ELSEVIER

CrossMark

The Spine Journal 15 (2015) 2188–2197

THE SPINE JOURNAL

Clinical Study

The fragility of statistically significant findings from randomized trials in spine surgery: a systematic survey

Nathan Evaniew, MD[a,b,*], Carly Files, BSc[a], Christopher Smith, MSc[a], Mohit Bhandari, MD, PhD, FRCSC[a,b], Michelle Ghert, MD, FRCSC[a], Michael Walsh, MD, MSc, PhD, FRCPC[b], Philip J. Devereaux, MD, PhD, FRCPC[b], Gordon Guyatt, MD, MSc, FRCPC[b]

[a]*Division of Orthopaedics, Department of Surgery, McMaster University, Hamilton, Ontario, Canada*
[b]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada*

Received 28 January 2015; revised 27 April 2015; accepted 1 June 2015

Abstract

BACKGROUND CONTEXT: Randomized controlled trials (RCTs) are the most trustworthy source for evaluating treatment effects, but RCTs of spine surgery interventions often produce discordant results. The Fragility Index is a novel metric to inform about the robustness of statistically significant results.

PURPOSE: The aim was to determine the robustness of statistically significant results from RCTs of spine surgery interventions.

STUDY DESIGN/SETTING: This was a systematic survey.

PATIENT SAMPLE: The sample included RCTs of spine surgery interventions.

OUTCOME MEASURES: The Fragility Index is the minimum number of patients in a trial whose status would have to change from a nonevent to an event to change a statistically significant result to a nonsignificant result. Events refer to the occurrence of any dichotomous outcome, such as successful fusion, incident fracture, adjacent segment degeneration, or achievement of a certain functional score. A small Fragility Index indicates that the statistical significance of a result hinges on only a few events, and a large Fragility Index increases one's confidence in the observed treatment effects.

METHODS: We systematically reviewed a database for evidence-based orthopedics and identified all the RCTs that reported at least one positive outcome (ie, $p<.05$). Two reviewers independently assessed eligibility and extracted data. We used the Fisher exact test to compute Fragility Index values and multivariable linear regression to evaluate potential associated factors.

RESULTS: We identified 40 eligible RCTs with a median sample size of 132 patients (interquartile range [IQR] 79–208) and a median total number of outcome events for the chosen outcome of 31 (IQR 13–63). The median Fragility Index was two (IQR 1–3), which means that adding two events to one of the trial's treatment arms eliminated its statistical significance. The Fragility Index was less than or equal to three events in 75% of the trials, and was less than or equal to the number of patients lost to follow-up in 65% of the trials. Fragility Index values correlated positively with total sample size ($r=0.35$; $p<.05$). When adjusted for losses to follow-up and risk of bias, increasing Fragility Index values were associated only with increasingly significant reported p values ($p<.01$).

CONCLUSIONS: Statistically significant results in spine surgery RCTs are frequently fragile. The addition of only a small number of outcome events can completely eliminate significance.

FDA device/drug status: Not applicable.

Author disclosures: ***NE:*** Nothing to disclose. ***CF:*** Employment: OrthoEvidence Inc.. ***CS:*** Other: OrthoEvidence (Employee). ***MB:*** Consultancy: Stryker (B). ***MG:*** Wright Medical (Consultant ad hoc—C). ***MW:*** Pfizer (D, Research grant); Isis Pharmaceuticals (Consulting, B). ***PJD:*** Nothing to disclose. ***GG:*** Nothing to disclose.

The disclosure key can be found on the Table of Contents and at www.TheSpineJournalOnline.com.

This study was not funded.

This study did not require local institutional ethics board approval.

* Corresponding author. Division of Orthopaedics, Department of Surgery, McMaster University, Hamilton, Ontario, Canada. Tel.: (905) 521-4322x44992; fax: (905) 523-8781.

E-mail address: nathan.evaniew@medportal.ca (N. Evaniew)

http://dx.doi.org/10.1016/j.spinee.2015.06.004

Surgeons, researchers, and other evidence users should exercise caution when interpreting the findings from RCTs with low Fragility Index values and applying these results to patient care.

Keywords: Spine surgery; Randomized controlled trials; Outcomes; Statistical significance; Fragility Index; Clinical epidemiology

Introduction

The most trustworthy results for addressing the impact of treatment effects and establishing causality come from rigorously conducted and adequately powered randomized controlled trials (RCTs), but RCTs of spine surgery interventions often provide discordant results [1]. Although each of study quality, sample size, and conflicts of interest have been previously explored as potential associated factors, little attention has focused on the importance of the number of outcome events in each arm [2–6]. Trials with small numbers of outcome events are at risk of producing implausibly large treatment effects, particularly when their sample sizes are also small [7,8].

The Fragility Index was recently developed as a novel metric to describe the robustness of statistically significant results, and it is intended to complement p values and 95% confidence intervals (CIs) [9]. The Fragility Index for a given study is defined as the minimum number of patients in the trial group with fewer outcome events whose status would have to change from a nonevent to an event to change a statistically significant result to a nonsignificant result. Events refer to the occurrence of any dichotomous outcome, such as successful fusion, incident fracture, adjacent segment degeneration, or achievement of a certain functional score. A small Fragility Index indicates that the statistical significance of a result hinges on only a few events, and a large Fragility Index increases one's confidence in the observed treatment effects.

For example, consider an RCT in which 100 patients with symptomatic spinal stenosis were randomized to either surgical treatment with a minimally invasive interspinous spacer or conventional open decompression [10]. In this trial, 13 patients in the interspinous spacer group underwent a subsequent reoperation in comparison to three patients in the conventional open decompression group. This difference was statistically significant (p=.04), but it would have been completely insignificant if just two additional patients in the conventional open decompression group had also undergone a reoperation (p=.07). Thus, the Fragility Index for this outcome is two events.

Investigators can apply the Fragility Index to any dichotomous outcome in a 1:1 parallel design RCT, and its application does not require specialized statistical expertise. In their review of 399 randomized trials from high-impact medical journals, Walsh et al. [9] reported a median Fragility Index of eight events; 24% percent of the included trials had a Fragility Index of three or less, and 53% had a Fragility Index less than the number of patients lost to follow-up.

Given that many trials in spine surgery are characterized by small sample sizes and few events [5,6,11], our primary objective was to determine the robustness of statistically significant results in RCTs of spine surgery interventions by systematically applying the Fragility Index. Our secondary objective was to identify potential factors associated with the Fragility Index.

Methods

Eligibility criteria

We performed a systematic survey of RCTs of spine surgery interventions published from January 2009 to September 2014. We included all trials that reported in their abstract at least one statistically significant dichotomous outcome (ie, p value <.05 under a null hypothesis that no difference existed or a 95% CI that excluded a null value), were randomized according to a 1:1 parallel two-arm design, and examined a preoperative, intraoperative, or postoperative intervention in patients undergoing spine surgery.

Identification of studies

We identified potential trials using a systematic database for evidence-based orthopedics [12]. The database search strategy includes search terms that identify relevant RCTs published in orthopedic surgery journals, neurosurgery journals, and general medical journals without any language restrictions (Appendix 1). The database search strategy is executed at the beginning of each month using MEDLINE, and its contents have been kept updated from January 1, 2009 onward. We queried the database on September 11, 2014, and two reviewers independently screened the titles and abstracts of all studies using piloted electronic forms. Reviewers resolved discrepancies through discussion of the rationale for their decisions.

Assessments of risk of bias and data extraction

Two reviewers independently assessed trial-level risk of bias and extracted study outcome data using piloted electronic forms. Risk of bias was assessed using the Cochrane Collaboration Risk of Bias tool [13], which includes items