THE ORTHOPAEDIC FORUM

Randomized Controlled Trials for Geriatric Hip Fracture Are Rare and Underpowered

A Systematic Review and a Call for Greater Collaboration

Joseph Bernstein, MD, Sara Weintraub, MD, Tyler Morris, MD, and Jaimo Ahn, MD, PhD

Investigation performed at the University of Pennsylvania, Philadelphia, Pennsylvania

Background: Geriatric hip fracture is a common condition, and there are many open questions regarding patient management. Among the various types of medical evidence, the prospective randomized controlled trial (RCT) is considered the best. Our primary hypothesis was that small sample size would be seen frequently among RCTs involving geriatric patients with hip fracture. A related hypothesis was that studies from the United States would have particularly large deficits in sample size. Therefore, we asked the following research questions: (1) What is the mean sample size of RCTs involving geriatric patients with hip fracture? (2) How do sample sizes for studies from the U.S. differ from those performed elsewhere?

Methods: Following the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines, a systematic review of hip fracture RCTs was conducted. The Embase and MEDLINE databases were searched. Additional data included the country of origin, the power of the study, and whether sample size calculations were performed. One hundred and forty-seven RCTs were identified.

Results: The mean sample size of the 147 RCTs was 134.9. The mean sample size for the 7 American trials was 110.3, and the mean sample size for all trials conducted outside of the United States was 136.1. A sample size that was sufficient to ensure 80% power was used in only 31.3% of the RCTs.

Conclusions: RCTs for hip fracture are small and underpowered. Moreover, <5% of the RCT studies have been conducted in the U.S., and they were smaller than those conducted elsewhere. The shortage of American trials may be a feature of the dispersion of geriatric hip fracture care across many hospitals in the United States. If so, better clinical research might require more centralized care (e.g., in specialized geriatric hip fracture centers) or greater collaboration among the many hospitals that provide care.

Geriatric hip fractures are increasingly common, with an estimated annual incidence of approximately 1 million cases across the globe[1]. Although medical and surgical treatments certainly help many patients with hip fracture, the overall clinical and functional outcomes following geriatric hip fracture are poor: only about one-third of patients return to their preinjury

Disclosure: The authors indicated that no external funding was received for any aspect of this work. The **Disclosure of Potential Conflicts of Interest** forms are provided with the online version of the article (http://links.lww.com/JBJS/F514).

functional status, and the remainder either lose a level of independence or die within the first year[2]. Simply put, there is room for improvement. Thus, the topic of geriatric hip fracture is worthy of intensive research.

Among the various types of evidence that could be used to improve the care of geriatric hip fractures, the prospective randomized controlled trial (RCT) is considered to be the highest quality[3]. The unique features that are described by its name (randomization, in particular) make it more resistant to bias. Still, because they are difficult to execute for surgical studies[4], RCTs are uncommon in orthopaedics[5] and when such trials are conducted, the sample size is often too small.

The primary hypothesis of this study was that the problem of small sample size would be seen frequently among RCTs involving geriatric patients with hip fracture (including trials that study nonsurgical aspects of care). A second and related hypothesis was that studies from the United States would have particularly large deficits in sample size.

To assess these hypotheses, a systematic review utilizing PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines[6] was undertaken. This review addressed the following questions: What is the mean sample size of RCTs involving geriatric patients with hip fracture? How does the mean sample size for studies emanating from the U.S. differ from those performed in other countries? Was sample size appropriately considered and described in the studies?

Materials and Methods

Registration

The study was preregistered at PROSPERO (international prospective register of systematic reviews (http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42016043545).

Search Strategy and Criteria

The Embase and MEDLINE databases were searched in December 2016. The following Boolean search strings were used: "intertrochanteric" OR "femoral neck" OR "proximal femur" OR "hip" AND ("randomized controlled study"/exp OR "randomized controlled study") AND "surgery" AND "fracture" AND [article]/lim).

No limits on the date of publication were used.

Inclusion and Exclusion

All English-language reports of RCTs involving the care of acute geriatric hip fracture were included. Geriatric patients were defined as those ≥60 years of age. Studies were excluded if they assessed patients <60 years of age or if the research was described as a pilot study.

In the instances where there were multiple manuscripts reporting on the same cohort, only 1 manuscript was included in order to not overcount. Inclusion and exclusion criteria were applied by the second and third authors of this review, and disparities were resolved by the first and fourth authors.

Assessment of Study Quality

Because the only data of interest were the sample size and the country of origin, no assessment of study quality was made. That is, the data were not pooled in any way and, therefore, there was no need to weight the studies according to quality.

Data Collection and Abstraction

From each study, 2 readers (the second and third authors of this review) recorded the following:

- In what country was the trial done?
- What was the sample size (noting the number of patients, facilities, and surgeons who participated)?
- Was a sample size/power calculation performed? If so, what was the study's power?
- Were significant outcomes reported?

Results

A total of 656 manuscripts were found using the structured search parameters. Of these, 509 were excluded as not pertaining to geriatric hip fracture, not available from our library system, duplicate reports, or pilot studies. A final group of 147 geriatric hip fracture RCT cohorts was available for analysis (Fig. 1). Although no limits on the date of publication were used, in point of fact, the earliest study was from 1981 and the most recent was from 2016.

Among the 147 geriatric hip fracture RCT cohorts, there were 30 countries of origin (Table I). There were 7 trials (4.8%) from the United States. Overall, the mean sample size for all trials was 134.9 (standard deviation [SD] = 110.4). The mean sample size for the 7 U.S. trials was 110.3 (SD = 75.1) compared with a mean sample size of 136.1 (SD = 112.0) for the trials conducted outside of the U.S.

Of the 147 included trials, a priori sample size calculations were described in 62 studies (42.2%), and 16 of these reported power of <0.80 despite the calculation. Thus, there were 46 trials (31.3%) that were adequately powered in advance. Fourteen (87.5%) of these 16 underpowered studies reported significant outcomes. Only 3 of the 7 U.S. papers reported an a priori

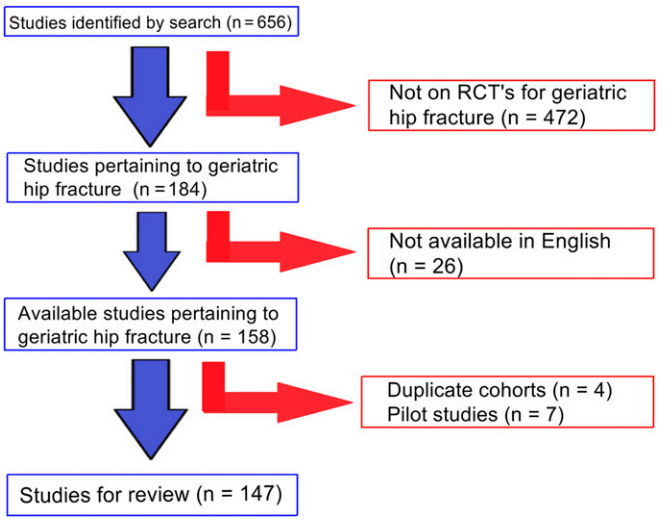

Fig. 1

Flowchart demonstrating the manuscript review process. RCT's = randomized controlled trials.

TABLE I Sources of RCTs by Country (Reporting ≥5 Cohorts)

Country	No. of RCT Cohorts	Mean Sample Size
Australia	5	101.0
China	7	125.1
Greece	6	106.5
Israel	5	68.4
Italy	5	54.6
Netherlands	6	194.2
Norway	5	346.4
Spain	9	172.9
Sweden	22	133.4
United Kingdom	23	166.2
United States	7	110.3

sample size calculation, and just 1 of the RCTs that did a calculation achieved a sample size sufficient to provide 80% power.

Considering only those studies with a priori sample size calculation, the mean number of subjects was 164.5, whereas for studies with no a priori calculation, the mean sample size was 113.3.

Discussion

Among the various types of evidence that can drive medical treatment decision-making, the prospective RCT is the most resistant to the various biases that can skew results. A prospective study is less likely to retrospectively define the outcomes of interest or fail to recognize lost data. Additionally, randomization helps to ensure that patients receiving different treatments are roughly alike, and the contrast offered by a control group defines the relative contribution of the intervention to the observed outcome.

Within this context, an RCT must be adequately powered if it is to correctly drive medical decision-making. A trial with 2 patients cannot teach us which treatment is better; depending on the distribution of the data and the effect size of interest, perhaps even hundreds of patients will be needed.

In this systematic review, we examined the hypothesis that RCTs pertaining to geriatric hip fractures are inadequately powered. We reviewed every cohort for geriatric hip fracture RCTs that had been reported in the Embase and MEDLINE databases and found that 68.7% of all reported RCTs assessing geriatric patients with hip fracture did not use a sample size that was sufficient to ensure 80% power a priori.

Using an adequate sample size is crucial if the correct inference is to be drawn from the trial report. The first and perhaps more obvious reason is that without adequate subjects, a study is at risk of not detecting significant differences and, therefore, wrongly concluding that there are no real differences—a problem known as a type-II error. Beyond that, a lack of power also can mar studies that do find significant differences: namely, if a study has too few subjects, it can attain significance only if the effect reported is (by chance) larger than the true difference. This gives rise to what can be termed a "type-M error"[7]: the magnitude of the effect will be overstated in the minds of readers who do not scrutinize the confidence intervals[8].

Because meaningful results can be produced only by adequately powered trials, conducting a trial of inadequate power is at best an invitation to futility. Trials are among the most expensive and labor-intensive studies, and to initiate one that is unlikely to produce meaningful results is a grievous mistake. Worse, a type-M error may not be recognized as a failure, and wrong conclusions might be drawn.

The results found herein are consistent with those reported in the literature. Freedman et al. found 33 RCTs in a single year in *The Journal of Bone & Joint Surgery* and in *Clinical Orthopaedics and Related Research*[5]. Among them, only 3 studies described calculations of sample size; of the 25 studies that reported negative results, none had adequate power to detect a small effect size, and 12 lacked the power necessary to detect a large effect size. More recently, Checketts et al. studied the RCTs that were cited in the American Academy of Orthopaedic Surgeons (AAOS) Clinical Practice Guidelines and found that 38 (53%) of 72 trials were underpowered[9].

We additionally discovered that there were relatively few studies conducted in the U.S. (<5%). Ranked by contribution and adjusting for population size (since larger countries should contribute more), the U.S. was twenty-fifth. By contrast, note that papers from the United States represent nearly half of all papers appearing in the 10 most frequently cited orthopaedic surgery journals. Thus, it is reasonable to conclude there is a distinct shortage of American-based trials.

Also, the trials in the U.S. were smaller than those conducted elsewhere. Indeed, only 1 American trial was sufficiently large to have 80% power to detect a so-called "medium effect" (defined as a difference in means of equal to one-half of the standard deviation)[10].

The shortage of American trials may be a feature of the dispersion of geriatric hip fracture care in the United States. For example, it was reported that in the Philadelphia metropolitan area, among hospitals offering hip fracture care, the mean number of cases per year was 26.8, with 66% of the cases performed at a hospital that treated <150 geriatric patients with hip fracture[11]. Thus, even with a very high recruitment rate, it would be difficult for a single institution to perform an adequately powered trial.

It is interesting to consider why RCTs may be found more commonly outside of the United States. Djurisic et al. have identified some barriers to conducting such trials, including lack of funding, excessive monitoring, restrictive privacy laws, complex regulatory requirements, and inadequate infrastructures[12]. These barriers may be applied disproportionately in the United States. Also, public health systems that use national personal identification numbers (e.g., Sweden and Norway) may facilitate clinical research. More to the point, public health systems may demand demonstration of cost-effectiveness and that requirement, in turn, demands more and better research.

Beyond the issues of methodology, the overall paucity of American RCTs examining geriatric hip fracture may be a feature of the condition itself. As noted by Lane in his commentary on Bernstein's proposal for geriatric hip fracture

centers, many patients presenting with geriatric hip fracture are inadequately lucid at the time of presentation to consent to participating in an RCT[13]. Also, RCTs regarding some questions pertaining to geriatric hip fracture (notably, the effect of surgical delay) are ethically complex and, thus, only pseudorandomized trials using instrumental variables[14] or other techniques may be truly feasible.

Limitations

We have identified the following limitations of our study. To start, only geriatric hip fracture trials were assessed. Thus, while we may claim with certainty that such trials are characterized by small samples, we may only speculate how this result compares to what might be found with other clinical problems. That is, we cannot claim that the problems of small sample size or the rarity of trials in general are worse in the case of geriatric hip fracture.

In addition, we did not calibrate the importance of the questions that were asked in the trial. It may be the case that when more important questions are addressed, better methods will be used[3]. Because journals may wish to publish RCTs in an effort to claim a higher level-of-evidence pedigree[15], underpowered trials may be overrepresented in the literature.

It also must be noted that although so-called Level-I medical evidence requires RCT data, there are some questions for which the RCT is not the best method of study. For example, the question of whether a urinary tract infection at the time of admission increases the likelihood of a late surgical site infection is best answered with a case-control study rather than an RCT. Indeed, medical evidence strong enough to guide clinical practice (e.g., that a given device is too prone to failure) may emerge from even a small case series. Along those lines, the finding that geriatric hip fracture RCTs are rare in the United States and underpowered overall does not mean that the overall quality of the literature on this topic is poor. After all, there are many questions that are best answered with other methods[16].

Our study sample was limited to only English-language studies. Although this routinely employed strategy[17] certainly did not lead to the exclusion of any American study, it may have limited our ability to contrast American RCTs with those performed elsewhere ("elsewhere" is defined as reporting research in languages besides English). Yet, just as one can assess the outcome of treatment even if some patients are lost to follow-up (by assuming the worst for all of those patients who were not found), we can create an upper bound on the magnitude of error that is introduced by study exclusions by assuming that all missing trials from outside the United States were as small as can be. Even if all of the excluded trials had only 2 subjects each (it would not be a trial if it had fewer), it still is the case that American trials are smaller overall. At present, there are 19,054 subjects in 140 non-U.S. trials. If there were 2 patients in each of the 26 trials that were not considered, there would be 19,106 subjects in a set of 166 trials. This yields a mean sample size of 115, which is still larger than the mean for U.S. trials (110 subjects).

It is also important to recognize that differences regarding RCTs in the United States and Europe may be based in part on differences regarding ethical considerations of enrolling patients with dementia—a substantial segment of the hip fracture population. Although presumed consent[18] may be used in the United States for treating the fracture, patients who cannot provide consent are apt to be excluded from research studies. By contrast, many studies from the United Kingdom and Scandinavia did, in fact, include patients with dementia. This difference not only potentially skews the results of the American trials but also is likely to yield a smaller sample size in them.

Finally, because there were only 147 geriatric hip fracture RCT cohorts, the sample was too small to make meaningful statements about temporal trends or differences between journals. The advent of pretrial registration along with more stringent requirements at high-quality journals combine to make it more likely that underpowered RCTs will be rarer moving forward.

Conclusions

RCTs for geriatric hip fracture are rare in the United States, and they are underpowered overall. Because of the high annual incidence of geriatric hip fractures and their high costs (both medical and financial), high-quality care is needed, and, in turn, high-quality medical evidence upon which that care depends also is needed. Better clinical research in the U.S. might require more centralized care (e.g., in specialized geriatric hip fracture centers) or greater collaboration among the many hospitals that proved care. The systematic review data presented herein indicate that a greater emphasis on well-powered RCTs for geriatric hip fracture is needed. ■

Joseph Bernstein, MD[1]
Sara Weintraub, MD[1]
Tyler Morris, MD[1]
Jaimo Ahn, MD, PhD[1]

[1]Department of Orthopaedic Surgery, University of Pennsylvania, Philadelphia, Pennsylvania

Email address for J. Bernstein: joseph.bernstein@uphs.upenn.edu

ORCID iD for J. Bernstein: 0000-0001-9052-2897
ORCID iD for S. Weintraub: 0000-0002-6669-1380
ORCID iD for T. Morris: 0000-0002-5328-7303
ORCID iD for J. Ahn: 0000-0001-8151-9987

References

1. Chen IJ, Chiang CY, Li YH, Chang CH, Hu CC, Chen DW, Chang Y, Yang WE, Shih HN, Ueng SW, Hsieh PH. Nationwide cohort study of hip fractures: time trends in the incidence rates and projections up to 2035. Osteoporos Int. 2015 Feb;26(2):681-8. Epub 2014 Oct 30.

2. von Friesendorff M, McGuigan FE, Wizert A, Rogmark C, Holmberg AH, Woolf AD, Akesson K. Hip fracture, mortality risk, and cause of death over two decades. Osteoporos Int. 2016 Oct;27(10):2945-53. Epub 2016 May 12.

3. Bhandari M, Devereaux PJ, Einhorn TA, Thabane L, Schemitsch EH, Koval KJ, Frihagen F, Poolman RW, Tetsworth K, Guerra-Farfán E, Madden K, Sprague S, Guyatt G; HEALTH Investigators. Hip fracture evaluation with alternatives of total hip arthroplasty versus hemiarthroplasty (HEALTH): protocol for a multicentre randomised trial. BMJ Open. 2015 Feb 13;5(2):e006263.

4. Farrokhyar F, Karanicolas PJ, Thoma A, Simunovic M, Bhandari M, Devereaux PJ, Anvari M, Adili A, Guyatt G. Randomized controlled trials of surgical interventions. Ann Surg. 2010 Mar;251(3):409-16.

5. Freedman KB, Back S, Bernstein J. Sample size and statistical power of randomised, controlled trials in orthopaedics. J Bone Joint Surg Br. 2001 Apr;83(3):397-402.

6. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg. 2010;8(5):336-41. Epub 2010 Feb 18.

7. Gelman A, Carlin J. Beyond power calculations: assessing Type S (sign) and Type M (magnitude) errors. Perspect Psychol Sci. 2014 Nov;9(6):641-51.

8. Porcher R. Reporting results of orthopaedic research: confidence intervals and p values. Clin Orthop Relat Res. 2009 Oct;467(10):2736-7. Epub 2009 Jun 30.

9. Checketts JX, Scott JT, Meyer C, Horn J, Jones J, Vassar M. The robustness of trials that guide evidence-based orthopaedic surgery. J Bone Joint Surg Am. 2018 Jun 20;100(12):e85.

10. Freedman KB, Bernstein J. Sample size and statistical power in clinical orthopaedic research. J Bone Joint Surg Am. 1999 Oct;81(10):1454-60.

11. Clement RC, Ahn J, Mehta S, Bernstein J. Economic viability of geriatric hip fracture centers. Orthopedics. 2013 Dec;36(12):e1509-14.

12. Djurisic S, Rath A, Gaber S, Garattini S, Bertele V, Ngwabyt SN, Hivert V, Neugebauer EAM, Laville M, Hiesmayr M, Demotes-Mainard J, Kubiak C, Jakobsen JC, Gluud C. Barriers to the conduct of randomised clinical trials within all disease areas. Trials. 2017 Aug 1;18(1):360.

13. Bernstein J. Not the last word: Bhandari's paradox. Clin Orthop Relat Res. 2018 Apr;476(4):674-7.

14. McGuire KJ, Bernstein J, Polsky D, Silber JH. The 2004 Marshall Urist Award: delays until surgery after hip fracture increases mortality. Clin Orthop Relat Res. 2004 Nov;428:294-301.

15. Hanzlik S, Mahabir RC, Baynosa RC, Khiabani KT. Levels of evidence in research published in The Journal of Bone and Joint Surgery (American volume) over the last thirty years. J Bone Joint Surg Am. 2009 Feb;91(2):425-8.

16. Baldwin KD, Bernstein J, Ahn J, McKay SD, Sankar WN. Level of evidence gap in orthopedic research. Orthopedics. 2012 Sep;35(9):e1416-9.

17. Butler M, Forte ML, Joglekar SB, Swiontkowski MF, Kane RL. Evidence summary: systematic review of surgical treatments for geriatric hip fractures. J Bone Joint Surg Am. 2011 Jun 15;93(12):1104-15.

18. Bernstein J, LeBrun D, MacCourt D, Ahn J. Presumed consent: licenses and limits inferred from the case of geriatric hip fractures. BMC Med Ethics. 2017 Feb 24;18(1):17.