Where do the (big) data come from?
… and why it matters

4. Comparing treatments: Limitations of Randomized Clinical Trials

Rod Little

Department of Biostatistics

University of Michigan

Problems with Randomized Clinical Trials

- Randomization does not solve all problems:
  - Not always ethically feasible
  - Inferences only for subjects willing to be randomized; may exclude subjects with strong treatment preferences, compromising external validity
  - A behavioral treatment may be more successful if subjects are allowed to choose, rather than being randomized
  - Noncompliance, missing data undermine randomization, complicate causal inferences
  - The primary outcome should really measure the effectiveness of the treatment….

Measurement: concepts versus measures

- What you want to measure versus what you get to measure
- Kidney function versus serum creatinine levels
- Genetic and social influences versus Race / Education
- Physiologic status versus discharged alive
- Health of immune system versus CD4 counts
- Improved quality of life versus 1 year probability of death
- Measuring the wrong thing can lead to disaster…

Prophylaxis of ventricular achyarrhythmias with intravenous and oral tocainide in patients with and recovering from acute myocardial infarction

Ryden et al. (1980), *American Heart Journal*, 100,6, 1006-1012

In a double-blind placebo controlled study, tocainide {dosage details} was administered to patients with acute myocardial infarction (AMI). Treatment was started as soon as possible following onset of symptoms; the follow-up period was 6 months.

The patient groups consisted of 56 tocainide and 56 placebo patients.

There was no significant effect on the incidence of ventricular fibrillation or symptomatic ventricular tachycardia. The mortality rates were similar and low in both groups.

Tocainide suppressed ventricular arrhythmias, including ventricular tachycardia, both in the acute stage of AMI and during convalescence. Tocainide also suppressed exercise-induced ventricular arrhythmias. Side effects were in general mild or moderate.

Comments on Tocainide and VTs

- Double-blind, placebo controlled
- Modest sample size (56 controls, 56 placebos, reduced to 26 tocainide, 24 placebo)
- Significant reductions in VPCs or VTs in first 24 hours (19% vs 47%, $P < 0.05$); but is this the right measure?
- Differential withdrawals – 22 in T group, 13 in placebo group, because of "failure of therapy" or "side effects". Five in T group developed significant VT.
- No significant differences in exercise-induced arrhythmias, or survival – but no significant differences is not the same as no differences!
- "Because of small n, not possible to conclude that tocainide lacks the ability to prevent VF, symptomatic VT, or sudden death …" {RL: or maybe it makes these worse…}

PRELIMARY REPORT: EFFECT OF TOCAINIDE AND FLECAINIDE ON MORTALITY IN A RANDOMIZED TRIAL OF ARRHYMIA SUPPRESSION AFTER MYOCARDIAL INFARCTION

CAST Trial Investigators

New England Journal of Medicine (1989), 321, 6, 406-412

- Randomized, stratified on center and measures of disease severity
- Initial dose titration
- Balanced on baseline characteristics
- Analysis: Kaplan-Meier survival curve, log-rank test
- Powered to assess differential survival (unlike earlier studies
- DSMB terminated study prematurely because of lower survival in treatment group

Summary survival data

Trial	N/Average Exposure	Control Sample size	Controls Deaths	Treatment Sample Size	Treatment Deaths
Tocainide	112/6 mos	56	5 (8.9%)	56	5 (8.9%)
CAST	1455/ 10 mos (planned 3yrs)	725	22 (3.0%)	730	56 (7.7%)

Surrogate markers

- Gold standard outcome for many clinical trials is survival (in a fixed interval, or the survival curve)
  - Requires long expensive studies when death is rare
  - Includes deaths unrelated to disease (excluding them creates its own set of problems)
- Surrogate markers: intermediate measures thought to allow quicker assessments of treatments:
  - CD4 counts for AIDS, reduced tumor size for cancer, ECG trace for cardiac arhythmias, BP for heart disease, genetic biomarkers
  - Some work, some are a disaster: definition requires careful biology, statistics

Effect modification and external validity

- X_2 is a confounding factor for effect of treatment X_1 on Y if it is not an outcome of treatment, its distribution differs between treatments, and it affects the outcome
  - Confounding is an important issue for *internal validity*
- X_2 is an effect modifier for treatment X_1 on Y if the mean treatment effect changes for different values of X_2. For example, X_2 = Age is an effect modifier if a treatment X_1 is effective when Age is low, ineffective when Age is high
  - Or, statisticians say X_1 and X_2 interact in their effects on Y – there is a 2-way $X_1 * X_2$ interaction.
  - Effect modification undermines *external validity* , since it suggests that treatment effects may vary depending on differences in how participants are recruited into studies

An example

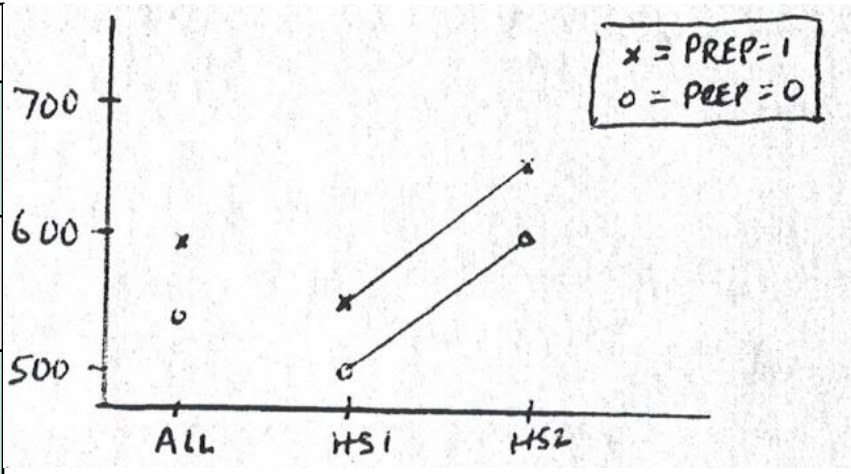
Y = SAT score (the outcome)

X_1 = SAT preparation course (Prep = 1 if taken, 0 if not taken) – the treatment

X_2 = high school (HS = 1 or 2) -- a potential confounding variable

•Table 1: Mean SAT Score (sample size), with no confounding, no effect modification

	PREP=0	PREP=1	ALL
HS=1	500 (240)	550 (120)	517 (360)
HS=2	600 (160)	650 (80)	617 (240)
ALL	540 (400)	590 (200)	557 (600)



PREP effect = 50, HS effect = 100, no effect of adjustment

Adjustment is not needed for bias, though adjustment tends to reduce SE

Example

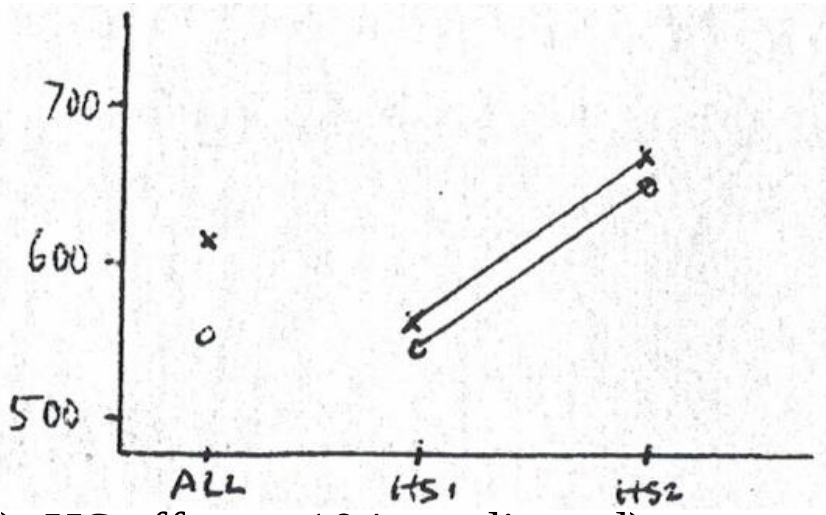
Y = SAT score

X_1 = SAT preparation course (Prep = 1 if taken, 0 if not taken)

X_2 = high school (HS = 1 or 2)

•Table 2: Mean SAT Score (sample size), with confounding, no effect modification

	PREP=0	PREP=1	ALL
HS=1	540 (320)	550 (80)	542 (400)
HS=2	640 (80)	650 (120)	646 (200)
ALL	560 (400)	610 (200)	577 (600)



PREP effect = 50 (unadjusted), 10 (adjusted); HS effect = 104 unadjusted), 100 (adjusted). Unadjusted effect of PREP is an overestimate. Adjustment corrects this bias.

Example

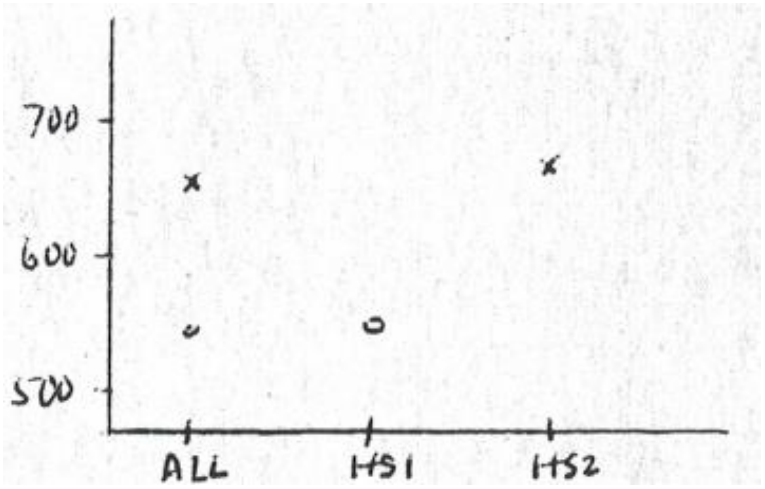
Y = SAT score

X_1 = SAT preparation course (Prep = 1 if taken, 0 if not taken)

X_2 = high school (HS = 1 or 2) a confounding variable

•Table 3: Mean SAT Score (sample size), with X_1 and X_2 completely confounded, no information on effect modification

	PREP=0	PREP=1	ALL
HS=1	540 (400)	? (0)	540 (400)
HS=2	? (0)	650 (200)	650 (200)
ALL	540 (400)	650 (200)	577 (600)



PREP effect = 110 (unadjusted),. Adjusted effects are inestimable since PREP and HS are complete confounded. Data do not allow us to conclude is difference is caused by PREP or HS.

Example

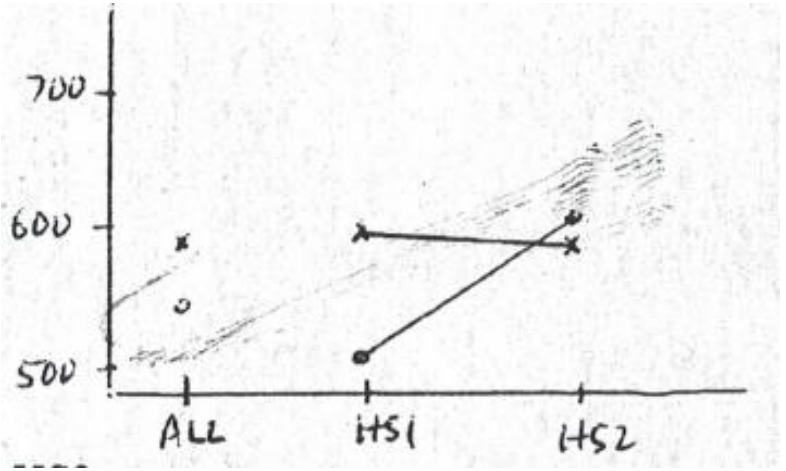
Y = SAT score

X_1 = SAT preparation course (Prep = 1 if taken, 0 if not taken)

X_2 = high school (HS = 1 or 2) a confounding variable

•Table 4: Mean SAT Score (sample size), with no confounding, and effect modification

	PREP=0	PREP=1	ALL
HS=1	500 (240)	590 (120)	530 (360)
HS=2	600 (160)	580 (80)	593 (240)
ALL	540 (400)	586 (200)	555 (600)



PREP effect = 46 (unadjusted), 90 for HS1, -20 for HS2. Overall effect is weighted average of effects for the two high schools. Need to report results by HS for full picture.

Example

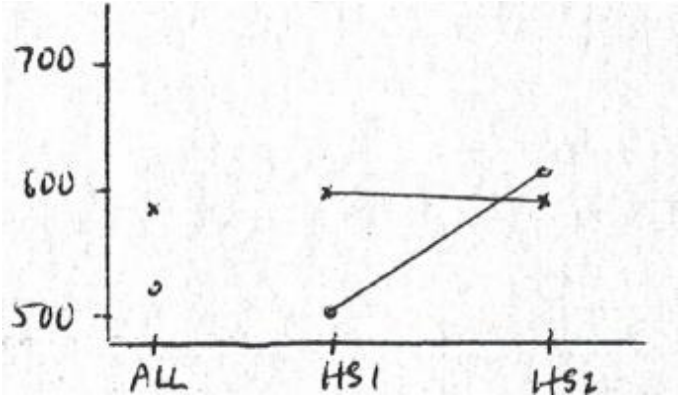
Y = SAT score

X_1 = SAT preparation course (Prep = 1 if taken, 0 if not taken)

X_2 = high school (HS = 1 or 2) a confounding variable

•Table 5: Mean SAT Score (sample size), with confounding and effect association

	PREP=0	PREP=1	ALL
HS= 1	500 (320)	590 (80)	518 (400)
HS= 2	600 (80)	580 (120)	588 (240)
ALL	520 (400)	584 (200)	541 (600)



PREP effect = 64 (unadjusted), 90 for HS1, -20 for HS2. Unadjusted effect or and effect from additive model are misleading. Need to report results by HS for full picture.

Confounding, effect modification and study design

- Randomized Clinical Trials tend to be
  - strong for avoiding confounding, ensuring internal validity
  - weak for detecting effect modification, since sample sizes tend to be small, so power for detecting effect modification is limited
- Clinical data bases / registries tend to be
  - Weak for confounding and internal validity, unless all important confounders are measured
  - Strong for detecting effect modification if confounders are recorded, since sample size tends to be large

Combining RCTs and data bases

- This suggests that RCTs and well-designed clinical data bases tend to have complementary strengths
- In combination, we might gain information for both internal and external validity
  - The RCT protects against confounding
  - If the clinical data base gives comparable estimates to RCT, suggests adequate control of confounders
  - The data base can then inform about potential effect modification, and hence provide information about external validity
  - Note however that RCTs still have a crucial role!

Final Example: Arthroscopic Debridement for Degenerative Knee Joint Disease

N.F. SPRAGUE (1981),Clin. Orthop. 160, 118-123.

SUMMARY

A series of 77 knees in 72 patients, ages ranging from 24 to 78 years (mean, 56 years), with moderate or severe degenerative arthritis were treated by percutaneous debridement of the joint under arthroscopic visualization. Three per cent had a previous meniscectomy, and 81% had a tear of at least one meniscus. Additional pathologic problems included loose bodies in 21%, absent anterior cruciate ligaments in 1396, adhesions in 9% and chondrocalcinosis in 9%.

Sixty-two patients with 68 knees were followed for at least six months, with a mean follow-up of 13.6 months. Subjectively, 84% of the patients were found to have a good or fair result. Complications were few and mild in nature, and there was little morbidity.

Arthroscopic debridement of the knee joint is recommended as a useful therapeutic modality in many patients with degenerative arthritis of the knee.

Sprague (1981)

“Patients were questioned on whether their knees had been improved by the surgery, whether they felt more functional than prior to surgery, and whether they had undergone or were planning additional knee surgery. The results were rated as good, fair or poor (Table 4). A good result was defined as one in which the patient reported that the knee was improved, and that they were equally as functional or more functional than prior to surgery. A fair result was defined as one in which the patient reported some improvement in the knee and was less functional, equally as functional or more functional than prior to surgery.”

Arthroscopic Debridement of the Arthritic Knee

M. Baumgaertner et al., (1990) Clin. Orthop., 252, 197-201

Abstract

Arthroscopic debridement was carried out in 49 knees of 44 patients. These patients, who had a primary diagnosis of arthritis, were older than 50 years of age. Two-thirds had roentgenographic evidence of severe arthritis. Age, weight, compartment location of arthritis, and presurgical range of motion did not affect surgical results. Symptoms of long duration, arthritic severity as evidenced by roentgenograms, and malalignment predicted poor results. Conversely, shorter duration of symptoms, mechanical symptoms, mild to moderate roentgenographic changes, and crystal deposition correlated with improved results.

Surgery offered no benefit for 39% of the patients. Another 9% had temporary improvement, averaging 15 months, but were judged failures at the final follow-up examination. Good or excellent results were achieved in 52% of the patients and maintained through the final follow-up examination in 40% of the patients. Of these, two-thirds had no visible deterioration within a 33-month average follow-up period.

Sprague (1981) and Baumgaertner (1990)

- Two of a number of similar studies reporting successful arthroscopic surgery for knee problems
- SOS Design!
  - No control group
  - Subjective outcomes
  - Regression to the mean? Placebo Effect?

A Controlled Trial of Arthroscopic Surgery for Osteoarthritis of the Knee. Moseley et al. (2002) N. Eng. J. Med. 347, 2, 81-88.

ABSTRACT

Background. Many patients report symptomatic relief after undergoing arthroscopy of the knee for osteoarthritis, but it is unclear how the procedure achieves this result. We conducted a randomized, placebo-controlled trial to evaluate the efficacy of arthroscopy for osteoarthritis of the knee.

Methods. A total of 180 patients with osteoarthritis of the knee were randomly assigned to receive arthroscopic débridement, arthroscopic lavage, or placebo surgery. Patients in the placebo group received skin incisions and underwent a simulated débridement without insertion of the arthroscope. Patients and assessors of outcome were blinded to the treatment group assignment. Outcomes were assessed at multiple points over a 24-month period with the use of five self-reported scores — three on scales for pain and two on scales for function — and one objective test of walking and stair climbing. A total of 165 patients completed the trial.

Moseley et al. (2002)

ABSTRACT

Results. At no point did either of the intervention groups report less pain or better function than the placebo group. For example, mean (±SD) scores on the Knee-Specific Pain Scale (range, 0 to 100, with higher scores indicating more severe pain) were similar in the placebo, lavage, and débridement groups: 48.9±21.9, 54.8±19.8, and 51.7±22.4, respectively, at one year (P=0.14 for the comparison between placebo and lavage; P=0.51 for the comparison between placebo and débridement) and 51.6±23.7, 53.7±23.7, and 51.4± 23.2, respectively, at two years (P=0.64 and P=0.96, respectively). Furthermore, the 95 percent confidence intervals for the differences between the placebo group and theintervention groups exclude any clinically meaningful difference.

Conclusions. In this controlled trial involving patients with osteoarthritis of the knee, the outcomes after arthroscopic lavage or arthroscopic débridement were no better than those after a placebo procedure.

Moseley et al. (2002): surgery cost, mechanism

- When medical therapy fails to relieve the pain of osteoarthritis of the knee, arthroscopic lavage or débridement is often recommended. More than 650,000 such procedures are performed each year at a cost of roughly $5,000 each. In uncontrolled studies of knee arthroscopy for osteoarthritis, about half the patients report relief from pain.
- However, the physiological basis for the pain relief is unclear. There is no evidence that arthroscopy cures or arrests the osteoarthritis. Therefore, we conducted a randomized, placebo-controlled trial to assess the efficacy of arthroscopic surgery of the knee in relieving pain and improving function in patients with osteoarthritis. Both patients and assessors of outcome were blinded to the treatment assignments.

Moseley et al. (2002): ethical issues

- All patients provided informed consent, which included writing in their chart, “On entering this study, I realize that I may receive only placebo surgery. I further realize that this means that I will not have surgery on my knee joint. This placebo surgery will not benefit my knee arthritis.” Of the 324 consecutive patients who met the criteria for inclusion, 144 (44 percent) declined to participate.

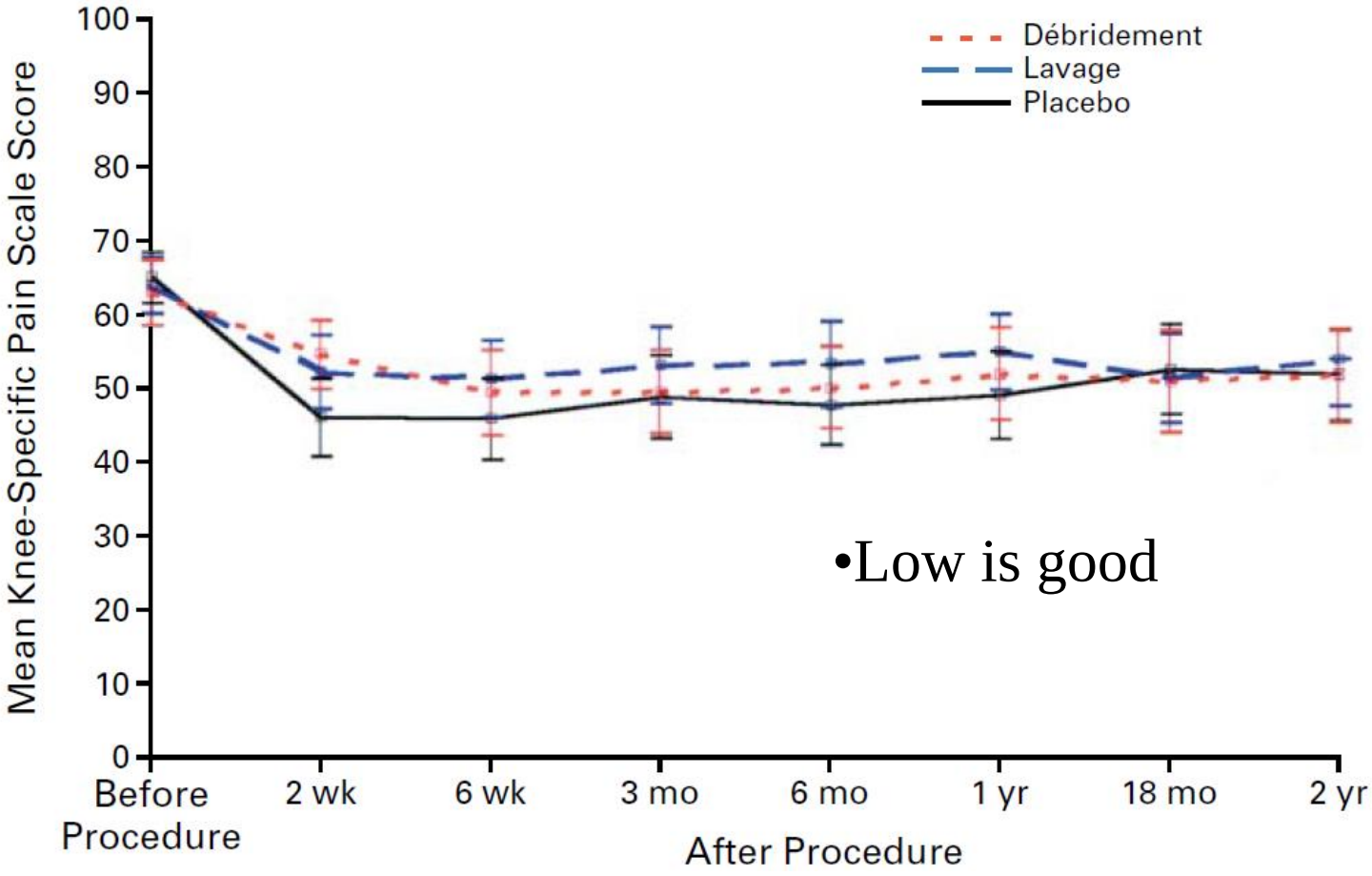
Moseley et al. (2002): stratified randomization

- Participants were stratified into three groups according to the severity of osteoarthritis (grade 1, 2, or 3; grade 4, 5, or 6; and grade 7 or 8). A stratified randomization process with fixed blocks of six was used. Sealed, sequentially numbered, stratum-specific envelopes containing treatment assignments were prepared and given to the research assistant. After the patient was in the operating suite, the surgeon was handed the envelope. The treatment assignment was not revealed to the patient.

Moseley et al. (2002): blinding

- To preserve blinding in the event that patients in the placebo group did not have total amnesia, a standard arthroscopic débridement procedure was simulated. After the knee was prepped and draped, three 1-cm incisions were made in the skin. The surgeon asked for all instruments and manipulated the knee as if arthroscopy were being performed. Saline was splashed to simulate the sounds of lavage. No instrument entered the portals for arthroscopy. The patient was kept in the operating room for the amount of time required for a débridement. Patients spent the night after the procedure in the hospital and were cared for by nurses who were unaware of the treatment-group assignment.

Moseley et al. (2002): results



•Low is good

Moseley et al. (2002): results

DISCUSSION

This study provides strong evidence that arthroscopic lavage with or without débridement is not better than and appears to be equivalent to a placebo procedure in improving knee pain and self-reported function. Indeed, at some points during follow-up, objective function was significantly worse in the débridement group than in the placebo group.

Moseley et al. (2002): surgeon skill

DISCUSSION

One surgeon performed all the procedures in this study. Consequently, his technical proficiency is critical to the generalizability of our findings. Our study surgeon is board-certified, is fellowship-trained in arthroscopy and sports medicine, and has been in practice for 10 years in an academic medical center. He is currently the orthopedic surgeon for a National Basketball Association team and was the physician for the men's and women's U.S. Olympic basketball teams in 1996.

Moseley et al. (2002): external validity

- The principal limitation of this study is that our participants may not be representative of all candidates for arthroscopic treatment of osteoarthritis of the knee. Almost all participants were men, because the study was conducted at a Veterans Affairs medical center. We do not know whether our findings may be generalized to women, although uncontrolled studies do not indicate that there are differences between the sexes in responses to arthroscopic procedures.
- A selection bias might have been introduced by the fact that 44 percent of the eligible patients declined to participate in the study…Patients who agreed to participate might have been so sure that an arthroscopic procedure would help that they were willing to take a one-in-three chance of undergoing the placebo procedure. Such patients might have had higher expectations of benefit or been more susceptible to a placebo effect than those who chose not to participate.

Moseley et al. (2002): final comments

- … This study has also shown the great potential for a placebo effect with surgery… Researchers should reconsider the best ways of testing the efficacy of surgical procedures performed purely for the improvement of symptoms.
- In the debate about placebo-controlled trials of surgery, the critical ethical considerations surround the choice of the placebo. Finally, health care researchers should not underestimate the placebo effect, regardless of its mechanism.

Summary

- REMEMBER:
  - Data come from somewhere …
  - Design matters …
  - When analyzing data, need to be constantly aware of possible biases that might lead to faulty conclusions