using the subgroup identification approach (G_i option), which is implemented with the package GUIDE obtained from http://pages.stat.wisc.edu/~loh/guide.html. We set the minimum sample size per treatment arm per subgroup to be n=15 and n=20 in respective iterations of the algorithm. We used GUIDE in an analysis that included only the two active treatments, and also in a second analysis that included all three treatment arms. Pruning in GUIDE was applied with cross validation.[19]

The conclusion of a QUINT implementation yields values of the baseline variables that define the subgroups, with mean differences and sample sizes for each of the treatments in each subgroup. An implementation of GUIDE yields the values of the baseline variables that define the subgroups, estimated mean differences between treatment groups adjusted for covariates in the model, and sample sizes for each treatment in each subgroup. The mean difference (4-month WOMAC score – Baseline WOMAC score) between treatment groups for both QUINT and GUIDE when only 2 active arms included are presented as *(IBET mean difference) – (PT mean difference)*, where a negative value indicates greater improvement for IBET over PT and a positive value indicates greater improvement for PT over IBET. The mean differences between treatment groups for GUIDE when all 3 arms are included are presented as *(IBET mean difference) – (WL mean difference)* and *(PT mean difference) – (WL mean difference)*, where a negative value indicates greater improvement for IBET or PT compared to WL and a positive value indicates greater improvement for WL compared to IBET or PT.

We then conducted an internal validation for both QUINT and GUIDE analyses via bootstrap resampling to estimate optimism or bias in the range of mean outcome differences between pairs of arms in the final selected tree (i.e., the 'apparent range') due to overfitting and to provide a bias-corrected estimate[30]. We followed the steps as outlined in Section C.2 of the web appendix of Dusseldorp and Mechelen (see Appendix for details)[26].

Results

Descriptive statistics for the 24 predictor variables, overall and by treatment arm, are provided in Table 1.

QUINT results

Figure 1 displays the pruned tree (unpruned tree was the same) from QUINT when the minimum subgroup size was n=40 total participants (i.e., at least n=20 in each arm), which contains four subgroups. For the 2 subgroups that had greater improvement in IBET than PT (red), mean differences in WOMAC scores were 7.2 and 6.3 points lower in IBET compared to PT. The first subgroup (n=44) was BMI ≤ 24.31 kg/m^2 and the second (n=57) was defined by a combination of BMI (> 24.31 kg/m^2), duration of OA symptoms (≤ 9.5) years and social support for exercise score (≤ 56.5 points). For the two subgroups with greater improvement in PT than IBET (green), mean differences in WOMAC scores were 3.1 points and 6.3 points lower in PT compared to IBET, respectively. The first subgroup (n=50) was defined by a combination of BMI (> 24.31 kg/m^2), duration of OA symptoms (≤ 9.5 years) and social support for exercise score (> 56.5 points), and the second subgroup (n=127) was defined by BMI (> 24.31 kg/m^2) and duration of OA symptoms (> 9.5 years).

Figure 2 displays the pruned tree (the unpruned tree had an additional split for subgroup 4) from QUINT when the minimum subgroup size was n=30 total participants (i.e., at least n=15 in each arm), which contains five subgroups. For the three subgroups that had greater improvement with IBET than PT (red), mean differences in WOMAC scores were 9.1, 8.0 points and 4.9 points lower in IBET than PT, respectively. One subgroup (n=38) was composed of individuals with lower BMI, with a cutoff similar to the QUINT analysis with a minimum subgroup size of n=40 (i.e., 23.94 kg/m^2). The second group (n=33) had higher BMI (>23.94 kg/m^2) and younger age (≤ 55.54 years), and the third subgroup (n=42) included older individuals (>72 years) with better performance on chair stands (>8.5 stands). For the two subgroups that had greater improvement with PT than IBET (green), mean differences in WOMAC scores were 9.6 and 4.7 points lower in PT than IBET, respectively. One group (n=86) included participants with higher BMI (> 23.94 kg/m^2) and worse performance on the 30-second chair stand (≤ 8.5 stands), and the other group (n=79) had higher BMI (> 23.94 kg/m^2), age between 55.5 and 72.0 years and better performance on the 30-second chair stand (>8.5 stands).

QUINT Internal validation

For the QUINT analysis internal validation when the minimum subgroup size was n=40 total participants (i.e., at least n=20 in each arm), the apparent range, the difference between the largest negative difference in means between arms in a subgroup (−7.2), and largest positive difference in means between arms in a subgroup (6.3), was −13.5 (Table A2 in Appendix). When this apparent range (−13.5) was corrected for estimated optimism, it was reduced to −4.5, well below the −8.0 that we would deem as a clinically meaningful difference in WOMAC change[31]. Similarly, for the QUINT procedure when the minimum subgroup size was n=30 participants (at least n=15 in each arm), we found a large reduction in the apparent range from −18.7 to −9.0 when corrected for estimated optimism (Table A2 in Appendix). However, this corrected apparent range of −9.0, was above the minimum range expected and was indicative of clinically meaningful difference in WOMAC change.

GUIDE results

Figure 3 displays the unpruned tree from GUIDE when applied to the two active arms when the minimum subgroup size was n=40 total participants (i.e., at least n=20 in each arm). The pruned tree was empty. Similar to QUINT, the tree contains four subgroups. However, one subgroup had greater improvement in WOMAC with IBET compared to PT (red), another subgroup had greater improvement in WOMAC with PT than IBET (green), and two subgroups showed no difference between IBET and PT (grey). The subgroup that had a larger improvement with IBET than PT (6.9 points lower, n=59) was composed of participants with duration of OA symptoms ≤ 9.5 years and BMI ≤ 29.45 kg/m^2. The subgroup that had greater improvement with PT than IBET (5.7 points lower, n=67) was composed of participants with duration of OA symptoms > 18.5 years.

Figure 4 displays the unpruned tree from GUIDE when applied to all three treatment arms when the minimum subgroup size was n=60 total participants (i.e., at least n=20 in each arm). The pruned tree was empty. The tree contains two subgroups, where one subgroup had greater improvement in IBET and PT compared to WL, and the other subgroup had greater improvement in WL compared to IBET and PT. The subgroup (n=205) that had lower mean differences in WOMAC for IBET (3.8 points) and PT (6.4 points), compared with WL, was composed of individuals not currently employed. The subgroup (n=140) for which there were larger mean differences in WOMAC for WL compared to IBET (0.7 points) or PT (1.2 points) was composed of individuals currently employed. Figure 5 displays the unpruned tree from GUIDE including all three arms when the minimum subgroup size was n=45 total participants (i.e., at least n=15 in each arm); the pruned tree was empty. This unpruned tree contains four subgroups, with the first split variable of employment status and then the two employment status groups further subdivided to obtain the four subgroups. There were two subgroups for which IBET and PT both had lower mean differences than WL control by 5.7, 6.8, 1.8 and 5.8 points, respectively; one subgroup (n=115) consisted of participants who were not currently employed, with duration of OA symptoms ≤ 10.5 years, and the other subgroup (n=90) included participants who were not currently employed and duration of symptoms > 10.5 years. For one subgroup (n=64), mean differences were lower for IBET than WL control but greater for WL than PT, though these differences were small; this group included individuals who were currently employed and had lower scores on the chair stand test (≤ 9.5 stands). For the last subgroup (n=76), mean differences were lower for WL than for either IBET (2.6 points) and PT (2.1 points); this subgroup was currently employed and had better chair stand scores (> 9.5 stands).

GUIDE Internal validation

For the GUIDE internal validation for two active arms, the apparent range was −12.6 (see Table A2 in Appendix), and when corrected based on our estimate for optimism, it was reduced to −7.2. For the internal validation of the GUIDE procedure for three arms, comparing PT to WL when the minimum subgroup size was set to n=45 total participants (i.e., at least n=15 in each arm), the apparent range was −8.9 and with optimism correction reduced to −5.0. In all cases for GUIDE, there were large reductions of the apparent range after correcting for optimism, but many remained on the border of clinically meaningful differences (Table A2 in Appendix).

Discussion

In this three-arm study, we addressed several research questions related to HTE for different exercise-based treatments among individuals with knee OA. We used QUINT to address qualitative subgroup interactions in the two active treatment arms, exploring which treatment worked better for which subgroups, and GUIDE to explore the more general question of whether some subgroups had larger improvements than others between the two active treatments. Based on results involving the two active arms, an overall observation was that BMI, age and disease duration seemed to be important factors regarding whether PT or IBET yielded greater improvement. Although some other factors contributed to subgroup identification, these three easily assessed patient characteristics could help to guide referrals in clinical situations. In particular, these results suggest patients who are older, have higher BMI, and have had knee OA symptoms for a longer period of time may particularly benefit from the personalized support and tailored exercise offered by a physical therapist vs. a more self-directed exercise program.

We used data from all three study arms to explore, using GUIDE, which subgroups showed the greatest improvement in each of the active treatment arms compared to usual care (WL). For the subgroup of participants who were not employed, both IBET and PT had greater improvements than WL (Figure 4); improvement relative to WL was somewhat larger for PT than IBET. However, for those who were employed, WL was associated with larger improvements in WOMAC than either PT or IBET after adjusting for covariates including baseline WOMAC score. It is notable that the primary driver of magnitude of effects (relative to WL) was employment status, a factor different from those involved in the comparisons of two active treatment arms. A likely explanation for the three-group GUIDE results is that participants who were not employed had more time to engage in the intervention, including adherence to home exercise recommendations. It is interesting that this pattern was observed for the IBET group (though less pronounced than for PT), given that no in-person visits were required for the intervention, and exercises could be completed at participants’ convenience. Individuals with knee OA who are still employed may need additional support or strategies to fit regular activity into daily routines. Based on the GUIDE model with a minimum of n=15 participants per arm, the PT intervention had a particularly strong impact for participants who had OA symptoms for a longer period of time (among those in the not employed subgroup). These results align with findings of GUIDE analyses of the two active groups, in which patients with the longest duration experienced greater benefit from PT (Figure 3).

An important aspect of both the QUINT and GUIDE procedures is pruning of trees to avoid overfitting. For the QUINT analysis there was little to no difference between pruned and unpruned trees for both analyses. In our GUIDE analysis, when we applied the pruning procedures, all trees were empty. Therefore, results should be interpreted with caution due to the potential for overfitting and instability of identified subgroups. As these are exploratory analyses in a relatively small sample size for HTE analyses (but typical for a clinical trial of a behavioral intervention), we presented unpruned trees for GUIDE. However, we applied pre-pruning procedures by specifying limits on how small subgroups could be, a step to prevent overfitting. Furthermore, we used internal validation to evaluate the bias in our effect size differences due to overfitting when applying both QUINT and GUIDE.

Another important aspect of our methods was the process of internal validation. In HTE analyses, validation provides guidance for interpreting and applying results, even when analyses are exploratory and in studies with small samples where overfitting can easily occur. We applied these procedures to estimate optimism or bias in the range of mean differences in outcomes for the final tree (i.e., the apparent range). Based on internal validation results, there were large corrections to all apparent ranges, reflecting potential bias. However, some of the optimism-corrected ranges still fell in the clinically meaningful range, indicating that meaningful differences in subgroups may apply beyond this sample. Specifically, in the QUINT analysis with two active treatments, there were large potential biases in the apparent range for both n=20 and n=15 per arm for each subgroup, indicating overfitting and instability of results. In the analyses including the two active arms only, an optimism-corrected range greater than −8 is indicative of subgroups with clinically meaningful differences that may apply beyond this sample [31]. In the analysis with n=15 per arm, the corrected range was above the threshold of −8, but was well below this threshold in the n=20 per arm analysis[31]. It could be that applying the n=20 per arm per subgroup is too stringent of a criterion, masking smaller subgroups with larger and more stable differences. In GUIDE with all three arms, there were large potential biases in the apparent range for both n=20 and n=15 per arm per subgroup. In this case, an optimism-corrected range greater than −4 is indicative of subgroups with clinically meaningful differences of treatment vs. control that may apply beyond this sample. Similar to the QUINT analysis, the n=15 per subgroup per arm yielded optimism-corrected ranges greater than this threshold, while the n=20 per subgroup per arm yielded optimism-corrected ranges below this threshold.

There are some limitations to this study. We have not focused the interpretation of results on uncertainty estimates of treatment differences in subgroups, as standard uncertainty estimates do not account for all the uncertainty due to the data-driven process; this is an area of active research in HTE analysis. In this pragmatic trial, we did not obtain radiographs, and information on radiographic severity was not systematically available within the electronic heath record. We also did not collect data on joint malalignment. Both radiographic severity and malalignment have some previous evidence for moderating effects of exercise-based interventions for OA and therefore would have been valuable to include in these analyses. However, current radiographic severity and precise measurement of joint alignment are not available at all clinical encounters; therefore, these variables may not be the most practical for use in guiding recommendations for exercise-based interventions.

In summary, these analyses highlight ways to use various data-driven HTE methods to address different research questions within randomized clinical trials, depending on whether the trial has two active treatments (where the QUINT method applies) or usual care and one or more treatment arms (where the GUIDE method applies). Problematic areas that need to be considered or addressed when applying these methods are multiple testing implications, potential for too much complexity, appropriate uncertainty estimation and reproducibility of subgroups[15]. While these analyses were exploratory in nature, they provide some evidence that there are subgroups for whom different exercise-based treatments (IBET vs. PT) were more efficacious than for others; even after correcting for optimism bias, some differences were in the clinically meaningful range. In particular, our results suggest that younger patients with lower BMI may be good candidates for self-guided exercise programs (e.g., our IBET intervention) and that regardless of the type of exercise-based intervention, individuals who are currently employed may need additional supportive strategies. However, additional studies are needed to further explore HTE in the context of exercise-based therapies for OA, including different programs and cohorts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The study team thanks all of the study participants, without whom this work would not be possible. We thank the following team members for their contributions to the research: Caroline Nagle, Kimberlea Grimm, Ashley Gwyn, Bernadette Benas, Alex Gunn, Leah Schrubbe, and Quinn Williams. The study team also expresses gratitude to the Stakeholder Panel for this project: Ms. Sandy Walker LPN (Chapel Hill Children’s Clinic), Ms. Susan Pedersen RN BSN, Ms. Sally Langdon Thomas, Mr. Ralph B. Brown, Ms. Frances Talton CDA RHS Retired, Dr. Katrina Donahue, MD, MPH (Department of Family Medicine at the University of North Carolina at Chapel Hill), Dr. Alison Brooks, MD, MPH (Department of Orthopedics & Rehabilitation at the University of Wisconsin-Madison), Dr. Anita Bemis-Dougherty, PT, DPT, MAS (American Physical Therapy Association), Dr. Teresa J. Brady, PhD (Centers for Disease Control and Prevention), Ms. Laura Marrow (Arthritis Foundation National Office), Ms. Megan Simmons Skidmore (American Institute of Healthcare and Fitness), and Dr. Maura Daly Iversen, PT, DPT, SD, MPH, FNAP, FAPTA (Department of Physical Therapy, Movement and Rehabilitation Sciences Northeastern University). The study team thanks study physical therapists and physical therapy assistants: Jennifer Cooke, PT, DPT, Jyotsna Gupta, PT, PhD and Carla Hill, PT, DPT, OCS, Cert MDT (Division of Physical Therapy, University of North Carolina at Chapel Hill), Bruce Buley, Andrew Genova, and Ami Pathak (Comprehensive Physical Therapy, Chapel Hill, NC), Chris Gridley and Aaron Kline (Pivot Physical Therapy, Smithfield, NC).

Role of the Funding Source

This study was funded through a Patient-Centered Outcomes Research Institute Award (CER-1306-02043). The statements and opinions presented in this manuscript are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute, its Board of Governors or Methodology Committee. CC and KDA receive support from the Center of Innovation to Accelerate Discovery and Practice Transformation (CIN 13-410) at the Durham VA Health Care System. KDA receives support from a VA Health Services Research and Development Research Career Scientist Award (19-332). KDA, LA, LFC, YMG, APG, and TAS receive support from National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) Core Center for Clinical Research P30AR072580.

References

1. Kolasinski SL, Neogi T, Hochberg MC, et al. 2019 American College of Rheumatology/Arthritis Foundation Guideline for the Management of Osteoarthritis of the Hand, Hip, and Knee. Arthritis Care Res (Hoboken). 2020.
2. Bannuru RR, Osani MC, Vaysbrot EE, et al. OARSI guidelines for the non-surgical management of knee, hip, and polyarticular osteoarthritis. Osteoarthritis Cartilage. 2019.
3. Fransen M, McConnell S, Harmer AR, Van der Esch M, Simic M, Bennell KL. Exercise for osteoarthritis of the knee: a Cochrane systematic review. Br J Sports Med. 2015;49(24):1554–1557. [PubMed: 26405113]
4. Pignato M, Arbeeva L, Schwartz TA, et al. Level of participation in physical therapy or an internet-based exercise training program: associations with outcomes for patients with knee osteoarthritis. BMC Musculoskelet Disord. 2018;19(1):238. [PubMed: 30025540]
5. Holden MA, Burke DL, Runhaar J, et al. Subgrouping and TargetEd Exercise pRogrammes for knee and hip OsteoArthritis (STEER OA): a systematic review update and individual participant data meta-analysis protocol. BMJ Open. 2017;7(12):e018971.
6. Ross R, Goodpaster BH, Koch LG, et al. Precision exercise medicine: understanding exercise response variability. Br J Sports Med. 2019;53(18):1141–1153. [PubMed: 30862704]
7. Genetics Home Reference. National Library of Medicine. What is Precision Medicine? https://ghr.nlm.nih.gov/primer/precisionmedicine/definition. Accessed January 9, 2020.
8. Wright AA, Cook CE, Flynn TW, Baxter GD, Abbott JH. Predictors of response to physical therapy intervention in patients with primary hip osteoarthritis. Physical therapy. 2011;91(4):510–524. [PubMed: 21310898]
9. French HP, Galvin R, Cusack T, McCarthy GM. Predictors of short-term outcome to exercise and manual therapy for people with hip osteoarthritis. Phys Ther. 2014;94(1):31–39. [PubMed: 23929827]
10. Bennell KL, Dobson F, Roos EM, et al. Influence of Biomechanical Characteristics on Pain and Function Outcomes From Exercise in Medial Knee Osteoarthritis and Varus Malalignment: Exploratory Analyses From a Randomized Controlled Trial. Arthritis Care Res (Hoboken). 2015;67(9):1281–1288. [PubMed: 25623617]
11. Kudo M, Watanabe K, Otsubo H, et al. Analysis of effectiveness of therapeutic exercise for knee osteoarthritis and possible factors affecting outcome. J Orthop Sci. 2013;18(6):932–939. [PubMed: 24085378]
12. Knoop J, Dekker J, van der Leeden M, et al. Is the severity of knee osteoarthritis on magnetic resonance imaging associated with outcome of exercise therapy? Arthritis Care Res (Hoboken). 2014;66(1):63–68. [PubMed: 23982988]
13. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. The Lancet. 2005;365(9454):176–186.
14. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. Bmj-Brit Med J. 2010;340.
15. Lipkovich I, Dmitrienko A, D’Agostino SBR Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat Med. 2017;36(1):136–196. [PubMed: 27488683]
16. Deveza LA, Nelson AE, Loeser RF. Phenotypes of osteoarthritis: current state and future implications. Clin Exp Rheumatol. 2019;37 Suppl 120(5):64–72.
17. Seibold H, Zeileis A, Hothorn T. Model-Based Recursive Partitioning for Subgroup Analyses. Int J Biostat. 2016;12(1):45–63. [PubMed: 27227717]
18. Dusseldorp E, Van Mechelen I. Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. Stat Med. 2014;33(2):219–237. [PubMed: 23922224]
19. Loh WY, Fu H, Man M, Champion V, Yu M. Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables. Stat Med. 2016;35(26):4837–4855. [PubMed: 27346729]
20. Loh WY, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects. Stat Med. 2015;34(11):1818–1833. [PubMed: 25656439]

21. Williams QI, Gunn AH, Beaulieu JE, et al. Physical therapy vs. internet-based exercise training (PATH-IN) for patients with knee osteoarthritis: study protocol of a randomized controlled trial. BMC Musculoskelet Disord. 2015;16:264. [PubMed: 26416025]

22. Allen KD, Arbeeva L, Callahan L, et al. Physical therapy vs. internet-based exercise training for patients with knee osteoarthritis: Results of a randomized controlled trial. Osteoarthritis and Cartilage. 2018;In Press.

23. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods. 2009;14(4):323–348. [PubMed: 19968396]

24. Seibold H, Zeileis A, Hothorn T. Model-Based Recursive Partitioning for Subgroup Analyses. Int J Biostat. 2016;12(1):45–63. [PubMed: 27227717]

25. Doove LL, Van Deun K, Dusseldorp E, Van Mechelen I. QUINT: A tool to detect qualitative treatment-subgroup interactions in randomized controlled trials. Psychother Res. 2016;26(5):612–622. [PubMed: 26169837]

26. Dusseldorp E, Van Mechelen I. Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. Stat Med. 2014;33(2):219–237. [PubMed: 23922224]

27. Loh WY, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects. Stat Med. 2015;34(11):1818–1833. [PubMed: 25656439]

28. Littell R, Milliken G, Stroup W, Wolfinger R, Schabenberger O. SAS for Mixed Models, Second Edition. Cary, NC: SAS Press; 2006.

29. Dusseldorp E, Doove L, Mechelen I. Quint: An R package for the identification of subgroups of clients who differ in which treatment alternative is best for them. Behav Res Methods. 2016;48(2):650–663. [PubMed: 26092391]

30. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd Edition. Springer Ser Stat. 2015.

31. Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. Arthritis Rheum. 2001;45(4):384–391. [PubMed: 11501727]

Significance and Innovation

- This study is the first to use statistical methods of QUalitative INteraction Trees (QUINT) and Generalized Unbiased Interaction Detection and Estimation (GUIDE) to examine heterogeneity of treatment effects for different exercise-based treatments among individuals with knee OA.
- Both QUINT and GUIDE indicate that for participants with lower body mass index, an internet-based training program (IBET) was better than physical therapy (PT); for those with higher body mass index and longer duration of knee OA symptoms, PT was better than IBET.
- GUIDE analysis indicated that participants who were not employed had greater improvements with PT or IBET, relative to usual care.
- In HTE analyses with small samples, internal validation provides guidance for interpreting and applying results.

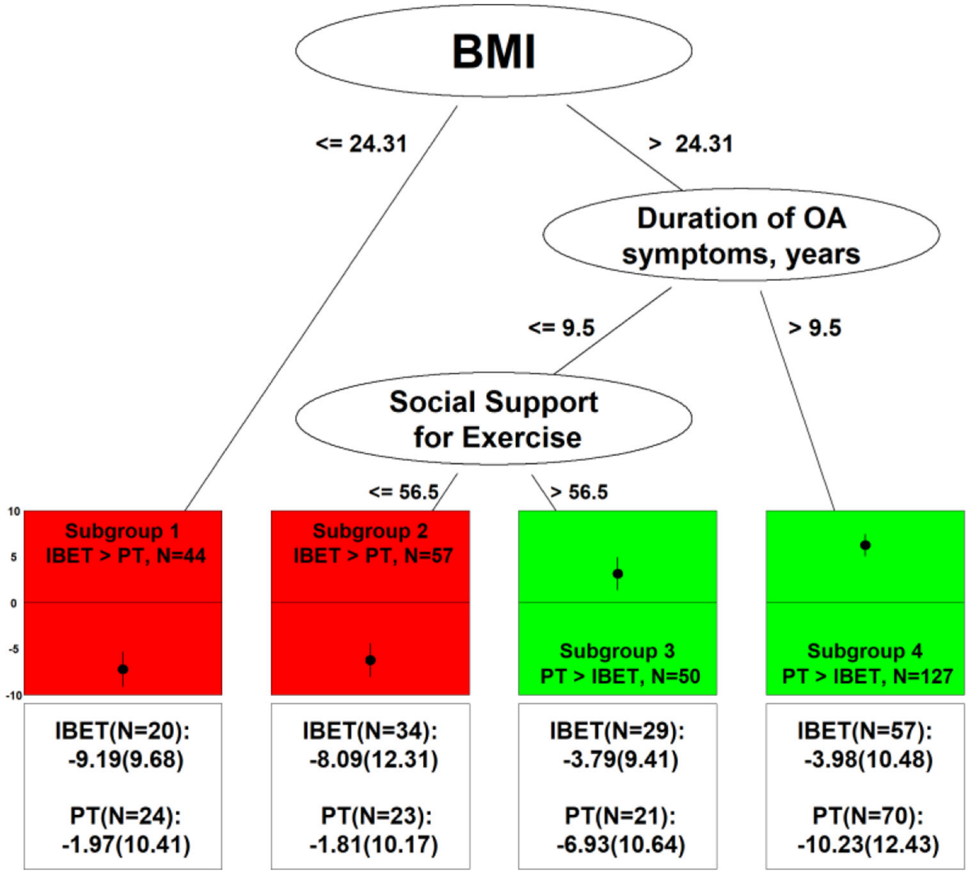

Figure 1.

QUINT subgroups with 2 active arms

Minimum sample size for subgroup is n=40 total participants (i.e., at least n=20 in each arm); The subgroups are as follows: subgroup 1: BMI ≤ 24.31 kg/m^2 (n=44; mean difference =−7.2 points); subgroup 2: BMI > 24.31 kg/m^2, OA Symptom duration ≤ 9.5 years, and score on Social Support for Exercise ≤ 56.5 (n=57; mean difference = −6.3 points); subgroup 3: BMI > 24.31 kg/m^2, OA Symptom duration ≤ 9.5 years, and score on Social Support for Exercise > 56.5 (n=50; mean difference = 3.1 points); and subgroup 4: BMI > 24.31 kg/m^2, OA Symptom duration > 9.5 years, (n=127; mean difference = 6.3 points).

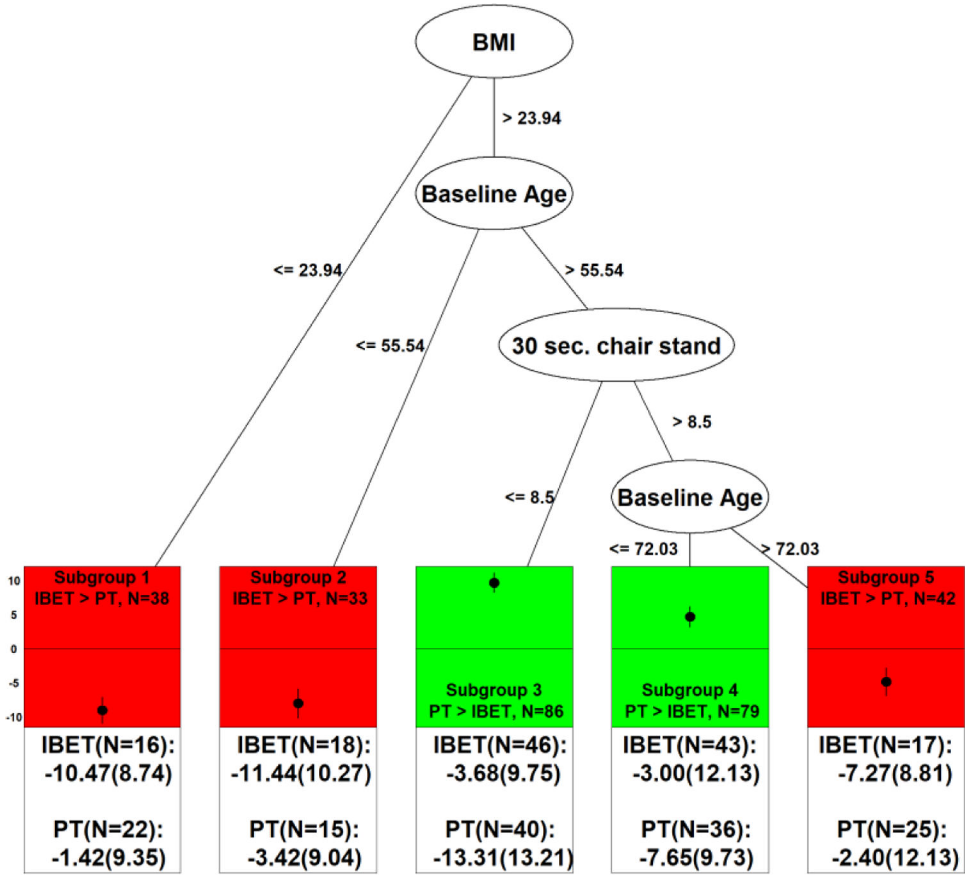

Figure 2.

QUINT subgroups with two active arms

Minimum sample size per subgroup is n=30 total participants (i.e., at least n=15 per arm); The subgroups are as follows: subgroup 1: BMI ≤ 23.94 kg/m^2 (n=38; mean difference =−9.1 points); subgroup 2: BMI > 23.94 kg/m^2, Age ≤ 55.54 years(n=33; mean difference = −8.0 points); subgroup 3: BMI > 23.94 kg/m^2, Age >55.54 years, and number of chair stands ≤ 8.5 (n=86; mean difference = 9.6 points); subgroup 4: BMI > 23.94 kg/m^2, Age >55.54 and ≤72.03 years, and number of chair stands > 8.5, (n=79; mean difference = 4.7 points), and subgroup 5: BMI > 23.94 kg/m^2, Age >72.03 years, and number of chair stands > 8.5, (n=42; mean difference = −4.9 points).

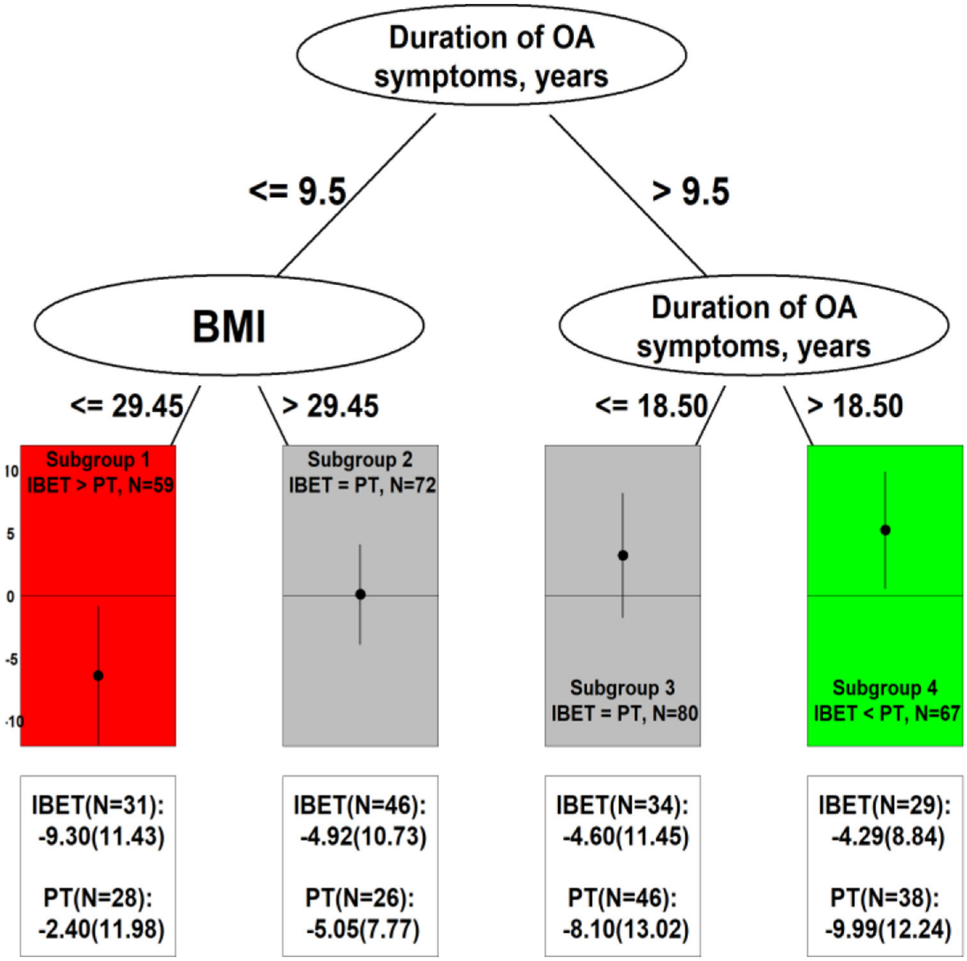

Figure 3.

GUIDE subgroups with two active arms

IBET and PT, n=278 with minimum sample size per subgroup of n=40 total participants (i.e., at least n=20 in each arm); The subgroups are as follows: subgroup 1: OA Symptom duration ≤ 9.5 years and BMI ≤ 29.45 kg/m^2 (n=59; unadjusted mean difference =−6.9 points, adjusted mean difference =−5.0 points); subgroup 2=:OA Symptom duration ≤ 9.5 years and BMI > 29.45 kg/m^2 (n=72; unadjusted mean difference =0.1 points, adjusted mean difference =−0.5 points); subgroup 3: OA Symptom duration > 9.5 years and OA Symptom duration ≤ 18.5 years (n=80; unadjusted mean difference = 3.5 points, adjusted mean difference=3.1 points); and subgroup 4: OA Symptom duration > 18.5 years (n=67; unadjusted mean difference = 5.7 points, adjusted mean difference=7.5 points).

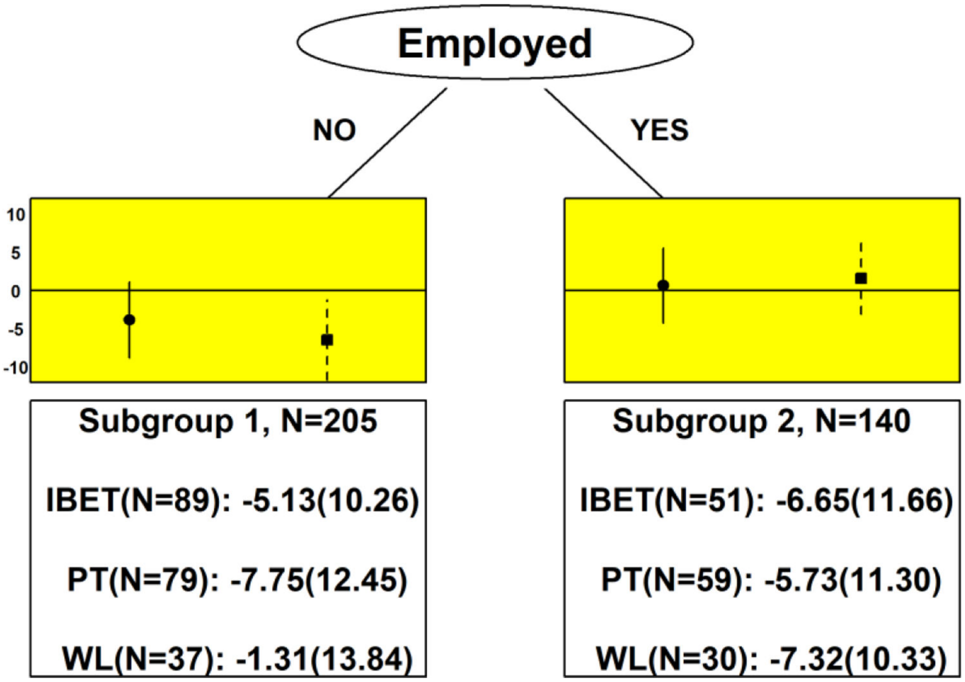

Figure 4.

GUIDE subgroups with all three arms

IBET, PT and WL. n=345 with minimum sample size per subgroup of n=60 total participants (i.e., at least n=20 in each arm); mean differences for subgroups (nodes) are IBET-Usual care followed by PT-WL with negative values indicating greater improvement in treatment arm (IBET or PT) compared to WL. The subgroups are as follows: subgroup 1: Not employed (n=205; unadjusted IBET- WL mean difference =−3.8 points, adjusted mean difference=−5.4; unadjusted PT-WL mean difference=−6.4, adjusted mean difference=−7.4); subgroup 2: Employed (n=140; unadjusted IBET- WL mean difference =0.7 points, adjusted mean difference=1.2; unadjusted PT-WL mean difference=1.6, adjusted mean difference=2.1).

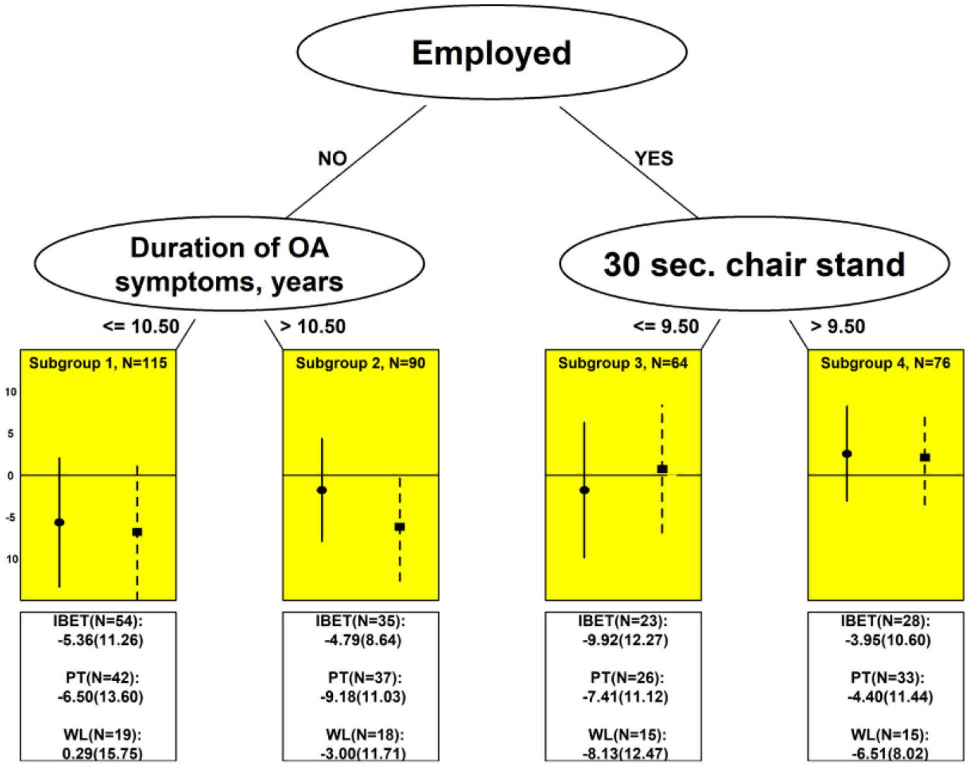

Figure 5.

GUIDE subgroups with all three arms

IBET, PT and WL n=345 with minimum sample size per subgroup of n=45 total participants (i.e., at least n=15 in each arm); Subgroups are as follows: subgroup 1: Not employed and OA symptom duration ≤ 10.5 years (n=115; unadjusted IBET- WL mean difference =−5.7 points, adjusted mean difference=−5.7; unadjusted PT-WL mean difference=−6.8, adjusted mean difference=−6.5); subgroup 2: Not employed and OA symptom duration > 10.5 years (n=90; unadjusted IBET-WL mean difference =−1.8 points, adjusted mean difference=−5.8; unadjusted PT-WL mean difference=−6.2, adjusted mean difference=−11.4); subgroup 3: Employed and number of chair stands ≤ 9.5 (n=64; unadjusted IBET-WL mean difference =−1.8 points, adjusted mean difference=−1.4; unadjusted PT-WL mean difference=0.7, adjusted mean difference=0.7); subgroup 4: Employed and number of chair stands > 9.5 (n=76; unadjusted IBET- WL mean difference =2.6 points, adjusted mean difference=3.7; unadjusted PT-WL mean difference=2.1, adjusted mean difference=2.3)

Table 1.

Descriptive statistics for baseline patient characteristics and explanatory variables (mean (SD) or N (%))

Characteristic	Total (N=345)	PT (N=138)	IBET (N=140)	WL (N=67)
Age, years	65.3 (11)	65.7 (10.3)	65.1 (11.4)	64.7 (11.7)
Women, N (%)	247 (71.6%)	99 (71.7%)	96 (68.6%)	52 (77.6%)
White Race, N (%)	251 (72.8%)	109 (79.0%)	93 (66.4%)	49 (73.1%)
Married or Living with Partner, N (%)	213 (61.7%)	78 (56.5%)	93 (66.4%)	42 (62.7%)
Bachelor's Degree or post-graduate work, N (%)	205 (59.4%)	84 (60.9%)	79 (56.4%)	42 (62.7%)
Fair or Poor Health, N (%)	48 (13.9%)	14 (10.1%)	22 (15.7%)	12 (17.9%)
Household Financial Status: Live Comfortably or Meet basic needs with a little left over for extras, N (%)	285 (82.6%)	118 (85.5%)	112 (80.0%)	55 (82.1%)
Employed full or part time, N (%)	140 (40.6%)	59 (42.8%)	51 (36.4%)	30 (44.8%)
Body Mass Index, kg/m^2	31.3 (8.0)	31.8 (8.6)	31.5 (7.7)	29.8 (6.8)
# Joints with OA Symptoms	5.3 (3.2)	5.5 (3.0)	5.1 (3.1)	5.4 (3.9)
Duration of OA Symptoms, years	13.1 (11.6)	13.9 (11.5)	11.6 (11.0)	14.4 (12.9)
History of Knee Injury, N (%)	173 (50.1%)	71 (51.4%)	69 (49.3%)	33 (49.3%)
# Problems Learning[#]	3.7 (0.7)	3.7 (0.6)	3.6 (0.8)	3.8 (0.5)
Filling Out Forms[$]	0.4 (0.7)	0.3 (0.8)	0.4 (0.8)	0.3 (0.6)
Internet Comfort[%]	4.1 (1.2)	4.1 (1.2)	4.1 (1.3)	4.1 (1.2)
Internet Frequency[^]	1.4 (1.2)	1.3 (1.1)	1.5 (1.2)	1.4 (1.3)
WOMAC Total Score, 0–96	31.9 (17.8)	32.0 (17.7)	31.3 (17.7)	33.1 (18.8)
PHQ-8 score, 0–24	3.8 (4.1)	4.0 (4.5)	3.7 (4.1)	3.5 (3.4)
PROMIS Fatigue Score	51.2 (8.8)	51.9 (9.1)	50.3 (9.0)	51.9 (7.9)
Self -Efficacy Exercise	56.2 (20.3)	57.3 (20.8)	56.8 (19.8)	52.8 (20.5)
Social Support Exercise	52.1 (18.4)	51.9 (17.4)	51.8 (19.6)	53.0 (18.3)
30 Second Chair Stand	9.6 (3.9)	9.5 (4.2)	9.6 (3.7)	9.6 (3.6)
Two Minute March Test	50.9 (29.6)	51.6 (31.0)	51.5 (29.5)	48.3 (26.8)
Unilateral Stand Test, seconds	7.3 (3.6)	7.3 (3.6)	7.4 (3.4)	6.8 (3.7)

[#] Problems learning were assessed via questionnaire consisting of one question: "How often do you have problems learning about your medical condition because of difficulty understanding written information" and reported as "always (0)", "often (1)", "sometimes (2)", "occasionally (3)", and "never (4)"

[$] Problems filling out forms were assessed via questionnaire consisting of one question: "How confident are you filling out forms by yourself?" and reported as "extremely (0)", "quite a bit (1)", "somewhat (2)", "a little bit (3)", and "not at all (4)"

[%] Comfort using internet, likert scale 1 (Not at all)…..5(Very)

[^] The frequency of internet use was reported as "every day (1)", "a few times a week (2)", "once a week (3)", "a few times a month (4)", "once a month (5)", "less than once a month (6), "not at all (7)".