Uncertain Certainty: A Scoping Review of GRADE Use in Systematic Reviews on Arthroplasty for Femoral Neck Fractures

Ashleigh Warner, BS[1], Holden Husbands, BS[1], Steven Stagg, BS[1], Nicholas B. Sajjadi, DO[2], Brandon Hull, MD[2]
[1]Office of Medical Student Research, University of Oklahoma College of Medicine, Oklahoma City, OK [2]Department of Orthopedic Surgery & Rehabilitation, OU Health, Oklahoma City, OK
Corresponding email: Ashleigh-warner@ou.edu

Disclosures: All authors declare no disclosures.

INTRODUCTION: Systematic reviews (SRs) and meta-analyses (MAs) are the highest level of evidence and increasingly guide clinical practice, especially in areas of uncertainty such as the surgical management of unstable femoral neck fractures. To help guide clinical decisions in the face of this uncertainty, SRs and MAs commonly compare total hip arthroplasty (THA) and hemiarthroplasty (HA) for the treatment of these fractures to evaluate outcomes including mortality, function, dislocation, and revision risk. The reliability of these conclusions depends on the study quality and rigorous evidence appraisal. The GRADE framework (Grading of Recommendations Assessment, Development and Evaluation) offers a standardized method to assess evidence certainty and recommendation strength, yet its use in orthopedic SRs appears inconsistent. This review examines how frequently GRADE is applied in SRs comparing THA and HA for the treatment of unstable femoral neck fractures and whether journal guidelines recommend or require its use, highlighting gaps and opportunities to improve evidence-based recommendations in orthopedic practice.

METHODS: A comprehensive literature search was conducted following Preferred Reporting Items and Systematic Reviews and Meta-Analyses (PRISMA) guidelines via PubMed and EMBASE for our initial search inquiry. Inclusion criteria were (1) SRs with or without MAs, (2) direct comparison in outcomes between THA and HA, (3) articles written in English, (4) articles in peer-reviewed sources. Two independent reviewers screened all titles and abstracts followed by a full-text screen to determine articles which met the inclusion and exclusion criteria, with a third reviewer settling inconsistencies. Data was then extracted in a blinded, duplicate fashion using Google Forms that were pilot tested to ensure high inter-rater reliability. Data extracted included GRADE inclusion, certainty of evidence, study types included, number of randomized controlled trials included, conclusion/recommendation of the SR, and publication information including year, country, funding, journal, and journal requirements. Descriptive statistics were calculated using R version 4.5.1.

RESULTS SECTION: The initial search resulted in 206 studies: 88 from PubMed, 118 from EMBASE, with 34 duplicates identified. 172 articles were screened by title and abstract for inclusion criteria, resulting in 51 studies. These 51 studies were screened using full text, with 12 reports being excluded due to wrong patient population and 9 reports being excluded due to wrong publication type. The resultant 30 articles were published in 19 distinct journals. Each article underwent data extraction and statistical analysis, as did the SR author submission guidelines for each journal. Only 1 journal was found to recommend the use of GRADE (5%), while none required it (0%) (**Figure 2**). 3 SRs were found to have included GRADE (10%), while 27 SRs did not include a GRADE assessment (90%) (**Figure 3**). In the 3 articles which used GRADE, a GRADE assessment was conducted on a compiled total of 20 articles, with 5 being very low certainty of evidence (25%), 3 being low certainty of evidence (15%), 9 being moderate certainty of evidence (45%), and 3 being high certainty of evidence (15%) (**Figure 1**). The certainty of evidence was downgraded most often for imprecision. The articles which used GRADE had a mean publication year of 2017 with a SD of 4.62, while the articles which did not use GRADE had a mean publication year of 2018 with a SD of 4.83. The 3 articles which included GRADE were each published in different journals, none of the 3 of which required or recommended the use of GRADE in SR author guidelines. The 1 article which was published in the journal recommending GRADE did not use a GRADE assessment (0%).

DISCUSSION: Our review demonstrates that the use of the GRADE framework in SRs comparing THA and HA is rare, with only 10% of reviews incorporating it and none of the publishing journals requiring its use. Even among reviews that applied GRADE, a substantial portion of included studies were rated as low or very low certainty of evidence, highlighting potential overconfidence in conclusions drawn from the literature. The near-universal omission of GRADE underscores a gap between recommended methodological standards and current orthopedic practice, which may limit the reliability of evidence informing high-stakes surgical decisions. The limited number of eligible SRs (n=30) and the very small subset using GRADE (n=3) restricted the scope of analysis and precluded the use of inferential statistics, limiting our ability to identify trends or associations and constraining generalizability across broader orthopedic literature. Despite these limitations, the findings display a clear gap in evidence appraisal within hip arthroplasty reviews. Increasing awareness and adoption of standardized frameworks such as GRADE could enhance reliability and clinical applicability of SR conclusions, ultimately supporting more evidence-based decision-making in orthopedic surgery.

SIGNIFICANCE/CLINICAL RELEVANCE: This review addresses a critical barrier in orthopedic surgery research by highlighting the underuse of the GRADE framework in SRs comparing THA and HA for the treatment of unstable femoral neck fractures, limiting the reliability of evidence informing crucial surgical decisions. By identifying these gaps in evidence appraisal, this study provides a foundation for improving methodological rigor, enhancing the clinical applicability of SR conclusions, and supporting more evidence-based decisions in the surgical management of unstable femoral neck fractures.

IMAGES AND TABLES:

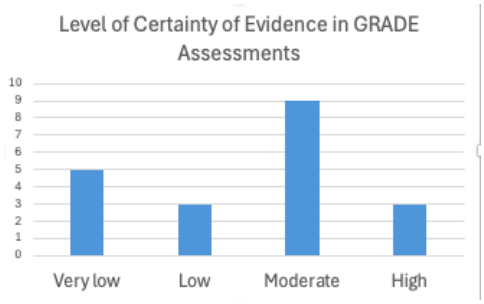

Figure 1. Level of certainty of evidence in GRADE assessments among articles using GRADE

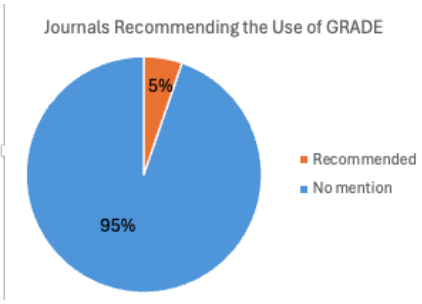

Figure 2. Proportion of publishing journals of articles studied which recommend GRADE

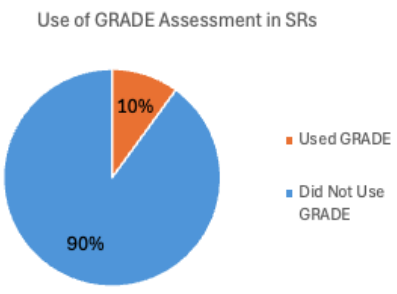

Figure 3. Proportion of SRs studied which completed a GRADE assessment