Definition of the Different Levels of Evidence (LoE)

**Articles on treatment**

<table>
<thead>
<tr>
<th>Level</th>
<th>Risk of bias</th>
<th>Study design</th>
<th>Criteria</th>
</tr>
</thead>
</table>
| I     | Low risk     | Good quality RCT | • Random sequence generation
|       |               |              | • Allocation concealment
|       |               |              | • Intent-to-treat analysis
|       |               |              | • Blind or independent assessment for important outcomes
|       |               |              | • Co-interventions applied equally
|       |               |              | • F/U rate of 80% +
|       |               |              | • Adequate sample size |
| II    | Moderately low risk | Moderate or poor quality RCT | • Violation of one of the criteria for good quality RCT
|       |               |              | • Blind or independent assessment in a prospective study, or use of reliable data\(^a\) in a retrospective study
|       |               |              | • Co-interventions applied equally
|       |               |              | • F/U rate of 80% -
|       |               |              | • Adequate sample size
|       |               |              | • Controlling for possible confounding\(^b\) |
| III   | Moderately high risk | Moderate or poor quality cohort | • Violation of any of the criteria for good quality cohort
|       |               |              | • Any case-control design |
| IV    | High risk    | Case-control | Any case series design |

\(^a\)Outcome assessment is independent of healthcare personnel judgment. Reliable data are data such as mortality or re-operation.

\(^b\)Authors must provide a description of robust baseline characteristics, and control for those that are unequally distributed between treatment groups.

**Determination of Overall Strength of Evidence (SoE)**

After individual article evaluation, the overall body of evidence with respect to each outcome is determined based on principles outlined by the Grades of Recommendation Assessment, Development and Evaluation (GRADE) Working Group and recommendations made by the Agency for Healthcare Research and Quality (AHRQ). Qualitative analysis is performed considering the AHRQ required and additional criteria. The table below provides an outline of the method used to determine the final SoE.

**Strength of Evidence for Existing Systematic Reviews**

Level of evidence ratings for Cochrane reviews and other systematic reviews are assigned a baseline score of HIGH (Level I). Level II/III reviews are assigned a baseline score of MODERATE (Level II). Level IV reviews are assigned a baseline score of LOW (Level III). The rating can be upgraded or downgraded based on adherence to the core criteria for qualitative, and quantitative analyses for systematic reviews (there is a reference/evaluation table for this).

The following four possible levels and their definitions are reported:

- **High**: High confidence that the evidence reflects the true effect. Further research is very unlikely to change the conclusion.
- **Moderate**: Moderate confidence that the evidence reflects the true effect. Further research may change the confidence in the estimate of effect and may change the estimate.
- **Low**: Low confidence that the evidence reflects the true effect. Further research is likely to change the confidence in the estimate of effect and likely to change the estimate.
- **Insufficient**: Evidence either is unavailable or does not permit a conclusion.

**Articles on prognosis or risk**

<table>
<thead>
<tr>
<th>Level</th>
<th>Risk of bias</th>
<th>Study design</th>
<th>Criteria</th>
</tr>
</thead>
</table>
| I     | Low risk     | Good quality RCT | • Prospective design
|       |               |              | • Patients at similar point in the course of their disease or treatment
|       |               |              | • F/U rate of 80% -
|       |               |              | • Patients followed long enough for outcomes to occur
|       |               |              | • Accounting for other prognostic factors\(^c\) |
| II    | Moderately low risk | Moderate or poor quality cohort | • Prospective design, with violation of one of the other criteria for good quality cohort study
|       |               |              | • Retrospective design, meeting all the rest of the criteria in level I |
| III   | Moderately high risk | Poor quality cohort | • Prospective design with violation of 2 or more criteria for good quality cohort, or
|       |               |              | • Retrospective design with violation of 1 or more criteria for good quality cohort
|       |               |              | • A good case-control study\(^d\)
|       |               |              | • A good cross-sectional study\(^d\) |
| IV    | High risk    | Poor quality- level III/A study | • Other than a good case-control study
|       |               |              | • Other than a good cross-sectional study
|       |               |              | • Any case series\(^e\) design |

\(^c\)Cohort studies follow individuals with the exposure of interest over time and monitor for occurrence of the outcome of interest.

\(^d\)Applies to cohort studies only.

\(^e\)Authors must consider other factors that might influence patient outcomes and should control for them if appropriate.

\(^f\)A good case-control study must have all of the following: all incident cases from the defined population over a specified time period, controls that represent the population from which the cases come, exposure that precedes an outcome of interest, and accounting for other prognostic factors.

\(^g\)A good cross-sectional study must have all of the following: a representative sample of the population of interest, an exposure that precedes an outcome of interest (e.g., sex, genetic factor), an accounting for other prognostic factors, and for surveys, at least a 80% return rate.

\(^h\)A case-series design for prognosis is one where all the patients in the study have the exposure of interest. Since all the patients have the exposure, risks of an outcome can be calculated only for those with the exposure, but cannot be compared with those who do not have the exposure. For example, a case-series evaluating the effect of smoking on spine fusion that only recruits patients who smoke can simply provide the risk of patients who smoke that result in pseudarthrosis but cannot compare this risk to those that do not smoke.

**Definitions of the Different Levels of Evidence for Reliability Studies**

<table>
<thead>
<tr>
<th>Level</th>
<th>Study type</th>
<th>Criteria</th>
</tr>
</thead>
</table>
| 1     | Good quality study | • Broad spectrum of persons with the expected condition
|       |             | • Adequate description of methods for replication
|       |             | • Blinded performance of tests, measurements or interpretation
|       |             | • Second test/interpretation performed independently of the first |
| 2     | Moderate quality | • Violation of any one of the criteria for a good quality study |
| 3     | Poor quality study | • Violation of any two of the criteria |
| 4     | Very poor quality study | • Violation of all three of the criteria |

\(\text{\textsuperscript{a}}\)Required domains: risk of bias, consistency, directness, precision. Plausible confounding that would decrease observed effect is accounted for in our baseline risk of bias assessment through individual article evaluation. Additional domains: dose-response, strength of association, publication bias.

\(\text{\textsuperscript{b}}\)Single study = “consistency unknown.”