EVIDENCE FOR DECISION
MAKING IN HEALTH CARE
BASEL INSTITUTE FOR CLINICAL EPIDEMIOLOGY AND BIOSTATISTICS

# Health technology assessment on spinal implants

Produced by:

Hannah Ewald[1], Dominik Glinz[1], Viktoria Gloy[1], Joris van Stiphout[2], Heike Raatz[1]

[1] Basel Institute for Clinical Epidemiology and Biostatistics (CEB), University Hospital Basel

[2] Institut für Pharmazeutische Medizin, University Hospital Basel

Date of submission

15 August 2016

Version

3.0.0

CEB head of institute

Prof. Dr. med  Heiner C. Bucher, MPH

ECPM head of research

PD Dr. Matthias Schwenkglenks, MPH

## Acknowledgement

The assessment team thanks Prof. M. Aebi for his support as clinical expert. The assessment team is responsible for the contents of this report.

## Contributions

**Scoping:** HR, HE and JVS; **Systematic review:** HE and HR wrote the protocol; HE, VG and HR developed and conducted the literature search; HE, JVS, DG and HR screened the literature; HE, JVS and DG were responsible for data extraction; DG, VG and HR graded the risk of bias and the quality of evidence; HE, DG and VG conducted data analysis; HE, DG ,VG and HR are responsible for data interpretation; HE, DG, VG, JVS and HR wrote this report.

## Table of contents

Acknowledgement 2

Contributions 2

Abbreviations 5

Executive summary 6

1. Medical background 9

2. Aim 9

3. Methods 10

3.1. Overview of the eligibility criteria 10

3.2. Eligibility criteria for systematic reviews on interspinous and pedicle-based stabilisation – PICO 1-3 11

3.2.1 Population – PICO 1 and 2 11

3.2.2 Population – PICO 3 12

3.2.3 Interventions – PICO 1-3 12

3.2.4 Comparators – PICO 1-3 12

3.2.5 Outcomes – PICO 1-3 12

3.2.6 Study design – PICO 1-3 13

3.2.7 Languages – PICO 1-3 13

3.3 Eligibility criteria for systematic reviews on disc prostheses – PICO 4 and 5 13

3.3.1 Populations – PICO 4 and 5 13

3.3.2 Intervention – PICO 4 and 5 14

3.3.3 Comparator – PICO 4 and 5 14

3.3.4 Outcomes – PICO 4 and 5 14

3.3.5 Study design – PICO 4 and 5 14

3.3.6 Languages – PICO 4 and 5 14

3.4 Literature search 14

3.5 Data extraction 15

3.6 Risk of bias and quality of evidence assessment 15

3.7 Data synthesis 16

3.7.1 Subgroup analyses 17

3.7.2 Sensitivity analyses 18

4 Results 19

4.1 Literature search 19

4.2 Interspinous and pedicle-based devices – PICO 1-3 19

4.2.1 Results – PICO 1 ... 19
4.2.2 Results – PICO 2 ... 33
4.2.3 Results – PICO 3 ... 42
4.3 Disc prostheses – PICO 4-5 ... 55
4.3.1 Results – PICO 4 ... 55
4.3.2 Results – PICO 5 ... 70
5 Ongoing studies ... 93
6 Summary of results ... 94
6.1 Interspinous and pedicle-based devices ... 94
6.1.1 Summary of results – PICO 1 ... 94
6.1.2 Summary of results – PICO 2 ... 94
6.1.3 Summary of results – PICO 3 ... 94
6.2 Disc prostheses ... 95
6.2.1 Summary of results – PICO 4 ... 95
6.2.2 Summary of results – PICO 5 ... 95
7 Discussion ... 96
7.1 General methodological issues ... 96
7.2 Interspinous and pedicle-based devices - PICO 1-3 ... 99
7.3 Disc prostheses - PICO 4-5 ... 100
8 Conclusions ... 104
Reference list ... 105
Appendices ... 108
I. Appendix – Search strategy for Pubmed/Medline and ClinicalTrials.gov ... 108
II. Appendix – PICO 1 short-term ... 110
III. Appendix – Risk of Bias with support for judgement ... 112
IV. Appendix – PICO 2 short-term ... 131
V. Appendix – PICO 3 short-term ... 131
VI. Appendix – PICO 4 short-term ... 133
VII. Appendix – PICO 5 short-term ... 134

## Abbreviations

| | |
|---|---|
| AE | Adverse events |
| ACDF | Anterior cervical discectomy and fusion |
| ASD | Adjacent segment disease |
| CI | Confidence Interval |
| EQ-5D | EuroQol 5-dimensional questionnaire |
| FU | Follow-up |
| GRADE | Grading of Recommendations Assessment, Development and Evaluation |
| HTA | Health Technology Assessment |
| ID | Identification |
| MCID | Minimal clinically important difference |
| MD | Mean difference |
| MRDQ | Modified Roland Disability Questionnaire |
| n | Number (of) |
| NDI | Neck Disability Index |
| n.r. | Not reported |
| NRS | Numeric Rating Scale |
| ODI | Oswestry Disability Index |
| PICO | Population, Intervention, Comparator and Outcomes |
| QoL | Quality of life |
| RCT | Randomised controlled trial |
| RoB | Risk of bias |
| RR | Relative risk ratio |
| SAE | Serious adverse events |
| SD | Standard deviation |
| SF-12 | 12-Item Short Form Health Survey |
| SF-36 | 36-Item Short Form Health Survey |
| SMD | Standardised mean difference |
| SoF | Summary of Findings (GRADE output) |
| VAS | Visual Analogue Scale |
| vs. | versus |
| ZCQ | Zurich Claudication Questionnaire |

## Executive summary

### Background

Spinal surgery, including fusion and dynamic stabilisation, is one of the available treatment options for patients with symptomatic degenerative changes of the lumbar or cervical spine. The various symptoms can include local pain, radicular pain due to nerve root compression, as well as other neurological symptoms such as muscular weakness and numbness.

Dynamic stabilisation has been suggested as an alternative to fusion to obtain better treatment results due to a preservation of movement in the spine. Three main types of implants can be distinguished disc prosthesis, interspinous spacers and pedicle based dynamic stabilisation. These implants can be inserted using different surgical approaches and in combination with other surgical interventions like direct decompression.

### Aim

The aim of the current Health Technology Assessment (HTA) report is to systematically review the evidence on the clinical effectiveness and safety of disc prostheses in the cervical and lumbar spine and of interspinous or pedicle-based stabilisation in patients with degenerative changes of the lumbar spine either compared to direct decompression only or to fusion.

### Methods

This systematic review covers five different research questions with varying populations, interventions, comparisons and outcomes (PICO). The following table shows the populations, interventions and comparators which were assessed.

Overview of all the PICO-questions

| | **Populations** | | | **Interventions** | | | | **Comparators** | |
|---|---|---|---|---|---|---|---|---|---|
| | Patients with symptoms due to degenerative changes of the **lumbar spine** | | Patients with symptoms due to degenerative changes of the **cervical spine** | Interspinous or pedicle-based stabilisation | | | Disc prosthesis | Direct decompression only | Fusion with implants |
| | with neurological symptoms | With or without neurological symptoms | With or without neurological symptoms | without direct decompression | with direct decompression | with or without direct decompression | with or without direct decompression | | |
| **1** | X | - | - | X | - | - | - | X | - |
| **2** | X | - | - | - | X | - | - | X | - |
| **3** | - | X | - | - | - | X* | - | - | X* |
| **4** | - | X | - | - | - | - | X* | - | X* |
| **5** | - | - | X | - | - | - | X* | - | X* |

*if decompression was done, it had to be done in both treatment arms

A literature search was performed in PubMed. The literature was screened and extracted by two reviewers with the second reviewer checking the extraction of the first. The assessment of the risk of bias was performed according to the Cochrane Handbook for Interventions and the quality of the evidence was evaluated according to GRADE.

**Results**

For PICO 1 - comparing interspinous stabilisation without direct decompression to direct decompression - in a population with neurological symptoms due to degenerative changes of the lumbar spine, three randomised controlled studies were included for long-term follow-up. For interspinous stabilisation without direct decompression compared to direct decompression quality of life (EQ-5D MD 0.04, 95% CI 0.02 to 0.06; important outcome, low quality of the evidence) was statistically significantly higher but there was also a statistically significantly higher relative risk of reoperations (RR 3.02, 95% CI 1.75 to 5.22; important outcome, low quality of the evidence). There was no statistically significant effect on back pain, radicular pain, spinal claudication, function and adverse events. No study reported revision rates or serious adverse events. The overall quality of evidence for the outcomes of PICO 1 was judged to be very low.

For PICO 2 - comparing interspinous stabilisation with direct decompression to direct decompression - in a population with neurological symptoms due to degenerative changes of the lumbar spine, one randomised controlled study was included for long-term follow-up. There was no statistically significant effect for interspinous devices with direct decompression compared to direct decompression only (PICO 2) on back pain (VAS MD -0.80, 95% CI -2.31 to 0.71; critical outcome, very low quality of the evidence) and function (ODI MD -8.70, 95% CI -19.91 to 2.51; important outcome, low quality of the evidence). Zero events were reported for complications. No studies were available for radicular pain, spinal claudication, quality of life, revision rate, reoperation rate and serious adverse events. No study examined pedicle-based stabilisation. The overall quality of evidence for the outcomes of PICO 2 was judged to be very low.

For PICO 3 - comparing interspinous or pedicle-based stabilisation to fusion with implants - in a population with symptoms due to degenerative changes of the lumbar spine with or without neurological symptoms, two studies were included for long-term follow-up. Only the outcome function was reported by both studies. There was no statistically significant effect for interspinous or pedicle-based stabilisation with direct decompression compared to fusion (PICO 3) on function based on two studies (Davis 2013, Madan 2003). Only one study (Davis 2013) reported back pain, radicular pain, spinal claudication, and reoperation rate. For all these outcomes no statistically significant effect was found. Only one study (Madan 2003) reported revision rate and adverse events. For both outcomes the effects of pedicle-based stabilisation compared to fusion with implants were not statistically significant. The quality of the evidence was low or very low for all outcomes in PICO 3. No studies were available for quality of life and serious adverse events. The overall quality of evidence for the outcomes of PICO 3 was judged to be low.

For PICO 4 - comparing lumbar disc prosthesis compared to fusion - in a population with symptoms due to degenerative changes of the lumbar spine with or without neurological symptoms, six studies were eligible. At long-term follow-up, there were statistically significant effects of disc prosthesis compared to fusion for back pain (VAS MD -5.60, 95% CI -10.47 to -0.73; important outcome, low quality of the evidence), quality of life physical component score (SF-36 MD 2.77, 95% CI 0.85 to 4.70; important outcome, low quality of the evidence), and function (ODI MD -5.19, 95% CI -7.67 to -2.71; important outcome, moderate quality of the evidence). There were no statistically significant effects for radicular pain, mental component of quality of life, reoperation rate, and serious adverse events. The quality of evidence was low or very low for these outcomes. There was as well no statistically significant effect on adverse events with moderate quality of evidence. Only one study (Gornet 2011) reported on revision rate, but the number of events was zero in both groups. The overall quality of evidence for the outcomes of PICO 4 was judged to be low.

For PICO 5 - comparing cervical disc prosthesis compared to cervical fusion - in a population with symptoms due to degenerative changes of the cervical spine with or without neurological symptoms, fourteen studies were eligible. At long-term follow-up, there were statistically significant effects for cervical disc prosthesis compared to fusion for radicular pain (VAS MD -3.76, 95% CI -6.34 to -1.17; critical outcome, moderate quality of the evidence), neck pain (SF-36 MD -6.35, 95% CI -9.03 to -3.67; important outcome, low quality of the evidence), quality of life physical component score (SF-36 MD 1.95, 95% CI 0.81 to 3.10; important outcomes, moderate quality of the evidence) and mental component scores (SF-36 MD 1.78, 95% CI 0.57 to 2.99; important outcomes, moderate quality of the evidence), and function (NDI MD -3.50, 95% CI -5.77 to -1.23; important outcome, moderate quality of the evidence). For PICO 5, there were no statistically significant effects for revision rate, reoperation rate, adverse events, and serious adverse events. The quality of evidence was low or very low for these outcomes. No study reported myelopathy. The overall quality of evidence for the outcomes of PICO 5 was judged to be very low.

**Conclusion**

Though the overall quality of the evidence is similar for all PICO-questions, considerably more studies were identified for PICO 4 and PICO 5 than for PICO 1, 2 and 3. Major limitations of the quality of the evidence included risk of bias, unexplained heterogeneity (inconsistency) and imprecision.

The evaluation of the quality of the evidence should be re-considered in the context of decision making where values and preferences regarding aspects like the balance of benefit and harm, and costs can affect the appraisal of the available evidence and its quality.

# 1. Medical background

Spinal surgery, including fusion and dynamic stabilisation, is one of the available treatment options for patients with symptomatic degenerative changes of the lumbar or cervical spine. Degenerative changes of the spine can lead to impingement of the nerve roots or the spinal cord, causing neurological symptoms like radicular pain, paraesthesia or numbness, as well as muscular weakness or paresis. In these patients, it is important to relieve the pressure either by direct surgical decompression where organic material, like material of the prolapsed disc or bone, is removed, or by indirect surgical decompression using dynamic stabilisation[1] or fusion. In other patients, the primary symptoms can be local back or neck pain without neurological symptoms.

The aim of spinal fusion is to improve those symptoms by joining two or more vertebral bodies while dynamic stabilisation has been suggested as an alternative treatment to fusion to obtain better treatment results due to a preservation of movement in the spine. The implants used for dynamic stabilisation can be inserted using different approaches (e.g. anterior or posterior approach) and in combination with other surgical devices.

Spinal fusion is considered a risk factor for adjacent segment disease (ASD) through the increased biomechanical stress on the segments adjacent to the fused vertebral bodies[2]. ASD can manifest in various ways, for example, as instability, discus hernia, scoliosis, or vertebral compression fracture[2]. However, the association between ASD and fusion surgery remains controversial as the symptoms can also occur secondary to degenerative changes[2].

Dynamic stabilisation has been suggested as an alternative treatment for fusion with a lower risk of ASD. The following forms of dynamic stabilisation can be differentiated:

- Dynamic stabilisation of the anterior spine via:
  - Disc prostheses
- Dynamic stabilisation of the posterior spine via:
  - Pedicle-based stabilisation
  - Interspinous stabilisation

These technologies were the object of a previous health technology assessment by the Swiss Federal Office of Public Health and were granted provisional reimbursement pending further evaluation and studies.

According to the 2014 American Association of Neurological Surgeons guideline update and a survey on the management of spinal stenosis amongst Dutch surgeons in the same year, there is still considerable uncertainty regarding the indications for fusions and dynamic stabilisation devices[3-8].

This Health Technology Assessment (HTA) report re-evaluates the technologies defined above as treatment options for patients with degenerative changes of the cervical and lumbar spine.

# 2. Aim

The aim of the report is to systematically review the evidence on the clinical effectiveness and safety of disc prostheses and interspinous stabilisation or pedicle-based devices in patients with degenerative changes of the spine, compared to direct decompression, fusion or a combination of both (for details see section 3.1).

# 3. Methods

## 3.1. Overview of the eligibility criteria

An overview of the relevant inclusion criteria (**P**opulation, **I**ntervention, **C**omparator, **O**utcomes, short: PICO) is provided in Table 1, Table 2, and the subsequent sections. Interspinous and pedicle-based stabilisations are the technologies investigated in PICO 1-3, disc prostheses are covered in PICO 4 and 5. The populations of interest include patients with symptoms due to degenerative changes of the spine. For PICO 1-4, the degenerative changes are in the lumbar spine and for PICO 5 in the cervical spine. The relevant populations with degenerative changes included only patients with neurological symptoms for PICO 1-2 and included patients with back pain with or without neurological symptoms for PICO 3-5 (see Table 1 and Table 2).

Table 1: Inclusion criteria for PICO 1-3 on interspinous or pedicle-based stabilisation

| **Population for PICO 1 and 2** | | **Described in** |
|---|---|---|
| **Population** | Patients **with** neurological symptoms due to degenerative changes of the lumbar spine | section 3.2.1 |
| **Population for PICO 3** | | |
| **Population** | Patients with low back pain **with or without** neurological symptoms due to degenerative changes of the lumbar spine | chapter 3.2.1 |
| **Intervention and Comparator for PICO 1** | | |
| **Intervention** | Interspinous or pedicle-based stabilisation of 1-2 affected levels **without** direct decompression | section 3.2.3 |
| **Comparator** | Direct decompression only of 1-2 affected levels | section 3.2.4 |
| **Intervention and Comparator for PICO 2** | | |
| **Intervention** | Interspinous or pedicle-based stabilisation of 1-2 affected levels **with** direct decompression | section 3.2.3 |
| **Comparator** | Direct decompression only of 1-2 affected levels | section 3.2.4 |
| **Intervention and Comparator for PICO 3** | | |
| **Intervention** | Interspinous or pedicle-based stabilisation of 1-2 affected levels (**with or without** direct decompression)* | section 3.2.3 |
| **Comparator** | Fusion with an implant at 1-2 affected levels (**with or without** direct decompression)* | section 3.2.4 |
| **Other inclusion criteria for PICO 1-3** | | |
| **Outcomes** | Outcomes on morbidity and quality of life | section 3.2.5 |
| **Study design** | Randomised controlled trials or quasi-randomised trials | section 3.2.6 |

| | | |
|---|---|---|
| **Languages** | English, German, French | section 3.2.7 |
| * if decompression was performed, it had to be performed in both treatment arms | | |

Table 2: Inclusion criteria for PICO 4-5 on disc prostheses

| **Population, PICO 4** | | **Described in** |
|---|---|---|
| Population | Patients with low back pain **with or without** neurological symptoms due to degenerative changes of the **lumbar** spine | section 3.3.1 |
| **Population, PICO 5** | | |
| Population | Patients with neck pain **with or without** neurological symptoms due to degenerative changes of the **cervical** spine | section 3.3.1 |
| **Other inclusion criteria for PICO 4 and 5** | | |
| Intervention | Disc prosthesis of 1 affected level (**with or without** direct decompression)* | section 3.3.2 |
| Comparator | Fusion of 1 affected level (**with or without** direct decompression)* | section 3.3.3 |
| Outcomes | Outcomes on morbidity and quality of life | section 3.3.4 |
| Study design | Randomised controlled trials or quasi-randomised trials | section 3.3.5 |
| Languages | English, German, French | section 3.3.6 |
| * if decompression was performed, it had to be performed in both treatment arms | | |

## 3.2. Eligibility criteria for systematic reviews on interspinous and pedicle-based stabilisation – PICO 1-3

### 3.2.1 Population – PICO 1 and 2

Eligible patients had symptomatic degenerative changes of the lumbar spine with neurological symptoms requiring the treatment of 1-2 affected levels. Relevant symptoms were neurological symptoms like radiculopathy or neurogenic claudication. The degenerative changes could be associated with spondylolisthesis ≤ grade 1. It was likely that the patients already had conservative treatment attempts prior to surgery. Studies in patients with prior dynamic stabilisation or fusion at the same level of the spine were excluded. Studies with patients with other causes for their symptoms than degenerative disease were not eligible.

Studies were eligible if at least 80% of the study population fulfilled the inclusion criteria. Reporting of the symptoms in the publications was poor and explicit information on the presence of neurological symptoms was often missing therefore it was decided a posteriori that any patients with diagnosed spinal stenosis were likely to suffer from neurological symptoms and fulfil the inclusion criteria.

### 3.2.2 Population – PICO 3

Eligible patients had symptomatic degenerative changes of the lumbar spine requiring the treatment of 1-2 affected levels. Relevant symptoms included local back pain but also neurological symptoms like radiculopathy or neurogenic claudication. Any combination of symptoms was accepted. As for PICO 1 and 2 the degenerative changes could be associated with spondylolisthesis ≤ grade 1 and it was likely that the patients already had conservative treatment attempts prior to surgery. Studies in patients with prior dynamic stabilisation or fusion at the same level of the spine were excluded. Studies with patients with other causes for their symptoms than degenerative disease were not eligible.

Studies were eligible if at least 80% of the study population fulfilled the inclusion criteria.

### 3.2.3 Interventions – PICO 1-3

Three different interventions were investigated separately:

a) Dynamic stabilisation of 1-2 affected levels **without** direct decompression (PICO 1).

b) Dynamic stabilisation of 1-2 affected levels **with** direct decompression (PICO 2)

c) Dynamic stabilisation of 1-2 affected levels **with** or **without** direct decompression as co-intervention. The practice regarding the co-intervention (i.e. direct decompression) had to be identical to that in the comparator arm (PICO 3).

For PICO 1 interspinous stabilisation is expected to the most relevant intervention but other types of implants, like pedicle-based stabilisation, were considered as well. Hence for all three PICO-questions (PICO 1-3) both interspinous and pedicle-based stabilisation were considered.

### 3.2.4 Comparators – PICO 1-3

a) **Direct decompression surgery only** was the relevant comparator for PICO 1 and 2.

b) **Fusion** with an implant **with or without** direct decompression as co-intervention was the relevant comparator for PICO 3. The practice regarding the co-intervention (i.e. direct decompression) had to be identical to that in the intervention arm (PICO 3). Fusion could be performed with bone grafts or cage; additional fixation could be performed with screws and/or plates. Any combination of these was considered. Any type of surgical approach (e.g. anterior, posterior) was considered.

### 3.2.5 Outcomes – PICO 1-3

The relevant outcomes classified according to GRADE as critical and important outcomes[9-23] were:

1. Back pain (critical)
2. Radicular pain* (critical)
3. Spinal claudication (Walking distance or Zurich Claudication Questionnaire) (critical)
4. Quality of life (QoL) (e.g.EuroQoL) (important)
5. Function (e.g. ODI) (important)
6. Revision rate (important)
7. Reoperation rate (important)

8. Complication rate and adverse events (important)
9. Serious adverse events (important)

*If authors did not explicitly report radicular pain but arm and leg pain, arm and leg pain were extracted instead.

Surgical revisions were defined as operations due to ineffective initial surgery while reoperations were performed due to complications following initial surgery. As the definitions of revision and reoperation in the publications varied, it was decided a posteriori to use the definitions provided by the authors.

As the included trials frequently did not report on walking distance, data from the Zurich Claudication Questionnaire was extracted instead.

### 3.2.6 Study design – PICO 1-3

Only randomised controlled trials (RCTs) and quasi-randomised trials were included for this topic.

### 3.2.7 Languages – PICO 1-3

Trials published in English, French, and German were eligible for inclusion.

## 3.3 Eligibility criteria for systematic reviews on disc prostheses – PICO 4 and 5

### 3.3.1 Populations – PICO 4 and 5

**PICO 4**

Eligible patients had symptomatic degenerative changes of the lumbar spine with or without neurological symptoms requiring the treatment of 1 affected level. Relevant symptoms included local back pain but also neurological symptoms like radiculopathy or neurogenic claudication. Any combination of symptoms was accepted. It was likely that the patients already had conservative treatment attempts prior to surgery. Studies in patients with prior disc prosthesis or fusion at the same level of the spine were excluded. Studies with patients with other causes for their symptoms than degenerative disease were not eligible.

Studies were eligible if at least 80% of the study population fulfilled the inclusion criteria.

**PICO 5**

Eligible patients had symptomatic degenerative changes of the cervical spine with or without neurological symptoms requiring the treatment of 1 affected level. Relevant symptoms included for example local neck pain but also neurological symptoms like radiculopathy or myelopathy. Any combination of symptoms was accepted. It was likely that the patients already had conservative treatment attempts prior to surgery. Studies in patients with prior disc prosthesis or fusion at the same level of the spine were excluded. Studies with patients with other causes for their symptoms than degenerative disease were not eligible.

Studies were eligible if at least 80% of the study population fulfilled the inclusion criteria.

### 3.3.2 Intervention – PICO 4 and 5

Any type of single-level disc prosthesis **with** or **without** direct decompression as co-intervention was eligible. The practice regarding the co-intervention (i.e. direct decompression) had to be identical to that in the comparator arm.

### 3.3.3 Comparator – PICO 4 and 5

Fusion with an implant **with** or **without** direct decompression as co-intervention was the relevant comparator. The practice regarding co-interventions (i.e. direct decompression) had to be identical to that in the intervention arm. Fusion could be performed with bone grafts or cage; additional fixation could be performed with screws and/or plates. Any combination of these was considered. Any type of surgical approach was considered.

As the removal of a prolapsed disc is synonymous with a direct decompression it was not mandatory that the included studies explicitly mentioned (direct) decompression as part of the treatment as long as the practice regarding the co-intervention was identical in the intervention and the comparator arm.

### 3.3.4 Outcomes – PICO 4 and 5

The relevant outcomes classified according to GRADE to distinguish between critical and important outcomes[9-23] were:

1. Radicular pain* (critical)
2. Myelopathy (PICO 5, critical)
3. Back /neck pain (PICO4 / PICO 5, important)
4. QoL (e.g. EuroQoL) (important)
5. Function (e.g. Neck disability questionnaire) (important)
6. Revision rate (important)
7. Reoperation rate (important)
8. Complication rate and adverse events (important)
9. Serious adverse events (important)

*If authors did not explicitly report radicular pain but arm and leg pain, arm and leg pain were extracted instead.

Surgical revisions were defined as operations due to ineffective initial surgery while reoperations were performed due to complications following initial surgery. As the definitions of revision and reoperation in the publications varied, it was decided to use the definitions provided by the authors.

### 3.3.5 Study design – PICO 4 and 5

Only RCTs and quasi-randomised trials were included for this topic.

### 3.3.6 Languages – PICO 4 and 5

Trials published in English, French, and German were eligible for inclusion

## 3.4 Literature search

The literature search comprised Medline via PubMed. Clinical experts and the producers of devices were given the opportunity to inform us about additional trials that fulfilled the inclusion criteria but had not been identified in the search.

The literature search strategy was combined with a search filter for RCTs: Cochrane Highly Sensitive Search Strategy for identifying randomised trials in MEDLINE: sensitivity- and precision-maximizing version (2008 revision), combined with the terms “randomised” and “random” (details in Appendix I).

Two reviewers independently screened titles/abstracts of records found in the literature search for potentially eligible studies. The full text articles of these were independently screened by two reviewers to identify eligible studies. Discrepant screening results were discussed and resolved by consensus or by third party arbitration.

## 3.5 Data extraction

Data on study characteristics and outcomes were extracted into a standardised form by one reviewer and checked by another. Discrepancies were resolved by discussion or third party arbitration.

Information on patient recruitment time, maximum follow-up time, setting and country, age, sex, eligibility criteria, and description of the study interventions were extracted.

If possible, outcome data for two different points in time were extracted, a short-term and a long-term follow-up. For the short-term follow-up, the relevant point in time was a follow-up of 1 year (where multiple time points were available the one closest to a follow-up of 1 year was taken, with a range of follow-up times considered from ≥ 1 year to a < 2 years). For the long-term follow-up, a follow-up ≥ 2 years was used and for each study and all the outcomes were extracted for the same time point. This time point was chosen based on the largest number of reported (and extractable) relevant outcomes. If the number of reported outcomes was the same for two different time points, the time point with the longest follow-up was used. The inclusion of results on outcomes assessed after the end of the official study period was considered if patients in both treatment arms had no special adjuvant treatments after their initial surgical intervention. Inclusion of these studies was decided on a case-by-case basis.

Only general adverse events or complications were extracted, i.e. data had to be termed “adverse event” or “complication” based on authors’ definitions. Surgery- or implant-related complications or similar descriptions were not extracted as the risk for adverse events may be underestimated.

## 3.6 Risk of bias and quality of evidence assessment

One reviewer assessed the internal validity (risk of bias assessment) of each trial and per endpoint. This was checked by a second reviewer. Discrepancies were resolved by discussion or third party arbitration.

To assess the risk of bias of individual trials the following criteria were used[9-24]:

- adequate random sequence generation
- adequate concealment of treatment allocation
- adequate blinding of patients and health carers
- adequate blinding of outcome assessors
- complete outcome data
- reporting bias

Blinding of outcome assessors and complete outcome data were judged at the outcome level. To judge the completeness of outcome data and the resulting risk of bias, the following operationalisation was used:

- The risk of bias was judged low if the proportion of patients with missing data was 0 - 10% in either study arm and comparable between the randomised treatment arms.
- The risk of bias was also judged low if the proportion of patients with missing data was between 10-20% per arm, was comparable between the randomised treatment arms, and was being addressed using adequate methods. In case of continuous data, methods considered to be adequate are so called multiple imputation methods but not simple replacement methods like "last observation carried forward" or "baseline carried forward". In case of binary data adequate methods to address missing data were so called conservative assumptions about missing data; i.e. those patients with missing data in the control arm are treated in the analysis as if they have beneficial outcome results.
- Missing data in the treatment arms were considered comparable if the difference between the groups was 5% or less.
- The risk of bias was judged high if more than 20% of the data were missing irrespective of how the missing data were addressed in the analysis.

Reporting bias was judged low if all outcomes relevant for the review were stated in both the methods section and the results section.

The quality of the evidence was judged by one reviewer and checked by another according to GRADE (Grading of Recommendations Assessment, Development and Evaluation) for the critical and important outcomes for the long-term follow-up (≥ 2 years; i.e. on the outcome level by considering all the available trials for the respective outcome). Discrepancies were resolved by consensus or third party arbitration. The following criteria were considered to judge the quality of the evidence[9-23].

*Criteria for rating down the quality of evidence:*

- risk of bias (internal validity)
- inconsistency
- indirectness
- imprecision
- publication bias

*Criteria for rating up the quality of evidence:*

- large magnitude of effect
- dose-response gradient
- all plausible confounders or other biases increase the confidence in the estimated effect

Using the GRADE software (GRADEprofiler Version 3.6.1) results of the judgement were presented in a summary of findings table.

## 3.7 Data synthesis

Study characteristics and results of the eligible trials were presented per study in tables and summarised descriptively.

The main focus of the analysis was on the long-term follow up (≥ 2 years).

Where possible, outcome results were summarised quantitatively in a meta-analysis by using inverse variance model assuming random effects[25]. The analyses were performed using Review Manager (Version 5.3.5).

In case the relevant outcome data were not available, they were calculated based on other relevant information in the publication. For data where it was unclear whether the mean or the median had been given, it was assumed that the data referred to the mean. Missing standard deviations were approximated by the median standard deviations of other included studies on the same outcome measure[24,26]. If that was not possible, other SDs reported in the publication were discussed for approximation and this was indicated in the analysis. Authors were contacted for information if it was unclear whether the publication was part of another trial. They were also contacted regarding the number of patients analysed at the long-term follow-up if this information could not be deducted from the publication. For the short-term follow-up it was assumed that the follow-up had been complete if no other information was available. If there is no data on long-term follow-up and authors did not reply, the number of patients randomised was used for long-term follow-up.

Continuous outcomes were presented as mean differences. For binary outcomes the absolute and relative risks were determined. Effect estimates (summary and single for each trial) with the corresponding 95% confidence interval were presented in forest plots.

If a continuous outcome was measured on different scales, mean differences of the individual trial results would have been standardised using the following formula:

Standard mean difference (SMD) = $(\text{mean}_{\text{surgery}} - \text{mean}_{\text{conservative}})/\text{SD}_{\text{pooled}}$

An effect size of 0.2 standard deviations (SD) corresponds to a small effect; effect sizes of 0.5 and 0.8 SDs correspond to medium and big effects, respectively[27,28].

The presence of heterogeneity among the pooled effect estimates was estimated using $I^2$. Estimates of $I^2$ were interpreted under the guidance of the Cochrane Handbook[24]. Heterogeneity with an $I^2$ of 0% to 40% was considered low, 41% to 60% was considered moderate, and 61 to 100% high. The importance of the observed $I^2$ value depends on (i) magnitude and direction of effects and (ii) strength of evidence for heterogeneity (e.g. p value from the chi-squared test, or a confidence interval for $I^2$)[24].

In case of substantial or considerable heterogeneity, methodological and clinical factors that might explain the heterogeneity were explored in explorative subgroup and sensitivity analyses.

### 3.7.1 Subgroup analyses

In order to answer questions regarding possible variations of the effects depending on the patients treated, the type of intervention and study design and to investigate possible heterogeneous results subgroup analyses were planned for the following pre-specified subsets.

Subgroup analyses will be performed depending on the number of available trials per PICO. The sequence of the subgroup analyses listed below corresponds to the sequence in which the subgroup analyses will be performed depending on the number of available trials. For subgroup analyses a minimum number of 5 studies had to be included per PICO.

**PICO 1, 2 and 3:**

1. Interspinous stabilisation vs. pedicle-based stabilisation
2. Studies including patients with neurologic symptoms vs. no neurologic symptoms
3. Studies including patients with 1 affected level vs. patients 2 affected levels
4. Anterior fusion vs. posterior fusion vs. other (PICO 3)
5. Comparator fusion with bone graft vs. fusion with cage (PICO 3)

Patients with failed prior surgical treatment were considered separately if surgery (i.e. fusion, disc prosthesis, dynamic stabilisation) in the trial was performed on a different level than the previous intervention. It was decided a posteriori to also include a trial where fusion and dynamic stabilisation on different levels was performed simultaneously. Subgroup analysis 1 will be performed irrespective of the number of available studies.

**PICO 4 and 5:**

1. 
    a. Patients only with radiculopathy of the cervical spine vs. patients with myelopathy with or without radiculopathy vs. patients without neurological symptoms (PICO 5)
    b. Patients only with radiculopathy of the lumbar spine vs. patients without neurological symptoms (PICO 4)
2. Anterior fusion vs. posterior fusion vs. other
3. Comparator fusion with bone graft vs. fusion with cage
4. Patients with failed prior surgical treatment were considered separately if surgery (i.e. fusion, disc prosthesis, dynamic stabilisation) in the included trial was performed on a different level than the previous intervention.

**All PICO-Questions: Subgroup analyses for methodological aspects for**

- Adequate vs. no adequate or unclear allocation concealment
- Adequate vs. inadequate or unclear randomization
- Adequate vs. inadequate or unclear blinding of patients, carers, and outcome assessors
- Complete vs. incomplete, imputed or unclear outcome data

### 3.7.2 Sensitivity analyses

In case of substantial or considerable heterogeneity, measured with $I^2$, explorative sensitivity analyses were conducted.

# 4 Results

## 4.1 Literature search

The electronic literature search yielded 2902 records (last search 19 April 2016) and clinical experts contributed one additional study that was not identified via Pubmed/Medline. These 2903 records were screened at title and abstract level and 148 potentially relevant records were screened in full-text. Finally, 72 journal articles were included, corresponding to 27 studies (RCTs).

Details regarding the search strategy and the number of studies and publications included per PICO are documented in Appendix I. The study selection process is presented in Figure 1.

As multiple publications were identified for some of the studies a unique study ID was assigned to each study throughout the report.

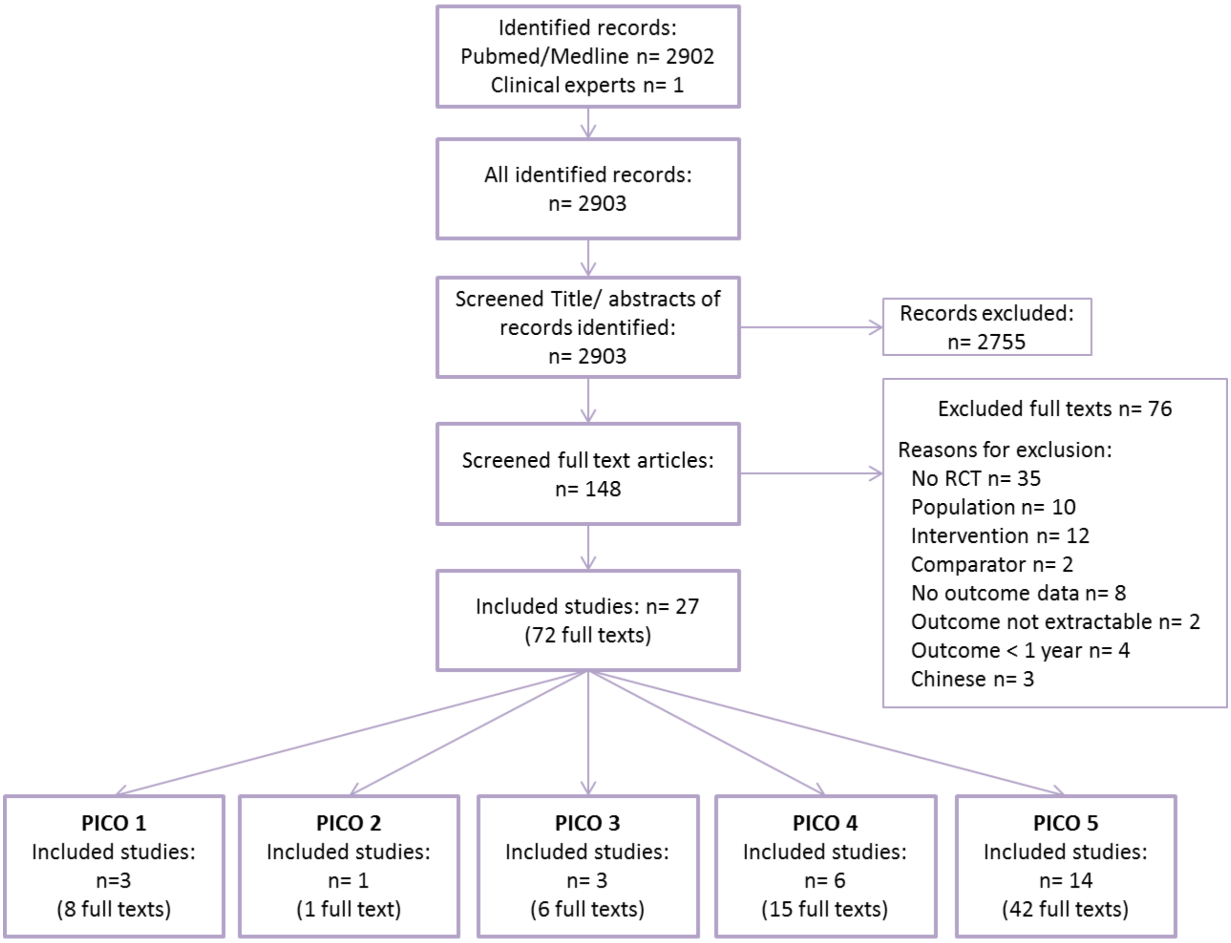


**Figure 1 Study selection process**

## 4.2 Interspinous and pedicle-based devices – PICO 1-3

### 4.2.1 Results – PICO 1

The following sections show first the study characteristics and risk of bias assessment and then the results for each outcome including GRADE results for PICO 1. Three relevant studies (Lønne 2015, Moojen 2013, Strömqvist 2013; see Table 3 for references) have been identified. All three studies have reported short- and long-term results. An overview of the outcomes analysed from each study is given in Table 4. Only the pooled long-term results are presented here. The pooled short-term results (1-year follow-up) for PICO 1 are presented in Appendix II. At short-term, no study examined pedicle-based stabilisation.

Table 3 Overview of included studies and their unique study IDs – PICO 1

| Study ID | Reference (main reference highlighted in colour) |
|---|---|
| **Lønne 2015** | Lønne G, Johnsen LG, Aas E, et al. Comparing cost-effectiveness of X-Stop with minimally invasive decompression in lumbar spinal stenosis: a randomized controlled trial. *Spine (Phila Pa 1976).* 2015;40(8):514-520. |
| | Lønne G, Johnsen LG, Rossvoll I, et al. Minimally invasive decompression versus x-stop in lumbar spinal stenosis: a randomized controlled multicenter study. *Spine (Phila Pa 1976).* 2015;40(2):77-85. |
| **Moojen 2013** | Moojen WA, Arts MP, Jacobs WC, et al. IPD without bony decompression versus conventional surgical decompression for lumbar spinal stenosis: 2-year results of a double-blind randomized controlled trial. *Eur Spine J.* 2015;24(10):2295-2305. |
| | Moojen WA, Arts MP, Jacobs WC, et al. Interspinous process device versus standard conventional surgical decompression for lumbar spinal stenosis: randomised controlled trial. *Br J Sports Med.* 2015;49(2):135. |
| | Moojen WA, Arts MP, Jacobs WC, et al. Interspinous process device versus standard conventional surgical decompression for lumbar spinal stenosis: randomized controlled trial. *BMJ.* 2013;347:f6415. |
| | Moojen WA, Arts MP, Brand R, Koes BW, Peul WC. The Felix-trial. Double-blind randomization of interspinous implant or bony decompression for treatment of spinal stenosis related intermittent neurogenic claudication. *BMC Musculoskelet Disord.* 2010;11:100. |
| | van den Akker-van Marle ME, Moojen WA, Arts MP, Vleggeert-Lankamp CL, Peul WC. Interspinous Process Devices versus Standard Conventional Surgical Decompression for Lumbar Spinal Stenosis: Cost Utility Analysis. *Spine J.* 2014. |
| **Strömqvist 2013** | Strömqvist BH, Berg S, Gerdhem P, et al. X-stop versus decompressive surgery for lumbar neurogenic intermittent claudication: randomized controlled trial with 2-year follow-up. *Spine (Phila Pa 1976).* 2013;38(17):1436-1442. |

Table 4 Overview of the outcomes analysed for PICO 1

| | Back pain | Radicular pain | Spinal claudication | Myelopathy | Quality of life | Function | Revision rate | Reoperation rate | Complications/AE | Serious AE |
|---|---|---|---|---|---|---|---|---|---|---|
| Lønne 2015 | 2 | 2 | 2 | | 2 | 2 | | 2 | 2 | |
| Moojen 2013 | 2 | 2 | 2 | | | 2 | | 2 | | |
| Strömqvist 2013 | 2 | | 2 | | 2 | | | 2 | | |

The number in the fields denote the analysed long-term follow-up in years

#### *4.2.1.1 Characteristics of the included studies – PICO 1*

General characteristics of studies for PICO-question 1 are summarised in Table 5. The three included RCTs for PICO 1 were multicentre studies conducted in Northern Europe and included a total of 355 participants. The enrolment period was not reported in one study and ranged from 2007 to 2011 in the other two. Maximum follow-up was two years in all three studies. All participants were affected and treated both on 1 and 2 levels of the lumbar spine. In one study, it was reported that participants had neurological symptoms (Lønne 2015). In the two other studies (Moojen 2013, Strömqvist 2013) participants were reported to have spinal stenosis with neurogenic claudication and therefore the presence of neurologic symptoms was assumed. Participants' mean age ranged from 64 to 71 years. The technology used as intervention in the three studies as intervention was interspinous stabilisation. In two studies, the comparator was decompression, and in one study, it was minimally invasive decompression.

Table 5 Study characteristics, PICO 1

| **Study ID** | **Country**<br>**Setting** | **Enrollment period**<br>**Maximum FU** | **Population**<br>Key condition<br>Affected levels | **Intervention**<br>○ n randomised<br>○ Male n (%)<br>○ Mean age (SD) | **Comparator**<br>○ n randomised<br>○ Male n (%)<br>○ Mean age (SD) |
|---|---|---|---|---|---|
| Lønne 2015 | Norway<br>Multicentre (6 sites) | Jun 2007-Sep 2011<br>2 years | Lumbar spinal stenosis with neurological symptoms<br>1-/2-level | Interspinous stabilisation (X-Stop)<br>○ 47<br>○ 17 (42%)<br>○ 67 (8.8) years | Minimally invasive decompression<br>○ 49<br>○ 23 (56%)<br>○ 67 (8.7) years |
| Moojen 2013 | The Netherlands<br>Multicenter (5 sites) | Oct 2008-Sep 2011<br>2 years | Degenerative lumbar canal stenosis with neurogenic claudication<br>1-/2-level | Interspinous stabilisation (Coflex)<br>○ 80<br>○ 49 (61%)<br>○ 66 (45-83)* years | Decompression<br>○ 79<br>○ 37 (47%)<br>○ 64 (47-83)* years |
| Strömqvist 2013 | Sweden<br>Multicentre (3 sites) | n.r.<br>2 years | Spinal stenosis with symptoms of neurogenic claudication<br>1-/2-level | Interspinous stabilisation (X-Stop)<br>○ 50<br>○ 30 (60%)<br>○ 67 (49-89)** years | Decompression<br>○ 50<br>○ 26 (52%)<br>○ 71 (57-84)** years |

Note: Under key condition the presence or absence of neurological symptoms is being described according to the inclusion criteria. If additional information based on the baseline characteristics would have been available this would have been documented in footnotes.
* median (range); ** mean (range);
Abbreviations: FU, follow-up; n, number; n.r., not reported; SD, standard deviation

#### *4.2.1.2 Risk of bias – PICO 1*

The method for the random sequence generation was adequate in two studies (Lønne 2015, Moojen 2013) and unclear in one study (Strömqvist 2013). Allocation concealment was adequate in two studies (Lønne 2015, Moojen 2013) and unclear in one study (Strömqvist 2013). The risk of performance and of detection bias was low in one study (Moojen 2013) and unclear in two studies (Lønne 2015, Strömqvist 2013). Risk of attrition bias for continuous and binary outcome data was rated to be low in one study (Strömqvist 2013) and to be high in two studies (Lønne 2015, Moojen 2013). Reporting bias was unclear in two studies (Lønne 2015, Moojen 2013) and was rated high in one study (Strömqvist 2013). A summarised overview of the risk of bias assessment is shown in Table 6 and a detailed summary with support of judgement in Appendix III.

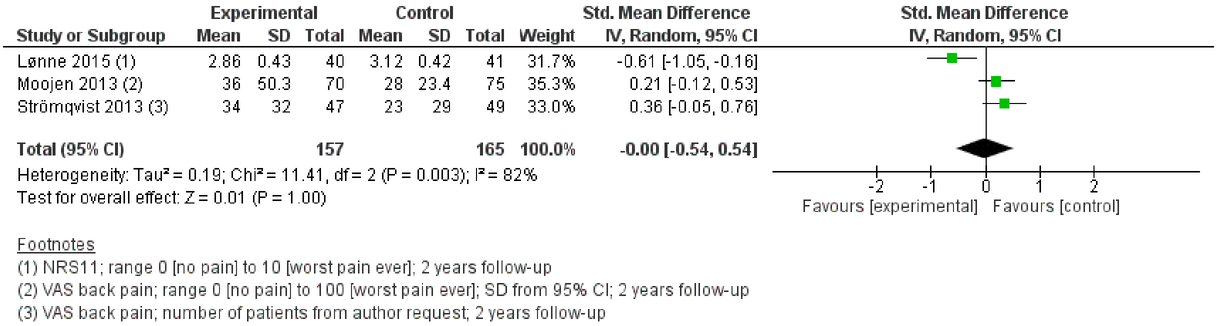
Table 6 Risk of bias, PICO 1

| Study | Random sequence generation (selection bias) | Allocation concealment (selection bias) | Blinding of participants and personnel (performance bias) | Blinding of outcome assessment, judgement did not differ among outcomes (detection bias) | Incomplete continuous outcome data (attrition bias) | Incomplete binary data (attrition bias) | Selective reporting (reporting bias) |
|---|---|---|---|---|---|---|---|
| Lønne 2015 | Low | Low | Unclear | Unclear | High | High | Unclear |
| Moojen 2013 | Low | Low | Low | Low | High | High | Unclear |
| Strömqvist 2013 | Unclear | Unclear | Unclear | Unclear | Low | Low | High |

### *4.2.1.3 Critical outcomes– PICO 1*

#### 4.2.1.3.1 Back pain

Three studies reported on back pain with a long-term follow-up of 2 years. Because different instruments were used to measure back pain, standardised mean differences (SMDs) were pooled. Two studies (Moojen 2013, Strömqvist 2013) used the Visual Analogue Scale (VAS, range 0 [no pain] to 100 [worst pain ever]). One study (Lønne 2015) used the Numeric Rating Scale (NRS11, range 0 [no pain] to 10 [worst pain ever]). There was no statistically significant difference between dynamic stabilisation, i.e. interspinous stabilisation, and direct decompression only (SMD -0.00, 95% CI -0.54 to 0.54; very low quality of evidence; Figure 2). Heterogeneity between studies was high ($I^2$=82%). Short-term (1 year) results for back pain showed a similar effect though heterogeneity for back pain was higher for the short-term than for the long-term follow-up.



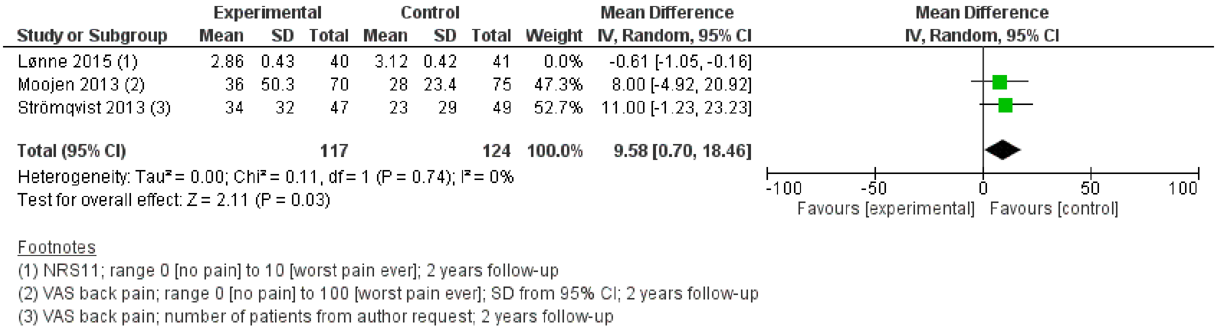
|  | Experimental |  |  | Control |  |  |  | Std. Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Lønne 2015 (1) | 2.86 | 0.43 | 40 | 3.12 | 0.42 | 41 | 31.7% | -0.61 [-1.05, -0.16] |
| Moojen 2013 (2) | 36 | 50.3 | 70 | 28 | 23.4 | 75 | 35.3% | 0.21 [-0.12, 0.53] |
| Strömqvist 2013 (3) | 34 | 32 | 47 | 23 | 29 | 49 | 33.0% | 0.36 [-0.05, 0.76] |
| **Total (95% CI)** |  |  | **157** |  |  | **165** | **100.0%** | **-0.00 [-0.54, 0.54]** |

Heterogeneity: Tau² = 0.19; Chi² = 11.41, df = 2 (P = 0.003); I² = 82%
Test for overall effect: Z = 0.01 (P = 1.00)

Footnotes
(1) NRS11; range 0 [no pain] to 10 [worst pain ever]; 2 years follow-up
(2) VAS back pain; range 0 [no pain] to 100 [worst pain ever]; SD from 95% CI; 2 years follow-up
(3) VAS back pain; number of patients from author request; 2 years follow-up

**Figure 2 PICO 1 long-term: Back pain**

A sensitivity analysis identified Lønne 2015 (using NRS11) as possible cause for heterogeneity. Exclusion of Lønne 2015 decreased heterogeneity to $I^2$=0%. Direct decompression then had a significantly greater effect on reduction of lower back pain (MD -9.58, 95% CI 0.70 to 18.46; Figure 3).



|  | Experimental |  |  | Control |  |  |  | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Lønne 2015 (1) | 2.86 | 0.43 | 40 | 3.12 | 0.42 | 41 | 0.0% | -0.61 [-1.05, -0.16] |
| Moojen 2013 (2) | 36 | 50.3 | 70 | 28 | 23.4 | 75 | 47.3% | 8.00 [-4.92, 20.92] |
| Strömqvist 2013 (3) | 34 | 32 | 47 | 23 | 29 | 49 | 52.7% | 11.00 [-1.23, 23.23] |
| **Total (95% CI)** |  |  | **117** |  |  | **124** | **100.0%** | **9.58 [0.70, 18.46]** |

Heterogeneity: Tau² = 0.00; Chi² = 0.11, df = 1 (P = 0.74); I² = 0%
Test for overall effect: Z = 2.11 (P = 0.03)

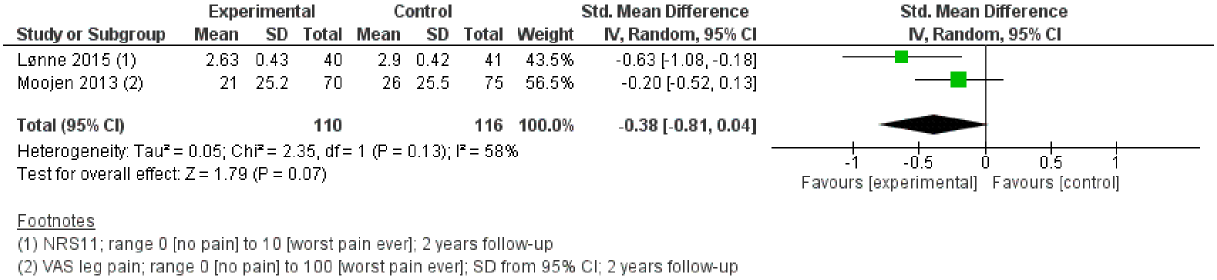
Footnotes
(1) NRS11; range 0 [no pain] to 10 [worst pain ever]; 2 years follow-up
(2) VAS back pain; range 0 [no pain] to 100 [worst pain ever]; SD from 95% CI; 2 years follow-up
(3) VAS back pain; number of patients from author request; 2 years follow-up

**Figure 3 PICO 1 long-term: Back pain - Sensitivity analysis excluding the study with different measure of back pain**

#### 4.2.1.3.2 Radicular pain

Three studies (Lønne 2015, Moojen 2013 and Strömqvist 2013) reported on leg pain with a long-term follow-up of 2 years. Two studies (Moojen 2013, Strömqvist 2013) used the VAS (range 0 [no pain] to 100 [worst pain ever]) while one (Lønne 2015) used NRS11 (range 0 [no pain] to 10 [worst pain ever]) (Figure 4). One study (Strömqvist 2013) reported the VAS for the left and right pain separately and therefore, could not be pooled with the other two studies. At 2 years, this study reported a mean (SD not reported) VAS score of 21 and 25.5 for the right and left leg, respectively, for the intervention group and a mean VAS score of 20.6 and 19 for the right and left leg, respectively, for the control group. In the other two studies (Lønne 2015 and Moojen 2013), there was no significant difference between interspinous stabilisation and direct decompression only (SMD -0.38, 95% CI -0.81 to 0.04;

low quality of evidence; Figure 4). Heterogeneity was high ($I^2$=58%). Short-term (1 year) results for radicular pain showed a similar effect.



|  | Experimental |  |  | Control |  |  |  | Std. Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Lønne 2015 (1) | 2.63 | 0.43 | 40 | 2.9 | 0.42 | 41 | 43.5% | -0.63 [-1.08, -0.18] |
| Moojen 2013 (2) | 21 | 25.2 | 70 | 26 | 25.5 | 75 | 56.5% | -0.20 [-0.52, 0.13] |
| **Total (95% CI)** |  |  | **110** |  |  | **116** | **100.0%** | **-0.38 [-0.81, 0.04]** |

Heterogeneity: Tau² = 0.05; Chi² = 2.35, df = 1 (P = 0.13); I² = 58%
Test for overall effect: Z = 1.79 (P = 0.07)

Footnotes
(1) NRS11; range 0 [no pain] to 10 [worst pain ever]; 2 years follow-up
(2) VAS leg pain; range 0 [no pain] to 100 [worst pain ever]; SD from 95% CI; 2 years follow-up

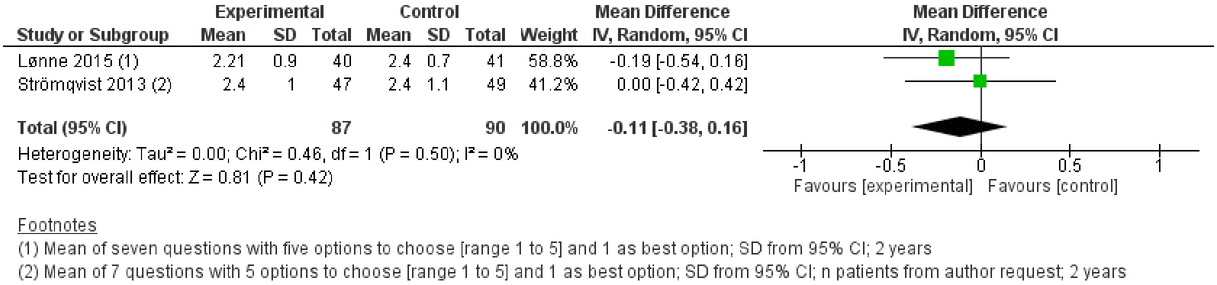
Figure 4 PICO 1 long-term: Radicular pain

### 4.2.1.3.3 Spinal claudication (Walking distance or Zurich Claudication Questionnaire)

No study reported on walking distance. Two studies (Lønne 2015, Strömqvist 2013) reported the subscales of the Zurich Claudication Questionnaire (ZCQ), namely symptom severity, physical function and patient satisfaction, at a long-term follow-up of 2 years.

**Zurich Claudication Questionnaire: Symptom severity**

Two studies (Lønne 2015, Strömqvist 2013) reported on symptom severity (mean of seven questions with five options to choose [range 1 to 5] and 1 as best option) at 2-year follow-up. There was no significant difference between interspinous stabilisation and direct decompression only (MD -0.11, 95% CI -0.38 to 0.16; low quality of evidence; Figure 5). Heterogeneity between studies was low ($I^2$=0%). Short-term (1 year) results for ZCQ symptom severity showed a similar effect though heterogeneity for ZCQ symptom severity was higher for the short-term than for the long-term follow-up.



|  | Experimental |  |  | Control |  |  |  | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Lønne 2015 (1) | 2.21 | 0.9 | 40 | 2.4 | 0.7 | 41 | 58.8% | -0.19 [-0.54, 0.16] |
| Strömqvist 2013 (2) | 2.4 | 1 | 47 | 2.4 | 1.1 | 49 | 41.2% | 0.00 [-0.42, 0.42] |
| **Total (95% CI)** |  |  | **87** |  |  | **90** | **100.0%** | **-0.11 [-0.38, 0.16]** |

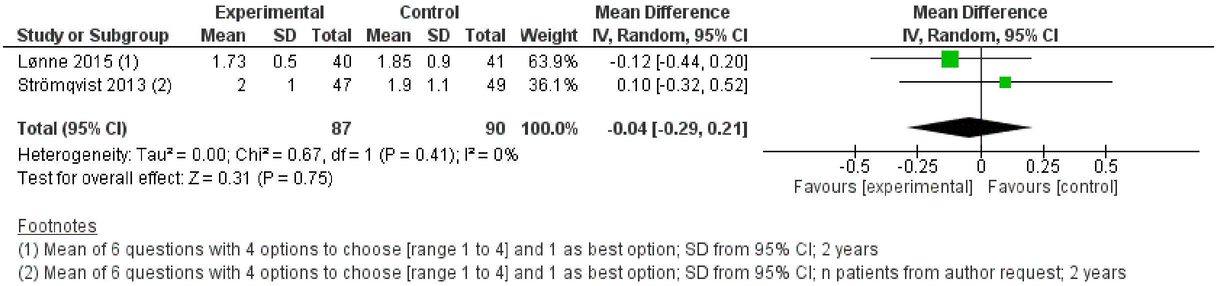
Heterogeneity: Tau² = 0.00; Chi² = 0.46, df = 1 (P = 0.50); I² = 0%
Test for overall effect: Z = 0.81 (P = 0.42)

Footnotes
(1) Mean of seven questions with five options to choose [range 1 to 5] and 1 as best option; SD from 95% CI; 2 years
(2) Mean of 7 questions with 5 options to choose [range 1 to 5] and 1 as best option; SD from 95% CI; n patients from author request; 2 years

Figure 5 PICO 1 long-term: Zurich Claudication Questionnaire, Symptom severity

**Zurich Claudication Questionnaire: Patient satisfaction**

Two studies (Lønne 2015, Strömqvist 2013) reported on patient satisfaction (mean of six questions with four options to choose [range 1 to 4] and 1 as best option) at 2 years follow-up. There was no significant difference between interspinous stabilisation and direct decompression only in patient satisfaction (MD -0.04, 95% CI -0.29 to 0.21; low quality of evidence; Figure 6). Heterogeneity between studies was low ($I^2$=0%). Short-term (1 year) results for ZCQ patient satisfaction showed a similar effect though heterogeneity for ZCQ patient satisfaction was higher for the short-term than for the long-term follow-up.



| Study or Subgroup | Experimental Mean | SD | Total | Control Mean | SD | Total | Weight | Mean Difference IV, Random, 95% CI |
|---|---|---|---|---|---|---|---|---|
| Lønne 2015 (1) | 1.73 | 0.5 | 40 | 1.85 | 0.9 | 41 | 63.9% | -0.12 [-0.44, 0.20] |
| Strömqvist 2013 (2) | 2 | 1 | 47 | 1.9 | 1.1 | 49 | 36.1% | 0.10 [-0.32, 0.52] |
| **Total (95% CI)** | | | **87** | | | **90** | **100.0%** | **-0.04 [-0.29, 0.21]** |

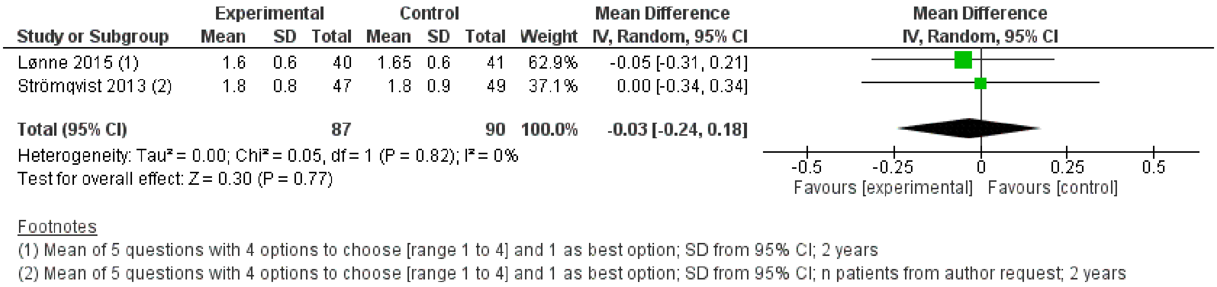
Heterogeneity: Tau² = 0.00; Chi² = 0.67, df = 1 (P = 0.41); I² = 0%
Test for overall effect: Z = 0.31 (P = 0.75)

Footnotes
(1) Mean of 6 questions with 4 options to choose [range 1 to 4] and 1 as best option; SD from 95% CI; 2 years
(2) Mean of 6 questions with 4 options to choose [range 1 to 4] and 1 as best option; SD from 95% CI; n patients from author request; 2 years

Figure 6 PICO 1 long-term: Zurich Claudication Questionnaire, Patient satisfaction

### Zurich Claudication Questionnaire: Physical function

Two studies (Lønne 2015, Strömqvist 2013) reported on physical function (mean of five questions with four options to choose [range 1 to 4] and 1 as best option) at 2 years follow-up. There was no difference between interspinous stabilisation and direct decompression only in physical function based on pooled results of two studies (MD -0.03, 95% CI -0.24 to 0.18; low quality of evidence; Figure 7). Heterogeneity between studies was low ($I^2$=0%). Short-term (1 year) results for ZCQ physical function showed a similar effect though heterogeneity for ZCQ physical function was higher for the short-term than for the long-term follow-up.



| Study or Subgroup | Experimental Mean | SD | Total | Control Mean | SD | Total | Weight | Mean Difference IV, Random, 95% CI |
|---|---|---|---|---|---|---|---|---|
| Lønne 2015 (1) | 1.6 | 0.6 | 40 | 1.65 | 0.6 | 41 | 62.9% | -0.05 [-0.31, 0.21] |
| Strömqvist 2013 (2) | 1.8 | 0.8 | 47 | 1.8 | 0.9 | 49 | 37.1% | 0.00 [-0.34, 0.34] |
| **Total (95% CI)** | | | **87** | | | **90** | **100.0%** | **-0.03 [-0.24, 0.18]** |

Heterogeneity: Tau² = 0.00; Chi² = 0.05, df = 1 (P = 0.82); I² = 0%
Test for overall effect: Z = 0.30 (P = 0.77)

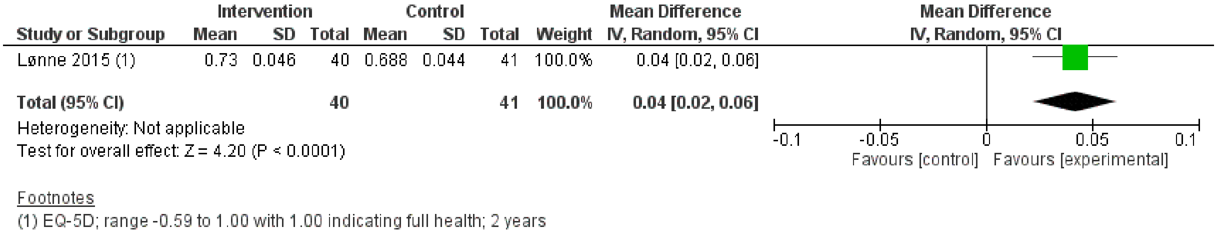
Footnotes
(1) Mean of 5 questions with 4 options to choose [range 1 to 4] and 1 as best option; SD from 95% CI; 2 years
(2) Mean of 5 questions with 4 options to choose [range 1 to 4] and 1 as best option; SD from 95% CI; n patients from author request; 2 years

Figure 7 PICO 1 long-term: Zurich Claudication Questionnaire, Physical function

### *4.2.1.4 Important outcomes – PICO 1*

#### 4.2.1.4.1 Quality of life

Two studies (Lønne 2015, Strömqvist 2013) reported on long-term results for quality of life with a follow-up of 2 years. One study (Strömqvist 2013) reported results for the SF-36 physical component score but no SDs. The SF-36 scales range from 1 to 100 with higher value indicating better quality of life. After 2 years of follow-up, Strömqvist reported 40 and 38 scores for the intervention and the comparator, respectively. Since missing SDs could not be imputed, these results could not be integrated in the meta-analysis. The other study (Lønne 2015) measured quality of life with the EuroQol 5-dimensional questionnaire (EQ-5D, range -0.59 to 1.00 with 1.00 indicating full health). Based on this one study, the mean difference between interspinous stabilisation and direct decompression was 0.04 (95% CI 0.02 to 0.06; low quality of evidence), favouring interspinous stabilisation (Figure 8). Short-term (1 year) results for quality of life showed a similar effect.



| | Intervention | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Lønne 2015 (1) | 0.73 | 0.046 | 40 | 0.688 | 0.044 | 41 | 100.0% | 0.04 [0.02, 0.06] |
| **Total (95% CI)** | | | **40** | | | **41** | **100.0%** | **0.04 [0.02, 0.06]** |

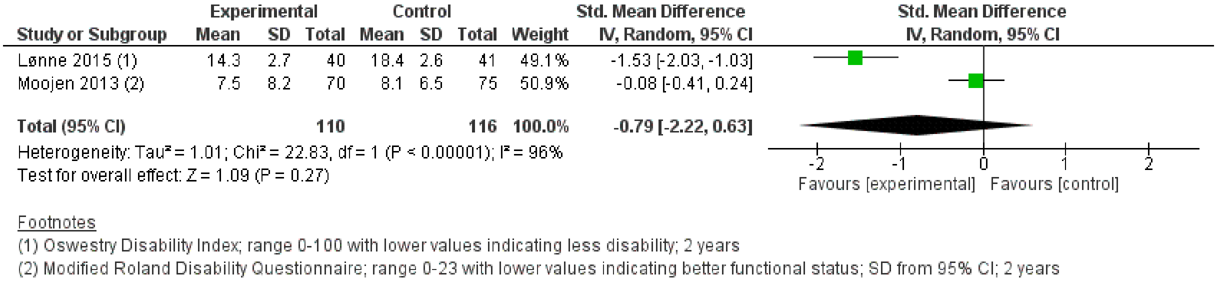
Heterogeneity: Not applicable
Test for overall effect: Z = 4.20 (P < 0.0001)

Footnotes
(1) EQ-5D; range -0.59 to 1.00 with 1.00 indicating full health; 2 years

Figure 8 PICO 1 long-term: Quality of life

### 4.2.1.4.2 Function

Two studies (Lønne 2015, Moojen 2013) reported long-term results for function with a follow-up of 2 years. One study (Lønne 2015) measured function with the Oswestry Disability Index (ODI, range 0-100% with lower values indicating better functional status). The second study (Moojen 2013) measured function with the Modified Roland Disability Questionnaire (range 0-23 with lower values indicating better functional status). There was no significant difference between interspinous stabilisation and direct decompression only (SMD -0.79, 95% CI -2.22 to 0.63; very low quality of evidence; Figure 9). Heterogeneity between the two studies using two different measures to assess function was high ($I^2$=96%). Short-term (1 year) results for function showed a similar effect.



| | Experimental | | | Control | | | | Std. Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Lønne 2015 (1) | 14.3 | 2.7 | 40 | 18.4 | 2.6 | 41 | 49.1% | -1.53 [-2.03, -1.03] |
| Moojen 2013 (2) | 7.5 | 8.2 | 70 | 8.1 | 6.5 | 75 | 50.9% | -0.08 [-0.41, 0.24] |
| **Total (95% CI)** | | | **110** | | | **116** | **100.0%** | **-0.79 [-2.22, 0.63]** |

Heterogeneity: Tau² = 1.01; Chi² = 22.83, df = 1 (P < 0.00001); I² = 96%
Test for overall effect: Z = 1.09 (P = 0.27)

Footnotes
(1) Oswestry Disability Index; range 0-100 with lower values indicating less disability; 2 years
(2) Modified Roland Disability Questionnaire; range 0-23 with lower values indicating better functional status; SD from 95% CI; 2 years

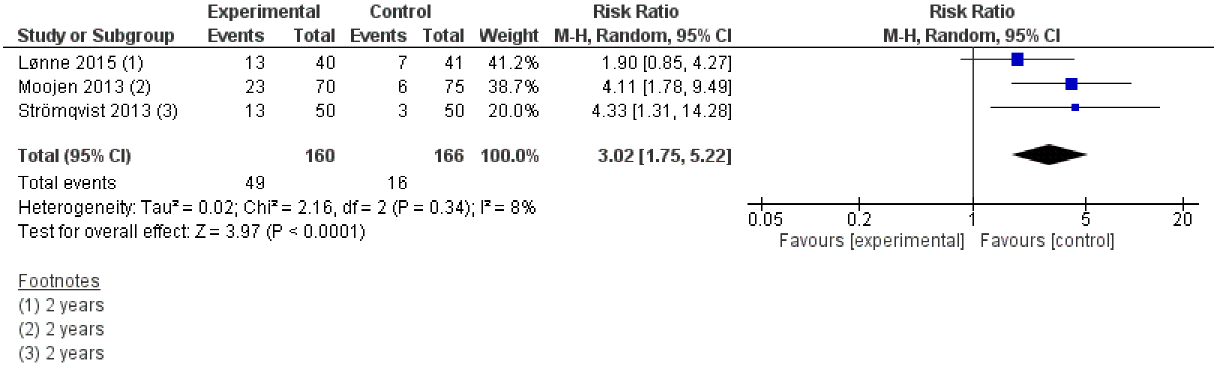
Figure 9 PICO 1 long-term: Function

### 4.2.1.4.3 Revision rate

No study reported on revision rate.

### 4.2.1.4.4 Reoperation rate

Reoperation rate was reported in three studies (Lønne 2015, Moojen 2013, Strömqvist 2013) at a follow-up of 2 years. Direct decompression resulted in significantly less reoperations than interspinous stabilisation (RR 3.02, 95% CI 1.75 to 5.22; low quality of evidence; Figure 10). Heterogeneity between studies was low ($I^2$=8%). Short-term (1 year) results for reoperation rate showed a similar effect.



| | Experimental | | Control | | | Risk Ratio |
|---|---|---|---|---|---|---|
| Study or Subgroup | Events | Total | Events | Total | Weight | M-H, Random, 95% CI |
| Lønne 2015 (1) | 13 | 40 | 7 | 41 | 41.2% | 1.90 [0.85, 4.27] |
| Moojen 2013 (2) | 23 | 70 | 6 | 75 | 38.7% | 4.11 [1.78, 9.49] |
| Strömqvist 2013 (3) | 13 | 50 | 3 | 50 | 20.0% | 4.33 [1.31, 14.28] |
| **Total (95% CI)** | | **160** | | **166** | **100.0%** | **3.02 [1.75, 5.22]** |
| Total events | 49 | | 16 | | | |

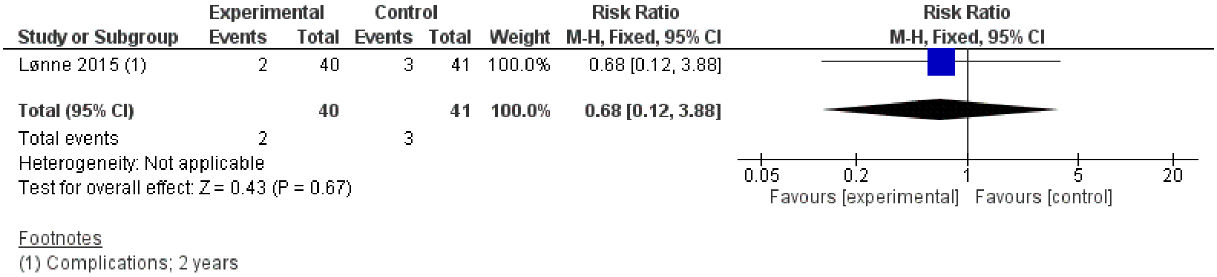
Heterogeneity: Tau² = 0.02; Chi² = 2.16, df = 2 (P = 0.34); I² = 8%
Test for overall effect: Z = 3.97 (P < 0.0001)

Footnotes
(1) 2 years
(2) 2 years
(3) 2 years

Figure 10 PICO 1 long-term: Reoperation rate

### 4.2.1.4.5 Complications or adverse events

Only one study (Lønne 2015) reported on complications at a follow-up of 2 years. There was no significant difference between interspinous stabilisation and direct decompression only (RR 0.68, 95% CI 0.12 to 3.88; very low quality of evidence; Figure 11). No study reported complications or adverse events at short-term follow-up.



| | Experimental | | Control | | | Risk Ratio |
|---|---|---|---|---|---|---|
| Study or Subgroup | Events | Total | Events | Total | Weight | M-H, Fixed, 95% CI |
| Lønne 2015 (1) | 2 | 40 | 3 | 41 | 100.0% | 0.68 [0.12, 3.88] |
| **Total (95% CI)** | | **40** | | **41** | **100.0%** | **0.68 [0.12, 3.88]** |
| Total events | 2 | | 3 | | | |

Heterogeneity: Not applicable
Test for overall effect: Z = 0.43 (P = 0.67)

Footnotes
(1) Complications; 2 years

Figure 11 PICO 1 long-term: Complications or adverse events

### 4.2.1.4.6 Serious adverse events

No study reported on serious adverse events.

### *4.2.1.5 Subgroup analyses – PICO 1*

Less than 5 RCTs were included for PICO 1 and all investigated devices were interspinous spacers. Hence, no subgroup analyses were performed.

### *4.2.1.6 GRADE – PICO 1*

**Table 7 PICO 1 long-term: Summary of findings (GRADE)**

**PICO 1 - Dynamic stabilisation without decompression compared to direct decompression for patients with neurological symptoms due to degenerative changes of the lumbar spine**

**Patient or population:** patients with neurological symptoms due to degenerative changes of the lumbar spine
**Settings:**
**Intervention:** PICO 1 - Dynamic stabilisation without decompression
**Comparison:** direct decompression

| Outcomes | Illustrative comparative risks* (95% CI) | | Relative effect (95% CI) | No of Participants (studies) | Quality of the evidence (GRADE) | Comments |
|---|---|---|---|---|---|---|
| | Assumed risk | Corresponding risk | | | | |
| | **Direct decompression** | **PICO 1 - Dynamic stabilisation without decompression** | | | | |
| **Back pain**<br>VAS etc | | The mean back pain in the intervention groups was<br>**0.00 standard deviations lower**<br>(0.54 lower to 0.54 higher) | | 322<br>(3 studies) | ⊕⊖⊖⊖<br>**very low**[1,2,3] | |
| **Radicular Pain** | | The mean radicular pain in the intervention groups was<br>**0.38 standard deviations lower**<br>(0.81 lower to 0.04 higher) | | 226<br>(2 studies) | ⊕⊕⊖⊖<br>**low**[4,5] | |
| **Quality of life**<br>EQ-5D | | The mean quality of life in the intervention groups was<br>**0.04 higher**<br>(0.02 to 0.06 higher) | | 81<br>(1 study) | ⊕⊕⊖⊖<br>**low**[6,7] | |
| **Function**<br>ODI and MRDQ | | The mean function in the intervention groups was<br>**0.79 standard deviations lower**<br>(2.22 lower to 0.63 higher) | | 226<br>(2 studies) | ⊕⊖⊖⊖<br>**very low**[4,8,9] | |
| **Revision rate** | See comment | See comment | Not estimable | 0<br>(0) | See comment | No RCT did report this outcome |
| **Reoperation rate** | **Study population** | | **RR 3.02**<br>(1.75 to 5.22) | 326<br>(3 studies) | ⊕⊕⊕⊖<br>**moderate**[1] | |
| | **96 per 1000** | **291 per 1000**<br>(169 to 503) | | | | |
| | **Moderate** | | | | | |
| **Complication rate and adverse events** | **Study population** | | **RR 0.68**<br>(0.12 to 3.88) | 81<br>(1 study) | ⊕⊖⊖⊖<br>**very low**[6,10] | |
| | **73 per 1000** | **50 per 1000**<br>(9 to 284) | | | | |
| | **Moderate** | | | | | |
| **Serious adverse events** | **Study population** | | Not estimable | 0 | See comment | No RCT did report this |

| | | | | | | |
|---|---|---|---|---|---|---|
| | See comment | See comment | | (0) | | outcome |
| | **Moderate** | | | | | |
| **Claudicatio spinalis, ZCQ-Symptom severity**<br>ZCQ | | The mean claudicatio spinalis, zcq-symptom severity in the intervention groups was<br>**0.11 lower**<br>(0.38 lower to 0.16 higher) | | 177<br>(2 studies) | ⊕⊕⊝⊝<br>**low**[11,12] | |
| **Claudicatio spinalis, ZCQ-Physical function**<br>ZCQ | | The mean claudicatio spinalis, zcq-physical function in the intervention groups was<br>**0.03 lower**<br>(0.24 lower to 0.18 higher) | | 177<br>(2 studies) | ⊕⊕⊝⊝<br>**low**[11,12] | |
| **Claudicatio spinalis, ZCQ-Satisfaction**<br>ZCQ | | The mean claudicatio spinalis, zcq-satisfaction in the intervention groups was<br>**0.04 lower**<br>(0.29 lower to 0.21 higher) | | 177<br>(2 studies) | ⊕⊕⊝⊝<br>**low**[11,12] | |

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).

**CI:** Confidence interval; **RR:** Risk ratio; **OR:** Odds ratio;

GRADE Working Group grades of evidence
**High quality:** Further research is very unlikely to change our confidence in the estimate of effect.
**Moderate quality:** Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
**Low quality:** Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
**Very low quality:** We are very uncertain about the estimate.

[1] Risk of selection bias (random sequence generation and allocation concealment) was unclear in one study; risk of performance bias was unclear in two studies; risk of detection bias was unclear in two studies; risk of attrition bias was high in two studies; risk of reporting bias was unclear in two and high in one studies.

[2] Inconsistency was downgraded by one level because heterogeneity (I2) was high, and there was minimal or no overlap of confidence intervals of the individual studies. Directions of effects were not similar and differences in effect sizes were large. Because of the limited number of studies it was not possible to perform sensitivity or subgroup analysis to assess reasons for heterogeneity.

[3] Imprecision was downgraded by two levels because the 95% CI of the SMD supported either the effectiveness of the intervention (assuming an MCID of 0.5) or the effectiveness of the comparator.

[4] Risk of performance bias was unclear in one study; risk of detection bias was unclear in one study; risk of attrition bias was high in two studies; risk of reporting bias was unclear in two studies.

[5] Imprecision was downgraded by one levels because total population was small. The 95% CI of the SMD included clinically relevant benefits and no effect for the intervention.

[6] Risk of performance bias, detection bias and selective reporting were unclear, and risk of attrition bias was high in one study

[7] Although the 95% CI of the effect estimate was narrow and excluded clinically important harm or benefit (assuming an MCID of 0.14), imprecision was downgraded by one level because the total sample size was lower than 400 (rule-of-thumb).

[8] Inconsistency was downgraded by one level because heterogeneity (I2) was high, and there was minimal or no overlap of confidence intervals of the individual studies. Directions of effects were not similar and differences in effect sizes were large. Because of the limited number of studies it was not possible to perform sensitivity or subgroup analysis to assess reasons for heterogeneity.

[9] Imprecision was downgraded by two levels because the 95% CI support either the effectiveness of the intervention (assuming an MCID of 0.5) or the effectiveness of the comparator, and because total sample size was low.

[10] Imprecision was downgraded by two levels because the 95% CI included appreciable benefit (greater than 25% relative risk reduction) and no benefit, and because the total sample size was lower than the optimal information size.

[11] Risk of selection bias (random sequence generation and allocation concealment) was unclear in one study; risk of performance bias and detection bias were unclear in two studies; risk for attrition bias

was high in one study; risk of reporting bias was high in one study.

[12] Although the 95% CI of the effect estimate was narrow and excluded clinically important harm or benefit (assuming an MCID of 0.5), imprecision was downgraded by one level because the total sample size was low.

### 4.2.2 Results – PICO 2

The following sections show first the study characteristics and risk of bias assessment and then the long-term results for each outcome including GRADE results for PICO 2. Only one study (Marsh 2014; see Table 8 for reference), comparing interspinous stabilisation and decompression with decompression alone, was identified. Data for meta-analysis were only available for the long-term follow-up of 4 years. An overview of the outcomes analysed from this study is given in Table 9. Information on short-term results (1-year follow-up) for PICO 2 are presented in Appendix IV. No study reported results at short-term follow-up.

**Table 8 Overview of included studies and their unique study IDs – PICO 2**

| Study ID | Reference |
|---|---|
| **Marsh 2014** | Marsh GD, Mahir S, Leyte A. A prospective randomised controlled trial to assess the efficacy of dynamic stabilisation of the lumbar spine with the Wallis ligament. *Eur Spine J.* 2014;23(10):2156-2160. |

**Table 9 Overview of the outcomes analysed for PICO 2**

| | **Back pain** | **Radicular pain** | **Spinal claudication** | **Myelopathy** | **Quality of life** | **Function** | **Revision rate** | **Reoperation rate** | **Complications/AE** | **Serious AE** |
|---|---|---|---|---|---|---|---|---|---|---|
| Marsh 2014 | 4 | | | | | 4 | | 4 | 4 | |

The number in the fields denote the analysed long-term follow-up in years

#### *4.2.2.1 Characteristics of the included studies – PICO 2*

General characteristics of the RCT for PICO-question 2 are summarised in Table 10. The RCT was a single centre study from the United Kingdom and included 60 participants. The enrolment period was not reported and the maximum follow-up was 4 years. Participants were reported to have symptomatic spinal stenosis and therefore the presence of neurologic symptoms was assumed. Participants were affected and treated both on 1 and on 2 levels of the lumbar spine. The mean age ranged from 56 to 60 years. The comparison was interspinous stabilisation versus decompression.

Table 10 Study characteristics, PICO 2

| **Study ID** | **Country**<br>**Setting** | **Enrollment period**<br>**Maximum FU** | **Population**<br>Key condition<br>Affected levels | **Intervention**<br>○ n randomised<br>○ Male n (%)<br>○ Mean age (SD) | **Comparator**<br>○ n randomised<br>○ Male n (%)<br>○ Mean age (SD) |
|---|---|---|---|---|---|
| Marsh 2014 | UK<br>Single Centre | n.r.<br>4 years | Symptomatic spinal stenosis with nerve root compression<br>1-/2-level | Interspinous stabilisation and decompression (Wallis implant)<br>○ 30<br>○ 11 (37%)<br>○ 59.6 (13.4) years | Decompression<br>○ 30<br>○ 14 (47%)<br>○ 56.4 (12.9) years |

Note: Under key condition the presence or absence of neurological symptoms is being described according to the inclusion criteria. If additional information based on the baseline characteristics would have been available this would have been documented in footnotes.
Abbreviations: FU, follow-up; n, number; n.r., not reported; SD, standard deviation

#### 4.2.2.2 Risk of bias – PICO 2

For the study (Marsh 2014) the risk of selection bias was rated to be low depending on adequate random sequence generation but the risk of selection bias was unclear due to allocation. No information about blinding was reported and therefore the risk of performance bias and detection bias were judged to be unclear. Attrition bias for binary and continuous outcome data and reporting bias were rated to be of low risk. A summarised overview of the risk of bias assessment is shown in Table 11 and a detailed summary with support of judgement in Appendix III.

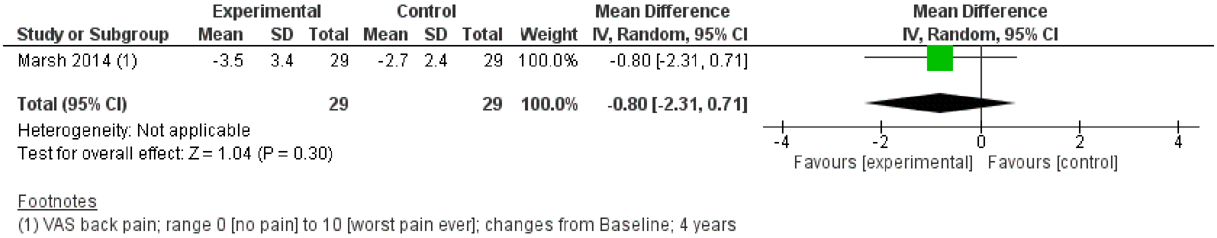
Table 11 Risk of bias, PICO 2

| Study | Random sequence generation (selection bias) | Allocation concealment (selection bias) | Blinding of participants and personnel (performance bias) | Blinding of outcome assessment, judgement did not differ among outcomes (detection bias) | Incomplete continuous outcome data (attrition bias) | Incomplete binary data (attrition bias) | Selective reporting (reporting bias) |
|---|---|---|---|---|---|---|---|
| Marsh 2014 | Low | Unclear | Unclear | Unclear | Low | Low | Low |

### *4.2.2.3 Critical outcomes – PICO 2*

#### 4.2.2.3.1 Back pain

One study (Marsh 2014) reported low back pain with a long-term follow-up of 4 years. The study reported the change measured on a Visual Analogue Scale (VAS, range 0 [no pain] to 10 [worst pain ever]). There was no significant difference between dynamic stabilisation, i.e. interspinous stabilisation plus decompression and direct decompression only (MD -0.80, 95% CI -2.31 to 0.71; very low quality of evidence; Figure 12).



| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| **Study or Subgroup** | **Mean** | **SD** | **Total** | **Mean** | **SD** | **Total** | **Weight** | **IV, Random, 95% CI** |
| Marsh 2014 (1) | -3.5 | 3.4 | 29 | -2.7 | 2.4 | 29 | 100.0% | -0.80 [-2.31, 0.71] |
| **Total (95% CI)** | | | **29** | | | **29** | **100.0%** | **-0.80 [-2.31, 0.71]** |

Heterogeneity: Not applicable
Test for overall effect: Z = 1.04 (P = 0.30)

Footnotes
(1) VAS back pain; range 0 [no pain] to 10 [worst pain ever]; changes from Baseline; 4 years

**Figure 12 PICO 2 long-term: Back pain**

#### 4.2.2.3.2 Radicular pain

No study reported on radicular pain.

#### 4.2.2.3.3 Spinal claudication (Walking distance or Zurich Claudication Questionnaire)

No study reported on spinal claudication.

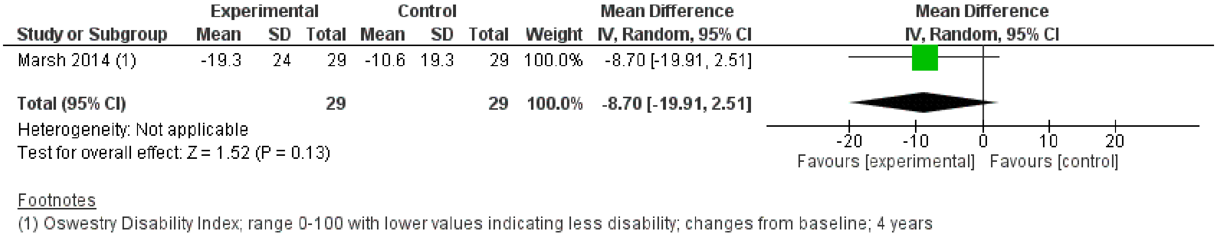
### *4.2.2.4 Important outcomes – PICO 2*

#### 4.2.2.4.1 Quality of life

No study reported on quality of life.

#### 4.2.2.4.2 Function

One study (Marsh 2014) reported long-term results for function with a follow-up of 4 years. The study measured function with the Oswestry Disability Index (ODI, range 0-100 with lower values indicating better functional status). There was no statistically significant difference between interspinous stabilisation plus decompression and direct decompression only (MD -8.70, 95% CI - 19.91 to 2.51; low quality of evidence; Figure 13).



| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| **Study or Subgroup** | **Mean** | **SD** | **Total** | **Mean** | **SD** | **Total** | **Weight** | **IV, Random, 95% CI** |
| Marsh 2014 (1) | -19.3 | 24 | 29 | -10.6 | 19.3 | 29 | 100.0% | -8.70 [-19.91, 2.51] |
| **Total (95% CI)** | | | **29** | | | **29** | **100.0%** | **-8.70 [-19.91, 2.51]** |

Heterogeneity: Not applicable
Test for overall effect: Z = 1.52 (P = 0.13)

Footnotes
(1) Oswestry Disability Index; range 0-100 with lower values indicating less disability; changes from baseline; 4 years

**Figure 13 PICO 2 long-term: Function**

#### 4.2.2.4.3 Revision rate

No study reported on revision rate.

#### 4.2.2.4.4 Reoperation rate

One study reported zero reoperations at a long-term follow-up of 4 years in 58 participants (Marsh 2014).

#### 4.2.2.4.5 Complications or adverse events

One study reported zero complications at a long-term follow-up of 4 years in 58 participants (Marsh 2014).

#### 4.2.2.4.6 Serious adverse events

No study reported on serious adverse events.

### *4.2.2.5 Subgroup analyses – PICO 2*

No subgroup and sensitivity analyses were conducted because only one study was included for this PICO-question.

### *4.2.2.6 GRADE – PICO 2*

Table 12 PICO 2 long-term: Summary of findings (GRADE)

**PICO 2 - Dynamic stabilisation with decompression compared to direct decompression for patients with neurological symptoms due to degenerative changes of the lumbar spine**

**Patient or population:** patients with neurological symptoms due to degenerative changes of the lumbar spine
**Intervention:** Dynamic stabilisation with decompression
**Comparison:** Direct decompression

| Outcomes | Illustrative comparative risks* (95% CI) | | Relative effect (95% CI) | No of Participants (studies) | Quality of the evidence (GRADE) | Comments |
|---|---|---|---|---|---|---|
| | Assumed risk | Corresponding risk | | | | |
| | **Direct decompression** | **PICO 2 - Dynamic stabilisation with decompression** | | | | |
| **Back pain**<br>VAS | | The mean back pain in the intervention groups was<br>**0.80 lower**<br>(2.31 lower to 0.71 higher) | | 58<br>(1 study) | ⊕⊖⊖⊖<br>**very low**[1,2] | |
| **Radicular Pain** | **Study population** | | Not estimable | 0<br>(0) | See comment | No RCT reported this outcome |
| | See comment | See comment | | | | |
| | **Moderate** | | | | | |
| **Quality of life** | **Study population** | | Not estimable | 0<br>(0) | See comment | No RCT reported this outcome |
| | See comment | See comment | | | | |
| | **Moderate** | | | | | |
| **Function**<br>ODI | | The mean function in the intervention groups was<br>**8.70 lower**<br>(19.19 lower to 2.51 higher) | | 58<br>(1 study) | ⊕⊕⊖⊖<br>**low**[1,3] | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Revision rate** | See comment | See comment | Not estimable | 0<br>(0) | See comment | No RCT reported this outcome |
| **Reoperation rate** | **Study population** | | Not estimable | 58<br>(1 study) | See comment | Single study, no events were reported |
| | See comment | See comment | | | | |
| | **Moderate** | | | | | |
| **Complication rate and adverse events** | **Study population** | | Not estimable | 58<br>(1 study) | See comment | Single study, no events were reported |
| | See comment | See comment | | | | |
| | **Moderate** | | | | | |
| **Serious adverse events** | **Study population** | | Not estimable | 0<br>(0) | See comment | No RCT reported this outcome |
| | See comment | See comment | | | | |
| | **Moderate** | | | | | |
| **Spinal claudication, ZCQ-Symptom severity**<br>ZCQ | **Study population** | | Not estimable | 0<br>(0) | See comment | No RCT reported this outcome |
| | See comment | See comment | | | | |
| | **Moderate** | | | | | |
| **Spinal claudication, ZCQ-Physical function**<br>ZCQ | **Study population** | | Not estimable | 0<br>(0) | See comment | No RCT reported this outcome |
| | See comment | See comment | | | | |
| | **Moderate** | | | | | |
| **Spinal claudication, ZCQ-Satisfaction**<br>ZCQ | **Study population** | | Not estimable | 0<br>(0) | See comment | No RCT reported this outcome |
| | See comment | See comment | | | | |
| | **Moderate** | | | | | |

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).

**CI:** Confidence interval; **RR:** Risk ratio;

GRADE Working Group grades of evidence
**High quality:** Further research is very unlikely to change our confidence in the estimate of effect.
**Moderate quality:** Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.

**Low quality:** Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
**Very low quality:** We are very uncertain about the estimate.

[1] Risk of selection bias (allocation concealment), performance bias and detection bias were unclear in one study

[2] Imprecision was downgraded by two levels because the 95% CI of the effect estimate included the possibility for a clinically relevant benefit or no effect assuming a MCID of 0.72, and because the total sample size was lower than the optimal information size.

[3] Imprecision was downgraded by one level because the 95% CI of the effect estimate included the possibility for a clinically relevant benefit or no effect assuming a MCID of 7.5 and because of small total population size.

### 4.2.3 Results – PICO 3

The following sections show first the study characteristics and risk of bias assessment and then the results for each outcome including GRADE results for PICO 3. Three relevant studies (Davis 2013, Madan 2003, Putzier 2010; see Table 13 for references) have been identified. All three studies have reported long-term results and two studies (Davis 2013, Putzier 2010) have reported short-term results. The study by Putzier 2010 reported on a slightly different intervention than pre-defined by the inclusion criteria and was therefore examined separately. Only the pooled long-term results are presented here. An overview of the outcomes analysed from each study is given in Table 14. The pooled short-term results (1-year follow-up) for PICO 3 are presented in Appendix V. At short-term, no study examined pedicle-based stabilisation.

**Table 13 Overview of included studies and their unique study IDs – PICO 3**

| Study ID | Reference (main reference highlighted in colour) |
|---|---|
| **Davis 2013** | Davis RJ, Errico TJ, Bae H, Auerbach JD. Decompression and Coflex interlaminar stabilization compared with decompression and instrumented spinal fusion for spinal stenosis and low-grade degenerative spondylolisthesis: two-year results from the prospective, randomized, multicenter, Food and Drug Administration Investigational Device Exemption trial. *Spine (Phila Pa 1976).* 2013;38(18):1529-1539. |
|  | Bae HW, Davis RJ, Lauryssen C, Leary S, Maislin G, Musacchio M, Jr. Three-Year Follow-up of the Prospective, Randomized, Controlled Trial of Coflex Interlaminar Stabilization vs Instrumented Fusion in Patients With Lumbar Stenosis. *Neurosurgery.* 2016. |
|  | Musacchio MJ, Lauryssen C, Davis RJ, et al. Evaluation of Decompression and Interlaminar Stabilization Compared with Decompression and Fusion for the Treatment of Lumbar Spinal Stenosis: 5-year Follow-up of a Prospective, Randomized, Controlled Trial. *Int J Spine Surg.* 2016;10:6. |
|  | Davis R, Auerbach JD, Bae H, Errico TJ. Can low-grade spondylolisthesis be effectively treated by either coflex interlaminar stabilization or laminectomy and posterior spinal fusion? Two-year clinical and radiographic results from the randomized, prospective, multicenter US investigational device exemption trial: clinical article. *J Neurosurg Spine.* 2013;19(2):174-184. |
| **Madan 2003** | Madan S, Boeree NR. Outcome of the Graf ligamentoplasty procedure compared with anterior lumbar interbody fusion with the Hartshill horseshoe cage. *Eur Spine J.* 2003;12(4):361-368. |
| **Putzier 2010** | Putzier M, Hoff E, Tohtz S, Gross C, Perka C, Strube P. Dynamic stabilization adjacent to single-level fusion: part II. No clinical benefit for asymptomatic, initially degenerated adjacent segments after 6 years follow-up. *Eur Spine J.* 2010;19(12):2181-2189. |

Table 14 Overview of the outcomes analysed for PICO 3

| | Back pain | Radicular pain | Spinal claudication | Myelopathy | Quality of life | Function | Revision rate | Reoperation rate | Complications/AE | Serious AE |
|---|---|---|---|---|---|---|---|---|---|---|
| Davis 2013 | 5 | 5 | 5 | | | 5 | | 5 | | |
| Madan 2003 | | | | | | 2.7 | 2.7 | | 2.7 | |
| Putzier 2010 | | | | | | 6.3 | | | | |

The number in the fields denote the analysed long-term follow-up in years

#### *4.2.3.1 Characteristics of the included studies – PICO 3*

General characteristics of the 3 studies included for PICO-question 3 are summarised in Table 15. One RCT was a multicentre study conducted in the USA (Davis 2013), the other two (Madan 2003, Putzier 2010) did not report the study setting but are possibly single centre studies conducted in Germany and probably the United Kingdom. The three studies included a total of 437 participants. The enrolment periods ranged from 1995 to 2010 and the follow-up between 2 and 6 years. One study included participants with neurologic symptoms who were affected and treated both on 1 and on 2 levels of the lumbar spine (Davis 2013). Participants in the two other studies had degenerative disc disease and no neurologic symptoms. The mean age across the three studies ranged from 44 to 64 years. In the two studies including participants without neurologic symptoms (Madan 2003, Putzier 2010), participants were younger in one study (mean age of 44 to 45 years versus mean age of 62 to 64) than in the other study. In the study by Davis 2013, interspinous stabilisation was compared with posterolateral spinal fusion with bone. The other two studies compared pedicle-based stabilisation with anterior lumbar interbody fusion with cage and bone (Madan 2003) and circumferential fusion with cage and bone (Putzier 2010), respectively. In one study, a fusion at an adjacent level was performed in addition to the pedicle-based stabilisation (Putzier 2010). The comparison in Putzier 2010 (dynamic stabilisation plus fusion on an adjacent level versus fusion) therefore differs slightly from the pre-defined PICO-question. For this reason, Putzier 2010 was excluded from the main analyses and the quality of evidence assessment and presented separately.

Table 15 Study characteristics, PICO 3

| **Study ID** | **Country**<br>**Setting** | **Enrollment period**<br>**Maximum FU** | **Population**<br>Key condition<br>Affected levels | **Intervention**<br>○ n randomised<br>○ Male n (%)<br>○ Mean age (SD) | **Comparator**<br>○ n randomised<br>○ Male n (%)<br>○ Mean age (SD) |
|---|---|---|---|---|---|
| Davis 2013 | USA<br>Multicentre (21 sites) | 2006-2010<br>5 years | Spinal stenosis or spondylolisthesis with neurogenic claudication<br>1-/2-level | Interspinous stabilisation (Coflex)<br>○ 215<br>○ n.r.<br>○ 62.1 (9.2) years | Posterolateral spinal fusion with bone<br>○ 107<br>○ n.r.<br>○ 64.1 (9.0) years |
| Madan 2003 | n.r. (probably UK)<br>n.r. | Apr 1995-Jun 1997 (treatment period)<br>2.7 years | Disc degeneration, including patients with leg pain<br>1-level | Pedicle-based stabilisation (Graf ligamentoplasty)<br>○ 28<br>○ 17 (61%)<br>○ 44 (26-70)* years | Anterior lumbar interbody fusion with cage and bone<br>○ 27<br>○ 12 (44%)<br>○ 45 (25-67)* years |
| Putzier 2010** | Germany<br>n.r. | Jan 2000-May 2002<br>6.3 years | DDD without radiculopathy<br>Unclear | Pedicle-based stabilisation (Dynesis) and adjacent circumferential fusion<br>○ 30<br>○ 17 (57%)<br>○ 44.9 (27-62)* years | Circumferential fusion with cage and bone<br>○ 30<br>○ 14 (47%)<br>○ 44.6 (27-63)* years |

Note: Under key condition the presence or absence of neurological symptoms is being described according to the inclusion criteria. If additional information based on the baseline characteristics would have been available this would have been documented in footnotes.
* mean (range);
**only considered for sensitivity analysis because the comparison in Putzier 2010 (dynamic stabilisation plus fusion on an adjacent level versus fusion) differs from the pre-defined PICO-question. For this reason, Putzier 2010 was excluded from the main analyses and the quality of evidence assessment.
Abbreviations: FU, follow-up; n, number; n.r., not reported; SD, standard deviation

#### *4.2.3.2 Risk of bias – PICO 3*

The method of random sequence generation was adequate in three studies (Davis 2013, Madan 2003 and Putzier 2010) and therefore the risk of selection bias was rated as low. The risk of selection bias depending on allocation concealment was rated as low for two studies (Davies 2013, Madan 2003) and as unclear for one study (Putzier 2010). The risk of performance bias depending on blinding of participants and personnel was rated to be high in two studies (Davis 2013, Putzier 2010) and to be unclear in one study (Madan 2003). Blinding of outcome assessment was unclear in two studies (Davies 2013, Madan 2003) and one study reported no blinding therefore the risk of detection bias was rated as high (Putzier 2010). Attrition bias for binary and continuous outcome data was rated as high in two studies (Davis 2013, Putzier 2010) and unclear in one study (Madan 2003). Selective reporting was unclear in two studies (Davis 2013, Madan 2003) and was rated as low in one study (Putzier 2010). A summarised overview of the risk of bias assessment is shown in Table 16 and a detailed summary with support of judgement in Appendix III.

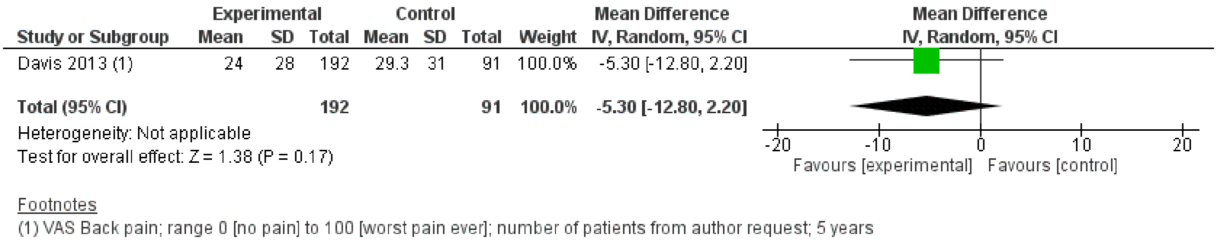
Table 16 Risk of bias, PICO 3

| Study | Random sequence generation (selection bias) | Allocation concealment (selection bias) | Blinding of participants and personnel (performance bias) | Blinding of outcome assessment, judgement did not differ among outcomes (detection bias) | Incomplete continuous outcome data (attrition bias) | Incomplete binary data (attrition bias) | Selective reporting (reporting bias) |
|---|---|---|---|---|---|---|---|
| Davis 2013 | Low | Low | High | Unclear | High | High | Unclear |
| Madan 2003 | Low | Low | Unclear | Unclear | Unclear | Unclear | Unclear |
| Putzier 2010 | Low | Unclear | High | High | High | High | Low |

### *4.2.3.3 Critical outcomes – PICO 3*

#### 4.2.3.3.1 Back pain

One study (Davis 2013) reported on low back pain at a 5-year follow-up. The study measured back pain on a Visual Analogue Scale (VAS, range 0 [no pain] to 100 [worst pain ever]). There was no statistically significant difference between interspinous stabilisation and fusion (MD -5.30, 95% CI -12.80 to 2.20; low quality of evidence; Figure 14). Short-term (1 year) results were only reported by Davis 2013 and showed a similar statistically non-significant effect on back pain.



| | Experimental | | | Control | | | | Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI | IV, Random, 95% CI |
| Davis 2013 (1) | 24 | 28 | 192 | 29.3 | 31 | 91 | 100.0% | -5.30 [-12.80, 2.20] | |
| **Total (95% CI)** | | | **192** | | | **91** | **100.0%** | **-5.30 [-12.80, 2.20]** | |

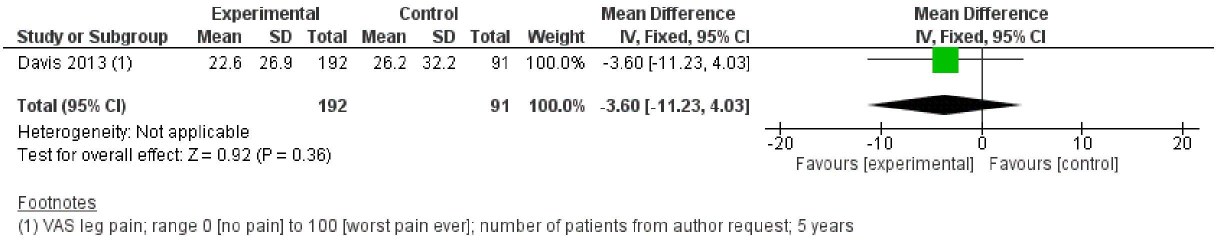
Heterogeneity: Not applicable
Test for overall effect: Z = 1.38 (P = 0.17)

Footnotes
(1) VAS Back pain; range 0 [no pain] to 100 [worst pain ever]; number of patients from author request; 5 years

**Figure 14 PICO 3 long-term: Back pain**

#### 4.2.3.3.2 Radicular pain

One study (Davis 2013) reported on radicular pain at a 5-year follow-up. The study measured radicular pain on a Visual Analogue Scale (VAS, range 0 [no pain] to 100 [worst pain ever]). There was no statistically significant difference between interspinous stabilisation and fusion (MD -3.60, 95% CI -11.23 to 4.03; low quality of evidence; Figure 15). Short-term (1 year) results were only reported by Davis 2013 and showed a similar statistically non-significant effect on radicular pain.



| | Experimental | | | Control | | | | Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Fixed, 95% CI | IV, Fixed, 95% CI |
| Davis 2013 (1) | 22.6 | 26.9 | 192 | 26.2 | 32.2 | 91 | 100.0% | -3.60 [-11.23, 4.03] | |
| **Total (95% CI)** | | | **192** | | | **91** | **100.0%** | **-3.60 [-11.23, 4.03]** | |

Heterogeneity: Not applicable
Test for overall effect: Z = 0.92 (P = 0.36)

Footnotes
(1) VAS leg pain; range 0 [no pain] to 100 [worst pain ever]; number of patients from author request; 5 years

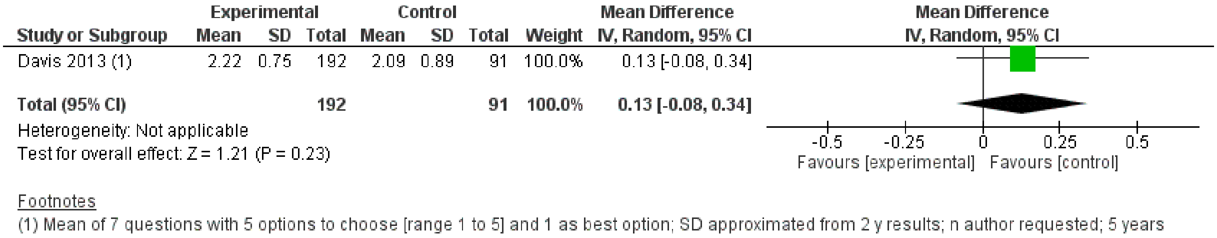
**Figure 15 PICO 3 long-term: Radicular pain**

#### 4.2.3.3.3 Spinal claudication (Walking distance or Zurich Claudication Questionnaire)

No study reported on walking distance. One study (Davis 2013) reported the subscales of the Zurich Claudication Questionnaire (ZCQ), namely symptom severity, physical function and patient satisfaction, at a long-term follow-up of 5 years in 283 participants. The study did not report standard deviations; therefore standard deviations were approximated from group specific SDs reported for 2-year results.

**Zurich Claudication Questionnaire: Symptom severity**

ZCQ Symptom severity is the mean of seven questions with five options to choose [range 1 to 5] with 1 as best option. There was no statistically significant difference between interspinous stabilisation and fusion (MD 0.13, 95% CI -0.08 to 0.34; low quality of evidence). No study reported spinal claudication at short-term (1 year) follow-up.



|  | Experimental |  |  | Control |  |  |  | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Davis 2013 (1) | 2.22 | 0.75 | 192 | 2.09 | 0.89 | 91 | 100.0% | 0.13 [-0.08, 0.34] |
| **Total (95% CI)** |  |  | **192** |  |  | **91** | **100.0%** | **0.13 [-0.08, 0.34]** |

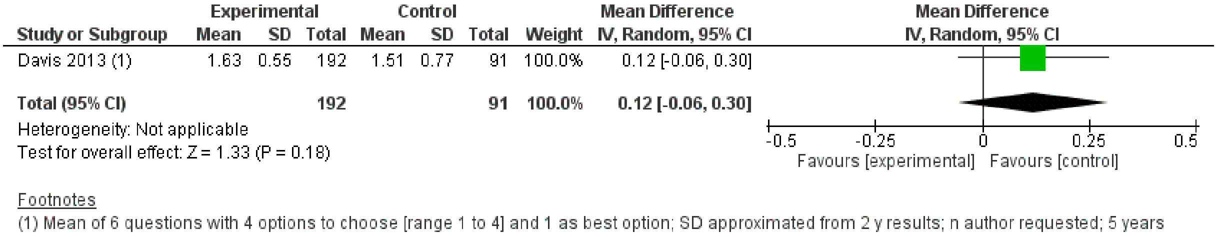
Heterogeneity: Not applicable
Test for overall effect: Z = 1.21 (P = 0.23)

Footnotes
(1) Mean of 7 questions with 5 options to choose [range 1 to 5] and 1 as best option; SD approximated from 2 y results; n author requested; 5 years

Figure 16 PICO 3 long-term: Zurich Claudication Questionnaire, Symptom severity

### Zurich Claudication Questionnaire: Patient satisfaction

ZCQ Patient satisfaction is the mean of six questions with four options to choose [range 1 to 4] and 1 as best option. There was no statistically significant difference between interspinous stabilisation and fusion (MD 0.12, 95% CI -0.06 to 0.30; low quality of evidence). No study reported spinal claudication at short-term (1 year) follow-up.



|  | Experimental |  |  | Control |  |  |  | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Davis 2013 (1) | 1.63 | 0.55 | 192 | 1.51 | 0.77 | 91 | 100.0% | 0.12 [-0.06, 0.30] |
| **Total (95% CI)** |  |  | **192** |  |  | **91** | **100.0%** | **0.12 [-0.06, 0.30]** |

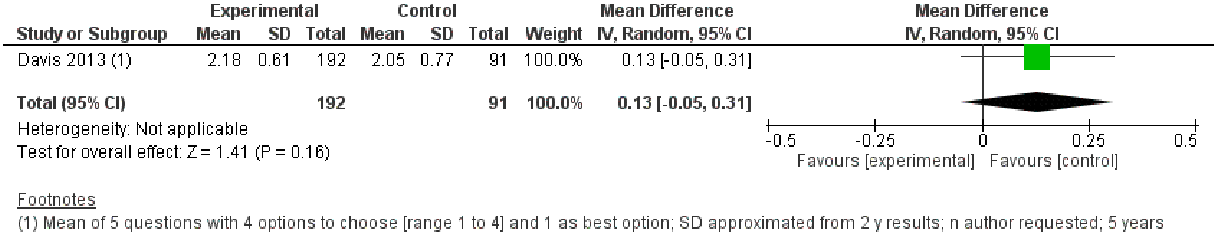
Heterogeneity: Not applicable
Test for overall effect: Z = 1.33 (P = 0.18)

Footnotes
(1) Mean of 6 questions with 4 options to choose [range 1 to 4] and 1 as best option; SD approximated from 2 y results; n author requested; 5 years

Figure 17 PICO 3 long-term: Zurich Claudication Questionnaire, Patient satisfaction

### Zurich Claudication Questionnaire: Physical function

ZCQ Physical function is the mean of five questions with four options to choose [range 1 to 4] and 1 as best option. There was no statistically significant difference between interspinous stabilisation and fusion (MD 0.13, 95% CI -0.05 to 0.31; low quality of evidence). No study reported spinal claudication at short-term (1 year) follow-up.



|  | Experimental |  |  | Control |  |  |  | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Davis 2013 (1) | 2.18 | 0.61 | 192 | 2.05 | 0.77 | 91 | 100.0% | 0.13 [-0.05, 0.31] |
| **Total (95% CI)** |  |  | **192** |  |  | **91** | **100.0%** | **0.13 [-0.05, 0.31]** |

Heterogeneity: Not applicable
Test for overall effect: Z = 1.41 (P = 0.16)

Footnotes
(1) Mean of 5 questions with 4 options to choose [range 1 to 4] and 1 as best option; SD approximated from 2 y results; n author requested; 5 years

Figure 18 PICO 3 long-term: Zurich Claudication Questionnaire, Physical function

### *4.2.3.4 Important outcomes – PICO 3*

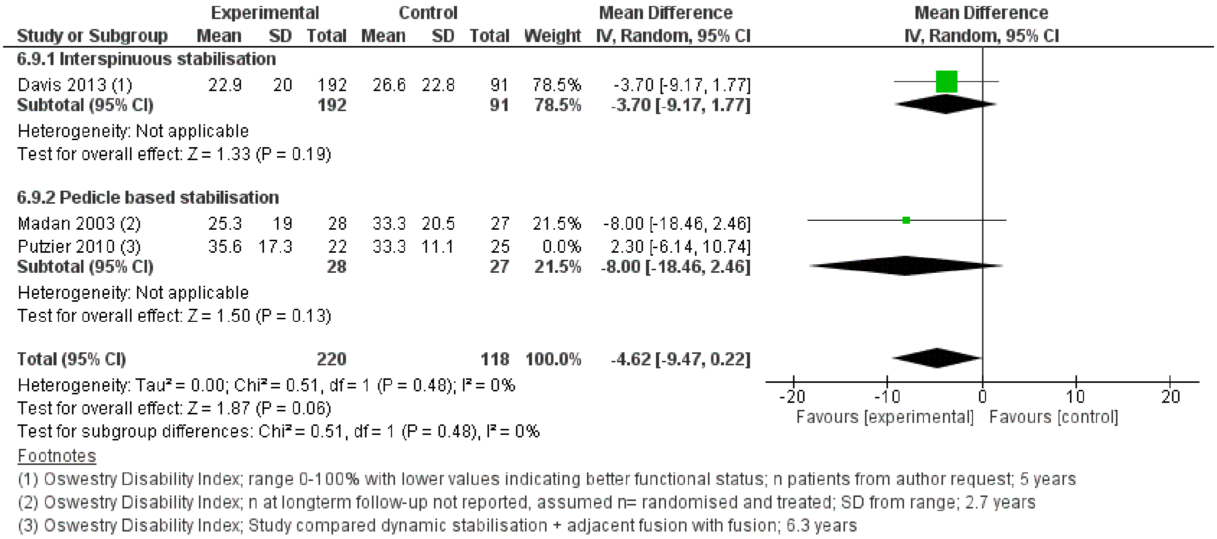
#### 4.2.3.4.1 Quality of life

No study reported on quality of life.

#### 4.2.3.4.2 Function

Two studies (Davis 2013, Madan 2003) reported long-term results for function. Both studies measured function with the Oswestry Disability Index (ODI, range 0-100% with lower values indicating better functional status). There was no statistically significant difference between dynamic stabilisation and direct decompression only (MD -4.62, 95% CI -9.47 to 0.22; low quality of evidence; Figure 19). Heterogeneity between studies was low ($I^2$=0%).

The stratification of the results per dynamic stabilisation technology showed no significant differences between the subgroups interspinous and pedicle-based stabilisation compared to fusion (Figure 19). Short-term (1 year) data were reported by Davis 2013 and showed a similar statistically non-significant effect on function.

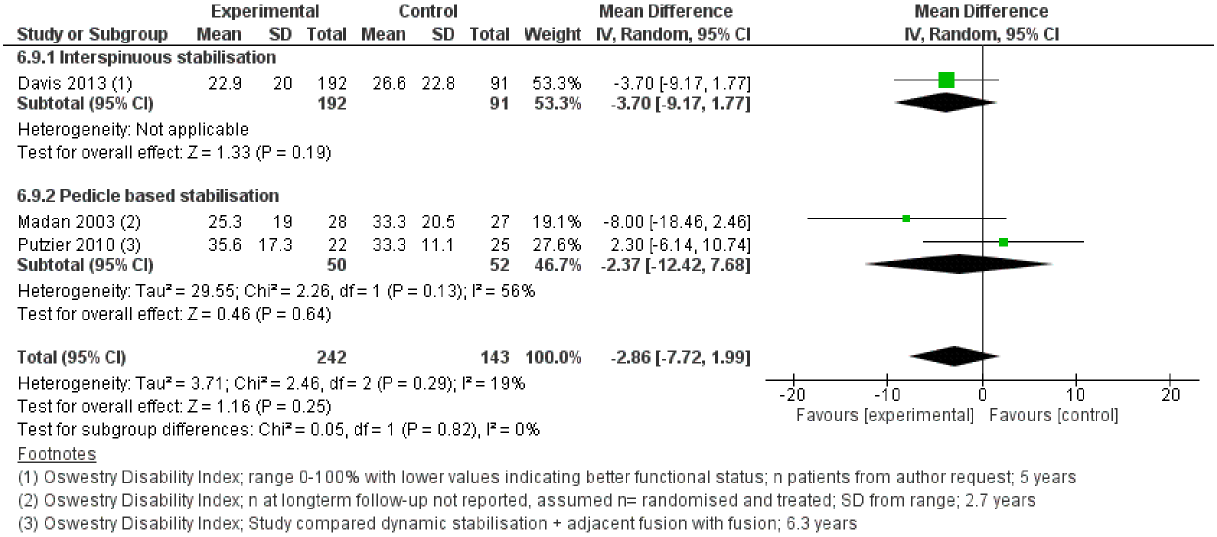


| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| **Study or Subgroup** | **Mean** | **SD** | **Total** | **Mean** | **SD** | **Total** | **Weight** | **IV, Random, 95% CI** |
| **6.9.1 Interspinuous stabilisation** | | | | | | | | |
| Davis 2013 (1) | 22.9 | 20 | 192 | 26.6 | 22.8 | 91 | 78.5% | -3.70 [-9.17, 1.77] |
| **Subtotal (95% CI)** | | | **192** | | | **91** | **78.5%** | **-3.70 [-9.17, 1.77]** |
| Heterogeneity: Not applicable | | | | | | | | |
| Test for overall effect: Z = 1.33 (P = 0.19) | | | | | | | | |
| **6.9.2 Pedicle based stabilisation** | | | | | | | | |
| Madan 2003 (2) | 25.3 | 19 | 28 | 33.3 | 20.5 | 27 | 21.5% | -8.00 [-18.46, 2.46] |
| Putzier 2010 (3) | 35.6 | 17.3 | 22 | 33.3 | 11.1 | 25 | 0.0% | 2.30 [-6.14, 10.74] |
| **Subtotal (95% CI)** | | | **28** | | | **27** | **21.5%** | **-8.00 [-18.46, 2.46]** |
| Heterogeneity: Not applicable | | | | | | | | |
| Test for overall effect: Z = 1.50 (P = 0.13) | | | | | | | | |
| **Total (95% CI)** | | | **220** | | | **118** | **100.0%** | **-4.62 [-9.47, 0.22]** |
| Heterogeneity: Tau² = 0.00; Chi² = 0.51, df = 1 (P = 0.48); I² = 0% | | | | | | | | |
| Test for overall effect: Z = 1.87 (P = 0.06) | | | | | | | | |
| Test for subgroup differences: Chi² = 0.51, df = 1 (P = 0.48), I² = 0% | | | | | | | | |

Footnotes
(1) Oswestry Disability Index; range 0-100% with lower values indicating better functional status; n patients from author request; 5 years
(2) Oswestry Disability Index; n at longterm follow-up not reported, assumed n= randomised and treated; SD from range; 2.7 years
(3) Oswestry Disability Index; Study compared dynamic stabilisation + adjacent fusion with fusion; 6.3 years

Figure 19 PICO 3 long-term: Function

An additional study was considered. Putzier 2010, compared the pedicle-based stabilisation combined with a fusion for the adjacent level to fusion only. It is arguable whether this comparison fits the inclusion criteria of the present report. Hence, this study was formally included in the report based on extended inclusion criteria, but was only pooled with the other studies in a sensitivity analysis. Putzier 2010 reported extractable data only for the outcome function. Putzier 2010 was not considered for the assessment of the quality of the evidence. In the sensitivity analysis including Putzier 2010 yielded similar results (MD -2.86, 95% CI -7.72 to 1.99) and low heterogeneity ($I^2$=19%) (Figure 20).



| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| **6.9.1 Interspinuous stabilisation** | | | | | | | | |
| Davis 2013 (1) | 22.9 | 20 | 192 | 26.6 | 22.8 | 91 | 53.3% | -3.70 [-9.17, 1.77] |
| **Subtotal (95% CI)** | | | **192** | | | **91** | **53.3%** | **-3.70 [-9.17, 1.77]** |
| Heterogeneity: Not applicable | | | | | | | | |
| Test for overall effect: Z = 1.33 (P = 0.19) | | | | | | | | |
| **6.9.2 Pedicle based stabilisation** | | | | | | | | |
| Madan 2003 (2) | 25.3 | 19 | 28 | 33.3 | 20.5 | 27 | 19.1% | -8.00 [-18.46, 2.46] |
| Putzier 2010 (3) | 35.6 | 17.3 | 22 | 33.3 | 11.1 | 25 | 27.6% | 2.30 [-6.14, 10.74] |
| **Subtotal (95% CI)** | | | **50** | | | **52** | **46.7%** | **-2.37 [-12.42, 7.68]** |
| Heterogeneity: Tau² = 29.55; Chi² = 2.26, df = 1 (P = 0.13); I² = 56% | | | | | | | | |
| Test for overall effect: Z = 0.46 (P = 0.64) | | | | | | | | |
| **Total (95% CI)** | | | **242** | | | **143** | **100.0%** | **-2.86 [-7.72, 1.99]** |

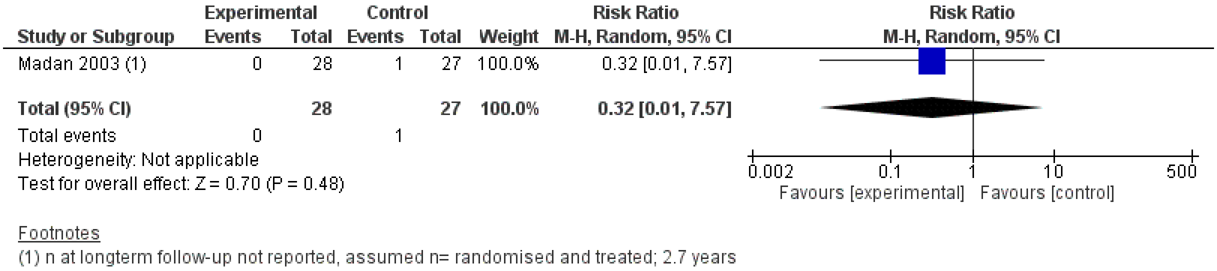
Heterogeneity: Tau² = 3.71; Chi² = 2.46, df = 2 (P = 0.29); I² = 19%
Test for overall effect: Z = 1.16 (P = 0.25)
Test for subgroup differences: Chi² = 0.05, df = 1 (P = 0.82), I² = 0%

Footnotes
(1) Oswestry Disability Index; range 0-100% with lower values indicating better functional status; n patients from author request; 5 years
(2) Oswestry Disability Index; n at longterm follow-up not reported, assumed n= randomised and treated; SD from range; 2.7 years
(3) Oswestry Disability Index; Study compared dynamic stabilisation + adjacent fusion with fusion; 6.3 years

**Figure 20 PICO 3 long-term: Function - Sensitivity analysis including Putzier 2010 which did not fit PICO-question**

#### 4.2.3.4.3 Revision rate

Revision rate was reported in one study (Madan 2003) at a follow-up of 2.7 years. There was no difference between pedicle-based stabilisation and fusion (RR 0.32, 95% CI 0.01 to 7.57; very low quality of evidence; Figure 21). No study reported revision rate at short-term (1 year) follow-up.



| | Experimental | | Control | | | Risk Ratio |
|---|---|---|---|---|---|---|
| Study or Subgroup | Events | Total | Events | Total | Weight | M-H, Random, 95% CI |
| Madan 2003 (1) | 0 | 28 | 1 | 27 | 100.0% | 0.32 [0.01, 7.57] |
| **Total (95% CI)** | | **28** | | **27** | **100.0%** | **0.32 [0.01, 7.57]** |
| Total events | 0 | | 1 | | | |

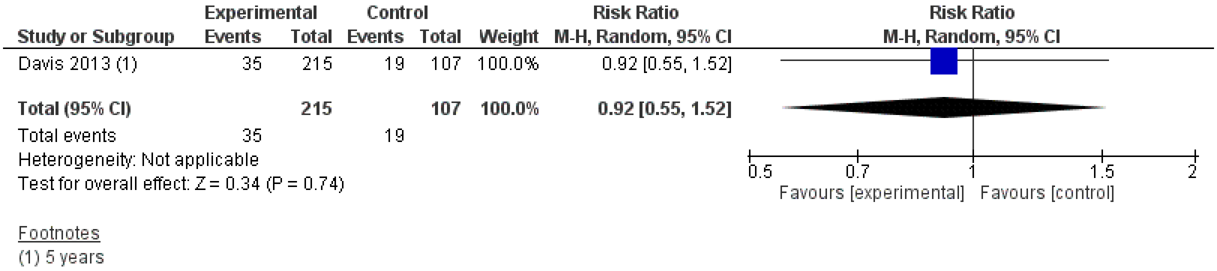
Heterogeneity: Not applicable
Test for overall effect: Z = 0.70 (P = 0.48)

Footnotes
(1) n at longterm follow-up not reported, assumed n= randomised and treated; 2.7 years

**Figure 21 PICO 3 long-term: Revision rate**

#### 4.2.3.4.4 Reoperation rate

Reoperation rate was reported in one study (Davis 2013) at a follow-up of 5 years. There was no statistically significant difference between interspinous stabilisation and fusion (RR 0.92, 95% CI 0.55 to 1.52; very low quality of evidence; Figure 22). No study reported reoperation rate at short-term (1 year) follow-up.



| | Experimental | | Control | | | Risk Ratio |
|---|---|---|---|---|---|---|
| Study or Subgroup | Events | Total | Events | Total | Weight | M-H, Random, 95% CI |
| Davis 2013 (1) | 35 | 215 | 19 | 107 | 100.0% | 0.92 [0.55, 1.52] |
| **Total (95% CI)** | | **215** | | **107** | **100.0%** | **0.92 [0.55, 1.52]** |
| Total events | 35 | | 19 | | | |

Heterogeneity: Not applicable
Test for overall effect: Z = 0.34 (P = 0.74)

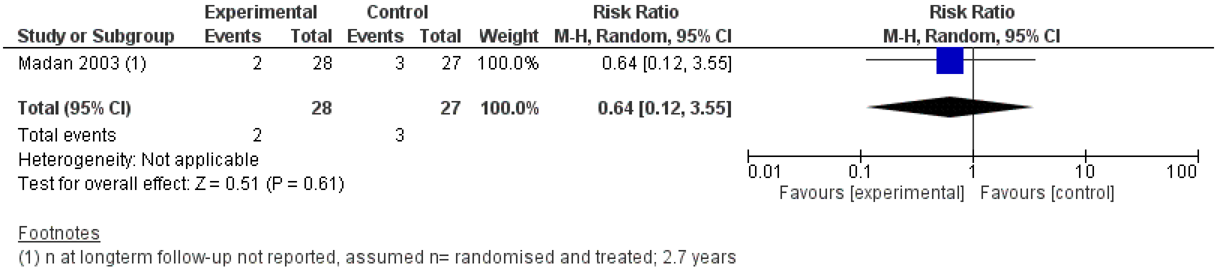
Footnotes
(1) 5 years

**Figure 22 PICO 3 long-term: Reoperation rate**

#### 4.2.3.4.5 Complications or adverse events

Complications were reported in one study (Madan 2003) at a follow-up of 5 years. There was no statistically significant difference between pedicle-based stabilisation and fusion (RR 0.64, 95% CI

0.12 to 3.55; very low quality of evidence; Figure 23). No study reported complications or adverse events at short-term (1 year) follow-up.



| | Experimental | | Control | | | Risk Ratio |
|---|---|---|---|---|---|---|
| Study or Subgroup | Events | Total | Events | Total | Weight | M-H, Random, 95% CI |
| Madan 2003 (1) | 2 | 28 | 3 | 27 | 100.0% | 0.64 [0.12, 3.55] |
| **Total (95% CI)** | | **28** | | **27** | **100.0%** | **0.64 [0.12, 3.55]** |
| Total events | 2 | | 3 | | | |

Heterogeneity: Not applicable
Test for overall effect: Z = 0.51 (P = 0.61)

Footnotes
(1) n at longterm follow-up not reported, assumed n= randomised and treated; 2.7 years

Figure 23 PICO 3 long-term: Complications or adverse events

#### 4.2.3.4.6 Serious adverse events

No study reported on serious adverse events.

### *4.2.3.5 Subgroup analyses – PICO 3*

Subgroup analyses were not conducted due to the small number of studies. The above results were stratified per interspinous or pedicle-based stabilisation – which was only possible for the outcome function (Section 4.2.3.4.2).

### 4.2.3.6 GRADE – PICO 3

Table 17 PICO 3 long-term: Summary of findings (GRADE)

**PICO 3 - Dynamic stabilisation (with or without decompression) compared to fusion with an implant (with or without decompression) for patients with neurological symptoms due to degenerative changes of the lumbar spine**

**Patient or population:** patients with low back pain with or without neurological symptoms due to degenerative changes of the lumbar spine
**Intervention:** Dynamic stabilisation (with or without decompression)
**Comparison:** fusion with an implant (with or without decompression)

| **Outcomes** | **Illustrative comparative risks* (95% CI)** | | **Relative effect (95% CI)** | **No of Participants (studies)** | **Quality of the evidence (GRADE)** | **Comments** |
|---|---|---|---|---|---|---|
| | Assumed risk | Corresponding risk | | | | |
| | **Fusion with an implant (with or without decompression)** | **PICO 3 - Dynamic stabilisation (with or without decompression)** | | | | |
| **Back pain**<br>VAS | | The mean back pain in the intervention groups was<br>**5.30 lower**<br>(12.80 lower to 2.20 higher) | | 283<br>(1 study) | ⊕⊕⊖⊖<br>**low**[1,2] | |
| **Radicular Pain** | | The mean radicular pain in the intervention groups was<br>**3.60 lower**<br>(11.23 lower to 4.03 higher) | | 283<br>(1 study) | ⊕⊕⊖⊖<br>**low**[1,2] | |
| **Quality of life**<br>Euroqol, SF-36 etc | **Study population** | | Not estimable | 0<br>(0) | See comment | No RCT reported this outcome |
| | See comment | See comment | | | | |
| | **Moderate** | | | | | |
| **Function**<br>ODI | | The mean function in the intervention groups was<br>**4.62 lower**<br>(9.47 lower to 0.22 higher) | | 338<br>(2 studies) | ⊕⊕⊖⊖<br>**low**[3,4] | Assessment of quality of evidence is based on only 2 studies. Putzier 2010 was only included in sensitivity analysis. |
| **Revision rate** | **37 per 1000** | **12 per 1000** | **RR 0.32** | 55 | ⊕⊖⊖⊖ | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | (0 to 280) | (0.01 to 7.57) | (1 study) | **very low**[5,6] | |
| **Reoperation rate** | **Study population** | | **RR 0.92**<br>(0.55 to 1.52) | 322<br>(1 study) | ⊕⊖⊖⊖<br>**very low**[1,7] | |
| | **178 per 1000** | **163 per 1000**<br>(98 to 270) | | | | |
| | **Moderate** | | | | | |
| **Complication rate and adverse events** | **Study population** | | **RR 0.64**<br>(0.12 to 3.55) | 55<br>(1 study) | ⊕⊖⊖⊖<br>**very low**[5,6] | |
| | **111 per 1000** | **71 per 1000**<br>(13 to 394) | | | | |
| | **Moderate** | | | | | |
| **Serious adverse events** | **Study population** | | Not estimable | 0<br>(0) | See comment | No RCT reported this outcome |
| | See comment | See comment | | | | |
| | **Moderate** | | | | | |
| **Spinal claudication, ZCQ-Symptom severity**<br>ZCQ | | The mean spinal claudication, ZCQ-symptom severity in the intervention groups was<br>**0.1 higher**<br>(0.08 lower to 0.34 higher) | | 283<br>(1 study) | ⊕⊕⊖⊖<br>**low**[1,8] | |
| **Spinal claudication, ZCQ-Physical function**<br>ZCQ | | The mean spinal claudication, ZCQ-physical function in the intervention groups was<br>**0.13 higher**<br>(0.05 lower to 0.31 higher) | | 283<br>(1 study) | ⊕⊕⊖⊖<br>**low**[1,8] | |
| **Spinal claudication, ZCQ-Satisfaction**<br>ZCQ | | The mean spinal claudication, ZCQ-satisfaction in the intervention groups was<br>**0.12 higher**<br>(0.06 to 0.30 higher) | | 283<br>(1 study) | ⊕⊕⊖⊖<br>**low**[1,8] | |

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).

**CI:** Confidence interval; **RR:** Risk ratio;

GRADE Working Group grades of evidence
**High quality:** Further research is very unlikely to change our confidence in the estimate of effect.

**Moderate quality:** Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.

**Low quality:** Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.

**Very low quality:** We are very uncertain about the estimate.

[1] Risk of performance bias and attrition bias were high, and risk of detection bias and reporting bias were unclear.

[2] Imprecision was downgraded by one level because the 95% CI of the effect estimate included the possibility for a clinically relevant benefit assuming a MCID of 7.2 or no effect.

[3] Risk of performance bias was unclear in one and high in one study; risk of detection bias was unclear in two studies; risk of attrition bias was unclear in one and high in one study; risk of reporting bias was unclear in two studies.

[4] Imprecision was downgraded by one level because the 95% CI of the effect estimate included the possibility for a clinically relevant benefit assuming a MCID of 7.5 or no effect.

[5] Risk of performance bias and attrition bias were high; risk of detection bias was unclear.

[6] Imprecision was downgraded by two levels because the 95% CI included appreciable harm (greater than 25% relative risk increase) and benefit (greater than 25% relative risk reduction) and because total sample size and event rate were low.

[7] Imprecision was downgraded by two levels because the 95% CI appreciable harm (greater than 25% relative risk increase) and benefit (greater than 25% relative risk reduction).

[8] Although the 95% CI of the effect estimate was narrow and excluded clinically important harm or benefit (assuming an MCID of 0.5), imprecision was downgraded by one level because the total sample size was lower than 400 (rule-of-thumb).

## 4.3 Disc prostheses – PICO 4-5

### 4.3.1 Results – PICO 4

The following sections show first the study characteristics and risk of bias assessment and then the results for each outcome including GRADE results for PICO 4. Six relevant studies have been identified. Four studies (Blumenthal 2005, Gornet 2011, Sasso 2008, Zigler 2007; see Table 18 for references) reported long-term and short-term results, one study (Moreno 2008) reported results only for long-term and one (Strube 2016) only at short-term. Only the pooled long-term results are presented here. An overview of the outcomes analysed from each study is given in Table 19. The pooled short-term results (1-year follow-up) for PICO 4 are presented in Appendix VI.

Table 18 Overview of included studies and their unique study IDs – PICO 4

| Study ID | Reference (main reference highlighted in colour) |
|---|---|
| **Blumenthal 2005** | Blumenthal S, McAfee PC, Guyer RD, et al. A prospective, randomized, multicenter Food and Drug Administration investigational device exemptions study of lumbar total disc replacement with the CHARITE artificial disc versus lumbar fusion: part I: evaluation of clinical outcomes. *Spine (Phila Pa 1976).* 2005;30(14):1565-1575; discussion E1387-1591. |
| | Geisler FH, Blumenthal SL, Guyer RD, et al. Neurological complications of lumbar artificial disc replacement and comparison of clinical results with those related to lumbar arthrodesis in the literature: results of a multicenter, prospective, randomized investigational device exemption study of Charite intervertebral disc. Invited submission from the Joint Section Meeting on Disorders of the Spine and Peripheral Nerves, March 2004. *J Neurosurg Spine.* 2004;1(2):143-154. |
| | Guyer RD, McAfee PC, Banco RJ, et al. Prospective, randomized, multicenter Food and Drug Administration investigational device exemption study of lumbar total disc replacement with the CHARITE artificial disc versus lumbar fusion: five-year follow-up. *Spine J.* 2009;9(5):374-386. |
| | Holt RT, Majd ME, Isaza JE, et al. Complications of Lumbar Artificial Disc Replacement Compared to Fusion: Results From the Prospective, Randomized, Multicenter US Food and Drug Administration Investigational Device Exemption Study of the Charite Artificial Disc. *SAS J.* 2007;1(1):20-27. |
| | McAfee PC, Cunningham B, Holsapple G, et al. A prospective, randomized, multicenter Food and Drug Administration investigational device exemption study of lumbar total disc replacement with the CHARITE artificial disc versus lumbar fusion: part II: evaluation of radiographic outcomes and correlation of surgical technique accuracy with clinical outcomes. *Spine (Phila Pa 1976).* 2005;30(14):1576-1583; discussion E1388-1590. |
| | Guyer RD, McAfee PC, Hochschuler SH, et al. Prospective randomized study of the Charite artificial disc: data from two investigational centers. *Spine J.* 2004;4(6 Suppl):252S-259S. |
| | McAfee PC, Geisler FH, Saiedy SS, et al. Revisability of the CHARITE artificial disc replacement: analysis of 688 patients enrolled in the U.S. IDE study of the CHARITE Artificial Disc. *Spine (Phila Pa 1976).* 2006;31(11):1217-1226. |
| **Gornet 2011** | Gornet MF, Burkus JK, Dryer RF, Peloza JH. Lumbar disc arthroplasty with Maverick disc versus stand-alone interbody fusion: a prospective, randomized, controlled, multicenter investigational device exemption trial. *Spine (Phila Pa 1976).* 2011;36(25):E1600-1611. |

| | |
|---|---|
| **Moreno 2008** | Moreno P, Boulot J. [Comparative study of short-term results between total artificial disc prosthesis and anterior lumbar interbody fusion]. *Rev Chir Orthop Reparatrice Appar Mot.* 2008;94(3):282-288. |
| **Sasso 2008** | Sasso RC, Foulk DM, Hahn M. Prospective, randomized trial of metal-on-metal artificial lumbar disc replacement: initial results for treatment of discogenic pain. *Spine (Phila Pa 1976).* 2008;33(2):123-131. |
| **Strube 2016** | Strube P, Putzier M, Streitparth F, Hoff EK, Hartwig T. Postoperative posterior lumbar muscle changes and their relationship to segmental motion preservation or restriction: a randomized prospective study. *J Neurosurg Spine.* 2016;24(1):25-31. |
| **Zigler 2007** | Zigler J, Delamarter R, Spivak JM, et al. Results of the prospective, randomized, multicenter Food and Drug Administration investigational device exemption study of the ProDisc-L total disc replacement versus circumferential fusion for the treatment of 1-level degenerative disc disease. *Spine (Phila Pa 1976).* 2007;32(11):1155-1162; discussion 1163. |
| | Zigler JE. Five-Year Results of the ProDisc-L Multicenter, Prospective, Randomized, Controlled Trial Comparing ProDisc-L With Circumferential Spinal Fusion for Single-Level Disabling Degenerative Disk Disease. *Seminars in Spine Surgery.* 2012;24(1):25-31. |
| | Zigler JE, Delamarter RB. Five-year results of the prospective, randomized, multicenter, Food and Drug Administration investigational device exemption study of the ProDisc-L total disc replacement versus circumferential arthrodesis for the treatment of single-level degenerative disc disease. *J Neurosurg Spine.* 2012;17(6):493-501. |
| | Zigler JE, Glenn J, Delamarter RB. Five-year adjacent-level degenerative changes in patients with single-level disease treated using lumbar total disc replacement with ProDisc-L versus circumferential fusion. *J Neurosurg Spine.* 2012;17(6):504-511. |

Table 19 Overview of the outcomes analysed for PICO 4

| | Back pain | Radicular pain | Spinal claudication | Myelopathy | Quality of life | Function | Revision rate | Reoperation rate | Complications/AE | Serious AE |
|---|---|---|---|---|---|---|---|---|---|---|
| Blumenthal 2005 | | | | | | 2 | | 2 | 2 | |
| Gornet 2011 | 2 | 2 | | | 2 | 2 | 2 | 2 | 2 | 2 |
| Moreno 2008 | | | | | | 2 | | | 2 | |
| Sasso 2008 | | | | | | 2 | | | | 2 |
| Strube 2016 | | | | | | | | | | |
| Zigler 2007 | | | | | 5 | 5 | | | 5 | |

The number in the fields denote the analysed long-term follow-up in years

### 4.3.1.1 Characteristics of the included studies – PICO 4

General characteristics of studies for PICO-question 4 are summarised in Table 20. Four of the six included RCTs were multicentre studies from the USA, the other two RCTs were likely single centre studies from Europe. The enrolment periods range from 2000 to 2005 with the longest follow-up at 5 years. The 1275 participants had degenerative disc disease and in 4 of 6 studies it was unclear how many patients had neurologic symptoms. In two (Blumenthal 2005, Morena 2008) of six studies, participants with radiculopathy were excluded. Participants' mean age ranged from 36 to 48 years. Disc prostheses were compared to anterior lumbar interbody fusion with cage and bone in 4 RCTs and to circumferential fusion with cage and/or bone in 2 RCTs.

Table 20 Study characteristics, PICO 4

| Study ID | Country<br>Setting | Enrollment period<br>Maximum FU | Population<br>Key condition | Intervention<br>o n randomised<br>o Male n (%)<br>o Mean age (SD) | Comparator<br>o n randomised<br>o Male n (%)<br>o Mean age (SD) |
|---|---|---|---|---|---|
| Blumenthal 2005 | USA<br>Multicentre (14 sites) | May 2000-Apr2002<br>2 years | DDD without nerve root compression (radiculopathy) | Disc prosthesis (Charité)<br>○ 205<br>○ 113 (55%)<br>○ 39.6 (8.16) years | Anterior lumbar interbody fusion with cage and bone<br>○ 99<br>○ 44 (44%)<br>○ 39.6 (9.07) years |
| Gornet 2011 | USA<br>Multicentre (31 sites) | April 2003 - Aug 2004<br>2 years | DDD with or without leg pain | Disc prosthesis (Maverick)<br>○ n.r. (405 treated)<br>○ 205 (50.6%)<br>○ 39.9 (18–70)* years | Anterior lumbar interbody fusion with cage and bone<br>○ n.r. (172 treated)<br>○ 86 (50.0%)<br>○ 40.2 (18–65)* years |
| Moreno 2008 | France<br>n.r. | Sep 2002-Apr 2005**<br>2 years | DDD without radiculopathy | Disc prosthesis (Charité III)<br>○ 14<br>○ 7 (50%)<br>○ 39 (33-53)* years | Anterior lumbar interbody fusion with cage and bone<br>○ 18<br>○ 11 (61%)<br>○ 44 (33-55)* years |
| Sasso 2008 | n.r. (likely USA)<br>Part of a multicentre study (2 sites) | n.r.<br>2 years | Discogenic pain due to DDD (greater percentage of axial than radicular pain) | Disc prosthesis (FlexiCore)<br>○ 50<br>○ 23/44 (52%)<br>○ 36 years | Circumferential fusion with cage and bone<br>○ 26<br>○ 10/23 (43%)<br>○ 41 years |
| Strube 2016 | Germany<br>n.r. | n.r.<br>1 year | DDD with or without radicular symptoms | Disc prosthesis (Maverick)<br>○ 25<br>○ 10 (23%)<br>○ 47.3 (35–59)* years | Anterior lumbar interbody fusion with cage and bone<br>○ 25<br>○ 10 (25%)<br>○ 48.4 (38–58)* years |
| Zigler 2007 | USA<br>Multicentre (17 sites) | Oct 2001-Jun 2003<br>5 years | DDD with back or leg pain | Disc prosthesis (ProDisc-L)<br>○ 161<br>○ 82 (51%) | Circumferential fusion with bone<br>○ 75<br>○ 34 (45%) |

| Study ID | Country<br>Setting | Enrollment period<br>Maximum FU | Population<br>Key condition | Intervention<br>○ n randomised<br>○ Male n (%)<br>○ Mean age (SD) | Comparator<br>○ n randomised<br>○ Male n (%)<br>○ Mean age (SD) |
|---|---|---|---|---|---|
| | | | | ○ 38.7 (8.0) years | ○ 40.4 (7.6) years |

Note: Under key condition the presence or absence of neurological symptoms is being described according to the inclusion criteria. If additional information based on the baseline characteristics would have been available this would have been documented in footnotes.

* mean (range);

** Treatment period

Abbreviations: DDD, degenerative disc disease; FU, follow-up; n, number; n.r., not reported; SD, standard deviation

#### *4.3.1.2 Risk of bias – PICO 4*

The method of random sequence generation was adequate in one study (Blumenthal 2005) and unclear in four studies (Gornet 2011, Moreno 2008, Sasso 2008, Zigler 2007). The risk of selection bias depending on allocation concealment was rated as low for three studies (Blumenthal 2005, Gornet 2011, Zigler 2007) and as unclear for two studies (Moreno 2008, Sasso 2008). The risk of performance bias depending on the blinding of participants and personnel was rated to be high in three studies (Blumenthal 2005, Gornet 2011, Zigler 2007) and to be unclear in two studies (Moreno 2008, Sasso 2008). Blinding of outcome assessment was unclear in three studies (Moreno 2008, Sasso 2008, Zigler 2007) and high in two studies (Blumenthal 2005, Gornet 2011). Attrition bias for continuous outcome data was rated as high in four studies (Blumenthal 2005, Gornet 2011, Sasso 2008, Zigler 2007) and unclear in one study (Moreno 2008). Attrition bias for binary outcome data was rated as low in two studies (Blumenthal 2005, Gornet 2011), unclear in two studies (Moreno 2008, Sasso 2008) and high in one study (Zigler 2007). Selective reporting was low in three studies (Gornet 2011, Sasso 2008, Zigler 2007) and was rated as unclear in two studies (Blumenthal 2005, Moreno 2008). A summarised overview of the risk of bias assessment is shown in Table 21 and a detailed summary with support of judgement in Appendix III.

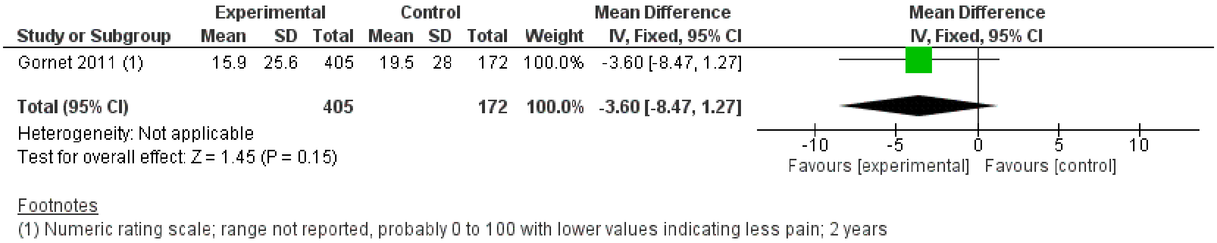
Table 21 Risk of bias, PICO 4

| Study | Random sequence generation (selection bias) | Allocation concealment (selection bias) | Blinding of participants and personnel (performance bias) | Blinding of outcome assessment, judgement did not differ among outcomes (detection bias) | Incomplete continuous outcome data (attrition bias) | Incomplete binary data (attrition bias) | Selective reporting (reporting bias) |
|---|---|---|---|---|---|---|---|
| Blumenthal 2005 | Low | Low | High | High | High | Low | Unclear |
| Gornet 2011 | Unclear | Low | High | High | High | Low | Low |
| Moreno 2008 | Unclear | Unclear | Unclear | Unclear | Unclear | Unclear | Unclear |
| Sasso 2008 | Unclear | Unclear | Unclear | Unclear | High | Unclear | Low |
| Zigler 2007 | Unclear | Low | High | Unclear | High | High | Low |

### *4.3.1.3 Critical outcomes– PICO 4*

#### 4.3.1.3.1 Radicular pain

One study (Gornet 2011) reported on leg pain at a 2-year follow-up. The study measured leg pain with a numeric rating scale (NRS, range was not reported, probably 0 to 100 with lower values indicating less pain). There was no statistically significant difference between disc prostheses and fusion (MD -3.60, 95% CI -8.47 to 1.27; low quality of evidence; Figure 24). Short-term (1 year) effect of radicular pain was only reported by Gornet 2011 and showed a similar effect, although the effect at short-term follow-up was statistically significant.



| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Fixed, 95% CI |
| Gornet 2011 (1) | 15.9 | 25.6 | 405 | 19.5 | 28 | 172 | 100.0% | -3.60 [-8.47, 1.27] |
| **Total (95% CI)** | | | **405** | | | **172** | **100.0%** | **-3.60 [-8.47, 1.27]** |

Heterogeneity: Not applicable
Test for overall effect: Z = 1.45 (P = 0.15)

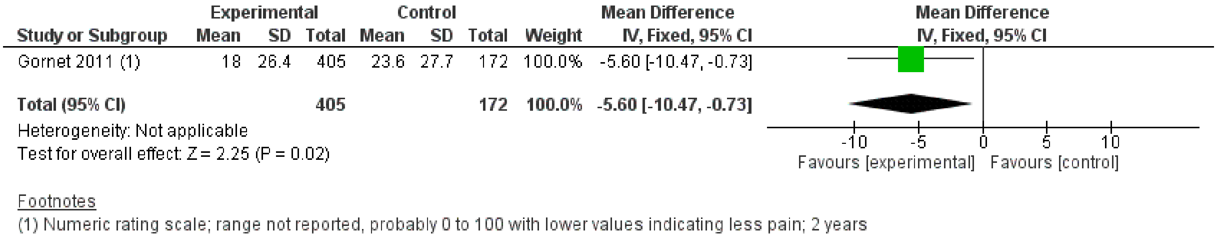
Footnotes
(1) Numeric rating scale; range not reported, probably 0 to 100 with lower values indicating less pain; 2 years

**Figure 24 PICO 4 long-term: Leg pain**

### *4.3.1.4 Important outcomes – PICO 4*

#### 4.3.1.4.1 Back pain

One study (Gornet 2011) reported on back pain at a 2-year follow-up. The study measured back pain with a numeric rating scale (NRS, range was not reported, probably 0 to 100 with lower values indicating less pain). Disc prostheses had a significantly greater reduction in back pain compared to fusion (SMD -5.60, 95% CI -10.47 to -0.73); low quality of evidence; Figure 25). At short-term (1 year) follow-up, back pain was reported by two studies and showed a similar statistically significant effect as at long-term follow-up.



| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Fixed, 95% CI |
| Gornet 2011 (1) | 18 | 26.4 | 405 | 23.6 | 27.7 | 172 | 100.0% | -5.60 [-10.47, -0.73] |
| **Total (95% CI)** | | | **405** | | | **172** | **100.0%** | **-5.60 [-10.47, -0.73]** |

Heterogeneity: Not applicable
Test for overall effect: Z = 2.25 (P = 0.02)

Footnotes
(1) Numeric rating scale; range not reported, probably 0 to 100 with lower values indicating less pain; 2 years

**Figure 25 PICO 4 long-term: Back pain**

#### 4.3.1.4.2 Quality of life

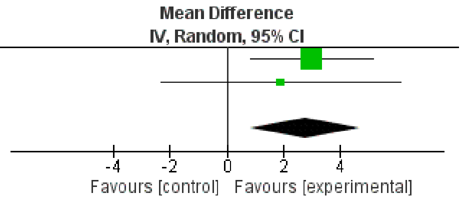
Two studies (Gornet 2011, Zigler 2007) reported on quality of life measured with the SF-36. One study reported the physical component score and mental component scores of SF-36 after 2 years and the other study reported only the physical component score after 5 years of follow-up. The SF-36 scales range from 1 to 100 with higher value indicating better quality of life.

**Short form 36: physical component score**

Two studies (Gornet 2011, Zigler) reported the SF-36 physical component score. Disc prostheses had a statistically significantly greater effect on quality of life than fusion (MD 2.77, 95% CI 0.85 to 4.70; low quality of evidence; Figure 26). Heterogeneity was low ($I^2$=0%). Short-term (1 year) data were only reported by Gornet 2011 and showed a similar statistically significant effect on quality of life physical component score.

| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| **Study or Subgroup** | **Mean** | **SD** | **Total** | **Mean** | **SD** | **Total** | **Weight** | **IV, Random, 95% CI** |
| Gornet 2011 (1) | 45.1 | 12.2 | 405 | 42.1 | 12.1 | 172 | 79.2% | 3.00 [0.84, 5.16] |
| Zigler 2007 (2) | 42 | 11.3 | 126 | 40.1 | 13.6 | 51 | 20.8% | 1.90 [-2.32, 6.12] |
| **Total (95% CI)** | | | **531** | | | **223** | **100.0%** | **2.77 [0.85, 4.70]** |

Heterogeneity: Tau² = 0.00; Chi² = 0.21, df = 1 (P = 0.65); I² = 0%
Test for overall effect: Z = 2.82 (P = 0.005)



Footnotes
(1) SF-36 physical component summary; range from 1 to 100 with higher value indicating better quality of life; 2 years
(2) SF-36 physical component summary; range from 1 to 100 with higher value indicating better quality of life; 5 years

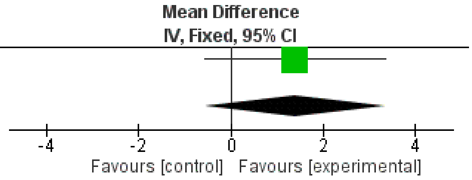
Figure 26 PICO 4 long-term: Quality of life, SF-36 physical component score

### Short form 36: mental component score

One study (Gornet 2011) reported the SF-36 mental component score. There was no significant difference between disc prostheses and fusion (MD 1.40, 95% CI -0.56 to 3.36; very low quality of evidence; Figure 27). Short-term (1 year) data were only reported by Gornet 2011 and showed similar no statistical significant effect on quality of life mental component score.

| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| **Study or Subgroup** | **Mean** | **SD** | **Total** | **Mean** | **SD** | **Total** | **Weight** | **IV, Fixed, 95% CI** |
| Gornet 2011 (1) | 51.4 | 11 | 405 | 50 | 11 | 172 | 100.0% | 1.40 [-0.56, 3.36] |
| **Total (95% CI)** | | | **405** | | | **172** | **100.0%** | **1.40 [-0.56, 3.36]** |

Heterogeneity: Not applicable
Test for overall effect: Z = 1.40 (P = 0.16)



Footnotes
(1) SF-36 mental health summary; range from 1 to 100 with higher value indicating better quality of life; 2 years

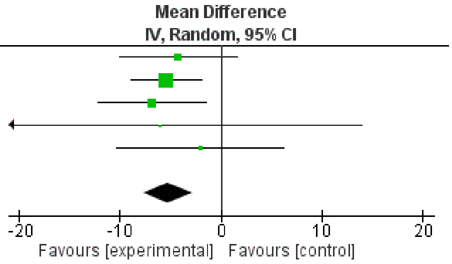
Figure 27 PICO 4 long-term: Quality of life, SF-36 mental component score

### 4.3.1.4.3 Function

Five studies (Blumenthal 2005, Gornet 2011, Moreno 2008, Sasso 2008, Zigler 2007) reported on function after 2 and 5 years of follow-up. All studies stud measured function with the Oswestry Disability Index (ODI, range 0-100% with lower values indicating better functional status). Disc prostheses had a significantly greater effect than fusion (MD -5.19, 95% CI -7.67 to -2.71; moderate quality of evidence; Figure 28). Heterogeneity was low ($I^2$=0%). Short-term (1 year) effect of function showed a similar statistical significant effect.

| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| **Study or Subgroup** | **Mean** | **SD** | **Total** | **Mean** | **SD** | **Total** | **Weight** | **IV, Random, 95% CI** |
| Blumenthal 2005 (1) | 26.3 | 21.7 | 185 | 30.5 | 22.6 | 82 | 18.2% | -4.20 [-10.01, 1.61] |
| Gornet 2011 (2) | 19.4 | 20.2 | 405 | 24.8 | 19.6 | 172 | 49.3% | -5.40 [-8.93, -1.87] |
| Moreno 2008 (3) | 16 | 5.05 | 14 | 22.8 | 9.93 | 18 | 21.9% | -6.80 [-12.10, -1.50] |
| Sasso 2008 (4) | 6 | 20.95 | 11 | 12 | 21.1 | 7 | 1.5% | -6.00 [-25.94, 13.94] |
| Zigler 2007 (5) | 34.2 | 24.3 | 126 | 36.2 | 25.7 | 51 | 9.1% | -2.00 [-10.23, 6.23] |
| **Total (95% CI)** | | | **741** | | | **330** | **100.0%** | **-5.19 [-7.67, -2.71]** |

Heterogeneity: Tau² = 0.00; Chi² = 1.06, df = 4 (P = 0.90); I² = 0%
Test for overall effect: Z = 4.11 (P < 0.0001)



Footnotes
(1) Oswestry Disability Index; range 0-100 with lower values indicating less disability; 2 years
(2) Oswestry Disability Index; range 0-100 with lower values indicating less disability; 2 years
(3) Oswestry Disability Index; range 0-100 with lower values indicating less disability; 2 years
(4) Oswestry Disability Index; range 0-100 with lower values indicating less disability; SD imputed; 2 years
(5) Oswestry Disability Index; range 0-100 with lower values indicating less disability; 5 years

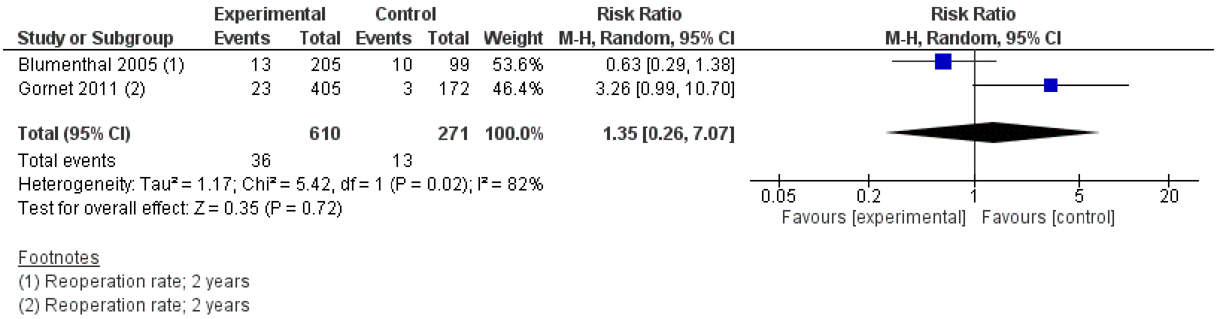
Figure 28 PICO 4 long-term: Function

### 4.3.1.4.4 Revision rate

One study (Gornet 2011) reported revision rates at 2-year follow-up. Zero revisions were reported for both treatment groups. The effect was therefore not estimable. No study reported revision rate at short-term (1 year) follow-up.

### 4.3.1.4.5 Reoperation rate

Two studies (Blumenthal 2005, Gornet 2011) reported reoperation rates at two years follow-up. There was no significant difference between disc prostheses and fusion (RR 1.35, 95% CI 0.26 to 7.07; low quality of evidence; Figure 29). Heterogeneity was high ($I^2$=82%). No study reported reoperation rate at short-term (1 year) follow-up.



| Study or Subgroup | Experimental Events | Experimental Total | Control Events | Control Total | Weight | Risk Ratio M-H, Random, 95% CI |
|---|---|---|---|---|---|---|
| Blumenthal 2005 (1) | 13 | 205 | 10 | 99 | 53.6% | 0.63 [0.29, 1.38] |
| Gornet 2011 (2) | 23 | 405 | 3 | 172 | 46.4% | 3.26 [0.99, 10.70] |
| **Total (95% CI)** | | **610** | | **271** | **100.0%** | **1.35 [0.26, 7.07]** |
| Total events | 36 | | 13 | | | |

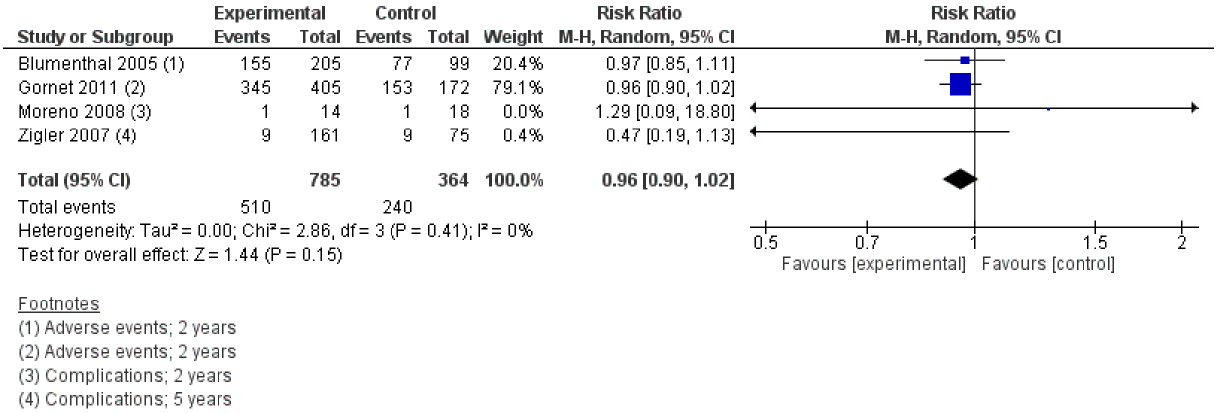
Heterogeneity: Tau² = 1.17; Chi² = 5.42, df = 1 (P = 0.02); I² = 82%
Test for overall effect: Z = 0.35 (P = 0.72)

Footnotes
(1) Reoperation rate; 2 years
(2) Reoperation rate; 2 years

Figure 29 PICO 4 long-term: Reoperation rate

### 4.3.1.4.6 Complications or adverse events

Four studies (Blumenthal 2005, Gornet 2011, Moreno 2008, Zigler 2007) reported on complications or adverse events after 2 and 5 years of follow-up. There was no statistically significant difference between disc prostheses and fusion (RR 0.96, 95% CI 0.90 to 1.02; moderate quality of evidence; Figure 30). Heterogeneity was low ($I^2$=0%). No study reported complications or adverse events at short-term (1 year) follow-up.



| Study or Subgroup | Experimental Events | Experimental Total | Control Events | Control Total | Weight | Risk Ratio M-H, Random, 95% CI |
|---|---|---|---|---|---|---|
| Blumenthal 2005 (1) | 155 | 205 | 77 | 99 | 20.4% | 0.97 [0.85, 1.11] |
| Gornet 2011 (2) | 345 | 405 | 153 | 172 | 79.1% | 0.96 [0.90, 1.02] |
| Moreno 2008 (3) | 1 | 14 | 1 | 18 | 0.0% | 1.29 [0.09, 18.80] |
| Zigler 2007 (4) | 9 | 161 | 9 | 75 | 0.4% | 0.47 [0.19, 1.13] |
| **Total (95% CI)** | | **785** | | **364** | **100.0%** | **0.96 [0.90, 1.02]** |
| Total events | 510 | | 240 | | | |

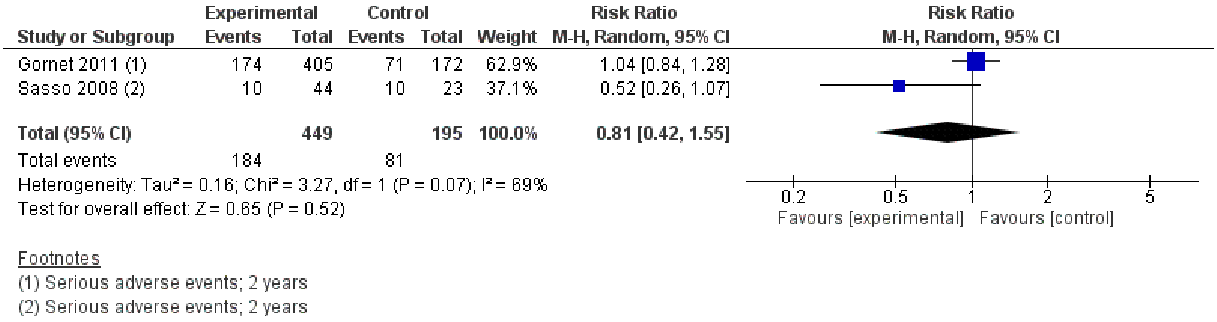
Heterogeneity: Tau² = 0.00; Chi² = 2.86, df = 3 (P = 0.41); I² = 0%
Test for overall effect: Z = 1.44 (P = 0.15)

Footnotes
(1) Adverse events; 2 years
(2) Adverse events; 2 years
(3) Complications; 2 years
(4) Complications; 5 years

Figure 30 PICO 4 long-term: Complications or adverse events

### 4.3.1.4.7 Serious adverse events

Two studies (Gornet 2011, Sasso 2008) reported serious adverse events at a 2-year follow-up. There was no significant difference between disc prostheses and fusion (RR 0.81, 95% CI 0.42 to 1.55; very low quality of evidence; Figure 31). Heterogeneity was high ($I^2$=69%). No study reported serious adverse events rate at short-term (1 year) follow-up.



| Study or Subgroup | Experimental Events | Experimental Total | Control Events | Control Total | Weight | Risk Ratio M-H, Random, 95% CI |
|---|---|---|---|---|---|---|
| Gornet 2011 (1) | 174 | 405 | 71 | 172 | 62.9% | 1.04 [0.84, 1.28] |
| Sasso 2008 (2) | 10 | 44 | 10 | 23 | 37.1% | 0.52 [0.26, 1.07] |
| **Total (95% CI)** | | **449** | | **195** | **100.0%** | **0.81 [0.42, 1.55]** |
| Total events | 184 | | 81 | | | |

Heterogeneity: Tau² = 0.16; Chi² = 3.27, df = 1 (P = 0.07); I² = 69%
Test for overall effect: Z = 0.65 (P = 0.52)

Footnotes
(1) Serious adverse events; 2 years
(2) Serious adverse events; 2 years

**Figure 31 PICO 4 long-term: Serious adverse events**

### *4.3.1.5 Subgroup analyses – PICO 4*

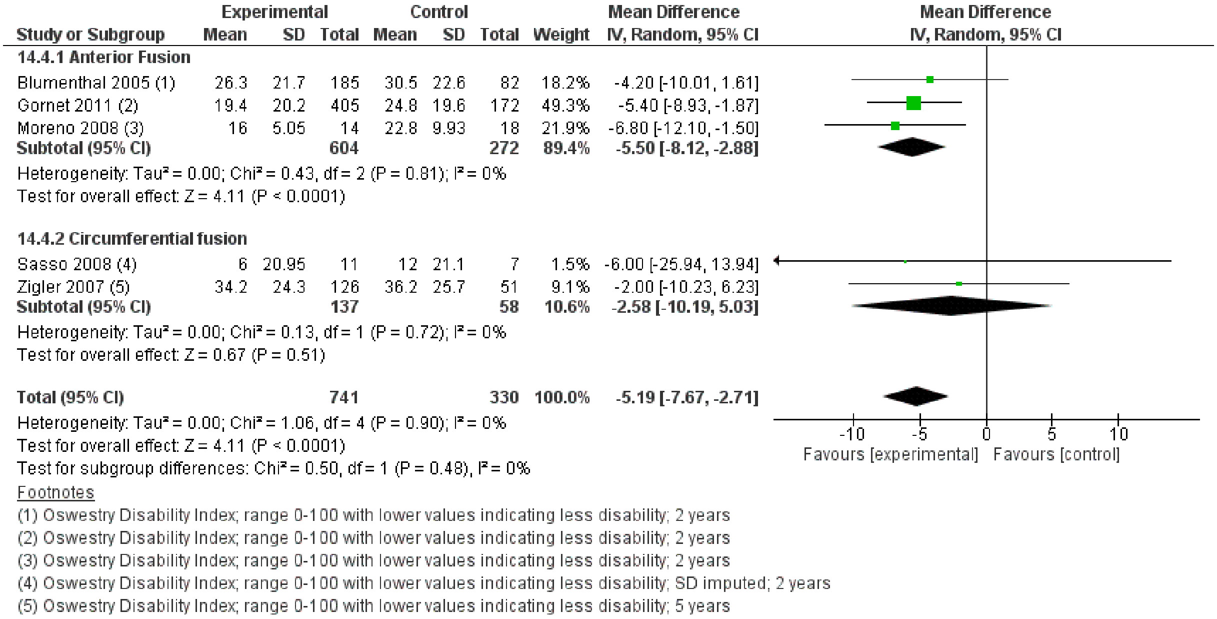
Subgroup and sensitivity analyses were only considered for the outcome function which was the only outcome reported by five studies. Due to the limited number of studies, only one subgroup analysis was possible. Consideration of possible subgroup analyses followed the a priori prioritised sequence.

*Subgroup 1: Patients only with radiculopathy of the lumbar spine vs. patients without neurological symptoms*

Two studies were in patients with no neurological symptoms at baseline (Blumenthal 2005, Moreno 2008) and in three studies the neurologic status was unclear, i.e. per eligibility criteria patients with and without neurological symptoms could be included but no information on their proportion in the included in the trials was provided by the authors (Zigler 2007, Sasso 2008, Gornet 2011). Hence it was not possible to distinguish between studies including patients with neurological symptoms and studies including patients without neurological symptoms and it was not possible to conduct this subgroup analysis.

*Subgroup 2: Anterior fusion vs. posterior fusion vs. other*

Anterior fusion was the comparator in three studies (Blumenthal 2005, Gornet 2011, Moreno 2008). In the other two studies, circumferential fusion was used (Sasso 2008, Zigler 2007) (Figure 32). No statistically significant difference was found between the two subgroups.



|  | Experimental |  |  | Control |  |  |  | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| **14.4.1 Anterior Fusion** |  |  |  |  |  |  |  |  |
| Blumenthal 2005 (1) | 26.3 | 21.7 | 185 | 30.5 | 22.6 | 82 | 18.2% | -4.20 [-10.01, 1.61] |
| Gornet 2011 (2) | 19.4 | 20.2 | 405 | 24.8 | 19.6 | 172 | 49.3% | -5.40 [-8.93, -1.87] |
| Moreno 2008 (3) | 16 | 5.05 | 14 | 22.8 | 9.93 | 18 | 21.9% | -6.80 [-12.10, -1.50] |
| **Subtotal (95% CI)** |  |  | **604** |  |  | **272** | **89.4%** | **-5.50 [-8.12, -2.88]** |
| Heterogeneity: Tau² = 0.00; Chi² = 0.43, df = 2 (P = 0.81); I² = 0% |  |  |  |  |  |  |  |  |
| Test for overall effect: Z = 4.11 (P < 0.0001) |  |  |  |  |  |  |  |  |
| **14.4.2 Circumferential fusion** |  |  |  |  |  |  |  |  |
| Sasso 2008 (4) | 6 | 20.95 | 11 | 12 | 21.1 | 7 | 1.5% | -6.00 [-25.94, 13.94] |
| Zigler 2007 (5) | 34.2 | 24.3 | 126 | 36.2 | 25.7 | 51 | 9.1% | -2.00 [-10.23, 6.23] |
| **Subtotal (95% CI)** |  |  | **137** |  |  | **58** | **10.6%** | **-2.58 [-10.19, 5.03]** |
| Heterogeneity: Tau² = 0.00; Chi² = 0.13, df = 1 (P = 0.72); I² = 0% |  |  |  |  |  |  |  |  |
| Test for overall effect: Z = 0.67 (P = 0.51) |  |  |  |  |  |  |  |  |
| **Total (95% CI)** |  |  | **741** |  |  | **330** | **100.0%** | **-5.19 [-7.67, -2.71]** |

Heterogeneity: Tau² = 0.00; Chi² = 1.06, df = 4 (P = 0.90); I² = 0%

Test for overall effect: Z = 4.11 (P < 0.0001)

Test for subgroup differences: Chi² = 0.50, df = 1 (P = 0.48), I² = 0%

Footnotes

(1) Oswestry Disability Index; range 0-100 with lower values indicating less disability; 2 years

(2) Oswestry Disability Index; range 0-100 with lower values indicating less disability; 2 years

(3) Oswestry Disability Index; range 0-100 with lower values indicating less disability; 2 years

(4) Oswestry Disability Index; range 0-100 with lower values indicating less disability; SD imputed; 2 years

(5) Oswestry Disability Index; range 0-100 with lower values indicating less disability; 5 years

Figure 32 PICO 4 - Subgroup analysis 2: Anterior fusion vs. posterior fusion vs. other

### *4.3.1.6 GRADE – PICO 4*

Table 22 PICO 4 long-term: Summary of findings (GRADE)

**PICO 4 - Disc prosthesis (with or without decompression) compared to fusion (with or without decompression) for patients with low back pain with or without neurological symptoms due to degenerative changes of the lumbar spine**

**Patient or population:** patients with low back pain with or without neurological symptoms due to degenerative changes of the lumbar spine
**Intervention:** Disc prosthesis (with or without decompression)
**Comparison:** Fusion (with or without decompression)

| **Outcomes** | **Illustrative comparative risks* (95% CI)** | | **Relative effect (95% CI)** | **No of Participants (studies)** | **Quality of the evidence (GRADE)** | **Comments** |
|---|---|---|---|---|---|---|
| | Assumed risk | Corresponding risk | | | | |
| | **Fusion (with or without decompression)** | **PICO 4 - Disc prosthesis (with or without decompression)** | | | | |
| **Radicular Pain** | | The mean radicular pain in the intervention groups was<br>**3.60 lower**<br>(8.47 lower to 1.27 higher) | | 577<br>(1 study) | ⊕⊕⊝⊝<br>**low**[1,2] | |
| **Back pain** | | The mean back pain in the intervention groups was<br>**5.60 lower**<br>(10.47 to 0.73 lower) | | 577<br>(1 study) | ⊕⊕⊝⊝<br>**low**[1,3] | |
| **Quality of life, physical component summary measure** | | The mean quality of life, physical component summary measure in the intervention groups was<br>**2.77 higher**<br>(0.85 to 4.70 higher) | | 754<br>(2 studies) | ⊕⊕⊝⊝<br>**low**[4,5] | |
| **Quality of life, mental component summary measure** | | The mean quality of life, mental component summary measure in the intervention groups was<br>**1.40 higher**<br>(0.56 lower to 3.36 higher) | | 577<br>(1 study) | ⊕⊝⊝⊝<br>**very low**[1,6] | |
| **Function** | | The mean function in the intervention groups was<br>**5.19 lower** | | 1071<br>(5 studies) | ⊕⊕⊕⊝<br>**moderate**[7] | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | (7.67 to 2.71 lower) | | | | |
| **Revision rate**<br>Follow-up: 2 years | See comment | See comment | Not estimable | 577<br>(1 study) | See comment | One study reported this outcome with no events. |
| **Reoperation rate** | **Study population** | | **RR 1.35** (0.26 to 7.07) | 881<br>(2 studies) | ⊕⊖⊖⊖<br>**very low**[8,9,10] | |
| | **48 per 1000** | **65 per 1000**<br>(12 to 339) | | | | |
| | **Moderate** | | | | | |
| **Complication rate and adverse events** | **Study population** | | **RR 0.96** (0.9 to 1.02) | 1149<br>(4 studies) | ⊕⊕⊕⊖<br>**moderate**[11] | |
| | **659 per 1000** | **633 per 1000**<br>(593 to 673) | | | | |
| | **Moderate** | | | | | |
| **Serious adverse events** | **Study population** | | **RR 0.81** (0.42 to 1.55) | 644<br>(2 studies) | ⊕⊖⊖⊖<br>**very low**[9,10,12] | |
| | **415 per 1000** | **336 per 1000**<br>(174 to 644) | | | | |
| | **Moderate** | | | | | |

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).
**CI:** Confidence interval; **RR:** Risk ratio;

GRADE Working Group grades of evidence
**High quality:** Further research is very unlikely to change our confidence in the estimate of effect.
**Moderate quality:** Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
**Low quality:** Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
**Very low quality:** We are very uncertain about the estimate.

[1] Risk of selection bias (random sequence generation and allocation concealment) was unclear in 1 study; risk of performance bias was high; risk of detection bias was high; risk of attrition bias was high; risk of reporting bias was unclear.

[2] Imprecision was downgraded by one level because the 95% CI of the effect estimate included the possibility for a clinically relevant benefit or no effect assuming a MCID of 7.2.

[3] Imprecision was downgraded by one level because the total sample size was lower than the optimal information size (OIS).

[4] Risk of selection bias (random sequence generation and allocation concealment) was unclear in 2 studies; risk of performance bias was high in 2 studies; risk of detection bias was unclear in 1 and high in 1 study; risk of attrition bias was high in 2 studies.

[5] Imprecision was downgraded by one level because the total sample size was lower than the optimal information size (OIS).

[6] Imprecision was downgraded by two levels because the 95% CI of the effect estimate included the possibility for a clinically relevant benefit or no effect assuming a MCID of 3, and because the total sample size was lower than the OIS.

[7] Risk of selection bias (random sequence generation and allocation concealment) was unclear in 4 studies; risk of performance bias was unclear in 2 and high in 3 studies; risk of detection bias was unclear in 3 and high in 2 studies; risk of attrition bias was unclear in 1 and high in 5 studies; risk of reporting bias was unclear in 2 studies.

[8] Risk of selection bias (random sequence generation and allocation concealment) was unclear in 1 study; risk of performance bias was high in 2 studies; risk of detection bias was high in 2 studies; risk of reporting bias was high in 1 study.

[9] Inconsistency was downgraded by one level because heterogeneity ($I^2$) was high. Directions of effects were not similar and differences in effect sizes were large. Because of the limited number of studies it was not possible to perform sensitivity or subgroup analysis to assess reasons for heterogeneity.

[10] Imprecision was downgraded by two levels because the 95% CI included appreciable harm (greater than 25% relative risk increase) and benefit (greater than 25% relative risk reduction), and because and the total number of events was <300.

[11] Risk of selection bias (random sequence generation and allocation concealment) was unclear in 3 studies; risk of performance bias was unclear in 1 and high in 3 studies; risk of detection bias was unclear in 2 and high in 2 studies; risk of attrition bias was unclear in 1 and high in 1 study; risk of reporting bias was unclear in 2 studies.

[12] Risk of selection bias (random sequence generation and allocation concealment) was unclear in 2 studies; risk of performance bias was high in 2 studies; risk of detection bias was unclear in 1 and high in 1 study; risk of attrition bias was unclear in 1 study.

### 4.3.2 Results – PICO 5

The following sections show first the study characteristics and risk of bias assessment and then the results for each outcome including GRADE results for PICO 5. Fourteen relevant studies have been identified. References can be found in Table 23. Only the pooled long-term results are presented here. An overview of the outcomes analysed from each study is given in Table 24. The pooled short-term results (1-year follow-up) for PICO 5 are presented in Appendix VII.

Table 23 Overview of included studies and their unique study IDs – PICO 5

| Study ID | Reference (main reference highlighted in colour) |
|---|---|
| **Coric 2011** | Coric D, Nunley PD, Guyer RD, et al. Prospective, randomized, multicenter study of cervical arthroplasty: 269 patients from the Kineflex\|C artificial disc investigational device exemption study with a minimum 2-year follow-up: clinical article. *J Neurosurg Spine.* 2011;15(4):348-358. |
| **Heller 2009** | Heller JG, Sasso RC, Papadopoulos SM, et al. Comparison of BRYAN cervical disc arthroplasty with anterior cervical decompression and fusion: clinical and radiographic results of a randomized, controlled, clinical trial. *Spine (Phila Pa 1976).* 2009;34(2):101-107. |
| | Sasso RC, Anderson PA, Riew KD, Heller JG. Results of cervical arthroplasty compared with anterior discectomy and fusion: four-year clinical outcomes in a prospective, randomized controlled trial. *J Bone Joint Surg Am.* 2011;93(18):1684-1692. |
| | Anderson PA, Sasso RC, Riew KD. Comparison of adverse events between the Bryan artificial cervical disc and anterior cervical arthrodesis. *Spine (Phila Pa 1976).* 2008;33(12):1305-1312. |
| | Coric D, Finger F, Boltes P. Prospective randomized controlled study of the Bryan Cervical Disc: early clinical results from a single investigational site. *J Neurosurg Spine.* 2006;4(1):31-35. |
| | Garrido BJ, Taha TA, Sasso RC. Clinical outcomes of Bryan cervical disc arthroplasty a prospective, randomized, controlled, single site trial with 48-month follow-up. *J Spinal Disord Tech.* 2010;23(6):367-371. |
| | Hacker RJ. Cervical disc arthroplasty: a controlled randomized prospective study with intermediate follow-up results. Invited submission from the joint section meeting on disorders of the spine and peripheral nerves, March 2005. *J Neurosurg Spine.* 2005;3(6):424-428. |
| | Sasso RC, Best NM. Cervical kinematics after fusion and bryan disc arthroplasty. *J Spinal Disord Tech.* 2008;21(1):19-22. |
| | Sasso RC, Anderson PA, Riew KD, Heller JG. Results of cervical arthroplasty compared with anterior discectomy and fusion: four-year clinical outcomes in a prospective, randomized controlled trial. *Orthopedics.* 2011;34(11):889. |
| | Sasso RC, Best NM, Metcalf NH, Anderson PA. Motion analysis of bryan cervical disc arthroplasty versus anterior discectomy and fusion: results from a prospective, randomized, multicenter, clinical trial. *J Spinal Disord Tech.* 2008;21(6):393-399. |
| | Sasso RC, Metcalf NH, Hipp JA, Wharton ND, Anderson PA. Sagittal alignment after Bryan cervical arthroplasty. *Spine (Phila Pa 1976).* 2011;36(13):991-996. |

| | |
|---|---|
| | Sasso RC, Smucker JD, Hacker RJ, Heller JG. Clinical outcomes of BRYAN cervical disc arthroplasty: a prospective, randomized, controlled, multicenter trial with 24-month follow-up. *J Spinal Disord Tech.* 2007;20(7):481-491. |
| | Sasso RC, Smucker JD, Hacker RJ, Heller JG. Artificial disc versus fusion: a prospective, randomized study with 2-year follow-up on 99 patients. *Spine (Phila Pa 1976).* 2007;32(26):2933-2940; discussion 2941-2932. |
| **Hisey 2014** | Bae HW, Kim KD, Nunley PD, et al. Comparison of Clinical Outcomes of 1- and 2-Level Total Disc Replacement: Four-Year Results From a Prospective, Randomized, Controlled, Multicenter IDE Clinical Trial. *Spine (Phila Pa 1976).* 2015;40(11):759-766. |
| | Hisey MS, Bae HW, Davis R, et al. Multi-center, prospective, randomized, controlled investigational device exemption clinical trial comparing Mobi-C Cervical Artificial Disc to anterior discectomy and fusion in the treatment of symptomatic degenerative disc disease in the cervical spine. *Int J Spine Surg.* 2014;8. |
| | Hisey MS, Bae HW, Davis RJ, et al. Prospective, Randomized Comparison of Cervical Total Disk Replacement Versus Anterior Cervical Fusion: Results at 48 Months Follow-up. *J Spinal Disord Tech.* 2015;28(4):E237-243. |
| | Jackson RJ, Davis RJ, Hoffman GA, et al. Subsequent surgery rates after cervical total disc replacement using a Mobi-C Cervical Disc Prosthesis versus anterior cervical discectomy and fusion: a prospective randomized clinical trial with 5-year follow-up. *J Neurosurg Spine.* 2016:1-12. |
| **Karabag 2014** | Karabag H, Cakmak E, Celik B, Iplikcioglu AC, Soran AF. Arthroplasty versus fusion for single-level cervical disc disease. *J Pak Med Assoc.* 2014;64(12):1348-1351. |
| **Mummaneni 2007** | Burkus JK, Haid RW, Traynelis VC, Mummaneni PV. Long-term clinical and radiographic outcomes of cervical disc replacement with the Prestige disc: results from a prospective randomized controlled clinical trial. *J Neurosurg Spine.* 2010;13(3):308-318. |
| | Burkus JK, Traynelis VC, Haid RW, Jr., Mummaneni PV. Clinical and radiographic analysis of an artificial cervical disc: 7-year follow-up from the Prestige prospective randomized controlled clinical trial: Clinical article. *J Neurosurg Spine.* 2014;21(4):516-528. |
| | Mummaneni PV, Burkus JK, Haid RW, Traynelis VC, Zdeblick TA. Clinical and radiographic analysis of cervical disc arthroplasty compared with allograft fusion: a randomized controlled clinical trial. *J Neurosurg Spine.* 2007;6(3):198-209. |
| | Riina J, Patel A, Dietz JW, Hoskins JS, Trammell TR, Schwartz DD. Comparison of single-level cervical fusion and a metal-on-metal cervical disc replacement device. *Am J Orthop (Belle Mead NJ).* 2008;37(4):E71-77. |
| **Murrey 2009** | Anakwenze OA, Auerbach JD, Milby AH, Lonner BS, Balderston RA. Sagittal cervical alignment after cervical disc arthroplasty and anterior cervical discectomy and fusion: results of a prospective, randomized, controlled trial. *Spine (Phila Pa 1976).* 2009;34(19):2001-2007. |
| | Delamarter RB, Zigler J. Five-year reoperation rates, cervical total disc replacement versus fusion, results of a prospective randomized clinical trial. *Spine (Phila Pa 1976).* 2013;38(9):711-717. |
| | Delamarter RB, Murrey D, Janssen ME, et al. Results at 24 months from the prospective, randomized, multicenter Investigational Device Exemption trial of ProDisc-C versus anterior cervical discectomy and fusion with 4-year follow-up and continued access patients. *SAS J.* 2010;4(4):122-128. |

| | |
|---|---|
| | Janssen ME, Zigler JE, Spivak JM, Delamarter RB, Darden BV, 2nd, Kopjar B. ProDisc-C Total Disc Replacement Versus Anterior Cervical Discectomy and Fusion for Single-Level Symptomatic Cervical Disc Disease: Seven-Year Follow-up of the Prospective Randomized U.S. Food and Drug Administration Investigational Device Exemption Study. *J Bone Joint Surg Am.* 2015;97(21):1738-1747. |
| | Kelly MP, Mitchell MD, Hacker RJ, Riew KD, Sasso RC. Single-level degenerative cervical disc disease and driving disability: results from a prospective, randomized trial. *Global Spine J.* 2013;3(4):237-242. |
| | Kelly MP, Mok JM, Frisch RF, Tay BK. Adjacent segment motion after anterior cervical discectomy and fusion versus Prodisc-c cervical total disk arthroplasty: analysis from a randomized, controlled trial. *Spine (Phila Pa 1976).* 2011;36(15):1171-1179. |
| | Loumeau TP, Darden BV, Kesman TJ, et al. A RCT comparing 7-year clinical outcomes of one level symptomatic cervical disc disease (SCDD) following ProDisc-C total disc arthroplasty (TDA) versus anterior cervical discectomy and fusion (ACDF). *Eur Spine J.* 2016. |
| | Murrey D, Janssen M, Delamarter R, et al. Results of the prospective, randomized, controlled multicenter Food and Drug Administration investigational device exemption study of the ProDisc-C total disc replacement versus anterior discectomy and fusion for the treatment of 1-level symptomatic cervical disc disease. *Spine J.* 2009;9(4):275-286. |
| | Murrey DB, Janssen ME, Odum SM, Gottlieb JR, Spector LR, Darden BV. Two-Year Results of a Randomized Controlled Clinical Trial Comparing ProDisc-C and Anterior Cervical Discectomy and Fusion. *SAS J.* 2008;2(2):76-85. |
| | Zigler JE, Delamarter R, Murrey D, Spivak J, Janssen M. ProDisc-C and anterior cervical discectomy and fusion as surgical treatment for single-level cervical symptomatic degenerative disc disease: five-year results of a Food and Drug Administration study. *Spine (Phila Pa 1976).* 2013;38(3):203-209. |
| **Nabhan 2007** | Nabhan A, Ahlhelm F, Shariat K, et al. The ProDisc-C prosthesis: clinical and radiological experience 1 year after surgery. *Spine (Phila Pa 1976).* 2007;32(18):1935-1941. |
| | Nabhan A, Steudel WI, Nabhan A, Pape D, Ishak B. Segmental kinematics and adjacent level degeneration following disc replacement versus fusion: RCT with three years of follow-up. *J Long Term Eff Med Implants.* 2007;17(3):229-236. |
| **Nabhan 2011** | Nabhan A, Ishak B, Steudel WI, Ramadhan S, Steimer O. Assessment of adjacent-segment mobility after cervical disc replacement versus fusion: RCT with 1 year's results. *Eur Spine J.* 2011;20(6):934-941. |
| **Phillips 2013** | Phillips FM, Geisler FH, Gilder KM, Reah C, Howell KM, McAfee PC. Long-term Outcomes of the US FDA IDE Prospective, Randomized Controlled Clinical Trial Comparing PCM Cervical Disc Arthroplasty With Anterior Cervical Discectomy and Fusion. *Spine (Phila Pa 1976).* 2015;40(10):674-683. |
| | Phillips FM, Lee JY, Geisler FH, et al. A prospective, randomized, controlled clinical investigation comparing PCM cervical disc arthroplasty with anterior cervical discectomy and fusion. 2-year results from the US FDA IDE clinical trial. *Spine (Phila Pa 1976).* 2013;38(15):E907-918. |
| **Porchet 2004** | Porchet F, Metcalf NH. Clinical outcomes with the Prestige II cervical disc: preliminary results from a prospective randomized clinical trial. *Neurosurg Focus.* 2004;17(3):E6. |
| **Rozankovic** | Rozankovic M, Marasanov SM, Vukic M. Cervical Disc Replacement With Discover Versus |

| | |
|---|---|
| **2014** | Fusion In A Single Level Cervical Disc Disease: A Prospective Single Center Randomized Trial With A Minimum Two-Year Follow - Up. *J Spinal Disord Tech.* 2014. |
| **Vaccaro 2013** | Vaccaro A, Beutler W, Peppelman W, et al. Clinical outcomes with selectively constrained SECURE-C cervical disc arthroplasty: two-year results from a prospective, randomized, controlled, multicenter investigational device exemption study. *Spine (Phila Pa 1976).* 2013;38(26):2227-2239. |
| **Zhang 2012** | Zhang X, Zhang X, Chen C, et al. Randomized, controlled, multicenter, clinical trial comparing BRYAN cervical disc arthroplasty with anterior cervical decompression and fusion in China. *Spine (Phila Pa 1976).* 2012;37(6):433-438. |
| **Zhang 2014** | Zhang HX, Shao YD, Chen Y, et al. A prospective, randomised, controlled multicentre study comparing cervical disc replacement with anterior cervical decompression and fusion. *Int Orthop.* 2014;38(12):2533-2541. |

Table 24 Overview of the outcomes analysed for PICO 5

| | **Neck pain** | **Radicular pain** | **Spinal claudication** | **Myelopathy** | **Quality of life** | **Function** | **Revision rate** | **Reoperation rate** | **Complications/AE** | **Serious AE** |
|---|---|---|---|---|---|---|---|---|---|---|
| Coric 2011 | | | | | | 2 | | 2 | 2 | |
| Heller 2009 | 4 | 4 | | | 4 | 4 | 4 | 4 | | 2 |
| Hisey 2014 | 4 | 4 | | | 4 | 4 | | 4 | | 4 |
| Karabag 2014 | | | | | | 2 | | | | |
| Mummaneni 2007 | 7 | 7 | | | 7 | 7 | 7 | 7 | 7 | |
| Murrey 2009 | 7 | 7 | | | 7 | 7 | | | 7 | |
| Nabhan 2007 | 3 | 3 | | | | | | | | |
| Nabhan 2011 | | | | | | | | | | |
| Phillips 2013 | 5 | 5 | | | 5 | 5 | | 5 | | |
| Porchet 2004 | | | | | | | | | | |
| Rozankovic 2014 | 2 | 2 | | | | 2 | | | | |
| Vaccaro 2013 | 2 | | | | 2 | 2 | | | | |
| Zhang 2012 | 2 | 2 | | | | 2 | | 2 | | |
| Zhang 2014 | | | | | | 4 | | | | |

The number in the fields denote the analysed long-term follow-up in years

### *4.3.2.1 Characteristics of the included studies – PICO 5*

The general characteristics of the studies for PICO-question 5 are summarised in Table 25. Fourteen studies with 3085 participants were included. Ten studies were multicentric and four were single centre studies. Seven studies were conducted in the USA, three in Asia and 4 in Europe. The enrolment period ranged from 2002 to 2010 and the maximum follow-up was 2 to 7 years. Participants had mostly diseased discs but in two studies the main condition was spondylosis. The mean age ranged from 41 to 47 years. Nine studies compared disc prostheses with anterior cervical disc fusion with bone, one study with cage and four studies with bone and cage.

Table 25 Study characteristics, PICO 5

| **Study ID** | **Country**<br>**Setting** | **Enrollment period**<br>**Maximum FU** | **Population**<br>Key condition | **Intervention**<br>○ n randomised<br>○ Male n (%)<br>○ Mean age (SD) | **Comparator**<br>○ n randomised<br>○ Male n (%)<br>○ Mean age (SD), years |
|---|---|---|---|---|---|
| Coric 2011 | USA<br>Multicentre (21 sites) | n.r.<br>2 years | Cervical disc disease with radiculopathy or myelopathy | Disc prosthesis (Kineflex\|C)<br>○ n.r. (136 treated)<br>○ 51 (37.5%)<br>○ 43.7 (7.76) years | ACDF with bone<br>○ n.r. (133 treated)<br>○ 59 (44.4%)<br>○ 43.9 (7.39) years |
| Heller 2009 | USA<br>Multicentre (30 sites) | May 2002-Oct 2004<br>4 years | Cervical disc disease with radiculopathy and/or myelopathy | Disc prosthesis (Bryan)<br>○ 290<br>○ 109 (45.5%)<br>○ 44.4 (25–78)* years | ACDF with bone<br>○ 292<br>○ 113 (51.1%)<br>○ 44.7 (27–68)* years |
| Hisey 2014 | USA<br>Multicentre (23 sites) | Apr 2006-Mar 2008<br>2 years | DDD with radiculopathy or myeloradiculopathy | Disc prosthesis (Mobi-C)<br>○ 169<br>○ 78 (47.6%)<br>○ 43.3 (9.2) years | ACDF with bone<br>○ 87<br>○ 36 (44.4%)<br>○ 44.0 (8.2) years |
| Karabag 2014 | Turkey<br>Single centre | Feb 2009-Jan 2010<br>2 years | Disc disorder; no information on neurologic symptoms reported | Disc prosthesis (n.r.)<br>○ n.r. (19 treated)<br>○ n.r.<br>○ 43.1 (6.1) years | ACDF with cage<br>○ n.r. (23 treated)<br>○ n.r.<br>○ 46.2 (4.7) years |
| Mummaneni 2007 | USA<br>Multicentre (32 sites) | Oct 2002-Aug 2004<br>7 years | DDD with radiculopathy and/or myelopathy | Disc prosthesis (PRESTIGE ST)<br>○ 276<br>○ 128 (46%)<br>○ 43.3 (25-72)* years | ACDF with bone<br>○ 265<br>○ 122 (46%)<br>○ 43.9 (22-73)* years |
| Murrey 2009 | USA<br>Multicentre (13 sites) | Aug 2003-Oct 2004<br>7 years | Disc disesase with neck or arm (radicular) pain | Disc prosthesis (ProDisc-C)<br>○ 111<br>○ 46 (44.7%) years<br>○ 42.1 (8.4) years | ACDF with bone<br>○ 117<br>○ 49 (46.2%) years<br>○ 43.5 (7.1) years |
| Nabhan 2007 | Germany<br>Single centre | Apr 2004-May 2005<br>3 years | Disc disesase with radiculopathy | Disc prosthesis (ProDisc C)<br>○ 25<br>○ n.r. | ACDF with bone<br>○ 24<br>○ n.r. |

| **Study ID** | **Country**<br>**Setting** | **Enrollment period**<br>**Maximum FU** | **Population**<br>Key condition | **Intervention**<br>○ n randomised<br>○ Male n (%)<br>○ Mean age (SD) | **Comparator**<br>○ n randomised<br>○ Male n (%)<br>○ Mean age (SD), years |
|---|---|---|---|---|---|
| | | | | ○ n.r. | ○ n.r. |
| Nabhan 2011 | Germany<br>Single centre | Jan 2006-Aug 2007**<br>1 years | DDD with radiculopathy | Disc prosthesis (ProDisc C)<br>○ 10<br>○ Overall: 13 (65%)<br>○ Overall: 43 (9) years | ACDF with cage and bone<br>○ 10<br>○ Overall: 13 (65%)<br>○ Overall: 43 (9) years |
| Phillips 2013 | USA<br>Multicentre (24 sites) | Jan 2005-Dec 2007**<br>5 years | Cervical spondylosis and radiculopathy and/or myelopathy*** | Disc prosthesis (Porous Coated Motion Cervical Disc)<br>○ 224<br>○ 113 (52%)<br>○ 45.3 (9.0) years | ACDF with bone<br>○ 192<br>○ 96 (52%)<br>○ 43.7 (8.3) years |
| Porchet 2004 | UK, Belgium, Australia, Switzerland<br>Multicentre (4 sites) | n.r.<br>1 year | DDD with radiculopathy or myelopathy | Disc prosthesis (Prestige II)<br>○ 27<br>○ 17 (63%)<br>○ 44.3 (8.9) years | ACDF with cage and bone<br>○ 28<br>○ 12 (43%)<br>○ 43.2 (6.9) years |
| Rozankovic 2014 | Croatia<br>Single centre | Oct 2008-Jun 2010<br>2 years | DDD with radiculopathy and/or myelopathy | Disc prosthesis (Discover)<br>○ 52<br>○ 25 (49%)<br>○ 41 (8.8) years | ACDF with cage and bone<br>○ 53<br>○ 25 (50%)<br>○ 42 (9.4) years |
| Vaccaro 2013 | USA<br>Multicentre (18 sites) | n.r.<br>4 years | Disc disease with neck or arm (radicular) pain | Disc prosthesis (SECURE-C)<br>○ 151<br>○ 81 (54%)<br>○ 43.4 (7.50) years | ACDF with bone<br>○ 140<br>○ 68 (49%)<br>○ 44.4 (7.86) years |
| Zhang 2012 | China<br>Multicentre (3 sites) | May 2004-May 2006<br>2 years | DDD with radiculopathy or myelopathy | Disc prosthesis (Bryan)<br>○ 60<br>○ 35 (58%)<br>○ 44.77 (5.60) years | ACDF with bone<br>○ 60<br>○ 32 (53%)<br>○ 45.57 (5.83) years |
| Zhang 2014 | China<br>Multicentre (11 sites) | Feb 2008-Nov 2009 | Degenerative cervical spondylosis | Disc prosthesis (Mobi-C)<br>○ n.r. (55 treated) | ACDF with cage and bone<br>○ n.r. (56 treated) |

| **Study ID** | **Country**<br>**Setting** | **Enrollment period**<br>**Maximum FU** | **Population**<br>Key condition | **Intervention**<br>○ n randomised<br>○ Male n (%)<br>○ Mean age (SD) | **Comparator**<br>○ n randomised<br>○ Male n (%)<br>○ Mean age (SD), years |
|---|---|---|---|---|---|
| | | 4 years | | ○ 25 (45%)<br>○ 44.8 (18 – 68)* years | ○ 26 (46%)<br>○ 46.7 (18 – 68)* years |

Note: Under key condition the presence or absence of neurological symptoms is being described according to the inclusion criteria. If additional information based on the baseline characteristics would have been available this would have been documented in footnotes.

* mean (range);

** Treatment period;

*** only 1 patient (0.4%) with diagnosed myelopathy was reported (80.3% only radiculopathy and 19.2% with both)

Abbreviations: ACDF, anterior cervical discectomy and fusion; DDD, degenerative disc disease; FU, follow-up; n, number; n.r., not reported; SD, standard deviation

#### 4.1.1.1 Risk of bias – PICO 5

The method of random sequence generation was adequate in six studies (Hisey 2014, Mummaneni 2007, Murrey 2009, Nabhan 2007, Rozankovic 2014, Zhang 2012) and unclear in six studies (Coric 2011, Heller 2009, Karabag 2014, Phillips 2013, Vaccaro 2013, Zhang 2014). The risk of selection bias depending on allocation concealment was rated as low for three studies (Heller 2009, Hisey 2014, Murrey 2009) and as unclear for nine studies (Coric 2011, Karabag 2014, Mummaneni 2007, Nabhan 2007, Phillips 2013, Vaccaro 2013, Zhang 2012, Zhang 2014). The risk of performance bias depending on blinding of participants and personnel was rated to be high in six studies (Heller 2009, Hisey 2014, Mummaneni 2007, Murrey 2009, Phillips 2013, Rozankovic 2014, Vaccaro 2013) and to be unclear in six studies (Coric 2011, Karabag 2014, Nabhan 2007, Rozankovic 2014, Zhang 2012, Zhang 2014). Blinding of outcome assessment was unclear in all twelve studies (Coric 2011, Heller 2009, Hisey 2014, Karabag 2014, Mummaneni 2007, Murrey 2009, Nabhan 2007, Phillips 2013, Rozankovic 2014, Vaccaro 2013, Zhang 2012, Zhang 2014). Attrition bias for continuous outcome data was rated as low in two studies (Rozankovic 2014, Zhang 2012), as high in seven studies (Coric 2011, Heller 2009, Hisey 2014, Mummaneni 2007, Murrey 2009, Nabhan 2007, Phillips 2013,) and unclear in three studies (Karabag 2014, Vaccaro 2013, Zhang 2014). Attrition bias for binary outcome data was rated as low in two studies (Rozankovic 2014, Zhang 2012, Zhang 2014), as high in five studies (Coric 2011, Heller 2009, Hisey 2014, Mummaneni 2007, Murrey 2009, Phillips 2013), as unclear in one study (Vaccaro 2013) and from two studies no binary data was extracted (Karabag 2014, Nabhan 2007). Selective reporting was low in three studies (Heller 2009, Karabag 2014, Mummaneni 2007, Murrey 2009, Nabhan 2007, Phillips 2013, Rozankovic 2014) and was rated as unclear in five studies (Coric 2011, Hisey 2014, Vaccaro 2013, Zhang 2012, Zhang 2014). A summarised overview of the risk of bias assessment is shown in Table 26 and a detailed summary with support of judgement in Appendix III.

Table 26 Risk of bias, PICO 5

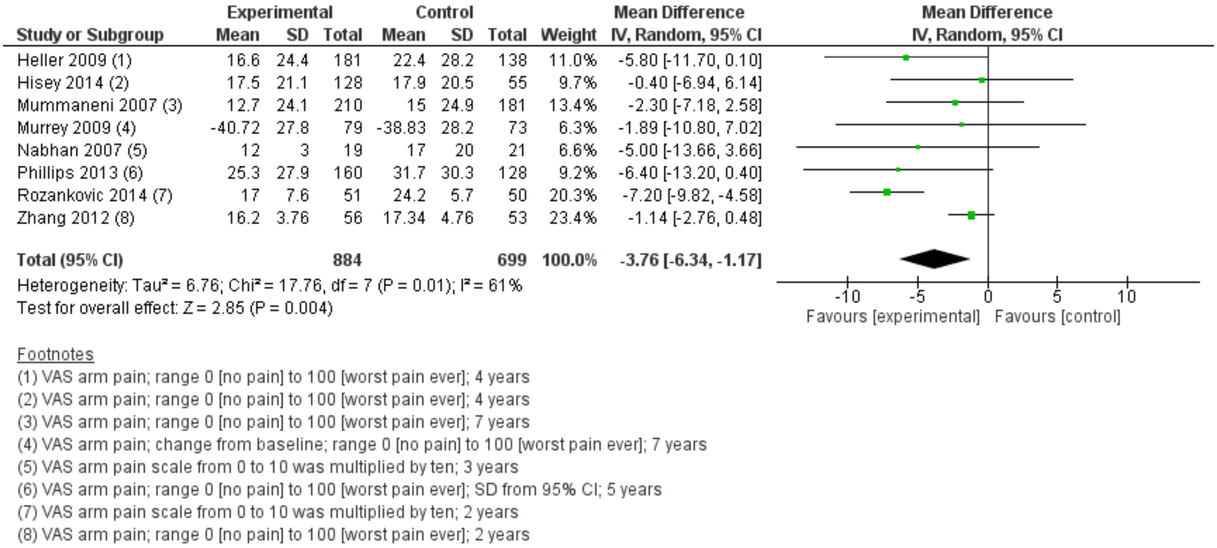
| Study | Random sequence generation (selection bias) | Allocation concealment (selection bias) | Blinding of participants and personnel (performance bias) | Blinding of outcome assessment, judgement did not differ among outcomes (detection bias) | Incomplete continuous outcome data (attrition bias) | Incomplete binary data (attrition bias) | Selective reporting (reporting bias) |
|---|---|---|---|---|---|---|---|
| Coric 2011 | Unclear | Unclear | Unclear | Unclear | High | High | Unclear |
| Heller 2009 | Unclear | Low | High | Unclear | High | High | Low |
| Hisey 2014 | Low | Low | High | Unclear | High | High | Unclear |
| Karabag 2014 | Unclear | Unclear | Unclear | Unclear | Unclear | n.a. | Low |
| Mummaneni 2007 | Low | Uncear | High | Unclear | High | High | Low |
| Murrey 2009 | Low | Low | High | Unclear | High | High | Low |
| Nabhan 2007 | Low | Unclear | Unclear | Unclear | High | n.a. | Low |
| Phillips 2013 | Unclear | Unclear | High | Unclear | High | High | Low |
| Rozankovic 2014 | Low | Unclear | Unclear | Unclear | Low | Low | Low |
| Vaccaro 2013 | Unclear | Unclear | High | Unclear | Unclear | Unclear | Unclear |
| Zhang 2012 | Low | Unclear | Unclear | Unclear | Low | Low | Unclear |
| Zhang 2014 | Unclear | Unclear | Unclear | Unclear | Unclear | Low | Unclear |

Abbreviation: n.a., not applicable.
*no relevant binary outcome identified

### *4.1.1.2 Critical outcomes – PICO 5*

#### 4.1.1.2.1 Radicular pain

Eight studies (Heller 2009, Hisey 2014, Mummaneni 2007, Murrey 2009, Nabhan 2007, Phillips 2013, Rozankovic 2014, Zhang 2012) reported on arm pain with a follow-up of 2 to 7 years. Five studies (Heller 2009, Hisey 2014, Mummaneni 2007, Phillips 2013, Zhang 2012) used the Visual Analogue Scale (VAS, range 0 [no pain] to 100 [worst pain ever]). One study (Murrey 2009) reported the change from baseline on a VAS after 7 years and was pooled with the end of follow-up measurements of the other studies. Two studies (Nabhan 2007, Rozankovic 2014) reported arm pain on a VAS with a range from 0 to 100. These results were multiplied with 10 and then pooled with the other studies. Compared to fusion, disc prostheses reduced arm pain statistically significantly (MD -3.76, 95% CI -6.34 to -1.17; moderate quality of evidence; Figure 33). Heterogeneity between studies was high ($I^2$=61%). Short-term (1 year) results were reported by ten studies and showed a similar statistically significant effect on radicular pain.



| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Heller 2009 (1) | 16.6 | 24.4 | 181 | 22.4 | 28.2 | 138 | 11.0% | -5.80 [-11.70, 0.10] |
| Hisey 2014 (2) | 17.5 | 21.1 | 128 | 17.9 | 20.5 | 55 | 9.7% | -0.40 [-6.94, 6.14] |
| Mummaneni 2007 (3) | 12.7 | 24.1 | 210 | 15 | 24.9 | 181 | 13.4% | -2.30 [-7.18, 2.58] |
| Murrey 2009 (4) | -40.72 | 27.8 | 79 | -38.83 | 28.2 | 73 | 6.3% | -1.89 [-10.80, 7.02] |
| Nabhan 2007 (5) | 12 | 3 | 19 | 17 | 20 | 21 | 6.6% | -5.00 [-13.66, 3.66] |
| Phillips 2013 (6) | 25.3 | 27.9 | 160 | 31.7 | 30.3 | 128 | 9.2% | -6.40 [-13.20, 0.40] |
| Rozankovic 2014 (7) | 17 | 7.6 | 51 | 24.2 | 5.7 | 50 | 20.3% | -7.20 [-9.82, -4.58] |
| Zhang 2012 (8) | 16.2 | 3.76 | 56 | 17.34 | 4.76 | 53 | 23.4% | -1.14 [-2.76, 0.48] |
| **Total (95% CI)** | | | **884** | | | **699** | **100.0%** | **-3.76 [-6.34, -1.17]** |

Heterogeneity: Tau² = 6.76; Chi² = 17.76, df = 7 (P = 0.01); I² = 61%
Test for overall effect: Z = 2.85 (P = 0.004)

Footnotes
(1) VAS arm pain; range 0 [no pain] to 100 [worst pain ever]; 4 years
(2) VAS arm pain; range 0 [no pain] to 100 [worst pain ever]; 4 years
(3) VAS arm pain; range 0 [no pain] to 100 [worst pain ever]; 7 years
(4) VAS arm pain; change from baseline; range 0 [no pain] to 100 [worst pain ever]; 7 years
(5) VAS arm pain scale from 0 to 10 was multiplied by ten; 3 years
(6) VAS arm pain; range 0 [no pain] to 100 [worst pain ever]; SD from 95% CI; 5 years
(7) VAS arm pain scale from 0 to 10 was multiplied by ten; 2 years
(8) VAS arm pain; range 0 [no pain] to 100 [worst pain ever]; 2 years

**Figure 33 PICO 5 long-term: Radicular pain**

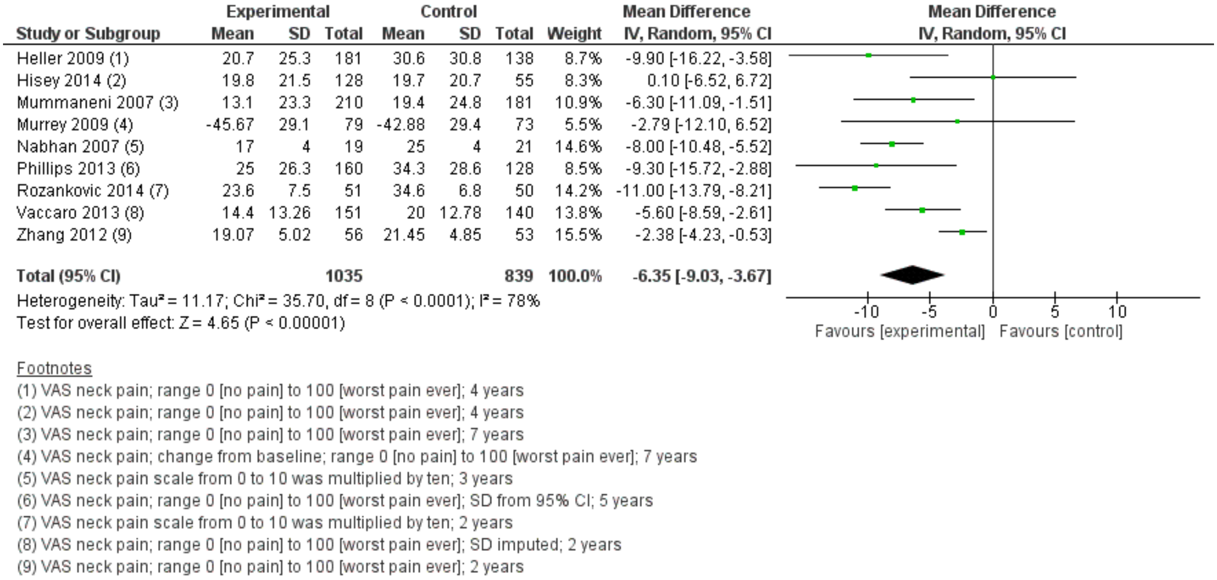
#### 4.1.1.2.2 Myelopathy

No study reported on myelopathy.

### *4.1.1.3 Important outcomes – PICO 5*

#### 4.1.1.3.1 Neck pain

Nine studies (Heller 2009, Hisey 2014, Mummaneni 2007, Murrey 2009, Nabhan 2007, Phillips 2013, Rozankovic 2014, Vaccaro 2013, Zhang 2012) reported neck pain with a follow-up of 2 to 7 years. Six studies (Heller 2009, Hisey 2014, Mummaneni 2007, Phillips 2013, Vaccaro 2013, Zhang 2012) used Visual Analogue Scale (VAS, range 0 [no pain] to 100 [worst pain ever]). One study (Murrey 2009) reported the change from baseline on a VAS after 7 years and was pooled with the end of follow-up measurements of the other studies. Two studies (Nabhan 2007, Rozankovic 2014) reported arm pain on a VAS with a range from 0 to 100. These results were multiplied with 10 and then pooled with the other studies. Compared to fusion, disc prostheses reduced neck pain statistically significantly (MD -6.35, 95% CI -9.03 to -3.67; low quality of evidence; Figure 34). Heterogeneity between studies was high ($I^2$=78%). Short-term (1 year) results were reported by eleven studies and showed a similar statistically significant effect on neck pain.



| Study or Subgroup | Experimental Mean | SD | Total | Control Mean | SD | Total | Weight | Mean Difference IV, Random, 95% CI |
|---|---|---|---|---|---|---|---|---|
| Heller 2009 (1) | 20.7 | 25.3 | 181 | 30.6 | 30.8 | 138 | 8.7% | -9.90 [-16.22, -3.58] |
| Hisey 2014 (2) | 19.8 | 21.5 | 128 | 19.7 | 20.7 | 55 | 8.3% | 0.10 [-6.52, 6.72] |
| Mummaneni 2007 (3) | 13.1 | 23.3 | 210 | 19.4 | 24.8 | 181 | 10.9% | -6.30 [-11.09, -1.51] |
| Murrey 2009 (4) | -45.67 | 29.1 | 79 | -42.88 | 29.4 | 73 | 5.5% | -2.79 [-12.10, 6.52] |
| Nabhan 2007 (5) | 17 | 4 | 19 | 25 | 4 | 21 | 14.6% | -8.00 [-10.48, -5.52] |
| Phillips 2013 (6) | 25 | 26.3 | 160 | 34.3 | 28.6 | 128 | 8.5% | -9.30 [-15.72, -2.88] |
| Rozankovic 2014 (7) | 23.6 | 7.5 | 51 | 34.6 | 6.8 | 50 | 14.2% | -11.00 [-13.79, -8.21] |
| Vaccaro 2013 (8) | 14.4 | 13.26 | 151 | 20 | 12.78 | 140 | 13.8% | -5.60 [-8.59, -2.61] |
| Zhang 2012 (9) | 19.07 | 5.02 | 56 | 21.45 | 4.85 | 53 | 15.5% | -2.38 [-4.23, -0.53] |
| **Total (95% CI)** | | | **1035** | | | **839** | **100.0%** | **-6.35 [-9.03, -3.67]** |

Heterogeneity: Tau² = 11.17; Chi² = 35.70, df = 8 (P < 0.0001); I² = 78%
Test for overall effect: Z = 4.65 (P < 0.00001)

Footnotes
(1) VAS neck pain; range 0 [no pain] to 100 [worst pain ever]; 4 years
(2) VAS neck pain; range 0 [no pain] to 100 [worst pain ever]; 4 years
(3) VAS neck pain; range 0 [no pain] to 100 [worst pain ever]; 7 years
(4) VAS neck pain; change from baseline; range 0 [no pain] to 100 [worst pain ever]; 7 years
(5) VAS neck pain scale from 0 to 10 was multiplied by ten; 3 years
(6) VAS neck pain; range 0 [no pain] to 100 [worst pain ever]; SD from 95% CI; 5 years
(7) VAS neck pain scale from 0 to 10 was multiplied by ten; 2 years
(8) VAS neck pain; range 0 [no pain] to 100 [worst pain ever]; SD imputed; 2 years
(9) VAS neck pain; range 0 [no pain] to 100 [worst pain ever]; 2 years

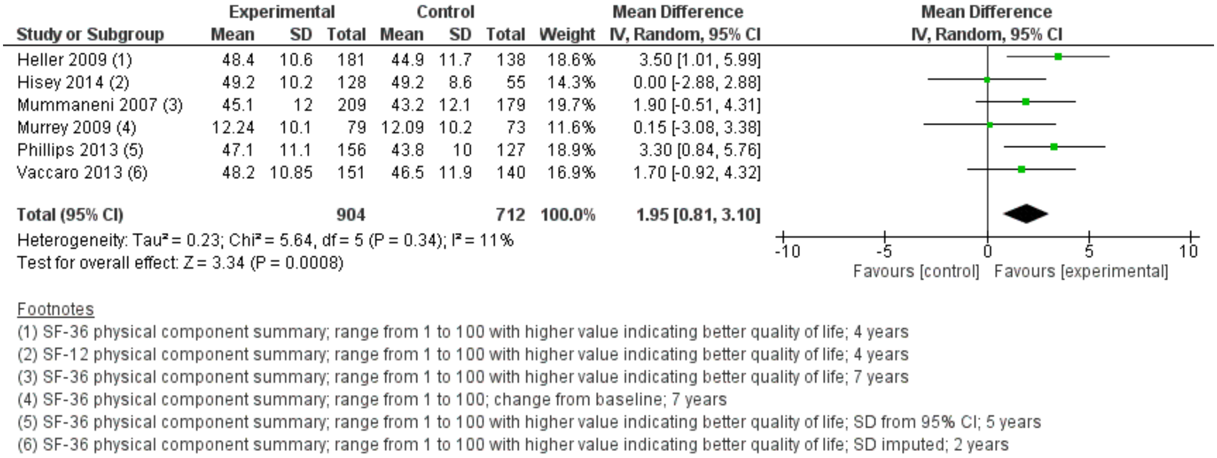
**Figure 34 PICO 5 long-term: Neck pain**

#### 4.1.1.3.2 Quality of life

Five studies (Heller 2009, Mummaneni 2007, Murrey 2009, Phillips 2013, Vaccaro 2013) reported on quality of life measured with the SF-36 and one study (Hisey 2014) with the SF-12. Four studies (Heller 2009, Murrey 2009, Phillips 2013, Vaccaro 2013) reported both the physical component score and mental component score of SF-36. One study (Mummaneni 2007) reported only the physical component score of SF-36. One study (Hisey 2014) reported the physical component score and mental component score of SF-12.

**Short form 36: physical component score**

Six studies (Heller 2009, Hisey 2014, Mummaneni 2007, Murrey 2009, Phillips 2013, Vaccaro 2013) reported the physical component score. There was a significant difference between disc prostheses and fusion (MD 1.95, 95% CI 0.81 to 3.10; moderate quality of evidence; Figure 35). Heterogeneity was low ($I^2$=11%). Short-term (1 year) results were reported by six studies and showed a similar statistically significant effect on quality of life physical component score.



| Study or Subgroup | Experimental Mean | SD | Total | Control Mean | SD | Total | Weight | Mean Difference IV, Random, 95% CI |
|---|---|---|---|---|---|---|---|---|
| Heller 2009 (1) | 48.4 | 10.6 | 181 | 44.9 | 11.7 | 138 | 18.6% | 3.50 [1.01, 5.99] |
| Hisey 2014 (2) | 49.2 | 10.2 | 128 | 49.2 | 8.6 | 55 | 14.3% | 0.00 [-2.88, 2.88] |
| Mummaneni 2007 (3) | 45.1 | 12 | 209 | 43.2 | 12.1 | 179 | 19.7% | 1.90 [-0.51, 4.31] |
| Murrey 2009 (4) | 12.24 | 10.1 | 79 | 12.09 | 10.2 | 73 | 11.6% | 0.15 [-3.08, 3.38] |
| Phillips 2013 (5) | 47.1 | 11.1 | 156 | 43.8 | 10 | 127 | 18.9% | 3.30 [0.84, 5.76] |
| Vaccaro 2013 (6) | 48.2 | 10.85 | 151 | 46.5 | 11.9 | 140 | 16.9% | 1.70 [-0.92, 4.32] |
| **Total (95% CI)** | | | **904** | | | **712** | **100.0%** | **1.95 [0.81, 3.10]** |

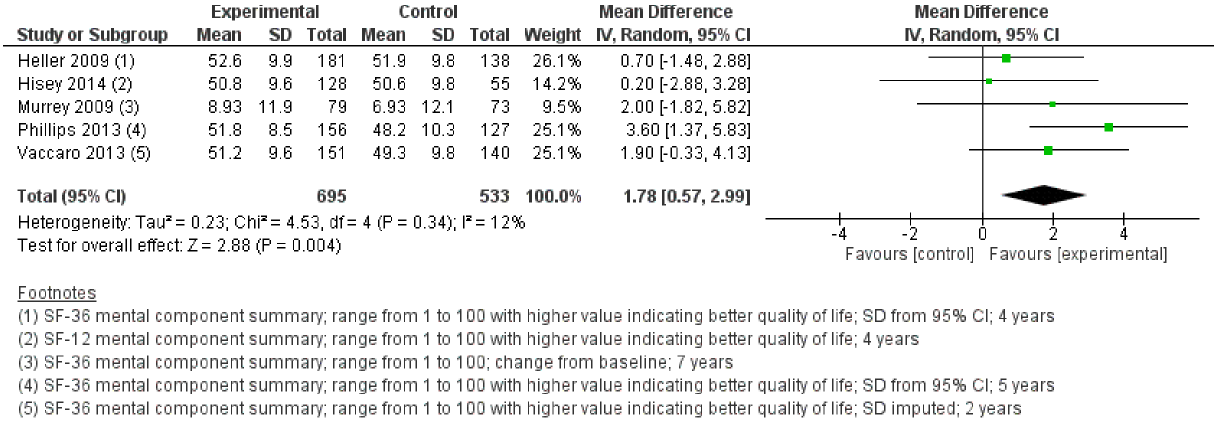
Heterogeneity: Tau² = 0.23; Chi² = 5.64, df = 5 (P = 0.34); I² = 11%
Test for overall effect: Z = 3.34 (P = 0.0008)

Footnotes
(1) SF-36 physical component summary; range from 1 to 100 with higher value indicating better quality of life; 4 years
(2) SF-12 physical component summary; range from 1 to 100 with higher value indicating better quality of life; 4 years
(3) SF-36 physical component summary; range from 1 to 100 with higher value indicating better quality of life; 7 years
(4) SF-36 physical component summary; range from 1 to 100; change from baseline; 7 years
(5) SF-36 physical component summary; range from 1 to 100 with higher value indicating better quality of life; SD from 95% CI; 5 years
(6) SF-36 physical component summary; range from 1 to 100 with higher value indicating better quality of life; SD imputed; 2 years

**Figure 35 PICO 5 long-term: Quality of life, SF-36 physical component score**

**Short form 36: mental component score**

Five studies (Heller 2009, Hisey 2014, Murrey 2009, Phillips 2013, Vaccaro 2013) reported the mental component score. There was a statistically significant difference between disc prostheses and fusion (MD 1.78, 95% CI 0.57 to 2.99; moderate quality of evidence; Figure 36). Heterogeneity was low ($I^2$=12%). Short-term (1 year) results were reported by six studies and showed a similar statistical significantly effect on quality of life mental component score.



| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Heller 2009 (1) | 52.6 | 9.9 | 181 | 51.9 | 9.8 | 138 | 26.1% | 0.70 [-1.48, 2.88] |
| Hisey 2014 (2) | 50.8 | 9.6 | 128 | 50.6 | 9.8 | 55 | 14.2% | 0.20 [-2.88, 3.28] |
| Murrey 2009 (3) | 8.93 | 11.9 | 79 | 6.93 | 12.1 | 73 | 9.5% | 2.00 [-1.82, 5.82] |
| Phillips 2013 (4) | 51.8 | 8.5 | 156 | 48.2 | 10.3 | 127 | 25.1% | 3.60 [1.37, 5.83] |
| Vaccaro 2013 (5) | 51.2 | 9.6 | 151 | 49.3 | 9.8 | 140 | 25.1% | 1.90 [-0.33, 4.13] |
| **Total (95% CI)** | | | **695** | | | **533** | **100.0%** | **1.78 [0.57, 2.99]** |

Heterogeneity: Tau$^2$ = 0.23; Chi$^2$ = 4.53, df = 4 (P = 0.34); I$^2$ = 12%

Test for overall effect: Z = 2.88 (P = 0.004)

Footnotes

(1) SF-36 mental component summary; range from 1 to 100 with higher value indicating better quality of life; SD from 95% CI; 4 years

(2) SF-12 mental component summary; range from 1 to 100 with higher value indicating better quality of life; 4 years

(3) SF-36 mental component summary; range from 1 to 100; change from baseline; 7 years

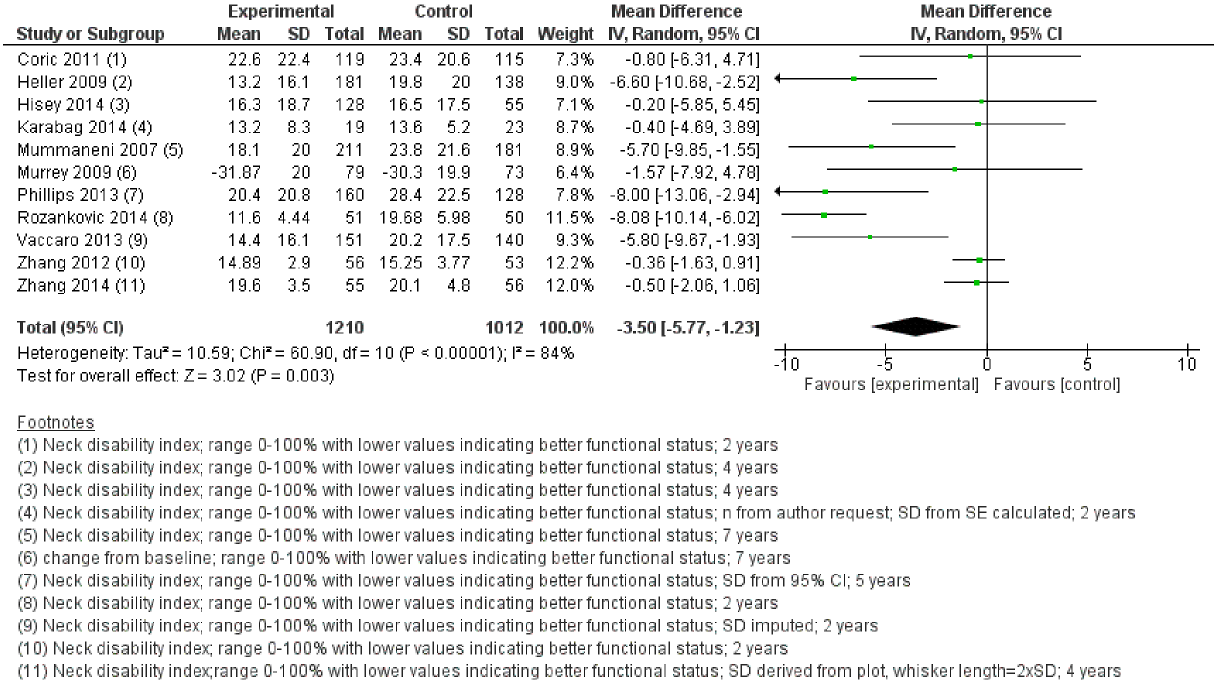
(4) SF-36 mental component summary; range from 1 to 100 with higher value indicating better quality of life; SD from 95% CI; 5 years

(5) SF-36 mental component summary; range from 1 to 100 with higher value indicating better quality of life; SD imputed; 2 years

**Figure 36 PICO 5 long-term: Quality of life, SF-36 mental component score**

### 4.1.1.3.3 Function

Eleven studies (Coric 2011, Heller 2009, Hisey 2014, Karabag 2014, Mummaneni 2007, Murrey 2009, Phillips 2013, Rozankovic 2014, Vaccaro 2013, Zhang 2012, Zhang 2014) reported on function after a follow-up of 2 to 7 years. All studies measured function with the Neck Disability Index (NDI, range 0-100 with lower values indicating better functional status). Compared to fusion, disc prostheses statistically significantly improved function (MD -3.50, 95% CI -5.77 to -1.23; moderate quality of evidence; Figure 37). Heterogeneity between studies was high ($I^2$=84%). Short-term (1 year) results were reported by eleven studies and showed a similar statistically significant effect on function.



| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Coric 2011 (1) | 22.6 | 22.4 | 119 | 23.4 | 20.6 | 115 | 7.3% | -0.80 [-6.31, 4.71] |
| Heller 2009 (2) | 13.2 | 16.1 | 181 | 19.8 | 20 | 138 | 9.0% | -6.60 [-10.68, -2.52] |
| Hisey 2014 (3) | 16.3 | 18.7 | 128 | 16.5 | 17.5 | 55 | 7.1% | -0.20 [-5.85, 5.45] |
| Karabag 2014 (4) | 13.2 | 8.3 | 19 | 13.6 | 5.2 | 23 | 8.7% | -0.40 [-4.69, 3.89] |
| Mummaneni 2007 (5) | 18.1 | 20 | 211 | 23.8 | 21.6 | 181 | 8.9% | -5.70 [-9.85, -1.55] |
| Murrey 2009 (6) | -31.87 | 20 | 79 | -30.3 | 19.9 | 73 | 6.4% | -1.57 [-7.92, 4.78] |
| Phillips 2013 (7) | 20.4 | 20.8 | 160 | 28.4 | 22.5 | 128 | 7.8% | -8.00 [-13.06, -2.94] |
| Rozankovic 2014 (8) | 11.6 | 4.44 | 51 | 19.68 | 5.98 | 50 | 11.5% | -8.08 [-10.14, -6.02] |
| Vaccaro 2013 (9) | 14.4 | 16.1 | 151 | 20.2 | 17.5 | 140 | 9.3% | -5.80 [-9.67, -1.93] |
| Zhang 2012 (10) | 14.89 | 2.9 | 56 | 15.25 | 3.77 | 53 | 12.2% | -0.36 [-1.63, 0.91] |
| Zhang 2014 (11) | 19.6 | 3.5 | 55 | 20.1 | 4.8 | 56 | 12.0% | -0.50 [-2.06, 1.06] |
| **Total (95% CI)** | | | **1210** | | | **1012** | **100.0%** | **-3.50 [-5.77, -1.23]** |

Heterogeneity: Tau$^2$ = 10.59; Chi$^2$ = 60.90, df = 10 (P < 0.00001); I$^2$ = 84%

Test for overall effect: Z = 3.02 (P = 0.003)

Footnotes

(1) Neck disability index; range 0-100% with lower values indicating better functional status; 2 years

(2) Neck disability index; range 0-100% with lower values indicating better functional status; 4 years

(3) Neck disability index; range 0-100% with lower values indicating better functional status; 4 years

(4) Neck disability index; range 0-100% with lower values indicating better functional status; n from author request; SD from SE calculated; 2 years

(5) Neck disability index; range 0-100% with lower values indicating better functional status; 7 years

(6) change from baseline; range 0-100% with lower values indicating better functional status; 7 years

(7) Neck disability index; range 0-100% with lower values indicating better functional status; SD from 95% CI; 5 years

(8) Neck disability index; range 0-100% with lower values indicating better functional status; 2 years

(9) Neck disability index; range 0-100% with lower values indicating better functional status; SD imputed; 2 years

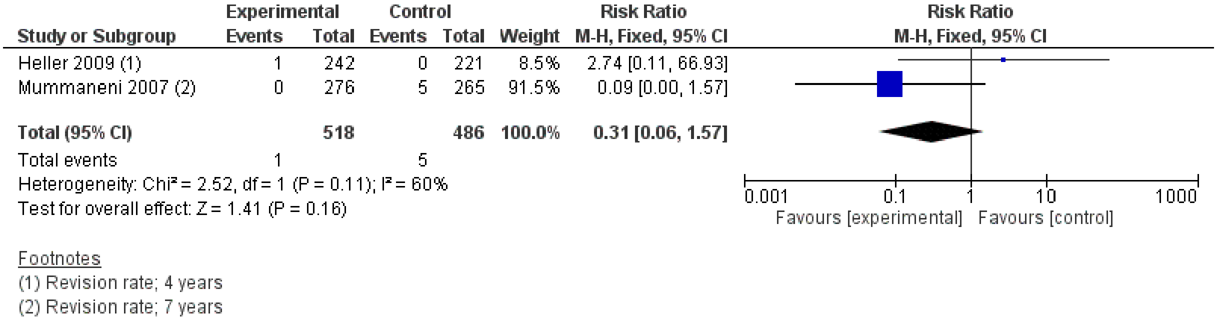
(10) Neck disability index; range 0-100% with lower values indicating better functional status; 2 years

(11) Neck disability index;range 0-100% with lower values indicating better functional status; SD derived from plot, whisker length=2xSD; 4 years

**Figure 37 PICO 5 long-term function: Neck disability index**

### 4.1.1.3.4 Revision rate

Two studies (Heller 2009, Mummaneni 2007) reported the revision rates after a follow-up of 4 and 7 years. There were 6 events in 1004 participants. There was no statistically significant difference between disc prostheses and fusion (RR 0.31, 95% CI 0.06 to 1.57; very low quality of evidence; Figure 38). Heterogeneity was moderate ($I^2$=60%). Short-term (1 year) results were reported by two studies and showed a similar effect on revision rate.



| | Experimental | | Control | | | Risk Ratio |
|---|---|---|---|---|---|---|
| Study or Subgroup | Events | Total | Events | Total | Weight | M-H, Fixed, 95% CI |
| Heller 2009 (1) | 1 | 242 | 0 | 221 | 8.5% | 2.74 [0.11, 66.93] |
| Mummaneni 2007 (2) | 0 | 276 | 5 | 265 | 91.5% | 0.09 [0.00, 1.57] |
| **Total (95% CI)** | | **518** | | **486** | **100.0%** | **0.31 [0.06, 1.57]** |
| Total events | 1 | | 5 | | | |

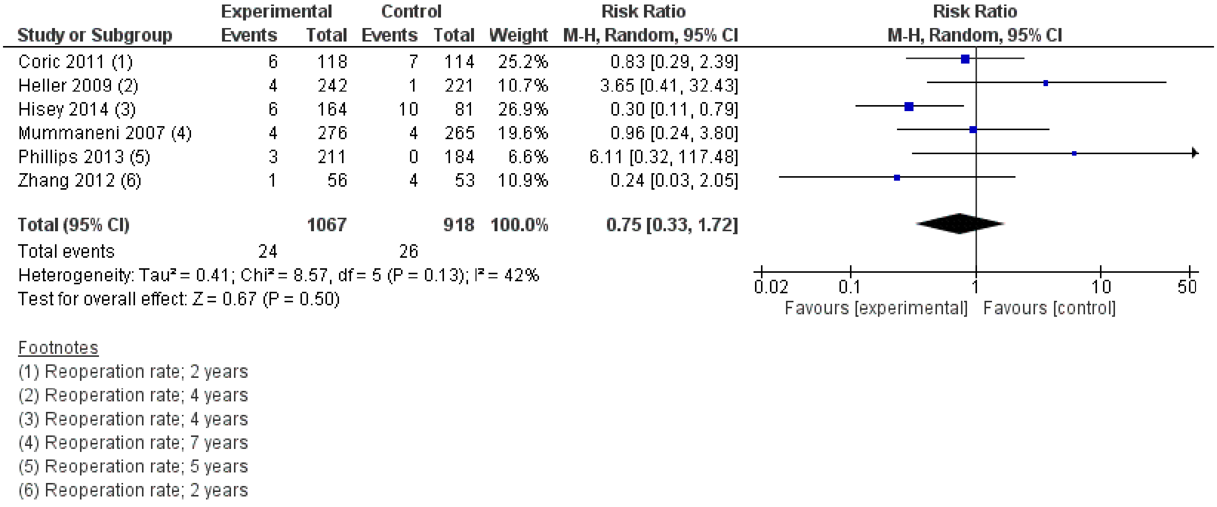
Heterogeneity: Chi² = 2.52, df = 1 (P = 0.11); I² = 60%
Test for overall effect: Z = 1.41 (P = 0.16)

Footnotes
(1) Revision rate; 4 years
(2) Revision rate; 7 years

**Figure 38 PICO 5 long-term: Revision rate**

### 4.1.1.3.5 Reoperation rate

Six studies (Coric 2011, Heller 2009, Hisey 2014, Mummaneni 2007, Phillips 2013, Zhang 2012) reported the reoperation rates after a follow-up of 2 to 7 years. In a population of 1985 patients 50 events were reported. There was no statistically significant difference between disc prostheses and fusion (RR 0.75, 95% CI 0.33 to 1. 72; very low quality of evidence; Figure 39). Heterogeneity was moderate ($I^2$=42%). Reoperation at short-term (1 year) was reported by only one study and showed a statistically non-significant result.



| | Experimental | | Control | | | Risk Ratio |
|---|---|---|---|---|---|---|
| Study or Subgroup | Events | Total | Events | Total | Weight | M-H, Random, 95% CI |
| Coric 2011 (1) | 6 | 118 | 7 | 114 | 25.2% | 0.83 [0.29, 2.39] |
| Heller 2009 (2) | 4 | 242 | 1 | 221 | 10.7% | 3.65 [0.41, 32.43] |
| Hisey 2014 (3) | 6 | 164 | 10 | 81 | 26.9% | 0.30 [0.11, 0.79] |
| Mummaneni 2007 (4) | 4 | 276 | 4 | 265 | 19.6% | 0.96 [0.24, 3.80] |
| Phillips 2013 (5) | 3 | 211 | 0 | 184 | 6.6% | 6.11 [0.32, 117.48] |
| Zhang 2012 (6) | 1 | 56 | 4 | 53 | 10.9% | 0.24 [0.03, 2.05] |
| **Total (95% CI)** | | **1067** | | **918** | **100.0%** | **0.75 [0.33, 1.72]** |
| Total events | 24 | | 26 | | | |

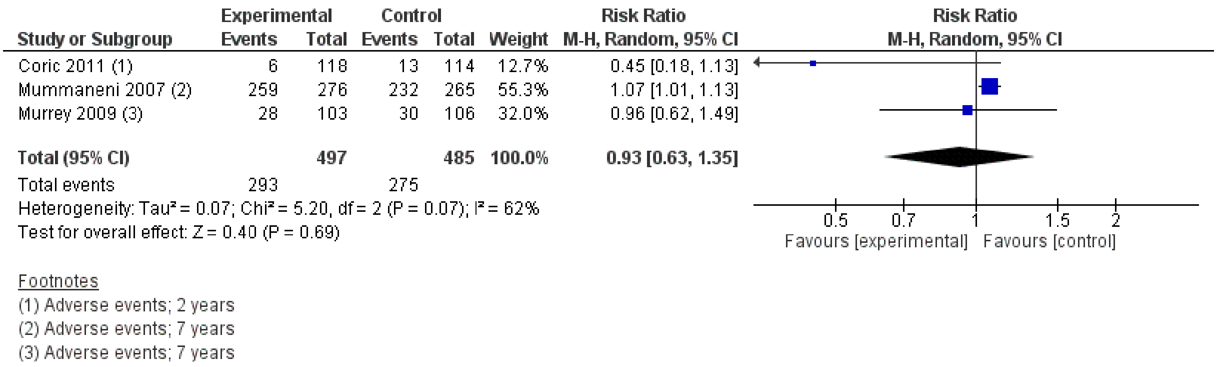
Heterogeneity: Tau² = 0.41; Chi² = 8.57, df = 5 (P = 0.13); I² = 42%
Test for overall effect: Z = 0.67 (P = 0.50)

Footnotes
(1) Reoperation rate; 2 years
(2) Reoperation rate; 4 years
(3) Reoperation rate; 4 years
(4) Reoperation rate; 7 years
(5) Reoperation rate; 5 years
(6) Reoperation rate; 2 years

**Figure 39 PICO 5 long-term: Reoperation rate**

### 4.1.1.3.6 Complications or adverse events

Three (Coric 2011, Mummaneni 2007, Murrey 2009) studies reported adverse event rates after a follow-up of 2 and 7 years. Overall, 568 events have been reported in 982 patients. There was no statistically significant difference between disc prostheses and fusion (RR 0.93, 95% CI 0.63 to 1.35; very low quality of evidence; Figure 40). Heterogeneity was high ($I^2$=62%). No study reported complications or adverse events at short-term (1 year) follow-up.



| Study or Subgroup | Experimental Events | Experimental Total | Control Events | Control Total | Weight | Risk Ratio M-H, Random, 95% CI |
|---|---|---|---|---|---|---|
| Coric 2011 (1) | 6 | 118 | 13 | 114 | 12.7% | 0.45 [0.18, 1.13] |
| Mummaneni 2007 (2) | 259 | 276 | 232 | 265 | 55.3% | 1.07 [1.01, 1.13] |
| Murrey 2009 (3) | 28 | 103 | 30 | 106 | 32.0% | 0.96 [0.62, 1.49] |
| **Total (95% CI)** | | **497** | | **485** | **100.0%** | **0.93 [0.63, 1.35]** |
| Total events | 293 | | 275 | | | |

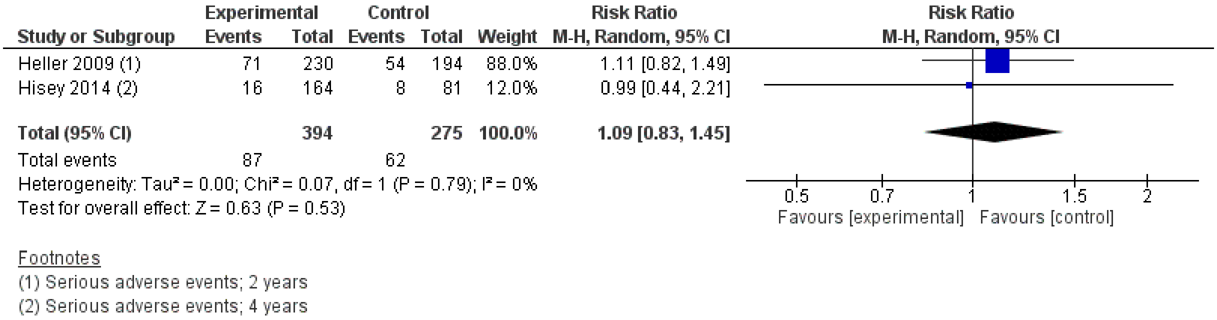
Heterogeneity: Tau² = 0.07; Chi² = 5.20, df = 2 (P = 0.07); I² = 62%
Test for overall effect: Z = 0.40 (P = 0.69)

Footnotes
(1) Adverse events; 2 years
(2) Adverse events; 7 years
(3) Adverse events; 7 years

Figure 40 PICO 5 long-term: Adverse events

#### 4.1.1.3.7 Serious adverse events

Two studies (Heller 2009, Hisey 2014) reported serious adverse event rates after a follow-up of 2 and 4 years. Overall, 149 events have been reported in 669 patients. There was no statistically significant difference between disc prostheses and fusion (RR 1.09, 95% CI 0.83 to 1.45; low quality of evidence; Figure 41). Heterogeneity was low ($I^2$=0%). No study reported serious adverse events at short-term (1 year) follow-up.



| Study or Subgroup | Experimental Events | Experimental Total | Control Events | Control Total | Weight | Risk Ratio M-H, Random, 95% CI |
|---|---|---|---|---|---|---|
| Heller 2009 (1) | 71 | 230 | 54 | 194 | 88.0% | 1.11 [0.82, 1.49] |
| Hisey 2014 (2) | 16 | 164 | 8 | 81 | 12.0% | 0.99 [0.44, 2.21] |
| **Total (95% CI)** | | **394** | | **275** | **100.0%** | **1.09 [0.83, 1.45]** |
| Total events | 87 | | 62 | | | |

Heterogeneity: Tau² = 0.00; Chi² = 0.07, df = 1 (P = 0.79); I² = 0%
Test for overall effect: Z = 0.63 (P = 0.53)

Footnotes
(1) Serious adverse events; 2 years
(2) Serious adverse events; 4 years

Figure 41 PICO 5 long-term: Serious adverse events

### *4.1.1.4 Subgroup analyses – PICO 5*

Subgroup and sensitivity analyses could only be considered for the outcomes radicular pain, neck pain, quality of life, function and reoperation rate which were reported by at least five studies. Only the two subgroup analyses prioritised highest a priori were considered.

*Subgroup 1: Patients only with radiculopathy of the lumbar spine vs. patients without neurological symptoms*

Overall, ten studies were in patients with neurologic symptoms at baseline (Coric 2011, Heller 2009, Hisey 2014, Mummaneni 2007, Nabhan 2007, Nabhan 2011, Phillips 2013, Porchet 2004, Rozankovic 2014, Zhang 2012). In four studies, the neurologic status was either not reported (Karabag 2014) or unclear, i.e. per eligibility criteria it was possible that patients had neurologic symptoms but proportions were not reported (Murrey 2009, Vaccaro 2013, Zhang 2014). A clear distinction between patients with or without neurological symptoms could not be made.

*Subgroup 2: Anterior fusion vs. posterior fusion vs. other*

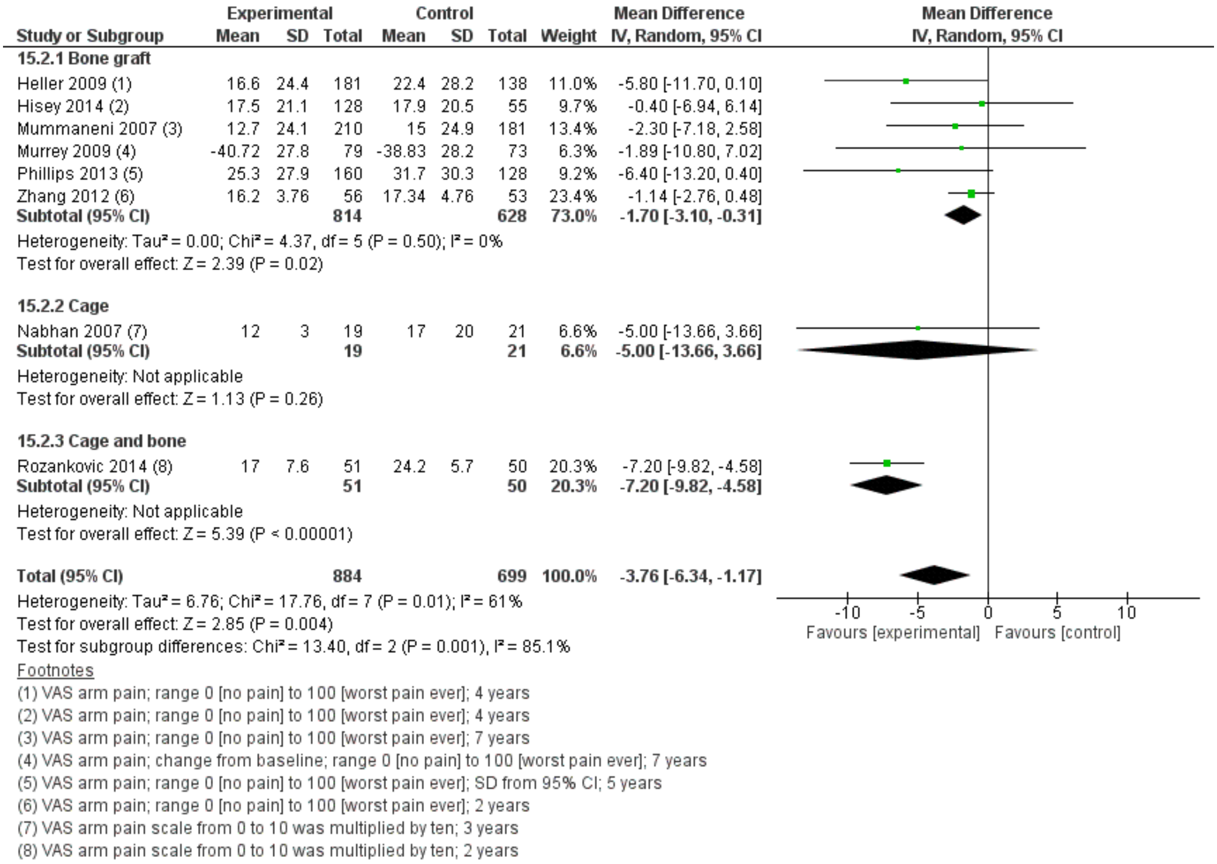
All fourteen studies used anterior discectomy and fusion (Coric 2011, Heller 2009, Hisey 2014, Karabag 2014, Mummaneni 2007, Murrey 2009, Nabhan 2007, Nabhan 2011, Philips 2013, Porchet

2004, Rozankovic 2014, Vaccaro 2013, Zhang 2012, Zhang 2014). A distinction between anterior and other types of fusion could not be made.

*Subgroup 3: Bone graft vs. fusion with cage*

Subgroup-analyses comparing bone graft vs. fusion with cage were possible for the outcomes radicular pain, neck pain, and function. The outcomes quality of life physical component score, quality of life mental component score, and reoperations were only reported in studies with bone grafts as comparator. Subgroup analyses for these outcomes were therefore not feasible.

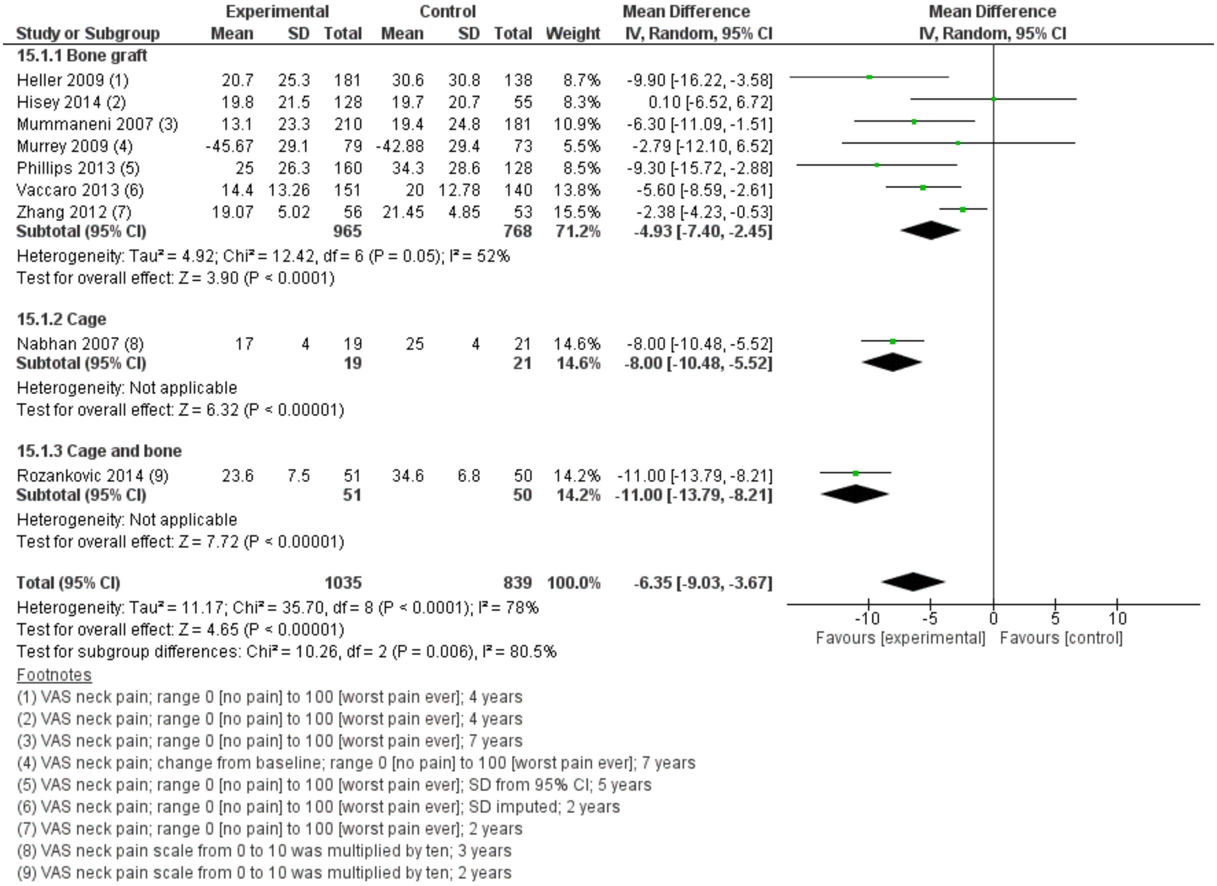
Some studies reported the use of cage and bone; this was added as a third subgroup. There were statistically significant differences between the three subgroups for the outcomes back pain and neck pain. These subgroup effects were based on one study for the subgroup cage and for the subgroup cage and bone. For the outcome function, there was no statistically significant difference between the three subgroups.



| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| **15.2.1 Bone graft** | | | | | | | | |
| Heller 2009 (1) | 16.6 | 24.4 | 181 | 22.4 | 28.2 | 138 | 11.0% | -5.80 [-11.70, 0.10] |
| Hisey 2014 (2) | 17.5 | 21.1 | 128 | 17.9 | 20.5 | 55 | 9.7% | -0.40 [-6.94, 6.14] |
| Mummaneni 2007 (3) | 12.7 | 24.1 | 210 | 15 | 24.9 | 181 | 13.4% | -2.30 [-7.18, 2.58] |
| Murrey 2009 (4) | -40.72 | 27.8 | 79 | -38.83 | 28.2 | 73 | 6.3% | -1.89 [-10.80, 7.02] |
| Phillips 2013 (5) | 25.3 | 27.9 | 160 | 31.7 | 30.3 | 128 | 9.2% | -6.40 [-13.20, 0.40] |
| Zhang 2012 (6) | 16.2 | 3.76 | 56 | 17.34 | 4.76 | 53 | 23.4% | -1.14 [-2.76, 0.48] |
| **Subtotal (95% CI)** | | | **814** | | | **628** | **73.0%** | **-1.70 [-3.10, -0.31]** |
| Heterogeneity: Tau² = 0.00; Chi² = 4.37, df = 5 (P = 0.50); I² = 0% | | | | | | | | |
| Test for overall effect: Z = 2.39 (P = 0.02) | | | | | | | | |
| **15.2.2 Cage** | | | | | | | | |
| Nabhan 2007 (7) | 12 | 3 | 19 | 17 | 20 | 21 | 6.6% | -5.00 [-13.66, 3.66] |
| **Subtotal (95% CI)** | | | **19** | | | **21** | **6.6%** | **-5.00 [-13.66, 3.66]** |
| Heterogeneity: Not applicable | | | | | | | | |
| Test for overall effect: Z = 1.13 (P = 0.26) | | | | | | | | |
| **15.2.3 Cage and bone** | | | | | | | | |
| Rozankovic 2014 (8) | 17 | 7.6 | 51 | 24.2 | 5.7 | 50 | 20.3% | -7.20 [-9.82, -4.58] |
| **Subtotal (95% CI)** | | | **51** | | | **50** | **20.3%** | **-7.20 [-9.82, -4.58]** |
| Heterogeneity: Not applicable | | | | | | | | |
| Test for overall effect: Z = 5.39 (P < 0.00001) | | | | | | | | |
| **Total (95% CI)** | | | **884** | | | **699** | **100.0%** | **-3.76 [-6.34, -1.17]** |
| Heterogeneity: Tau² = 6.76; Chi² = 17.76, df = 7 (P = 0.01); I² = 61% | | | | | | | | |
| Test for overall effect: Z = 2.85 (P = 0.004) | | | | | | | | |
| Test for subgroup differences: Chi² = 13.40, df = 2 (P = 0.001), I² = 85.1% | | | | | | | | |

Footnotes
(1) VAS arm pain; range 0 [no pain] to 100 [worst pain ever]; 4 years
(2) VAS arm pain; range 0 [no pain] to 100 [worst pain ever]; 4 years
(3) VAS arm pain; range 0 [no pain] to 100 [worst pain ever]; 7 years
(4) VAS arm pain; change from baseline; range 0 [no pain] to 100 [worst pain ever]; 7 years
(5) VAS arm pain; range 0 [no pain] to 100 [worst pain ever]; SD from 95% CI; 5 years
(6) VAS arm pain; range 0 [no pain] to 100 [worst pain ever]; 2 years
(7) VAS arm pain scale from 0 to 10 was multiplied by ten; 3 years
(8) VAS arm pain scale from 0 to 10 was multiplied by ten; 2 years

Figure 42 PICO 5 Subgroup analysis 3: Radicular pain, bone graft vs. fusion with cage



| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| **Study or Subgroup** | **Mean** | **SD** | **Total** | **Mean** | **SD** | **Total** | **Weight** | **IV, Random, 95% CI** |
| **15.1.1 Bone graft** | | | | | | | | |
| Heller 2009 (1) | 20.7 | 25.3 | 181 | 30.6 | 30.8 | 138 | 8.7% | -9.90 [-16.22, -3.58] |
| Hisey 2014 (2) | 19.8 | 21.5 | 128 | 19.7 | 20.7 | 55 | 8.3% | 0.10 [-6.52, 6.72] |
| Mummaneni 2007 (3) | 13.1 | 23.3 | 210 | 19.4 | 24.8 | 181 | 10.9% | -6.30 [-11.09, -1.51] |
| Murrey 2009 (4) | -45.67 | 29.1 | 79 | -42.88 | 29.4 | 73 | 5.5% | -2.79 [-12.10, 6.52] |
| Phillips 2013 (5) | 25 | 26.3 | 160 | 34.3 | 28.6 | 128 | 8.5% | -9.30 [-15.72, -2.88] |
| Vaccaro 2013 (6) | 14.4 | 13.26 | 151 | 20 | 12.78 | 140 | 13.8% | -5.60 [-8.59, -2.61] |
| Zhang 2012 (7) | 19.07 | 5.02 | 56 | 21.45 | 4.85 | 53 | 15.5% | -2.38 [-4.23, -0.53] |
| **Subtotal (95% CI)** | | | **965** | | | **768** | **71.2%** | **-4.93 [-7.40, -2.45]** |
| Heterogeneity: Tau² = 4.92; Chi² = 12.42, df = 6 (P = 0.05); I² = 52% | | | | | | | | |
| Test for overall effect: Z = 3.90 (P < 0.0001) | | | | | | | | |
| **15.1.2 Cage** | | | | | | | | |
| Nabhan 2007 (8) | 17 | 4 | 19 | 25 | 4 | 21 | 14.6% | -8.00 [-10.48, -5.52] |
| **Subtotal (95% CI)** | | | **19** | | | **21** | **14.6%** | **-8.00 [-10.48, -5.52]** |
| Heterogeneity: Not applicable | | | | | | | | |
| Test for overall effect: Z = 6.32 (P < 0.00001) | | | | | | | | |
| **15.1.3 Cage and bone** | | | | | | | | |
| Rozankovic 2014 (9) | 23.6 | 7.5 | 51 | 34.6 | 6.8 | 50 | 14.2% | -11.00 [-13.79, -8.21] |
| **Subtotal (95% CI)** | | | **51** | | | **50** | **14.2%** | **-11.00 [-13.79, -8.21]** |
| Heterogeneity: Not applicable | | | | | | | | |
| Test for overall effect: Z = 7.72 (P < 0.00001) | | | | | | | | |
| **Total (95% CI)** | | | **1035** | | | **839** | **100.0%** | **-6.35 [-9.03, -3.67]** |

Heterogeneity: Tau² = 11.17; Chi² = 35.70, df = 8 (P < 0.0001); I² = 78%

Test for overall effect: Z = 4.65 (P < 0.00001)

Test for subgroup differences: Chi² = 10.26, df = 2 (P = 0.006), I² = 80.5%

Footnotes

(1) VAS neck pain; range 0 [no pain] to 100 [worst pain ever]; 4 years

(2) VAS neck pain; range 0 [no pain] to 100 [worst pain ever]; 4 years

(3) VAS neck pain; range 0 [no pain] to 100 [worst pain ever]; 7 years

(4) VAS neck pain; change from baseline; range 0 [no pain] to 100 [worst pain ever]; 7 years

(5) VAS neck pain; range 0 [no pain] to 100 [worst pain ever]; SD from 95% CI; 5 years

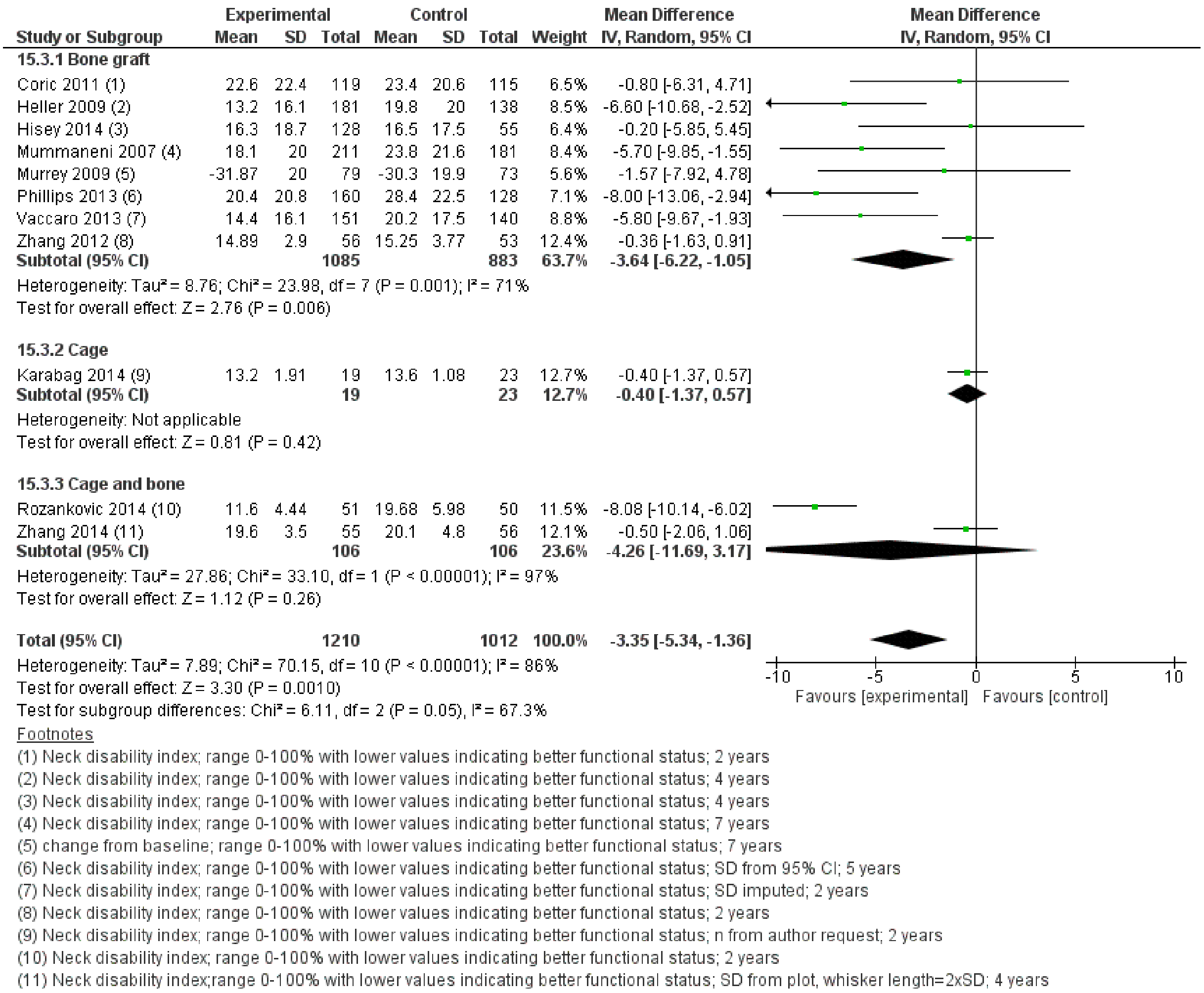
(6) VAS neck pain; range 0 [no pain] to 100 [worst pain ever]; SD imputed; 2 years

(7) VAS neck pain; range 0 [no pain] to 100 [worst pain ever]; 2 years

(8) VAS neck pain scale from 0 to 10 was multiplied by ten; 3 years

(9) VAS neck pain scale from 0 to 10 was multiplied by ten; 2 years

Figure 43 PICO 5 Subgroup analysis 3: Neck pain, bone graft vs. fusion with cage



|  | Experimental |  |  | Control |  |  |  | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| **15.3.1 Bone graft** |  |  |  |  |  |  |  |  |
| Coric 2011 (1) | 22.6 | 22.4 | 119 | 23.4 | 20.6 | 115 | 6.5% | -0.80 [-6.31, 4.71] |
| Heller 2009 (2) | 13.2 | 16.1 | 181 | 19.8 | 20 | 138 | 8.5% | -6.60 [-10.68, -2.52] |
| Hisey 2014 (3) | 16.3 | 18.7 | 128 | 16.5 | 17.5 | 55 | 6.4% | -0.20 [-5.85, 5.45] |
| Mummaneni 2007 (4) | 18.1 | 20 | 211 | 23.8 | 21.6 | 181 | 8.4% | -5.70 [-9.85, -1.55] |
| Murrey 2009 (5) | -31.87 | 20 | 79 | -30.3 | 19.9 | 73 | 5.6% | -1.57 [-7.92, 4.78] |
| Phillips 2013 (6) | 20.4 | 20.8 | 160 | 28.4 | 22.5 | 128 | 7.1% | -8.00 [-13.06, -2.94] |
| Vaccaro 2013 (7) | 14.4 | 16.1 | 151 | 20.2 | 17.5 | 140 | 8.8% | -5.80 [-9.67, -1.93] |
| Zhang 2012 (8) | 14.89 | 2.9 | 56 | 15.25 | 3.77 | 53 | 12.4% | -0.36 [-1.63, 0.91] |
| **Subtotal (95% CI)** |  |  | **1085** |  |  | **883** | **63.7%** | **-3.64 [-6.22, -1.05]** |
| Heterogeneity: Tau² = 8.76; Chi² = 23.98, df = 7 (P = 0.001); I² = 71% |  |  |  |  |  |  |  |  |
| Test for overall effect: Z = 2.76 (P = 0.006) |  |  |  |  |  |  |  |  |
| **15.3.2 Cage** |  |  |  |  |  |  |  |  |
| Karabag 2014 (9) | 13.2 | 1.91 | 19 | 13.6 | 1.08 | 23 | 12.7% | -0.40 [-1.37, 0.57] |
| **Subtotal (95% CI)** |  |  | **19** |  |  | **23** | **12.7%** | **-0.40 [-1.37, 0.57]** |
| Heterogeneity: Not applicable |  |  |  |  |  |  |  |  |
| Test for overall effect: Z = 0.81 (P = 0.42) |  |  |  |  |  |  |  |  |
| **15.3.3 Cage and bone** |  |  |  |  |  |  |  |  |
| Rozankovic 2014 (10) | 11.6 | 4.44 | 51 | 19.68 | 5.98 | 50 | 11.5% | -8.08 [-10.14, -6.02] |
| Zhang 2014 (11) | 19.6 | 3.5 | 55 | 20.1 | 4.8 | 56 | 12.1% | -0.50 [-2.06, 1.06] |
| **Subtotal (95% CI)** |  |  | **106** |  |  | **106** | **23.6%** | **-4.26 [-11.69, 3.17]** |
| Heterogeneity: Tau² = 27.86; Chi² = 33.10, df = 1 (P < 0.00001); I² = 97% |  |  |  |  |  |  |  |  |
| Test for overall effect: Z = 1.12 (P = 0.26) |  |  |  |  |  |  |  |  |
| **Total (95% CI)** |  |  | **1210** |  |  | **1012** | **100.0%** | **-3.35 [-5.34, -1.36]** |

Heterogeneity: Tau² = 7.89; Chi² = 70.15, df = 10 (P < 0.00001); I² = 86%
Test for overall effect: Z = 3.30 (P = 0.0010)
Test for subgroup differences: Chi² = 6.11, df = 2 (P = 0.05), I² = 67.3%

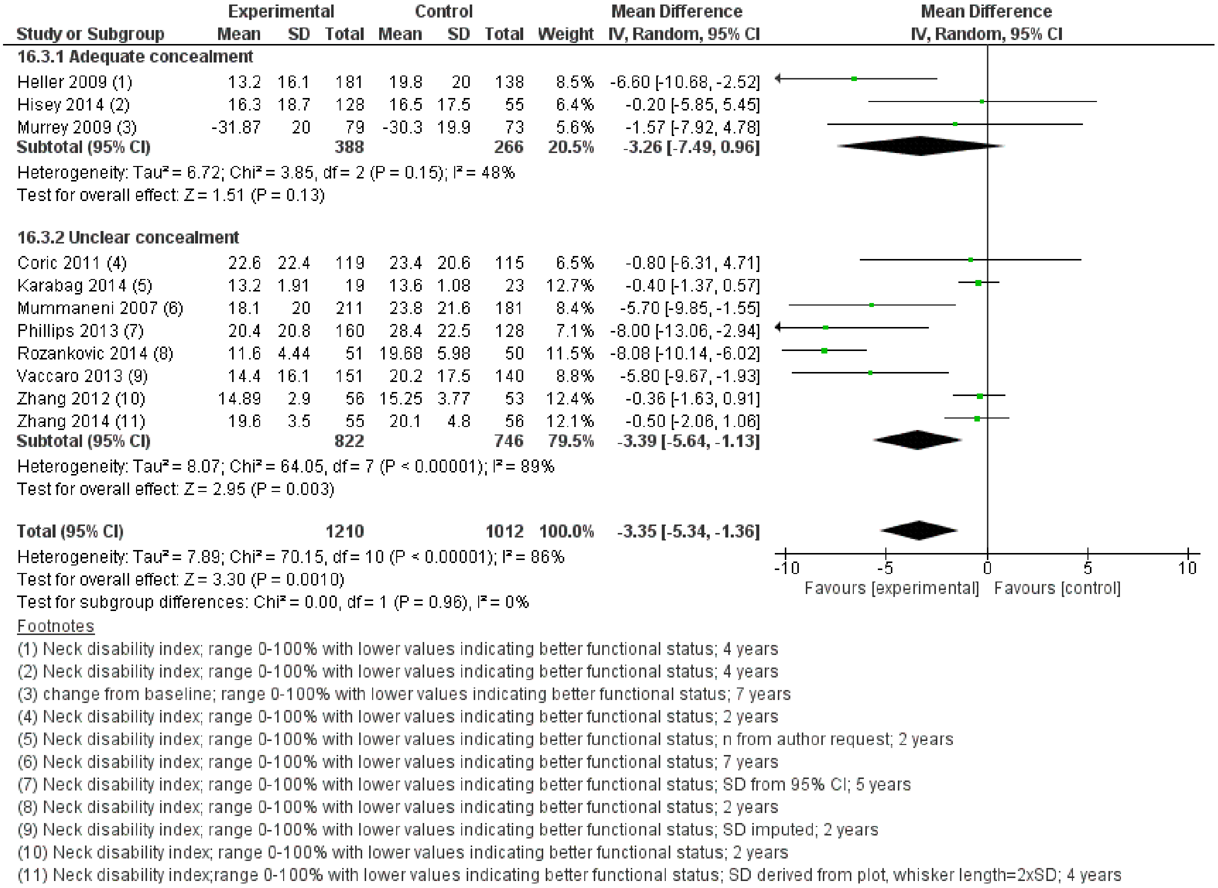
Footnotes

(1) Neck disability index; range 0-100% with lower values indicating better functional status; 2 years
(2) Neck disability index; range 0-100% with lower values indicating better functional status; 4 years
(3) Neck disability index; range 0-100% with lower values indicating better functional status; 4 years
(4) Neck disability index; range 0-100% with lower values indicating better functional status; 7 years
(5) change from baseline; range 0-100% with lower values indicating better functional status; 7 years
(6) Neck disability index; range 0-100% with lower values indicating better functional status; SD from 95% CI; 5 years
(7) Neck disability index; range 0-100% with lower values indicating better functional status; SD imputed; 2 years
(8) Neck disability index; range 0-100% with lower values indicating better functional status; 2 years
(9) Neck disability index; range 0-100% with lower values indicating better functional status; n from author request; 2 years
(10) Neck disability index; range 0-100% with lower values indicating better functional status; 2 years
(11) Neck disability index;range 0-100% with lower values indicating better functional status; SD from plot, whisker length=2xSD; 4 years

Figure 44 PICO 5 Subgroup analysis 3: Function, bone graft vs. fusion with cage

### *Subgroup 4: Adequate vs. no adequate or unclear allocation concealment*

Subgroup-analyses for adequate vs. no adequate or unclear allocation concealment were performed for function because more than 10 studies reported this outcome; and for quality of life physical component score, quality of life mental component score, and reoperation rate because the analysis for Subgroup 3 was not possible. As no study had no adequate allocation concealment, only adequate vs. unclear concealment were compared. No statistically significant differences were found between the two subgroups irrespective of the outcome.

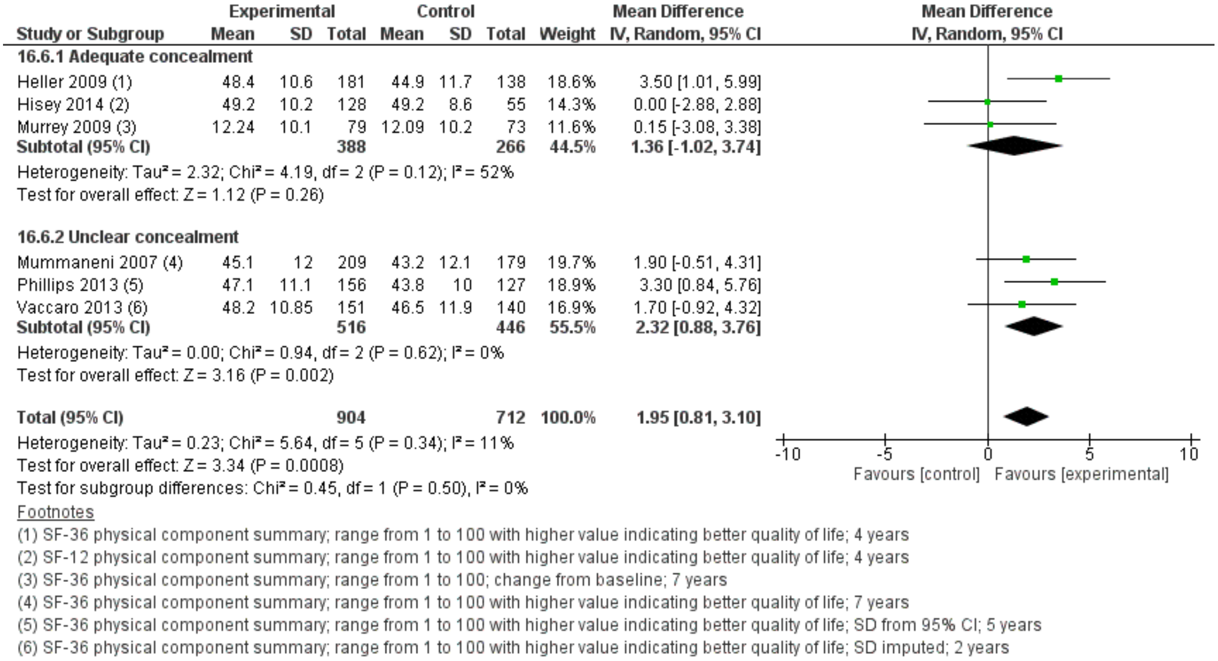


| Study or Subgroup | Experimental Mean | Experimental SD | Experimental Total | Control Mean | Control SD | Control Total | Weight | Mean Difference IV, Random, 95% CI |
|---|---|---|---|---|---|---|---|---|
| **16.3.1 Adequate concealment** | | | | | | | | |
| Heller 2009 (1) | 13.2 | 16.1 | 181 | 19.8 | 20 | 138 | 8.5% | -6.60 [-10.68, -2.52] |
| Hisey 2014 (2) | 16.3 | 18.7 | 128 | 16.5 | 17.5 | 55 | 6.4% | -0.20 [-5.85, 5.45] |
| Murrey 2009 (3) | -31.87 | 20 | 79 | -30.3 | 19.9 | 73 | 5.6% | -1.57 [-7.92, 4.78] |
| **Subtotal (95% CI)** | | | **388** | | | **266** | **20.5%** | **-3.26 [-7.49, 0.96]** |
| Heterogeneity: Tau² = 6.72; Chi² = 3.85, df = 2 (P = 0.15); I² = 48% | | | | | | | | |
| Test for overall effect: Z = 1.51 (P = 0.13) | | | | | | | | |
| **16.3.2 Unclear concealment** | | | | | | | | |
| Coric 2011 (4) | 22.6 | 22.4 | 119 | 23.4 | 20.6 | 115 | 6.5% | -0.80 [-6.31, 4.71] |
| Karabag 2014 (5) | 13.2 | 1.91 | 19 | 13.6 | 1.08 | 23 | 12.7% | -0.40 [-1.37, 0.57] |
| Mummaneni 2007 (6) | 18.1 | 20 | 211 | 23.8 | 21.6 | 181 | 8.4% | -5.70 [-9.85, -1.55] |
| Phillips 2013 (7) | 20.4 | 20.8 | 160 | 28.4 | 22.5 | 128 | 7.1% | -8.00 [-13.06, -2.94] |
| Rozankovic 2014 (8) | 11.6 | 4.44 | 51 | 19.68 | 5.98 | 50 | 11.5% | -8.08 [-10.14, -6.02] |
| Vaccaro 2013 (9) | 14.4 | 16.1 | 151 | 20.2 | 17.5 | 140 | 8.8% | -5.80 [-9.67, -1.93] |
| Zhang 2012 (10) | 14.89 | 2.9 | 56 | 15.25 | 3.77 | 53 | 12.4% | -0.36 [-1.63, 0.91] |
| Zhang 2014 (11) | 19.6 | 3.5 | 55 | 20.1 | 4.8 | 56 | 12.1% | -0.50 [-2.06, 1.06] |
| **Subtotal (95% CI)** | | | **822** | | | **746** | **79.5%** | **-3.39 [-5.64, -1.13]** |
| Heterogeneity: Tau² = 8.07; Chi² = 64.05, df = 7 (P < 0.00001); I² = 89% | | | | | | | | |
| Test for overall effect: Z = 2.95 (P = 0.003) | | | | | | | | |
| **Total (95% CI)** | | | **1210** | | | **1012** | **100.0%** | **-3.35 [-5.34, -1.36]** |
| Heterogeneity: Tau² = 7.89; Chi² = 70.15, df = 10 (P < 0.00001); I² = 86% | | | | | | | | |
| Test for overall effect: Z = 3.30 (P = 0.0010) | | | | | | | | |
| Test for subgroup differences: Chi² = 0.00, df = 1 (P = 0.96), I² = 0% | | | | | | | | |

Footnotes

(1) Neck disability index; range 0-100% with lower values indicating better functional status; 4 years
(2) Neck disability index; range 0-100% with lower values indicating better functional status; 4 years
(3) change from baseline; range 0-100% with lower values indicating better functional status; 7 years
(4) Neck disability index; range 0-100% with lower values indicating better functional status; 2 years
(5) Neck disability index; range 0-100% with lower values indicating better functional status; n from author request; 2 years
(6) Neck disability index; range 0-100% with lower values indicating better functional status; 7 years
(7) Neck disability index; range 0-100% with lower values indicating better functional status; SD from 95% CI; 5 years
(8) Neck disability index; range 0-100% with lower values indicating better functional status; 2 years
(9) Neck disability index; range 0-100% with lower values indicating better functional status; SD imputed; 2 years
(10) Neck disability index; range 0-100% with lower values indicating better functional status; 2 years
(11) Neck disability index;range 0-100% with lower values indicating better functional status; SD derived from plot, whisker length=2xSD; 4 years

Figure 45 PICO 5 Subgroup analysis 3: Function, adequate vs. no adequate or unclear allocation concealment

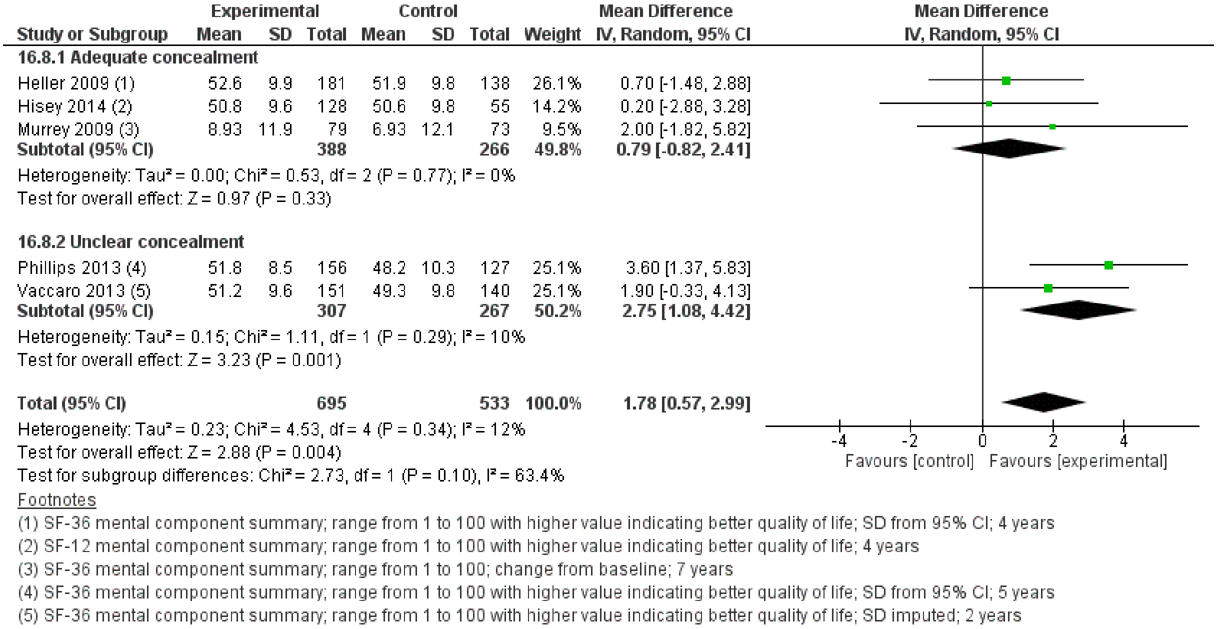


| Study or Subgroup | Experimental Mean | Experimental SD | Experimental Total | Control Mean | Control SD | Control Total | Weight | Mean Difference IV, Random, 95% CI |
|---|---|---|---|---|---|---|---|---|
| **16.6.1 Adequate concealment** | | | | | | | | |
| Heller 2009 (1) | 48.4 | 10.6 | 181 | 44.9 | 11.7 | 138 | 18.6% | 3.50 [1.01, 5.99] |
| Hisey 2014 (2) | 49.2 | 10.2 | 128 | 49.2 | 8.6 | 55 | 14.3% | 0.00 [-2.88, 2.88] |
| Murrey 2009 (3) | 12.24 | 10.1 | 79 | 12.09 | 10.2 | 73 | 11.6% | 0.15 [-3.08, 3.38] |
| **Subtotal (95% CI)** | | | **388** | | | **266** | **44.5%** | **1.36 [-1.02, 3.74]** |
| Heterogeneity: Tau² = 2.32; Chi² = 4.19, df = 2 (P = 0.12); I² = 52% | | | | | | | | |
| Test for overall effect: Z = 1.12 (P = 0.26) | | | | | | | | |
| **16.6.2 Unclear concealment** | | | | | | | | |
| Mummaneni 2007 (4) | 45.1 | 12 | 209 | 43.2 | 12.1 | 179 | 19.7% | 1.90 [-0.51, 4.31] |
| Phillips 2013 (5) | 47.1 | 11.1 | 156 | 43.8 | 10 | 127 | 18.9% | 3.30 [0.84, 5.76] |
| Vaccaro 2013 (6) | 48.2 | 10.85 | 151 | 46.5 | 11.9 | 140 | 16.9% | 1.70 [-0.92, 4.32] |
| **Subtotal (95% CI)** | | | **516** | | | **446** | **55.5%** | **2.32 [0.88, 3.76]** |
| Heterogeneity: Tau² = 0.00; Chi² = 0.94, df = 2 (P = 0.62); I² = 0% | | | | | | | | |
| Test for overall effect: Z = 3.16 (P = 0.002) | | | | | | | | |
| **Total (95% CI)** | | | **904** | | | **712** | **100.0%** | **1.95 [0.81, 3.10]** |
| Heterogeneity: Tau² = 0.23; Chi² = 5.64, df = 5 (P = 0.34); I² = 11% | | | | | | | | |
| Test for overall effect: Z = 3.34 (P = 0.0008) | | | | | | | | |
| Test for subgroup differences: Chi² = 0.45, df = 1 (P = 0.50), I² = 0% | | | | | | | | |

Footnotes

(1) SF-36 physical component summary; range from 1 to 100 with higher value indicating better quality of life; 4 years
(2) SF-12 physical component summary; range from 1 to 100 with higher value indicating better quality of life; 4 years
(3) SF-36 physical component summary; range from 1 to 100; change from baseline; 7 years
(4) SF-36 physical component summary; range from 1 to 100 with higher value indicating better quality of life; 7 years
(5) SF-36 physical component summary; range from 1 to 100 with higher value indicating better quality of life; SD from 95% CI; 5 years
(6) SF-36 physical component summary; range from 1 to 100 with higher value indicating better quality of life; SD imputed; 2 years

Figure 46 PICO 5 Subgroup analysis 3: Quality of life physical component score, adequate vs. no adequate or unclear allocation concealment



| Study or Subgroup | Experimental Mean | Experimental SD | Experimental Total | Control Mean | Control SD | Control Total | Weight | Mean Difference IV, Random, 95% CI |
|---|---|---|---|---|---|---|---|---|
| **16.8.1 Adequate concealment** | | | | | | | | |
| Heller 2009 (1) | 52.6 | 9.9 | 181 | 51.9 | 9.8 | 138 | 26.1% | 0.70 [-1.48, 2.88] |
| Hisey 2014 (2) | 50.8 | 9.6 | 128 | 50.6 | 9.8 | 55 | 14.2% | 0.20 [-2.88, 3.28] |
| Murrey 2009 (3) | 8.93 | 11.9 | 79 | 6.93 | 12.1 | 73 | 9.5% | 2.00 [-1.82, 5.82] |
| **Subtotal (95% CI)** | | | **388** | | | **266** | **49.8%** | **0.79 [-0.82, 2.41]** |
| Heterogeneity: Tau² = 0.00; Chi² = 0.53, df = 2 (P = 0.77); I² = 0% | | | | | | | | |
| Test for overall effect: Z = 0.97 (P = 0.33) | | | | | | | | |
| **16.8.2 Unclear concealment** | | | | | | | | |
| Phillips 2013 (4) | 51.8 | 8.5 | 156 | 48.2 | 10.3 | 127 | 25.1% | 3.60 [1.37, 5.83] |
| Vaccaro 2013 (5) | 51.2 | 9.6 | 151 | 49.3 | 9.8 | 140 | 25.1% | 1.90 [-0.33, 4.13] |
| **Subtotal (95% CI)** | | | **307** | | | **267** | **50.2%** | **2.75 [1.08, 4.42]** |
| Heterogeneity: Tau² = 0.15; Chi² = 1.11, df = 1 (P = 0.29); I² = 10% | | | | | | | | |
| Test for overall effect: Z = 3.23 (P = 0.001) | | | | | | | | |
| **Total (95% CI)** | | | **695** | | | **533** | **100.0%** | **1.78 [0.57, 2.99]** |
| Heterogeneity: Tau² = 0.23; Chi² = 4.53, df = 4 (P = 0.34); I² = 12% | | | | | | | | |
| Test for overall effect: Z = 2.88 (P = 0.004) | | | | | | | | |
| Test for subgroup differences: Chi² = 2.73, df = 1 (P = 0.10), I² = 63.4% | | | | | | | | |

Footnotes

(1) SF-36 mental component summary; range from 1 to 100 with higher value indicating better quality of life; SD from 95% CI; 4 years

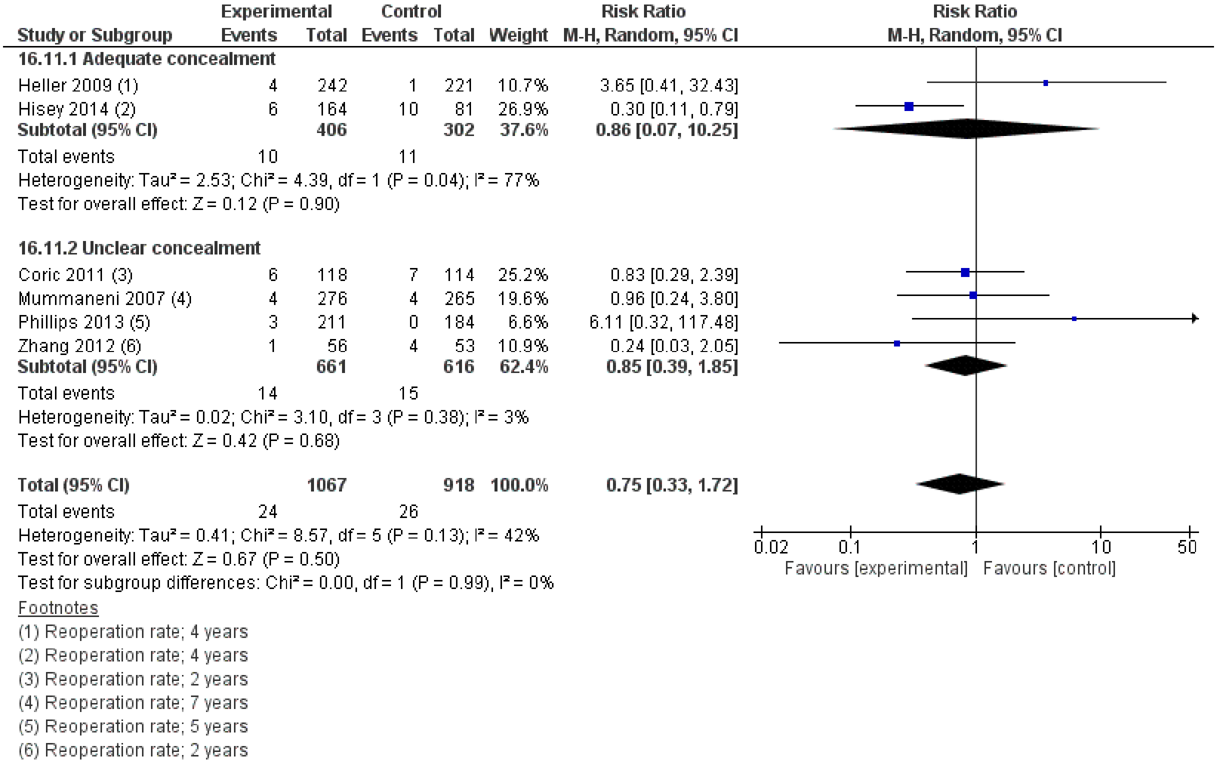
(2) SF-12 mental component summary; range from 1 to 100 with higher value indicating better quality of life; 4 years

(3) SF-36 mental component summary; range from 1 to 100; change from baseline; 7 years

(4) SF-36 mental component summary; range from 1 to 100 with higher value indicating better quality of life; SD from 95% CI; 5 years

(5) SF-36 mental component summary; range from 1 to 100 with higher value indicating better quality of life; SD imputed; 2 years

**Figure 47 PICO 5 Subgroup analysis 3: Quality of life mental component score, adequate vs. no adequate or unclear allocation concealment**



| Study or Subgroup | Experimental Events | Experimental Total | Control Events | Control Total | Weight | Risk Ratio M-H, Random, 95% CI |
|---|---|---|---|---|---|---|
| **16.11.1 Adequate concealment** | | | | | | |
| Heller 2009 (1) | 4 | 242 | 1 | 221 | 10.7% | 3.65 [0.41, 32.43] |
| Hisey 2014 (2) | 6 | 164 | 10 | 81 | 26.9% | 0.30 [0.11, 0.79] |
| **Subtotal (95% CI)** | | **406** | | **302** | **37.6%** | **0.86 [0.07, 10.25]** |
| Total events | 10 | | 11 | | | |
| Heterogeneity: Tau² = 2.53; Chi² = 4.39, df = 1 (P = 0.04); I² = 77% | | | | | | |
| Test for overall effect: Z = 0.12 (P = 0.90) | | | | | | |
| **16.11.2 Unclear concealment** | | | | | | |
| Coric 2011 (3) | 6 | 118 | 7 | 114 | 25.2% | 0.83 [0.29, 2.39] |
| Mummaneni 2007 (4) | 4 | 276 | 4 | 265 | 19.6% | 0.96 [0.24, 3.80] |
| Phillips 2013 (5) | 3 | 211 | 0 | 184 | 6.6% | 6.11 [0.32, 117.48] |
| Zhang 2012 (6) | 1 | 56 | 4 | 53 | 10.9% | 0.24 [0.03, 2.05] |
| **Subtotal (95% CI)** | | **661** | | **616** | **62.4%** | **0.85 [0.39, 1.85]** |
| Total events | 14 | | 15 | | | |
| Heterogeneity: Tau² = 0.02; Chi² = 3.10, df = 3 (P = 0.38); I² = 3% | | | | | | |
| Test for overall effect: Z = 0.42 (P = 0.68) | | | | | | |
| **Total (95% CI)** | | **1067** | | **918** | **100.0%** | **0.75 [0.33, 1.72]** |
| Total events | 24 | | 26 | | | |
| Heterogeneity: Tau² = 0.41; Chi² = 8.57, df = 5 (P = 0.13); I² = 42% | | | | | | |
| Test for overall effect: Z = 0.67 (P = 0.50) | | | | | | |
| Test for subgroup differences: Chi² = 0.00, df = 1 (P = 0.99), I² = 0% | | | | | | |

Footnotes

(1) Reoperation rate; 4 years

(2) Reoperation rate; 4 years

(3) Reoperation rate; 2 years

(4) Reoperation rate; 7 years

(5) Reoperation rate; 5 years

(6) Reoperation rate; 2 years

**Figure 48 PICO 5 Subgroup analysis 3: Reoperation rate, adequate vs. no adequate or unclear allocation concealment**

### 4.1.1.5 GRADE – PICO 5

Table 27 PICO 5 long-term: Summary of findings (GRADE)

**PICO 5 - Disc prosthesis (with or without decompression) compared to fusion (with or without decompression) for patients with neck pain with or without neurological symptoms due to degenerative changes of the cervical spine**

**Patient or population:** patients with neck pain with or without neurological symptoms due to degenerative changes of the cervical spine
**Intervention:** Disc prosthesis (with or without decompression)
**Comparison:** Fusion (with or without decompression)

| Outcomes | Illustrative comparative risks* (95% CI) | | Relative effect (95% CI) | No of Participants (studies) | Quality of the evidence (GRADE) | Comments |
|---|---|---|---|---|---|---|
| | Assumed risk | Corresponding risk | | | | |
| | **Fusion (with or without decompression)** | **PICO 5 - Disc prosthesis (with or without decompression)** | | | | |
| **Radicular Pain** | | The mean radicular pain in the intervention groups was **3.76 lower** (6.37 to 1.17 lower) | | 1583 (8 studies) | ⊕⊕⊕⊝ **moderate**[1,2,3] | |
| **Myelopathy** | **Study population** | | Not estimable | 0 (0) | See comment | No RCT reported this outcome |
| | See comment | See comment | | | | |
| | **Moderate** | | | | | |
| **Neck pain**<br>VAS | | The mean neck pain in the intervention groups was **6.35 lower** (9.03 to 3.67 lower) | | 1874 (9 studies) | ⊕⊕⊝⊝ **low**[4,5,6] | |
| **Quality of life - physical component score**<br>SF-36 and SF-12 | | The mean quality of life - physical component score in the intervention groups was **1.95 higher** (0.81 to 3.1 higher) | | 1616 (6 studies) | ⊕⊕⊕⊝ **moderate**[7,8] | |
| **Quality of life - mental** | | The mean quality of life - mental component score in the | | 1228 | ⊕⊕⊕⊝ | |

| | | | | | |
|---|---|---|---|---|---|
| **component score**<br>SF-36 and SF-12 | | intervention groups was<br>**1.78 higher**<br>(0.57 to 2.99 higher) | | (5 studies) | **moderate**[9,10] |
| **Function**<br>ODI etc | | The mean function in the intervention groups was<br>**3.50 lower**<br>(5.77 to 1.23 lower) | | 2222<br>(11 studies) | ⊕⊕⊕⊖<br>**moderate**[5,11,12] |
| **Revision rate** | **10 per 1000** | **3 per 1000**<br>(1 to 16) | **RR 0.31**<br>(0.06 to 1.57) | 1004<br>(2 studies) | ⊕⊖⊖⊖<br>**very low**[13,14,15] |
| **Reoperation rate** | **Study population** | | **RR 0.75**<br>(0.33 to 1.72) | 1985<br>(6 studies) | ⊕⊖⊖⊖<br>**very low**[5,16,17] |
| | **28 per 1000** | **21 per 1000**<br>(9 to 49) | | | |
| | **Moderate** | | | | |
| **Complication rate and adverse events** | **Study population** | | **RR 0.93**<br>(0.63 to 1.35) | 982<br>(3 studies) | ⊕⊖⊖⊖<br>**very low**[5,18,19] |
| | **567 per 1000** | **527 per 1000**<br>(357 to 765) | | | |
| | **Moderate** | | | | |
| **Serious adverse events** | **Study population** | | **RR 1.09**<br>(0.83 to 1.45) | 669<br>(2 studies) | ⊕⊕⊖⊖<br>**low**[20,21] |
| | **225 per 1000** | **246 per 1000**<br>(187 to 327) | | | |
| | **Moderate** | | | | |

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).

**CI:** Confidence interval; **RR:** Risk ratio;

GRADE Working Group grades of evidence
**High quality:** Further research is very unlikely to change our confidence in the estimate of effect.

**Moderate quality:** Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.

**Low quality:** Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.

**Very low quality:** We are very uncertain about the estimate.

[1] Risk of selection bias (random sequence generation and allocation concealment) was unclear in 7 studies; risk of performance bias was unclear in 4 and high in 5 studies; risk of detection bias was unclear in 9 studies; risk of attrition bias was unclear in 1 and high in 6 studies; risk of reporting bias was unclear in 3 studies.

[2] Inconsistency was downgraded because heterogeneity was high.

[3] Imprecision was not downgraded because OIS was sufficient.[4] Risk of selection bias (random sequence generation and allocation concealment) was unclear in 6 studies; risk of performance bias was unclear in 3 and high in 5 studies; risk of detection bias was unclear in 8 studies; risk of attrition bias was high in 6 studies; risk of reporting bias was unclear in 2 studies.

[5] Inconsistency was downgraded by one level because heterogeneity ($I^2$) was high, and there was minimal or no overlap of confidence intervals of the individual studies. Heterogeneity could not be explained by sensitivity analysis.

[6] Imprecision was not downgraded because OIS was sufficient.

[7] Risk of selection bias (random sequence generation and allocation concealment) was unclear in 4 studies; risk of performance bias was high in 6 studies; risk of detection bias was unclear in 6 studies; risk of attrition bias was unclear 1 and high in 5 studies; risk of reporting bias was unclear in 2 studies.

[8] Imprecision was not downgraded because OIS was sufficient.

[9] Risk of selection bias (random sequence generation and allocation concealment) was unclear in 3 studies; risk of performance bias was high in 5 studies; risk of detection bias was unclear in 5 studies; risk of attrition bias was unclear in 1 and high in 4 studies; risk of reporting bias was unclear in 2 studies.

[10] Imprecision was not downgraded because OIS was sufficient

[11] Risk of selection bias (random sequence generation and allocation concealment) was unclear in 9 studies; risk of performance bias was unclear in 5 and high in 6 studies; risk of detection bias was unclear in 11 studies; risk of attrition bias was unclear in 3 and high in 6 studies; risk of reporting bias was unclear in 5 studies.

[12] Imprecision was not downgraded because OIS was sufficient

[13] Risk of selection bias (allocation concealment) was unclear in 2 studies; risk of performance bias was high in 2 studies; risk of detection bias was unclear in 2 studies; risk of attrition bias was high in 2 studies.

[14] Inconsistency was downgraded by one level because heterogeneity ($I^2$) was high. Because of the limited number of studies it was not possible to perform sensitivity or subgroup analysis to assess reasons for heterogeneity.

[15] Imprecision was downgraded by two levels because the 95% CI included appreciable harm (greater than 25% relative risk increase) and benefit (greater than 25% relative risk reduction).

[16] Risk of selection bias (random sequence generation and allocation concealment) was unclear in 5 studies; risk of performance bias was unclear in 2 and high in 4 studies; risk of detection bias was unclear in 6 studies; risk of attrition bias was unclear in 1 and high in 5 studies; risk of reporting bias was unclear in 3 studies.

[17] Imprecision was downgraded by two levels because the 95% CI included appreciable harm (greater than 25% relative risk increase) and benefit (greater than 25% relative risk reduction).

[18] Risk of selection bias (random sequence generation and allocation concealment) was unclear in 2 studies; risk of performance bias was unclear in 1 and high in 2 studies; risk of detection bias was unclear in 3 studies; risk of attrition bias was high in 3 studies; risk of reporting bias was unclear in 1 studies.

[19] Imprecision was downgraded by two levels because the 95% CI included appreciable harm (greater than 25% relative risk increase) and benefit (greater than 25% relative risk reduction).

[20] Risk of selection bias (random sequence generation and allocation concealment) was unclear in 1 study; risk of performance bias was high in 2 studies; risk of detection bias was unclear in 2 studies; risk of attrition bias was high in 2 studies; risk of reporting bias was unclear in 1 study.

[21] Imprecision was downgraded by one level because the 95% CI included appreciable harm (greater than 25% relative risk increase) or no effect.

# 5 Ongoing studies

CinicalTrials.gov was searched for any ongoing (recruiting) eligible RCTs (date of search 20 June 2016; search details in Appendix I). Two reviewers independently screened the results. Of 62 registered trials, 1 matched the eligibility criteria. This trial is eligible for PICO 5. It is registered under the identifier NCT02417272. They openly randomise total cervical disc replacement with degenerative CP ESP® and anterior cervical decompression and fusion in adult patients with symptomatic cervical degenerative disc disease and neurological symptoms (planned enrolment of 110 patients). Pre-specified outcomes relevant for our research questions are complication rate, function (Neck Disability Index), quality of life, cervical and radicular pain. The planned follow-up is 2 years. The estimated study completion date is May 2019.

# 6 Summary of results

## 6.1 Interspinous and pedicle-based devices

### 6.1.1 Summary of results – PICO 1

For PICO 1 - comparing interspinous stabilisation without direct decompression to direct decompression - in a population with neurological symptoms due to degenerative changes of the lumbar spine, three randomised controlled studies were included for long-term follow-up. For interspinous stabilisation without direct decompression compared to direct decompression quality of life (EQ-5D MD 0.04, 95% CI 0.02 to 0.06; important outcome, low quality of the evidence) was statistically significantly higher but there was also a statistically significantly higher risk of reoperations (RR 3.02, 95% CI 1.75 to 5.22; important outcome, low quality of the evidence). There was no statistically significant effect on back pain, radicular pain, spinal claudication, function and adverse events. No study reported revision rates or serious adverse events.

The **overall quality of evidence** for the outcomes of PICO 1 was judged to be very low.

### 6.1.2 Summary of results – PICO 2

For PICO 2 - comparing interspinous stabilisation with direct decompression to direct decompression - in a population with neurological symptoms due to degenerative changes of the lumbar spine, one randomised controlled study was included for long-term follow-up. There was no statistically significant effect for interspinous devices with direct decompression compared to direct decompression only (PICO 2) on back pain (VAS MD -0.80, 95% CI -2.31 to 0.71; critical outcome, very low quality of the evidence) and function (ODI MD -8.70, 95% CI -19.91 to 2.51; important outcome, low quality of the evidence). Zero events were reported for complications. No studies were available for radicular pain, spinal claudication, quality of life, revision rate, reoperation rate and serious adverse events. No study examined pedicle-based stabilisation.

The **overall quality of evidence** for the outcomes of PICO 2 was judged to be very low.

### 6.1.3 Summary of results – PICO 3

For PICO 3 - comparing interspinous or pedicle-based stabilisation to fusion with implants - in a population with symptoms due to degenerative changes of the lumbar spine with or without neurological symptoms, two studies were included for long-term follow-up. Only the outcome function was reported by both studies. There was no statistically significant effect for interspinous or pedicle-based stabilisation with direct decompression compared to fusion (PICO 3) on function based on two studies (Davis 2013, Madan 2003). Only one study (Davis 2013) reported back pain, radicular pain, spinal claudication, and reoperation rate. For all these outcomes no statistically significant effect was found. One study (Madan 2003) reported revision rate and adverse events. For both outcomes the effects of pedicle-based stabilisation compared to fusion with implants were not statistically significant. The quality of the evidence was low or very low for all outcomes in PICO 3. No studies were available for quality of life and serious adverse events.

The **overall quality of evidence** for the outcomes of PICO 3 was judged to be low.

## 6.2 Disc prostheses

### 6.2.1 Summary of results – PICO 4

For PICO 4 - comparing lumbar disc prosthesis compared to fusion - in a population with symptoms due to degenerative changes of the lumbar spine with or without neurological symptoms, six studies were eligible. At long-term follow-up, there were statistically significant effects of disc prosthesis compared to fusion for back pain (VAS MD -5.60, 95% CI -10.47 to -0.73; important outcome, low quality of the evidence), quality of life physical component score (SF-36 MD 2.77, 95% CI 0.85 to 4.70; important outcome, low quality of the evidence), and function (ODI MD -5.19, 95% CI -7.67 to -2.71; important outcome, moderate quality of the evidence). There were no statistically significant effects for radicular pain, mental component of quality of life, reoperation rate, and serious adverse events. The quality of evidence was low or very low for these outcomes. There was as well no statistically significant effect on adverse events with moderate quality of evidence. Only one study (Gornet 2011) reported on revision rate, but the number of events was zero in both groups.

The **overall quality of evidence** for the outcomes of PICO 4 was judged to be low.

### 6.2.2 Summary of results – PICO 5

For PICO 5 - cervical disc prosthesis compared to cervical fusion - in a population with symptoms due to degenerative changes of the cervical spine with or without neurological symptoms, fourteen studies were eligible. At long-term follow-up, there were statistically significant effects for cervical disc prosthesis compared to fusion for radicular pain (VAS MD -3.76, 95% CI -6.34 to -1.17; critical outcome, moderate quality of the evidence), neck pain (SF-36 MD -6.35, 95% CI -9.03 to -3.67; important outcome, low quality of the evidence), quality of life physical component score (SF-36 MD 1.95, 95% CI 0.81 to 3.10; important outcomes, moderate quality of the evidence) and mental component scores (SF-36 MD 1.78, 95% CI 0.57 to 2.99; important outcomes, moderate quality of the evidence), and function (NDI MD -3.50, 95% CI -5.77 to -1.23; important outcome, moderate quality of the evidence). For PICO 5, there were no statistically significant effects for revision rate, reoperation rate, adverse events, and serious adverse events. The quality of evidence was low or very low for these outcomes. No study reported myelopathy.

The **overall quality of evidence** for the outcomes of PICO 5 was judged to be very low because of no evidence for the critical outcome myelopathy.

# 7 Discussion

This report addressed five different PICO-questions. Three questions assessed the clinical effectiveness and safety of interspinous and pedicle-based stabilisation devices in patients with symptoms due to degenerative changes of the lumbar spine. The other two questions of this report assessed disc prostheses in patients with symptoms due to degenerative changes of the lumbar and cervical spine, respectively.

Dynamic stabilisation has been suggested as an alternative to fusion in order to avoid or reduce adjacent segment disease (ASD) that can develop as a consequence of spinal fusion due to mechanical stress[2]. ASD can manifest in various ways, for example as instability, discus hernia, scoliosis, vertebral compression fracture[2]. ASD was not assessed for this report as it is difficult to establish the diagnosis because the associated changes can also occur secondary to degenerative changes[2]. The presence or absence of radiological findings also does not necessarily constitute a patient relevant outcome, as these changes may be asymptomatic and never require treatment. Instead this report focussed on patient-relevant outcomes like pain and function rather than trying to establish the presence or absence of ASD.

## 7.1 General methodological issues

The confidence in an effect estimate decreases with study limitations (risk of bias, inconsistency, indirectness, imprecision or publication bias). The following paragraphs discuss the five categories of reasons rating down quality of evidence according to GRADE.

**Risk of bias**

When assessing risk of bias, several limitations were noted for all five PICO-questions, especially inadequate blinding and attrition bias. Inadequate blinding of participants and outcome assessors were major concerns in almost all studies included in this report. The feasibility of blinding in surgical RCTs is challenging but not impossible[29] and inadequate blinding contributed to substantial uncertainty of the empirical evidence[30]. Attrition bias also contributed to high uncertainty of the evidence. In several studies, the number of patients with missing data was unclear or not reported. Reporting of the number of patients randomised was absent in seven of all included studies. Moreover, eight studies (five of them in PICO 5) reported missing data of more than 20% at the long-term follow-up. Furthermore, at long-term follow-up, substantial differences in missing data between treatment arms (differences of more than 5% between arms) were a problem in six of twelve studies in PICO 5 and in four of five studies in PICO 4 and it cannot be excluded that this is related to the inadequate blinding of the randomised treatment assignment.

**Inconsistency**

Unexplained heterogeneity decreased the confidence in the effect estimates. Heterogeneity (high $I^2$) was a lesser problem among the studies in PICO 2 and PICO 3 where most outcomes were based on only one study. In PICO 1, 4 and 5 it seemed that some studies tended to be responsible for heterogeneity for one or more outcomes even though no study characteristic could be identified that was responsible for the differences of effects. For example, in PICO 1 the study by Lønne 2015 seemed to favour dynamic stabilisation over direct decompression while the two other studies tended to favour direct decompression (see outcomes: back pain, radicular pain, spinal claudication and function). This caused substantial heterogeneity which may be due by differences in the populations, interventions, outcome assessment but could also be due to differences in study methods (performance bias). Possible reasons for heterogeneity are discussed in more detail in the discussions of PICO 1-3 and of PICO 4-5.

**Indirectness**

No serious indirectness was identified within this health technology assessment.

**Imprecision**

Imprecision was judged to be serious if the total sample size was lower than the optimal information size (OIS), if the number of events was less than 300 (only binary outcomes), or if the 95% confidence interval included the possibility of both the "no effect line" and "clinically relevant benefit"[14]. The same applies for harm. If the confidence interval included both, the possibility of "clinically relevant benefit" and "clinically relevant harm", imprecision was judged to be very serious.

The IQWiG in Germany recently published a paper suggesting that the additional benefit of an intervention should be judged mainly based on the relative risk for binary data in their reports rather than focussing on absolute risks as suggested by GRADE[31]. So far the Swiss Federal Office of Public Health has no pre-defined criteria for the definition of clinically relevant benefit or harm and hence simple measures suggested in the methodological literature were used in the assessment of imprecision.

Depending on the effect measures obtained, different criteria were used to estimate clinically relevant benefit and harm. In the case of standardised mean differences, a cut off of 0.5 was used[24], in the case of risk ratios, effects were judged to be clinically relevant if they were ≥ 1.25 or ≤ 0.75[14]. For non-standardised continuous data, minimal clinically important differences (MCIDs) were based on cut-offs in the published literature on research on spinal disease and back pain. The used MCIDS were found based on a non-systematic search in the literature. All these estimates can only serve as rule of thumb.

Particular care is necessary in the interpretation of MCIDs. The definition of MCIDs is not straightforward and a variety of methods have been suggested to define cut-offs for specifying minimally clinically important difference (MCID) or minimally important difference yielding estimates that can vary widely[32-35]. In addition, the used MCIDs from the literature have been developed in order to determine the clinical relevance of treatment effects in individual patients compared to baseline and not to determine the relevance of the difference in treatment effect between groups at follow-up. One reason is that it is much more difficult to determine it than clinically important differences in individual patients compared to baseline[36].

Therefore, it has to be noted that the judgment based on the MCIDs should not be used as definitive solution and the usage of different MCIDs is possible. It is possible that in a study where two effective treatments are being compared, like in this report, the MCID between study groups could be smaller, because it is only the incremental improvement of one over the other intervention.

Therefore when MCIDs are being used to judge the clinical relevance of differences of effects instead then one cannot simply assume that all the patients profit if the effect estimate lies above the MCID and no patient profits if the effect estimate lies below the MCID[33,36-39]. Even when the estimate of a group difference lies below the MCID, it is still possible that a considerable number of patients profits from a treatment. More accurate estimates of the clinical relevance of differences between groups could be obtained by interpreting MCIDs in conjunction with responder rates or using other methods to estimate the proportion of patients (see Johnston et al. and Dworkin et al. for a more detailed discussion)[36-39].

Depending on the threshold used to define clinical relevance for a specific outcome, the rating of imprecision could change. Imprecision was judged per outcome without considering the balance of effects in comparison with other outcomes in order to avoid making value judgments, which fall into the remit of the decision makers (see also paragraph "Important considerations for the decision makers and guideline developers")[24].

For most outcomes in PICO 1-3, imprecision was judged serious because total sample size was lower than the optimal information size (OIS). In PICO 4, the OIS was insufficient or large confidence intervals included clinically important effects as well as no effects, and therefore imprecision was judged serious for most outcomes. The number of events for the outcomes revision, reoperation and serious adverse events, if reported for a PICO-question, were generally small. Therefore, for these outcomes, imprecision was judged serious in all PICOs. In contrast, the number of adverse events reported for PICO 4 and 5 was sufficiently high. Thus, the confidence intervals were sufficiently narrow and it was not rated down for imprecision.

**Publication bias**

For PICO 1-4 the number of available studies was too small to assess the risk for publication bias. For PICO 5, no indication of publications bias was found.

**Limitations of the methods used in this report**

For the long-term data, all the outcomes were extracted for the time point for which the largest number of relevant outcomes was reported. If the number of reported outcomes was the same for two different time points, the time point with the longest follow-up was taken. This approach was chosen due to time constraints and in view of the large number of publications per study, which would have required more in depth evaluation if the data for the longest available time point had been taken for each outcome. As a consequence, it is possible that data for a later time point were ignored or that data for additional outcomes would have been available but at other time points.

**Important considerations for the decision makers and guideline developers**

The quality of evidence was assessed with GRADE for each specific outcome from the perspective of a systematic review author. Decision makers and guideline authors, however, are advised to reassess the quality of evidence for all important and critical outcomes to make an overall rating of the quality of evidence as this is an iterative process. An overall rating may differ from the outcome specific ratings[19] as presented in this report. Importantly, systematic review authors defined outcome specific thresholds (minimal clinically important difference) to rate imprecision[19]. These thresholds should be carefully evaluated by decision makers and may need to be adapted based on the balance and magnitude of the effects of other outcomes based on their values and preferences (for example, if for a PICO-question one outcome had a clinically important benefit but at the same time another outcome for the same PICO-question had a clinically important harm)[40,41].

## 7.2 Interspinous and pedicle-based devices - PICO 1-3

Heterogeneity between selected studies, due to differences in populations, interventions and outcome assessments may influence the pooled results. Some aspects that may have contributed to these differences are described in the following sections.

**Population**

In PICO 1 to 3, the populations had slightly different diagnoses and mean age. Patients in studies for PICO 1 and 2 all had spinal stenosis (respectively 3 studies and 1 study) while patients in the studies for PICO 3 either suffered from spinal stenosis (1 study) or disc degeneration (2 studies). Reporting regarding the presence of neurological symptoms was very poor. Therefore, for PICO 1 and 2, where the presence of these symptoms was mandatory, it was assumed that a sufficient proportion of the patients fitted the inclusion criteria though this cannot be proven. Age differed more between studies for PICO 3 than for PICO 1 and 2. Both interspinous devices and pedicle-based stabilisation is used in patients with spinal disease with the aim of reducing adjacent segment disease. Interspinous devices can be implanted using minimally invasive techniques and are therefore particularly interesting in elderly patients[42]. In keeping with this indication the mean age in the studies on interspinous devices tended to be higher (range 56-71 years) than in the studies on pedicle-based stabilisation (mean age 45 years). The indication for surgery differed between the two studies: in the study on pedicle-based stabilisation patients had disc degeneration, including patients with leg pain, (Madan 2003), and in the study with interspinous stabilisation, the patients had spinal stenosis (Davis 2013). These differences could influence the pooled results and should be taken into account when interpreting them.

**Intervention and comparator**

All studies included for PICO 1 and 2 used interspinous devices. For PICO 3, there was only one outcome (function) with evidence on both interspinous and pedicle-based stabilisation devices. Furthermore, it may be important to consider that the type of devices used differed between studies. In PICO 1, interspinous stabilisation was done with the Coflex or X-stop devices. In PICO 2, interspinous stabilisation was done with the Wallis implant. In PICO 3, Coflex was used for interspinous stabilisation and the Graf ligamentoplasty was used for pedicle-based stabilisation.

In the only study for PICO 2 (Marsh 2014) and the study on pedicle-based stabilisation for PICO 3 (Madan 2003), all surgical interventions were performed by the same senior surgeon. This could limit the external validity of these studies. The other studies were multicentre RCTs and so a limitation of the external validity is less likely for these studies.

At least one of the included studies (Davis 2013) reported surgeon training, i.e. the surgeons were learning the new surgical technique prior to the study. In another study, the surgeons were already “experienced in both techniques” (Moojen 2013). Surgical interventions, especially new techniques, have steep learning curves[43,44]. One of the included studies discussed that these training cases might affect their results and in particular the need for reoperation and revisions (Davis 2013) and that with increasing experience the latter were declining. The need for training cases may be different though depending on the device and surgical technique used, and whether surgeons are already familiar and experienced with the surgical technique. While in the study by Davis 2013 the authors felt that the need to learn affected their results in the intervention arm (treatment with Coflex), they did not observe such a learning curve with the comparator arm (fusion with pedicle screws). On the other hand, the authors of the Strömqvist 2013 study felt that surgery with the X-Stop device was so simple that no training beyond the supervision during the first intervention was needed. Information on the handling of training or number of training cases in the different studies was not systematically searched or extracted as the interpretation of this kind of information would be difficult – even if it were available: Irrespective of possibly varying need for training depending on the device, it is possible that despite training cases surgeons did not reach the plateau of the learning curve required for the respective surgery in each study and that in other cases where no training cases were reported, surgeons already had the required experience with the technique[43,44].

**Outcomes**

Pain was assessed separately depending on whether it was local (lower back) or radiated into an extremity, i.e. the leg. Although all the included studies reported on pain, this meant that the results from Madan 2003 where pain was reported irrespective of its location were not included. Pooling scores on overall pain with data on pain depending on location would have been difficult though as pain data depending on location would then have to be considered as clustered data and the question would have arisen of how they compare to an "overall" pain score.

The pooled effect estimate of radicular pain included only data on leg pain. However, leg pain is not necessarily the same as radicular pain and it is possible to have pain radiating into the leg without nerve root compression. Therefore, when authors reported on "leg pain" it was not clear whether this corresponded to "radicular pain" or not, but was interpreted as such.

To assess the harm of an intervention, general adverse events or complications were pooled. By definition this does not necessitate causality between the adverse outcome and the intervention. In contrast, adverse effects and other like surgery-related complications assume causality and this can be quite subjective and thereby more strongly affected by bias – especially in unblinded studies. Only one study for each PICO-question reported on adverse events and only very a small absolute number of events. Serious adverse events were reported for none of the three PICO-questions. For each PICO-question the number of reoperations was reported though not by all included studies. The definitions of adverse events seemed to vary between the included studies. For instance reoperations were not always counted as adverse event. For example, the study by Lønne et al. for PICO 1 reported a higher number of reoperations, than adverse events. In most instances (apart from the Study by Marsh et al. for PICO 2, which found neither adverse events nor reoperations) different studies reported on adverse events than on reoperations. Only one study reported on revisions (Madan 2003 for PICO 3); for the other PICO-questions no data on revisions were available. Hence the evidence on adverse effects, reoperations and revisions was rather poor (see also Section 6.1).

## 7.3 Disc prostheses - PICO 4-5

Differences in populations, interventions and outcome assessment between the included studies may cause heterogeneity which would influence the pooled results. Some aspects that may have contributed to these differences are described in the following section.

**Population**

The inclusion criteria for PICO 4 and 5 were broad including patients with and without neurological symptoms. Often, the included studies, too, allowed for patients with or without neurological symptoms and an exact number of patients with neurologic symptoms could not be determined due to of lack of reporting.

**Intervention and comparator**

The prosthesis types differed between studies. Broadly two types of devices could be distinguished based on the materials used: devices made out of metal and polyethylene and metal-on-metal devices. In PICO 4, four different devices were used in the six included studies (Charité III, Maverick, FlexiCore, or ProDiscL). One of study investigating the Maverick prosthesis (Gornet 2011) had a strong impact on the overall results as it had a very big weight in the statistical analysis, when pooled with other studies and in many instances was the only study reporting an outcome. This study used a metal-on-metal device. In PICO 5, nine different devices were investigated in the fourteen included studies (Kineflex C, Bryan, Mobi C, PRESTIGE ST, ProDisc C, Porous Coated Motion Cervical Disc, Prestige II, Discover, SECURE-C). The number of studies was small compared to the number of devices investigated preventing the investigation of device dependent effects.

The fusion in the control arm was done using cages and bone in PICO 4 and bone, or cage, or cage and bone in PICO 5. In PICO 4, one of the included studies (Gornet 2011) stated that fusion techniques using iliac bone graft did not correspond to the current standard because of the second surgical site leading to extra-pain and longer disability. However, more recent literature suggested that there is no evidence that other methods, like fusions with bone from the surgical site or morphogenetic proteins, are superior[45,46].

Similar to PICO 1-3, the issue of prior or in-study training was also relevant for PICO 4 and 5. Information on training was not systematically extracted but a brief check regarding this issue only revealed the study by Moreno 2008 for PICO 4 who attributed differences in the duration of the operation and the complication rate between their study and referred to another study (Blumenthal 2005) to the fact that in this study the surgeons still needed to learn the technique. While Zigler 2007 described that surgeons had training cases before patients were randomised into the study, Gornet 2011 described that surgical manuals as well as hands-on cadaver training was provided and that the study design did not include a surgical training phase, i.e. that the results are based on all the patients treated. Blumenthal et al. described that adequate training of the surgeons is a pre-requisite for the reproducibility of their results in clinical practice. In PICO 5, training cases were reported by Hisey 2014, and Mummaneni 2007 reported that surgeons received training.

**Outcomes**

The instruments used to assess outcomes differed less among studies on disc prostheses compared to the studies on interspinous or pedicle-based devices, and therefore it was not necessary to calculate standardised mean differences for the pooled outcomes. Radicular or neck pain were always assessed with a visual analogue scale (VAS) ranging either from 0 to 100 or from 0 to 10. Function was measured with the Oswestry Disability Index for PICO 4 and the Neck Disability Index for PICO 5. If quality of life was reported, it was reported by SF-12 or SF-36, always presenting at least one of the two component scores.

Pain was only extracted when it was reported separately for local (neck or back) or pain in the extremities (leg or arm). The same issues discussed for PICO 1-3 regarding pain reporting applies for PICO 4 and 5. For PICO 4 only one of six studies reported on long-term radicular pain and back pain

(Gornet 2011) even though all studies assessed pain as an outcome but either they reported only short-term data (Strube 2016) or assessed overall pain, without differentiating between the different types of pain (Blumenthal 2005, Moreno 2008, Sasso 2008, Ziegler 2007). It was not formally assessed whether all the information would have been available in order to pool these data but at least in some instances information on the variation of the data (confidence interval, standard deviations or standard errors) seemed to be missing. In PICO 5, the proportion of separate reporting for neck and neck pain is higher with eight and nine studies of 14 included studies respectively reporting long-term data for the outcomes arm pain and neck pain that were poolable. The outcome data of studies reporting arm pain were pooled for the outcome radicular pain though strictly speaking arm pain and radicular pain are not necessarily the same. It is possible to have pain radiating into the arm without having nerve compression (radicular pain). However, all studies in PICO 5 reported arm pain, without specifying whether this corresponded to radicular pain. Given the context and the difficulty to correctly identify the source of arm pain, arm pain was interpreted as radicular pain.

Quality of life, measured with the SF-36-Questionnaire, was only reported as mental and physical component score. An overall summary score was not reported. Therefore, the mental and physical component scores were pooled separately. Interestingly, one of two studies for PICO 4, and one of six studies for PICO 5 reported only the physical component score although the mental component score at baseline was assessed. Possibly, the authors decided to report only the physical component score because they considered it to be more important, or maybe no differences were measured in the mental health component score. In theory this could be judged as selective reporting, but we were not that strict in this case. It is unclear how this might influence the pooled results and any conclusions on quality of life, as both physical and mental components are vital when measuring this outcome[47].

For function in PICO 5 there is a substantial uncertainty because effect estimates of the individual studies seemed inconsistent and heterogeneity ($I^2$=86%) was high. From the forest plot, it was apparent that some studies clearly showed a significant effect (Heller 2009, Mummaneni 2007, Phillips 2013, Rozankovic 2014, Vaccaro 2013) while others did not (Coric 2011, Hisey 2014, Karabag 2014, Murrey 2009, Zhang 2012, Zhang 2014). All variables (country, enrolment period, setting, follow-up, eligibility criteria, neurologic symptoms, device technology, cage/bone, affected levels, sex, age, bias) extracted for this PICO-question would have been considered for post-hoc sensitivity analysis but no obvious groups were identified. Also, the pre-specified sub-groups (comparator fusion with bone graft vs. fusion with cage; adequate vs. no adequate or unclear allocation concealment; adequate vs. inadequate or unclear randomization; complete vs. incomplete, imputed or unclear outcome data) were applied to assess the heterogeneity of function. No explanations for the high heterogeneity were identified, but this does not exclude differences in effects in so far unidentified subgroups for the outcome function. The uncertainty due to the high $I^2$ affects our confidence in the effect estimate which does or does not include a clinically relevant effect.

As only general adverse events or complications were pooled, no conclusions can be drawn for surgery- and implant-related adverse events as these are different outcomes. In the present report, a non-significant relative risk (RR 0.93, 95% CI 0.63 to 1.35) was reported. The RR correspond to an odds ratio of 1.02, 95% CI 0.43 to 2.39. In contrast, a recent meta-analysis[48] pooled results from 15 RCTs on surgery- and implant-related adverse events. The authors found a statistically significant and clinical relevant lower rate of surgery related adverse events in the cervical disc arthroplasty group compared to anterior cervical discectomy and fusion (odds ratio 0.58, 95% CI 0.46 to 0.73)[48]. Official criteria have been developed in order to judge the causality between interventions and adverse events[49] but the advantage of the assessment of the overall number of adverse effects is that they neither depend on the adequate use of those criteria nor on the existing knowledge of the causality. The adverse event and complication rates differed greatly from study to study. For example, in PICO 4, two studies reported that 76% or 86%, respectively, of the patients had an adverse event whereas the other two studies reported that only 6% and 8%, respectively, had an adverse event. A similar picture was observed in PICO 5. This large discrepancy may be explained by different definitions used for adverse events and complications. As definitions are usually not reported no such explanations could be further assessed.

Serious adverse events were extracted if they have been termed as such. For serious adverse events, too, only general serious adverse events were extracted as these are usually well defined. Specific serious adverse events like surgery- and implant-related adverse events were not extracted as these are different outcomes and prone to subjectivity. Hence, this might be an explanation for the differing numbers of serious adverse events between studies.

Subgroup analyses were conducted for patients with fusion with bone graft, with cage, or with cage and bone graft for the outcomes radicular pain, neck pain and function. Although there was an indication for a difference in effect between the three subgroups (statistically significant for radicular and local neck pain), the analysis was based on only one study in the subgroups cage and cage and bone and is therefore not meaningful.

The between-study heterogeneity was high for the three outcomes radicular pain, neck pain and function. For the outcome radicular pain, the stratification according to subgroups reduced $I^2$ to 0 in the studies in patients with bone grafts suggesting that the studies here were less heterogeneous. However, the heterogeneity in the other strata and outcomes could either not be assessed, as only one study was available, or remained high. Hence, the reduced statistical heterogeneity for the outcome radicular pain could just be due to chance.

# 8 Conclusions

For surgical interventions inserting interspinous or pedicle-based devices without decompression compared to direct decompression, a statistically significant better improvement of quality of life was found but also a statistically significant increase of the relative risk of reoperations (PICO 1).

For interspinous or pedicle-based devices with direct decompression either compared to direct decompression only (PICO 2) or to fusion (PICO 3), the effects for any outcome were either not statistically significant or no data were available.

The overall quality of the evidence was very low for PICO 1 and 2 and low for PICO 3.

In patients with degenerative changes of the lumbar (PICO 4) or cervical (PICO 5) spine, a statistically significant improvement was found for back or neck pain, function and physical quality of life for disc prosthesis compared to fusion with implants. In addition, for PICO 5, a statistically significant improvement in radicular pain and mental quality of life was found that could not be observed for PICO 4. For all other outcomes, effects were either not statistically significantly different or had not been reported. The overall quality of the evidence was low for PICO 4 and very low for PICO 5.

The overall quality of the evidence (based on the quality of the evidence for the critical outcomes) is similar for all PICO-questions. However, considerably more studies were identified for PICO 4 and in particular for PICO 5 than for PICO 1, 2 and 3. Major limitations of the quality of the evidence were most frequently due to risk of bias, unexplained heterogeneity (inconsistency) and imprecision.

The evaluation of the quality of the evidence should be re-considered in the context of decision making where values and preferences regarding aspects like the balance of benefit and harm, and costs can affect the appraisal of the available evidence and its quality.

## Reference list

1. Moojen WA, Arts MP, Bartels RH, Jacobs WC, Peul WC. Effectiveness of interspinous implant surgery in patients with intermittent neurogenic claudication: a systematic review and meta-analysis. *Eur Spine J.* 2011;20(10):1596-1606.
2. Virk SS, Niedermeier S, Yu E, Khan SN. Adjacent segment disease. *Orthopedics.* 2014;37(8):547-555.
3. Eck JC, Sharan A, Ghogawala Z, et al. Guideline update for the performance of fusion procedures for degenerative disease of the lumbar spine. Part 7: lumbar fusion for intractable low-back pain without stenosis or spondylolisthesis. *J Neurosurg Spine.* 2014;21(1):42-47.
4. Wang JC, Dailey AT, Mummaneni PV, et al. Guideline update for the performance of fusion procedures for degenerative disease of the lumbar spine. Part 8: lumbar fusion for disc herniation and radiculopathy. *J Neurosurg Spine.* 2014;21(1):48-53.
5. Resnick DK, Watters WC, 3rd, Mummaneni PV, et al. Guideline update for the performance of fusion procedures for degenerative disease of the lumbar spine. Part 10: lumbar fusion for stenosis without spondylolisthesis. *J Neurosurg Spine.* 2014;21(1):62-66.
6. Resnick DK, Watters WC, 3rd, Sharan A, et al. Guideline update for the performance of fusion procedures for degenerative disease of the lumbar spine. Part 9: lumbar fusion for stenosis with spondylolisthesis. *J Neurosurg Spine.* 2014;21(1):54-61.
7. Groff MW, Dailey AT, Ghogawala Z, et al. Guideline update for the performance of fusion procedures for degenerative disease of the lumbar spine. Part 12: pedicle screw fixation as an adjunct to posterolateral fusion. *J Neurosurg Spine.* 2014;21(1):75-78.
8. Overdevest GM, Moojen WA, Arts MP, Vleggeert-Lankamp CL, Jacobs WC, Peul WC. Management of lumbar spinal stenosis: a survey among Dutch spine surgeons. *Acta Neurochir (Wien).* 2014;156(11):2139-2145.
9. Guyatt G, Oxman AD, Akl E, et al. GRADE guidelines 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol.* 2010.
10. Guyatt G, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. *J Clin Epidemiol.* 2010.
11. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol.* 2011.
12. Guyatt G, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence-study limitations (risk of bias) and publication bias. *J Clin Epidemiol.* 2011.
13. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *J Clin Epidemiol.* 2011;64(12):1277-1282.
14. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. *J Clin Epidemiol.* 2011;64(12):1283-1293.
15. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *J Clin Epidemiol.* 2011;64(12):1294-1302.
16. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *J Clin Epidemiol.* 2011;64(12):1303-1310.
17. Guyatt GH, Oxman AD, Sultan S, et al. GRADE guidelines: 9. Rating up the quality of evidence. *J.Clin.Epidemiol.* 2011;64(12):1311-1316.
18. Brunetti M, Shemilt I, Pregno S, et al. GRADE guidelines: 10. Considering resource use and rating the quality of economic evidence. *J Clin Epidemiol.* 2012.
19. Guyatt G, Oxman AD, Sultan S, et al. GRADE guidelines 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. *J Clin Epidemiol.* 2012.

20. Guyatt GH, Oxman AD, Santesso N, et al. GRADE guidelines 12. Preparing Summary of Findings tables-binary outcomes. *J Clin Epidemiol.* 2012.
21. Guyatt GH, Thorlund K, Oxman AD, et al. GRADE guidelines: 13. Preparing Summary of Findings tables and evidence profiles - Continuous outcomes. *Journal of Clinical Epidemiology.* 2013;66(2):173-183.
22. Andrews J, Guyatt G, Oxman AD, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. *J Clin Epidemiol.* 2013;66(7):719-725.
23. Andrews JC, Schunemann HJ, Oxman AD, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. *J Clin Epidemiol.* 2013;66(7):726-735.
24. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions.* Vol Version 5.1.0. www.cochrane.org/training/cochrane-handbook The Cochrane Collaboration last edited 20 March 2011
25. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled clinical trials.* 1986;7(3):177-188.
26. Reichenbach S, Sterchi R, Scherer M, et al. Meta-analysis: chondroitin for osteoarthritis of the knee or hip. *Ann Intern Med.* 2007;146(8):580-590.
27. Juni P, Reichenbach S, Dieppe P. Osteoarthritis: rational approach to treating the individual. *Best Pract Res Clin Rheumatol.* 2006;20(4):721-740.
28. Cohen J. *Statistical power analysis for the behavioral sciences. 2nd edition. Hillsdale, NJ: Lawrence Earlbaum Associates.* 1988.
29. Boutron I, Guittet L, Estellat C, Moher D, Hrobjartsson A, Ravaud P. Reporting methods of blinding in randomized trials assessing nonpharmacological treatments. *PLoS Med.* 2007;4(2):e61.
30. 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(21):1660-1673.
31. Skipka G, Wieseler B, Kaiser T, et al. Methodological approach to determine minor, considerable, and major treatment effects in the early benefit assessment of new drugs. *Biom J.* 2016;58(1):43-58.
32. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA.* 1994;271(5):389-391.
33. Cook JA, Hislop J, Altman DG, et al. Specifying the target difference in the primary outcome for a randomised controlled trial: guidance for researchers. *Trials.* 2015;16:12.
34. Hislop J, Adewuyi TE, Vale LD, et al. Methods for specifying the target difference in a randomised controlled trial: the Difference ELicitation in TriAls (DELTA) systematic review. *PLoS Med.* 2014;11(5):e1001645.
35. Treadwell J, Uhl S, Tipton K, et al. *Assessing Equivalence and Noninferiority*. Rockville (MD)2012.
36. Dworkin RH, Turk DC, McDermott MP, et al. Interpreting the clinical importance of group differences in chronic pain clinical trials: IMMPACT recommendations. *Pain.* 2009;146(3):238-244.
37. Johnston BC, Thorlund K, Schunemann HJ, et al. Improving the interpretation of quality of life evidence in meta-analyses: the application of minimal important difference units. *Health Qual Life Outcomes.* 2010;8:116.
38. Johnston BC, Patrick DL, Busse JW, Schunemann HJ, Agarwal A, Guyatt GH. Patient-reported outcomes in meta-analyses--Part 1: assessing risk of bias and combining outcomes. *Health Qual Life Outcomes.* 2013;11:109.

39. Johnston BC, Patrick DL, Thorlund K, et al. Patient-reported outcomes in meta-analyses-part 2: methods for improving interpretability for decision-makers. *Health Qual Life Outcomes.* 2013;11:211.

40. Schünemann HJ. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? *J Clin Epidemiol.* 2016;75:6-15.

41. Anttila S, Persson J, Vareman N, Sahlin NE. Conclusiveness resolves the conflict between quality of evidence and imprecision in GRADE. *J Clin Epidemiol.* 2016;75:1-5.

42. Moojen WA, Arts MP, Jacobs WC, et al. Interspinous process device versus standard conventional surgical decompression for lumbar spinal stenosis: randomized controlled trial. *BMJ.* 2013;347:f6415.

43. Pannucci CJ, Wilkins EG. Identifying and Avoiding Bias in Research. *Plastic and reconstructive surgery.* 2010;126(2):619-625.

44. Devereaux PJ, Bhandari M, Clarke M, et al. Need for expertise based randomised controlled trials. *BMJ : British Medical Journal.* 2005;330(7482):88-88.

45. Fu R, Selph S, McDonagh M, et al. Effectiveness and harms of recombinant human bone morphogenetic protein-2 in spine fusion: a systematic review and meta-analysis. *Ann Intern Med.* 2013;158(12):890-902.

46. Simmonds MC, Brown JV, Heirs MK, et al. Safety and effectiveness of recombinant human bone morphogenetic protein-2 for spinal fusion: a meta-analysis of individual-participant data. *Ann Intern Med.* 2013;158(12):877-889.

47. Post MWM. Definitions of Quality of Life: What Has Happened and How to Move On. *Topics in Spinal Cord Injury Rehabilitation.* 2014;20(3):167-180.

48. Rao MJ, Nie SP, Xiao BW, Zhang GH, Gan XR, Cao SS. Cervical disc arthroplasty versus anterior cervical discectomy and fusion for treatment of symptomatic cervical disc disease: a meta-analysis of randomized controlled trials. *Arch Orthop Trauma Surg.* 2015;135(1):19-28.

49. Ioannidis JP, Evans SJ, Gotzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med.* 2004;141(10):781-788.

# Appendices

## I. Appendix – Search strategy for Pubmed/Medline and ClinicalTrials.gov

### Search strategy for Pubmed/MEDLINE

| Search string | Hits<br>19.04.2016 |
|---|---|
| (((Spinal diseases[mh] OR "spondylolisthesis"[MeSH Terms] OR spondylolisthesis [Title/Abstract] OR "Adjacent segment disease" [Title/Abstract] OR Spondylarthrosis [Title/Abstract] OR Spondyloarthrosis [Title/Abstract] OR Spondylarthropathy [Title/Abstract] OR Spondyloarthropathy [Title/Abstract] OR "spondylarthropathies"[MeSH Terms] OR Cervicoarthrosis[Title/Abstract] OR "Osteoarthritis, Spine"[Mesh] OR "intervertebral disc degeneration"[MeSH Terms] OR spondylosis [Title/Abstract] OR "spondylosis"[MeSH Terms] OR myelopathy [Title/Abstract] OR (neurogenic[Title/Abstract] AND claudication[Title/Abstract]) OR Pseudoclaudication [Title/Abstract] OR pseudo-claudication [Title/Abstract] OR "spinal stenosis"[MeSH Terms] OR Radiculopathy [Title/Abstract] OR radiculopathies [Title/Abstract] OR sciatic [Title/Abstract] OR sciatica [Title/Abstract] OR "sciatica"[MeSH Terms] OR Ischialgia [Title/Abstract] OR Dorsalgia [Title/Abstract] OR Cervicobrachialgia [Title/Abstract] OR Backache [Title/Abstract] OR "back pain"[MeSH Terms] OR Lumbago [Title/Abstract] OR ((referred [Title/Abstract] OR radiating[Title/Abstract] OR radiated[Title/Abstract] OR radicular [Title/Abstract]) AND pain [tiab] ) OR ((lumbar [Title/Abstract] OR lumbal [Title/Abstract] OR lumbo [Title/Abstract] OR sacral [Title/Abstract] OR lumbosacral [Title/Abstract] OR lumbo-sacral [Title/Abstract] OR intervertebral [Title/Abstract] OR vertebral [Title/Abstract] OR vertebra [Title/Abstract] OR cervical [Title/Abstract] OR cervicobrachial [Title/Abstract] OR neck [Title/Abstract] OR back [Title/Abstract] OR leg [Title/Abstract] OR arm [Title/Abstract] OR spinal [Title/Abstract] OR spine) AND (syndrome [Title/Abstract] OR pain [Title/Abstract] OR arthritis [Title/Abstract] OR (nerve [Title/Abstract] AND irritat* [Title/Abstract]) OR degenerated [Title/Abstract] OR degeneration [Title/Abstract] OR degenerative [Title/Abstract])) OR ((spinal [Title/Abstract] OR spine [Title/Abstract] OR root [Title/Abstract] OR canal [Title/Abstract] OR lateral recess [Title/Abstract] OR lateral recesses [Title/Abstract] OR foraminal [Title/Abstract] OR foramina [Title/Abstract] OR foramen[Title/Abstract] ) AND (stenosis [Title/Abstract] OR stenoses [Title/Abstract] OR constriction [Title/Abstract] OR constrictions [Title/Abstract] OR constricted [Title/Abstract] OR compression [Title/Abstract] OR compressed[Title/Abstract] )) OR "intervertebral disc displacement"[MeSH Terms] OR ((disc [Title/Abstract] OR discs [Title/Abstract] OR disk[Title/Abstract] OR disks[Title/Abstract]) AND (hernia [Title/Abstract] OR herniated [Title/Abstract] OR slip[Title/Abstract] OR slipped [Title/Abstract] OR prolapse [Title/Abstract] OR prolapsed [Title/Abstract] OR sclerosis[Title/Abstract] OR rupture[Title/Abstract] OR ruptured[Title/Abstract] OR displaced [Title/Abstract] OR displacement [Title/Abstract])))) AND ((Dynamic[tiab] AND stabili*[tiab]) OR ((Interspinous [Title/Abstract] OR spinal [Title/Abstract] OR spine[Title/Abstract]) AND (spacer [Title/Abstract] OR spacers [Title/Abstract] OR device [Title/Abstract] OR devices [Title/Abstract] OR decompression [Title/Abstract] OR process [Title/Abstract] OR processes [Title/Abstract])) OR ((Pedicle [Title/Abstract] OR bone[Title/Abstract]) AND (screws[Title/Abstract] OR screw [Title/Abstract] OR plate[Title/Abstract] OR plates[Title/Abstract])) OR ((Disc [Title/Abstract] OR discs [Title/Abstract] OR disk [Title/Abstract] OR disks[Title/Abstract]) AND (artificial [Title/Abstract] OR replaced [Title/Abstract] OR replacement [Title/Abstract] OR prosthesis[tiab] OR prostheses[tiab] OR implant [Title/Abstract] OR implants [Title/Abstract] OR implantation [Title/Abstract] OR implantations [Title/Abstract])) OR "total disc replacement"[MeSH Terms] OR "arthroplasty"[MeSH Terms] OR arthroplasty [Title/Abstract] OR "prosthesis implantation"[MeSH Terms] OR "diskectomy"[MeSH Terms] OR discectom* [Title/Abstract] OR diskectom* [Title/Abstract] OR "internal fixators"[MeSH Terms] OR Internal fixators [Title/Abstract] OR Internal fixator [Title/Abstract] OR internal fixation[Title/Abstract] OR Laminectom*[Title/Abstract] OR "laminectomy"[MeSH Terms] OR laminotom*[Title/Abstract] OR "Decompression, Surgical "[Mesh])) AND ((randomized controlled trial[pt] OR controlled clinical trial[pt] OR randomized[tiab] OR placebo[tiab] OR "clinical trials as topic"[MeSH | 2902 |

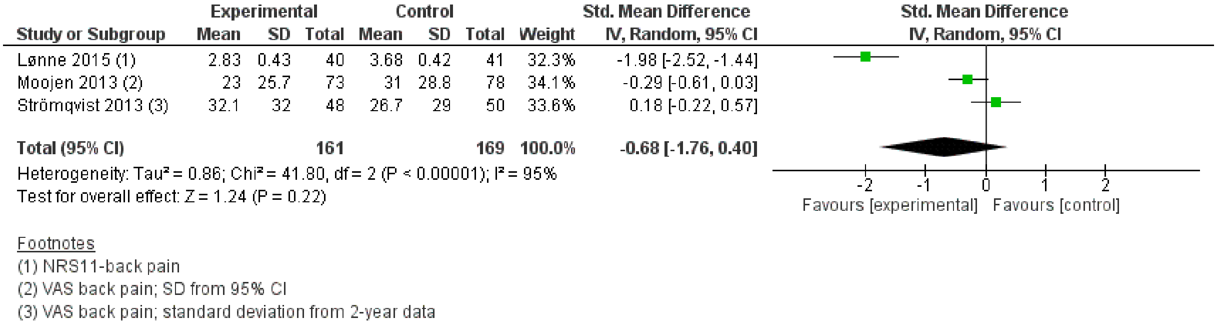
| | |
|---|---|
| Terms] OR randomly[tiab] OR trial[ti] OR “randomised” [tiab] OR “random” [tiab]) NOT ("animals"[MeSH Terms] NOT "humans"[MeSH Terms])) | |

Search strategy for currently recruiting randomised controlled trials registered on ClinicalTrials.gov

| **Search terms** | **Hits** |
|---|---|
| dynamic stabilisation OR dynamic stabilization | 8 |
| interspinous | 2 |
| pedicle AND spine | 23 |
| disc replacement OR disk replacement | 8 |
| Disc implant OR disk implant | 10 |
| Disc prosthesis OR disc prostheses OR disk prosthesis OR disk prostheses | 3 |
| (Dynamic AND stabilization) OR (Dynamic AND stabilisation) OR (Interspinous AND spacer) OR (Interspinous AND spacers) OR (Interspinous AND device) OR (Interspinous AND devices) OR (Interspinous AND decompression) OR (Interspinous AND process) | 8 |
| **Total** | **62** |

# II. Appendix – PICO 1 short-term

## PICO 1 short-term: Back pain (1 year)

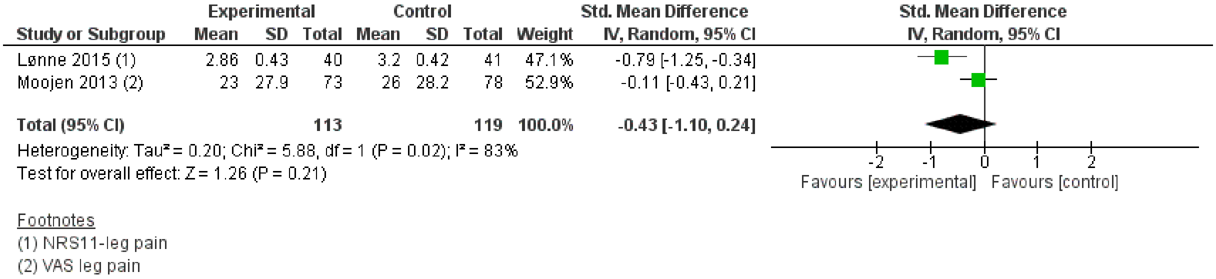


| | Experimental | | | Control | | | | Std. Mean Difference |
|---|---|---|---|---|---|---|---|---|
| **Study or Subgroup** | **Mean** | **SD** | **Total** | **Mean** | **SD** | **Total** | **Weight** | **IV, Random, 95% CI** |
| Lønne 2015 (1) | 2.83 | 0.43 | 40 | 3.68 | 0.42 | 41 | 32.3% | -1.98 [-2.52, -1.44] |
| Moojen 2013 (2) | 23 | 25.7 | 73 | 31 | 28.8 | 78 | 34.1% | -0.29 [-0.61, 0.03] |
| Strömqvist 2013 (3) | 32.1 | 32 | 48 | 26.7 | 29 | 50 | 33.6% | 0.18 [-0.22, 0.57] |
| **Total (95% CI)** | | | **161** | | | **169** | **100.0%** | **-0.68 [-1.76, 0.40]** |

Heterogeneity: Tau² = 0.86; Chi² = 41.80, df = 2 (P < 0.00001); I² = 95%
Test for overall effect: Z = 1.24 (P = 0.22)

Footnotes
(1) NRS11-back pain
(2) VAS back pain; SD from 95% CI
(3) VAS back pain; standard deviation from 2-year data

## PICO 1 short-term: Radicular pain (1 year)



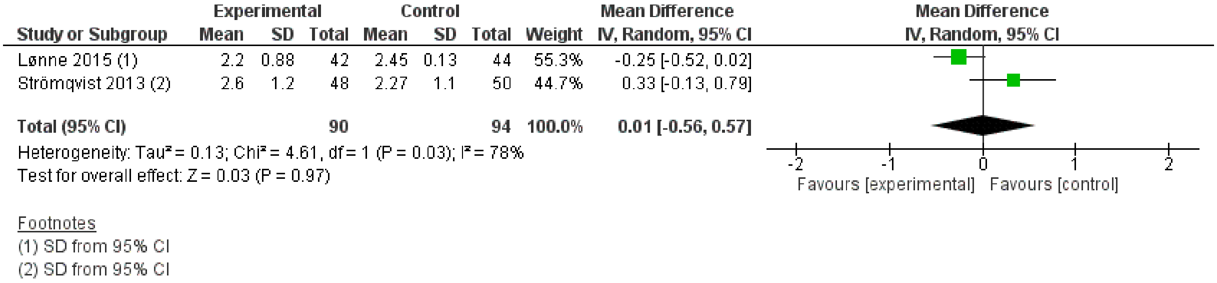
| | Experimental | | | Control | | | | Std. Mean Difference |
|---|---|---|---|---|---|---|---|---|
| **Study or Subgroup** | **Mean** | **SD** | **Total** | **Mean** | **SD** | **Total** | **Weight** | **IV, Random, 95% CI** |
| Lønne 2015 (1) | 2.86 | 0.43 | 40 | 3.2 | 0.42 | 41 | 47.1% | -0.79 [-1.25, -0.34] |
| Moojen 2013 (2) | 23 | 27.9 | 73 | 26 | 28.2 | 78 | 52.9% | -0.11 [-0.43, 0.21] |
| **Total (95% CI)** | | | **113** | | | **119** | **100.0%** | **-0.43 [-1.10, 0.24]** |

Heterogeneity: Tau² = 0.20; Chi² = 5.88, df = 1 (P = 0.02); I² = 83%
Test for overall effect: Z = 1.26 (P = 0.21)

Footnotes
(1) NRS11-leg pain
(2) VAS leg pain

## PICO 1 short-term: Spinal claudication (Walking distance or Zurich Claudication Questionnaire) (1 year)

### Zurich Claudication Questionnaire: Symptom severity

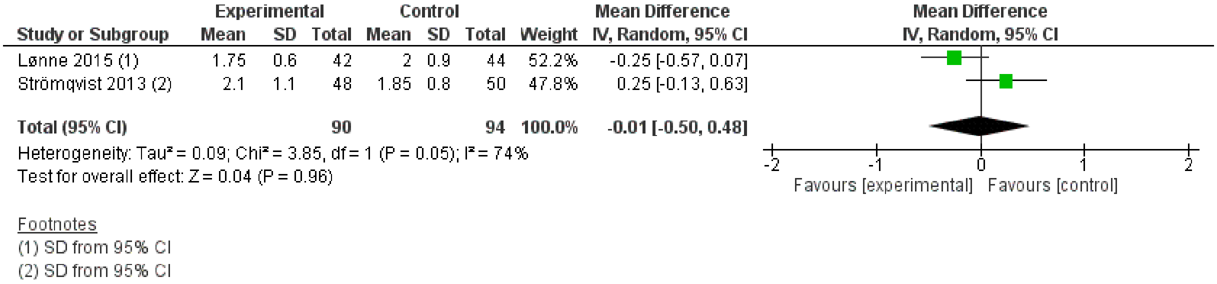


| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| **Study or Subgroup** | **Mean** | **SD** | **Total** | **Mean** | **SD** | **Total** | **Weight** | **IV, Random, 95% CI** |
| Lønne 2015 (1) | 2.2 | 0.88 | 42 | 2.45 | 0.13 | 44 | 55.3% | -0.25 [-0.52, 0.02] |
| Strömqvist 2013 (2) | 2.6 | 1.2 | 48 | 2.27 | 1.1 | 50 | 44.7% | 0.33 [-0.13, 0.79] |
| **Total (95% CI)** | | | **90** | | | **94** | **100.0%** | **0.01 [-0.56, 0.57]** |

Heterogeneity: Tau² = 0.13; Chi² = 4.61, df = 1 (P = 0.03); I² = 78%
Test for overall effect: Z = 0.03 (P = 0.97)

Footnotes
(1) SD from 95% CI
(2) SD from 95% CI

### Zurich Claudication Questionnaire: Satisfaction

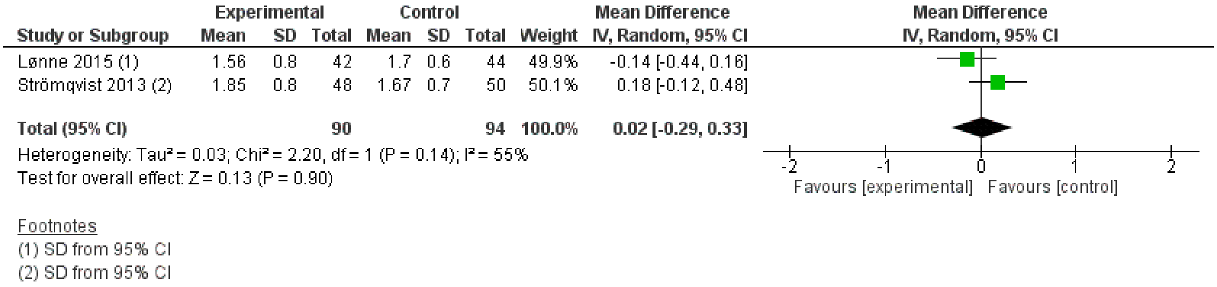


| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| **Study or Subgroup** | **Mean** | **SD** | **Total** | **Mean** | **SD** | **Total** | **Weight** | **IV, Random, 95% CI** |
| Lønne 2015 (1) | 1.75 | 0.6 | 42 | 2 | 0.9 | 44 | 52.2% | -0.25 [-0.57, 0.07] |
| Strömqvist 2013 (2) | 2.1 | 1.1 | 48 | 1.85 | 0.8 | 50 | 47.8% | 0.25 [-0.13, 0.63] |
| **Total (95% CI)** | | | **90** | | | **94** | **100.0%** | **-0.01 [-0.50, 0.48]** |

Heterogeneity: Tau² = 0.09; Chi² = 3.85, df = 1 (P = 0.05); I² = 74%
Test for overall effect: Z = 0.04 (P = 0.96)

Footnotes
(1) SD from 95% CI
(2) SD from 95% CI

### Zurich Claudication Questionnaire: Physical function

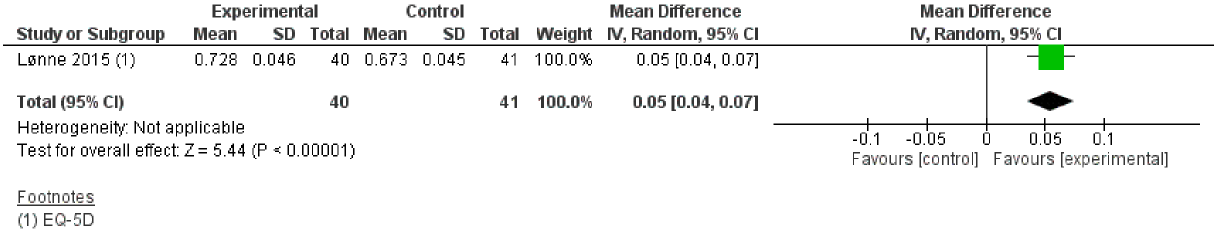


| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Lønne 2015 (1) | 1.56 | 0.8 | 42 | 1.7 | 0.6 | 44 | 49.9% | -0.14 [-0.44, 0.16] |
| Strömqvist 2013 (2) | 1.85 | 0.8 | 48 | 1.67 | 0.7 | 50 | 50.1% | 0.18 [-0.12, 0.48] |
| **Total (95% CI)** | | | **90** | | | **94** | **100.0%** | **0.02 [-0.29, 0.33]** |

Heterogeneity: Tau² = 0.03; Chi² = 2.20, df = 1 (P = 0.14); I² = 55%
Test for overall effect: Z = 0.13 (P = 0.90)

Footnotes
(1) SD from 95% CI
(2) SD from 95% CI

## PICO 1 short-term: Quality of life (1 year)

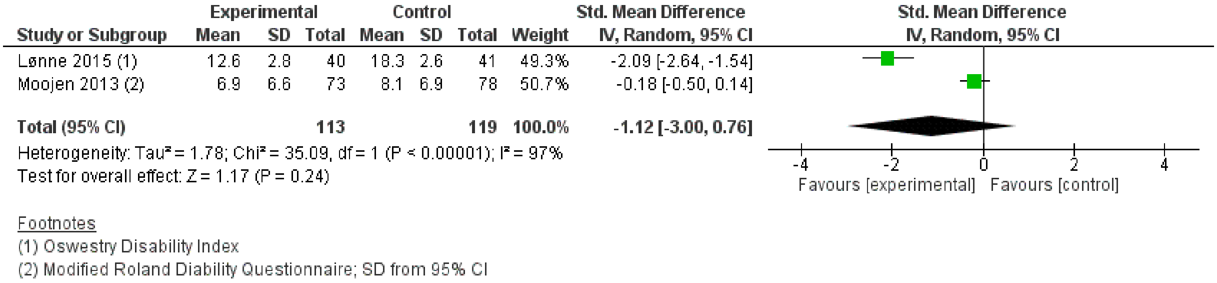


| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Lønne 2015 (1) | 0.728 | 0.046 | 40 | 0.673 | 0.045 | 41 | 100.0% | 0.05 [0.04, 0.07] |
| **Total (95% CI)** | | | **40** | | | **41** | **100.0%** | **0.05 [0.04, 0.07]** |

Heterogeneity: Not applicable
Test for overall effect: Z = 5.44 (P < 0.00001)

Footnotes
(1) EQ-5D

## PICO 1 short-term: Function (1 year)



| | Experimental | | | Control | | | | Std. Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Lønne 2015 (1) | 12.6 | 2.8 | 40 | 18.3 | 2.6 | 41 | 49.3% | -2.09 [-2.64, -1.54] |
| Moojen 2013 (2) | 6.9 | 6.6 | 73 | 8.1 | 6.9 | 78 | 50.7% | -0.18 [-0.50, 0.14] |
| **Total (95% CI)** | | | **113** | | | **119** | **100.0%** | **-1.12 [-3.00, 0.76]** |

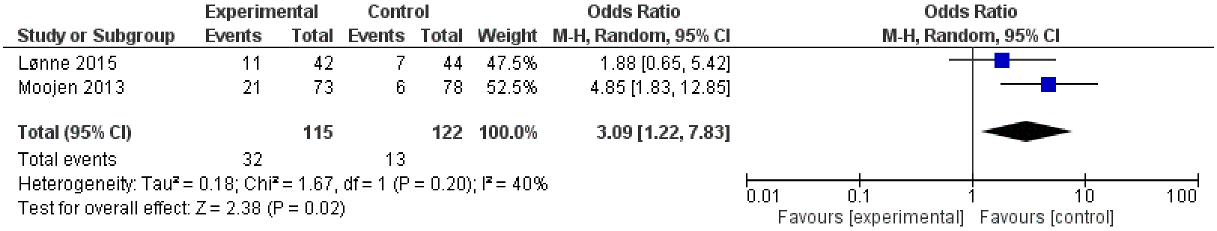
Heterogeneity: Tau² = 1.78; Chi² = 35.09, df = 1 (P < 0.00001); I² = 97%
Test for overall effect: Z = 1.17 (P = 0.24)

Footnotes
(1) Oswestry Disability Index
(2) Modified Roland Diability Questionnaire; SD from 95% CI

## PICO 1 short-term: Revision rate (1 year)

No data available.

## PICO 1 short-term: Reoperation rate (1 year)



| | Experimental | | Control | | | Odds Ratio |
|---|---|---|---|---|---|---|
| Study or Subgroup | Events | Total | Events | Total | Weight | M-H, Random, 95% CI |
| Lønne 2015 | 11 | 42 | 7 | 44 | 47.5% | 1.88 [0.65, 5.42] |
| Moojen 2013 | 21 | 73 | 6 | 78 | 52.5% | 4.85 [1.83, 12.85] |
| **Total (95% CI)** | | **115** | | **122** | **100.0%** | **3.09 [1.22, 7.83]** |
| Total events | 32 | | 13 | | | |

Heterogeneity: Tau² = 0.18; Chi² = 1.67, df = 1 (P = 0.20); I² = 40%
Test for overall effect: Z = 2.38 (P = 0.02)

## PICO 1 short-term: Complications or adverse events (1 year)

No data available.

## PICO 1 short-term: Serious adverse events (1 year)

No data available.

## III. Appendix – Risk of Bias with support for judgement

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PICO 1 | | | | | | | | | | | | | | |
| Lønne 2015 | Low | "Patients were randomized with randomly selected block sizes by a computer-based web solution hosted by the medical faculty at the Norwegian University of Science and Technology." | Low | "Patients were randomized with randomly selected block sizes by a computer-based web solution hosted by the medical faculty at the Norwegian University of Science and Technology. | Unclear | Not reported | Unclear | Unclear if blinded: "The data were collected and entered by independent observers, and permission to store the data was granted by the Norwegia | High | missing data 10% - 20 %, addressed using inadequate methods or not addressed | High | missing data 10% - 20 %, addressed using inadequate methods or not addressed | Unclear | Outcomes complications and revision rate were not prespecified in the methods section |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Details from all hospitals were available to a coordinating secretary not involved in the treatment." | | | | n data inspector ate." | | | | | | |
| Moojen 2013 | Low | "The randomization was prepared by the study statistician and the principle data manager at the department of Biostatistics. " | Low | "Allocations were stored in prepared opaque, coded, and sealed envelopes." | Low | " Patients, nurses on the hospital wards, and research nurses remained blind to the allocated treatment during the follow-up period of one year. " | Low | "observer and patient blinded" | High | missing data 10% - 20 %, addressed using adequate methods, but not comparable among | High | missing data 10% - 20 %, addressed using inadequate methods or not addressed | Unclear | Outcomes revision rate and complictaions were not prespecified in the methods section. Primary outcome ZQC only reported as |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | and "The average subscale scores were obtained at every follow-up moment by blinded research nurses" | | | | study arms | | | | "success" and no results per scale reported. |
| Strömqvist 2013 | Unclear | Unclear how the envelopes were used for randomisation, i.e. if they were shuffled; "Randomization was performed by using envelopes." | Unclear | Not reported | Unclear | Not reported | Unclear | Not reported | Low | missing data ≤10% and comparable among study arms | Low | missing data ≤10% and comparable among study arms | high | Walking distance, Euroqol and ODI measured, but results not reported |
| PICO 2 | | | | | | | | | | | | | | |
| Marsh 2014 | Low | "random number | Unclear | Not reported | Unclear | Not reported | Unclear | Not reported | Low | missing data | Low | missing data | low | results were |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | generator" | | | | | | | | ≤10% and comparable among study arms | | ≤10% and comparable among study arms | | given for all outcomes mentioned in method section |
| PICO 3 | | | | | | | | | | | | | | |
| Davis 2013 | Low | See Davis 2013 | Low | See Davis 2013 | High | "The study was not blinded during follow-up." | Unclear | Not reported | High | missing data 10% - 20 %, addressed using inadequate methods or not addressed | High | missing data 10% - 20 %, addressed using inadequate methods or not addressed | Unclear | Outcomes revision rate and complications were not prespecified in the methods section |
| Madan 2003 | Low | "The patients were assigned numbers after a decision was | Low | "The patients were assigned | Unclear | Not reported | Unclear | Not reported | Unclear | not reported | Unclear | not reported | Unclear | Outcome complication not prespecifi |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | made to operate on them. A chit was drawn blindly from a box, with Graf ligament operation designated by “1” and Hartshill horseshoe fusion designated by “2”. The draw was done a day before the operation, after which the patient was consented for the appropriate surgery. By picking up the chit from the box after shaking it well, | | numbers after a decision was made to operate on them. A chit was drawn blindly from a box, with Graf ligament operation designated by “1” and Hartshill horseshoe fusion designated by “2”. The draw was done a day before the operation, after which the patient | | | | | | | | | | ed. Most of the comparisons were baseline vs end-of-follow and not between studies, although, this did not influence our results and hence was not judge as "high". |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | we thought that the process was sufficiently random for there to be a 0.5 probability that the patient would have one of the two procedures." | | was consented for the appropriate surgery. By picking up the chit from the box after shaking it well, we thought that the process was sufficiently random for there to be a 0.5 probability that the patient would have one of the two procedures." | | | | | | | | | | |
| Putzier | Low | "Randomization | Uncl | "Randomizat | High | "… were | High | not | High | missing | High | missing | low | Outcome |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2010 | | was performed by the Randlist Software (DataInf GmbH, Tuebingen, Germany)." | ear | ion was performed by the Randlist Software (DataInf GmbH, Tuebingen, Germany)." | | enrolled in this prospective, randomized, non-blind study." | | blinded trial | | data 10% - 20 %, addressed using inadequate methods or not addressed | | data 10% - 20 %, addressed using inadequate methods or not addressed | | s prespecified in the method section were reported (complications was not extractable |
| PICO 4 | | | | | | | | | | | | | | |
| Blumenthal 2005 | Low | "A contract research organization generated the random allocation sequence using SAS software in a ratio of 2:1 (investigational: control). A fixed blocking method | low | "Each site was provided with sequentially numbered sealed envelopes that contained the treatment | High | "The investigator, key office staff, and operating room staff were nonblinded to group assignment. Patients | High | stated as non blinded Trial | high | missing data 10% - 20 %, and not comparable among study arms | **low** | missing data ≤10% and comparable among study arms | unclear | SF-36 mean scores not reported but pre-specified in the methods; |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | of randomization was used with 6 assignments per block" | | assignments for their site. The site opened the next sequential envelope the day before surgery and only when a subject satisfied inclusion and exclusion criteria, and signed the informed consent form." | | were not blinded throughout their 2 year course within the study because blinding all patient records including radiographs, other postoperative images such as computed tomography scans, and third-party records would have been | | | | | | | | |

| Study | **Random sequence generation (selection bias) and support for judgement** | | **Allocation concealment (selection bias) and support for judgement** | | **Blinding of participants and personnel (performance bias) and support for judgement** | | **Blinding of outcome assessment (detection bias) and support for judgement** | | **Incomplete continuous outcome data (attrition bias) and support for judgement** | | **Incomplete binary data (attrition bias) and support for judgement** | | **Selective reporting (reporting bias) and support for judgement** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | exceedingly difficult. Patients experiencing postoperative bone graft donor site pain would be unblinded. " | | | | | | | | |
| Gornet 2011 | unclear | "randomized according to a schedule centrally generated by the study sponsor with a fixed block size of six. The block size was unknown to study investigators | Low | "randomized according to a schedule centrally generated by the study sponsor with a fixed block size of six. The block size was unknown to study | High | "Both the investigator and the patient were blinded to the randomization before informed consent but were not blinded | High | "Both the investigator and the patient were blinded to the randomization before informed consent | high | missing data 10% - 20 %, adressed using inadequate methods or not addressed | **low** | missing data ≤10% and comparable among study arms | low | pre-specified outcomes in the methods section were reported; not sufficient information on the back and |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | and patients during the study. Treatment randomization was 2:1 (investigational to control) on a site basis with sequentially numbered, sealed envelopes provided by the study sponsor. | | investigators and patients during the study. Treatment randomization was 2:1 (investigational to control) on a site basis with sequentially numbered, sealed envelopes provided by the study sponsor. | | after the opening of the treatment envelope" | | but were not blinded after the opening of the treatment envelope" | | | | | | leg pain scale (range not reported, but probably 0-100) |
| Moreno 2008 | Unclear | Not reported | Unclear | Not reported | Unclear | Not reported | Unclear | Not reported | unclear | not reported | unclear | not reported | unclear | not prespecified outcomes in methods |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | section |
| Sasso 2008 | Unclear | Not reported | Unclear | Not reported | Unclear | Not reported | Unclear | Not reported | high | missing data >20% in either study arm | unclear | not reported | low | all pre-specified outcomes were reported |
| Zigler 2007 | Unclear | Separate randomization schedules were generated for each of the 17 sites using a fixed block size of 6, with the randomization performed external to the site after individual patient enrollment | low | Separate randomization schedules were generated for each of the 17 sites using a fixed block size of 6, with the randomization performed external to the site after individual patient enrolment | High | "Patients were blinded to randomization until immediately postsurgery" | Unclear | Not reported | high | missing data >20% in either study arm | High | missing data >20% in either study arm | low | all pre-specified outcomes were reported |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PICO 5 | | | | | | | | | | | | | | |
| Coric 2011 | Unclear | Not reported | Unclear | Not reported | Unclear | Not reported | Unclear | Not reported | High | missing data 10% - 20 %, addressed using inadequate methods or not addressed | High | missing data 10% - 20 %, addressed using inadequate methods or not addressed | Unclear | Outcomes Complications and revision rate were not prespecified in the method section |
| Heller 2009 | Unclear | "The randomization schedule was centrally generated by the study's sponsor, stratified by site and by using a fixed block size of 4." | Low | "The randomization schedule was centrally generated by the study's sponsor, stratified by site and by using a fixed | High | "Blinding for investigators and patients was maintained through confirmation of eligibility | Unclear | not reported | High | missing data >20% in either study arm | High | missing data >20% in either study arm | Low | Outcomes pre-specified |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | block size of 4." "Blinding for investigators and patients was maintained through confirmation of eligibility and informed consent." | | and informed consent." BUT: "Because of this difference between the treatment groups and issues related to patient care, further blinding was not practical or ethical." | | | | | | | | |
| Hisey 2014 | Low | "patients were randomized to groups by an Interactive Voice Randomization | Low | "patients were randomized to groups by an | High | "Due to the fact that the implant was evident to | Unclear | not reported | High | missing data >20% in either study | High | missing data >20% in either study | Unclear | Outcomes Complications were not |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | System (IVRS). The investigator or study coordinator called the IVRS after the pre-operative inclusion/exclusion checklist confirmed eligibility " "Patients were assigned to the TDR or control group by IVRS according to a stratified randomization schedule (by baseline Neck Disability Index (NDI) score) with institutional balancing" | | Interactive Voice Randomization System (IVRS). The investigator or study coordinator called the IVRS after the pre-operative inclusion/exclusion checklist confirmed eligibility " "Patients were assigned to the TDR or control group by IVRS according to | | the surgeon, blinding the physician to treatment was not possible. Patients remained blinded to the treatment group assignment until after surgery had been performed to minimize the potential for | | | | arm | | arm | | prespecified in the methods section |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | a stratified randomization schedule (by baseline Neck Disability Index (NDI) score) with institutional balancing" | | disproportionate patient dropouts." | | | | | | | | |
| Karabag 2014 | Unclear | Not reported | Unclear | Not reported | Unclear | Not reported | Unclear | Not reported | Unclear | not reported | Not applicable | No binary outcome extracted | low | Prespecified outcomes were reported |
| Mummaneni 2007 - Prestige ST | Low | "Patients were randomly assigned according to a randomization schedule using the Plan Procedure in Statistical | Unclear | Not reported | High | "It was not practical, however, to blind the patients and the surgeons as to the type of | Unclear | not reported | High | missing data >20% in either study arm | High | missing data >20% in either study arm | Low | All outcomes mentioned in the methods. |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Analysis System (version 6.12 or higher, SAS)" | | | | surgery that was performed " | | | | | | | | |
| Murrey 2009 | Low | "fixed block randomization sequence of four subjects per block generated by the contract research organization and executed at each site with use of sequenced opaque sealed envelopes." | Low | "fixed block randomization sequence of four subjects per block generated by the contract research organization and executed at each site with use of sequenced opaque sealed envelopes." | High | "The surgeon and surgical staff were not blinded to group assignment because of surgery preparation requirements. The subject remained blinded until immediately following surgery." | Unclear | not reported | High | missing data >20% in either study arm | High | missing data >20% in either study arm | Low | All outcomes prespecified in the methods section |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Nabhan 2007 | Low | "Randomization was carried out by drawing cards in sealed envelopes. " | Unclear | Randomization was carried out by drawing cards in sealed envelopes. | Unclear | Not reported | Unclear | Not reported | High | missing data 10% - 20 %, adressed using inadequate methods or not addressed | Not applicable | No binary outcome extracted | Low | All outcomes prespecified in the methods section |
| Phillips 2013 | Unclear | Not reported | unclear | Concealemnt not reported, but "The investigator and surgical staff were not masked to the treatment assignment; however, the patient | High | "The investigator and surgical staff were not masked to the treatment assignment; however, the patient remained | Unclear | Not reported | High | missing data >20% in either study arm | High | missing data >20% in either study arm | Low | All outcomes prespecified in the methods section (**HE: we could do unclear, some results reported |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | remained masked to the random treatment assignment until after surgery." | | masked to the random treatment assignment until after surgery." | | | | | | | | for 7-y and some for 5 but not for 7) |
| Rozankovic 2014 | Low | " Randomizer (www.randomizer.org) was used for patient randomization" | Unclear | Not reported | Unclear | Not reported | Unclear | Not reported | Low | missing data ≤10% and comparable among study arms | Low | missing data ≤10% and comparable among study arms | low | All prespecified outcomes in the methods were reported |
| Vaccaro 2013 | Unclear | Not reported | Unclear | Not reported | high | "Patients were blinded to randomization (1:1) before surgery." | Unclear | Not reported | Unclear | not reported | Unclear | not reported | Unclear | Adverse events were no prespecified in Methods |
| Zhang 2012 | Low | "A list of sequential | Unclear | "A list of sequential | Unclear | not reported | Unclear | not reported | Low | missing data | Low | missing data | Unclear | Outcome revision |

| Study | Random sequence generation (selection bias) and support for judgement | | Allocation concealment (selection bias) and support for judgement | | Blinding of participants and personnel (performance bias) and support for judgement | | Blinding of outcome assessment (detection bias) and support for judgement | | Incomplete continuous outcome data (attrition bias) and support for judgement | | Incomplete binary data (attrition bias) and support for judgement | | Selective reporting (reporting bias) and support for judgement | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | numbers was generated using a simple randomization in SPSS 15.0 (SPSS Inc, Chicago, IL), with each number randomly assigned to 1 group." | | numbers was generated using a simple randomization in SPSS 15.0 (SPSS Inc, Chicago, IL), with each number randomly assigned to 1 group." | | | | | | ≤10% and comparable among study arms | | ≤10% and comparable among study arms | | rate was not prespecified in the methods section |
| Zhang 2014 | Unclear | Not reported | Unclear | Not reported | Unclear | Not reported | Unclear | Not reported | Unclear | not reported | Low | missing data ≤10% and comparable among study arms | Unclear | Outcomes were not prespecified in the methods section |

## IV. Appendix – PICO 2 short-term

### PICO 2 short-term: Back pain (1 year)

No data available.

### PICO 2 short-term: Radicular pain (1 year)

No data available.

### PICO 2 short-term: Spinal claudication (Walking distance or Zurich Claudication Questionnaire) (1 year)

No data available.

### PICO 2 short-term: Quality of life (1 year)

No data available.

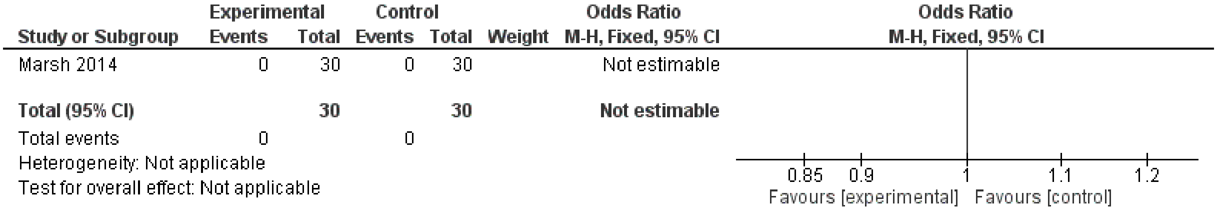
### PICO 2 short-term: Function (1 year)

No data available.

### PICO 2 short-term: Revision rate (1 year)

No data available.

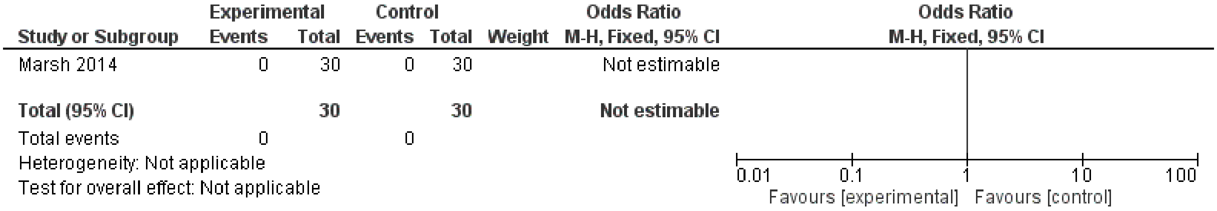
### PICO 2 short-term: Reoperation rate (1 year)



| | Experimental | | Control | | | Odds Ratio |
|---|---|---|---|---|---|---|
| Study or Subgroup | Events | Total | Events | Total | Weight | M-H, Fixed, 95% CI |
| Marsh 2014 | 0 | 30 | 0 | 30 | | Not estimable |
| **Total (95% CI)** | | **30** | | **30** | | **Not estimable** |
| Total events | 0 | | 0 | | | |

Heterogeneity: Not applicable
Test for overall effect: Not applicable

### PICO 2 short-term: Complications or adverse events (1 year)



| | Experimental | | Control | | | Odds Ratio |
|---|---|---|---|---|---|---|
| Study or Subgroup | Events | Total | Events | Total | Weight | M-H, Fixed, 95% CI |
| Marsh 2014 | 0 | 30 | 0 | 30 | | Not estimable |
| **Total (95% CI)** | | **30** | | **30** | | **Not estimable** |
| Total events | 0 | | 0 | | | |

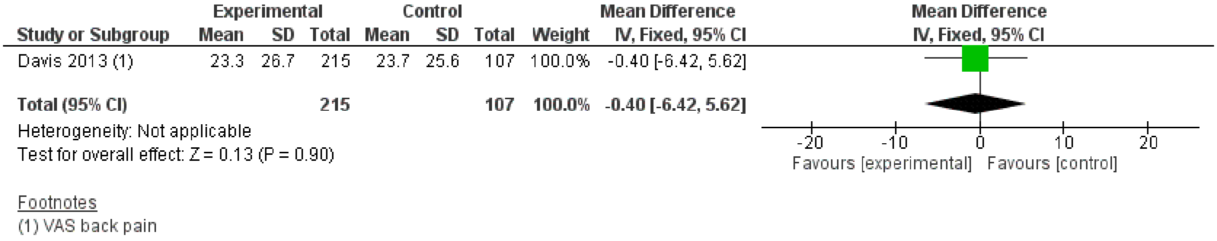
Heterogeneity: Not applicable
Test for overall effect: Not applicable

### PICO 2 short-term: Serious adverse events (1 year)

No data available.

## V. Appendix – PICO 3 short-term

### PICO 3 short-term: back pain (1 year)

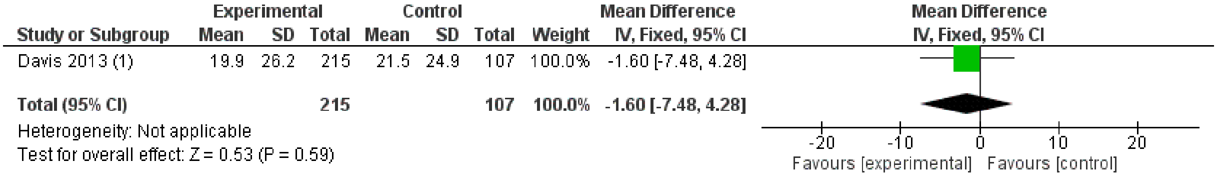


| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Fixed, 95% CI |
| Davis 2013 (1) | 23.3 | 26.7 | 215 | 23.7 | 25.6 | 107 | 100.0% | -0.40 [-6.42, 5.62] |
| **Total (95% CI)** | | | **215** | | | **107** | **100.0%** | **-0.40 [-6.42, 5.62]** |

Heterogeneity: Not applicable
Test for overall effect: Z = 0.13 (P = 0.90)

Footnotes
(1) VAS back pain

### PICO 3 short-term: Radicular pain (1 year)



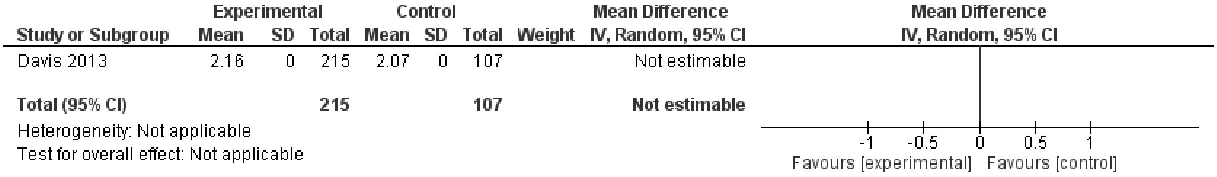
| | Experimental | | | Control | | | | Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Fixed, 95% CI | IV, Fixed, 95% CI |
| Davis 2013 (1) | 19.9 | 26.2 | 215 | 21.5 | 24.9 | 107 | 100.0% | -1.60 [-7.48, 4.28] | |
| **Total (95% CI)** | | | **215** | | | **107** | **100.0%** | **-1.60 [-7.48, 4.28]** | |

Heterogeneity: Not applicable
Test for overall effect: Z = 0.53 (P = 0.59)

Footnotes
(1) VAS leg pain

## PICO 3 short-term: Spinal claudication (Walking distance or Zurich Claudication Questionnaire) (1 year)

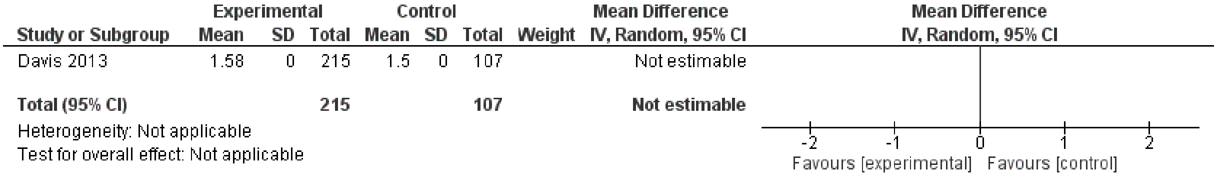
### Zurich Claudication Questionnaire: Symptom severity



| | Experimental | | | Control | | | | Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI | IV, Random, 95% CI |
| Davis 2013 | 2.16 | 0 | 215 | 2.07 | 0 | 107 | | Not estimable | |
| **Total (95% CI)** | | | **215** | | | **107** | | **Not estimable** | |

Heterogeneity: Not applicable
Test for overall effect: Not applicable

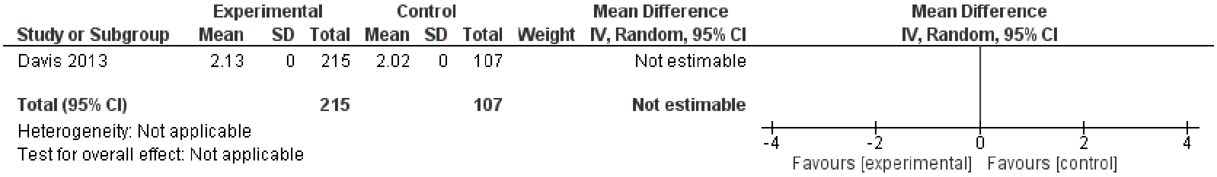
### Zurich Claudication Questionnaire: Satisfaction



| | Experimental | | | Control | | | | Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI | IV, Random, 95% CI |
| Davis 2013 | 1.58 | 0 | 215 | 1.5 | 0 | 107 | | Not estimable | |
| **Total (95% CI)** | | | **215** | | | **107** | | **Not estimable** | |

Heterogeneity: Not applicable
Test for overall effect: Not applicable

### Zurich Claudication Questionnaire: Physical function



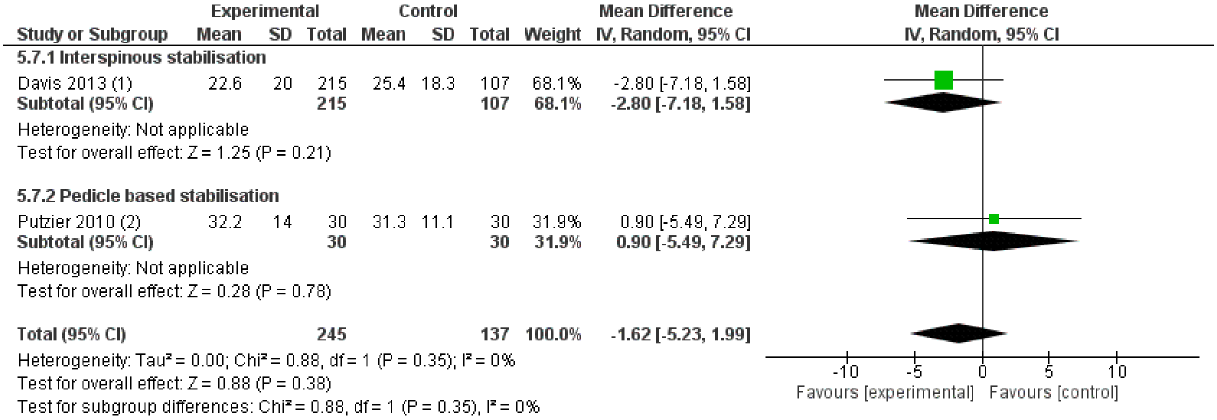
| | Experimental | | | Control | | | | Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI | IV, Random, 95% CI |
| Davis 2013 | 2.13 | 0 | 215 | 2.02 | 0 | 107 | | Not estimable | |
| **Total (95% CI)** | | | **215** | | | **107** | | **Not estimable** | |

Heterogeneity: Not applicable
Test for overall effect: Not applicable

## PICO 3 short-term: Quality of life (1 year)

No data available.

## PICO 3 short-term: Function (1 year)



| | Experimental | | | Control | | | | Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI | IV, Random, 95% CI |
| **5.7.1 Interspinous stabilisation** | | | | | | | | | |
| Davis 2013 (1) | 22.6 | 20 | 215 | 25.4 | 18.3 | 107 | 68.1% | -2.80 [-7.18, 1.58] | |
| **Subtotal (95% CI)** | | | **215** | | | **107** | **68.1%** | **-2.80 [-7.18, 1.58]** | |
| Heterogeneity: Not applicable | | | | | | | | | |
| Test for overall effect: Z = 1.25 (P = 0.21) | | | | | | | | | |
| **5.7.2 Pedicle based stabilisation** | | | | | | | | | |
| Putzier 2010 (2) | 32.2 | 14 | 30 | 31.3 | 11.1 | 30 | 31.9% | 0.90 [-5.49, 7.29] | |
| **Subtotal (95% CI)** | | | **30** | | | **30** | **31.9%** | **0.90 [-5.49, 7.29]** | |
| Heterogeneity: Not applicable | | | | | | | | | |
| Test for overall effect: Z = 0.28 (P = 0.78) | | | | | | | | | |
| **Total (95% CI)** | | | **245** | | | **137** | **100.0%** | **-1.62 [-5.23, 1.99]** | |

Heterogeneity: Tau² = 0.00; Chi² = 0.88, df = 1 (P = 0.35); I² = 0%
Test for overall effect: Z = 0.88 (P = 0.38)
Test for subgroup differences: Chi² = 0.88, df = 1 (P = 0.35), I² = 0%

Footnotes
(1) Oswestry Disability Index
(2) Oswestry Disability Index; Dynamic stabilisation + adjacent fusion vs fusion

## PICO 3 short-term: Revision rate (1 year)

No data available.

## PICO 3 short-term: Reoperation rate (1 year)

No data available.

### PICO 3 short-term: Complications or adverse events (1 year)

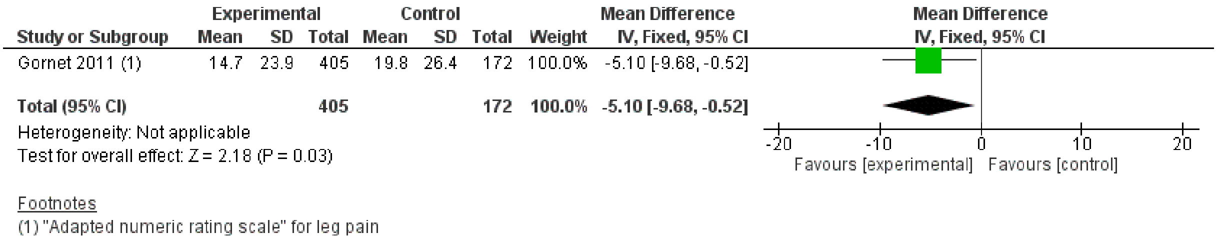
No data available.

### PICO 3 short-term: Serious adverse events (1 year)

No data available.

## VI. Appendix – PICO 4 short-term

### PICO 4 short-term: Radicular pain (1 year)



| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Fixed, 95% CI |
| Gornet 2011 (1) | 14.7 | 23.9 | 405 | 19.8 | 26.4 | 172 | 100.0% | -5.10 [-9.68, -0.52] |
| **Total (95% CI)** | | | **405** | | | **172** | **100.0%** | **-5.10 [-9.68, -0.52]** |

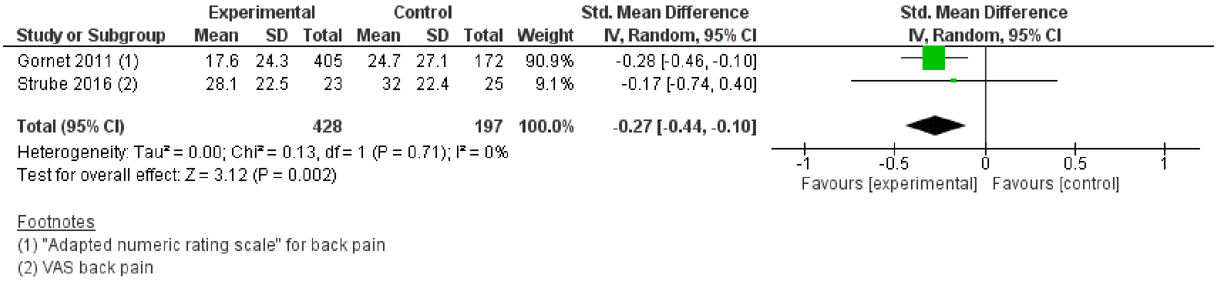
Heterogeneity: Not applicable

Test for overall effect: Z = 2.18 (P = 0.03)

Footnotes

(1) "Adapted numeric rating scale" for leg pain

### PICO 4 short-term: Back pain (1 year)



| | Experimental | | | Control | | | | Std. Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Gornet 2011 (1) | 17.6 | 24.3 | 405 | 24.7 | 27.1 | 172 | 90.9% | -0.28 [-0.46, -0.10] |
| Strube 2016 (2) | 28.1 | 22.5 | 23 | 32 | 22.4 | 25 | 9.1% | -0.17 [-0.74, 0.40] |
| **Total (95% CI)** | | | **428** | | | **197** | **100.0%** | **-0.27 [-0.44, -0.10]** |

Heterogeneity: Tau² = 0.00; Chi² = 0.13, df = 1 (P = 0.71); I² = 0%

Test for overall effect: Z = 3.12 (P = 0.002)

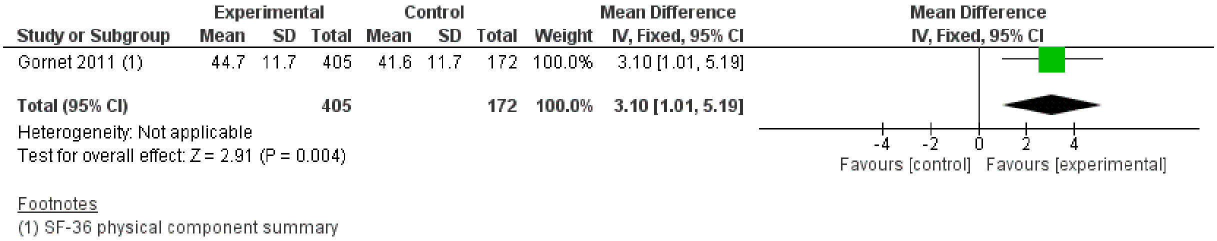
Footnotes

(1) "Adapted numeric rating scale" for back pain

(2) VAS back pain

### PICO 4 short-term: Quality of life (1 year)

## Short form 36: physical component score



| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Fixed, 95% CI |
| Gornet 2011 (1) | 44.7 | 11.7 | 405 | 41.6 | 11.7 | 172 | 100.0% | 3.10 [1.01, 5.19] |
| **Total (95% CI)** | | | **405** | | | **172** | **100.0%** | **3.10 [1.01, 5.19]** |

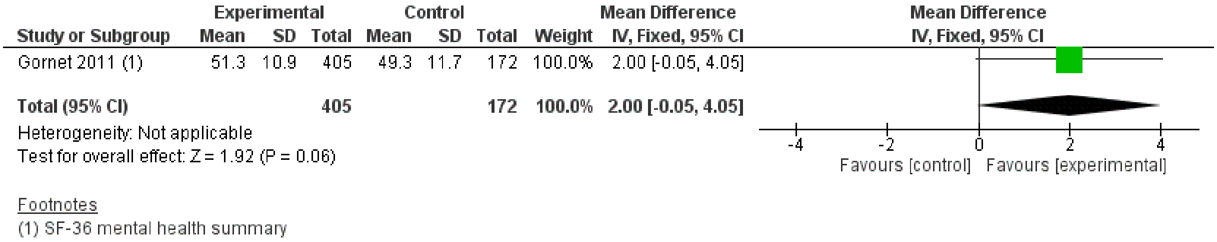
Heterogeneity: Not applicable

Test for overall effect: Z = 2.91 (P = 0.004)

Footnotes

(1) SF-36 physical component summary

## Short form 36: mental component score



| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Fixed, 95% CI |
| Gornet 2011 (1) | 51.3 | 10.9 | 405 | 49.3 | 11.7 | 172 | 100.0% | 2.00 [-0.05, 4.05] |
| **Total (95% CI)** | | | **405** | | | **172** | **100.0%** | **2.00 [-0.05, 4.05]** |

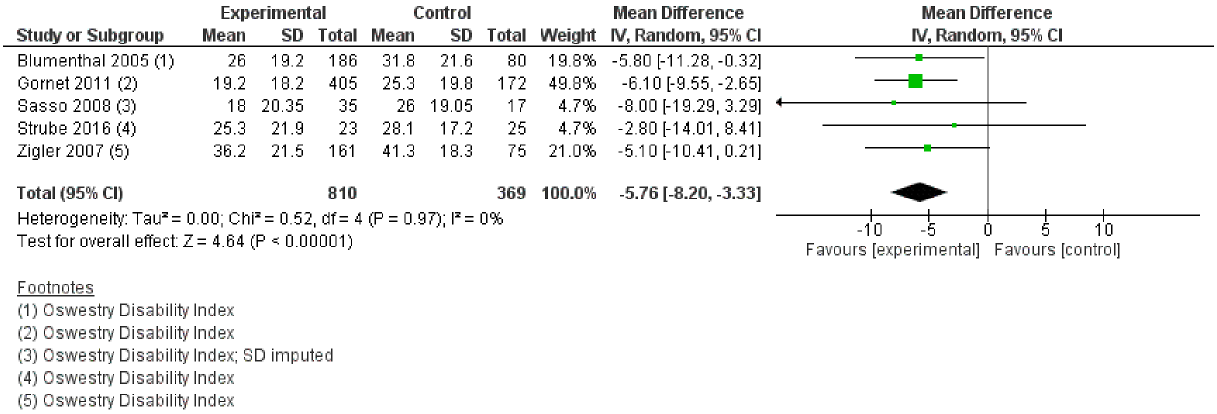
Heterogeneity: Not applicable

Test for overall effect: Z = 1.92 (P = 0.06)

Footnotes

(1) SF-36 mental health summary

### PICO 4 short-term: Function (1 year)



|  | Experimental |  |  | Control |  |  |  | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Blumenthal 2005 (1) | 26 | 19.2 | 186 | 31.8 | 21.6 | 80 | 19.8% | -5.80 [-11.28, -0.32] |
| Gornet 2011 (2) | 19.2 | 18.2 | 405 | 25.3 | 19.8 | 172 | 49.8% | -6.10 [-9.55, -2.65] |
| Sasso 2008 (3) | 18 | 20.35 | 35 | 26 | 19.05 | 17 | 4.7% | -8.00 [-19.29, 3.29] |
| Strube 2016 (4) | 25.3 | 21.9 | 23 | 28.1 | 17.2 | 25 | 4.7% | -2.80 [-14.01, 8.41] |
| Zigler 2007 (5) | 36.2 | 21.5 | 161 | 41.3 | 18.3 | 75 | 21.0% | -5.10 [-10.41, 0.21] |
| **Total (95% CI)** |  |  | **810** |  |  | **369** | **100.0%** | **-5.76 [-8.20, -3.33]** |

Heterogeneity: Tau$^2$ = 0.00; Chi$^2$ = 0.52, df = 4 (P = 0.97); I$^2$ = 0%
Test for overall effect: Z = 4.64 (P < 0.00001)

Footnotes
(1) Oswestry Disability Index
(2) Oswestry Disability Index
(3) Oswestry Disability Index; SD imputed
(4) Oswestry Disability Index
(5) Oswestry Disability Index

### PICO 4 short-term: Revision rate (1 year)

No data available.

### PICO 4 short-term: Reoperation rate (1 year)

No data available.

### PICO 4 short-term: Complications or adverse events (1 year)

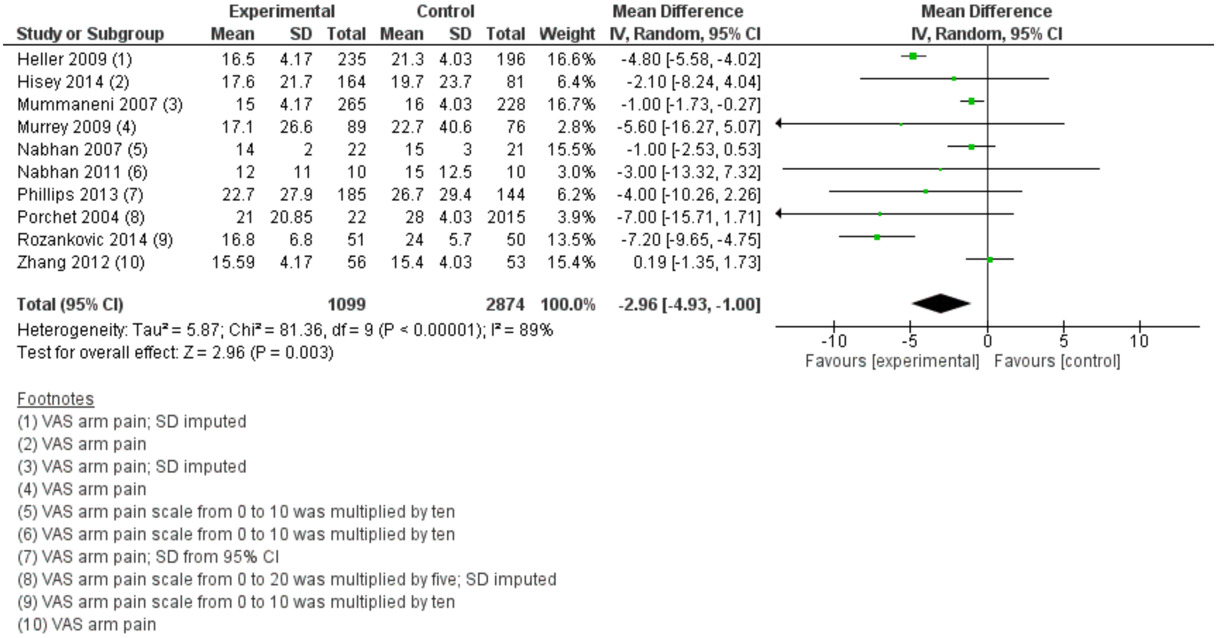
No data available.

### PICO 4 short-term: Serious adverse events (1 year)

No data available.

## VII. Appendix – PICO 5 short-term

### PICO 5 short-term: Radicular pain (1 year)



|  | Experimental |  |  | Control |  |  |  | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Heller 2009 (1) | 16.5 | 4.17 | 235 | 21.3 | 4.03 | 196 | 16.6% | -4.80 [-5.58, -4.02] |
| Hisey 2014 (2) | 17.6 | 21.7 | 164 | 19.7 | 23.7 | 81 | 6.4% | -2.10 [-8.24, 4.04] |
| Mummaneni 2007 (3) | 15 | 4.17 | 265 | 16 | 4.03 | 228 | 16.7% | -1.00 [-1.73, -0.27] |
| Murrey 2009 (4) | 17.1 | 26.6 | 89 | 22.7 | 40.6 | 76 | 2.8% | -5.60 [-16.27, 5.07] |
| Nabhan 2007 (5) | 14 | 2 | 22 | 15 | 3 | 21 | 15.5% | -1.00 [-2.53, 0.53] |
| Nabhan 2011 (6) | 12 | 11 | 10 | 15 | 12.5 | 10 | 3.0% | -3.00 [-13.32, 7.32] |
| Phillips 2013 (7) | 22.7 | 27.9 | 185 | 26.7 | 29.4 | 144 | 6.2% | -4.00 [-10.26, 2.26] |
| Porchet 2004 (8) | 21 | 20.85 | 22 | 28 | 4.03 | 2015 | 3.9% | -7.00 [-15.71, 1.71] |
| Rozankovic 2014 (9) | 16.8 | 6.8 | 51 | 24 | 5.7 | 50 | 13.5% | -7.20 [-9.65, -4.75] |
| Zhang 2012 (10) | 15.59 | 4.17 | 56 | 15.4 | 4.03 | 53 | 15.4% | 0.19 [-1.35, 1.73] |
| **Total (95% CI)** |  |  | **1099** |  |  | **2874** | **100.0%** | **-2.96 [-4.93, -1.00]** |

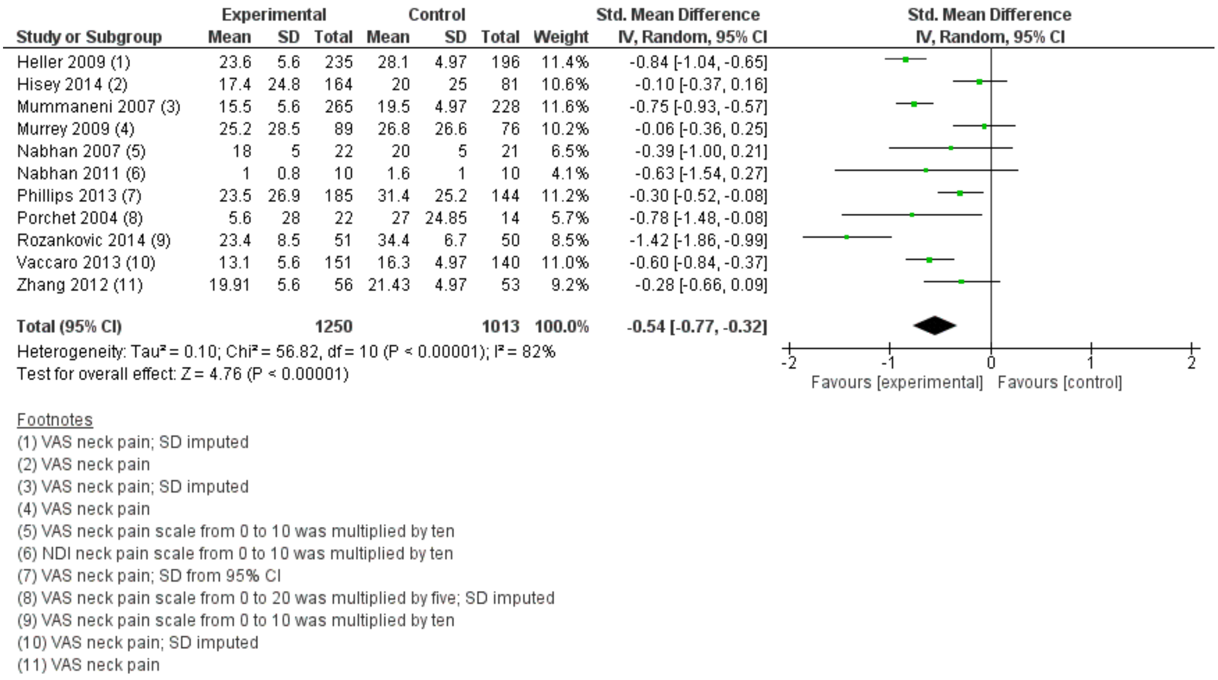
Heterogeneity: Tau$^2$ = 5.87; Chi$^2$ = 81.36, df = 9 (P < 0.00001); I$^2$ = 89%
Test for overall effect: Z = 2.96 (P = 0.003)

Footnotes
(1) VAS arm pain; SD imputed
(2) VAS arm pain
(3) VAS arm pain; SD imputed
(4) VAS arm pain
(5) VAS arm pain scale from 0 to 10 was multiplied by ten
(6) VAS arm pain scale from 0 to 10 was multiplied by ten
(7) VAS arm pain; SD from 95% CI
(8) VAS arm pain scale from 0 to 20 was multiplied by five; SD imputed
(9) VAS arm pain scale from 0 to 10 was multiplied by ten
(10) VAS arm pain

### PICO 5 short-term: Myelopathy (1 year)

No data available.

### PICO 5 short-term: Neck pain (1 year)



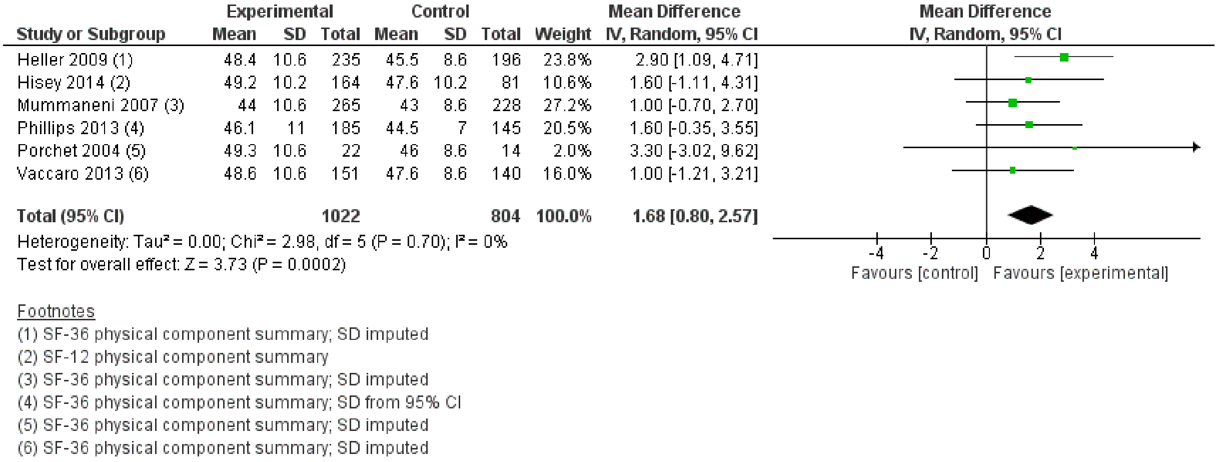
| | Experimental | | | Control | | | | Std. Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Heller 2009 (1) | 23.6 | 5.6 | 235 | 28.1 | 4.97 | 196 | 11.4% | -0.84 [-1.04, -0.65] |
| Hisey 2014 (2) | 17.4 | 24.8 | 164 | 20 | 25 | 81 | 10.6% | -0.10 [-0.37, 0.16] |
| Mummaneni 2007 (3) | 15.5 | 5.6 | 265 | 19.5 | 4.97 | 228 | 11.6% | -0.75 [-0.93, -0.57] |
| Murrey 2009 (4) | 25.2 | 28.5 | 89 | 26.8 | 26.6 | 76 | 10.2% | -0.06 [-0.36, 0.25] |
| Nabhan 2007 (5) | 18 | 5 | 22 | 20 | 5 | 21 | 6.5% | -0.39 [-1.00, 0.21] |
| Nabhan 2011 (6) | 1 | 0.8 | 10 | 1.6 | 1 | 10 | 4.1% | -0.63 [-1.54, 0.27] |
| Phillips 2013 (7) | 23.5 | 26.9 | 185 | 31.4 | 25.2 | 144 | 11.2% | -0.30 [-0.52, -0.08] |
| Porchet 2004 (8) | 5.6 | 28 | 22 | 27 | 24.85 | 14 | 5.7% | -0.78 [-1.48, -0.08] |
| Rozankovic 2014 (9) | 23.4 | 8.5 | 51 | 34.4 | 6.7 | 50 | 8.5% | -1.42 [-1.86, -0.99] |
| Vaccaro 2013 (10) | 13.1 | 5.6 | 151 | 16.3 | 4.97 | 140 | 11.0% | -0.60 [-0.84, -0.37] |
| Zhang 2012 (11) | 19.91 | 5.6 | 56 | 21.43 | 4.97 | 53 | 9.2% | -0.28 [-0.66, 0.09] |
| **Total (95% CI)** | | | **1250** | | | **1013** | **100.0%** | **-0.54 [-0.77, -0.32]** |

Heterogeneity: Tau$^2$ = 0.10; Chi$^2$ = 56.82, df = 10 (P < 0.00001); I$^2$ = 82%
Test for overall effect: Z = 4.76 (P < 0.00001)

Footnotes
(1) VAS neck pain; SD imputed
(2) VAS neck pain
(3) VAS neck pain; SD imputed
(4) VAS neck pain
(5) VAS neck pain scale from 0 to 10 was multiplied by ten
(6) NDI neck pain scale from 0 to 10 was multiplied by ten
(7) VAS neck pain; SD from 95% CI
(8) VAS neck pain scale from 0 to 20 was multiplied by five; SD imputed
(9) VAS neck pain scale from 0 to 10 was multiplied by ten
(10) VAS neck pain; SD imputed
(11) VAS neck pain

### PICO 5 short-term: Quality of life (1 year)

## Short form 36: physical component score

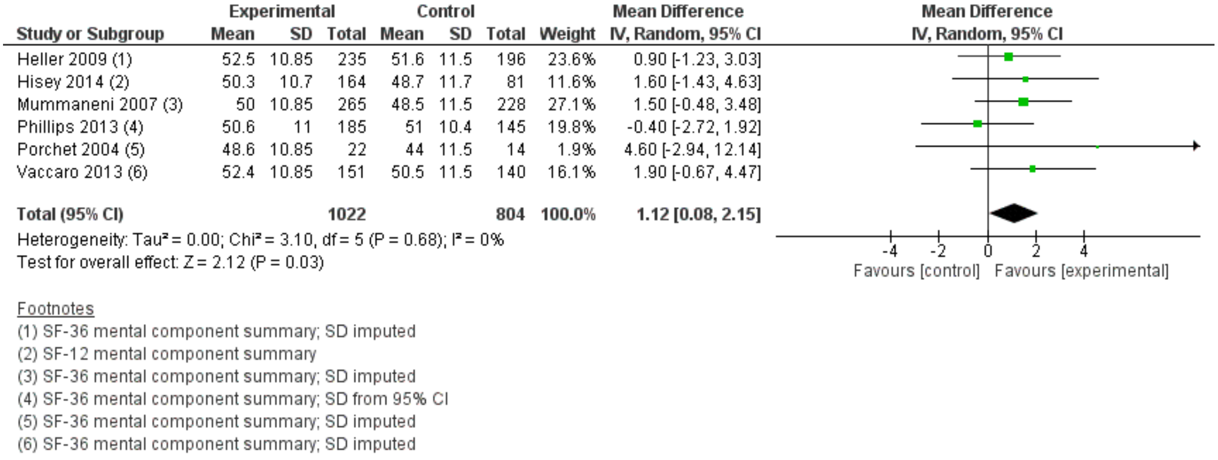


| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Heller 2009 (1) | 48.4 | 10.6 | 235 | 45.5 | 8.6 | 196 | 23.8% | 2.90 [1.09, 4.71] |
| Hisey 2014 (2) | 49.2 | 10.2 | 164 | 47.6 | 10.2 | 81 | 10.6% | 1.60 [-1.11, 4.31] |
| Mummaneni 2007 (3) | 44 | 10.6 | 265 | 43 | 8.6 | 228 | 27.2% | 1.00 [-0.70, 2.70] |
| Phillips 2013 (4) | 46.1 | 11 | 185 | 44.5 | 7 | 145 | 20.5% | 1.60 [-0.35, 3.55] |
| Porchet 2004 (5) | 49.3 | 10.6 | 22 | 46 | 8.6 | 14 | 2.0% | 3.30 [-3.02, 9.62] |
| Vaccaro 2013 (6) | 48.6 | 10.6 | 151 | 47.6 | 8.6 | 140 | 16.0% | 1.00 [-1.21, 3.21] |
| **Total (95% CI)** | | | **1022** | | | **804** | **100.0%** | **1.68 [0.80, 2.57]** |

Heterogeneity: Tau$^2$ = 0.00; Chi$^2$ = 2.98, df = 5 (P = 0.70); I$^2$ = 0%
Test for overall effect: Z = 3.73 (P = 0.0002)

Footnotes
(1) SF-36 physical component summary; SD imputed
(2) SF-12 physical component summary
(3) SF-36 physical component summary; SD imputed
(4) SF-36 physical component summary; SD from 95% CI
(5) SF-36 physical component summary; SD imputed
(6) SF-36 physical component summary; SD imputed

## Short form 36: mental component score

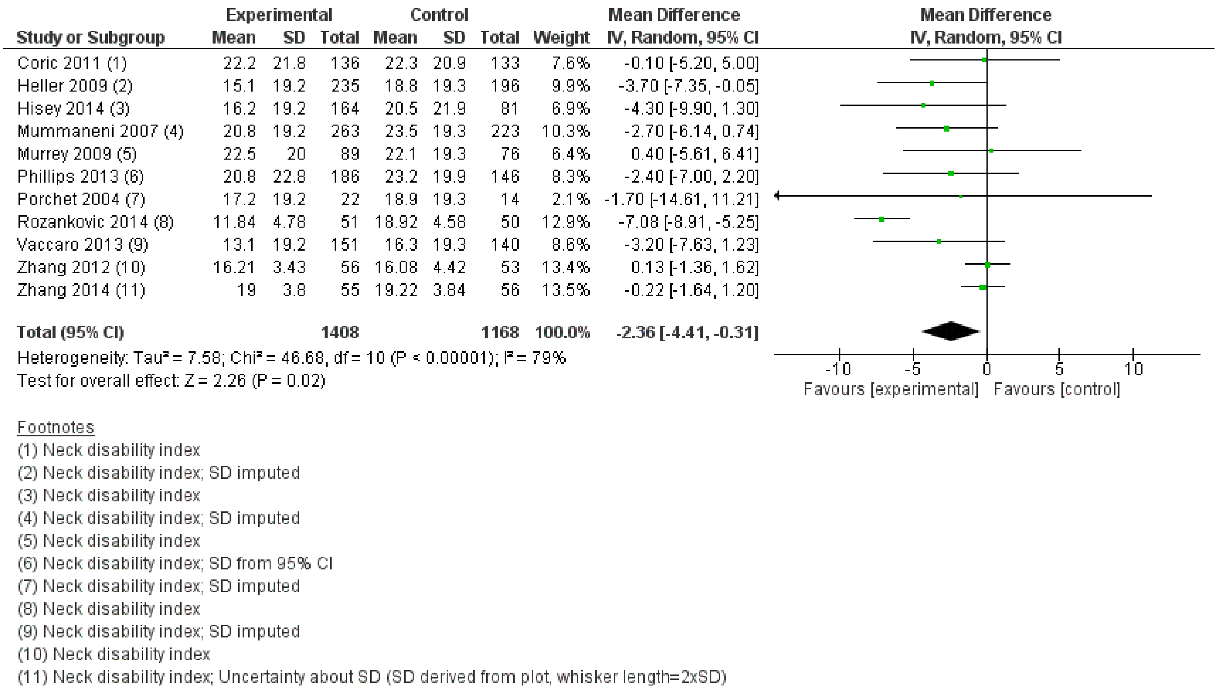


| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Heller 2009 (1) | 52.5 | 10.85 | 235 | 51.6 | 11.5 | 196 | 23.6% | 0.90 [-1.23, 3.03] |
| Hisey 2014 (2) | 50.3 | 10.7 | 164 | 48.7 | 11.7 | 81 | 11.6% | 1.60 [-1.43, 4.63] |
| Mummaneni 2007 (3) | 50 | 10.85 | 265 | 48.5 | 11.5 | 228 | 27.1% | 1.50 [-0.48, 3.48] |
| Phillips 2013 (4) | 50.6 | 11 | 185 | 51 | 10.4 | 145 | 19.8% | -0.40 [-2.72, 1.92] |
| Porchet 2004 (5) | 48.6 | 10.85 | 22 | 44 | 11.5 | 14 | 1.9% | 4.60 [-2.94, 12.14] |
| Vaccaro 2013 (6) | 52.4 | 10.85 | 151 | 50.5 | 11.5 | 140 | 16.1% | 1.90 [-0.67, 4.47] |
| **Total (95% CI)** | | | **1022** | | | **804** | **100.0%** | **1.12 [0.08, 2.15]** |

Heterogeneity: Tau$^2$ = 0.00; Chi$^2$ = 3.10, df = 5 (P = 0.68); I$^2$ = 0%
Test for overall effect: Z = 2.12 (P = 0.03)

Footnotes
(1) SF-36 mental component summary; SD imputed
(2) SF-12 mental component summary
(3) SF-36 mental component summary; SD imputed
(4) SF-36 mental component summary; SD from 95% CI
(5) SF-36 mental component summary; SD imputed
(6) SF-36 mental component summary; SD imputed

### PICO 5 short-term: Function (1 year)

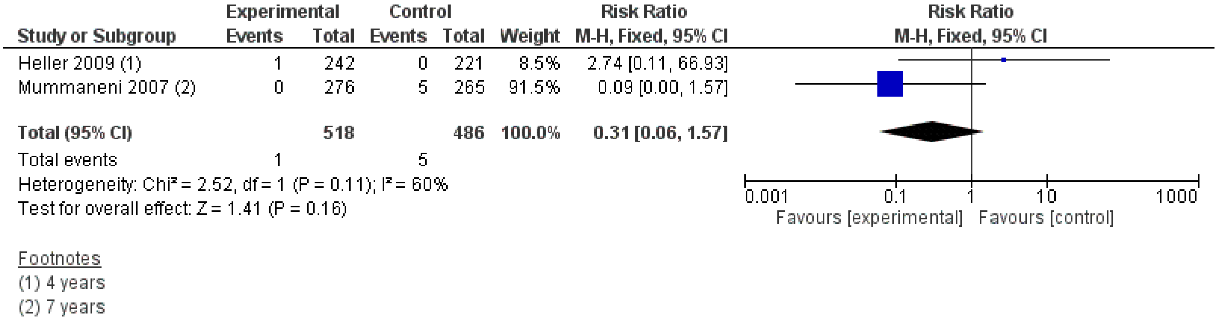


| | Experimental | | | Control | | | | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| Study or Subgroup | Mean | SD | Total | Mean | SD | Total | Weight | IV, Random, 95% CI |
| Coric 2011 (1) | 22.2 | 21.8 | 136 | 22.3 | 20.9 | 133 | 7.6% | -0.10 [-5.20, 5.00] |
| Heller 2009 (2) | 15.1 | 19.2 | 235 | 18.8 | 19.3 | 196 | 9.9% | -3.70 [-7.35, -0.05] |
| Hisey 2014 (3) | 16.2 | 19.2 | 164 | 20.5 | 21.9 | 81 | 6.9% | -4.30 [-9.90, 1.30] |
| Mummaneni 2007 (4) | 20.8 | 19.2 | 263 | 23.5 | 19.3 | 223 | 10.3% | -2.70 [-6.14, 0.74] |
| Murrey 2009 (5) | 22.5 | 20 | 89 | 22.1 | 19.3 | 76 | 6.4% | 0.40 [-5.61, 6.41] |
| Phillips 2013 (6) | 20.8 | 22.8 | 186 | 23.2 | 19.9 | 146 | 8.3% | -2.40 [-7.00, 2.20] |
| Porchet 2004 (7) | 17.2 | 19.2 | 22 | 18.9 | 19.3 | 14 | 2.1% | -1.70 [-14.61, 11.21] |
| Rozankovic 2014 (8) | 11.84 | 4.78 | 51 | 18.92 | 4.58 | 50 | 12.9% | -7.08 [-8.91, -5.25] |
| Vaccaro 2013 (9) | 13.1 | 19.2 | 151 | 16.3 | 19.3 | 140 | 8.6% | -3.20 [-7.63, 1.23] |
| Zhang 2012 (10) | 16.21 | 3.43 | 56 | 16.08 | 4.42 | 53 | 13.4% | 0.13 [-1.36, 1.62] |
| Zhang 2014 (11) | 19 | 3.8 | 55 | 19.22 | 3.84 | 56 | 13.5% | -0.22 [-1.64, 1.20] |
| **Total (95% CI)** | | | **1408** | | | **1168** | **100.0%** | **-2.36 [-4.41, -0.31]** |

Heterogeneity: Tau² = 7.58; Chi² = 46.68, df = 10 (P < 0.00001); I² = 79%
Test for overall effect: Z = 2.26 (P = 0.02)

Footnotes
(1) Neck disability index
(2) Neck disability index; SD imputed
(3) Neck disability index
(4) Neck disability index; SD imputed
(5) Neck disability index
(6) Neck disability index; SD from 95% CI
(7) Neck disability index; SD imputed
(8) Neck disability index
(9) Neck disability index; SD imputed
(10) Neck disability index
(11) Neck disability index; Uncertainty about SD (SD derived from plot, whisker length=2xSD)

## PICO 5 short-term: Revision rate (1 year)

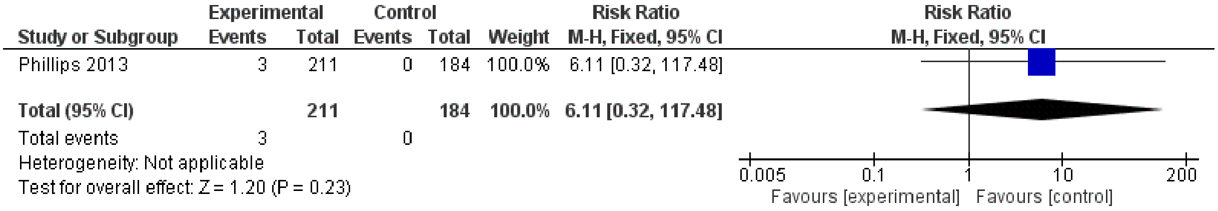


| | Experimental | | Control | | | Risk Ratio |
|---|---|---|---|---|---|---|
| Study or Subgroup | Events | Total | Events | Total | Weight | M-H, Fixed, 95% CI |
| Heller 2009 (1) | 1 | 242 | 0 | 221 | 8.5% | 2.74 [0.11, 66.93] |
| Mummaneni 2007 (2) | 0 | 276 | 5 | 265 | 91.5% | 0.09 [0.00, 1.57] |
| **Total (95% CI)** | | **518** | | **486** | **100.0%** | **0.31 [0.06, 1.57]** |
| Total events | 1 | | 5 | | | |

Heterogeneity: Chi² = 2.52, df = 1 (P = 0.11); I² = 60%
Test for overall effect: Z = 1.41 (P = 0.16)

Footnotes
(1) 4 years
(2) 7 years

## PICO 5 short-term: Reoperation rate (1 year)



| | Experimental | | Control | | | Risk Ratio |
|---|---|---|---|---|---|---|
| Study or Subgroup | Events | Total | Events | Total | Weight | M-H, Fixed, 95% CI |
| Phillips 2013 | 3 | 211 | 0 | 184 | 100.0% | 6.11 [0.32, 117.48] |
| **Total (95% CI)** | | **211** | | **184** | **100.0%** | **6.11 [0.32, 117.48]** |
| Total events | 3 | | 0 | | | |

Heterogeneity: Not applicable
Test for overall effect: Z = 1.20 (P = 0.23)

## PICO 5 short-term: Complications or adverse events (1 year)

No data available.

## PICO 5 short-term: Serious adverse events (1 year)

No data available.