Where do the (big) data come from?
… and why it matters
2. Comparing treatments

Rod Little

Department of Biostatistics

University of Michigan

Evidence-based medicine

- The idea that choices between different treatments or behaviors should be based on empirical evidence, rather than opinions of “experts”
- Plausible theories can often be provided for effectiveness of many treatments – see e.g. the Cameron and Pauling arguments for Vitamin C as a treatment of cancer
- While scientific plausibility is important, empirical evidence is key, since “plausible” does not necessarily mean “right”

Data, Data, everywhere!

- We use data to answer public health questions
  - Effectiveness of diets to control weight
  - Relationships between pollutants and health outcomes
- How strong is the evidence?
  - Many studies have conflicting conclusions
  - Design: How were the data collected? What are the strengths and weaknesses of various studies?
  - GIGO (Garbage In, Garbage Out). Clever statistical analysis can't rescue an inherently flawed study.
- Statistical analysis
  - Distinguish real from chance differences.
  - Are real differences "causal" or attributable to other factors (confounders)?

Badly designed studies can do serious harm!

- Vaccines and autism
- “In recent years the antivaccine movement has focused on the claim that vaccines are linked to neurological injury, and specifically to the neurological disorder autism, now referred to as autism spectrum disorder (ASD). However the scientific evidence overwhelmingly shows no correlation between vaccines in general, the MMR vaccine specifically, or thimerosal (a mercury-based preservative) in vaccines with ASD or other neurodevelopmental disorders.” https://www.sciencebasedmedicine.org/reference/vaccines-and-autism/

The source of the vaccine-autism link is this (very poorly designed) study

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

Summary

Background We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods 12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

Findings Onset of behavioural symptoms was associ[ated] by the parents, with measles, mumps, and rub[ella] vaccination in eight of the 12 children, with measl[es] infection in one child, and otitis media in an[other]. All 12 children had intestinal abnormalities, rangi[ng] from lymphoid nodular hyperplasia to aphthoid ul[cer]ation. Histology showed patchy chronic inflam[mation] in 11 children and reactive ileal lymphoid hyperplasia in seven, but no granulomas. Behavioural disorders included autism (nine), disintegrative psychosis (one), and possible postviral or vaccinal encephalitis (one). There were no focal neurological abnormalities and MRI and EEG tests were normal. Abnormal laboratory results were significantly raised urinary methylmalonic acid compared with age-matched controls ($p=0.003$), low haemoglobin in four children, and low serum IgA in four children.

Interpretation We identified associated gastrointestinal disease and developmental regression in a group of previously normal children, which was generally associated in time with possible environmental triggers.

Lancet 1998; **351**: 637–41
See Commentary page

Inflammatory Bowel Disease Study Group, University Departments of Medicine and Histopathology (A J Wakefield FRCS, A Anthony MB, J Linnell PhD, A P Dhillon MRCPath, S E Davies MRCPath) **and the University Departments of Paediatric Gastroenterology** (S H Murch MB, D M Casson MRCP, M Malik MRCP, M A Thomson FRCP, J A Walker-Smith FRCP,), **Child and Adolescent Psychiatry** (M Berelowitz FRCPsych), **Neurology** (P Harvey FRCP), **and Radiology** (A Valentine FRCR), **Royal Free Hospital and School of Medicine, London NW3 2QG, UK**

Correspondence to: Dr A J Wakefield

Introduction

We saw several children who, after a period of apparent normality, lost acquired skills, including communication. They all had gastrointestinal symptoms, including abdominal pain, diarrhoea, and bloating and, in some cases, food intolerance. We describe the clinical findings, and gastrointestinal features of these children.

Patients and methods

12 children, consecutively referred to the department of paediatric gastroenterology with a history of a pervasive developmental disorder with loss of acquired skills and intestinal symptoms (diarrhoea, abdominal pain, bloating and food intolerance), were investigated. All children were admitted to the ward for 1 week, accompanied by their parents.

Clinical investigations

We took histories including details of immunisations and exposure to infectious diseases, and assessed the children. In 11 cases the history was obtained by the senior clinician (JW-S). Neurological and psychiatric assessments were done by consultant staff (PH, MB) with HMS-4 criteria.[1] Developmental history included a review of prospective developmental records from parents, health visitors, and general practitioners. Four children did not undergo psychiatric assessment in hospital; all had been assessed professionally elsewhere, so these assessments were used as the basis for their behavioural diagnosis.

After bowel preparation, ileocolonoscopy was performed by SHM or MAT under sedation with midazolam and pethidine. Paired frozen and formalin-fixed mucosal biopsy samples were taken from the terminal ileum; ascending, transverse, descending, and sigmoid colons, and from the rectum. The procedure was recorded by video or still images, and were compared with images of the previous seven consecutive paediatric colonoscopies (four normal colonoscopies and three on children with ulcerative colitis), in which the physician reported normal appearances in the terminal ileum. Barium follow-through radiography was possible in some cases.

Also under sedation, cerebral magnetic-resonance imaging (MRI), electroencephalography (EEG) including visual, brain stem auditory, and sensory evoked potentials (where compliance made these possible), and lumbar puncture were done.

Laboratory investigations

Thyroid function, serum long-chain fatty acids, and cerebrospinal-fluid lactate were measured to exclude known causes of childhood neurodegenerative disease. Urinary methylmalonic acid was measured in random urine samples from eight of the 12 children and 14 age-matched and sex-matched normal controls, by a modification of a technique described previously.[2] Chromatograms were scanned digitally on computer, to analyse the methylmalonic-acid zones from cases and controls. Urinary methylmalonic-acid concentrations in patients and controls were compared by a two-sample t test. Urinary creatinine was estimated by routine spectrophotometric assay.

Children were screened for antiendomysial antibodies and boys were screened for fragile-X if this had not been done

Conflicting conclusions -- example

- Two articles on treatment of advanced cancer using Vitamin C yield conflicting conclusions:
- Cameron, E. and Pauling, L. (1976). Supplemental ascorbate in the supportive treatment of cancer: prolongation of survival times in terminal human cancer Proc. Natl. Acad. Sci. USA Vol. 73, No. 10, pp. 3685-3689,1976.
- Creagan. E. et al (1979). Failure of High Dose Vitamin C (ascorbic acid) therapy to benefit patients with advanced cancer. New. Eng. J. Med. 301: 687-690, 1979.

Questions

- Cameron & Pauling: large effect of Vit C
- Creagan et al.: no effect of Vit C
- Why do these studies give such different results, and which should we believe?
  - More on this later
- The U.S. Food and Drug Administration (FDA) (and international equivalents) decide when treatments should be approved for widespread use
  - a big responsibility not to sanction treatments that are harmful, or stand in the way of treatments that are beneficial
- Major role of study design

Key concepts

We focus on the following key concepts:

1. Defining a causal effect – the Rubin/Neyman causal model
2. Confounding and internal validity
3. Effect-modification and external validity
4. Alternative study designs and their strengths and weaknesses – in particular, the role of randomization in the assignment of treatments

Goals of Research Design

- Internal validity: are the estimated effects of the treatments valid for the individuals in the study?
  - A crucial component — avoiding bias of all kinds
- External validity/Generalizability: are the estimated effects valid for the target population of to which the treatments are to be applied
  - internal validity is a prerequisite
  - Individuals in a study are usually volunteers, not randomly sampled from the target population -- does that matter?
  - There's a tendency to leap to inference far beyond the targeted population.

When is a treatment effect causal?

- How do we know the improvement is caused by the treatment and not something else?
- This gets to a central question: how do we define a causal effect? Phenomena have multiple causes, often hard to disentangle…
- E.g. what “causes” mass shootings
  - Ready access to guns, lack of gun training, mental health of shooters, etc. etc.

Defining causal effects

- Association is not causation: We are interested in causal effects of treatments/etiologic factors.
  - How do we define a “causal effect”?
- “Rubin Causal Model” – causal effect of treatment for subject is difference in outcome under active treatment and under control.
- Estimation of causal effects is basically a missing data problem: We only get to see the outcome from one treatment, the treatment actually received!
- How the treatments are assigned is a crucial issue – randomization plays a key role in avoiding bias

Numerical example

Y(j) = depression score given treatment j
(high = more depressed)

Subject	Y(A)	Y(B)	Y(A)-Y(B)
1	[1]	6 6	[-5]
2	[3]	12 12	[-9]
3	9 9	[10]	[-1]
4	11 11	[12]	[-1]
Mean	10* [6]	9* [10]	1* [-4]

- Assignment mechanism is confounded:

Sicker (more depressed) subjects got treatment A!

Confounding

- 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*: whether a treatment effect is causal for the individuals in a study.
  - In numerical example, baseline depression is a confounding variable

Assignment mechanism

$$T = \begin{cases} \text{A, if assigned to treatment A} \\ \text{B, if assigned to treatment B} \end{cases}$$

$Y(\text{A}) =$ Outcome if assigned A

$Y(\text{B}) =$ Outcome if assigned B

- Assignment mechanism is called *unconfounded* if

  $T \wedge [Y(\text{A}), Y(\text{B})]$, $\wedge$ = independent

  Otherwise assignment mechanism is confounded
- Average causal effects can be estimated as difference in observed means if *assignment* mechanism is *unconfounded*

$$E[Y \mid T = j] = E[Y(j)]$$

Alternative Designs

- Suppose we have a new treatment, and we want to assess its effectiveness
- (Or: we are interested in whether an environmental factor is causally related to disease)
- Consider alternative designs:
  - "Snake Oil Salesman" (SOS)
  - Other observational designs
  - Randomized Clinical Trial (RCT)

The SOS Design

- Give someone the treatment and see if they get better
- Seems logical
- I call this the “Snake-Oil Salesman” (SOS) design
- Much seen in “before and after” commercials on TV

Need for comparison group

- Why not simply assign the new treatment to everyone in study and see if they improve?
  - Do not observe outcome under “no treatment”
  - Implicitly makes dubious assumption of no change under no treatment
  - Better designs have a comparison group.

Three more problems with SOS

- Selection bias: even if the treatment does nothing, if the outcome is variable, we can cherry-pick the cases where the outcome improved
  - E.g. weight loss on a diet – after the diet starts, some people lose weight, some gain weight, some don't change much. Select the ones that lose weight
  - Investment managers etc.: the ones that flog books on TV are the ones that made money, but it could be they were not smart, just lucky
  - History is written by the winners…
  - see "Fooled by Randomness" by Nassim Taleb

Three more problems with SOS

- Regression to the mean: if the outcome is change in a measure (e.g. depression) and that measure fluctuates naturally, then people who start high on the measure will tend finish lower, and people who start low on the measure will tend to finish higher, without any treatment
- E.g. baseball: after 20 at bats, some players are batting .100 (2 hits and some are batting .600 (12 hits)
- After 200 bats, those batting .100 will in all likelihood end up higher, and those batting .600 will end up lower
- If we select individuals batting .100, and give them a magic “batting snake oil” they’ll surely improve, even though the improvement has nothing to do with the oil

Three more problems with SOS

- Placebo effect: even in the absence of any active ingredient, people report an improvement.
- If a treatment involves an investment, we want to believe the investment has been worthwhile – not throwing time or money down the drain – hence believe the treatment has worked
- Particularly a problem with subjective responses, like pain scores; objective measures are less vulnerable

Case Reports and Case Series

- Similar in nature to the SOS design are reports of unusual medical occurrences or associations:
  - Led to early identification of the AIDS epidemic
  - Useful in identifying unusual clusters of disease
- Hypothesis generating
- Anecdotal; not valid statistical evidence
- Sometimes it's real:
  - Vinyl chloride and liver disease
- Sometimes it's not:
  - Breast implants and scleroderma

Example: Disease clusters

- Newspaper reports that 4 out of 8 pregnant female secretaries in a large office with extended exposure to electromagnetic radiation from computer monitors had spontaneous abortions!
- Causality or coincidence?
- Worrying, but newspaper could be reporting a chance event in the tail of the distribution — what about the thousands of offices where this surprising number of abortions did not occur?
- Need prospective clinical study to avoid selection bias

Cross-sectional Surveys

- Exposure and disease status are assessed at a single survey. For example:
  - Assessing fluoride history and number of dental cavities at a single visit
  - National health and nutrition examination survey (NHANES)
- Such studies often find associations between disease and exposure.
- But, is the association truly causation?
  - E.g., did the exposure precede the disease?
  - E.g., does sedentary lifestyle cause CHD, or do people with developing CHD feel too ill to exercise?

Prospective Observational Studies

- The problems with the SOS design suggest that we need a comparator – a placebo, or an existing treatment
  - Some individuals are assigned the new treatment, and some are assigned the comparator treatment.
- Compare two groups with respect to an appropriate outcome, e.g. five year survival rates, and see which group does better
- BUT: If assignment to treatment/comparator is not random, there may be confounding factors.

What's an observational study?

- Assignment of the treatment or etiological factor is natural and not under the control of the investigator
  - Environmental factors are not randomly assigned
  - Smoking is choice of the study participant
  - Treatments in clinical data bases are assigned by clinicians, not controlled by the researcher
  - Review of historical case records

Confounding in Observational Studies

- Inference from every observational study depends on eliminating bias and adjusting for all confounding factors.
  - Confounding factors: age, gender, income, disease severity, etc. may be correlated with the treatment assignment and predict the outcome
- Analysis methods can (multiple and logistic regression, propensity adjustment) can adjust for observed confounders.
- But unobserved confounders remain a problem

Example: learning health systems

- An administrative health system captures data for 200 patients with a rare disorder – 100 are taking Drug A and 100 drug B. 70 people taking Drug A are “cured” and 30 people taking Drug B are “cured”
- The naïve conclusion is that Drug A is more effective. [Note: this difference too large to be attributable to chance]
- But we can’t conclude that Drug A is better – maybe something other than the effect of the drug – a confounding factor -- explains the difference…
- For valid inference, need to record and adjust for potential confounders in the analysis

Crossover designs

- An approximation to observing outcome under both treatments is achieved in crossover designs
  - Individuals receive both treatments A and B, and outcome is recorded for both.
  - Need to guard against spillover effects by suitable "washout period" between treatments
  - Good when feasible, but only possible for short-term, treatment of chronic conditions
  - Randomizing the order of treatments (A then B or B then A) is a good idea to reduce "order effects".
  - Still short of ideal — conditions under which treatments are given are still not identical.

Case-Control Studies

- Cases with disease are identified; controls are selected from the same population that gave rise to the cases.
- The proportions exposed among cases and controls are compared.
  - E.g., compare the proportion of smokers among lung cancer patients and non-cancer controls.
- An efficient design for rare diseases
  - In a simple random sample, lung cancer cases would be quite rare, so a huge sample size would be needed to make the same comparison.
- Assignment not at random, may be confounded

Selecting Controls

- The hardest and most important design issue. Controls are selected from the population that gave rise to the cases.
- Hospital controls: convenient, cheap
  - Use other patients, without the target disease.
  - Because they are ill, they have been shown to be different from the general population (e.g., more likely to smoke and be heavy drinkers).
- Population controls: the gold standard
  - RDD or canvassing households
- Friend / neighbor / relative controls

Potential Bias in Exposure Ascertainment

- Information from record reviews
  - May have missing or incorrect information
  - Case info may be more completely documented.
- Patient interviews
  - Different response rates in cases and controls
    - Cases may be more willing to participate
  - Recall bias
    - Differential reporting of exposure in cases and controls
    - For long-ago exposures, memory helpers (e.g., concurrent residential history) may be helpful.
    - Make sure the exposure pre-dated the disease

Randomized Clinical Trials

- Random assignment of subjects to treatments yields an unconfounded assignment mechanism
  - Facilitates causal inference.
  - Eliminates selection bias from choosing the “best” patients for the preferred treatment

RCT's vs. Observational Studies

- Randomized clinical trials
  - Assignment is random, hence unconfounded
- Observational studies (e.g., registries)
  - Assignment of treatment is uncontrolled, potentially confounded
  - Easier to conduct
  - Good for hypothesis generation
  - Necessary when randomization cannot be performed

Randomized assignment

- All participants are treated the same, except for the treatment assigned
- Unconfounded assignment mechanism, eliminates observed and unobserved confounding factors
  - including the investigator's conflict of interest in favor of new treatment
  - Blinding to treatment, if feasible, removes potential bias in whether or not participants are included

Blinding / Masking

- Single-blind: The patient does not know which treatment s/he is receiving.
- Double blind: Both patient and investigator do not know the treatment assignment.
- Triple blind: The person analyzing the data is also masked to the treatment assignment.
- The evaluator may be a different person, and blinding of this person is crucial.

Blinded Studies (cont'd)

- Blinding removes or equalizes biases due to patients' desire to please and investigator enthusiasm.
- Logistics:
  - Blinded studies of drugs are simple because placebo pills can usually be made.
  - Blinded studies of surgery vs. medical management are hard, sometimes not possible. (But see later).

Levels of evidence

- Several groups have attempted to provide “levels of evidence” for medical study designs. See for example https://en.wikipedia.org/wiki/Levels_of_evidence
- http://www.ahfsdruginformation.com/levels-of-evidence-rating-system/

Double-blind RCT’s are generally considered the gold standard, when feasible

Article Critique 1

- The following outline serves as a framework for evaluating articles in the public health literature.
- 1. General
  - Experiment or survey?
  - What are the authors seeking to demonstrate? Are they consistent?
- 2. Sample Selection
  - To what population (are/can) their results to be generalized?
  - Biases introduced by selection of cases? (nonresponse, excluded cases)
  - Sample large enough? Sufficient statistical power to detect differences of substantive interest?

Article Critique 2

- 3. Treatment Allocation
  - Sufficient documentation ?
  - What evidence is there that treatment arms are equal except for treatments applied:
    - Randomized allocation of treatments?
    - Stratification?
    - Treatment groups compared on observed factors?
    - Might unobserved factors explain the difference in outcomes?
    - Blinding (masking) (of subjects, treatment administrators, investigators)? Possible? Done?
    - Placebo effect?

Article Critique 3

- 4. Outcome Measures
  - Appropriate?
  - Clearly defined and reproducible?
  - Affect all treatment arms equally?
- 5. Analysis of Results
  - Adequate presentation of data?
  - Appropriate statistical analyses?
  - Arithmetic errors? Do the results look right?
  - Appropriate inferences from the analysis?
  - Balanced conclusions?
- 6. Constructive Criticism

Example: Vitamin C and Cancer

- Two articles on treatment of advanced cancer using Vitamin C yield conflicting conclusions:
- Cameron, E. and Pauling, L. (1976). Supplemental ascorbate in the supportive treatment of cancer: prolongation of survival times in terminal human cancer Proc. Natl. Acad. Sci. USA Vol. 73, No. 10, pp. 3685-3689,1976.
- Creagan. E. et al (1979). Failure of High Dose Vitamin C (ascorbic acid) therapy to benefit patients with advanced cancer. New. Eng. J. Med. 301: 687-690, 1979.

Example: Vitamin C and Cancer

- Cameron and Pauling: not randomized; retrospective chart review
  - Raises doubts about comparability of groups
  - Doubts about equal treatment
- Creagan et al: randomized prospective study
  - Evidence that groups are comparable
  - Blinding reduces chance that groups are treated differently

Cameron & Pauling (1976)

ABSTRACT Ascorbic acid metabolism is associated with a number of mechanisms known to be involved in host resistance to malignant disease. Cancer patients are significantly depleted of ascorbic acid, and in our opinion this demonstrable biochemical characteristic indicates a substantially increased requirement and utilization of this substance to potentiate these various host resistance factors.

The results of a clinical trial are presented in which 100 terminal cancer patients were given supplemental ascorbate as part of their routine management. Their progress is compared to that of 1000 similar patients treated identically, but who received no supplemental ascorbate.

The mean survival time is more than 4.2 times as great for the ascorbate subjects (more than 210 days) as for the controls (50 days) Analysis of the survival-time curves indicates that deaths occur for about 90% of the ascorbate-treated patients at one-third the rate for the controls and that the other 10% have a much greater survival time, averaging more than 20 times that for the controls.

The results clearly indicate that this simple and safe form of medication is of definite value in the treatment of patients with advanced cancer.

Creagan et al. (1979)

ABSTRACT. 150 patients with advanced cancer participated in a controlled double blind study to evaluate the effects of high-dose vitamin C on symptoms and survival.

Patients were divided randomly into a group that received Vitamin C (10 g per day) and one that received a comparatively flavored lactose placebo. 60 evaluable patients received vitamin C and 63 received a placebo.

Both groups were similar in age, sex, type of primary tumor, performance score, tumor grade and previous chemotherapy.

The two groups showed no appreciable difference in changes of symptoms, performance status, appetite and weight. The median survival for all patients was about 7 weeks, and the survival times essentially overlapped.

In this selected group of patients, we were unable to show a therapeutic benefit of high-dose vitamin C.

Kaplan-Meier Survival Curve

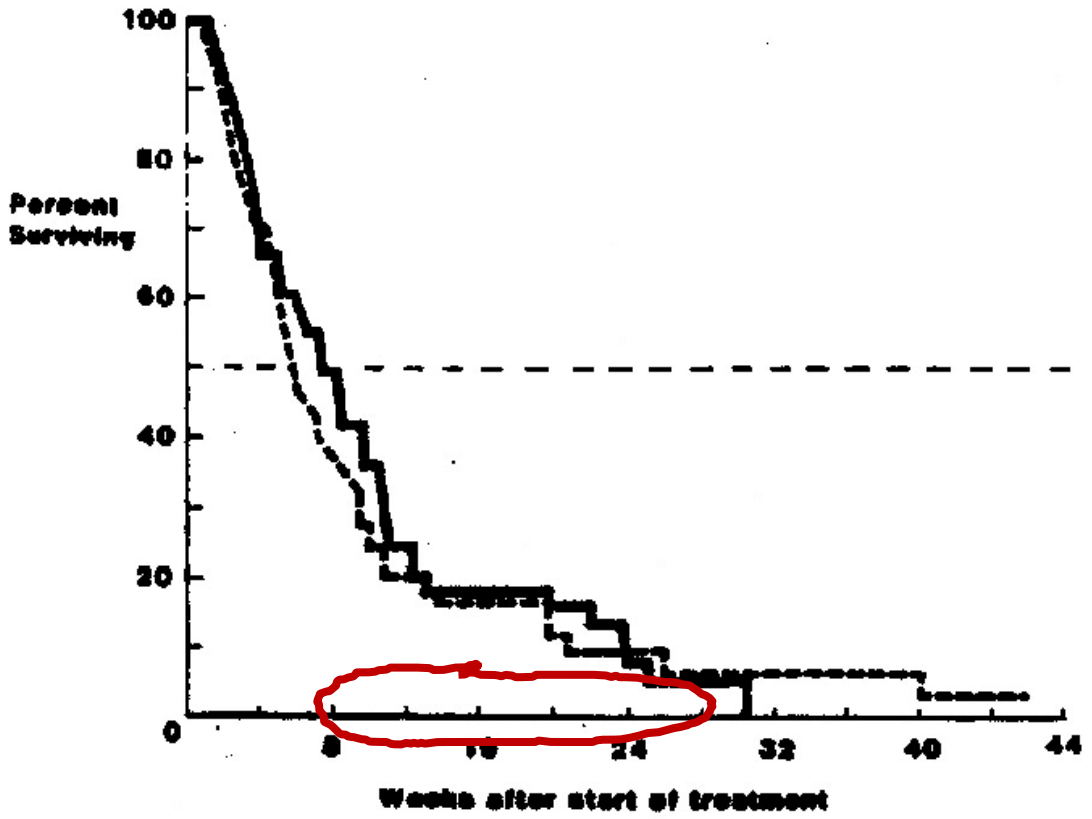


Figure 1. High-Dose Vitamin C versus Placebo and Survival Results in Patients with Advanced Cancer.

The solid line shows survival in 60 patients given vitamin C. The dashed line shows survival in 63 patients given the lactose placebo.

Discussion

- Strengths and weaknesses of Cameron and Pauling?
- Strengths and weaknesses of Creagan et al.?
- Which result do you believe?

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

Measurement: concepts versus measures

- Good measures are :
  - Objective, Verifiable, Meaningful/Clinically relevant
- Example: Alzheimer's Disease
  - Probable AD: clinical observation, neuropsych tests
  - Classification involves subjective elements, potential for inconsistency between clinicians, sites, race of patient, overlap with non-AD dementias
  - Date of diagnosis vs. date of disease onset
  - Confirmed AD: definition based on pathology
  - Authoritative (maybe?), but retrospective, requires autopsy consent
- Next example shows that measuring the wrong thing can be disastrous….

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

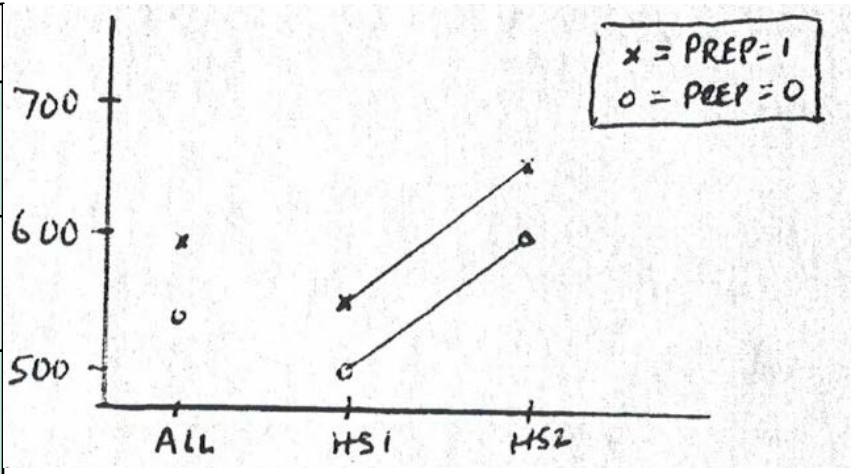
Y = SAT score (the outcome)

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

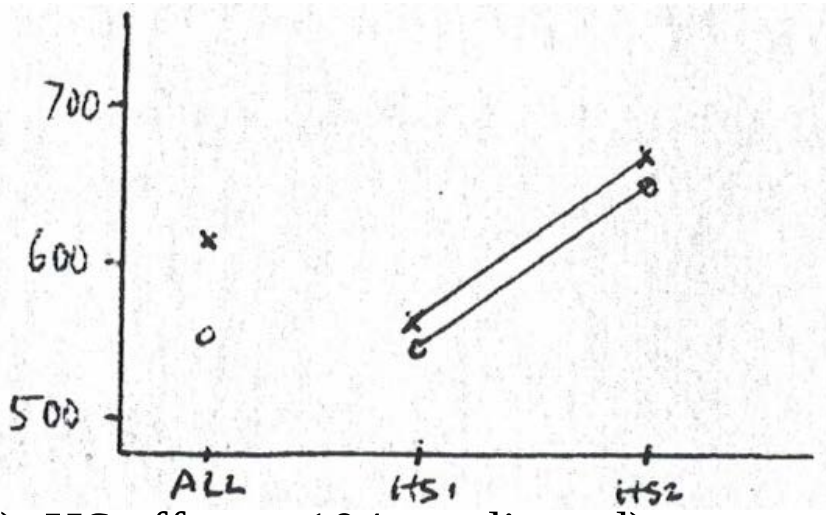
Y = SAT score

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

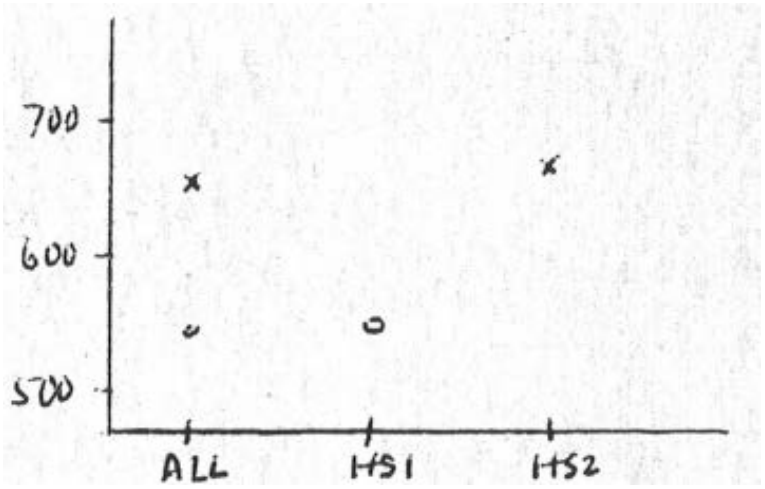
Y = SAT score

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

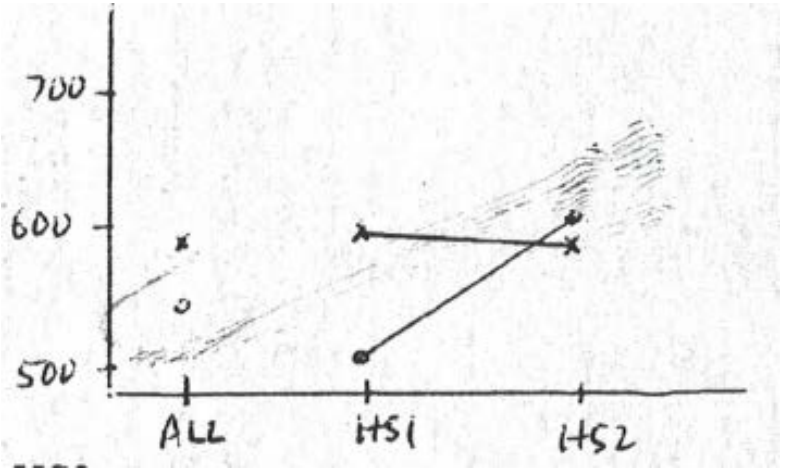
Y = SAT score

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

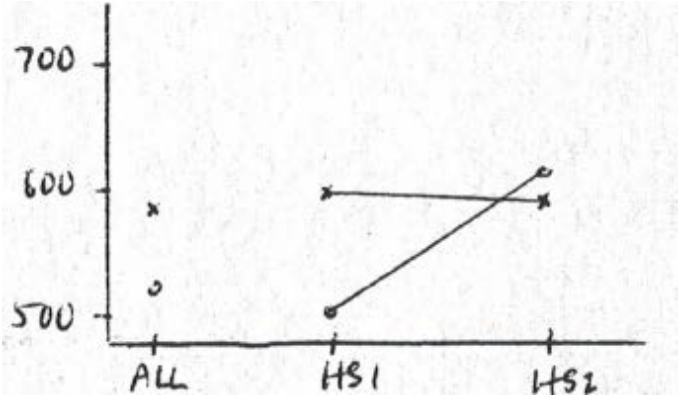
Y = SAT score

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.

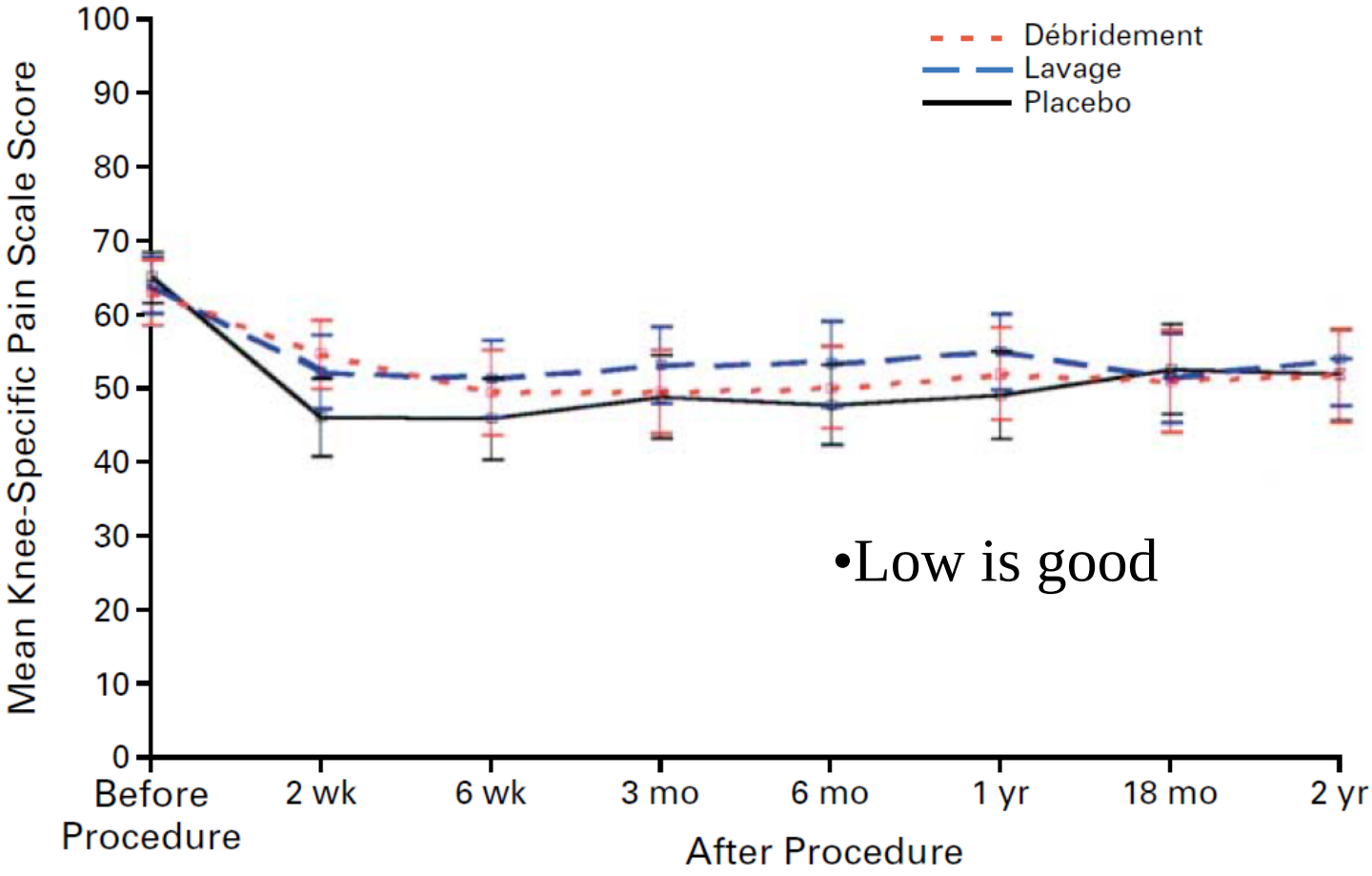
Moseley et al. (2002): stratified randomization

- 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