Journal of The Korean Society of Integrative Medicine, 2023, 11(2), 231~242
https://doi.org/10.15268/ksim.2023.11.2.231
ISSN 2288-1174(Print)
ISSN 2383-9651(Online)

국내 물리치료분야에 대한 질적 평가와 근거 수준 및 권고 등급 모형 개발 방안

조성현[1] · 이정우[2†]

[1]남부대학교 물리치료학과 교수, [2†]광주여자대학교 물리치료학과 교수

Qualitative Assessment and Development of Level of Evidence and Strength of Recommendation Models in the Field of Physical Therapy in Korea

Sung-Hyoun Cho, PT, Ph.D[1] · Jeong-Woo Lee, PT, Ph.D[2†]

[1]Dept. of Physical Therapy, Nambu University, Professor

[2†]Dept. of Physical Therapy, Kwangju Women's University, Professor

Abstract

Purpose : This study aimed to identify ways to improve the quality of physical therapy research and ultimately review the current situation to improve evidence-based decision-making in physical therapy.

Methods : For better evidence-based decision-making in physical therapy, researchers should review the quality assessment of articles in more detail and report their findings for valid and appropriate level of evidence and strength of recommendations. The level of evidence affects how well the findings are derived from well-designed literature. The evaluation of the evidence focuses primarily on the study design and the degree of bias that may compromise the validity of the findings. The final recommendation is based on a combination of the study design and literature quality. To uncover gems of information in each paper, a risk of bias assessment should be performed after the literature has been initially selected.

Results : Researchers should consider the complexity of the intervention, appropriate grouping, and calculation of effect sizes for the intervention. Researchers conducting systematic reviews should provide a detailed description of the quality assessment performed and present a detailed analysis of their interpretation of the results. The results of systematic reviews and meta-analyses should be interpreted with caution and include a risk of bias assessment. Guidelines for the level of evidence and strength of recommendations should be developed and utilized more broadly to improve reporting practices in physical therapy.

Conclusion : Researchers should be knowledgeable about the strengths and limitations of each study design and methodology. In the future, researchers will also need to improve their ability to critically evaluate their findings, given the potential for their results to influence clinical practice.

Key Words : development, evidence level, physical therapy, qualitative Assessment, recommendation

†교신저자 : 이정우, jwlee@kwu.ac.kr

* 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행되었음(No. 2022R1F1A1067604).

제출일 : 2023년 4월 17일 | 수정일 : 2023년 5월 11일 | 게재승인일 : 2023년 5월 19일

Ⅰ. 서 론

1. 연구의 배경 및 필요성

근거 기반 실무(evidence based-practice; EBP)는 환자를 위한 최선의 의사 결정 방법이다(Jette 등, 2003). 물리치료사들은 근거 기반 실무에서 연구 결과를 사용하는 데 익숙해지고 있다. 체계적 문헌고찰은 임상 전문 지식과 환자가 보고한 결과를 포함한 최고의 근거 및 과학적 연구의 원천 중 하나이다(Salbach 등, 2007). 연구 결과를 사용하여 환자에게 최적의 물리치료 중재를 결정하는 데 관심이 있는 물리치료사는 높은 질적 수준의 체계적 문헌고찰 및 메타분석을 찾을 수 있다(Bagg 등, 2018). 물리치료사는 환자를 위한 더 나은 의사결정에 도움이 되는 체계적 문헌고찰 및 메타분석을 작성하고 이해할 수 있는 지식과 기술을 갖추어야 한다(Kurichi & Sonnad, 2006).

체계적 문헌고찰은 연구 주제와 관련된 모든 연구를 수집, 종합적으로 분석, 평가하는 문헌고찰이다(Ahn & Kang, 2018). 따라서 하나의 임상연구가 아닌 동일한 치료 효과에 대한 연구를 통해 여러 임상연구 논문이 동일한 결과를 보인다면, 하나의 개별 임상연구 논문보다 치료 효과에 대한 보다 정확한 근거를 제시할 수 있다(Harbour & Miller, 2001). 임상적 중재는 과학적인 증거에 근거해야 하며, 중재는 타당하고 투명한 절차와 방법을 통해 시행되어야 한다. 임상 적응증에 대해 사용 가능한 많은 또는 모든 중재의 비교 효과를 평가하는 것은 어려운 일이다(Bafeta 등, 2013). 과학적 절차의 비뚤림(bias)은 부적절한 임상적 관행으로 이어질 수 있다. 체계적 문헌고찰을 위한 코크란 핸드북은 문헌고찰을 완료한 연구자의 평가 프로세스를 지원한다(Higgins 등, 2022a). 근거 수준 결정에 객관적인 기준을 사용하는 GRADE 방법론(the grading of recommendations, assessment, development and evaluation) 기준에는 비뚤림 위험, 불일치, 간접성, 부정확성, 출판 편향의 5가지 질 평가가 포함된다(Schünemann 등, 2022).

물리치료 연구에 대한 체계적 문헌고찰에는 무작위배정연구, 관찰 연구와 같은 다양한 연구 디자인이 포함되며, 연구자는 물리치료 연구에서 검증된 검토를 탐색하는 적절한 방법을 고려해야 한다. 체계적 문헌고찰을 위한 코크란 핸드북 버전 6.3이 2022년에 발표되었다(Higgins 등, 2022b). ROB 2.0 도구는 RCT 편향의 질을 평가하기 위한 골드 스탠다드(gold standard)이며(Luchini 등, 2021), 코크란 핸드북 버전 6.3에서 비무작위 배정 중재 연구를 평가하기 위해 개발된 비무작위 배정 중재 연구의 비뚤림 위험(the risk of bias in non-randomized studies of interventions; ROBINS-I)도 있다. 주요 차이점은 검토자가 중재 후 효과에 초점을 맞춰 편향성('low', 'moderate', 'serious', 'critical' 사용)을 쉽게 비판할 수 있다는 점이다(Sterne 등, 2022).

연구자는 문헌을 선택할 때 최소의 질 기준을 결정하거나 연구결과들이 이질성이 있을 때 문헌 간의 질 차이를 확인하고 싶을 때 문헌의 질 평가를 실시한다. 또한 메타분석에서 연구의 질에 따라 연구결과에 의미를 더 부여하거나, 추론의 강도를 결정하는 것을 도와서 결과 해석을 가이드 할 때 문헌의 질 평가를 실시한다. 그리고 체계적 문헌고찰이나 의료기술 평가 과정에서 근거 수준과 권고등급 도출에 활용하기 위한 목적으로 질 평가를 수행하기도 한다(Kim 등, 2020). 물리치료 연구자는 연구의 연구 설계와 목표를 반영하고 중재의 복잡성, 적절한 그룹화 및 과학적인 효과 크기 계산을 고려하기 위해 적절하고 업데이트된 질 평가 도구를 사용해야 한다.

2. 연구의 목적

본 연구의 목적은 현재 물리치료 연구의 질적 평가 도구를 점검하고 의학 분야에서 사용되고 있는 질적 평가 도구와 근거 수준 및 권고 등급의 방법들이 물리치료 분야에서 적용 시의 문제점을 분석하고 이를 보완할 수 있는 새로운 모형 개발방안의 방향을 제안하고자 한다.

Ⅱ. 물리치료 연구에 대한 질적 평가 도구

체계적 문헌 고찰 연구를 진행하면서 문헌분류와 비뚤림 위험 평가 과정이 필요하다. 특히 연구 설계에 적합한 비뚤림 평가도구를 적용하여 문헌의 질 평가를 하

는 것이 매우 중요하다(Kim 등, 2020). 1차적으로 대상 문헌이 선정된 다음 시행해야 할 과정은 각각의 논문이 주는 정보의 옥석을 가리는 작업이 비뚤림 위험 평가이다. 문헌의 질 평가는 훌륭한 문헌을 골라내는 절차라기보다 비뚤림 위험에 대한 평가를 의미한다(Kim 등, 2021). 비뚤림은 체계적인 오류로 결과나 추정에 있어 참값으로부터 벗어남을 의미한다. 비뚤림은 중재 효과를 과소추정 혹은 과다추정하게 하는 요소가 될 수 있어 중요하다(Higgins 등, 2022a). 체계적 문헌고찰 시 연구의 질에 대한 평가가 반드시 이루어져야 하는 이유는 분석에 사용된 연구들이 실제로는 질이 낮은 연구임에도 불구하고 분석 결과를 마치 높은 수준의 연구를 통해 얻어진 결과인 것처럼 보일 수 있기 때문이다. 이보다 더 중요한 문제는 질이 낮은 연구들을 결합하게 되면 비뚤림이 증가하게 되고 결과가 왜곡되는 결합 추정치가 얻어지기 때문이다(Kim 등, 2021).

의학 분야에 있어서 연구들에 대한 질적 평가 도구들은 다양하게 이용되고 있으나 가장 일반적으로 널리 알려진 평가도구들을 살펴보면, 체계적 문헌고찰 연구에 대한 질적 평가는 AMSTAR 2(assessment of multiple systematic reviews 2)가 널리 사용되고 있다(Shea 등, 2017). 중재의 효과를 평가하는 무작위배정 비교임상시험(randomized controlled trials; RCT) 설계의 비뚤림 위험 평가도구로서 코크란 비뚤림 위험평가(ROB)는 2008년에 처음 발표되었고 2011년에 업데이트되었다. 코크란 ROB 2.0(Cochrane's risk of bias 2.0)도구는 2019년 6월에 개정되었고, 비뚤림 위험 평가도구로 가장 널리 알려져 있고 활용되고 있다(Higgins 등, 2022a). 체계적 문헌고찰을 위한 코크란 핸드북 버전 6.3이 2022년 2월에 발표되었다(Higgins 등, 2022b). 코크란 ROB 2.0 도구의 주요 특징은 연구자가 검토된 연구에 편향이 존재하는지 간단히 판단할 수 있고, RCT 설계를 넘어 RCT 설계 내에서 특정 결과 결과에 대한 편향성을 평가할 수 있다는 것이다(McGuinness & Higgins, 2021). 코크란 ROB 2.0 도구는 신호 질문 응답에 따라 판단을 매핑하는 알고리즘을 제시하여 평가자간 신뢰도를 높이는 방향으로 수정되었다(Sterne 등, 2019). 코크란에서는 RCT 연구 외에는 많은 비뚤림 요소를 가지고 있어 RCT 연구만을 합성하도록 권고하고 있다(Higgins 등, 2022a).

PEDro(physiotherapy evidence database)는 무작위 임상실험 논문, 문헌고찰 그리고 근거중심 임상 가이드라인, PEDro 평가 등급을 사용하여 논문의 수준을 평가한다(Ma 등, 2020). 물리치료학 분야에서는 PEDro 평가도구가 가장 많이 활용되고 있다. PEDro Scale은 0~10점 체계로 되어 있다. 임상시험의 신뢰성을 평가하기 위해서 무작위 배정(random allocation), 배정 숨김(concealment of allocation), 그룹의 기준점(baseline) 비교, 환자/치료사/평가자의 blind 유무, 치료의 목적을 분석, 그리고 적절한 후속평가(follow-up)와 같은 기준을 확인한다(de Morton, 2009).

비무작위 임상실험 연구들은 코크란 ROBINS-I(the risk of bias in non-randomized studies of interventions)(Sterne 등, 2022), 국내의 RoBANS(risk of bias assessment tool for non-randomized studies) 평가 도구가 가장 많이 활용되고 있다(Kim 등, 2013). 메타분석 연구에서는 ROB2 도구/ROBINS-I 도구를 사용하여 비뚤림 평가를 의미 있게 해석할 수 있다.

뉴캐슬-오타와 척도(Newcastle-Ottawa scale)는 비무작위 관찰연구의 질 평가를 위해서 개발된 도구로 환자-대조군, 전향적 연구, 코호트 연구를 포함한 관찰 연구를 검토할 때 사용된다(Luchini 등, 2021). 뉴캐슬-오타와 척도는 세 가지 영역(연구 참여자, 사례/대조군 간 비교, 결과 평가)으로 구성되어 있으며, 이는 질 개념과 연계되어 있으며, 검토가 상당히 주관적인 특성을 가지고 있으므로 최소 2명의 검토자를 두는 것이 좋다(Luchini 등, 2021). 세 영역의 총 8개 문항에서 '★'의 점수범위는 최소 0점에서 최대 9점까지이며, 평가대상 연구의 질 평가는 '★'의 수가 많을수록 질적인 논문으로 평가한다(Lim 등, 2011; Wells 등, 2000).

관찰 연구의 메타분석에 대한 보고 도구로는 MOOSE(meta-analysis of observational studies in epidemiology)를 사용한다(Stroup 등, 2000). 관찰연구를 위한 보고지침(the strengthening the reporting of observational studies in epidemiology; STROBE)은 관찰 연구의 정확하고 완전한 보고에 포함되어야 할 사항에 대한 권장 사항을 개발하였다. STROBE 체크리스트는 메타분석을 위한 질 평가 도구가 아닌 개별 관찰연구에 대한 보고 기준이다(Alsalaheen 등, 2017).

질 평가 도구와 관련하여, 통합 임상시험 보고 표준지침(consolidated standards of reporting trials; CONSORT)은 무작위 배정 임상시험의 개별 연구 보고 방법에 대한 일련의 권고사항일 뿐, 실제로 메타분석의 내부 타당성을 확립하기 위한 질 평가 도구는 아니다(Li 등, 2015). 메타분석 연구는 가장 높은 수준의 과학적 근거로 간주되지만, 내부 타당도에 영향을 미치는 선택, 성과, 소모, 검출 편향 등 많은 잠재적 편향의 원인이 있다(Jadad 등, 1996; Murad 등, 2014). 따라서 종합된 결과의 품질을 보장하기 위해서는 메타분석 연구에 포함된 개별 연구에 대해 유효하고 표준화된 방법론적 품질 평가 도구를 사용하는 것이 중요하다. 연구유형에 따른 질 평가도구는 다양한 기관 또는 연구자에 의해 지속적으로 개발되고 있다. Ma 등(2020)의 연구에는 다양한 연구유형에 따른 질 평가도구가 잘 소개되어 각 도구의 내용을 검토한 후 최선의 도구를 선택하는 것이 연구자에게 도움이 된다(Ma 등, 2020).

Ⅲ. 물리치료 연구의 근거 수준과 권고등급 개발 방안

물리치료 연구자들은 실험, 준실험, 비실험 등 다양한 연구 설계를 사용하여 연구를 수행해 왔다. 그러나 대부분의 근거 등급은 무작위배정 비교임상시험(randomized controlled trials; RCT)의 독립변수의 효과를 파악하는 데 편향되어 있어 이제는 연구 대상, 환경, 설계의 적합성에 대한 평가 범위로 확장되어야 한다(Frieden, 2017; Shin, 2017). 무작위배정 비교임상시험은 가장 높은 수준의 근거로 간주되지만, 내적 타당도는 우수하더라도 외적 타당도는 떨어질 수 있다(Deaton & Cartwright, 2018; Frieden, 2017). 외적 타당도는 결과의 일반화 가능성을 의미하며, 연구결과들을 어떤 인구, 장소, 치료변수, 측정변수에 적용할 수 있는지 적용 가능성이다. 내적 타당도는 해당 연구결과가 원래 연구하고자 했던 상황을 얼마나 반영하는지에 관한 정도이며, 방법론적인 질에 대한 평가를 의미한다(Kim 등, 2020).

근거수준(level of evidence)이란 입증된 근거의 강도(strength)이며, 현재까지의 근거를 바탕으로 특정 중재의 효과에 대해 확신하는 정도를 의미한다(Guyatt 등, 2011). 근거수준은 일반적으로 연구 설계, 문헌의 질, 근거의 양, 근거의 일관성, 근거의 직접성의 다섯 가지 요소를 일부 혹은 전체를 이용하여 평가한다(Kim 등, 2015a, Kim 등, 2015b). 권고등급(strength of recommendation)이란 권고 대상 환자에게 해당 중재를 시행하였을 때 위해(harm)보다 이득(benefit)이 더 클 것으로 혹은 작을 것으로 확신하는 정도이다. 권고등급은 근거수준, 임상적 이득 및 위해, 건강결과와 관련된 가치, 건강결과의 이득 및 자원사용 편익을 고려하여 결정한다(Kim 등, 2020).

기존에는 연구 설계에 따라서 단순히 권고의 등급을 결정하는 경우가 있었다(Fig 1). 최근에는 GRADE 등의 권고체계를 통해 단순히 연구 설계에 따라서 등급을 결정하지 않고 연구결과가 얼마나 일반적으로 적용할 수 있을 만큼 확실한 근거인지를 더 반영하여 체계를 내리고 있다(Fig 2)(Kim 등, 2017; Kim 등, 2020).

Fig 1. Levels of evidence before GRADE
Adapted from "Murad MH, Asi N, Alsawas M, et al(2016). New evidence pyramid. BMJ Evidence-Based Medicine, 21(4), 125-127."

근거의 수준과 권고의 등급을 살펴보면, SIGN(scottish intercollegiate guideline network)의 8개 근거 수준(1++, 1+, 1-, 2++, 2+, 2-, 3, 4), SIGN의 4개 권고등급(A, B, C, D), USPSTF(U.S. department of health and human services)의 근거수준(높음, 적정, 낮음), USPSTF의 5개 권고등급(A, B, C, D, I), GRADE의 4개 근거수준(high, moderate, low, very low)(Balshem 등, 2011), GRADE의 4개 권고등급(강하게 권고하지 않음, 약하게/조건부 권고하지 않음,

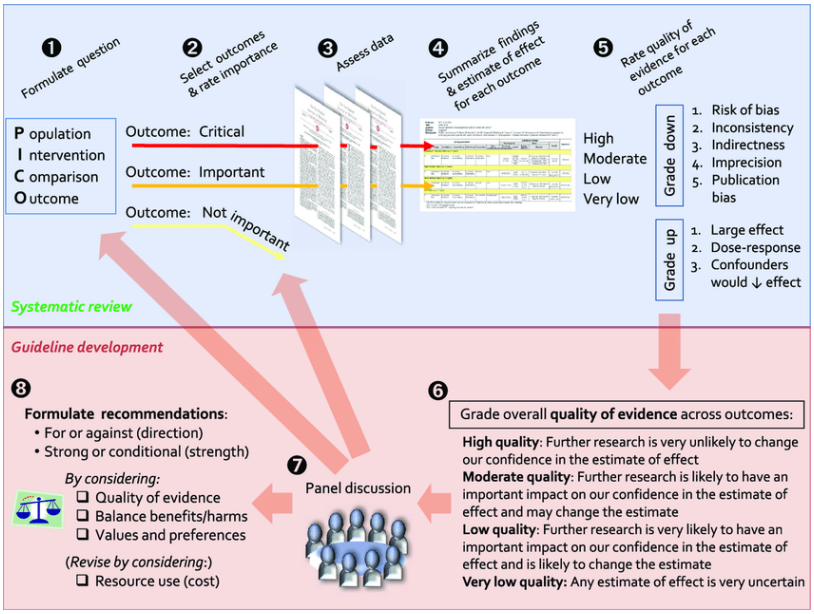

Fig 2. Schematic view of the GRADE approach (https://www.researchgate.net/figure/Schematic-representation-of-the-GRADE-approach-for-synthesizing-evidence-and-developing_fig1_323626314/download)

약하게/조건부 권고, 강하게 권고)으로 가장 널리 활용되고 있다(Andrews 등, 2013). 그러나 특정 질환에 대한 치료법의 근거수준을 알아보려고 할 때 각각의 질적 평가 도구들은 서로 평가 체계가 다르다. 그러므로 연구 설계의 질적수준 피라미드와 관계없이 통합적으로 비교분석할 수 있는 물리치료 연구의 질적 평가도구가 필요하다. 또한 현재 코크란 리뷰에서 적용하고 있는 GRADE 접근법에 의한 연구들의 질적 평가는 매우 엄격하게 평가하고 있다. 물리치료 분야의 연구들은 실제 임상에서 대규모의 연구가 어려운 실정이기에 대부분 소규모의 연구가 주로 이루어지고 있는 실정이다. 따라서 코크란 리뷰에서 보고되는 물리치료 분야의 연구들은 GRADE 접근법에 의한 질적 평가에서 높은 수준으로 평가 받기 어렵기 때문에 강한 근거로 인정받는 치료방법들은 극히 드문 실정이다.

국내의 건강보험에도 적용받는 경피신경전기자극(transcutaneous electrical nerve stimulation; TENS)의 경우에도 현재 코크란 리뷰의 체계적 문헌 고찰 연구들에서는 급성통증(Johnson 등, 2015), 만성 목통증(Martimbianco 등, 2019), 만성 허리통증(Khadilkar 등, 2008), 신경인성(neuropathic) 통증(Gibson 등, 2017)에 대한 경피신경전기자극의 임상적 효과는 아직 근거가 불충분한 것으로 보고되고 있다. 여러 개별적 임상실험 연구들은 효과적으로 보고되고 있음에도 대부분 대규모의 질적 수준이 높은 연구들의 수가 부족하기 때문으로 설명하고 있다(Gibson 등, 2017; Johnson 등, 2015; Khadilkar 등, 2008; Martimbianco 등, 2019). 만성 허리통증에 대한 경피신경전기자극의 효과는 적은 수의 연구들에서도 효과적이지 못한 것으로 보고되었다(Khadilkar 등, 2008).

또한 뇌성마비(cerebral palsy) 환자에 대한 기능향상을 위해 현재 임상에서 시행되고 건강보험에도 적용 중인 신경발달치료(neuro-developmental treatment; NDT)도 코크란 연합에서 제공하는 비뚤림 위험평가 및 GRADE 접근법에 의한 체계적 문헌고찰 연구(Novak 등, 2013)에서는 기능향상 측면에서의 효과가 불확실하며, Zanon 등(2019)의 체계적 문헌고찰 연구에서도 전통적인 물리치료 방법과 비교해서 효과가 불충분하다고 보고되고 있다(Zanon 등, 2019).

그런데 이때 생각해 볼 점은 통계적 오류에서 유의수준을 낮추게 될 경우 2종 오류의 위험성이 높아지게 되

는 것을 생각해 볼 때, 과연 연구들의 현행의 코크란 ROB 2.0 등의 엄격한 평가도구에 의해 질적 수준이 낮게 평가된 연구의 경우 이러한 연구들이 질적 수준이 낮음에도 효과가 있을 확률에 대한 부분을 검토할 필요성이 있다고 생각한다. 따라서 현행의 코크란 ROB 2.0 등의 엄격한 평가도구를 적용하더라도 이러한 2종 오류의 위험성이 높아질 수 있는 점을 고려한 새로운 평가도구의 개발이 필요하다.

질적 평가 수행에 있어서는 연구 설계의 차이를 포함하여 비뚤림 위험이 연구결과에 미치는 영향과 방향에 대한 충분한 고찰이 필요하다. 연구에 포함된 대상자가 연구결과에 어느 정도의 영향을 미칠지, 연구에 포함되지 않은 대상자와 얼마나 같은 결과를 가져올 수 있을지 등 다양한 것들을 고려해야 한다(Kim 등, 2021; Kim 등, 2020).

또한 물리치료 분야의 연구들은 일반 의학 분야와는 연구 설계 형태가 다르기 때문에 물리치료학 연구 분야들의 연구 설계 형태에 따른 근거수준 피라미드를 개발하고 각각의 근거수준에 따른 가중치 부여 및 개별 연구들의 점수화된 질적 평가도구를 통해 개별연구의 질적 평가 및 근거등급을 하나의 점수체계로 비교할 수 있도록 하는 새로운 질적 평가 및 근거수준 및 권고등급에 대한 모형개발이 필요한 실정이다.

질적 평가에 선택된 문헌들은 각각의 체크리스트를 활용하여 문헌에 대한 타당성 평가를 수행한다. 이 결과를 토대로 문헌의 근거수준을 결정하고 권고 등급에 영향을 미치게 된다. 근거 평가는 주로 연구 설계에 초점을 두며, 연구결과의 타당성을 손상시키는 비뚤림이 얼마나 존재하는지에 의해 결정된다. 권고등급은 연구 설계와 문헌의 질 평가를 함께 고려하여 최종 결정한다.

1. 질적평가 및 근거등급 모형 개발 1안

질적평가 및 근거등급 모형 개발 1안은 GRADE 시스템을 활용하여 평가한 후 이를 근거 교통 신호등(traffic

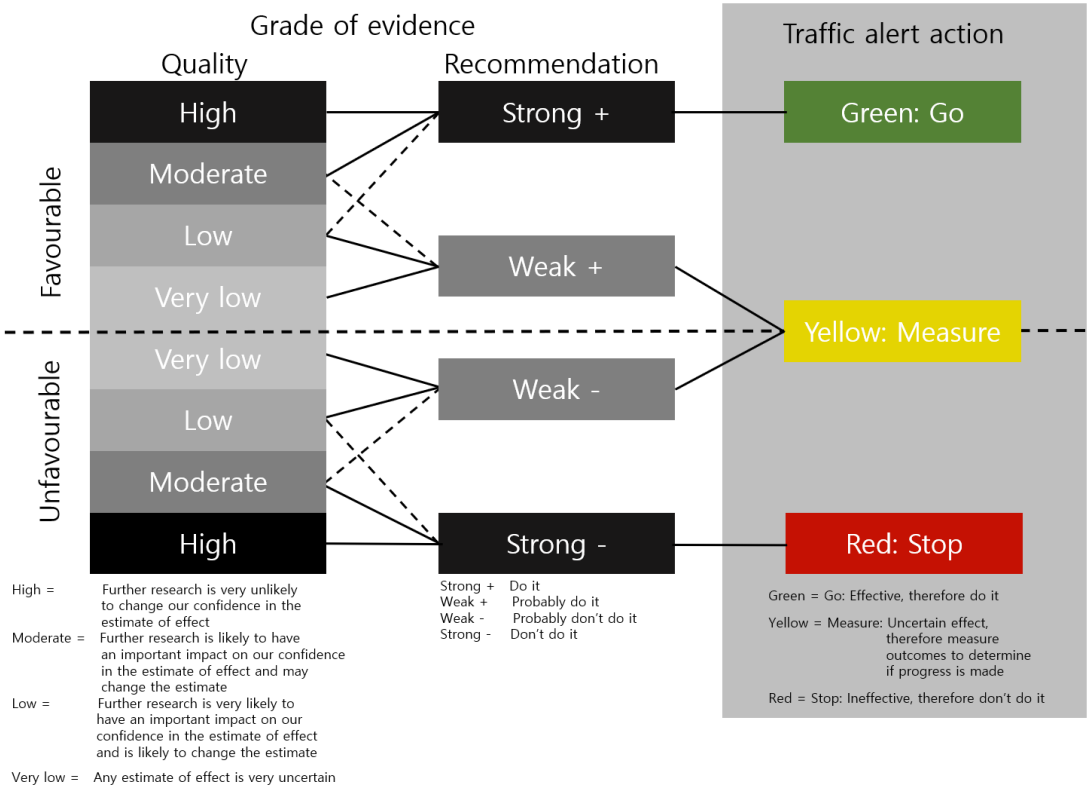

Fig 3. Relationship between the GRADE and traffic light system Adapted from "Novak I, Mcintyre S, Morgan C, et al(2013). A systematic review of interventions for children with cerebral palsy: state of the evidence. Dev Med Child Neurol, 55(10), 885-910. https://doi.org/10.1111/dmcn.12246."

light system)으로 변환하여 이를 통해 치료 적용에 대한 판단을 하도록 한 방법이다(Fig 3)(Novak 등, 2013). 근거 교통 신호등은 근거의 수준 및 강도에 따라 녹색등(go)은 근거의 수준과 질이 높은 치료방법을 의미하며, 노란등(measure)은 치료 효과에 대한 근거 수준과 질이 낮거나 부족한 치료 방법을 의미하고, 빨간등(stop)은 효과가 없는 것으로 치료에 권고되지 않음을 의미한다. 이러한 교통 신호등은 치료 권고에 대해 직관적으로 판단할 수 있도록 해주는 장점이 있다(Table 1).

Table 1. Model draft 1 for a quality assessment and evidence grade

Quality	Recommendation	Traffic alert action
Ⅰ(high)	A(strong+)	A(Green: go) = Ⅰa, Ⅱa, Ⅲa B(Yellow: measure) = Ⅱb, Ⅲb, Ⅳb, Ⅰc, Ⅱc, Ⅲc C(Red: stop) = Ⅱd, Ⅲd, Ⅳd
Ⅱ(moderate)	B(weak+)	
Ⅲ(low)	C(weak-)	
Ⅳ(very low)	D(strong-)	

2. 질적평가 및 근거등급 모형 개발 2안

질적 평가 및 근거등급 모형개발 2안은 기존의 근거 수준에 따른 피라미드에서 각 근거 수준 연구들의 질적 평가도구를 이용하여 평가한다. 평가를 한 이후에 각 근거 수준에 따라 일정 점수를 의미하는 적정 가중치(weight)를 적용하고 이에 따라 순위를 나열하여 SIGN 근거 등급에 따라 매우 강한 근거(very strong), 강한 근거(strong), 적당한 근거(medium), 약한 근거(weak), 매우 약한 근거(very weak), 근거 없음(no evidence)으로 분류하는 방안이다. 이 방안은 근거 수준과 질적 수준을 분석하여 점수화 혹은 등급화 하여 나열하고 이에 따른 치료의 근거 강도를 제시할 수 있다. 앞으로 이 방안을 적용하는데 해결해야 할 사항은 적정 가중치에 대한 사전 연구가 필요하다.

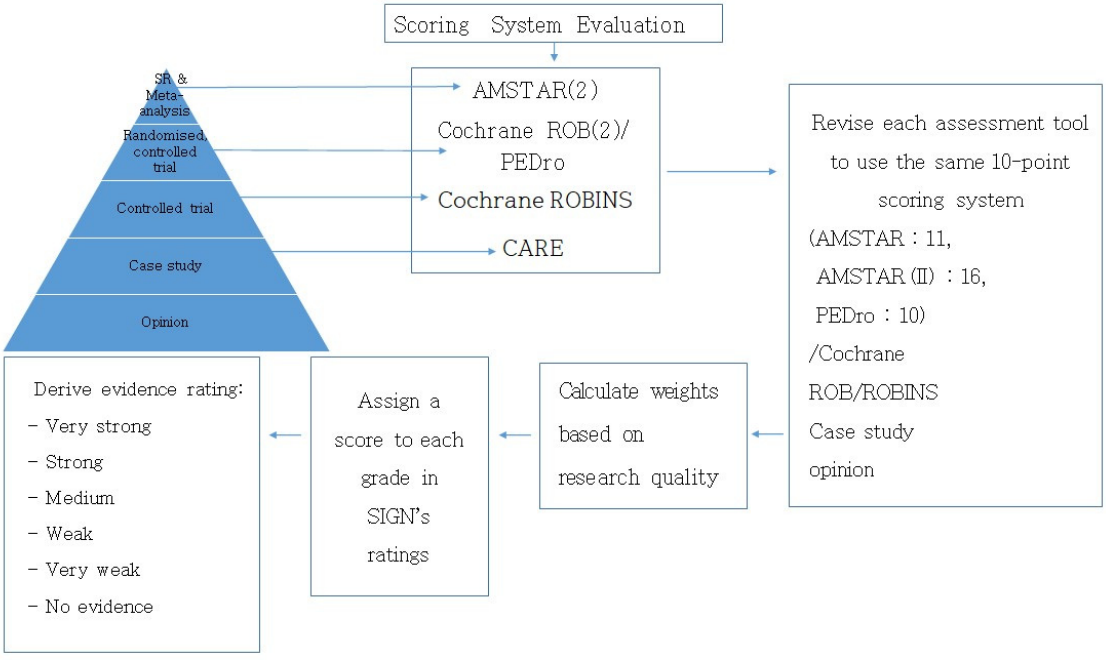

Fig 4. Model draft 2 for a quality assessment and evidence grade

3. 질적평가 및 근거등급 모형 개발 3안

질적평가 및 근거등급 모형개발 3안은 근거 수준 피라미드에서 각 근거 수준에 해당하는 연구들의 질적 평가 도구들을 이용하여 평가한다. 평가를 한 후 1수준은 1++, 1+, 1-로, 2수준은 2++, 2+, 2-로 등급화하며 3과 4수준의 등급으로 분류하고 근거 등급에서 이 중에서 1++는 매우 강한 근거(very strong), 1+는 강한 근거(strong), 1-, 2++는 적당한 근거(medium), 2+, 2-는 약한 근거(weak), 3과 4는 매우 약한 근거(very weak), 그 이하는 근거 없음으로 총 6등급화 한다는 방안이다.

첫째, AMSTAR Ⅱ의 각 항목 평가 시 '1, 0.5, 0'로 평가 후 총점에서 특정 점수들을 기준으로 α 1 이상, α 2를 결정하고 이들 점수에 따라 α 1 이상, α 2 ≤ SR/MA < α 1, α 2 이하로 구분하여 평가한다. 둘째, ROB 2의 최종 평가에 따른 3등급(high, unclear, low risk)을 기준으로 β 1 (low risk) 이상, β 2 (unclear) ≤ SR/MA < β 1 (low risk), β 2 (unclear) 이하로 구분하여 평가한다. 셋째, ROBINS의 최종 평가(5등급)에 따른 3등급(high, unclear, low risk)을 기준으로 θ 1 (low risk) 이상, Q2 (medium) ≤ SR/MA < θ 1 (low risk), θ 2 (medium) 이하로 구분하여 평가한다.

이 방안은 부작용의 위험성이 높은 치료들은 적절하지 않고 부작용의 위험성이 낮은 치료들의 치료 선택의 폭을 넓힐 수 있다는 장점이 있다. 앞으로 이 방안을 적용하기 위해 해결해야 할 사항은 1~2수준의 근거 수준에 따른 질적 평구도구로 평가된 점수 혹은 등급을 등급 내에서 다시 ++, +, -로 등급을 분류하기 위한 점수 또는 기준에 대한 타당성 연구가 필요하다.

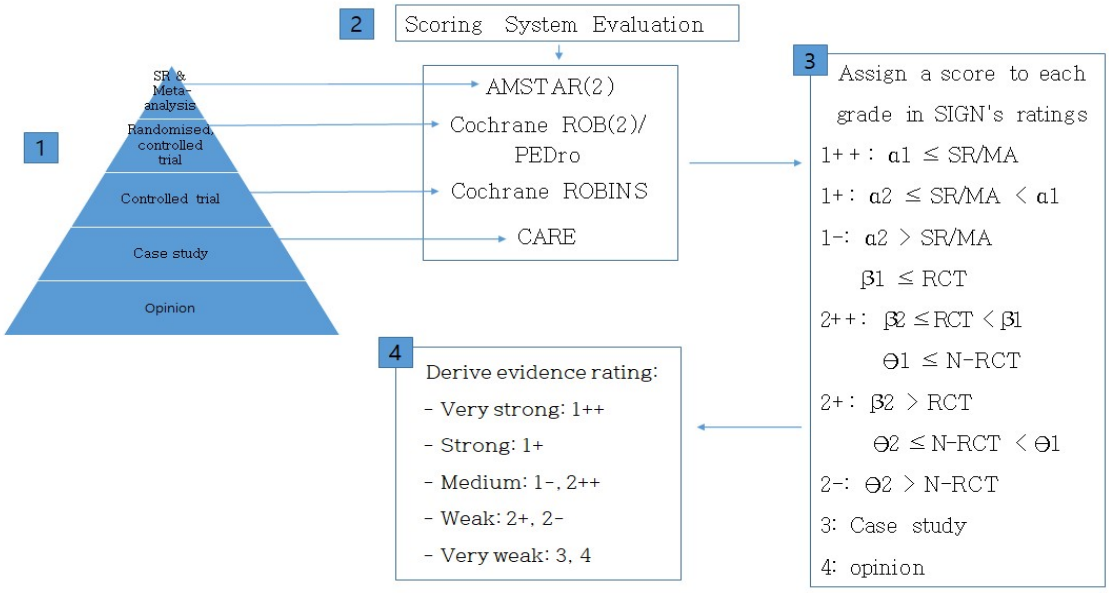

Fig 5. Model draft 3 for a quality assessment and evidence grade α , β , θ : Scores evaluated and graded on three qualitative levels (high, moderate, low) in each assessment tools

Ⅳ. 각 질적 평가 도구들에 대한 평가 시 고려할 점

현행 체계적 문헌고찰 및 메타분석 연구의 질적 평가는 AMSTAR 또는 AMSTAR 2를 사용하여 실시하는데 AMSTAR 2의 평가가 엄격하여 AMSTAR 보다 낮은 평가점수가 나올 수 있다(Shea 등, 2017). 메타분석 연구를 평가할 때 가중치 부여 방안도 고려해야 할 것이다. 가중치 요인으로는 이질성 정도, 분석 논문들의 질적 수준

정도, 조절변수 효과 검증 여부, 출판편의 여부(깔때기 그림, Egger's regression, trim and fill 분석 등 적용여부), 민감성, 누적 메타분석 적용 여부가 고려되어야 할 필요성이 있다(Table 1)(Fig 3).

무작위배정 비교임상시험에 대한 질적 평가는 코크란 ROB2로 하는 방안과 PEDro로 하는 방안을 구분해서 생각해볼 수 있다(de Morton, 2009; Luchini 등, 2021). 물리치료 분야 연구들에서는 PEDro 각 항목에서 GRADE 또는 AMSTRA 2를 참고하여 점수를 1, 0이 아닌 1, 0.5, 0으로 3단계화 방안을 고려해 볼 수 있다(Fig 4, 5).

비무작위 논문에 대한 평가는 국내의 RoBANS 도구와 코크란 ROBINS-I 평가에 대해서 점수화 방안을 향후 연구할 필요성이 있다(Kim 등, 2013; Sterne 등, 2022). Case study 및 전문가 의견에 대한 점수화 방안도 해당 분야의 전문가 집단을 통해 체계적인 질적 평가 도구 안을 마련해야 할 것이다.

근거수준과 권고 등급에 대한 평가 방안에 대해서는 첫 번째, SIGN의 근거수준(8단계)과 권고등급(A~D의 4단계), 두 번째로 GRADE의 권고등급 A~D(high: ++++, moderate: +++, low: ++, very low: +), 세 번째로 SR/MA, RCT/N-RCT 등에 대한 질적 평가 점수를 근거수준 혹은 등급화하기 위한 점수기준 부여 방안을 개발할 필요성이 있다(Fig 4, 5)(Andrews 등, 2013; McGuinness & Higgins, 2021; Sterne 등, 2022).

물리치료 연구에서 같은 종속변인 혹은 평가가 각 연구물마다 결과 변인의 범주가 다르게 분석되는 경우가 많다. 향후 연구를 위한 측정변인들의 범주를 명확히 분류해야 되며, 이에 관련된 가이드라인이 구축되어야 한다고 생각된다. 그리고 물리치료 연구에서는 주요 중재방법 외에 환자의 특성상 연구윤리적인 문제로 인하여 일반적인 물리치료를 함께 실시해야 하기 때문에 실질적인 새로운 중재의 효과검증에 대하여 많은 제한점이 있다.

본 연구를 통하여 각각의 임상에서의 연구 문제와 설계에 맞게 연구가 과학적으로 타당하게 시행되어야 하며 앞으로 중재프로그램에 대한 종속변수에 대한 연구가 보다 체계적으로 이루어져야 물리치료 연구의 근거수준과 권고등급도 향상된다고 생각된다.

Ⅴ. 결 론

물리치료사는 근거기반 실무에 대한 정확하고 타당한 판단을 내리는 것을 우선시해야 한다. 체계적 문헌고찰은 물리치료 연구에서 사용해야 하는 가장 높은 수준의 근거이다. 따라서 적절하고 타당한 연구 설계에 맞는 질적 평가 도구를 선택하고 적용하는 것은 문헌고찰에서 매우 중요한 문제이다. 연구자는 연구의 연구 설계와 목표를 반영하여 적절하고 업데이트된 질 평가 도구를 사용해야 하며, 개입의 복잡성, 적절한 그룹화, 중재에 대한 효과 크기 계산을 반드시 고려해야 한다. 연구자는 질적 평가 내용을 자세히 서술하고, 문헌고찰 저자의 판단을 기술해야 한다. 결과 해석 시 비뚤림 위험의 가능한 영향을 고려하고 적절한 주의를 기울어야 한다.

향후 물리치료 연구에서는 비뚤림 위험 평가를 더 자세히 검토하고 타당하고 적절한 근거 수준 및 등급에 대한 결과를 보고해야 한다. 연구자들은 각 연구 설계와 연구방법에 대한 장점 및 제한점에 대해 자세히 알고 있어야 한다. 논문에서의 결과가 실제 임상 실무에 영향을 미칠 수 있다는 가능성을 감안하여 연구결과를 신중하게 살펴볼 필요성이 있다. 또한, 물리치료 연구 분야에서 대규모의 연구, 연구윤리 상으로 인한 새로운 중재방법의 어려움 등의 요인으로 인해 낮은 질적 수준으로 평가받은 연구들에 대한 세부적이고 새로운 질적 수준을 등급화 할 수 있는 모형 개발과 이에 대한 타당성 관련 연구들이 지속적으로 필요할 것이다.

참고문헌

Ahn E, Kang H(2018). Introduction to systematic review and meta-analysis. Korean J Anesthesiol, 71(2), 103-112. https://doi.org/10.4097/kjae.2018.71.2.103.

Alsalaheen B, Stockdale K, Pechumer D, et al(2017). A comparative meta-analysis of the effects of concussion on a computerized neurocognitive test and self-reported symptoms. J Athl Train, 52(9), 834-846. https://doi.org/10.4085/1062-6050-52.7.05.

Andrews JC, Schünemann HJ, Oxman AD, et al(2013). GRADE guidelines: 15. Going from evidence to recommendation - determinants of a recommendation's direction and strength. J Clin Epidemiol, 66(7), 726-735. https://doi.org/10.1016/j.jclinepi.2013.02.003.

Bafeta A, Trinquart L, Seror R, et al(2013). Analysis of the systematic reviews process in reports of network meta-analyses: methodological systematic review. BMJ, 347(July), 1-12. https://doi.org/10.1136/bmj.f3675.

Bagg MK, Salanti G, McAuley JH(2018). Research note: comparing interventions with network meta-analysis. J Physiother, 64(2), 128-132. https://doi.org/10.1016/j.jphys.2018.02.014.

Balshem H, Helfand M, Schünemann HJ, et al(2011). GRADE guidelines: 3. rating the quality of evidence. J Clin Epidemiol, 64(4), 401-406. https://doi.org/10.1016/J.JCLINEPI.2010.07.015.

de Morton NA(2009). The PEDro scale is a valid measure of the methodological quality of clinical trials: a demographic study. Aust J Physiother, 55(2), 129-133. https://doi.org/10.1016/s0004-9514(09)70043-1.

Deaton A, Cartwright N(2018). Understanding and misunderstanding randomized controlled trials. Soc Sci Med, 210, 2-21. https://doi.org/10.1016/J.SOCSCIMED.2017.12.005.

Frieden TR(2017). Evidence for health decision making - beyond randomized, controlled trials. N Engl J Med, 377(5), 465-475. https://doi.org/10.1056/NEJMRA1614394.

Gibson W, Wand BM, O'Connell NE(2017). Transcutaneous electrical nerve stimulation (TENS) for neuropathic pain in adults. Cochrane Database Syst Rev, 9(9), Printed Online. https://doi.org/10.1002/14651858.CD011976.PUB2.

Guyatt G, Oxman AD, Akl EA, et al(2011). GRADE guidelines: 1. introduction - GRADE evidence profiles and summary of findings tables. J Clin Epidemiol, 64(4), 383-394. https://doi.org/10.1016/j.jclinepi.2010.04.026.

Harbour R, Miller J(2001). A new system for grading recommendations in evidence based guidelines. BMJ, 323(7308), 334-336. https://doi.org/10.1136/bmj.323.7308.334.

Jadad AR, Moore RA, Carroll D, et al(1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary?. Control Clin Trials, 17(1), 1-12. https://doi.org/10.1016/0197-2456(95)00134-4.

Jette DU, Bacon K, Batty C, et al(2003). Evidence-based practice: beliefs, attitudes, knowledge, and behaviors of physical therapists. Phys Ther, 83(9), 786-805. https://doi.org/10.1093/ptj/83.9.786.

Johnson MI, Paley CA, Howe TE, et al(2015). Transcutaneous electrical nerve stimulation for acute pain. Cochrane Database Syst Rev, 2015(6), Printed Online. https://doi.org/10.1002/14651858.CD006142.pub3.

Khadilkar A, Odebiyi DO, Brosseau L, et al(2008). Transcutaneous electrical nerve stimulation (TENS) versus placebo for chronic low-back pain. Cochrane Database Syst Rev, 2008(4), Printed Online. https://doi.org/10.1002/14651858.CD003008.pub3.

Kim SY(2017). Applying the GRADE methodology in evidence-based nursing practice. Korean Soc Evidence-Based Nur, 5(1), 1-3.

Kim SY, Park JE, Lee YJ, et al(2013). Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. J Clin Epidemiol, 66(4), 408-414. https://doi.org/10.1016/J.JCLINEPI.2012.09.016.

Kim SY, Choi MY, Shin SS, et al(2015a). NECA's handbook for clinical practice guideline developer, National Evidence-based Healthcare Collaborating Agency, Seoul, 2015.

Kim SY, Park DA, Seo HJ, et al(2020). Health technology assessment methodology: systematic review. National Evidence-based Healthcare Collaborating Agency, Seoul, 2020.

Kim SY, Park DA, Seo HJ, et al(2021). NECA's guidance for assessing tools of risk of bias, National Evidence-based Healthcare Collaborating Agency, Seoul, 2021.

Kim SY, Park JE, Seo HJ, et al(2015b). NECA's guidance for Systematic reviews manuals and clinical practice guideline manuals Developer. National Evidence-based Healthcare Collaborating Agency, Seoul, 2015.

Kurichi JE, Sonnad SS(2006). Statistical methods in the surgical literature. J Am Coll Surg, 202(3), 476-484. https://doi.org/10.1016/j.jamcollsurg.2005.11.018.

Li HC, Wang HH, Chou FH, et al(2015). The effect of music therapy on cognitive functioning among older adults: a systematic review and meta-analysis. J Am Med Dir Assoc, 16(1), 71-77. https://doi.org/10.1016/j.jamda.2014.10.004.

Lim SM, Shin ES, Lee SH, et al(2011). Tools for assessing quality and risk of bias by levels of evidence. J Korean Med Assoc, 54(4), 419-429. https://doi.org/10.5124/jkma.2011.54.4.419.

Luchini C, Veronese N, Nottegar A, et al(2021). Assessing the quality of studies in meta-research: review/guidelines on the most important quality assessment tools. Pharm Stat, 20(1), 185-195. https://doi.org/10.1002/PST.2068.

Ma LL, Wang YY, Yang ZH, et al(2020). Methodological quality (Risk of Bias) assessment tools for primary and secondary. Mil Med Res, 7(1), 1-11.

Martimbianco ALC, Porfírio GJ, Pacheco RL, et al(2019). Transcutaneous electrical nerve stimulation (TENS) for chronic neck pain. Cochrane Database Syst Rev, 12(12), Printed Online. https://doi.org/10.1002/14651858.CD011927.pub2.

McGuinness LA, Higgins JPT(2021). Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods, 12(1), 55-61. https://doi.org/10.1002/jrsm.1411.

Murad MH, Montori VM, Ioannidis JPA, et al(2014). How to read a systematic review and meta-analysis and apply the results to patient care: Users' guides to the medical literature. JAMA, 312(2), 171-179. https://doi.org/10.1001/JAMA.2014.5559.

Murad, MH, Asi N, Alsawas M, et al(2016). New evidence pyramid. BMJ Evidence-Based Medicine, 21(4), 125-127.

Novak I, Mcintyre S, Morgan C, et al(2013). A systematic review of interventions for children with cerebral palsy: state of the evidence. Dev Med Child Neurol, 55(10), 885-910. https://doi.org/10.1111/dmcn.12246.

Salbach NM, Jaglal SB, Korner-bitensky N, et al(2007). Practitioner and organizational barriers to evidence-based practice of physical therapists for people with stroke. Phys Ther, 87(10), 1284-1303. https://doi.org/10.2522/ptj.20070040.

Shea BJ, Reeves BC, Wells G, et al(2017). AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ, 358, 1-9. https://doi.org/10.1136/bmj.j4008.

Shin IS(2017). Recent research trends in meta-analysis. Asian Nurs Res, 11(2), 79-83. https://doi.org/10.1016/j.anr.2017.05.004.

Sterne JAC, Savović J, Page MJ, et al(2019). RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ, 366, 1-8. https://doi.org/10.1136/ bmj.l4898.

Stroup DF, Berlin JA, Morton SC, et al(2000). Meta-analysis of observational studies in epidemiology: a proposal for reporting. JAMA, 283(15), 2008-2012. https://doi.org/10.1001/jama.283.15.2008.

Zanon MA, Pacheco RL, Latorraca COC, et al(2019). Neurodevelopmental treatment (Bobath) for children with cerebral palsy: a systematic review. J Child Neurol, 34(11), 679-686. https://doi.org/10.1177/0883073819852237.

Higgins JPT, Savović J, Page MJ, et al(2022a). Chapter 8: assessing risk of bias in a randomized trial. In: Higgins JPT, Thomas J, Chandler J, et al(editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022). Cochrane, 2022. Available at https://training.cochrane.org/handbook/current/chapter-08. Accessed April 14, 2023.

Higgins JPT, Thomas J, Chandler J, et al(2022b). Cochrane

handbook for systematic reviews of interventions version 6.3 (updated February 2022). Cochrane, 2022. Available at https://training.cochrane.org/handbook/current. Accessed April 14, 2023.

Schünemann HJ, Higgins JPT, Vist GE, et al(2022). Chapter 14: completing 'Summary of findings' tables and grading the certainty of the evidence. In: Higgins JPT, Thomas J, Chandler J, et al(editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated. Cochrane). Available at https://training.cochrane.org/handbook/current/chapter-14. Accessed April 14, 2023.

Sterne JAC, Hernán MA, McAleenan A, et al(2022). Chapter 25: assessing risk of bias in a non-randomized study. In: Higgins JPT, Thomas J, Chandler J, et al (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022). Cochrane, 2022. Available at https://training.cochrane.org/handbook/current/chapter-25. Accessed April 14, 2023.

Wells GA, Shea B, O'Connell D, et al(2000). The Newcastle-Ottawa scale (NOS) for assessing the quality of non-randomized studies in meta-analysis. Ottawa: Ottawa Hospital Research Institute. Available at https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed April 14, 2023.