Validity and generalizability of findings of randomized controlled trials on arthroscopic partial meniscectomy of the knee

A. Malmivaara

1Centre for Health and Social Economics, National Institute for Health and Welfare, Helsinki, Finland
2Orton Orthopaedic Hospital and Orton Research Institute, Orton Foundation, Helsinki, Finland

Correspondence
Antti Malmivaara, Centre for Health and Social Economics, National Institute for Health and Welfare, Helsinki, Finland.
Email: antti.malmivaara@thl.fi

The validity and generalizability of evidence from trials on arthroscopic partial meniscectomy (APM) for a ruptured meniscus of the knee has been debated in scientific journals. The aim of this paper was to assess the validity and generalizability of evidence for trials on effectiveness of APM by a novel Benchmarking method; to compare this assessment with established assessment methods; and to make an overall assessment of the current evidence on validity and generalizability of findings. Literature search was undertaken to find all randomized controlled trials. The benchmarking method, the Cochrane method for the assessment of validity of RCTs, and CONSORT method for the assessment of generalizability of findings were used. The data were extracted and checked for accuracy twice. Seven RCTs were found, of which 2 were double blinded. The benchmarking method indicated that only one trial, assessing real-world effectiveness, had recruited patients comprehensively from the catchment area and showed clinically important effectiveness for APM. All trials had deficiencies in reporting of baseline characteristics and adherence to interventions. In 5 trials, the crossover to surgery from conservative treatment arm was between 19 and 36 percent. The benchmarking method indicates that APM may be an effective treatment for meticulously selected patients, or when APM is targeted to those patients who do not respond to the conservative treatment. There is a need for more sham-controlled and real-world effectiveness trials reporting comprehensively patient characteristics and adherence to interventions, preferably in a representative sample of patients living in the recruitment area.

KEYWORDS
arthroscopic partial meniscectomy, double blinding, generalizability, randomized controlled trial, real-world effectiveness, risk of bias, validity

1 | INTRODUCTION

The effectiveness of arthroscopic partial meniscectomy (APM) for a ruptured meniscus of the knee has been under debate in scientific journals.1-9 The conclusions of this debate are based on assessments of risk of bias and generalizability of findings in randomized controlled trials (RCTs).10-13 However, a more comprehensive description of study characteristics can be obtained using the methodology for the benchmarking controlled trials (BCTs), which are intended for the assessment of validity and generalizability of findings in observational effectiveness studies.14 In observational effectiveness studies, a very detailed description of study characteristics is most important, as there are always baseline differences between treatment groups, and one must be able to control for these differences statistically.14-16 It seemed indicated that the issues of validity of the RCTs and of generalizability of findings should be studied further by...
complementing the description of study characteristics by those intended for the benchmarking controlled trials.

The aim of this study was to evaluate the validity and generalizability of the findings in the RCTs assessing effectiveness of arthroscopic surgery among patients having a painful knee and a meniscal rupture using the benchmarking method and the established assessment methods, and to make an overall assessment of the validity and generalizability of findings.

2 | METHODS

Literature search was undertaken to find all randomized controlled trials published in peer-reviewed journals assessing effectiveness of arthroscopic partial meniscectomy of the knee in comparison to any other nonpharmacological treatment, including sham surgery, among patients having knee pain, with at least 1 year follow-up. Trials focusing on knee osteoarthrosis were excluded. The following key words were used: arthroscopic partial meniscectomy, randomized controlled trial, systematic review. Cochrane CENTRAL, Ovid MEDLINE, and Web of Science databases till October 2017 were used to find the eligible articles by the author, who checked the search findings to exclude misclassifications (Appendix 1). Finally, the very recently published RCT was added to the material.17

Assessment method intended for the benchmarking controlled trials, which is called the benchmarking method (BM), was used to assess validity and generalizability of findings; and Cochrane handbook’s tool for the assessment of validity for RCTs, and CONSORT statement’s recommendation for reporting of RCTs were used as reference standards.13,14,18 The BM is based on 5 categories and several subcategories outlining the trial question and its actualization in randomized controlled trials. There are 6 subcategories describing selection of patients, eight subcategories on baseline characteristics of patients, 7 subcategories on process data, 2 subcategories on outcome data, and 2 subcategories on statistical analysis (Table 1).

The descriptive information was extracted by the author concerning selection of patients, completeness of data of baseline characteristics, treatment processes, and outcomes; and statistical analysis, including validity items relevant to RCTs. Also the supplementary materials (protocols and other supplements) of the RCTs were studied. The accuracy of the extracted data was rechecked twice.

3 | RESULTS

Altogether, 7 randomized controlled trials (RCTs) fulfilling the inclusion criteria were found (Table 2).17,19-24 Of the 7 RCTs, 5 have been carried out in the Nordic countries, one in USA, and one in South Korea.

TABLE 1 Categories and subcategories outlining the study design of randomised controlled trials (RCTs). Modified from an article on observational effectiveness studies, the benchmarking controlled trials (BCTs); some items omitted as denoting only to observational effectiveness studies and some issues added as important in RCTs

| 1 | Selection of patients/population to the study |
| 1.1 | Description of patients’ clinical eligibility criteria |
| 1.2 | Description of patients’ clinical path before being eligible for the study |
| 1.3 | Pre-intervention therapy |
| 1.4 | Comprehensiveness of patient population of the catchment area |
| 1.5 | Place and time of recruitment. Number of patients per recruiting unit per year |
| 1.6 | Number of patients declining participation |
| 2 | Completeness of baseline data |
| 2.1 | Number of patients |
| 2.2 | Clinically important data relevant to the particular disorder/disease (eg, age, gender, severity by outcome variables) |
| 2.3 | General health/risk status |
| 2.4 | Comorbid conditions |
| 2.5 | Behavioral factors (eg, on health-related lifestyle) |
| 2.6 | Environmental factors (eg, work conditions) |
| 2.7 | Inequality (eg, socioeconomic status) |
| 2.8 | Other potential predictors (eg, genetic factors), confounders, and effect modifiers |
| 3 | Completeness of process data |
| 3.1 | Content of the index treatment |
| 3.2 | Content of the control intervention |
| 3.3 | Staff competence |
| 3.4 | Healthcare system features (eg, resources, clinical paths) |
| 3.5 | Adherence to index treatments |
| 3.6 | Adherence to comparison treatments |
| 3.7 | Use of other healthcare services |
| 4 | Completeness of outcome data |
| 4.1 | Primary and secondary outcomes |
| 4.2 | Percentage of and reasons for dropping out of follow-up |
| 5. | Statistical analysis |
| 5.1 | Description of power calculations |
| 5.2 | Description and appropriateness of all primary and secondary statistical analyses |
| Year, study, country | Herrlin et al (2007 and 2013) Sweden | Katz et al (2013) USA | Siivonen et al (2013) Finland | Yim et al (2013) South Korea | Gauffin et al (2014) Sweden | Kise et al (2016) Norway | Roos et al (2018) Denmark |
|----------------------|-------------------------------------|----------------------|-------------------------------|-----------------------------|----------------------------|-------------------------|--------------------------|
| 1. Selection of patients; healthcare system features; randomization (Yes/No/Unclear) |
| 1.1. Selection of patients described | No | No | No | No | Yes | No | No |
| 1.2. Healthcare system features described | No | No | No | No | Yes | No | No |
| 1.3. Random and concealed allocation | Unclear | Yes | Yes | Unclear | Unclear | Yes | Yes |
| 2. Baseline characteristics (incl. primary outcomes); (Yes/No/Unclear) |
| 2.1. Clinically important baseline data described comprehensively | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 2.2. Other important baseline data described comprehensively | No | No | No | No | No | No | No |
| 2.3. Baseline comparability | Some uncertainty a | Some uncertainty b | Some uncertainty c | Some uncertainty c | Some uncertainty b | Some uncertainty c | Some uncertainty c |
| 3. Interventions; (Yes/No/Unclear) |
| 3.1. Adherence to the surgical intervention described | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 3.2. Adherence to the nonsurgical intervention described | No | Yes | Yes | No | No | Yes | Yes |
| 3.3. Reporting of co-interventions | No | No | No | No | No | No | No |
| 3.4. Reporting of numbers of patients crossing over to APM | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 3.5. Reporting of reasons for patients crossing over to APM | No | No | No | No | No | No | No |
| 3.7. Description of surgeons’ competence | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 3.8. Description of the system features, for example, resources, clinical pathways | No | No | No | No | No | No | No |
| 3.9. Inferences on effectiveness estimates based on actualized experimental treatments | No | No | No | No | No | No | No |
| 3.10. Double blinding successful | Not indicated | Not indicated | Yes | Not indicated | Not indicated | Not indicated | Partially d |
| 4. Outcome variables (Yes/No/Unclear) |
| 4.1. All effectiveness outcomes and adverse effects reported | No | Yes | Yes | No | No | Yes | Yes |
| 4.2. Follow-up percentage high (dropouts <10%) | Yes | Yes | Yes | Yes | No | Yes | Yes |

(Continues)
Gauffin et al. described characteristics of the healthcare system within the catchment area. Four trials reported appropriately how the randomization and concealment of allocation had been performed. All trials reported clinically important baseline characteristics of patients, but none reported comprehensively data on general health status, comorbid conditions, health behavior, environmental factors, and socioeconomic status.

All 7 trials reported adherence to the surgical intervention, but only 4 reported adherence to the nonsurgical therapy, and none reported co-interventions during follow-up (Table 2). All studies reported the proportion of patients crossing over to surgical treatment, but none of the studies reported reasons for the crossover. In 5 trials, the proportions of patients crossing over to surgery were highest and varied between 19 and 36 percent. No trial reported prespecified criteria for crossing over to surgical treatment. All trials reported that APM was performed by experienced knee surgeons, but only Gauffin et al. reported that all physiotherapists were experienced in knee rehabilitation. No trial reported system features during follow-up (eg, resources, clinical pathways). None of the studies based their primary conclusions on the actualized intervention contrast between the treatment arms. Double blinding was indicated in the trial by Sihvonen et al. and by Roos et al. for the assessment of biological effectiveness of APM, while other trials assessed effectiveness in real-world healthcare circumstances, where the patients and practitioners are not blinded.

All effectiveness outcomes were reported in all the trials (Table 2). One trial had more than 10% loss to follow-up. Only one trial used an outcome validated for patients with a meniscal rupture, the WOMET scale. Other trials used outcomes primarily intended for patients with knee osteoarthritis or trauma. The trials reported outcomes according to the protocols published a priori at the ClinicalTrials.gov or according to the methods sections of their papers. All trials used similar follow-up time-points between groups, and appropriate statistical analyses.

The results using the Cochrane’s assessment of validity of the 7 RCTs are shown in Table 3. Four trials showed low risk of bias in their ability to produce appropriately both randomization sequence and conceal allocation of patients, while in 3 other trials, fulfillment of one or both of these items remained unclear. The double-blinded trial by Sihvonen et al. had low risk of bias in blinding of patients, healthcare providers, and outcomes, while the other trials showed unclear or high risk of bias. Six RCTs had low risk of bias due to incomplete outcome data and one trial a high degree of bias. None of the trials showed indications of selective reporting of outcomes.

The generalizability of results based on CONSORT criteria is shown in Table 4. All 7 trials had reported appropriately all items required for generalizability of results according to the CONSORT criteria, except that only the trial by Sihvonen...
### Table 3: Assessing risk of bias according to the Cochrane method in the 7 randomized controlled trials assessing effectiveness of arthroscopic partial meniscectomy (+ = low risk of bias; = high risk of bias; ? = unclear risk of bias)

| Domain                              | Herrlin et al (2007) and 2013 Sweden | Katz et al (2013) USA | Sihvonen et al (2013) Finland | Yim et al (2013) South Korea | Gauffin et al (2014) Sweden | Kise et al (2016) Norway | Roos et al (2018) Denmark |
|-------------------------------------|-------------------------------------|------------------------|-------------------------------|-----------------------------|-----------------------------|--------------------------|---------------------------|
| Random sequence generation          | ?                                   | +                      | +                             | ?                           | ?                           | +                        | +                         |
| Allocation concealment              | ?                                   | +                      | +                             | ?                           | +                           | +                        | +                         |
| Blinding of participantsa           | −                                   | −                      | +                             | −                           | −                           | −                        | ?b                        |
| Blinding of personnela              | −                                   | −                      | +                             | −                           | −                           | −                        | ?b                        |
| Blinding of outcome assessmenta     | −                                   | −                      | +                             | −                           | −                           | −                        | ?b                        |
| Incomplete outcome dataa            | +                                   | +                      | +                             | +                           | −                           | +                        | +                         |
| Selective reporting                 | +                                   | +                      | +                             | +                           | +                           | +                        | +                         |

- Assessments should be made for each main outcome (or class of outcomes).
- Blinding failed for 36% of the patients.

### Table 4: Assessing generalizability of findings of randomized controlled trials according to the CONSORT statement in the 7 randomized controlled trials assessing effectiveness of arthroscopic partial meniscectomy (+ = allows generalizability; = does not allow generalizability, or intervention is not beneficial; ? = unclear)

| Domain                              | Herrlin et al (2007) and 2013 Sweden | Katz et al (2013) USA | Sihvonen et al (2013) Finland | Yim et al (2013) South Korea | Gauffin et al (2014) Sweden | Kise et al (2016) Norway | Roos et al (2018) Denmark |
|-------------------------------------|-------------------------------------|------------------------|-------------------------------|-----------------------------|-----------------------------|--------------------------|---------------------------|
| Eligibility criteria of patients    | +                                   | +                      | +                             | +                           | +                           | +                        | +                         |
| Setting and location                | +                                   | +                      | +                             | +                           | +                           | +                        | +                         |
| Interventions and how they were administered | +                                   | +                      | +                             | +                           | +                           | +                        | +                         |
| Definition of outcomes              | ±a                                  | ±a                     | +                             | ±a                          | ±a                          | ±a                       | ±a                        |
| Period of recruitment and follow-up | +                                   | +                      | +                             | +                           | +                           | +                        | +                         |

- Use of knee osteoarthritis or knee trauma measure not specific for meniscal lesions.
et al\(^{21}\) had used a meniscus-specific outcome measure (the WOMET tool).

Table 5 shows the conclusions made in each RCT, and the deficiencies of validity and generalizability of results indicated by the benchmarking method. The validity and generalizability concerns were related to the lack of reporting of behavioral, environmental, and socioeconomic factors. All trials lacked description of co-interventions, and some did not report adherence to the nonoperative treatment. Only one study reported on the relevant healthcare system features.

| Study Details | Conclusions | Deficiencies in Validity when Assessed by the Benchmarking Method | Deficiencies in Generalizability when Assessed by the Benchmarking Method |
|---------------|-------------|---------------------------------------------------------------|---------------------------------------------------------------------|
| Herrlin et al (2007\(^{19}\) and 2013) Sweden | Arthroscopic partial medial meniscectomy followed by supervised exercise was not superior to supervised exercise alone in terms of reduced knee pain, improved knee function and improved quality of life | Some uncertainty in baseline comparability | Selection of patients unclear |
| | | Lack of description of adherence to the nonsurgical intervention | | |
| | | Lack of description of co-interventions | | |
| | | Outcome measure not specific for meniscal lesions | | |
| Katz et al (2013)\(^{24}\) USA | In the intention-to-treat analysis, no significant differences between the study groups in functional improvement 6 mo after randomization. However, 30% of the patients who were assigned to physical therapy alone underwent surgery within 6 mo | Some uncertainty in baseline comparability | Selection of patients unclear |
| | | Lack of description of co-interventions | | |
| | | Outcome measure not specific for meniscal lesions | | |
| Sihvonen et al (2013)\(^{21}\) Finland | Arthroscopic partial medial meniscectomy provides no significant benefit over sham surgery in patients with a degenerative meniscal tear and no knee osteoarthritis | Some uncertainty in baseline comparability | Selection of patients unclear |
| | | Lack of description of co-interventions | | |
| Yim et al (2013)\(^{22}\) South Korea | No significant differences between arthroscopic meniscectomy and nonoperative management with strengthening exercises in terms of relief in knee pain, improved knee function, or increased satisfaction in patients after 2 y of follow-up | Some uncertainty in baseline comparability | Selection of patients unclear |
| | | Lack of description of co-interventions | | |
| | | Outcome measure not specific for meniscal lesions | | |
| Gauffin et al (2014)\(^{23}\) Sweden | Middle-aged patients with meniscal symptoms may benefit from arthroscopic surgery in addition to a structured exercise program | Some uncertainty in baseline comparability | Incomplete baseline reporting |
| | | Lack of description of adherence to the nonsurgical intervention | | |
| | | Lack of description of co-interventions | | |
| | | Outcome measure not specific for meniscal lesions | | |
| Kise et al (2016)\(^{20}\) Norway | The difference in treatment effect was minute after 2 y of follow-up, and the trial’s inferential uncertainty was sufficiently small to exclude clinically relevant differences | Some uncertainty in baseline comparability | Selection of patients unclear |
| | | Lack of description of adherence to the nonsurgical intervention | | |
| | | Lack of description of co-interventions | | |
| | | Outcome measure not specific for meniscal lesions | | |
| Roos et al (2018)\(^{17}\) Denmark | Arthroscopic partial meniscectomy showed greater improvement compared with skin incisions only at 2 y, with the statistical uncertainty of the between-group difference including what could be considered clinically relevant | Some uncertainty in baseline comparability | Selection of patients unclear |
| | | Lack of description of co-interventions | | |
| | | Outcome measure not specific for meniscal lesions | | |
This study shows that there is a lack of important information in the 7 RCTs published on effectiveness arthroscopic partial meniscectomy hampering evaluation of generalizability of the results of individual trials. Furthermore, the RCTs have had different study questions leading to clinical heterogeneity between the RCTs, and consequently, a meta-analysis is not justified to assess the overall effectiveness and to produce general recommendations.

When assessed using the benchmarking method, only one of the 7 RCTs provided information on the selection of patients, and this was also the only trial which provided information on the healthcare system characteristics relevant for patient selection. In all the 7 trials, clinical data relevant to the disorder were reported, but other potentially important predictors of outcomes, such as comorbid conditions, behavioral factors (eg, degree of physical activity, smoking), environmental factors (eg, work conditions), and patients’ socioeconomic status, were reported poorly.

The BM shows that 5 trials had assessed effectiveness of APM in real healthcare circumstances. Of these trials, only one had recruited from the whole catchment area and thus the patients represented the population; this trial showed clinically important effectiveness for APM. Also, one double-blinded trial with selected patient population showed effectiveness of APM. Both these trials showed effectiveness for surgery in the intention-to-treat analysis despite considerable crossover from conservative treatment group to surgical intervention. Altogether, 5 of the 7 trials showed major crossover from conservative treatment arm to surgery. The effectiveness estimates from these trials must be interpreted as effectiveness when one-fifth to one-third of patients who have not responded to the conservative treatment and have undergone the APM are compared with a group where majority or all patients have undergone the APM. The interpretation that the comparisons in these trials have between surgery and no-surgery is not tenable.

This study utilizes methodology of observational effectiveness studies, the benchmarking controlled trials, as well as earlier studies and recommendations on how to describe the PICO in randomized controlled trials, and how to assess validity and generalizability of findings; the new method is named as benchmarking method (BM). Cochrane’s method for the assessment of validity of RCTs, and the CONSORT criteria for reporting RCTs to allow generalizability of results were used as a reference standard. As the category of “Other biases” is often not used in Cochrane reviews, and as it is not well defined in the Cochrane handbook, it was not included.

The BM proposes, as illustrated in Figure 1, the 3 main validity issues in RCTs (domain, cause, and effect), the means for reaching each validity issue, and the interpretations of the main validity issues.

The Cochrane method and the BM showed similar assessments in the ability of the 7 trials to produce appropriately the randomization sequence and conceal allocation of patients, although the BM assessed the 2 validity items together. The main differences in the assessment of validity between the method by the Cochrane method and the BM were in blinding, baseline comparability, intention-to-treat analysis and in validity of the outcome parameters. The BM considers baseline comparability as the aim of the randomization and concealment of allocation, and the latter as the means for reaching comparability. Furthermore, blinding is considered valid only when study question is on effectiveness of an intervention per se, but blinding is contraindicated (the 5 other APM trials) when the study question is on effectiveness in the nonblinded circumstances of ordinary health care, where patients and caregivers are not blinded. In addition, the BM proposes that intention-to-treat analysis must be interpreted according to how the experiment actually happened, that is, what was the adherence to the intervention in each of the treatment arms taking into consideration also the occurrence of co-interventions. Finally, the BM includes validity of outcome parameters as an essential validity item.

The benchmarking method proposes that baseline comparability is the aim of randomization and is the first main validity criterion (Figure 1). Randomization and concealment of allocation are means for this purpose. To truly verify comparability, a comprehensive list of characteristics of patient groups should be documented. Comprehensive data are needed also to enable the assessment of generalizability of findings. Due to deficiencies in reporting of patients’ comorbid conditions, health behavior, environment, and socioeconomic status, the baseline comparability remained to some extent unclear in all 7 trials.

Blinding (patient, provider, and outcome) is an important means for validity when assessing effectiveness of the intervention per se, here in the studies by Sihvonen et al and Roos et al. When assessing effectiveness of interventions in routine health care, the aim is to evaluate the combined effect of the intervention per se, and the concomitant placebo effect plus the unspecified treatment effect related to the interaction between patient and care providers, for example, information, advice, and encouragement. Even if no specific effective intervention is available, the empathetic care from a competent healthcare professional may carry both placebo effect and nonspecific effects for the patient. In routine healthcare circumstances, both care providers and patients are aware of the given treatments, and to answer this study question, blinding should not be used. Consequently, when the study question is on real-world effectiveness, as in the 5 APM trials here, blinding should not be considered a validity criterion in the assessment of RCTs.

A valid measurement of the causal factor is essential for avoiding risk of bias. The causal factor in RCTs is primarily
the actualized contrast (determined by adherence to interventions and crossover between treatment arms) between the experimental interventions, and secondarily also in the (non-intended) co-interventions. In addition, a major causal factor is the competence of the staff, which should be reported. As intervention contrast is the causal factor, it is the second main validity criterion (Figure 1). Intervention contrast encompasses all the differences between the study groups in terms of interventions under study, for example, percentages of patients who actually obtained surgery vs percentages of patients who actually obtained physiotherapy; and also percentages of patients who obtained additional medical or rehabilitation interventions (co-interventions) in the 2 treatment arms. When based on the actualized intervention contrast, the analysis can be made only by intention-to-treat. It is not necessary at all to do a secondary on-treatment-analysis, but if secondary analyses are carried out, the best available statistical methodology is suggested.25

Co-interventions reflect behavior of patients and service providers, and may modify the apparent intervention effect. This modification should be taken into account as a potential additional causal factor for the treatment effects. The outcome estimates are valid for the degree of actualized interventions and co-interventions in the index and control treatment arms.

The actualized intervention contrast, that is, the percentage of patients operated before the primary outcome assessment time-point, was not explicitly considered as the causal factor for the effects in any of the trials, although mentioned in the

---

**FIGURE 1** The 3 main issues (domain, cause, effect) for pursuing high validity (low risk of bias) in randomized controlled trials (RCTs), the means for reaching low risk of bias, and interpretation of the main validity issues.
conclusions of the trial by Katz et al.24 Rather, the intention-to-treat analyses, and the intended intervention contrast in the protocol, were inferred as the main effectiveness findings of the trial, even when the actual experiment resulted in marked crossover between groups, particularly crossover to surgery from the nonoperative treatment arm. Consequently, no study made explicitly conclusions based on the actualized experiment. Co-interventions in the treatment arms were not reported in any of the trials; however, it may not be probable that co-interventions would have had a major effect on outcomes in these trials. None of the studies described reasons for crossing over to the other intervention arm.

Intention-to-treat analysis follows naturally when conclusions are based on what actually happened in the experiment, and not on full adherence, which is usually written in the protocol. How the interventions did actualize; that is, adherence and degree of crossover between intervention groups may reflect in part the effectiveness and safety of the intervention, and in part the behavior of service providers and patients. Not only numbers, but also the reasons for crossing over from one treatment arm to another should be reported. This information is important, because besides definite medical reasons, also time- and place-dependent reasons, for example, treatment cultures, may have an effect on crossover.

Comprehensive reporting of outcome variables on effectiveness and safety of the index intervention is needed to avoid biased interpretations of treatment benefits. Outcome represents the effect for the cause (contrast between all interventions in the index and control groups) and is the third main validity criterion for RCTs (Figure 1). The challenge is to report comprehensively all primary and secondary outcome variables using validated outcome measures, and consider the influence of loss to follow-up for the credibility of the effect estimates. Loss to follow-up may lead to differences in characteristics of the patient populations across treatment arms and pose thus a major risk for bias in effect estimates. Timing of the outcome assessment is a matter of validity if there are between-group differences and weakens applicability of the results if there is no definite outcome time-point for the estimates.

Selective reporting of outcomes increases risk for biased outcomes and is included in the Cochrane risk of bias assessment. However, selective reporting is a matter of human risk of bias, and there are also other human risk of biases, for example, due to financial or intellectual conflict of interests.26,27 The CONSORT criteria were fulfilled in all the 6 RCTs, except for the validated outcome measure. However, according to the BM, generalizability of the findings was hampered by several factors. Only one trial had recruited almost all patients from the catchment area,23 and thus, these findings can be considered generalizable also elsewhere with similar clinical processes and patient characteristics. Other studies did not report selection of patients, and all studies lacked important baseline data which may modify the treatment effect. Only one trial23 described characteristics of the healthcare system within the catchment area.

The selection of patients and data on healthcare system features is needed for assessing generalizability of trial findings. Full and valid description of patient populations at baseline ensures the assessment of comparability between the treatment arms as well as generalizability of the trial findings. Valid data of the actualized intervention contrast (causal factor for effects) and co-interventions during course of the trial (potential additional causal factors for effects) are necessary for valid conclusions. Valid and comprehensive data of outcomes (effect factors), low loss to follow-up, similar follow-up time-points, and appropriate statistical analysis are needed to form a reliable and inclusive view of the effects.

In conclusion, there were many similarities between the Cochrane method and the benchmarking method in assessing validity and generalizability of results of the 7 RCTs. However, the BM proposes that when assessing effectiveness in routine healthcare circumstances blinding is contraindicated, as patients and caregivers are not blinded in ordinary healthcare. When the aim is to assess the pure biological effect, blinding of patients and healthcare personnel is the only alternative. However, double-blinded study designs produce effectiveness estimates that are abstract knowledge in that sense that pure biological effectiveness does not exist in ordinary healthcare circumstances, where patients and healthcare personnel are always aware of the treatments provided. The results of the intention-to-treat analysis should be interpreted based on how the interventions have been actualized, not on how the interventions were intended to occur in the study protocol. The validity of the outcome measures should be included as an essential appraisal item. In contrast to the results when applying the CONSORT statement, the BM indicated that all the 7 trials had deficiencies in reporting of patient selection, baseline characteristics, and actualization of interventions; and that the characteristics needed for the assessment of generalizability of findings should be described more comprehensively than hitherto.

To sum up, when assessed with the BM, the randomized controlled trials on arthroscopic partial meniscectomy of the knee show deficiencies hampering the assessment of validity and generalizability of findings, calling for more RCTs. The BM proposes that blinding is a validity criterion only when the study question is on intervention effect per (biological effect); interpretation of the intention-to-treat analysis should be based on actualized interventions; and validity of the outcome parameters should be documented as an essential validity item. A detailed description of patient selection, study setting, patients, interventions, and outcome assessments is necessary for the assessment of generalizability of findings of RCTs.

The benchmarking method indicates that APM may be an effective treatment for meticulously selected patients, and
for those patients (around 20%-33% with the indication in the RCTs published) who do not respond to the conservative treatment.

5 | PERSPECTIVE

The benchmarking method shows that 5 trials had assessed effectiveness of APM in real healthcare circumstances. Of these trials, only one had recruited from the whole catchment area and thus the patients represented the population; this trial showed clinically important effectiveness for APM. Also, one small trial with selected patient population showed effectiveness of APM. Both these 2 trials showed effectiveness for surgery despite considerable crossover from conservative treatment group to surgical intervention. Altogether, 5 of the 7 trials showed considerable crossover from conservative treatment arm to surgery. These effectiveness estimates must be interpreted as one-fifth to one-third of patients operated vs majority of patients operated. One of the sham-controlled trials found no biological effectiveness for APM, but this trial had a highly selected patient population. All trials lacked a comprehensive description of patient populations. The benchmarking method indicates that APM may be an effective treatment for meticulously selected patients, and for those patients who do not respond to the conservative treatment. These findings call for more sham-controlled and real-world effectiveness trials with comprehensive description of study features. The BM can be used for planning clinical trials and for assessing validity and generalizability of evidence on effectiveness of interventions within sports medicine and other fields.

CONFLICTS OF INTEREST

The author declares no support from any organization for the submitted work; no financial relationships with any organization that might have an interest in the submitted work; and no other relationships or activities that could appear to have influenced the submitted work.

AUTHORS’ CONTRIBUTION

The author has developed the idea for the paper and written the manuscript solely. Seija Puro is acknowledged for the graphical presentation of the figure; and Ritva Miikki for the literature search.

ORCID

A. Malmivaara http://orcid.org/0000-0002-9042-4450

REFERENCES

1. Monk P, Garfjeld Roberts P, Palmer AJ, et al. The urgent need for evidence in arthroscopic meniscal surgery: a systematic review of the evidence for operative management of meniscal tears. Am J Sports Med. 2016;44:965-973.
2. Swart NM, van Oudenaarde K, Reijnierse M, et al. Effectiveness of exercise therapy for meniscal lesions in adults: a systematic review and meta-analysis. J Sci Med Sport. 2016;19:990-998.
3. vande Graaf VA, Wolterbeek N, Mutsaerts EL, et al. Arthroscopic partial meniscectomy or conservative treatment for nonobstructive meniscal tears: a systematic review and meta-analysis of randomized controlled trials. Arthroscopy. 2016;32:1855-1865. e4.
4. Järvinen TL, Guyatt GH. Arthroscopic surgery for knee pain. BMJ. 2016;354:i3934.
5. Lamplot JD, Brophy RH. The role for arthroscopic partial meniscectomy in knees with degenerative changes: a systematic review. Bone Joint J. 2016;98-B:934-938.
6. Thorlund JB, Juhl CB, Roos EM, Lohmander LS. Arthroscopic surgery for degenerative knee: systematic review and meta-analysis of benefits and harms. Br J Sports Med. 2015;49:1229-1235.
7. Khan M, Evaniew N, Bedi A, Ayeni OR, Bhandari M. Arthroscopic surgery for degenerative tears of the meniscus: a systematic review and meta-analysis. CMAJ. 2014;186:1057-1064.
8. Katz JN, Losina E. Surgery versus physical therapy for meniscal tear and osteoarthritis. N Engl J Med. 2013;369:677-678.
9. Beaufils P, Becker R, Kopf S, et al. Surgical management of degenerative meniscus lesions: the 2016 ESSKA meniscus consensus. Knee Surg Sports Traumatol Arthrosc. 2017;25:335-346.
10. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med. 2009;151:W65-W94.
11. Malmivaara A, Koes BW, Bouter LM, van Tulder MW. Applicability and clinical relevance of results in randomized controlled trials: the Cochrane review on exercise therapy for low back pain as an example. Spine (Phila Pa 1976). 2006;31:1405-1409.
12. Furlan AD, Malmivaara A, Chou R, et al. 2015 Updated method guideline for systematic reviews in the Cochrane Back and Neck Group. Spine (Phila Pa 1976). 2015;40:1660-1673.
13. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869.
14. Malmivaara A. Benchmarking Controlled Trial-a novel concept covering all observational effectiveness studies. Ann Med. 2015;47:332-340.
15. Malmivaara A. System impact research - increasing public health and health care system performance. Ann Med. 2016;48:211-215.
16. Malmivaara A. Assessing validity of observational intervention studies - the Benchmarking Controlled Trials. Ann Med. 2016;48:440-443.
17. Roos EM, Hare KB, Nielsen SM, Christensen R, Lohmander LS. Better outcome from arthroscopic partial meniscectomy than skin
incisions only? A sham-controlled randomised trial in patients aged 35-55 years with knee pain and an MRI-verified meniscal tear. *BMJ Open*. 2018;8:e019461.

18. Higgins J, Altman D, Sterne J. Chapter 8: assessing risk of bias in included studies. In: Higgins JPT, Churchill R, Chandler J, Cumpston MS, eds. *Cochrane Handbook for Systematic Reviews of Interventions version 5.2.0.* (updated June 2017), Cochrane 2017 www.training.cochrane.org/handbook

19. Herrlin S, Hallander M, Wange P, Weidenhielm L, Werner S. Arthroscopic or conservative treatment of degenerative medial meniscal tears: a prospective randomised trial. *Knee Surg Sports Traumatol Arthrosc.* 2007;15:393-401.

20. Kise NJ, Risberg MA, Stensrud S, Ranstam J, Engebretsen L, Roos EM. Exercise therapy versus arthroscopic partial meniscectomy for degenerative meniscal tear in middle aged patients: randomised controlled trial with two year follow-up. *BMJ*. 2016;354:i3740.

21. Sihvonen R, Paavola M, Malmivaara A, et al. Arthroscopic partial meniscectomy versus sham surgery for a degenerative meniscal tear. *N Engl J Med*. 2013;369:2515-2524.

22. Yim JH, Seon JK, Song EK, et al. A comparative study of meniscectomy and nonoperative treatment for degenerative horizontal tears of the medial meniscus. *Am J Sports Med*. 2013;41:1565-1570.

23. Gauffin H, Tagesson S, Meunier A, Magnusson H, Kvist J. Knee arthroscopic surgery is beneficial to middle-aged patients with meniscal symptoms: a prospective, randomised, single-blinded study. *Osteoarthritis Cartilage*. 2014;22:1808-1816.

24. Katz JN, Brophy RH, Chaisson CE, et al. Surgery versus physical therapy for a meniscal tear and osteoarthritis. *N Engl J Med*. 2013;368:1675-1684.

25. Sussman JB, Hayward RA. An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials. *BMJ*. 2010;340:c2073.

26. Page MJ, McKenzie JE, Kirkham J, et al. Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of healthcare interventions. *Cochrane Database Syst Rev*. 2014:MR000035.

27. Lieb K, Osten-Sacken J, Stoffers-Winterling J, Reiss N, Barth J. Conflicts of interest and spin in reviews of psychological therapies: a systematic review. *BMJ Open*. 2016;6:e016066.

**How to cite this article:** Malmivaara A. Validity and generalizability of findings of randomized controlled trials on arthroscopic partial meniscectomy of the knee. *Scand J Med Sci Sports*. 2018;28:1970–1981. [https://doi.org/10.1111/sms.13215](https://doi.org/10.1111/sms.13215)

---

**APPENDIX 1**

**Literature search strategy**

**Searched databases**

CENTRAL - Cochrane Library/Wiley Online Library
Web of Science - Core collection (Indexes=SCI-EXPANDED)
Ovid MEDLINE(R) - Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) <1946 to Present>

**Search strategies**

**CENTRAL - Cochrane Library/Wiley Online Library**

| ID  | Search                                                                 | Hits   |
|-----|------------------------------------------------------------------------|--------|
| #1  | MeSH descriptor: [Meniscus] explode all trees with qualifier(s): [Surgery - SU] | 145    |
| #2  | MeSH descriptor: [Arthroscopy] explode all trees                        | 1454   |
| #3  | MeSH descriptor: [Knee] explode all trees                               | 706    |

**Web of science - core collection**

| ID  | Search                                                                 | Hits   |
|-----|------------------------------------------------------------------------|--------|
| #4  | #2 and #3                                                              | 112    |
| #5  | meniscal tear:ti,ab,kw (Word variations have been searched)            | 138    |
| #6  | arthroscopic meniscectomy:ti,ab,kw (Word variations have been searched) | 167    |
| #7  | arthroscopic debridement:ti,ab,kw (Word variations have been searched) | 104    |
| #8  | #1 or #4 or #5 or #6 or #7 in Trials                                     | 498    |

1. TS=(menisc* AND (surgery OR surgical* OR operative* OR arthroscopy* OR injury OR injuries OR tear* OR torn)) OR TS=(knee* NEAR/3 arthroscopy*) OR TS=(‘Arthroscopic debridement’*) OR TS=(arthroscopy* NEAR/3 meniscectom*)
2. TS=(random* OR rct OR crt OR “controlled trial”* OR “control group”* OR double-blind* OR “double blind”* OR single-blind* OR “single blind”* OR cross-over OR crossover OR “cross over” OR placebo)

Results: 1649

#2 AND #1
Refined by: DOCUMENT TYPES: (ARTICLE OR CORRECTION OR REVIEW OR RETRACTED PUBLICATION OR PROCEEDINGS PAPER)
Results: 1606
Indexes = SCI-EXPANDED Timespan = All years (from Web of Science Core Collection)

Database: Ovid MEDLINE(R) Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) <1946 to Present>

Search Strategy: 20/10/2017

1. exp Meniscus/su [Surgery] (3976)
2. exp Arthroscopy/(22577)
3. exp Knee/(13573)
4. 2 and 3 (450)
5. (menisc* adj5 (surgery or surgical* or operati* or arthroscop* or injury or injuries or tear* or torn)).ti,ab. (7325)
6. (knee* adj3 arthroscop*).ti,ab. (3229)
7. Arthroscopic debridement*.ti,ab. (840)
8. (arthroscop* adj3 meniscectom*).ti,ab. (780)
9. 1 or 4 or 5 or 6 or 7 or 8 (12430)
10. Randomized Controlled Trials as Topic/(121907)
11. randomized controlled trial/(497319)
12. Random Allocation/(99691)
13. Double Blind Method/(157599)
14. Single Blind Method/(26594)
15. clinical trial/(548114)
16. clinical trial, phase i.pt. (20030)
17. clinical trial, phase ii.pt. (32310)
18. clinical trial, phase iii.pt. (15306)
19. clinical trial, phase iv.pt. (1629)
20. controlled clinical trial.pt. (99268)
21. randomized controlled trial.pt. (497319)
22. multicenter study.pt. (249307)
23. clinical trial.pt. (548114)
24. exp Clinical Trials as topic/(32342)
25. or/10-24 (1316051)
26. (clinical adj trial$).tw. (321698)
27. ((singl$ or doubl$ or treb$ or tripl$) adj (blind$3 or mask$3)).tw. (167825)
28. PLACEBOS/(36433)
29. placebo$tw. (209223)
30. randomly allocated.tw. (25083)
31. (allocated adj2 random$).tw. (28243)
32. or/26-31 (583883)
33. 25 or 32 (1542791)
34. case report.tw. (275614)
35. letter/(1023624)
36. historical article/(356118)
37. or/34-36 (1640796)
38. 33 not 37 (1507650)
39. 9 and 38 (1437)

The Number of References by Database

| Database | References |
|----------|------------|
| CENTRAL | 498 |
| Web of Science | 1606 |
| Ovid MEDLINE(R) Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) <1946 to Present> | 1437 |
| Total | 3541 |
| References, duplicates excluded | 2375 |