Published in final edited form as:
J Clin Epidemiol. 2011 March ; 64(3): 280–285. doi:10.1016/j.jclinepi.2010.04.008.

Departures from community equipoise may lead to incorrect inference in randomized trials

Jeffrey N. Katz, MD, MSc, **John Wright, MD**, **Bruce A. Levy, MD**, **John A. Baron, MD, MSc**, and **Elena Losina, PhD**
Orthopaedic and Arthritis Center for Outcomes Research, Department of Orthopaedic Surgery (JNK, JW, EL) and Division of Rheumatology, Immunology and Allergy (JNK, EL), Brigham and Women's Hospital, Harvard Medical School; Department of Epidemiology, Harvard School of Public Health (JNK), Department of Biostatistics, Boston University School of Public Health (EL) Boston; Department of Medicine, Dartmouth Medical School, Lebanon NH (JAB); and Department of Orthopedic Surgery, Mayo Clinic, Rochester, MN (BAL)

Abstract

Objective—To assess the impact of selective enrollment on the results of randomized, controlled trials (RCTs).

Study Design and Setting—We simulated a RCT of arthroscopic partial meniscectomy vs. nonoperative therapy in patients with meniscal tear and osteoarthritis (OA). We estimated efficacy with the risk ratio (RR) comparing the likelihood of clinically important improvement following surgery with that following nonoperative therapy. We assumed that efficacy differs by extent of OA. We simulated four scenarios: 1) non-selective enrollment; 2) higher likelihood of enrolling subjects with mild OA; 3) higher likelihood of enrolling subjects with severe OA; 4) much higher likelihood of enrolling subjects with severe OA. For each scenario we simulated 100 trials with sample size 340.

Results—With non-selective enrollment, reflecting community equipoise, the results in 100 trials were consistent with those in the underlying population (mean RR = 1.87 (95% CI 1.57, 2.14). Selective enrollment of subjects with much higher likelihood of severe OA resulted in 28% lower efficacy of surgery (mean RR=1.34 (95% CI 0.93, 2.15)), with 95% CI containing the true efficacy in just 25% of trials and empirical power of 44%.

Conclusion—Selective enrollment with respect to factors associated with efficacy may affect trial results and lead to inaccurate conclusions.

Keywords

equipoise; randomized controlled trial; meniscectomy; selection bias; simulation; generalizability; arthroscopy

Correspondence: Jeffrey N. Katz, MD, MSc, Orthopaedic and Arthritis Center for Outcomes Research, Brigham and Women's Hospital, OBC – 4, 75 Francis Street, Boston, MA 02115, Tel: 617 732 5338; Fax 617 525 7900, jnkatz@partners.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

The randomized controlled trial (RCT) is widely recognized as the most rigorous method for establishing the efficacy of health care interventions. Problems may arise, however, if certain patients are reluctant to enroll in a trial, or if physicians are reluctant to recommend patients for a trial, based on specific clinical characteristics..

Clinicians' beliefs about the optimal management of individual patients may create a dilemma when clinicians enroll patients into a randomized controlled trial. On the one hand, if suitably planned, the trial's entry and exclusion criteria reflect 'community equipoise' – defined as the set of circumstances in which the community of clinicians and other scientists that designed the trial are comfortable with either treatment option.[1–3] On the other hand, the individual clinician may have strong beliefs about the optimal management for a particular patient, even though the patient meets eligibility criteria. In these circumstances, individual equipoise – defined as the individual clinician's comfort with both options under study – may not be congruent with community equipoise.[1–5]

When a clinician investigator endorses the eligibility criteria for a trial and yet is reluctant to randomize specific patients who meet these criteria, the clinician's preferences and community equipoise come into tension. Should the standard for physician conduct in randomized trials be community equipoise (clinician offers randomization to all eligible patients) or individual equipoise (clinician only offers randomization when he or she feels comfortable recommending both options)?

We do not intend to resolve this ethical debate but rather to quantify what is at stake. We seek to determine whether deviations from community equipoise have implications for the findings and interpretation of randomized trials. We perform a simulation study based upon a common clinical dilemma, the use of arthroscopic partial meniscectomy versus nonoperative therapy in subjects with symptomatic meniscal tear and underlying osteoarthritis. This is the focus of an ongoing RCT, the MeTeOR Trial (Meniscal Tear in Osteoarthritis Research; Clinicaltrials.gov NCT00597012). The efficacy of surgery in this setting is uncertain.[6, 7] Our specific concern is whether selective enrollment with respect to the extent of underlying osteoarthritis would affect trial results. Observational cohort studies have shown that patients with more severe underlying osteoarthritis tend to have more pain and functional loss than patients with less severe underlying arthritis if they are managed surgically.[8] However, patients with more severe OA also may have a worse outcome following nonoperative therapy; this question has not been examined rigorously.

We simulate several alternative enrollment scenarios ranging from community equipoise (all eligible patients enrolled) to selective enrollment with respect to knee OA severity, a factor that may be related to treatment outcomes. The question we address is whether randomization of certain subgroups of eligible patients, and failure to randomize other eligible patients, affects trial results.

Methods

Overview—We simulated trials of the same size as MeTeOR under alternative subject enrollment criteria reflecting selective surgeon enrollment preferences that deviate from community equipoise. The enrollment preferences are based upon a clinical factor – radiographic severity of underlying osteoarthritis – that, for the sake of this simulation, we assume to be associated with the efficacy of surgery. To quantify the consequences of selective enrollment, we estimated bias, empirical power and 'coverage' of 95% confidence intervals.

Design of the simulated trial

Underlying assumptions—We assumed that surgery is considerably more effective than nonoperative therapy in patients with minimal osteoarthritis and symptomatic tear (improvement in WOMAC of 20 points with surgery and 8 points with nonoperative therapy; Table 1), whereas surgery is just slightly more effective than nonoperative therapy in patients with moderate osteoarthritis and symptomatic meniscal tear (improvement in WOMAC of 3 points with surgery and 0 with nonoperative therapy). This reflects the clinical observation that meniscal surgery is most effective in reducing mechanical symptoms and less effective in reducing symptoms due to osteoarthritis.[8] Whether evidence of effect modification is in fact demonstrated awaits larger trials with prespecified subgroup analyses.

Details of the simulated trials—The sample consisted of patients with osteoarthritis of the knee and symptomatic meniscal tear. The extent of underlying osteoarthritis was reflected in the Kellgren-Lawrence radiographic scale, which rates radiographic osteoarthritis as: 0=none; 1=questionable osteophytes; 2=definite osteophytes; 3=definite joint space narrowing with loss of up to 50% of the joint space; 4= > 50% joint space narrowing.[9] Subjects with K-L grade 4 osteoarthritis were excluded from this simulated trial (and from MeTeOR) because they are more appropriately referred for total joint replacement than for arthroscopic meniscectomy. Subjects with normal plain radiographs (K-L grade 0), but osteoarthritis documented on MRI, were eligible.

The primary outcome variable was the Western Ontario McMaster Osteoarthritis Index (WOMAC) pain scale, a measure of lower extremity pain that can vary from 0 (most pain) to 100 (no pain).[10] We calculated the proportion of subjects in each group (surgical and nonoperative) that improved by 10 points or more, considered a clinically important difference.[11]

Key prognostic variables in this setting include the baseline value of the WOMAC score and the extent of osteoarthritis on the baseline radiograph. We based our assumptions of baseline WOMAC pain scores upon reports of preoperative scores in patients undergoing arthroscopic partial meniscectomy.[6] We assumed that subjects with joint space narrowing (KL-3) have somewhat worse baseline score than the other subjects. We also assumed that subjects with more advanced Kellgren-Lawrence grades are likely to have worse outcomes than patients with less severe radiographic findings following either surgery or nonoperative therapy, as noted above. These assumptions are documented in Table 1.[8]

Table 1 demonstrates that the baseline, follow up and change in WOMAC scores differ across the four groups, defined by K-L grade. Further, the extent of improvement in surgical vs. nonoperative therapy also differs across the groups. We assumed that surgery is associated with considerable improvement in patients with minimal osteoarthritis and minor improvement in patients with more advanced osteoarthritis. Nonoperatively treated patients also have somewhat better outcomes if they have less osteoarthritis on their radiographs, but the extent of improvement is much less in the nonoperative group than in the surgical group. This assumption implies effect modification, with the efficacy of surgery, as compared with nonoperative therapy, depending on the extent of osteoarthritis.

Enrollment scenarios

We first compared surgery and nonoperative therapy in the source population on two hundred thousand subjects, half undergoing surgery and half nonoperative therapy, providing an extremely large population to estimate the ‘true’ efficacy of surgery. The distribution of K-L scores in these patients precisely reflects that of the source population:

50% had K-L 0, 10% K-L 1, and 20% each K-L 2 and 3. This reflects the community based prevalence of K-L grades among persons with meniscal tear and knee OA.[12] Thus, the results of this extremely large simulated comparison conducted in the source population are regarded as the true efficacy of surgery.

We then simulated enrollment of subjects into each of four clinical trial enrollment scenarios. These scenarios are documented in Table 2. Each scenario is characterized by a different mixture of subjects from each K-L defined group. The first scenario assumed that the distribution of K-L grades in study participants was the same as that in the source population. We refer to this scenario as 'population based.' It assumes that 50% of subjects are in the KL=0 group, 10% in the KL = 1, and 20% each in the KL = 2 and KL=3 groups. The three alternative scenarios reflect variation across K-L grade distributions: one in which surgeons prefer to enroll patients with less severe osteoarthritis (Scenario 2) and two in which surgeons prefer to enroll patients with more severe osteoarthritis (Scenarios 3 and 4).

Trial simulation

For each of the scenarios noted in Table 1, we simulated 100 trials, providing reasonably stable simulation results. In each trial, we randomly selected 340 subjects with a distribution of K-L scores as specified on Table 2. A random half of these subjects received surgery and half nonoperative therapy, with randomization stratified within K-L defined group. WOMAC scores at baseline and follow up were derived with random sampling from a normal distribution of scores with means as shown in Table 1. The WOMAC mean pain scale scores were calculated for, each simulated trial population, including the proportion of subjects in each treatment arm who achieved a clinically meaningful improvement (10 points) in WOMAC score. For each trial, we calculated a risk ratio representing the likelihood of achieving a clinically meaningful difference with surgery as compared with nonoperative therapy. Thus, this risk ratio is an estimate of the efficacy of arthroscopic partial meniscectomy (APM). The 100 replications of trials with hypothetical study participants randomly selected from source population produced a distribution of efficacy (risk ratio) estimates. We calculated the mean efficacy value across all such trials with 95% confidence intervals. We calculated the proportion of the 100 trials simulated under each of four scenarios in which the 95% confidence intervals around the efficacy estimates contained the population based efficacy value. We called this criteria 'coverage probability'. We also calculated the proportion of the 100 trials in which the 95% CI around the risk ratio did not include 1.0. This latter statistic is an empirical estimate of power. All analyses were performed in SAS (Cary, North Carolina).

Results

Population-based values

We generated a source population and simulated the baseline characteristics and outcomes of two hundred thousand subjects treated with surgery and two hundred thousand treated conservatively. The surgical arm of the simulated trial improved in WOMAC score by a mean of 10.3 points (sd 14.5) and the nonoperative arm improved by 2.6 points (sd 13.5). The difference in means was 7.7 points (sd 14.0), favoring surgical treatment. The operative treatment arm had a greater proportion of patients achieving a clinically meaningful improvement of 10 points or more than the nonoperative cohort (66% vs. 34%, relative risk = 1.95).

Results of simulation studies

Population-based (community equipoise) Scenario—We simulated 100 trials with 340 subjects in each trial. Subjects were chosen from the K-L distribution concordant with the source population (Table 2). The mean increase in relative risk (representing the likelihood of clinically important difference in WOMAC score) was 1.87 (95% CI 1.57, 2.15). In all 100 trials, 95% CI around the relative risk excluded 1.0 (empirical power = 100%) and in all 100 trials the 95% CI around the relative risk included the ‘true’ parameter 1.95 (100% coverage).

Scenarios of Selective enrollment—The findings from simulated trials in which surgeons selectively enroll are shown in Table 3.

Selective enrollment of patients with less severe OA: In this scenario, the risk ratio reflecting the greater likelihood of achieving a meaningful improvement with surgery, as compared with nonoperative therapy, was 1.94 (95% CI 1.68, 2.45). In 100% of these trials the 95% CI around the risk ratio included the true risk ratio and excluded 1.0. Thus empirical power and coverage were both 100%.

Selective enrollment of patients with more severe OA: We simulated two scenarios in which patients with more severe OA were selectively enrolled. The distributions of KL grades in patients under these scenarios are shown in Table 2. In the first, less restrictive scenario, the mean risk ratio was 1.45 (95% 1.15, 1.93). The 95% CI around the risk ratio included the true effect in 46% of the simulations. The 95% CI around the risk ratio excluded 1.0 in 73% of the simulations (empirical power = 73%).

In the last, more restrictive scenario, with highly selective enrollment of persons with advanced disease, the efficacy was even lower, mean RR=1.34 (95% CI 0.93, 2.15). The 95% CI around the risk ratio included the true effect in 25% of the simulations (25% coverage). The mean 95% CI around the efficacy parameter excluded 1.0 in 44% of the simulations (empirical power=44%).

Discussion

We performed a series of simulations to test the hypothesis that when providers use clinical judgment (individual equipoise) as they enroll subjects into randomized trials the results may differ from those observed when providers are guided solely by the entry and exclusion criteria for the trial (community equipoise). We found that, indeed, as provider enrollment behavior departs from community equipoise, the trial results become increasingly distorted. In fact, in the scenario in which clinicians enroll very few patients with mild osteoarthritis (best surgical prognosis), only 44% of the simulated trials show a statistically significant advantage for surgery. Variability in results can be appreciated by the range of efficacy estimates (Table 2), from 1.94 to 1.34.

These findings have important implications for the conduct and interpretation of randomized controlled trials. If providers choose not to enroll patients who ultimately do especially well – or especially poorly - with one of the treatments, the distribution of subjects across prognostic groups will differ from the distribution in the target population. As a result, the conclusions of the trial may differ from and even contradict the conclusions that would be reached if all eligible patients were enrolled. This distortion occurs if there is effect modification (differential efficacy of the intervention across strata). As our results suggest, if providers selectively enroll patients with especially favorable surgical prognosis (mild OA in our example), the trial favors surgical therapy by a greater extent than observed in the base case scenario. If, alternatively, providers selectively enroll patients with less favorable prognoses (more severe OA in our example), the trial may conclude incorrectly that surgery is not useful. We note that if patients themselves selectively refuse enrollment, the same phenomenon would occur. Thus, this issue is not driven solely by provider behavior.

Selective enrollment also influences pre-planned subgroup analyses. Point estimates of intervention efficacy within subgroups are not affected by under- or over-enrollment into these groups. But the power to detect differences across these subgroups may be diminished if particular subgroups are underrepresented.

We are aware of little prior literature that tests the effects of selective enrollment on the validity of RCTs. Warlow et al proposed that in trials involving multiple clinician investigators, the differing preferences of individual investigators may neutralize each other and therefore not distort trial results.[13] This is a strong supposition, especially in smaller trials. A more prudent approach to the problem is to remain suspicious that departures from community equipoise can indeed distort results, even in multi-investigator trials.

RCTs typically raise concerns about generalizability. The pool of eligible patients may differ systematically from the overall population affected by the disease. Our data raise concerns about a more subtle aspect of generalizability: systematic differences between those who are eligible for the trial and those who eventually enroll. These issues are complemented by the effects on trial results that may arise from differences in inclusion and exclusion criteria, which themselves can vary among trials, influencing the particular results obtained.[14]

There is robust debate about the appropriate ethical framework for enrollment in clinical trials. Some advocate the principal of community equipoise – that it is acceptable to randomize any subject who meets the eligibility criteria established by a community of clinical experts. Others argue for individual equipoise, insisting that randomization is ethically justified when the investigator is comfortable with both options for the particular subject. Still others argue that the key ethical principle is non-coercion; that subjects should be permitted to make informed, non-coerced choices about participation in trials.[4, 5, 15, 16] We cannot resolve this debate but suggest that the problem is muted when trial participation is restricted to investigators who routinely randomize across the entire spectrum of eligibility.

The simulation approach used in this study provides a robust methodology for estimating the results of randomized controlled trials under different enrollment scenarios. However, the validity of the simulations hinges on the accuracy of the input data. Our data on the prevalence of symptomatic meniscal tears in osteoarthritis subgroups defined by Kellgren Lawrence score are robust.[12] The key assumption that patients with more severe osteoarthritis have worse surgical outcomes is well established.[8] However, there are no firm data on the differential efficacy of surgery across different KL strata. If in fact the efficacy is uniform across strata, the distortions we demonstrate would not occur.

Better reporting on the extent of selective enrollment would help to estimate the extent that the issues raised in this paper affect RCTs in practice. The magnitude of the effect of selective enrollment will depend upon the efficacy of the intervention, the strength of the interaction between intervention effect; distribution of the prognostic factor and the actual distribution of the prognostic factor. We suggest that trial investigators report on reasons that providers did not offer the trial to eligible patients. These data are not routinely gathered at present nor required in standard reporting of trials,[17] but would enable readers to appreciate the extent of selective enrollment.

The conceptual issues highlighted by this analysis are salient even if some of the assumptions prove to be incorrect when better data become available. We have shown that selective enrollment of subjects into trials across enrollment strata can distort trial results when treatment efficacy differs across these strata. This phenomenon compromises generalizability. We urge randomized trial teams to discuss these issues at length before embarking on a trial and to monitor the success of enrollment with respect to prognostically salient factors. If the trial investigators enroll selectively, trial results may be distorted.

Acknowledgments

Supported by: NIH/NIAMS R01 AR55557, NIH/NIAMS K24 AR02123

References

1. Alderson P. Equipoise as a means of managing uncertainty: personal, communal and proxy. J Med Ethics 1996;22:135–9. [PubMed: 8798934]
2. Chard JA, Lilford RJ. The use of equipoise in clinical trials. Soc Sci Med 1998;47:891–8. [PubMed: 9722109]
3. Freedman B. Equipoise and the ethics of clinical research. N Engl J Med 1987;317:141–5. [PubMed: 3600702]
4. Veatch RM. Indifference of subjects: an alternative to equipoise in randomized clinical trials. Soc Philos Policy 2002;19:295–323. [PubMed: 12678091]
5. Veatch RM. The irrelevance of equipoise. J Med Philos 2007;32:167–83. [PubMed: 17454421]
6. Herrlin S, Hallander M, Wange P, Weidenhielm L, Werner S. Arthroscopic or conservative treatment of degenerative medial meniscal tears: a prospective randomised trial. Knee Surg Sports Traumatol Arthrosc 2007;15:393–401. [PubMed: 17216272]
7. Merchan EC, Galindo E. Arthroscope-guided surgery versus nonoperative treatment for limited degenerative osteoarthritis of the femorotibial joint in patients over 50 years of age: a prospective comparative study. Arthroscopy 1993;9:663–7. [PubMed: 8305102]
8. Meredith DS, Losina E, Mahomed NN, Wright J, Katz JN. Factors predicting functional and radiographic outcomes after arthroscopic partial meniscectomy: a review of the literature. Arthroscopy 2005;21:211–23. [PubMed: 15689872]
9. Guermazi A, Hunter DJ, Roemer FW. Plain radiography and magnetic resonance imaging diagnostics in osteoarthritis: validated staging and scoring. J Bone Joint Surg Am 2009;91 (Suppl 1):54–62. [PubMed: 19182026]
10. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988;15:1833–40. [PubMed: 3068365]
11. Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. Arthritis Rheum 2001;45:384–91. [PubMed: 11501727]
12. Englund M, Guermazi A, Gale D, et al. Incidental meniscal findings on knee MRI in middle-aged and elderly persons. N Engl J Med 2008;359:1108–15. [PubMed: 18784100]
13. Warlow C. Advanced issues in the design and conduct of randomized clinical trials: the bigger the better? Stat Med 2002;21:2797–805. [PubMed: 12325095]
14. Concato J, Horwitz RI. Beyond randomised versus observational studies. Lancet 2004;363:1660–1. [PubMed: 15158623]
15. Marquis D. How to resolve an ethical dilemma concerning randomized clinical trials. N Engl J Med 1999;341:691–3. [PubMed: 10460824]
16. Miller PB, Weijer C. Rehabilitating equipoise. Kennedy Inst Ethics J 2003;13:93–118. [PubMed: 14569997]
17. Freemantle N, Mason JM, Haines A, Eccles MP. CONSORT: an important step toward evidence-based health care. Consolidated Standards of Reporting Trials. Ann Intern Med 1997;126:81–3. [PubMed: 8992927]

Table 1

Distribution of clinical characteristics based on Kellgren-Lawrence radiographic grade.

Kellgren-Lawrence Grade	**0**	**1**	**2**	**3**
Correlation between pre and postoperative WOMAC pain score	0.5	0.5	0.5	0.4
Baseline WOMAC pain score	65	65	65	60
Improvement in WOMAC pain score for surgically treated subjects	20	15	5	3
Improvement in WOMAC pain score for nonoperatively treated subjects	8	5	3	0

WOMAC pain score has theoretical range from 0 (worst pain) and 100 (no pain)

Table 2

Enrollment Scenarios: The four enrollment scenarios are characterized by a different distribution of subjects from each of the Groups defined by Kellgren-Lawrence (KL) score.

Enrollment scenario	**(KL = 0)**	**(KL = 1)**	**(KL = 2)**	**(KL = 3)**
1. Base Case: community equipoise	50%	10%	20%	20%
2. Selective enrollment of subjects with less severe OA	60%	30%	5%	5%
3. Selective enrollment of subjects with more severe OA	10%	10%	30%	50%
4. Highly selective enrollment of subjects with severe OA	5%	5%	20%	70%

See Table 1 for a description of the baseline and follow up WOMAC scores for each KL grade, both for surgically and nonoperatively treated subjects.

Table 3

Summary results of 100 simulated trials performed for each for four enrollment scenarios.

Enrollment scenario	% of simulated trials that cover true effect	% simulated trials showing APM significantly better (empirical power)	Mean RR (95% CI) for detecting clinically important difference
Community equipoise	100%	100%	1.87 (1.57, 2.14)
Selective enrollment of subjects with less severe OA	100%	100%	1.94 (1.68, 2.45)
Selective enrollment of subjects with more severe OA	46%	73%	1.45 (1.15, 1.93)
Highly selective enrollment of subjects with severe OA	25%	44%	1.34 (0.93, 2.15)