6-11-2020 9:00 AM

Issues Related To Framing And Interpretation Of Studies In The Orthopaedic Literature

Shgufta Docter, *The University of Western Ontario*

Supervisor: Dianne Bryant, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Health and Rehabilitation Sciences

Abstract

In research, appropriate statistical interpretation and methodology are essential to conduct quality work. To interpret results, p-values are frequently used in isolation, but this is insufficient as treatment effects, confidence intervals (CIs), and clinically important thresholds should also be reported. Further, the equality, superiority, non-inferiority, and equivalence frameworks have critical differences not well delineated in current literature. We conducted a systematic review of studies published in high-impact orthopaedic journals and examined a) how well studies interpreted the results of patient-reported outcome measures, and b) whether a consistent framework was used throughout studies. We found that the majority of studies do not report CIs around between-group differences and do not define a clinically meaningful difference. Half of studies reporting sample size calculations had inconsistency between framing of their research question, sample size calculation, and conclusion. Authors should report results with clinical context and maintain framework consistency to prevent misleading treatment recommendations.

Keywords

Patient-Reported Outcome Measures; Confidence Intervals; P-values; Minimal Clinically Important Thresholds; Superiority; Non-Inferiority; Equivalence; Equality

Summary for Lay Audience

A general understanding of important concepts such as basic statistics and methods are needed to conduct research, however, published research may still contain misinterpreted results. For example, authors rely on the widely used p-value statistic to measure the difference between groups. However, p-values only tell us that two treatment differ but not how large that difference is. The size of the difference is best communicated through providing the treatment effect, confidence intervals (CIs), and a threshold of clinical importance. Clinical importance indicates whether the effect of a treatment is meaningful from a clinician's perspective. Researchers are also interested in knowing whether their findings are applicable to patients, which requires the use of correct study methods, particularly the right framework (i.e. equality, superiority, non-inferiority, and equivalence).

The purpose of our study was to review studies published in top journals in the field of orthopaedic surgery and evaluate whether studies correctly reported their results and whether authors followed a consistent framework throughout their study. We looked at studies published in 2017 and 2019 in five journals that compared two different treatments and assessed patient-reported outcome measures, which are tools used to gain the patient's perspective. We found that the majority of studies relied on a p-value statistic, and only approximately one in five studies reported treatment effect with CIs. We also found that 52.2% of studies switched the framework throughout their study, which led to the wrong sample size being used and too few study participants to make treatment recommendations. Overall, when statistics are misinterpreted and the inappropriate methodology is applied, the study findings can lead clinicians into making misleading treatment recommendations to patients. We encourage journal editors and authors to work on ensuring that the results of their research are interpreted with clinical relevance and the correct framework is used. We believe that this will improve the quality of orthopaedic literature moving forward.

Co-Authorship Statement

The idea for Chapter 2 was developed by Dr. Dianne Bryant. In collaboration with Dr. Bryant, Zina Fathallah, Zi Dong, and Shi-Hsuan Li began the first iteration of the study. They determined inclusion criteria, developed the data extraction form, and conducted the first search. I, Shgufta Docter, acquired the study, refined the research questions, and completed an update of the search. Screening and reviewing studies and extracting data was done by myself and my co-authors, Michael J. Lukacs, Zina Fathalla, Michaela Khan, Morgan Jennings, Zi Dong, Shi-Hsuan Li, and Dr. Bryant. I solely analyzed all data and wrote the original draft of the manuscript. All co-author (Dr. Bryant, Dr. Alan Getgood, Michael J. Lukacs, Zina Fathalla, Michaela Khan, Morgan Jennings, Susan Dong, and Shi-Hsuan Li) reviewed and provided suggestions for the final manuscript. This manuscript was submitted to the Journal of Bone and Joint Surgery on March 27th, 2020.

The idea for Chapter 3 was also developed by Dr. Bryant. My responsibilities were the same as above, as this study uses information collected from the first study to answer a different research question. I wrote the original draft of the manuscript, and all co-author (Dr. Bryant, Dr. Alan Getgood, Michael J. Lukacs, Zina Fathalla, Michaela Khan, Morgan Jennings, Susan Dong, and Shi-Hsuan Li) reviewed and provided suggestions for the final manuscript. This manuscript has not yet been submitted for publication.

I wrote the draft of this thesis and Dr. Bryant provided me with feedback to improve the final submission. An estimated 75% of the work in this thesis was solely completed by me.

Acknowledgements

I have immense gratitude for the people without whom this work would not be possible. Thank you to my supervisor, Dr. Dianne Bryant, for your unwavering support and invaluable mentorship. I am grateful for the education and opportunities for professional growth you have given me. I am lucky to have trained with a leader who values integrity, authenticity, and respect; I have learned an incredible amount at Western on what it means to be a noble scientist that I will always carry with me. Thank you, also, to Dr. Alan Getgood for your mentorship, encouragement, and clinical perspective on our work.

Thank you to Dr. Brent Lanting and Dr. Ryan Degen for collaborating with me on research projects, sharing your subspecialty expertise, and offering advice on what lies ahead; the future is bright.

Thank you to my co-authors for their timely and exceptional work. An extended thank you to Mike and Michaela for your assistance with data extraction and, more so, your friendship.

Thank you to my fellow graduate students; I do not take for granted your guidance, encouragement, and camaraderie. A sincere thank you to Holly for a life-long friendship filled with laughs, support, and constancy.

Thank you to my family and friends for your unquestionable support and patience. To my sisters, brother, and dad; thank you for listening to me talk about high impact journals. You are, quite simply, the best.

Finally, thank you to my mom, Bilkis. Thank you for building a life for me free from the struggles you endured, and a character in me that moves others. I miss you every day and owe this all to you. The lessons you imparted of honesty and accountability have guided me through this very work. I look forward to honouring you every day in the work that I do and the person that I am.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

1 Introduction

1.1 Introduction: Background and Rationale

Scholarly experts and journal editors assess manuscript submissions to peer-reviewed journals in order to determine the quality of research and whether the manuscript is suitable for publication. Despite this rigorous review process, manuscripts with methodological and statistical issues are often accepted for publication[1–3]. Improper analyses or misinterpretation of results may lead to erroneous conclusions, which can negatively impact patient care if treatments that do not provide benefit are accepted into practice or a treatment is discarded early due to negative findings[4]. The peer-review process should identify minor and major issues with the submission, but the process itself is inconsistent. Sprowson et al. (2013) state that in orthopaedics, hundreds of reviewers are recruited and therefore formal training is difficult to achieve, and most reviewers learn to review by practice[5]. Some journals may have statistical experts to assist with the review process, but not all have the resources to include a reviewer with statistical expertise on each submission. Providing clinicians with resources related to appropriate research methodology and statistics can assist them when reviewing manuscripts and conducting their own research.

Common mistakes authors make when analyzing or interpreting study results in orthopaedic journals include making comparisons of p-values, missing measures of precision and estimate of effects, and failing to differentiate between statistical significance and clinical importance[6]. To determine clinical importance, patient-reported outcome measures (PROMs) are widely used in orthopedics[7], however, the interpretation of PROMs is challenging because it is difficult to assign meaning to differences between groups in units of a measured health outcome. PROMs are often reported using p-values, however, the threshold p-value<0.05 is arbitrary and does not represent clinical importance[8]. Rather, studies should be reporting results beyond p-values and should include treatment effect(s) and measures of precision (i.e. confidence intervals (CIs)) to

provide an interpretation of results that are more clinically relevant[9]. In addition, the treatment effect and CIs should be interpreted in light of a clinically important threshold to provide clinicians with context; this can be the minimal clinically important difference (MCID), expected difference, superiority margin, or non-inferiority margin[10–12]. Correct methodology and a clinically relevant interpretation of results is necessary for authors seeking to make treatment recommendations.

Other common methodological errors of studies published in orthopedic journals involve concepts such as the failure to follow the principles of trial design and the lack of justification for power analysis[6]. An important component of trial design is the identification of a framework that is consistent with the trial's objective to test the equality, superiority, non-inferiority, or equivalence between interventions[13]. An equality framework is explanatory and is generally used to test the safety and feasibility of a treatment prior to its implementation in a large scale trial[14]. This framework relies on p-values to determine if a study should be pursued further, requires fewer patients, and uses surrogate or lab-based measures to evaluate patient outcomes[15,16]. On the other hand, the superiority, non-inferiority, and equivalence frameworks relies on CIs and thresholds for clinical importance when interpreting results, making it possible to make clinical recommendations; these *a priori* thresholds are, importantly, used to calculate sample size[15]. As a result, these frameworks require more patients and investigate patient important outcomes[10,17]. Defining a framework and using the appropriate methodology is essential to prevent authors from overstating clinical conclusions based on underpowered studies. In current clinical health research, the quality of reporting of these frameworks is poor[10,18].

The issues identified in the literature reflect our own anecdotal experiences with the manuscript submission and peer review process. As part of my coursework, I took my supervisor's (Dr. Dianne Bryant) course "Advanced Quantitative Research Methods"; this taught me the ways that authors may misinterpret data and how to determine the correct study methodology based on the research question being asked. Further, my experience conducting systematic reviews and using quality rating tools such as Grading of Recommendations, Assessment, Development, and Evaluations (GRADE)[19] and

Cochrane's Risk of Bias tool (ROB)[20] trained me to critically appraise study findings and determine evidence quality. After expressing my interest in data misinterpretation and inconsistent study methods, my supervisor presented me with a project that aimed to evaluate both.

The past 20 years in orthopaedic research have given rise to landmark papers whose findings led to changes in clinical practice; it is clear that the results of research are important for the progress of medicine[21,22]. The quality of orthopaedic research has improved over time, but there is a gap in the current literature on reviews that analyze the interpretability of the results of PROMs and analyze the use of all four frameworks in published studies. To address this gap, we sought to publish two systematic reviews that seek to answer two questions, 1) Were studies reporting and interpreting the results of their PROMs appropriately; and 2) Were authors following a consistent methodological framework throughout their study?

1.2 Thesis Outline

This introduction is followed by three chapters (Chapters 2-4). Chapter 2 is a systematic review assessing clinical studies published in five high impact orthopaedic journals to evaluate the reporting and interpretation of the results of PROMs by determining the proportion that, (1) only report a p-value, (2) report a treatment effect, CI, or MCID, and (3) offer an interpretation of the results beyond interpreting a p-value. Chapter 3 is a systematic review that evaluates the same studies from Chapter 2 to answer whether (1) studies follow a consistent framework between their research question, sample size calculation, and conclusion, and (2) studies should have been framed differently based on the compared interventions. Chapter 4 comprises the general conclusion and future directions.

1.3 References

1. Lee S, Kang H. Statistical and methodological considerations for reporting RCTs in medical literature. *Korean J Anesthesiol*. 2015;68(2):106-115. doi:10.4097/kjae.2015.68.2.106

2. Jia P, Tang L, Yu J, et al. Risk of bias and methodological issues in randomised controlled trials of acupuncture for knee osteoarthritis: a cross-sectional study. *BMJ Open*. 2018;8(3):e019847. doi:10.1136/bmjopen-2017-019847

3. Yi D, Ma D, Li G, et al. Statistical Use in Clinical Studies: Is There Evidence of a Methodological Shift? Waldorp LJ, ed. *PLoS ONE*. 2015;10(10):e0140159. doi:10.1371/journal.pone.0140159

4. Ioannidis JPA. Why Most Published Research Findings Are False. *PLoS Med*. 2005;2(8):e124. doi:10.1371/journal.pmed.0020124

5. Sprowson AP, Rankin KS, McNamara I, Costa ML, Rangan A. Improving the peer review process in orthopaedic journals. *Bone & Joint Research*. 2013;2(11):245-247. doi:10.1302/2046-3758.211.2000224

6. Petrie A. Statistics in orthopaedic papers. *The Journal of Bone and Joint Surgery British volume*. 2006;88-B(9):1121-1136. doi:10.1302/0301-620X.88B9.17896

7. Christensen DL, Dickens JF, Freedman B, et al. Patient-Reported Outcomes in Orthopaedics: *The Journal of Bone and Joint Surgery*. 2018;100(5):436-442. doi:10.2106/JBJS.17.00608

8. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337-350. doi:10.1007/s10654-016-0149-3

9. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 2. Interpreting study results: confidence intervals. *CMAJ*. 1995;152(2):169-173.

10. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJW, CONSORT Group for the. Reporting of Noninferiority and Equivalence Randomized Trials: An Extension of the CONSORT Statement. *JAMA*. 2006;295(10):1152. doi:10.1001/jama.295.10.1152

11. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. *Controlled Clinical Trials*. 1989;10(4):407-415. doi:10.1016/0197-2456(89)90005-6

12. Golish SR, Groff MW, Araghi A, Inzana JA. Superiority Claims for Spinal Devices: A Systematic Review of Randomized Controlled Trials. *Global Spine Journal*. Published online June 7, 2019:219256821984104. doi:10.1177/2192568219841046

13. Chan A-W, Tetzlaff JM, Gotzsche PC, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ*. 2013;346(jan08 15):e7586-e7586. doi:10.1136/bmj.e7586

14. Selby P, Brosky G, Oh PI, Raymond V, Ranger S. How pragmatic or explanatory is the randomized, controlled trial? The application and enhancement of the PRECIS tool to the evaluation of a smoking cessation trial. *BMC Med Res Methodol*. 2012;12:101. doi:10.1186/1471-2288-12-101

15. Chow S-C, Shao J, Wang H. *Sample Size Calculations in Clinical Research*. Marcel Dekker; 2003. Accessed April 2, 2019. http://www.crcnetbase.com/isbn/9780824709709

16. Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ*. 2015;350(may08 1):h2147-h2147. doi:10.1136/bmj.h2147

17. Shafiq N, Malhotra S. Superiority trials: raising the bar of null hypothesis statistical testing. *Evid Based Med*. 2015;20(5):154-155. doi:10.1136/ebmed-2015-110280

18. Wangge G, Klungel OH, Roes KCB, de Boer A, Hoes AW, Knol MJ. Interpretation and Inference in Noninferiority Randomized Controlled Trials in Drug Research. *Clin Pharmacol Ther*. 2010;88(3):420-423. doi:10.1038/clpt.2010.134

19. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*. 2011;64(4):383-394. doi:10.1016/j.jclinepi.2010.04.026

20. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. Published online August 28, 2019:l4898. doi:10.1136/bmj.l4898

21. Amin NH, Hussain W, Ryan J, Morrison S, Miniaci A, Jones MH. Changes Within Clinical Practice After a Randomized Controlled Trial of Knee Arthroscopy for Osteoarthritis. *Orthop J Sports Med*. 2017;5(4):2325967117698439. doi:10.1177/2325967117698439

22. Akhter S, Mundi R, Bhandari M. The Impact of Evidence in Surgery of the Musculoskeletal System. *World J Surg*. 2020;44(4):1020-1025. doi:10.1007/s00268-019-04955-7

Chapter 2

2 Interpreting Patient-Reported Outcome Measures in Orthopaedic Surgery: A Systematic Review

2.1 Introduction

The Consolidated Standards of Reporting Trials (CONSORT) Statement includes a 25-item checklist recommending that studies report results beyond p-values. The statement proposes that treatment effect(s) and measures of precision (i.e. 95% confidence intervals (CI)) be included to facilitate the interpretation of results[1]. However, a 2013 review of medical and surgical literature found less than 40% of included studies reported treatment effects with CIs[2]. This suggests that authors, peer-reviewers, and journal editors may not appreciate the extreme limitations of p-values to interpret findings[2]. A p-value in isolation only describes whether the outcomes of two or more treatments differ statistically, but does not sufficiently describe the magnitude of, or certainty around, the estimate of the effect[3]. In addition, a study that reports a statistically significant difference should carry much less influence over clinical decision-making than a study that reports a clinically important difference. Specifically, the effects of treatment can be statistically significant but not clinically important; or the effects of treatment can be not statistically significant, which may mean that the treatment is ineffective, that the study lacks precision, or that there has been a random sampling error[4] (Figure 2-1).

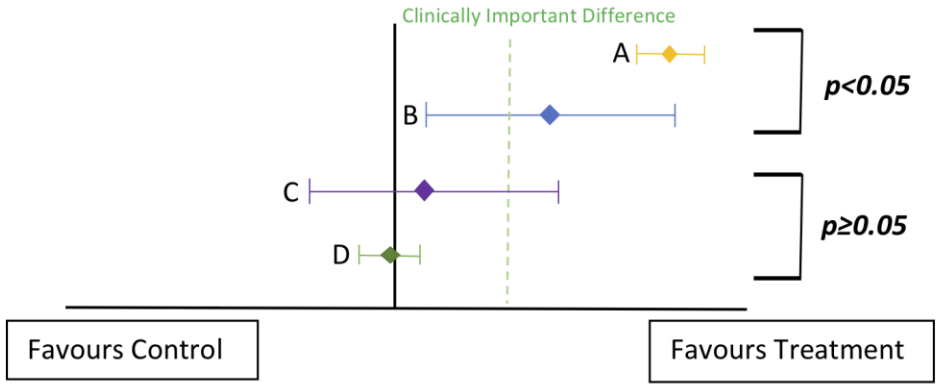

Figure 2-1. A, B, C, and D represent four examples of study treatment effects with associated 95% confidence intervals (CI). Results from study A and D provide a consistent message that clinicians should feel confident acting upon (due to narrow CIs that fall completely to the right (A) or completely to the left (D) of a clinically important threshold). The results from study B and C do not provide a consistent message (due to wide CIs which include both the possibility that between-group difference surpass a clinically important threshold and that it does not).

A treatment effect is a measure of the magnitude of the difference between groups and may be expressed as a mean difference, odds ratio, relative risk, Cohen's *d* effect size, risk difference, median, or mean change[5]. Further, the associated CI provides valuable information regarding the variability of the data and the precision of the effect. A 95% (the conventionally used confidence level) CI represents the range of values where the true value of a parameter lies 95% of the time and provides a degree of confidence for the interval of the estimate[3,6]. Narrow CIs indicate more precise results due to a large sample size (continuous outcome), large number of events (dichotomous outcome), or low variability between groups[7]. CIs should be interpreted with respect to a threshold that defines a clinically important difference to provide clinicians with meaningful context. Common threshold include the minimal clinically important difference (MCID) (the threshold representing the smallest meaningful benefit/value[8]), a superiority margin (a pre-determined value used to declare that one treatment is better than another) or a non-inferiority margin (a pre-determined value used to declare that one treatment is not worse than another)[10]. For example, Smekal et al. randomized patients with a displaced

midshaft clavicular fracture to receive elastic stable intramedullary nailing (ESIN) (n=30) or non-operative treatment (n=30)[9]. They found statistically significant ($p<0.05$) differences on the Disability of the Shoulder and Arm (DASH) scores in favour of ESIN, and conclude that ESIN should be an alternative to non-operative treatment[9]. However, they did not report the difference between groups (treatment effect), CIs, or any means to interpret the CIs, such as a MCID, making it difficult to interpret whether this difference is likely to be meaningful to patients or clinicians.

Patient-reported outcome measures (PROMs) are frequently used in orthopaedics to quantify a patient's perspective of their quality of life, function, and pain. However, the interpretability of the results of a study reporting PROMs is challenging because it is difficult to assign meaning to differences between groups in units of quality of life. To date, no studies have evaluated the quality of reporting and interpretation of the results of PROMs in orthopaedic literature. The objective of this systematic review was to assess clinical studies from five high impact orthopaedic journals to evaluate the reporting and interpretation of the results of PROMs by determining the proportion that, (1) only report a p-value, (2) report a treatment effect, CI, or MCID, and (3) offer an interpretation of the results beyond interpreting a p-value.

2.2 Methods

This systematic review was conducted in accordance with the Cochrane Handbook of Systematic Reviews and the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines[13,14].

Literature Search and Eligibility Criteria. We selected five orthopaedic journals with high impact factors[15] including: The American Journal of Sports Medicine (AJSM), Journal of Bone and Joint Surgery American Edition (JBJS), Arthroscopy: The Journal of Arthroscopic and Related Surgery, Osteoarthritis and Cartilage, and The Journal of Arthroplasty (JOA). We systematically searched the electronic database MEDLINE to identify eligible clinical studies published in 2017. We later updated this study to include eligible clinical studies published in 2019.

Study Selection. We imported all references to Covidence (www.covidence.org). Four pairs of reviewers independently screened titles and abstracts to exclude irrelevant studies and then reviewed full text. We included clinical studies that compared at least two intervention groups and evaluated at least one PROM. For any disagreement, a third reviewer was consulted. We evaluated the agreement of eligibility criteria between pairs of reviewers for both titles and abstracts screening and full text review using a kappa co-efficient (k). Agreement was interpreted as follows: almost perfect agreement (k= 0.81-1.00), substantial agreement (k=0.61-0.80), moderate agreement (k=0.41-0.6), and fair agreement (k=0.21-0.40)[16].

Statistical Analyses. We used proportions to report our findings and a Fisher's chi-square test to compare results between studies published in 2017 and 2019. We used $p<0.05$ to declare statistical significance and set no margin of importance for proportions. SPSS software (version 25, IBM) was used for all statistical analyses.

Data Collection and Outcomes of Interest. Eight reviewers independently extracted data using a standardized web-based data extraction form (Empower Health Research Inc., http://www.empowerhealthresearch.ca/). Reviewers collected the following citation information: study title, author, journal name, volume, page number, and study design. Reviewers extracted the following values: p-values, estimates of the treatment effect (mean difference, mean change, odds ratio, Cohen's *d* effect size, relative risk, median, risk difference), standard deviation or standard error or CIs, MCID, and whether a threshold was used to interpret the importance of results (MCID, Cohen's *d* effect size, or superiority/non-inferiority margin).

2.3 Results

Our search yielded 2363 studies. The full text of 334 studies were reviewed and 228 studies were included for analysis (Figure 2-2), including: randomized control trials (RCT) (n=126), prospective cohorts (n=35), retrospective cohorts (n=61), mixed cohorts (n=1), and case controls (n=5). Reviewers demonstrated substantial agreement for screening (k=0.71) and almost perfect agreement for full text review (k=0.83).

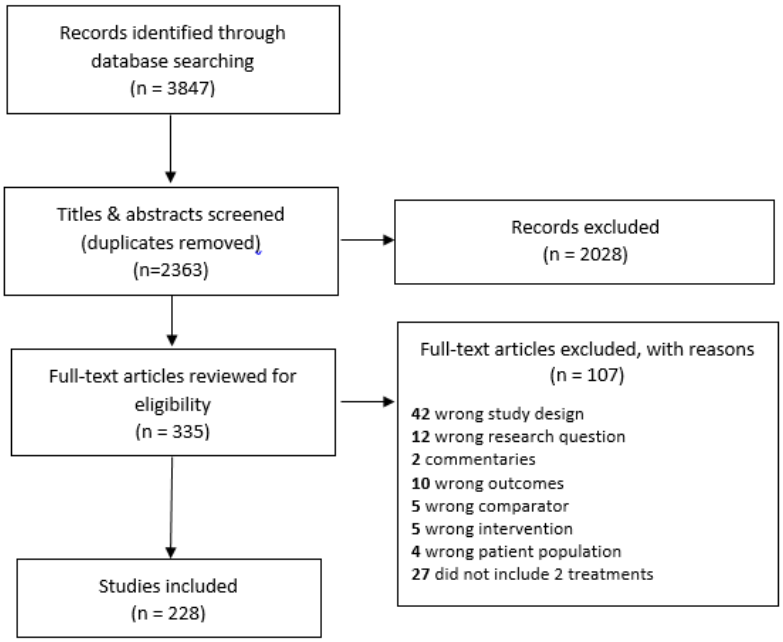

Figure 2-2. Preferred Reporting Items for Systematic Reviews and Meta-Analyses Flow Diagram.

Overall, 99.9% (227 of 228) of included studies presented a p-value when reporting the results of PROMs. Of these, 31.3% (71 of 227) reported a significant p-value (<0.05) and 68.7% (156 of 227) reported a non-significant p-value (>0.05). Overall, 76.3% (174 of 228) used p-values exclusively to evaluate between group differences; 86 of 126 RCTs reported p-values exclusively. Of the 54 (of 228) studies reporting a treatment effect, over half (32 of 54) interpreted their results using an MCID (24 of 54), Cohen's *d* effect size (5 of 54), or a non-inferiority margin (2 of 54) (Table 2-1). Only 22.4% (51 of 228) reported a treatment effect with associated 95% CIs and of these, three studies interpreted CIs in the context of an MCID or non-inferiority margin (Table 2-1). Analysis of all studies (n=228) revealed 35.5% (81 of 228) reported an MCID and 91.3% (74 of 81) were a within-group MCID and 8.6% (7 of 81) were a between-group MCID.

Table 2-1: Treatment Effect Reporting and Interpretation of Included Studies (n=228)

Table 1. Treatment Effect Reporting and Interpretation of Included Studies (n=228)	
	Frequency, n (%)
Between-Group Treatment Effect Reported	54 of 228 (23.7)
Between-Group Treatment Effect Reported (n=54)	
Mean Difference	46 of 54 (85.2)
Mean Change	5 of 54 (9.3)
Odds Ratio	2 of 54 (3.7)
Cohen's *d* effect size†	1 of 54 (1.9)
Authors used a threshold to interpret the importance of results	
MCID	24 of 54 (44.4)
Cohen's *d* effect size††	5 of 54 (9.3)
Non-inferiority margin	2 of 54 (3.7)
Between-Group Treatment Effect with Confidence Intervals Reported	51 of 228 (22.4)
Confidence Intervals Interpreted	3 of 51 (5.9)

MCID: Minimal Clinically Important Difference.

†: Cohen's *d* effect size is a standardized effect size equal to the mean difference divided by the pooled standard deviation.

†† These 6 studies reported the treatment effect of their patient reported outcome measure as mean difference but interpreted findings in light of a Cohen's d effect size, interpreting the findings based on a small (0.2), medium (0.5), or large (0.8) treatment effect.

Variables reported in the results were found to be not significantly different ($p \geq 0.05$) between studies published in 2017 and 2019 with the exception of the reporting of an MCID. We found a mean difference in proportions of 35.5% (95% CI: 20.9, 48.4, $p<0.001$) indicating the true improvement in the reporting of between-group differences using an MCID likely falls between 21% and 48%.

2.4 Discussion

Our findings reveal that the majority of comparative clinical studies published in five high impact factor orthopaedic journals in the years 2017 and 2019 use only p-values to report the results of PROMs. Only approximately one in five of studies reported treatment effects with 95% CIs and slightly more than half interpreted their findings in light of a clinically important threshold. The reporting of MCIDs was low and the majority were within-group MCIDs, which is inappropriate for a between-group comparison. Since evidence-based practice requires clinicians to remain up-to-date with scientific literature, it is important that we move away from the use of arbitrary p-value thresholds and towards reporting values that provide clinically relevant information[3,17].

Conclusions in clinical studies are frequently based on a p-value in isolation, using the threshold of 0.05 to determine whether one treatment is more effective than another[18]. However, the significance level of $p<0.05$ is arbitrary and does not represent clinical importance[18]. The p-value is influenced by sampling error, sample size, and variability: the smaller a sample, the less likely you are to achieve statistical significance that is reproducible[19]. The more variable a population, the less likely you are to achieve statistical significance even when treatments truly offer different outcomes to patients, unless the sample is quite large[19]. Conversely, the larger and more homogenous a sample, the more likely you are to achieve reproducible statistical significance[19] (Figure 2-1). Given the highly influential nature of the study sample in achieving statistical significance, an assessment of the likelihood of random sampling error, the precision of the results, and readiness for uptake into practice should be a requirement of all studies evaluating the effects of two or more interventions.

Many of the evaluated studies reported a non-significant p-value. In most cases, these results were misinterpreted as “no difference between treatments”. However, a $p \geq 0.05$ indicates that the null hypothesis is consistent with observed results, but it does not prove that there is no difference between treatments[20]. There are many reasons for a $p \geq 0.05$, including lack of power, imprecise or invalid measurement, poor study design (type two error), and erroneous statistical analyses[21]. Because these reasons are rarely given appropriate consideration, it is likely that potentially beneficial treatments have been discarded based on non-statistically significant findings[22,23].

Abdullah et al. conducted a systematic review of orthopaedic literature from 2012 to 2013 and found a number of RCTs (21.5%) concluded “no difference between study groups” but were underpowered to do so[23]. This problem is not specific to orthopaedics. In an analysis of five journals of various disciplines, Amrhein et al. report that 51% of published articles incorrectly interpreted statistically non-significant results as having “no effect”[24]. The authors call for the abandonment of p-values and have collected over 800 signatories from scientists in over 50 countries[24].

Conversely, Bhandari et al. surveyed orthopaedic clinicians and found that when p-values were statistically significant, clinicians perceived the study results to be more important[25]. However, considering that there are at least three possible explanations for statistically significant p-values (i.e. there is a difference in outcomes, there is a random sampling error, or there is insufficient power to be certain) and at least three possible explanations for results that are not statistically significant (i.e. there is no difference in outcomes, there is a random sampling error, or there is insufficient power to be certain), it is essential that we report results that communicate the magnitude of the treatment effect, precision of results, and ultimately, clinical importance[3,26,27].

We assessed the proportion of studies that reported results using more than just a p-value and found that only 22.4% of studies reported the magnitude of treatment effects and associated 95% CIs. Similarly, Vavken et al. conducted a systematic review evaluating orthopaedic journals from 2000 to 2006 and found only 22% reported CIs[28]. Despite over a decade of research, disappointingly, our reporting methods remain unchanged. When

comparing reporting methods to other surgical specialties, a 2015 systematic review by Karadaghy et al. of otolaryngology literature found that 54% of studies reported a treatment effect but only 27% reported associated CIs and even fewer (8%) interpreted the CIs for the reader[29]. In our review, 57.4% of the studies that reported a treatment effect interpreted their results with respect to a threshold (MCID, Cohen's *d* effect size, or non-inferiority margin), but only three studies interpreted findings using CIs. As researchers, we must remember that the effect size represents the average effect for that specific sample only; that there is a distribution of possible effect sizes where the true effect size is most likely found within one standard deviation of this average value and that the 95% CI describes possible values that are two standard deviations on either side of the average. Researchers should evaluate both the upper and lower bounds of the CI with respect to a clinically important threshold to ensure that the certainty around results is clearly articulated[30]. For example, readers should consider if the study, 1) has ruled out the probability of a clinically important difference (the study is not statistically different and the clinically important threshold falls outside of the study CIs), 2) found a high probability that differences are clinically important (the study is statistically significant and the threshold falls outside (is smaller than) of the study CIs), or 3) is underpowered to make definitive conclusions (the threshold falls within the study CIs). This will allow clinicians to make more informed, accurate and robust decisions regarding patient care.

We found that only 35.5% of studies mentioned an MCID which is a small proportion for a value that is essential to relay clinical importance when reporting the results of a PROM. Copay et al. evaluated MCID reporting trends in orthopaedic journals from 2014 to 2016, and found that only 129 of 1709 articles used or referenced an MCID, where the majority (86.1%-90.4%) used previously published MCIDs[31,32]. A challenge with using an MCID to interpret study results is that the majority of MCIDs are determined using a within-group study design; here, participants are asked to comment on whether they have experienced small, meaningful change following an intervention and the average change between a pre- to post- intervention score, in patients who claim to have experienced a small but important change, is proclaimed the MCID. In terms of similarities between within-group versus between-group studies, a pre- to post- intervention study (within-group) might be considered similar to an unblinded no treatment versus active treatment

comparator, in terms of measuring similar amounts of change[33]. Unfortunately, measured change between two active comparators or an active comparator versus a blinded placebo will be much smaller since the control group in both scenarios will also experience and report change[34]. A 1993 study by Goldsmith et al. reported that, on average, the value of the between-group MCID was 20-40% of the within-group MCID[35] (Figure 2-3). For example, Warby et al. randomized patients with multidirectional instability (MDI) of the shoulder to the Watson MDI (n=18) or Rockwood (n=23) program[11]. Groups were compared using the Melbourne Instability Shoulder Score (MISS) at 24 months. The authors reported a mean between-groups difference of 15.4 MISS points (95%CI 5.9 to 24.8) but did not interpret the results for the readers. Specifically, since the 95% CIs excluded a between-group MCID of 2.0 (approximately 40% of the within-group MCID of 5.0), the authors could have concluded with certainty that patients who undergo the Watson program will experience superior results to those who undergo a Rockwood program. Assuming no selection bias, exclusion of the between-group MCID from the 95%CI means that the findings are precise.

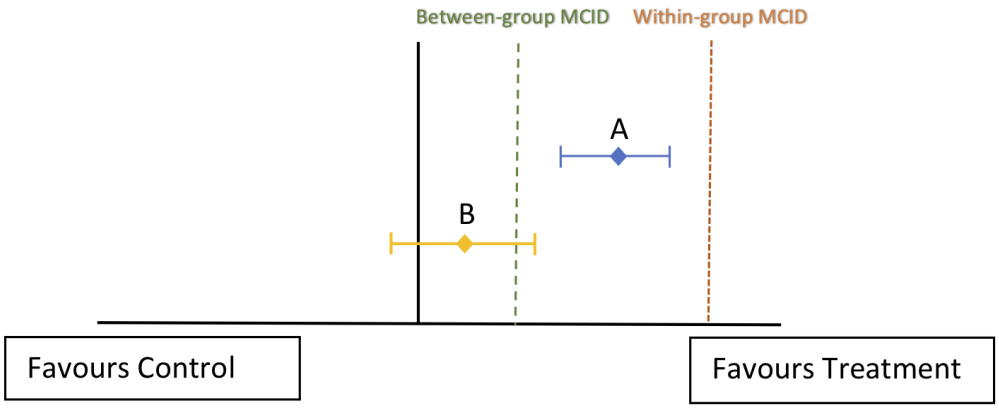

Figure 2-3. Study A is statistically significant and is clinically important according to the between-group minimal clinically important difference (MCID) (conclusive) but is not clinically important according to a within-group MCID (conclusive). The clinical interpretation is conclusive but opposite depending which MCID is used. The results of study B are not statistically different and not clinically important based on a within-group MCID (conclusive) but there is still a possibility

that the difference in outcome between treatments is clinically important if using a between-group MCID (inconclusive), since the upper boundary of the 95% confidence interval includes the between-group threshold.

Being able to detect a smaller difference between treatments requires a larger sample size (i.e. greater power). The denominator of the equation used to estimate sample size is defined by the value of the expected difference between the two treatment groups (squared). Given that the most common value that researchers use to represent the expected difference is the within-group MCID, it is no wonder why the majority of orthopaedic trials are underpowered. It is simply unreasonable to expect that the difference between two active treatment groups would be as large as the difference from pre- to post- intervention. This also means that study results interpreted using a within-group MCID, where the expected between-groups differences are much larger than is reasonable, may have falsely concluded that the outcomes were definitively not different between the two groups (E.g. Study A in Figure 2-3). For example, Kvalvaag et al. randomized patients with subacromial shoulder pain to receive supervised exercises with radial extracorporeal shock wave therapy (rESWT) (n=69) or sham rESWT (n=74)[12]. They reported a between-group difference of 0.7 (95%CI -6.9 to 8.3) on the Shoulder Pain and Disability Index (SPADI) at 24 months[12]. Given that an important between-group MCID likely falls around 4 points (40% of the 10 point within-group MCID), if they had set a superiority margin of 2 points (given that an intervention can still be useful even if not all patients will experience an important change), they would be unable to definitively conclude that rESWT is not more effective than sham because the CIs still include 2 points. If, on the other hand, the authors had used the within-group MCID of 10 points (as is commonly done in error) they would have erroneously concluded that we can be certain that rESWT is not better than sham.

To improve the quality of reporting in the field of orthopaedics, instructions to authors may need to be improved and the vetting process be more comprehensive. Specifically, we found there was inconsistency between the instructions for authors of journals included in this review with respect to whether the reporting of treatment effects, CIs, and

MCID were required. Further, limited availability of statistical support in some orthopaedic groups may help explain our findings.

Our study is not without limitations. We evaluated high impact orthopaedic journals based on the annual Journal Citation Report which is the ratio between citations and recent citable items published which does not necessarily reflect quality[36]. Further, we only evaluated studies in five journals and thus were also unable to capture the complete breadth of the orthopaedic literature. The generalizability of our findings to other fields is limited since the issues of significance in clinical trials are not the same as in basic sciences, where consistency of statistically significant results between different samples is emphasized (i.e. clinical importance is not relevant).

2.5 Conclusion

The majority of interventional studies reporting PROMs do not report CIs around between-group differences in outcome and do not define a clinically meaningful difference. A p-value, which cannot effectively communicate the clinical meaning of the results, is insufficient and may be misleading. Reporting requirements should be expanded to require authors to define and provide a rationale for between-group clinically important difference thresholds and the study findings should be communicated by comparing the CIs to these thresholds.

2.6 References

1. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol*. 2010;63(8):834-840. doi:10.1016/j.jclinepi.2010.02.005

2. Nagendran M, Harding D, Teo W, et al. Poor adherence of randomised trials in surgery to CONSORT guidelines for non-pharmacological treatments (NPT): a cross-sectional study. *BMJ Open*. 2013;3(12):e003898. doi:10.1136/bmjopen-2013-003898

3. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J Clin Res Ed*. 1986;292(6522):746-750.

4. Harris JD, Brand JC, Cote MP, Faucett SC, Dhawan A. Research Pearls: The Significance of Statistics and Perils of Pooling. Part 1: Clinical Versus Statistical Significance. *Arthrosc J Arthrosc Relat Surg*. 2017;33(6):1102-1112. doi:10.1016/j.arthro.2017.01.053

5. Sullivan GM, Feinn R. Using Effect Size-or Why the P Value Is Not Enough. *J Grad Med Educ*. 2012;4(3):279-282. doi:10.4300/JGME-D-12-00156.1

6. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 2. Interpreting study results: confidence intervals. *CMAJ Can Med Assoc J J Assoc Medicale Can*. 1995;152(2):169-173.

7. Kamper SJ. Confidence Intervals: Linking Evidence to Practice. *J Orthop Sports Phys Ther*. 2019;49(10):763-764. doi:10.2519/jospt.2019.0706

8. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. *Control Clin Trials*. 1989;10(4):407-415. doi:10.1016/0197-2456(89)90005-6

9. Smekal V, Irenberger A, Struve P, Wambacher M, Krappinger D, Kralinger FS. Elastic Stable Intramedullary Nailing Versus Nonoperative Treatment of Displaced

Midshaft Clavicular Fractures-A Randomized, Controlled, Clinical Trial: *J Orthop Trauma*. 2009;23(2):106-112. doi:10.1097/BOT.0b013e318190cf88

10. Wang B, Wang H, Tu XM, Feng C. Comparisons of Superiority, Non-inferiority, and Equivalence Trials. *Shanghai Arch Psychiatry*. 2017;29(6):385-388. doi:10.11919/j.issn.1002-0829.217163

11. Warby SA, Ford JJ, Hahne AJ, et al. Comparison of 2 Exercise Rehabilitation Programs for Multidirectional Instability of the Glenohumeral Joint: A Randomized Controlled Trial. *Am J Sports Med*. 2018;46(1):87-97. doi:10.1177/0363546517734508

12. Kvalvaag E, Brox JI, Engebretsen KB, et al. Effectiveness of Radial Extracorporeal Shock Wave Therapy (rESWT) When Combined With Supervised Exercises in Patients With Subacromial Shoulder Pain: A Double-Masked, Randomized, Sham-Controlled Trial. *Am J Sports Med*. 2017;45(11):2547-2554. doi:10.1177/0363546517707505

13. J. P T Higgins, S Green. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0.*; 2011. www.handbook.cochrane.org. Accessed October 30, 2018.

14. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Int J Surg*. 2010;8(5):336-341. doi:10.1016/j.ijsu.2010.02.007

15. InCites Journal Citation Report - Web of Science Group. 2017. https://jcr.clarivate.com/JCRJournalHomeAction.action?

16. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Medica*. 2012;22(3):276-282.

17. Page P. Beyond statistical significance: clinical interpretation of rehabilitation research literature. *Int J Sports Phys Ther*. 2014;9(5):726-736.

18. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond " $p < 0.05$." *Am Stat*. 2019;73(sup1):1-19. doi:10.1080/00031305.2019.1583913

19. Thiese MS, Ronna B, Ott U. P value interpretations and considerations. *J Thorac Dis*. 2016;8(9):E928-E931. doi:10.21037/jtd.2016.08.16

20. Goodman S. A Dirty Dozen: Twelve P-Value Misconceptions. *Semin Hematol*. 2008;45(3):135-140. doi:10.1053/j.seminhematol.2008.04.003

21. Nahm FS. What the P values really tell us. *Korean J Pain*. 2017;30(4):241-242. doi:10.3344/kjp.2017.30.4.241

22. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311(7003):485.

23. Abdullah L, Davis DE, Fabricant PD, Baldwin K, Namdari S. Is There Truly "No Significant Difference"?: Underpowered Randomized Controlled Trials in the Orthopaedic Literature. *J Bone Jt Surg-Am Vol*. 2015;97(24):2068-2073. doi:10.2106/JBJS.O.00012

24. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;567(7748):305-307. doi:10.1038/d41586-019-00857-9

25. Bhandari M, Bhandari M, Montori VM, et al. The undue influence of significant p-values on the perceived importance of study results. *Acta Orthop*. 2005;76(3):291-295. doi:10.1080/00016470510030724

26. Wasserstein RL, Lazar NA. The ASA's Statement on *p* -Values: Context, Process, and Purpose. *Am Stat*. 2016;70(2):129-133. doi:10.1080/00031305.2016.1154108

27. Gagnier JJ, Morgenstern H. Misconceptions, Misuses, and Misinterpretations of P Values and Significance Testing: *J Bone Jt Surg*. 2017;99(18):1598-1603. doi:10.2106/JBJS.16.01314

28. Vavken P, Heinrich KM, Koppelhuber C, Rois S, Dorotka R. The use of confidence intervals in reporting orthopaedic research findings. *Clin Orthop*. 2009;467(12):3334-3339. doi:10.1007/s11999-009-0817-7

29. Karadaghy OA, Hong H, Scott-Wittenborn N, et al. Reporting of Effect Size and Confidence Intervals in *JAMA Otolaryngology–Head & Neck Surgery*. *JAMA Otolaryngol Neck Surg*. 2017;143(11):1075. doi:10.1001/jamaoto.2017.1504

30. Finch S, Cumming G. Putting Research in Context: Understanding Confidence Intervals from One or More Studies. *J Pediatr Psychol*. 2009;34(9):903-916. doi:10.1093/jpepsy/jsn118

31. Copay AG, Chung AS, Eyberg B, Olmscheid N, Chutkan N, Spangehl MJ. Minimum Clinically Important Difference: Current Trends in the Orthopaedic Literature, Part I: Upper Extremity: A Systematic Review. *JBJS Rev*. 2018;6(9):e1. doi:10.2106/JBJS.RVW.17.00159

32. Copay AG, Eyberg B, Chung AS, Zurcher KS, Chutkan N, Spangehl MJ. Minimum Clinically Important Difference: Current Trends in the Orthopaedic Literature, Part II: Lower Extremity: A Systematic Review. *JBJS Rev*. 2018;6(9):e2. doi:10.2106/JBJS.RVW.17.00160

33. Kamper SJ. Interpreting Outcomes 1—Change and Difference: Linking Evidence to Practice. *J Orthop Sports Phys Ther*. 2019;49(5):357-358. doi:10.2519/jospt.2019.0703

34. Musahl V, Karlsson J, Hirschmann MT, et al. *Basic Methods Handbook for Clinical Orthopaedic Research: A Practical Guide and Case Based Research Approach*.; 2019. https://doi.org/10.1007/978-3-662-58254-1. Accessed February 19, 2020.

35. Goldsmith CH, Boers M, Bombardier C, Tugwell P. Criteria for clinically important changes in outcomes: development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. OMERACT Committee. *J Rheumatol*. 1993;20(3):561-565.

36. Eugene Garfield. *The Clarivate Analytics Impact Factor*. https://clarivate.com/webofsciencegroup/essays/impact-factor/.

Chapter 3

3 Inconsistencies in Methodological Framework Throughout Published Studies in Top Orthopaedic Journals: A Systematic Review

3.1 Introduction

The Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) Statement recommends that authors provide both the type and framework of the trial in the study protocol[1]. The framework of a study refers to its overall objective to test the equality, superiority, non-inferiority, or equivalence of one intervention against another[1]. Depending on the framework selected, there are important differences in design, analysis, sample size estimation, and interpretation[2].

An equality framework uses a two-sided statistical test to determine the probability that the observed differences in outcomes between a treatment and a control group are due to chance[3]. This is called a null hypothesis testing framework[3,4]. Although we can, and should, produce confidence intervals (CIs) around the estimate of the treatment effect (i.e., the difference between groups), the CIs are not used to indicate anything more than the precision of the estimate. Generally speaking, if the lower and the upper boundary of the CI fall on opposite sides of no between-group difference (mean difference (MD) or risk difference (RD) = 0, or relative risk (RR) or odds ratio (OR) = 1), then the statistical test will produce a probability value greater than 5% ($p>0.05$) or not statistically different[4]. Under this framework, studies with non-statistically significant results with very imprecise CIs reach the same conclusion as studies with precise CIs, that the observed difference between groups is not greater than that which might occur by chance. The same is true of studies that reaches statistical significance; CIs could be imprecise, range from a very small effect to a very large effect in favour of the new treatment, or could be precise and include a smaller range of plausible effect sizes[5].

Given that interpretation of the range of plausible effect sizes is not part of the equality framework, it is appropriate for feasibility or proof of concept studies that seek to

demonstrate that the intervention *can* affect change, or efficacy studies that use surrogate outcomes (i.e., proxy measures of patient important outcomes), where the intention is to demonstrate that change in the surrogate is possible and provide evidence to support a more pragmatic next study (or not). The recommendation should not include making changes to clinical care.

Conversely, a superiority framework uses a one-sided statistical test to declare whether one treatment is better than another and makes reference to a pre-determined superiority margin to make this declaration[3,6]. Studies that should use a superiority framework generally involve introducing a new intervention to replace an existing intervention. Here, less may be understood about the adverse event profile for the new intervention and there may be costs associated with bringing the intervention into routine practice. As such, clinicians should insist on certainty around these conclusions before adopting the new treatment into practice.

A superiority framework is also appropriate when we are adding resources to an existing intervention (e.g. providing the intervention more frequently or for a longer duration, requiring additional equipment or time to perform a procedure, etc.). The value assigned as the superiority margin is informed by the cost of the new intervention relative to old, including any costs associated with retraining the clinician, replacement or retooling of equipment, and the ability of the existing intervention to achieve desired rates and standards of outcomes. For studies that include a patient-reported outcome measure (PROM), the demonstrated minimally clinically important difference (MCID) for that outcome may also factor into decisions around the magnitude of the superiority margin[7]. Under a superiority framework, the superiority of one intervention over another is declared if the CIs rule out the possibility that the true between-group difference is unimportant.

Next, a non-inferiority framework uses a one-sided test to declare whether a treatment is "no worse" than its control[3]. This framework involves defining a non-inferiority margin, which is the maximum difference between treatments that one is willing to accept before declaring one of the treatments inferior to the other (i.e., causing unacceptably worse

outcomes for its recipients)[8]. A non-inferiority framework is appropriate for studies evaluating whether an existing intervention, or parts of that intervention, can be removed such that any negative affect on outcomes is within acceptable limits. The reason for removing an intervention is likely wrapped up in reducing resource use or adding efficiencies around existing protocols. Therefore, the value that defines the non-inferiority margin is informed by the likelihood and severity of worse outcomes, and the cost to the individual and the health care system associated with suffering a worse outcome.

Finally, an equivalence framework seeks to assess whether two interventions are interchangeable, offering equivalent outcomes[3]. This framework requires defining both a superiority and non-inferiority margin. To be justified in declaring that two interventions are equivalent, the CIs around the between-group difference would be completely contained with both margins[3].

Systematic reviews of medical literature evaluating the reporting and interpretation of superiority, non-inferiority, and equivalence trials have identified deficiencies in design and inconsistencies in methodology[9,10]. To address this issue, an extension of the Consolidated Standards of Reporting Trials (CONSORT) recommendations was developed to outline the differences between frameworks and improve the reporting of non-inferiority and equivalence trials[11]. With numerous studies trying to demonstrate superiority, non-inferiority, and equivalence of various orthopaedic interventions and make clinical recommendations, an investigation into the reporting and interpretation of these frameworks is warranted. Thus, the objective of this systematic review was to assess studies published in five of the top orthopaedic journals and evaluate the proportion of studies that (1) demonstrated consistency between the framing of their research question, sample size calculation, and conclusion, and (2) should have framed their research question differently based on the compared interventions.

3.2 Methods

This systematic review was conducted in accordance with the Cochrane Handbook of Systematic Reviews and the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines[12,13].

Literature Search and Eligibility Criteria. We selected five orthopaedic journals with high impact factors[14], including: The American Journal of Sports Medicine (AJSM), Journal of Bone and Joint Surgery American Edition (JBJS), Arthroscopy: The Journal of Arthroscopic and Related Surgery, Osteoarthritis and Cartilage, and The Journal of Arthroplasty (JOA). We systematically searched the electronic database MEDLINE to identify eligible clinical studies published in 2017. We later updated this study to include eligible clinical studies published in 2019.

Study Selection. We imported all references to Covidence (www.covidence.org). Four pairs of reviewers independently screened titles and abstracts in stage one to exclude irrelevant studies and reviewed full-text studies in stage two. We included clinical studies published in 2017 and 2019 that compared at least two interventions and evaluated at least one PROM. The reviewers discussed any disagreements and consulted a third reviewer when necessary until consensus was reached. We evaluated the agreement of study eligibility between pairs of reviewers for both titles and abstracts screening and full text review using a kappa co-efficient (k). Agreement was interpreted as follows: almost perfect agreement (k= 0.81-1.00), substantial agreement (k=0.61-0.80), moderate agreement (k=0.41-0.6), and fair agreement (k=0.21-0.40)[15].

Statistical Analyses. We used frequencies and percentages to report all categorical variables. SPSS software (version 25, IBM) was used for all statistical analyses.

Data Collection and Outcomes of Interest. Four reviewer pairs independently extracted data using a standardized web-based data extraction form (Empower Health Research Inc., http://www.empowerhealthresearch.ca/). Reviewers collected the following citation information: study title, author, journal name, volume, page number, study design, and sample size. For each study, reviewers assessed the framework of the research question,

sample size calculation, and conclusion. Each section was classified as either: equality, superiority, non-inferiority, or equivalence. Justification for each classification was noted. Finally, reviewers assessed studies that were framed as equality and determined whether they should have used a different framework based on the interventions being compared. Information regarding how studies were reframed is detailed in Appendix A.

3.3 Results

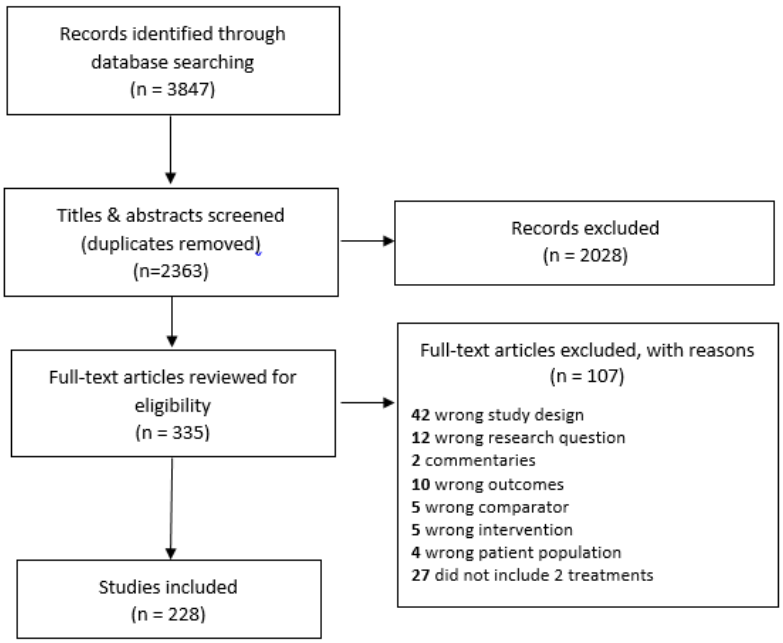

Figure 3-1. Preferred Reporting Items in Systematic Reviews and Meta-Analyses Flow Diagram.

Our search yielded 2363 studies **(**Figure 3-1**)**. The full texts of 335 studies were reviewed, and 228 studies were included for analysis **(**Table 3-1). Agreement at the titles and abstracts stage was substantial (k=0.71) and at the full text screening stage was almost perfect (k=0.83).

Table 3-1. Type of Included Study (n=228)

Type of Study	Frequency, n (%)
Randomized Control Trial	126 (55.3)

Prospective Cohort	35 (15.4)
Retrospective Cohort	61 (26.8)
Mixed Cohort	1 (0.4)
Case Control	5 (2.2)
Note: A mixed cohort is when one group has been followed prospectively and compared to a group collected retrospectively.	

Of studies that reported a sample size calculation (60.5%, n=138), 52.2% (n=72) demonstrated inconsistency between the framing of the research question, sample size calculation, and conclusion. Of the 137 studies that reported an equality sample size calculation, only 56.2% (n=77) were consistent with this approach in framing their concluding statements; the remaining 43.8% (n=60) studies incorrectly concluded superiority (n=49), non-inferiority (n=3), and equivalence (n=8).

Of studies that did not report a sample size calculation (39.5%, n=90), 42.2% (n=38) were inconsistent between the framing of the research question and the conclusion. Overall, 81.6% (n=186) of studies framed their research question as equality (Table 3-2). Based on the interventions being compared, we determine that 129 studies should have been framed as superiority, 52 as equivalence, and three as non-inferiority. Only two studies correctly framed their research question as equality.

Table 3-2. Inconsistency within published studies regarding the alignment of the research question, sample size calculation, and conclusion.

	Research question (n=228)	**Sample size calculation* (n=138)**	**Conclusion (n=228)**
	Frequency, n	Frequency, n	Frequency, n
Equality	186	137	128
Superiority	39	0	79

Non-Inferiority	2	1	8
Equivalence	1	0	13
*Sample size calculation was not provided in 90 studies.			

3.4 Discussion

We found that approximately half (52.2%) of the studies published in five high impact orthopaedic journals demonstrated inconsistency between the framing of their research question, sample size calculation, and conclusion. The majority (81.6%) of studies framed their research question as equality instead of superiority, non-inferiority, or equivalence, based on the interventions that were being compared. Of the studies that reported a sample size calculation, nearly all used an equality calculation despite almost half concluding that one intervention was superior, non-inferior, or equivalent to another. This pattern of inconsistency is problematic, as authors may be misinterpreting research findings and making unsubstantiated clinical recommendations based on statistical results.

The decision as to the appropriateness of each framework can be informed by the location of the study on the continuum of trial designs (Figure 3-2). The Pragmatic Explanatory Continuum Indicator Summary (PRECIS) tool states that explanatory trials aim to evaluate efficacy and safety in a highly controlled setting, whereas pragmatic trials aim to test effectiveness and apply findings to clinical practice[16]. Pilot studies are a type of explanatory trial used during the planning phase for large, expensive, pragmatic trials. The appropriate framework for pilot studies is equality, since they aim to assess the feasibility of the protocol, evaluate eligibility criteria, and examine safety[17]. In our study, only two articles were appropriately framed as equality since they intended only to evaluate efficacy using lab-based or surrogate outcomes[17]. Researchers who intend to make clinical recommendations based on their findings should be conducting a pragmatic study and use a superiority, non-inferiority, or equivalence framework[18].

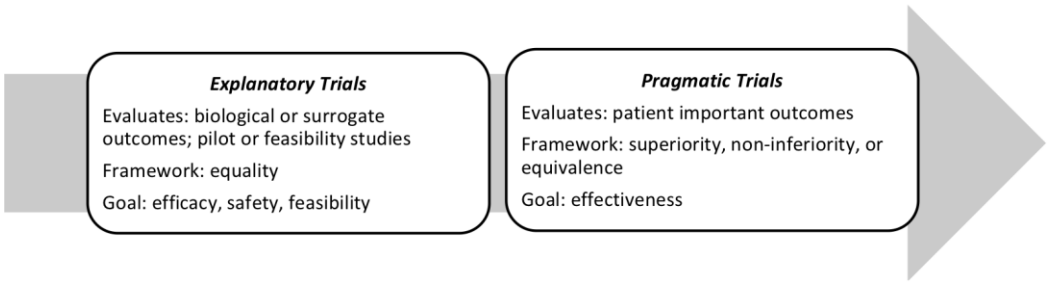

Figure 3-2. The explanatory-pragmatic continuum of trial design.

It could be argued that changing the framework as a study progresses is a form of outcome reporting bias, a bias that arises when the dissemination of research findings are influenced by the nature and direction of results[19]. Greene et al. evaluated clinical studies from 1992 to 1996 and found that 67% declared equivalence following a failed superiority test[20]. Paesmans et al. found 11 out of 23 non-inferiority oncology studies did not communicate their initial trial design and conclusion using the same framework[10]. Finally, Shafiq and Mahlotra argue that failed non-inferiority trials claiming superiority may be engaging in research misconduct or statistical trickery as no pre-specified superiority margin was identified[21].

As researchers, we are trying to estimate a parameter. For studies evaluating the effect of an intervention, the parameter we are trying to estimate is the size of the difference in treatment effect between two or more groups. A CI represents a range of values so defined that there is a specified probability that the value of a parameter lies within it (e.g., 99%, 95%, 80%, etc.)[22]. A 95% CI is an estimate of *plausible* values for the population parameter[23]. As researchers, we infer that the results observed in our sample apply to the population, but intuitively we know that the smaller the sample, the less likely it is to represent the population (i.e., random sampling error or sampling bias)[24]. This is why it is so important that researchers responsibly recognize the range of effect sizes that remains plausible for the population, and this requires more than the reliance on the p-value.

Our study shows that there is still some uncertainty about how to frame a research study and interpret the results correctly. For example, 33% of the studies in our review concluded that one treatment was superior to the other. However, no studies used a superiority margin in their sample size calculation or interpreted their CIs against a superiority threshold. Only one study consistently and properly used a non-inferiority framework. To declare superiority or non-inferiority, authors must stipulate a margin or threshold that delineates the magnitude of the between-groups difference required to declare superiority, use this margin when calculating sample size, and relate the findings to the margin when interpreting the results (Table 3-3, Figure 3-3).

Table 3-3. Characteristics of the four methodological frameworks.

Framework	Research Question/Aim	Calculation to Estimate Sample Size*	Concluding Statement	Key Elements
Equality	"The aim of our study was to statistically compare… "	Only includes an expected difference Denominator: δ^2	"Treatment A produced outcomes that are not statistically different from Treatment B."	No margin
Superiority	"The aim of our study was to determine if treatment A offers better outcomes than treatment B…"	Includes an expected difference and superiority margin Denominator: $(\delta - M)^2$	"Treatment A is better than Treatment B."	Superiority margin must be determined a priori
Non-Inferiority	"The aim of our study was to determine if treatment A is no worse than treatment B…"	Includes an expected difference and non-inferiority margin Denominator: $(\delta - M)^2$	"Treatment A is no worse than Treatment B."	Non-inferiority margin must be determined a priori
Equivalence	"The aim of our study was to determine if treatment A is equivalent to treatment B…"	Includes an expected difference, non-inferiority margin and superiority margin Denominator: $(M - \lvert\delta\rvert)^2$	"Treatment A is interchangeable with, comparable to, or equal to Treatment B."	Both superiority and non-inferiority margin must be determined a priori

*The numerator in all sample size calculations remains the same: n/group=$2(Z+ Z_{\beta})^2\sigma^2$ for outcomes that use a continuous scale and $(Z+ Z_{\beta})^2((po(1-po)) + (p1(1-p1)))$ for outcomes that are dichotomous. Note that Z is $Z_{\alpha/2}$ is 1.96 for a two-sided test where the Type 1 error rate is 5% (equality and equivalence) and Z is $Z_{\alpha/2}$ is 1.64 for a one-sided test where the Type 1 error rate is 5% (superiority and non-inferiority). Note that for dichotomous outcomes, the sample size calculation is different from continuous but similarly, only the denominator changes across frameworks.

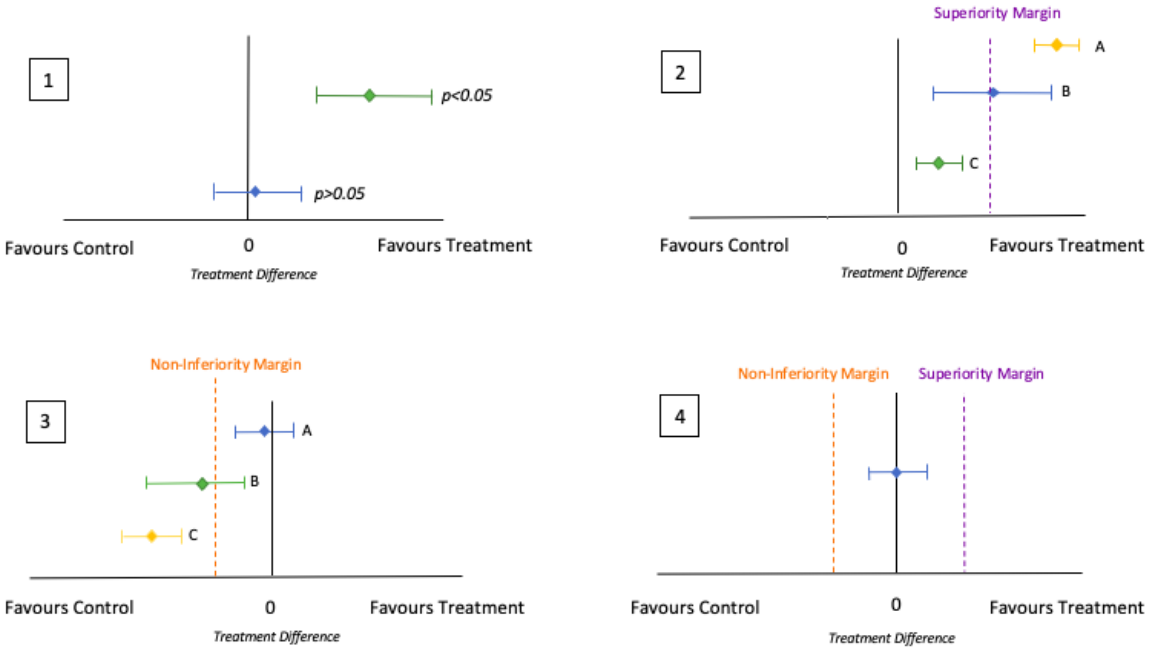

Figure 3-3. Forest plots labelled 1, 2, 3 and 4 represents the average between-group difference (diamond shape) with its associated 95% confidence interval (CI). In plot 1, the studies use the equality framework where the results of two separate studies show the relationship between the CIs, no difference (0), and achieving statistical significance. In plot 2, the studies use a superiority framework where A and C represent definitive results, whereas the results of study B cannot offer the same level of certainty. In plot 3, the studies are using a non-inferiority framework where A and C represent definitive results, whereas the results of study B are inconclusive. In plot 4, the study can conclude that the two treatments offer identical outcomes and can be used interchangeably. If the CIs around the between-group difference cross one or both margins, the study is inconclusive.

The placement of the non-inferiority margin will depend on the seriousness of experiencing worse outcomes than offered by usual care. For example, if switching to the treatment will result in a greater number of life-threatening or irreversible outcomes, we are unlikely to accept a large non-inferiority margin and the sample size requirements are likely to be large even if there are substantial savings to the institution or health system. On the hand, a more liberal non-inferiority margin may be acceptable if changing treatment will mean an increase in the number of patients with minor adverse events, inconveniences, or reversible, rare harmful events, especially if there are substantial savings to the institution or health system[8]. The degree of subjectivity and controversy attached to the specification of the margin has the potential to impact the uptake of findings. Wangge et al. found wide variations in non-inferiority margins for studies evaluating novel oral anti-coagulants after orthopaedic surgery, which led to inconsistent conclusions about the efficacy of the novel drug[25]. Therefore, a transparent description of how the clinicians arrived at the superiority or non-inferiority margin will provide readers with the context against which recommendations from the research team stem[26].

Green et al. evaluated clinical studies in a range of medical journals finding that only 23% of equivalence studies reported a pre-set margin[27]. In our study, we found 13 studies that concluded that two treatments were “interchangeable”, “similar”, “not different”, or “equivalent”. These studies interpreted a statistically non-significant result ($p \geq 0.05$) as evidence that there was no difference between treatments. This interpretation is incorrect, as a non-significant p-value only reveals that the null hypothesis (no difference between groups) is consistent with the observed results, but not that the null hypothesis is true[28]. Remember, we have only sampled the population and random sampling error is possible; to be able to confidently declare equivalence or comparability between treatments requires a definition of each margin, with justification, and a large sample size.

Reito et al. reviewed studies published in seven orthopaedic journals between 2016 and 2017, and found that the proportion of studies adequately powered to detect a clinically important difference ranged from 0% to 53% across different subspecialties[29]. Of the 60.5% of studies in our review that reported a sample size calculation, nearly every study used an equality calculation despite 43.8% claiming that one intervention was either

superior, non-inferior, or equivalent to another. A superiority, non-inferiority, and equivalence framework require a larger sample size than an equality framework because the margin sets restraints on the width of the CI[30]. Thus, the closer the defined margin is to the expected difference between groups, the larger the sample size requirements. We have provided two tables that illustrate sample size requirements for equality, superiority, and non-inferiority for varying values of the expected difference and the margins (Table 3-4, Table 3-5).

From these tables, one can see that the sample size requirements increase as the distance becomes smaller between the margin and expected between-group difference. We have also included the sample size requirements when no margin is instilled, where the CI just has to remain greater than 0 (if using MD or RD) or 1 (if using RR or OR) to achieve superiority. We have also identified values of effect size that represent within- and between-group MCIDs, which are commonly, but incorrectly, used to inform sample size estimates.

Specifically, the majority of MCIDs are determined using a within-group (pre- to post-design) where the MCID is the value of the average pre-to-post change in participants who claimed to have changed by a small but important amount following an intervention. However, the amount of change experienced within a group, whose members are aware that an intervention has been applied, will be larger than the difference that can be expected between two groups who have received an active control[31]. This is especially true when participants are blind to treatment group, because both groups are expected to demonstrate change. Thus, a within-groups MCID, which is approximately equal to an effect size of 0.5 standard deviation units[32], may be an unreasonably optimistic value to inform the expected difference in a sample size estimation.

Using an overly estimate of the expected difference between groups means that the study will be underpowered to make precise estimates of the effect of the treatment compared to control. It also increases the risk of a random sampling error, which means that by chance, the sample is not representative of the population[33]. The chance that studies obtain a representative sample increases with a larger sample size and multiple centres[34].

As a result, the larger sample required for a superiority, non-inferiority, or equivalence framework improves the applicability of findings, which is ultimately the goal of clinicians seeking to apply research findings to clinical decision-making.

Table 3-4. Sample size estimates for a superiority study (n per group).

Superiority Margin (M)	**Delta (δ)**								
	0.55	**0.50**	**0.45**	**0.40**	**0.35**	**0.30**	**0.25**	**0.20**	**0.15**
None*	52	63	**77**	98	128	174	251	392	697
0	41	49	61	77	100	137	197	308	547
0.05	49	61	77	100	137	197	308	547	1230
0.10	61	77	100	137	197	308	547	1230	4920
0.15	77	100	137	197	308	547	1230	4920	
0.20	100	137	197	308	547	1230	4920		
0.25	137	197	308	547	1230	4920			
0.30	197	308	547	1230	4920				
0.35	308	547	1230	4920					
0.40	547	1230	4920						
0.45	1230	4920							

Unless otherwise noted, sample size calculations include a one-sided alpha of 5% ($Z_{\alpha/2}$=1.64), a beta of 20% (β=0.84), and a standard deviation of 1.0.

*This represents an equality framework, where the statistical test is two-sided ($Z_{\alpha/2}$ = 1.96).

Table 3-5. Sample size estimates for a non-inferiority study (n per group).

Non-inferiority Margin (M)	Delta (δ)				
	-0.18	**-0.15**	**-0.10**	**-0.05**	**0.0**
None*	484	697	1568	6272	NA
-0.19	123008	7688	1519	628	341
-0.20	30752	4920	1230	547	308
-0.25	2510	1230	547	308	197
-0.30	854	547	308	197	137
-0.35	426	308	197	137	100
-0.40	254	197	137	100	77
-0.45	169	137	100	77	61
-0.50	120	100	77	61	49

Unless otherwise noted, sample size calculations include a one-sided alpha of 5% ($Z_{\alpha/2}$=1.64), a beta of 20% (β=0.84), and a standard deviation of 1.0.

*This represents an equality framework, where the statistical test is two-sided ($Z_{\alpha/2}$ = 1.96).

NA: Not applicable.

Our study is not without limitations. We evaluated high impact orthopaedic journals based on the annual Journal Citation Report, which is the ratio between citations and recent citable items published which may not always reflect quality[35]. Since we only evaluated five journals, we were also unable to capture the complete breadth of orthopaedic literature. When discussing explanatory trials, we have oversimplified the definition because our focus was on clinical applicability and treatment

recommendations. Further, it is likely that a large proportion of orthopaedic equality studies using surrogate or lab-based outcomes (i.e. basic science or biomechanical studies) are appropriately framing their research, but since our inclusion criteria evaluated only studies with a PROM, this number is low in our review. Lastly, because we focused on PROMs, we did not discuss this topic as it relates to dichotomous outcomes such as failure rates, although the majority of considerations are identical.

3.5 Conclusion

In conclusion, we found that the majority of published studies in top orthopaedic journals made conclusions based only on statistical findings and that recommendations for uptake into practice suggest an unjustified level of certainty, given the range of plausible treatment effects if CIs were used in place of p-values to interpret the study results. Researchers should state and justify their methodological framework (i.e., equality, superiority, non-inferiority, or equivalence) and choice of margin(s) in their protocol as it has implications for sample size and the applicability of conclusions. Editors should mandate the reporting and justification of a trial's framework and evaluate the consistency between the stated research question(s), sample size estimate, design (efficacy versus pragmatic), outcomes (surrogate versus patient important), interpretation of findings (p-values alone versus CIs), and whether recommendations to change practice are appropriate. This level of accountability will improve the quality of clinical trials in orthopaedics and the validity of their conclusions.

3.6 References

1. Chan A-W, Tetzlaff JM, Gotzsche PC, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ*. 2013;346(jan08 15):e7586-e7586. doi:10.1136/bmj.e7586

2. Dunn DT, Copas AJ, Brocklehurst P. Superiority and non-inferiority: two sides of the same coin? *Trials*. 2018;19(1):499. doi:10.1186/s13063-018-2885-z

3. Chow S-C, Shao J, Wang H. *Sample Size Calculations in Clinical Research*. Marcel Dekker; 2003. Accessed April 2, 2019. http://www.crcnetbase.com/isbn/9780824709709

4. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337-350. doi:10.1007/s10654-016-0149-3

5. Kamper SJ. Confidence Intervals: Linking Evidence to Practice. *J Orthop Sports Phys Ther*. 2019;49(10):763-764. doi:10.2519/jospt.2019.0706

6. Shafiq N, Malhotra S. Superiority trials: raising the bar of null hypothesis statistical testing. *Evid Based Med*. 2015;20(5):154-155. doi:10.1136/ebmed-2015-110280

7. Bigirumurame T, Kasim AS. Can testing clinical significance reduce false positive rates in randomized controlled trials? A snap review. *BMC Res Notes*. 2017;10(1):775. doi:10.1186/s13104-017-3117-4

8. Althunian TA, de Boer A, Groenwold RHH, Klungel OH. Defining the noninferiority margin and analysing noninferiority: An overview: Methods used to choose the margin and analyse noninferiority. *Br J Clin Pharmacol*. 2017;83(8):1636-1642. doi:10.1111/bcp.13280

9. Rehal S, Morris TP, Fielding K, Carpenter JR, Phillips PPJ. Non-inferiority trials: are they inferior? A systematic review of reporting in major medical journals. *BMJ Open*. 2016;6(10):e012594. doi:10.1136/bmjopen-2016-012594

10. Paesmans M, Grigoriu B, Ocak S, et al. Systematic qualitative review of randomised trials conducted in nonsmall cell lung cancer with a noninferiority or equivalence design. *Eur Respir J*. 2015;45(2):511-524. doi:10.1183/09031936.00092814

11. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJW, CONSORT Group for the. Reporting of Noninferiority and Equivalence Randomized Trials: An Extension of the CONSORT Statement. *JAMA*. 2006;295(10):1152. doi:10.1001/jama.295.10.1152

12. J. P T Higgins, S Green. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0.*; 2011. Accessed October 30, 2018. www.handbook.cochrane.org

13. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *International Journal of Surgery*. 2010;8(5):336-341. doi:10.1016/j.ijsu.2010.02.007

14. InCites Journal Citation Report - Web of Science Group. Published online 2017. https://jcr.clarivate.com/JCRJournalHomeAction.action?

15. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276-282.

16. Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ*. 2015;350(may08 1):h2147-h2147. doi:10.1136/bmj.h2147

17. Brooks D, Stratford P. Pilot Studies and Their Suitability for Publication in *Physiotherapy Canada*. *Physiotherapy Canada*. 2009;61(2):66-66. doi:10.3138/physio.61.2.66

18. Treweek S, Zwarenstein M. Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials*. 2009;10:37. doi:10.1186/1745-6215-10-37

19. Dickersin K, Chalmers I. Recognizing, investigating and dealing with incomplete and biased reporting of clinical research: from Francis Bacon to the WHO. *J R Soc Med*. 2011;104(12):532-538. doi:10.1258/jrsm.2011.11k042

20. Greene WL, Concato J, Feinstein AR. Claims of Equivalence in Medical Research: Are They Supported by the Evidence? *Ann Intern Med*. 2000;132(9):715. doi:10.7326/0003-4819-132-9-200005020-00006

21. Shafiq N, Malhotra S. Superiority trials: statistical trickery or mass blindness? *Postgrad Med J*. 2016;92(1084):118-119. doi:10.1136/postgradmedj-2015-133769

22. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 2. Interpreting study results: confidence intervals. *CMAJ*. 1995;152(2):169-173.

23. Hazra A. Using the confidence interval confidently. *J Thorac Dis*. 2017;9(10):4124-4129. doi:10.21037/jtd.2017.09.14

24. Lin L. Bias caused by sampling error in meta-analysis with small sample sizes. Chen Z, ed. *PLoS ONE*. 2018;13(9):e0204056. doi:10.1371/journal.pone.0204056

25. Wangge G, Roes KCB, de Boer A, Hoes AW, Knol MJ. The challenges of determining noninferiority margins: a case study of noninferiority randomized controlled trials of novel oral anticoagulants. *Canadian Medical Association Journal*. 2013;185(3):222-227. doi:10.1503/cmaj.120142

26. Rehal S, Morris TP, Fielding K, Carpenter JR, Phillips PPJ. Non-inferiority trials: are they inferior? A systematic review of reporting in major medical journals. *BMJ Open*. 2016;6(10):e012594. doi:10.1136/bmjopen-2016-012594

27. Greene WL, Concato J, Feinstein AR. Claims of Equivalence in Medical Research: Are They Supported by the Evidence? *Ann Intern Med*. 2000;132(9):715. doi:10.7326/0003-4819-132-9-200005020-00006

28. Bland JM, Altman DG. Statistics notes: Transformations, means, and confidence intervals. *BMJ*. 1996;312(7038):1079-1079. doi:10.1136/bmj.312.7038.1079

29. Reito A, Raittio L, Helminen O. Revisiting the Sample Size and Statistical Power of Randomized Controlled Trials in Orthopaedics After 2 Decades: *JBJS Reviews*. 2020;8(2):e0079. doi:10.2106/JBJS.RVW.19.00079

30. Walker E, Nowacki AS. Understanding Equivalence and Noninferiority Testing. *J GEN INTERN MED*. 2011;26(2):192-196. doi:10.1007/s11606-010-1513-8

31. Kamper SJ. Interpreting Outcomes 1—Change and Difference: Linking Evidence to Practice. *J Orthop Sports Phys Ther*. 2019;49(5):357-358. doi:10.2519/jospt.2019.0703

32. Norman GR, Sloan JA, Wyrwich KW. Interpretation of Changes in Health-related Quality of Life: The Remarkable Universality of Half a Standard Deviation. *Medical Care*. 2003;41(5):582-592. doi:10.1097/01.MLR.0000062554.74615.4C

33. Sampling Error. In: *Encyclopedia of Research Design*. SAGE Publications, Inc.; 2010. doi:10.4135/9781412961288.n401

34. Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?" *The Lancet*. 2005;365(9453):82-93. doi:10.1016/S0140-6736(04)17670-8

35. Eugene Garfield. *The Clarivate Analytics Impact Factor*. https://clarivate.com/webofsciencegroup/essays/impact-factor/

Chapter 4

4 General Conclusion and Future Directions

4.1 General Conclusion

Understanding important statistical and methodological concepts is essential when conducting clinical research and making informed treatment recommendations. To date, no studies have evaluated the reporting quality of patient-reported outcome measures (PROMs) or the use of all four methodological frameworks (equality, superiority, non-inferiority, and equivalence) in orthopaedic literature. In our first study, we found that most published studies rely solely on p-values to draw conclusions about between-groups differences and few (approximately one in five) report treatment effects with confidence intervals (CIs). In our second study, we found that half of the studies that reported a sample size calculation had inconsistency between the framing of their research question, sample size calculation, and conclusion. Our findings are problematic because p-values do not provide information on the magnitude or clinical relevance of the difference between treatment groups and inconsistencies in framework methodologies can lead to inaccurate sample size calculations and misinterpreted results.

Reporting results using treatment effects with CIs and applying the appropriate framework are essential concepts that must be considered together. P-values are often misinterpreted; for example, authors may interpret a significant p-value as evidence for superiority of a treatment or a non-significant p-value as evidence for no difference between treatments. However, a p-value only indicates if a difference between groups exists and is used as part of an equality framework to assess whether a meaningful change is likely to occur and make a decision on whether the study may move forward to a larger trial to determine its clinical effectiveness. A p-value, therefore, is insufficient evidence to declare superiority, non-inferiority, or equivalence between groups. To correctly declare superiority, equivalence, or non-inferiority, authors must report and interpret treatment effects, CIs, and clinically important thresholds. Further, the interpretation of treatment effects and CIs should align with the appropriate framework

because clinically important thresholds will differ depending on the framework used and the research question being asked. Additionally, since the precision of CIs depends, in part, on the sample size, the appropriate framework ensures that CIs are more likely to be precise and allows authors to be more confident in their conclusions. Compared to previous literature, our results underscore existing concerns of p-value misuse and inconsistent frameworks that have been identified in other biomedical literature. However, our investigation in the field in the orthopaedic surgery is unique and provides a novel perspective on these topics. We suspect that the issues we have identified can be attributed to varied statistical and methodological training of authors and reviewers and differing journal requirements; improvements in training combined with unified journal requirements will improve the quality of published research.

4.2 Future Directions

As part of our knowledge translation plan, Chapter 2 has been submitted to the Journal of Bone and Joint Surgery and the findings of both studies were presented at a number of events (Fowler Kennedy Sports Medicine Clinic Research Rounds, The Bone and Joint Trainee Lunch and Learn, and Western Research Forum in London, Ontario). Chapter 2 was also accepted as a podium presentation at the 2020 Canadian Orthopaedic Association Annual Meeting. Publishing our results and presenting at conferences are means to engage researchers, which is our target audience, and disseminate our findings. Additionally, it is known that statistical and methodological training quality during medical and graduate school can vary by institution[1,2]. We suggest that academic programs consider developing consistent and comprehensive statistics, methods, and critical appraisal training as part of their curriculum. As several orthopaedic journals have recently published papers highlighting the importance of sound research methodology and critical appraisal[3–6], our results provide evidence of the issues of statistical and methodological concepts in current orthopaedic literature and similarly support the need for improvements. Disseminating our findings and amendments to teaching at academic programs will facilitate critical appraisal of published literature and improve research quality within the orthopaedic community and beyond.

We also encourage journal editors to mandate the reporting of important information for manuscript submissions such as between-group treatment effects with CIs, clinically important thresholds, and frameworks to provide readers with a meaningful context. To improve the quality of manuscripts without compromising time and resources, we make several suggestions that can be implemented. Firstly, journal editors can modify their author instructions webpages to mandate the inclusion of these important key values. For example, upon submission, artificial intelligence can be used to screen for these values. If the authors do not satisfy the requirements, the submission engine would automatically return the manuscript to authors with a notification to include these values and ask them to revise before resubmitting their manuscript. For peer-reviewers, the increase in volume of sub-specializations in orthopaedic surgery and varying degrees of expertise provide different skillsets to address statistical concerns[7]. We propose that reviewers be trained to critically appraise whether these values are included and interpreted appropriately and notify authors. For example, some journals have implemented reviewer training through annual reviewer training days to uphold journal standards[7]. Direct communication with reviewers can help highlight important concepts and provide individuals with appropriate appraisal tools. Alternatively, comprehensive online module training may be provided to a select number of reviewers (similar to a pool of statistical experts), and editors can ensure that one of these trained reviewers reviews every submitted manuscript. Change will take time, but a commitment to adhere to high statistical and methodological standards from authors, editors, and reviewers will improve the quality of evidence in orthopaedic surgery and, ultimately, help the patients we serve.

4.3 References

1. Miles S, Price GM, Swift L, Shepstone L, Leinster SJ. Statistics teaching in medical school: opinions of practising doctors. *BMC Med Educ*. 2010;10:75. doi:10.1186/1472-6920-10-75

2. MacDougall M, Cameron HS, Maxwell SRJ. Medical graduate views on statistical learning needs for clinical practice: a comprehensive survey. *BMC Med Educ*. 2020;20(1):1. doi:10.1186/s12909-019-1842-1

3. Harris JD, Brand JC, Cote MP, Faucett SC, Dhawan A. Research Pearls: The Significance of Statistics and Perils of Pooling. Part 1: Clinical Versus Statistical Significance. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*. 2017;33(6):1102-1112. doi:10.1016/j.arthro.2017.01.053

4. Hohmann E, Feldman M, Hunt TJ, Cote MP, Brand JC. Research Pearls: How Do We Establish the Level of Evidence? *Arthroscopy: The Journal of Arthroscopic & Related Surgery*. 2018;34(12):3271-3277. doi:10.1016/j.arthro.2018.10.002

5. Harris JD, Brand JC, Cote MP, Dhawan A. Research Pearls: The Significance of Statistics and Perils of Pooling. Part 3: Pearls and Pitfalls of Meta-analyses and Systematic Reviews. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*. 2017;33(8):1594-1602. doi:10.1016/j.arthro.2017.01.055

6. Morshed S, Tornetta P, Bhandari M. Analysis of Observational Studies: A Guide to Understanding Statistical Methods: *The Journal of Bone and Joint Surgery-American Volume*. 2009;91(Suppl 3):50-60. doi:10.2106/JBJS.H.01577

7. Twaij H, Oussedik S, Hoffmeyer P. Peer review. *The Bone & Joint Journal*. 2014;96-B(4):436-441. doi:10.1302/0301-620X.96B4.33041

Appendices

Appendix A: Instructions for Reframing for Reviewers

*Classification should be based on the characteristics of the **interventions** as opposed to the hypothesis or purpose of the trial that authors have provided. Standard of care can be no treatment, wait and see, a conservative treatment and operative treatment and active treatment, etc.*

A study should be classified as **superiority** if:

- An intervention is being added to the standard of care
- An intervention will replace the standard of care
- I.e. conservative (usual care) vs operative (since operative carries more risks/most costly, operative would need to be shown to be superior to conservative)

A study should be classified as **non-inferiority** if:

- An existing intervention (or parts of) is being taken away (i.e. in-person visit (usual care) being replaced with an eHealth app, inpatient total hip arthroplasty (THA) (usual care) v outpatient THA)
- A treatment that is less costly (but may have more adverse events) is compared to the standard of care
- A treatment is expected to be less effective (but may cost less or have fewer side effects) compared to standard of care

A study should be classified as **equivalence** if:

- Two similar treatments for the same disease are compared (i.e. two common elbow surgeries for the same elbow problem, plating vs no plating in clavicle surgery, bone marrow aspirate concentrate (BMAC) vs. platelet rich plasma (PRP), anterior vs. posterior THA) and the intention is to recommend them as interchangeable, offering identical outcomes, risk profiles, etc.

A study should be classified as **equality** if:

- the study is a feasibility study only
- the study endpoints are surrogate outcomes or lab-based outcome (not patient important outcomes)

Curriculum Vitae

Name	Shgufta Docter
Post-secondary Education and Degrees	University of Western Ontario London, Ontario, Canada 2017-2020 M.Sc. Health and Rehabilitation Sciences, Measurements and Methods *Collaborative Specialization in Musculoskeletal Health Research* University of Toronto Toronto, Ontario, Canada 2013-2017 B.Kin., *High Honours*
Honours and Awards	Musculoskeletal Health Rehabilitation Research Network Award **(**2017) Transdisciplinary Bone and Joint Training Award (2017) R. Tait Mackenzie High Honours Society Inductee (2016)
Related Work Experience	Teaching Assistant, Counseling in Audiology University of Western Ontario Jan 2020-April 2020 Teaching Assistant, Advanced Quantitative Research Methods University of Western Ontario Sept 2019-Dec 2019 Teaching Assistant, Systemic Approaches to Functional Anatomy University of Western Ontario Sept 2018-April 2019 Laboratory Demonstrator, Elementary Human Anatomy University of Toronto Sept 2015-April 2017

Publications

1. **Docter S**, Philpott H, Godkin L, Bryant D, Somerville L, Jennings M, Marsh J, Lanting B. (2020). Comparison of Intra and Post-Operative Complication Rates Among Surgical Approaches in Total Hip Arthroplasty: A Systematic Review and Meta-Analysis. *Journal of Orthopaedics.* In Press, May 4, 2020.

2. **Docter S**, Khan M, Gohal C, Ravi B, Bhandari M, Gandhi R, Leroux T. (2020). Cannabis Use and Sport: A Systematic Review. *Sports Health: A multidisciplinary approach.* doi: 10.1177/1941738120901670. Feb 5, 2020.

3. **Docter S**, Khan M, Ekhtiari S, Veillette C, Paul R, Henry P, Leroux T. (2019). The Relationship between the Critical Shoulder Angle and the Incidence of Full-Thickness Rotator Cuff Tears and Outcomes after Rotator Cuff Repair: A Systematic Review. *Arthroscopy: The Journal of Arthroscopic and Related Surgery.* doi: 10.1016/j.arthro.2019.05.044. Nov 5, 2019.

4. Lebedeva K, Bryant D, **Docter S**, Litchfield RB, Getgood A, Degen RM. (2019). The Impact of Resident Involvement on Surgical Outcomes following Anterior Cruciate Ligament Reconstruction. *Journal of Knee Surgery*. doi: 10.1055/s-0039-1695705. Aug 28, 2019.

5. Paul R, Maldonado-Rodriguez N, **Docter S,** Khan M, Veillette C, Verma N, Nicholson G, Leroux T. (2019). Glenoid Bone Grafting in Reverse Total Shoulder Arthroplasty: A Systematic Review. *Journal of Shoulder and Elbow Surgery*. doi: 0.1016/j.jse.2019.05.011. Aug 8, 2019.

6. Siddiqi A, Forte S, **Docter S**, Bryant D, Chen A, Sheth N. (2019). Perioperative Antibiotic Prophylaxis in Total Joint Arthroplasty: A Systematic Review & Meta-Analysis. *The Journal of Bone and Joint Surgery.* doi: 10.2106/JBJS.18.00990. May 1, 2019.

Conference Presentations

1. **Docter S**, Fathalla Z, Lukacs M, Khan M, Jennings M, Dong S, Liu S, Getgood A, Bryant D. (June 2020). Interpreting Patient Reported Outcome Measures in Orthopaedic Surgery: A Systematic Review. Podium Presentation. Canadian Orthopaedic Association Annual Meeting. Halifax, Canada. **Virtual due to COVID-19*

2. **Docter S,** Fathalla Z, Lukacs M, Khan M, Jennings M, Dong S, Liu S, Bryant D. (March 2019). Interpreting Patient Reported Outcome Measures in Orthopaedic Surgery: A Systematic Review. Poster Presentation. Western Research Forum. London, Canada.

3. **Docter S**, Tamminen K. On the hunt: Searching for and finding information and policies regarding parent and athlete concerns in youth hockey. (March 2017). Podium Presentation. The 18th Annual Bertha Rosenstadt National Undergraduate Research Conference in Kinesiology and Physical Education. Toronto, Canada.