Name of Journal: *World Journal of Orthopedics*

Manuscript NO: 72695

Manuscript Type: SYSTEMATIC REVIEWS

Fragility of statistically significant findings from randomized clinical trials of surgical treatment of humeral shaft fractures: a systematic review

Humeral shaft fracture fragility index

S. Craig Morris, Anirudh K. Gowd, Avinesh Agarwalla, Wesley P Phipatanakul, Nirav H Amin, Joseph N. Liu
Abstract

BACKGROUND
Despite recent meta-analyses of randomized controlled trials (RCTs), there remains no consensus regarding the preferred surgical treatment for humeral shaft fractures. The fragility index (FI) is an emerging tool used to evaluate the robustness of RCTs by quantifying the number of participants in a study group that would need to switch outcomes in order to reverse the study conclusions.

AIM
Applying the FI to humeral shaft fractures will aid in clinical decision making regarding treatment of these injuries.

METHODS
We completed a systematic review of RCTs evaluating the surgical treatment of humeral shaft fractures. Inclusion criteria included: articles published in English; patients randomized and allotted in 1:1 ratio to 2 parallel arms; and dichotomous outcome variables. The FI was calculated for total complications, each complication individually, and secondary surgeries using the Fisher exact test, as previously published.

RESULTS
Fifteen RCTs were included in the analysis comparing open reduction plate osteosynthesis (ORPO) with dynamic compression plate (DCP) or locking compression plate (LCP), intramedullary nail (IMN), and minimally invasive plate osteosynthesis (MIPO). The median FI was 0 for all parameters analyzed. Regarding individual outcomes, the FI was 0 for 81/91 (89%) of outcomes. The FI exceeded the number lost to follow up in only 2/91 (2%) outcomes.

CONCLUSION
The FI shows that data from RCTs regarding operative treatment of humeral shaft fractures are fragile and does not demonstrate superiority of any particular surgical technique.

**Key Words:** Humerus fracture; open reduction internal fixation; intramedullary nail; fragility index; complications; fragility index


**Core Tip:** Humerus shaft fractures have been managed with intramedullary nail fixation and plate osteosynthesis. Multiple randomized control trials have been performed to compare outcomes, complications, reoperations, and union rates between both treatment modalities. Despite multiple randomized control trials, there remains a lack of consensus from the existing literature regarding surgical treatment of humeral shaft fractures. This manuscript aims to further assess the quality of the literature that guides treatment decisions by employing a new metric, the fragility index.

**INTRODUCTION**

Humeral shaft fractures represent approximately 3% of all long-bone fractures\(^1\) with an incidence around 13 per 100,000 people per year\(^2\). While the vast majority may be managed nonoperatively\(^1-5\), surgical treatment is generally indicated for open fractures, polytrauma patients, ipsilateral humeral shaft and forearm fractures (floating elbow), segmental fractures, and cases of failed treatment in functional brace\(^3\). However, it is important to note that there are currently no defined gold standards for the treatment of humeral shaft fractures\(^6-7\). Surgical treatment options include external fixation, open reduction and plate osteosynthesis (ORPO), minimally invasive plate osteosynthesis (MIPO), and intramedullary nail (IMN). Implant options
for both ORPO and MIPO include dynamic compression plate (DCP) and locking compression plate (LCP). Numerous recent systematic reviews, meta-analysis, and network meta-analysis (NMA) review papers have been published aiming to determine the efficacy of these treatment options in order to provide reliable evidence to guide clinical decision making[6,8-13]. Based on the lack of consensus from the existing literature regarding surgical treatment of humeral shaft fractures, this manuscript aims to further assess the quality of the literature that guides treatment decisions by employing a new metric, the fragility index (FI). The FI has been introduced to further evaluate the robustness (or fragility) of randomized control trial (RCT) results[14,15].

The evaluation of randomized controlled trials (RCTs) via systematic review, meta-analysis, or NMA represents level I evidence; however, the fact remains that many RCTs in orthopaedics, despite demonstrating statistically significant effects, are limited by small sample sizes and few outcome events[16-19]. Clinical studies are classically evaluated for statistical significance in the form of P-values, and 95% confidence intervals, which help determine how likely observed effects would occur based solely on chance[20-22]. The FI represents the required number of participants in the RCT whose outcome would have to change from nonevent to event in order to convert a statistically significant result to nonsignificant. The FI is calculated by sequentially calculating the P-value using the Fisher exact test while changing an outcome from nonevent to event between cycles until the calculated P-value is not significant, or P>0.05. Basically, the FI quantifies how many patients would be required to switch outcomes in order to change the study conclusions. In the case where a study reports a statistically significant result, but the FI is calculated to be zero, this would indicate that the Fisher's exact test did not find a p-value < 0.05 whereas the statistical method used in the paper did. In addition, the FI may be lower than the number of patients lost to follow-up, limiting the confidence one may have in the study conclusion[14]. The higher the FI the more confidence the reader can have that the result is robust. While there is no defined cut off for the FI value, if the FI is zero or less than the number of patients lost to follow up, then any statistically significant result should be considered fragile
and interpreted with caution. By applying the FI metric to RCTs evaluating surgical outcomes in humeral shaft fractures we can determine how much confidence these studies should be given in guiding treatment decisions.

Due to this added value, the FI has been gaining traction in the literature with studies published across numerous medical specialties[15,23,32-35,24-31], in addition to orthopaedic subspecialties[25,36-40]. This valuable, new tool, the FI, can serve to increase our understanding of the literature regarding treatment of humeral shaft fractures, aiding in clinical decision making. Our primary objective was to determine the robustness of statistically significant findings in RCTs of the surgical treatment of humeral shaft fractures by systematically applying the FI. We sought to accomplish this objective by testing our hypothesis that the median FI in these RCTs would be less than the number lost to follow up and therefore would indicate fragile results.

MATERIALS AND METHODS

The systematic review was completed, and results reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines statement[41]. We began by evaluating all review articles about humeral shaft fractures published from 2000 to 2019[6,8-12] and extracting from those studies all included RCTs for analysis. We then performed a systematic review of the literature to identify randomized controlled trials dealing with surgical treatment of humeral shaft fractures that had been published since the most recent review articles. Medline and EMBASE databases were searched for the dates of 01/01/2016 to 04/01/2019 using the following Medical Subject Headings terms: “humeral fractures”. These dates were selected to identify new RCTs that would not have been included in prior systematic review articles. Titles and abstracts were screened, and full text manuscripts reviewed. Inclusion criteria included the following: patients randomized to 2 parallel arms, articles published in English, patient allocated to treatment and control arms in 1:1 ratio, reported statistical significance for dichotomous variables. Exclusion criteria included: published abstract only; studies without available full text, (non-English
manuscripts; studies reporting patient data published previously; retrospective studies; and prospective studies that were not randomized.

Data was extracted from the included studies by individual review of each study by the primary author. Accuracy of data extraction was confirmed by independent review by the remaining authors separately, with any discrepancy resolved by group consensus. An electronic data form was developed and the following data were extracted for each included study: First author, journal, publication year, comparison groups, randomization parameters, initial sample size, total patients lost to follow up, final sample size, patients in study group 1, patients in study group 2, patients lost to follow up in group 1, patients lost to follow up in group 2, presence of power analysis, as well as the number of events for each outcome in each group and reported p-value for dichotomous outcomes (delayed union/nonunion, iatrogenic radial nerve palsy, infection, malunion, shoulder impingement, elbow stiffness, secondary surgeries). For our study lost to follow up included any patients initially enrolled in the study but not included in final analysis for any reason. The total number of events for all complications was defined as the sum of delayed union/nonunion, iatrogenic radial nerve palsy, infection, malunion, shoulder impingement, and elbow stiffness. The total number of events for all complications was calculated for each study group within each included study.

For each study the FI was then calculated for all complications, secondary surgeries, as well as each complication individually. The FI was calculated via method described previously by Walsh et al[14] using a free, publicly available calculator found at http://clincalc.com/Stats/FragilityIndex.aspx. After inputting the total number patients in the control group, experimental group, control group with primary endpoint, and experimental group with primary endpoint, the tool calculates the P-value using the Fisher exact test. If the P-value is significant (<0.05), the tool incrementally converts 1 outcome from nonevent to event and recalculates P-value until the P-value increases above 0.05 and the result becomes insignificant. The
methodological quality of each RCT was also assessed by calculating the Jadad scale\cite{42}, also known as the Oxford Quality Scoring System, for each trial.

**RESULTS**

Our review of RCTs from recent review articles as well as systemic search strategy produced 415 records screened and 28 full text articles assessed (Figure 1). Of these, 15 studies met inclusion criteria (Table I)\cite{43-57}. The primary outcome was only defined in two studies, shoulder function defined by University of California, Los Angeles (UCLA) scoring system in one study\cite{46} and shoulder function defined by the American Shoulder and Elbow Surgeons (ASES) score in the other\cite{56}. Table II contains summary characteristics for these trials. The mean initial sample size was 52.4 (range 30 – 89), mean lost to follow up of 2.7 (range 0 – 9), while the mean final sample size was 49.7 (range 30 – 84). The mean Jadad scale score was 2.5 (range 1 – 3). Power analysis was only reported in 4 studies (26.7%).

The most common comparison was between ORPO with DCP and IMN, found in 8 studies (53.3%). ORPO with LCP vs IMN, MIPO vs IMN, and ORPO with DCP vs IMN were the comparison groups of 2 studies each (13.3% each), and 1 study (6.7%) compared ORPO with DCP and MIPO. All 15 studies evaluated both the outcomes of delayed union/nonunion and iatrogenic radial nerve palsy. The majority of studies also reported incidence for infection (14 studies, 93.3%), secondary surgeries (11 studies, 73.3%), and shoulder impingement (10 studies, 66.7%). Malunion was a reported outcome in 7 studies (46.7%), while only 4 studies (26.7%) reported the outcome of elbow stiffness.

The cumulative FI values for each outcome within each study are listed in Table III and presented graphically (Figure 2). The FI was found to be 0 for all individual outcomes except for iatrogenic nerve palsy in 1 out of 14 studies (higher rate with DCP compared with IMN), malunion in 1 of 7 studies (higher rate in IMN compared with LCP), shoulder impingement in 4 of 10 studies (higher rate in IMN compared with MIPO or DCP), elbow stiffness in 1 of 4 studies (higher rate in DCP compared with
IMN), and secondary surgeries in 1 of 11 studies (higher rate with IMN compared with DCP). When totaling all complications for each study, the FI was >0 in 2 out of the 15 studies, with higher complication rates in IMN compared with MIPO or DCP. Overall, the FI was greater than 0 in only 9.8% (9/91) and was greater than the number lost to follow up in 2% (2/91) of outcomes studied.

The relationship between enrolled initial sample size and FI for all complications (Figure 3) was calculated using the Spearman correlation coefficient and was found to not be significant with a P-value of 0.830.

The majority of included RCTs reported continuous variable outcomes such as operative time, radiation exposure time, operative blood loss, length of hospital stay, time to union, and functional outcome scores such as the UCLA scoring system, Mayo elbow performance index, and the ASES score. The outcomes with reported differences between groups are summarized in Table IV.

DISCUSSION

Our systemic review looked at randomized control trials (RCTs) of the surgical treatment of humeral shaft fractures and discovered that the median FI for all outcomes was 0. In the studies with data leading to FI > 0, the FI exceeded the number lost to follow up in only two instances (2%): lower incidence of iatrogenic radial nerve palsy with IMN compared with ORPO[45], and lower rate of overall total complication with ORPO compared with IMN[56]. Therefore, all evaluated outcomes (nonunion, radial nerve palsy, infections, malunion, malrotation, shoulder impingement, elbow stiffness, secondary surgeries, and overall complications) were extremely fragile and did not demonstrate superiority of one intervention (ORPO, MIPO, IMN) over another.

In analyzing all outcomes individually for humeral shaft fractures, the median FI was 0, and remained so when calculating median FI for all outcomes combined. This result is not surprising given the median FI ≤ 3 reported in the orthopaedic literature previously[36-40]. A recent study used FI to explore the literature on the treatment of clavicular fractures and found the median FI to be 2, with 46.7% of trials reporting the
number of patients lost to follow-up exceeded the FI\textsuperscript{[40]}. Sample sizes in an operative population are inherently lower. In addition, the cost, time, and resources required to complete RCTs with sufficiently large sample sizes often pose a significant challenge in orthopaedics, where the incidence of desired exposures and events can be low\textsuperscript{[14,58]}. Simply increasing sample size alone, however, is not sufficient to guarantee increased FI values, as even very large sample size studies can still have fragile results if the between-group difference is very small\textsuperscript{[14]}.

While the FI was found to be > 0 in 9 outcomes total, the fact that the number lost to follow up exceeded the FI in 89/91 (98\%) instances further confirms that those outcomes are quite fragile, and the significance of those conclusions should be called into question. When the number lost to follow up exceeds the FI this indicates that inclusion of the patients lost to follow up alone could have resulted in a nonsignificant p-value. Kesemenli \textit{et al}\textsuperscript{[45]} reported significantly higher rate of iatrogenic radial nerve palsy among the DCP group compared with the IMN group. Of note, this study reported no patients lost to follow up. While this suggests a robust outcome, the fact remains that the other 14/15 studies showed no difference among treatment groups regarding iatrogenic radial nerve palsy. Regarding all complications combined, two studies\textsuperscript{[53,56]} resulted in $FI > 0$, but the FI exceeded the number lost to follow up in only one\textsuperscript{[56]}.

The difference between treatment options may possibly be captured only by continuous variables, and not by dichotomous variables. There is precedence for this in the orthopaedic literature, as Bhandari \textit{et al}\textsuperscript{[58]} recommended that when orthopaedic surgeons anticipate small sample sizes they can optimize their study’s statistical power by choosing a continuous outcome variable. In reviewing 76 orthopaedic RCTs, these authors found significantly greater study power in RCTs reporting continuous variables compared with studies reporting dichotomous variables ($P = 0.042$), despite similar mean sample size in each group ($p>0.05$). The difference in treatment options for humeral shaft fractures, however, has been reported and analyzed by continuous variables previously. As summarized in this review in Table IV, the majority of
included RCTs reported on continuous variable outcomes. The FI is not designed to evaluate continuous variables, and therefore all these continuous outcomes fell outside the scope of our review. As such, application of the FI does not add to the commentary favoring any one treatment over the others on the basis of these continuous variables.

Our study has potential weaknesses, with some inherent to the requirements of the FI. In order to calculate an FI a study must compare 2 treatment arms, randomize patients to those arms in 1:1 ratio, and report dichotomous outcomes. These inclusion criteria limit both the number of studies that can be included for analysis, as well as the number of outcomes or results that can be analyzed from the included studies. Another requirement of the FI is that a study must be a prospective, randomized trial. Due to this requirement, we excluded 3 retrospective studies and another 9 prospective studies that were not randomized. While this represents a loss in the number of included studies, and associated decrease in number of included patient outcomes, we do not feel this represents a significant loss as it means that the included studies represent the highest level of data availability.

Another potential weakness relates to the FI itself, which is not without inherent weakness or controversy. RCTs with small samples and in which the event of interest is rare, are common in orthopaedics and tend to be inherently fragile. The FI revolves around the statistical threshold of using $p < 0.05$ as a strict criterion of correct inference. While this cutoff is necessary for making statistical determinations, the actual judging of the quality of inference is a complex activity with more nuance than is afforded in having a p-value slightly greater of less than 0.05\textsuperscript{[59]}\textsuperscript{[59]}. The misinterpretation of statistical tests extends beyond just the FI\textsuperscript{[60]}.

**CONCLUSION**

The FI represents a valuable tool that can aid in the interpretation of results from RCTs. Along with P-value and confidence intervals, the FI provides a quantitative metric regarding the robustness of the reportedly significant results. In applying the FI to RCTs comparing surgical treatment options for humeral shaft fractures, this study
has shown that there is a significant lack of robust data to recommend one treatment option over another on the basis of delayed union/nonunion, iatrogenic radial nerve palsy, infection, malunion, shoulder impingement, elbow stiffness, or secondary surgeries. The results published in the literature for treatment of humeral shaft fractures should be interpreted cautiously. This study, while limited in the analysis of functional outcome, suggests no clear benefit of one surgical technique over another with respect to dichotomous outcomes. Plate and nail techniques should both be considered as options for surgical treatment of humeral shaft fractures.

**ARTICLE HIGHLIGHTS**

**Research background**

Humeral shaft fractures are a common injury which could be managed non-operatively or operatively. There is a lack of clear evidence to support open reduction internal fixation as intramedullary nail fixation.

**Research motivation**

Identify the fragility index, which identifies the number of patients have a change in outcome from a significant to non-significant. This is important as higher level studies guide management in orthopedics.

**Research objectives**

Applying the fragility index to humeral shaft fractures will aid in clinical decision making on treatment of humeral shaft fractures.

**Research methods**

A systematic review of RCTs evaluating the surgical treatment of humeral shaft fractures was conducted. The FI was calculated for total complications, each complication individually, and secondary surgeries using the Fisher exact test, as previously published.
Research results
Fifteen RCTs were included in the analysis comparing open reduction plate osteosynthesis (ORPO) with dynamic compression plate (DCP) or locking compression plate (LCP), intramedullary nail (IMN), and minimally invasive plate osteosynthesis (MIPO). The median FI was 0 for all parameters analyzed. Regarding individual outcomes, the FI was 0 for 81/91 (89%) of outcomes. The FI exceeded the number lost to follow up in only 2/91 (2%) outcomes.

Research conclusions
The FI shows that data from RCTs regarding operative treatment of humeral shaft fractures are fragile and does not demonstrate superiority of any particular surgical technique.

Research perspectives
Further research is needed to delineate whether open reduction internal fixation or intramedullary nail fixation is superior in the management of humeral shaft fractures.
<table>
<thead>
<tr>
<th>PRIMARY SOURCES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
</tr>
<tr>
<td><strong>2</strong></td>
</tr>
<tr>
<td><strong>3</strong></td>
</tr>
<tr>
<td><strong>4</strong></td>
</tr>
<tr>
<td><strong>5</strong></td>
</tr>
<tr>
<td><strong>6</strong></td>
</tr>
</tbody>
</table>

Crossref