Society for Clinical Trials 31\textsuperscript{st} Annual Meeting

Workshop P10
Considerations in the Design of Clinical Trials for Comparative Effectiveness Research

Sunday, May 16, 2010
1:00 PM - 5:00 PM
Harborview Ballroom D
WORKSHOP 10 - Considerations in the Design of Clinical Trials for Comparative Effectiveness Research

1. Overall, did the subject context of this workshop meet your expectations and needs?  
   Yes ( )  No ( )  
   If yes, in what way?  If no, why not?  ________________________________

2. Was the content of this workshop of value to you personally or on the Job?  
   Yes ( )  No ( )

3. Was the content of the workshop:  
   New ( )  New/Review ( )  Review ( )

4. The level and complexity of this workshop was:  
   Too elementary ( )  Correct ( )  Too advanced ( )

Please complete the following questions by circling the appropriate description using the rating scale listed below.  
1 = excellent  2 = very good  3 = good  4 = fair  5 = poor

5. Rate the extent to which this workshop:  
   a. Presented content clearly  1 2 3 4 5  
   b. Allowed sufficient time for discussion and audience participation  1 2 3 4 5  
   c. Provided useful information  1 2 3 4 5  
   d. Utilized appropriate teaching methods, i.e., audiovisual, handouts, lectures  1 2 3 4 5

6. Please rate each workshop faculty member:

<table>
<thead>
<tr>
<th>Name</th>
<th>Knowledge of Subject</th>
<th>Organization/Delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huang Grant</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>Walter Koroshetz</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>Michael Lauer</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>Dennis Wallace</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
</tbody>
</table>
1. Are you currently working in a clinical trial? (Yes) (No)

2. What is your job title? __________________________________________________________

3. Do you have any suggested topics for workshops at future meetings? If so, please list below:
   _____________________________________________________________________________
   _____________________________________________________________________________

4. What aspect of the workshop did you like best?
   _____________________________________________________________________________
   _____________________________________________________________________________

5. What aspect of the workshop would you change if this workshop were offered again?
   _____________________________________________________________________________
   _____________________________________________________________________________

6. Additional Comments: _____________________________________________________________________________
   _____________________________________________________________________________

May 16, 2010, 1pm

Title: Considerations in the design of clinical trials for comparative effectiveness research (CER)
Chair: Carmen Rosa (Center for the Clinical Trials Network, National Institute on Drug Abuse)

Agenda

**Introduction**: Michael Lauer, NIH (NHLBI)

**Key factors impacting the design of comparative effectiveness trials**

**Clinical and Practical Considerations to Enhance Structural and Operational Efficiency**

Some of the areas that can be addressed are: (1) participant eligibility (target populations); (2) intervention and control groups flexibility and expertise (using local sites staff and usual treatment vs. highly trained research staff or placebo); (3) choice of primary outcome measure; (4) economic evaluation (not usually done in traditional RCTs).

**Statistical Considerations**: RCT designed for Decision Makers:

This segment will address key issues that the statistician must address in working with other researchers to design and implement CER designs. The session will specifically discuss subtle ways in which research questions and outcome measures may be distinct for CER clinical trials and discuss how extensions to traditional/classic subject-randomized designs may provide advantages for CER trials. Examples of such designs that will be discussed are cluster-randomized designs, Bayesian designs and adaptive designs.

**Case Studies: Small group break outs** will provide participants a hands-on exercise in design.

**Discussion** of case studies: designs and suggestions to improve the application of the key factors

**Closing remarks**: what is next? Michael Lauer, NIH (NHLBI)
Comparative Effectiveness Research and Clinical Trials

Michael S Lauer, MD, FACC, FAHA
Director, Division of Cardiovascular Sciences
NHLBI/NIH
May 16, 2010

Disclosure: None

Disclosures

My immediate family and I have NOT received anything of value related to the technologies and topics being presented. I will present my views, which are not necessarily those of NHLBI/NIH/DHHS. I am neither an economist nor a politician.

Just Over One Year Ago...
ARRA and CER

Allocation: $1.1 billion
- NIH: $400 million
- AHRQ: $300 million
- HHS: $400 million

Federal coordinating counsel for CER
- Advice on CER Federal infrastructure needs
- Federal officials including AHRQ, NIH, VA

IOM issued report on priorities June 2009

So What is CER?

C = Comparative
- Real contest
- Existing options for clinicians or systems

Effectiveness = Outcomes
- Clinical outcomes: mortality, morbidity, major clinical events, costs
- Systems outcomes: adherence to guidelines

Research = Science
- Observations, experiments, syntheses

What to Do?

http://www.opencongress.org/bill/111-h1/text

Thanks to David Sackett for inspiring this content
Debates about Comparative Effectiveness

Come down to the contest…

“Come down to the contest ye Humorists: Let us take out of the Hospitals or the camps or elsewhere, 200, or 500 poor People, that have Fevers etc. Let us divide them in Halves, let us cast lots, that one half of them may fall to my share and the other to yours; I will cure them without bloodletting…; but do you do as ye know. We shall see how many Funerals both of us shall have: But let the reward of the contention or wager, be 300 Florens, deposited on both sides: Here your business is decided.”

Van Helmont JA. Oriatike. London: Lodovick-Cloyd, 1662, p.526

Chalmers I. Int J Epidemiol 2001;30:1156-64
The Trial is Reported…

“It had been so arranged, that this number was admitted, alternately, in such a manner that each of us had one third of the whole. The sick were indiscriminately received, and were attended as nearly as possible with the same care and accommodated with the same comforts.

Neither Mr. Anderson nor I ever once employed the lancet. He lost two, I four cases; whilst out of the other third [treated with bloodletting by the third surgeon] thirty five patients died.”

Over 100 Years Later…

“During the last decades we have certainly bled too little.”

William Osler, MD

Modern Examples of Bloodletting

Thalidomide (birth defects)
Hormone replacement (cancer, strokes)
Oxygen for premature infants (blindness)
Anti-arrhythmic drugs (higher death rate)
Bone marrow transplantation for breast cancer (higher death rate)
PSA for prostate cancer (over-diagnosis)

Thank you to Andrew Epstein
Why Do We Need CER?

"Only a limited amount of evidence is available about which treatments work best for which patients and whether the added benefits of more-effective but more-expensive services are sufficient to warrant their added costs—yet current practice tends to adopt more-expensive treatments even in the absence of rigorous assessments of their impacts...."

Peter Orszag

Is This So?

Scientific Evidence Underlying the ACC/AHA Clinical Practice Guidelines

Nearly 50% of recommendations are based on expert opinion. Only 11% are based on multiple randomized trials.

Illustrative Story: AMI and the Occluded Artery

http://www.circulation.or.kr/info/case/200904/fig1.gif

http://www.indiastudychannel.com/attachments/Resources/82666-221135-Coronary%20Angiogram.jpg
The Real Story

When researchers tried to organize a randomized study of the benefits of angioplasty for patients who had suffered a heart attack three days or more before, they ran into a problem. Many doctors were so convinced of the value of this procedure...that they thought it would be unethical to assign any patients to the control group, which would get all the best medicines for this condition but not the artery-reopening procedure.

But the researchers persisted, with heavy support from the National Heart, Lung, and Blood Institute. After four years of work examining 2,166 patients, they came to an unexpected conclusion....

The Findings

What is Comparative Effectiveness Research?

CBO: “...a rigorous evaluation of the impact of different options that are available for treating a medical condition...

- ...may compare similar treatments, such as competing drugs— or analyze different approaches
- ...may focus on medical risks/benefits, or weigh costs.”

IOM: “The purpose is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels.”

Debate about Randomized Trials

“As currently designed and conducted, many RCTs are ill suited to meet the evidentiary needs implicit in the IOM definition of CER: comparison of interventions among patients in typical patient care settings with decisions tailored to individual patient needs. Without major changes...the nation risks spending large sums of money ... to answer the wrong questions.”
Pragmatic – Explanatory Continuum

**Pragmatic**
- Broad eligibility
- Flexible interventions
- Typical practitioners
- "Usual care" comparison
- No follow-up visits
- Objective clinical outcome
- Usual compliance
- Intent-to-treat

**Explanatory**
- Narrow eligibility
- Strict instructions
- Expert practitioners
- Placebo/other comparator
- Frequent follow-up visits
- Surrogate outcomes
- Close monitoring
- ITT plus per protocol

---

**PRECIS Tool**

---

**PRECIS Tool**
Final Thoughts to Initiate Our Dialogue

Why CER?
- True choices
- Variable practices with similar outcomes
- Adoption without evidence
- Research to directly inform practice and policy

Practical trials
- Real patients
- Real settings ➔ Real discipline!
- Real outcomes
Congressional Budget Office definition for CER

“As applied in the health care sector, an analysis of Comparative Effectiveness is simply a rigorous evaluation of the impact of different options that are available for treating a given medical condition for a particular set of patients. Such a study may compare similar treatments, such as competing drugs, or it may analyze very different approaches, such as surgery and drug therapy. The analysis may focus only on the relative medical benefits and risks of each option, or it may also weigh both the costs and the benefits of those options. In some cases, a given treatment may prove to be more effective clinically or more cost-effective for a broad range of patients, but frequently a key issue is determining which specific types of patients would benefit most from it.”

Why are we here?
Projected Federal Spending on Medicare, as share of GDP
Can Comparative Effectiveness Research live up to it’s promise?

Regional variations in spending
What have we learned?
What do higher spending regions get?
Lower Quality?

Health Expenditures Unsustainable
Cost crisis has been inexorably linked to Comparative Effectiveness research (CER).
- Assumption is that we are paying for items that are not as effective as less costly alternatives and research will guide use of resources.
- In fact the CER linked to cost is Cost-Effectiveness-‐Research measuring the effect of an intervention in relation to the resources it consumes. Is it worth it?

Efficacy and Effectiveness
• Efficacy research is generally thought of as looking at “what can work”; i.e., is a treatment safe and does it have a real effect?
• Effectiveness research, looks at what works “for whom and under what conditions.” Requires large numbers, “real world setting”, clinically important outcomes, accounting for co-morbidities.
• Efficacy research also generally involves comparison of an active agent to a placebo, whereas comparative effectiveness studies

[Diagram and figures are not transcribed as they are not legible or relevant to the text.]
The trillion dollar questions

- How to collect evidence on value and then incorporate this evidence into decisions on coverage, reimbursement, and payment for healthcare services?
- How to develop value-based, cost effective healthcare that is trusted and not perceived as only cost cutting for profit/balancing federal budget.
- Lack of evidence is a real impediment to value-based healthcare. Collecting this data will take time. Policy decisions based on incomplete data is subject to serious negative consequences. Half truth sometimes more dangerous than nothing.

Major Gaps in CER

- Coordination across the CER framework
- Limited CER works across the Federal Government, but coordination is necessary to capture full value
- Research
  - Study comparison, patient-centered research questions remain unanswered
- Bureaucracy and Scientific Capital
  - Bureaucracy exists for conducting CER
- Clinical Relevance
  - Information and data
    - Data must be in terms of clinical attributes of data and longitudinal data capture
    - Data capture and feedback loop amongst of care often lacking
- Dissemination and Translation of CER
  - Suboptimal dissemination and translation of CER findings to patients and clinicians
  - Limited links between CER findings and directly improving patient outcomes
- Priority populations
  - Limited information on minority populations and subgroups
- Private Interest
  - Less information as certain comparative assessments such as behavioral change, procedures, devices, delivery certain examples, and prevention

Data Wars
Real Targets

• Overuse, Misuse and Underuse
  – Appropriate use requires evidence
• Danger is decision making in absence of outcome evidence. Cost will drive decision making.
  Performance measure substitute for outcomes.
  – Difficult to design a system without complete body of evidence and consensus on all of medicine.
• Difficult to accommodate a more expensive, more effective, new products in a system that is cost-driven. True health outcome data is key (QALYs).

National Institute for Health and Clinical Excellence (NICE)

• “An independent organization responsible for providing national guidance on promoting good health and preventing and treating ill health”.
• Secretary of State directs that NHS provide funds for medicines recommended by NICE appraisals usually within 3 months.
  • http://www.nice.org.uk/

National Institute for Health and Clinical Excellence

• Advisory to the British National Health System
• Produces “guidance” documents after consultation with practitioners and review of data, primarily published data but also pharma data. Pharma submits economic evaluations. Includes economic model-cost/QALY
Transforming medicine and health through Comparative Effectiveness Research

Long and Continuing Tradition

Sample NIH CER Projects

- Drug versus drug
- Surgery versus medical
- Lifestyle versus medical
- Surgery versus surgery
- Screening versus usual care
- Observational analyses based on EHR
Types of CER investments and activities can be grouped into four major categories:

- Research (e.g., comparing medicines for a specific condition or discharge process A to discharge process B for readmissions)
- Human and Scientific Capital (e.g., training new researchers to conduct CER, developing CER methodology)
- CER Data Infrastructure (e.g., developing a distributed practice-based data network, longitudinal linked administrative or Electronic Health Record (EHR) databases, or patient registries)
- Dissemination and Translation of CER (e.g., building tools and methods to disseminate CER findings to clinicians and patients and translate CER into practice)

Personalized medicine and patient subgroups

- Understanding the variability that underlies who benefits and who does not benefit from a given treatment.
- Limit the number of persons treated with an intervention to which they don’t respond, but which best “on average”

Drug versus Drug: CATIE

![Graph](image-url)
NIH CER Coordinating Committee Charge

Provide advice to NIH Director on:

- Priorities for ARRA CER funds
- Implementation of CER rules and definitions
- Optimal collaboration with AHRQ and other agencies
- NIH CER portfolio analysis
- CER communication and dissemination
- Long-term CER efforts

NIH-AHRQ CER Workgroup

Charge:

- Coordinate dialogue with AHRQ
- Report to NIH CER Coordinating Committee

AHRQ Members

- Jean Slutsky
- Yen-pin Chiang
- Lia Hotchkiss

NIH Members

- Michael Lauer (NHLBI): Chair
- Richard Suzman (NIA)
- Philip Wang (NIMH)
- Nancy Miller (OSP)
- Lynn Hudson (OSP)

The NIH is Fully Committed to CER

Our goals:

- Work closely with our DHHS colleagues
- Involve our scientific community, practitioners, consumers, industry, policymakers, IOM and other stakeholders
- Demonstrate our value to the public
- Work fast while ensuring accountability, transparency, and accessibility
- Develop infrastructure platforms to enhance CER capacity
Comparative Effectiveness Research at AHRQ

- Created in 2005, authorized by Section 1013 of the Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003
- AHRQ shall conduct and support research on:
  - “the outcomes, comparative clinical effectiveness, and appropriateness of health care items and services (including prescription drugs)”
- Goal: to provide patients, clinicians and policy makers with reliable, evidence-based healthcare information

Center for Medicare and Medicaid Services

- Operates under regulations on coverage.
  - Require payment not be made for interventions that are not "reasonable and necessary"
  - Default has been to provide coverage if there is no evidence of harm.
  - Has regulation to allow it in limited circumstances to cover only the "least costly alternative" for durable medical equipment and IV drugs.
- Reimbursement based on RVU, not clinical benefit.
- Would need new legal authority to apply CER principles to coverage decisions.
- Fee for service model incentivizes "capacity" expansion.

Coverage with Evidence Development

- Links payment to requirement for prospective data collection
- Intent is to guide clinical research to address questions of interest to Medicare
  - Medicare must approve study design
- Goal to support evidence and innovation
  - Lower evidence threshold with commitment to generate better information later
Need for pragmatic trials. The Pragmatic Randomized Control Trial Workshop was held on March 31st through April 2nd, 2008 at the University of Toronto Conference Centre.

- There is an unmet need for pragmatic clinical trials because of misaligned incentives:
  - Patients are interested in access, physicians do not like to admit uncertainty,
  - product manufacturers are only interested in meeting regulatory requirements,
  - not all decision makers are concerned with the same gaps in evidence.

Pragmatic Clinical Trial Characteristics

- The design of a pragmatic trial reflects variations between patients that occur in real clinical practice and aims to inform choices between treatments. To ensure generalisability pragmatic trials should, so far as possible, represent the patients to whom the treatment will be applied. The need for purchasers and providers of health care to use evidence from trials in policy decisions has increased the focus on pragmatic trials.
- Outcome measures differ between explanatory and pragmatic approaches. In explanatory trials intermediate outcomes are often used, which may relate to understanding the biological basis of the response to the treatment—for example, a reduction in blood pressure. In pragmatic trials they should represent the full range of health gains—for example, a reduction in stroke and improvement in quality of life.

Pragmatic vs. Explanatory Trials

- Pragmatic clinical trials are those prospective trials designed to address the evidence needs of healthcare decision makers (payers, patients, clinicians, and policy-makers). At the opposite end of the spectrum are explanatory trials which are designed to address scientific questions.
Assessing degree of “pragmaticism”
Scott Tunis

Research Methods & Reporting
Improving the reporting of pragmatic trials: an extension of the CONSORT statement
BMJ 2008;337:a2390

<table>
<thead>
<tr>
<th>Question</th>
<th>Efficacy—can the intervention work?</th>
<th>Effectiveness—does the intervention work when used in normal practice?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setting</td>
<td>Well resourced, “ideal” setting</td>
<td>Normal practice</td>
</tr>
<tr>
<td>Participants</td>
<td>Highly selected. Poorly adherent participants and those with conditions which might obscure the effect are often excluded</td>
<td>Little or no selection beyond the clinical indication of interest</td>
</tr>
<tr>
<td>Intervention</td>
<td>Strictly enforced and adherence to protocol is monitored closely</td>
<td>Applied flexibly as it would be in normal practice</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Often short term surrogates or process measures</td>
<td>Directly relevant to participants, funders, communities, and healthcare practitioners</td>
</tr>
<tr>
<td>Relevance to practice</td>
<td>Trial is designed to meet needs of those making decisions about treatment options in setting in which intervention will be implemented</td>
<td>Trial is designed to meet needs of those making decisions about treatment options in setting in which intervention will be implemented</td>
</tr>
</tbody>
</table>

Efficacy vs. Effectiveness
North American Symptomatic Carotid Endarterectomy Trial (NASCET)

| Participants | Asymptomatic patients selected for carotid stenosis severity. Exclusions included mental incompetence or other diseases likely to have been fatal within 5 years. Patients were temporarily ineligible if they had any of seven transient medical conditions (e.g., uncontrolled hypertension or diabetes). Anyone aged 50-69 with 50% or greater stenosis of one carotid artery. Exclusions included left atrial thrombus or other serious medical conditions. |
|-------------|-------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| Intervention | Surgeons had to be approved by an expert panel, and were restricted to those who had performed more than 50 carotid endarterectomies with a post-operative complication rate (stroke or death within 30 days) of less than 6%. Acupuncturists determined the content and number of treatments according to patients' needs. Residen determins the content and number of treatments according to patients' needs. |
| Outcomes    | The primary outcome was time to ipsilateral stroke, the outcome most likely to be affected by carotid endarterectomy. Secondary outcomes: all strokes, major strokes, and death. Primary outcome was mortality or non-fatal stroke. | Primary outcome was mortality or non-fatal stroke. |
| Relevance to practice | Patients and clinicians are highly selected and it isn't clear how widely applicable the results are. General practitioners and patients can immediately use the trial results in their decision making. | Patients and clinicians are highly selected and it isn't clear how widely applicable the results are. General practitioners and patients can immediately use the trial results in their decision making. |
Design of Studies of diagnostic tests.

- Estimates of precision around test accuracy are often lacking.
- The performance of test in practice may be misjudged (overestimated) by studies with too many patients who are disease free or too many with disease.
- Selection bias- if reluctance to enroll in a study due to invasive nature of the confirmatory "gold standard" test, then likely will end with patients with high pre-test probability of disease. Overestimate accuracy of the test.
- Verification bias-preferential referral for verificatio of dx. Using more invasive gold std such that verification is incomplete- lead to overestimation if those with low pretest probability are not verified.
- Unblinded reading.

First documented CT?

- The Old Testament book of Daniel records the world's first-ever controlled intervention study. It had a clear purpose, a concurrent comparison group, a pre-defined clinical endpoint and a useful conclusion, despite lacking a written protocol, treatment group blinding and having inadequate duration (10 days) and group sample sizes (four versus an unstated number):
- Then Daniel said to the steward whom the chief of the eunuchs had appointed over Daniel, Hananiah, Mishael, and Azariah: 'Test your servants for ten days; let us be given vegetables to eat and water to drink. Then let our appearance and the appearance of the youths who eat the kings rich food be observed by you, and according to what you see deal with your servants'. ...At the end of ten days it was seen that they were better in appearance ... than all the youths who ate the king's rich food. So the steward took away their rich food and the wine they were to drink, and gave them vegetables. Daniel 1:11-16.

First RCT in Medicine

  - Sample size calculation relying on specifying a supposed clinically relevant difference in treatment effects, and type I and type II error rates (with 0.05 set almost universally as the level of significance and 80–90 per cent as the power
The Bayesian Statistic: An approach fitted to the clinic.
Meyer, Vinicius Goichot. La Revue de Medicine Interne; Volume 30, Issue 3, March 2009, Pages 242-249

* Bayes' theorem on which this paradigm relies is frequently used by the clinicians. There is a direct link between the routine diagnostic test and the Bayesian statistic. This link is the Bayes' theorem which allows one to compute positive and negative predictive values of a test. The principle of this theorem is extended to simple statistical situations as an introduction to Bayesian statistic. The conceptual simplicity of Bayesian statistic should make for a greater acceptance in the biomedical world.

Practical Bayesian adaptive randomisation in clinical trials.

* While randomisation is the established method for obtaining scientifically valid treatment comparisons in clinical trials, it sometimes is at odds with what physicians feel is good medical practice. If a physician favours one treatment over another based on personal experience or published data, it may be more appropriate ethically for that physician to use the favoured treatment, rather than enrolling patients on a randomised trial. Still, the randomised trial may later show the physician's favoured treatment to be inferior. This paper reviews a statistical method, Bayesian adaptive randomisation, that provides a practical compromise between the scientific ideal of conventional randomisation and choosing each patient's treatment based on a personal preference that may prove to be incorrect.

Practical Bayesian Adaptive Randomization in Clinical Trials
Thall PF, Walthen JK. Eur J Cancer 2007 43:859

* An alternative statistical method for comparing treatments that provides a practical compromise between the scientific ideal of conventional randomization, which essentially bases treatment selection on a coin flip, and choosing the patient's treatment based on a personal preference that may turn out to be wrong.
  - The purpose of randomization is to ensure that whatever the unknown latent variables may be, on average their effects will be the same in the two treatment arms.
  - Comparing effects of two interventions tested in different trials is problematic because the inter trial differences are often larger than the treatment effect.
Practical Bayesian Adaptive Randomization in Clinical Trials
Hall, Kyle Eur J Cancer 2007 43:859

- Frequentist – the unknown parameters are fixed
- Bayesian - parameters are random
  - Includes a prior probability distribution (prior(θ))
  - Computes a posterior probability distribution
    - Likelihood (data|θ) x prior(θ)/ prob (data)
    - Uses data to convert a prior to a posterior
    - Used repeatedly so that posterior at one stage used as prior for next stage

Bayesian adaptive randomization
Hall, Kyle Eur J Cancer 2007 43:859

- Uses current data to compute randomization probability for each patient at time of enrollment.
- Helpful to simulate the trials behavior under variety of possibilities prior to start.
- Involves a trade-off between the variability in the estimator of the comparative treatment effect for the ethical desirability of treating as few patients as possible with the inferior treatment.
Bayesian Adaptive Randomization

• Changing characteristics of the patients enrolled over time (drift) can be a major confounder. Include known covariates in the model.

Conventional vs. Bayesian Randomization. A means to increase participation in clinical research?

• "Ms. B, I have two possible treatments for your cancer, A and B, but I do not know which is better. So I would like to enroll you in a clinical trial aimed at comparing these treatments to each other. If you agree to enter the trial, your treatment will be chosen randomly by a computer, based on the data that we have so far on how well these two treatments have done with previous patients in the trial.
• Or
• Ms B, I have two possible treatments for your cancer, A and B, but I do not know which is better. So I would like to enroll you in a clinical trial aimed at comparing these treatments to each other. If you agree to enter the trial, your treatment will be chosen by flipping a coin." Hall, Kyle Eur J Cancer 2007 43:859

• The methodology is for applications having two or three treatments with binary responses, best suited to the case of outcomes occurring rapidly relative to the patient accrual rate and when the disease involved is serious or life threatening.

Implementing a decision-theoretic design in clinical trials: why and how?


• Perhaps unsurprisingly, it takes fewer patients to come to a conclusion that treatment A is better than treatment B than it does to estimate the difference in their success rates with high precision. Pragmatists argue that it does not matter, other things being equal, if the success rate on A outperforms B by 0.20 or 0.02, merely that B is not as effective as A.
• Objections to the subjective nature of prior distributions involved in a Bayesian framework can be alleviated, and in parallel a claim of robustness made,...by implementing a design that has desirable properties across a range of priors and that will not lead to unduly early stopping. Appeal instead to a notion of 'worstcase scenario'.
Implementing a decision-theoretic design in clinical trials: why and how?


- Clinical trial design as put into practice has not changed fundamentally since Fisher introduced randomization in the 1920s into agricultural field studies, nor since Bradford Hill translated this defining feature into medical research some two decades thereafter. However, a wealth of theory, much ethically motivated, has been developed subsequently, yet is not routinely implemented in today's clinical research.
- Medical progress is based on research which ultimately must rest in part on experimentation involving human subjects. ... In medical research on human subjects, considerations related to the well-being of the human subject should take precedence over the interests of science and society. Research on humans, then, is necessary, but matters of individual ethics are to be given higher priority than matters of collective (or research) ethics.

The value of information and optimal clinical trial design


- Type I error, the probability of rejecting the null hypothesis of no difference when it is true, is most often set to 0.05, regardless of the cost of such an error. In addition, the traditional use of 0.2 for the type II error means that the money and effort spent on the trial will be wasted 20 per cent of the time, even when the true treatment difference is equal to the smallest clinically important one and, again, will not reflect the cost of making such an error.
- An effectiveness trial (otherwise known as a pragmatic trial or management trial) is essentially an effort to inform decision-making, i.e. should treatment be adopted over standard and sample size should be determined to maximize the difference between the cost of doing the trial and the value of the information gained from the results.
  - Or: Which of two wavelengths of light to treat macular degeneration vs. C section vs. vaginal delivery in breech babies. If really both treatments the same then not much hazard in Type I error in one but big cost for the other.

Cluster Randomized Trials:

- Examination of interventions at the level of the health service organizational unit or geography.
- Randomize groups as opposed to individuals
- Intervention may be designed to be delivered to groups.
- Intervention requires health professionals to change behavior which would be difficult to prevent leak in a practice.
- Intervention requires introduction of expensive equipment so economical to supply randomized sites.
- Attractive design to test effectiveness of intervention in practice setting (pragmatic)
Cluster Randomized Trials

- Randomization of individuals assumes outcomes are independent of one another.
- Cluster randomization characteristics of individuals in a group are more similar to each other than individuals in other groups.
- Outcomes may differ due to environmental differences as well as the intervention.
- Need to account for clustering effects in calculating sample size and in data analysis.

Cluster RCT

- Cluster RCT with reduced power due to increased chance of a Type 2 error.
- Intraclass coefficient (p) is the proportion of the true total variation in outcome attributable to the differences between clusters.
- Sample size = 1 + (m-1)p, where m = average cluster size
- Will always require larger sample size than a comparable non-cluster RCT.
- Nailing down the covariates that account for between cluster variation decreases sample size.
- Blinding may be difficult in cluster trials.

Cluster RCT

- Ethical issues
  - Consent of individuals is difficult as randomizing groups, little leeway for individual choice.
  - Consent obtained after cluster is randomized but post randomization selection bias an issue.
    - Drop outs after randomization may be due to assignment (individuals or entire clusters)
  - Decisions made by gatekeepers or caregivers.
  - How to ensure the interests of individual patients are met?
Shortfalls in currently published Cluster RCTs

• 42% of 149 cluster RCTs of implementation research did not account for the clustering in design.
• 42% of 54 studies of decision support systems accounted for clustering in analysis and none in their calculation of sample size.
• Only 25% of 16 explained the rationale for choosing cluster design.

CONSORT Recommendations for Cluster RCTs

• Explain rationale for choosing cluster RCT
• How clustering incorporated into design and sample size calculations.
• How effects of clustering were incorporated into the analysis.
  – Cluster outcome vs. individual patient outcome
• Flow of clusters from randomization to analysis.
  – Method of randomizing the assignment
    • Use of blocking, stratification, matching to minimize imbalance (potential for systematic bias)
    • Was the sequence of randomization concealed until assignment was made?

The delayed-start study design.
Considerations in the Design of Clinical Trials for Comparative Effectiveness Research

Grant D. Huang, MPH, PhD
VA Cooperative Studies Program
Society for Clinical Trials Pre-Conference Workshop
Baltimore, MD
May 16, 2010

OUTLINE

• Challenges facing clinical research
• Establishing and refining the question
• Methodological / operational considerations

The views, opinions, and content of this presentation are those of the speaker and do not necessarily reflect the views, opinions, or policies of the Department of Veterans Affairs.

Challenges in Clinical Research

• Burden of meeting regulatory requirements
• Limited resources
• Greater # of competing priorities
• Rapidly changing environment
  • Scientific & technological
• Contracting
• Hiring trained personnel
• More expectations for scientific process & product

RCTs take too long, cost too much, and don’t readily translate to the bedside.
The Problem

How does one conduct a clinical trial in comparative effectiveness research...
...that addresses an important clinical need
...which will be easily understood by patients and clinicians
...and can be completed quickly?

Can a clinical trial on comparative effectiveness be done “simply”?

Comparative Effectiveness Research

• Several definitions for CER are available (IOM, 2009):
  • Congressional Budget Office
  • IOM Roundtable on Evidence-Based Medicine
  • American College of Physicians
  • IOM Committee on Reviewing Evidence to Identify Highly Effective Clinical Services
  • Medicare Payment Advisory Commission
  • Agency for Healthcare Research and Quality

“Ultimately, CER aims to provide data that can influence clinical decisions for the better.”
- Initial National Priorities for CER – IOM, 2009

What is the Question?

• “Begin with the end in mind” Steven Covey, 7 Habits of Highly Effective People
  • What is the product at the end of the study?

• Comparative vs. explanatory study
  • What is the comparison - intervention, setting, population?
  • “while we’re at it…”

• Informing clinicians/patients vs. informing policy
  • A study aimed at changing policy may increase its complexity.
  • Policy is often the result of several factors.

• Consider the answer
  • Thinking about what is to be disseminated may help frame the question.
Refining the Question

• IOM list of priorities provide some guidance on important areas of investigation.

• What question do clinicians wrestle with and why?

• Are there existing biases that have to be taken into account?
  • E.g. - cancer screening

• What are the interests of the stakeholder(s)?

Methodological Considerations

• Design
  • Is an RCT the best approach? Does a study have to be blinded?

• Patient population
  • How do inclusion/exclusion criteria affect understanding of the comparison?
  • Do co-morbid conditions impact intervention?
  • Is managing risk factors important?

• “Cultural” factors
  • Is there clinical equipoise among clinicians?
  • Do differences in interventions affect implementation?
  • What are patient preferences?

• Outcomes
  • What is meaningful for the end user?
  • How often is follow-up needed? Justification?
Operational Considerations

- Do protocol requirements increase time needed for contracting, hiring, getting approvals?
- Do protocol requirements potentially complicate IRB reviews, consent form language, etc.?
  - E.g., variation in describing behavioral strategy vs. drug intervention vs surgical technique
- How much time will fulfilling GCP requirements, resolving data queries, adjudication, etc. add to the effort?

Conclusions

- CER has important potential to inform health care providers, recipients, and decision makers.
  - RCTs represent a major tool for providing strong evidence.
- Designing RCTs in CER should consider how the study will inform and be translated.
  - While attempting to keep things "simple"
- Methodological considerations in CER should give high priority to patients and clinician factors.

QUESTIONS?
Overview: This segment will address key issues that the statistician must address in working with other researchers to design and implement CER designs. The session will specifically discuss subtle ways in which research questions and outcome measures may be distinct for CER clinical trials and discuss how extensions to traditional/classic subject-randomized designs may provide advantages for CER trials. Examples of such designs that will be discussed are cluster-randomized designs, Bayesian designs and adaptive designs. An outline of the presentation is provided below:

1. The statistician and CER Clinical Trials: Some Background Thoughts
   a. Key Premise: The statistical aspects of the design of a CER RCT are not different than for any other SCT, but some of the issues differ subtly
   b. Review the key elements of the IOM definition of CER from a statisticians perspective
   c. Key elements in the statistical design of the RCT
2. What is the research question
   a. Overall goal of the study
   b. Treatment measures (key issue for CER, fixed or variable)
   c. Population of interest (address the CER issue of population versus individual risk)
3. Outcome measures (All of the normal issues plus the issue of patient-level or population-level inference plus the IOM definition)
4. Where does SCT fit within the CER evidentiary profile?
5. Selecting the study design in the face of research question and outcomes
   a. Limitations of the classic design
   b. Potential alternate study design approaches
6. Cluster-randomized designs
   a. Potential benefits for CER RCTs
   b. Key design issues
   c. Sample size calculations for cluster-randomized designs
7. Bayesian RCTs
   a. Potential benefits for CER RCTs
   b. Key design issues
   c. Sample size calculations for cluster-randomized designs
8. Adaptive designs for Arm trimming or addition
   a. Potential benefits for CER RCTs
   b. Key design issues
   c. Sample size calculations for cluster-randomized designs
9. Example study designs
   a. Cluster-randomized comparison of treatment for premature births
   b. Bayesian design for use of systematic review for study design
10. Final Comments
Statistical Issues in the Design of RCTs for Comparative Effectiveness Research
Dennis Wallace, PhD
Society for Clinical Trials Pre-Conference Workshop
Baltimore, MD
May 16, 2010

Designing RCTs for CER: Key Premise

• The statistical aspects of design of a randomized control trial (RCT) for comparative effectiveness research (CER) are fundamentally no different than those associated with any RCT, but the statistician must be ready to address specific CER concerns
• As always, the trial must be designed to answer the research questions of interest to the investigators within the available resource constraints and concerns for ethical treatment of research subjects.

The IOM Definition of CER from a Statistical Perspective

• **CER** is the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels.
• The key elements of this definition are the direct comparison of effective interventions, the study of patients in typical day-to-day clinical care, and the aim of tailoring decisions to the needs of individual patients.
Anatomy of a Study Design

Define the Research Questions and Hypotheses

Select the Study Design

Who are the Study Subjects?

What measurements do we need?

Predictors or Confounders

Outcomes

How will we analyze the data?

How many subjects do we need (sample size)?

What is the Research Question?

- What are the broad goals of the study in terms of clinical understanding and scientific inference?
- What treatment regimens are to be evaluated and how will those treatment regimens be implemented for this particular trial?
- What clinical population is the target of the inference for this trial?

Study Inference Goals

- In designing any randomized clinical trial, we must establish the framework in which the information generated from the study will be used to inform clinical understanding.
- In most classical designs, the clinical understanding is typically framed statistically as an estimation problem or a formal hypothesis testing problem.
- In designs for CER, clinical understanding may fall within the framework of evidence
  - Patient-specific decisions?
  - Evidence-based guidance?
Study Inference Goals—CER Evidence

Levels of evidence

- **1++** High quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias
- **1+** Well conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias
- **1** Meta-analyses, systematic reviews or RCTs, or RCTs with a high risk of bias
- **2++** High quality systematic reviews of case-control or cohort studies or High quality case-control or cohort studies with a very low risk of confounding, bias, or chance and a high probability that the relationship is causal
- **2+** Well conducted case-control or cohort studies with a low risk of confounding, bias, or chance and a moderate probability that the relationship is causal
- **2** Case-control or cohort studies with a high risk of confounding, bias, or chance and a significant risk that the relationship is not causal
- **3** Non-analytic studies, eg case reports, case series
- **4** Expert opinion


Treatment Regimens

- “Treatment” may be defined more ambiguously in CER RCT’s
  - Classicly the treatment regimen is very strictly defined through the protocol, with clear documentation of protocol deviations in the CRFs;
  - For the CER goal of “study of patients in typical day-to-day clinical care” regimens may adapted or modified to address patient comorbidities or potential drug-drug interactions.
  - Study designs should clearly specify how regimens are to be implemented or adapted and addressed in analyses.
  - Provisions may be needed for dropping or adding treatment regimens as evidence evolves or becomes available from external sources as the trial progresses.

Population of Inference

- Because a primary goal of CER is the study of patients in typical day-to-day clinical care, the populations included in CER-driven trials are likely to be more diverse than classic trials.
  - Provides greater generalizability
  - May create concerns of selection bias if inference population doesn’t match enrolled population
- Careful consideration will need to be given to the potentially conflicting goals of improving health care at both the population level and tailoring decisions to the needs of individual patients
  - With the CER focus on decisions that meet the need of individual patients, subgroup analyses will be an issue:
    - Key subgroups of interest should be defined a priori
    - Strategies for analysis re-post-hoc defined subgroups, particularly those defined by post-randomization factors should be addressed
What are the Study Outcome Measures?

- Luce in discussing pragmatic trials and CER notes: “Primary and secondary outcomes are patient-centered, chosen to reflect what matters most to patients and clinicians.”
- While focus on such items is laudable, this goal coupled with the goal of studying a diverse patient population in large simple trials will yield a number of study design issues for statisticians and clinicians:
  - In “patients with common comorbid conditions and diverse demographic characteristics” key patient-centered outcomes may differ across subpopulations of patients (composite outcome?)
  - For any given individual both the patient and clinician will be concerned with multiple measures of clinical effectiveness and with safety in the face of comorbidities (multiplicity)
  - Patient-centered outcomes are likely to involve quality of life measures, which may vary substantially with comorbidities and demography; these measures are historically difficult to standardize.

Study Design

- Goals of the study design are:
  - Yield clinically relevant estimates of treatment effects and the precision of these estimates
  - Minimize bias in both effect estimates and inference
  - Provide a high level of reproducibility and external validity
  - “Getting the correct design for the question being posed is critical since no amount of statistical analysis can adjust for an inadequate or inappropriate design” (Cook and Demets, 2008)
- Furthermore, to facilitate both design and interpretation, the trial should be designed as simply as possible given the complexities of the research question, the treatment algorithm, the outcomes of interest and their measurement, and the inferential requirements. (Adapted from Piantadosi, 1997)

Historic Limitations of RCTs for CER

- Narrowly defined inclusion/exclusion criteria limit the generalizability of the results to a broad clinical population
- Outcome measures selected are not those of greatest interest to patients and clinicians
- Trials tend to be expensive and long
  - Far too much data collected on parameters of limited value
  - Narrow criteria complicate recruitment
  - Dealing with subgroup differences and associated multiplicity issues increases sample size requirements
- Trials do not use external information efficiently for mid-trial decisions
Possible Alternative Approaches for CER RCT Design

- Cluster randomized trials
- Bayesian approaches for randomized clinical trial design and analysis
- Adaptive designs for trimming or addition of treatment regimens

Cluster-Randomized Trials: Potential Benefits for CER RCTs

- Randomization at the practice-level rather than the individual level generally provides a more heterogeneous patient population
- Cluster-randomized trials typically enroll larger numbers of subjects more quickly than individually randomized trials
- Potential for simplified data collection in that substantial portions of data can be collected from practice-level records rather than from individual patient visits.

Cluster-Randomized Trials: Key Design Issues

- Unit of Inference
  - Population-average effects based on use of the cluster (practice) as the unit of analysis and inference
  - Patient-level inference with the individual subject as the unit of analysis and inference
- Selection of the cluster-level randomization scheme
  - Simple, completely randomized design
  - Stratified randomization
  - Constrained allocation
- Accounting for within-cluster correlation of subjects in design and analysis
Cluster-Randomized Trials: Alternative Analytic Approaches

- Population-level inference with cluster as the unit of analysis
  - Because clusters are independent, use any standard approach like ANOVA or t-tests, possibly accounting for heterogeneity of variance across clusters
  - Permutation or randomization tests (possibly accounting for individual differences)
- Patient-level inference with subject as the unit of analysis (accounting for correlation among subjects)
  - Linear mixed models for continuous (normal) outcomes
  - Random coefficient models for categorical outcomes

Cluster-Randomized Trials: What is Intraclass Correlation and How Do I Determine it?

- The intraclass or intracluster correlation coefficient (ICC) is simply a measure of the "commonness" or correlation among individuals within the same cluster that makes those individuals have outcomes that are more similar than the outcomes for individuals in different clusters.
  - For continuous (normal) outcomes, the ICC is simply the ratio of the between cluster variance to the total sum of between and within cluster variance
  - For binary (binomial) and count (Poisson) outcomes, the ICC can also be viewed in a similar fashion, but it can also be viewed as a function of the between cluster variability in event rates that exceeds that expected by chance due to the underlying binomial or Poisson distribution

Cluster-Randomized Trials: Sample Size

- As with any clinical trial, selection of sample size is a function of the statistical properties of the trial and the resources available for analysis
  - For trials in which the cluster is the unit of analysis, sample size can be generated using the usual approaches for individually randomized trials, with the cluster considered to be the "individual"
  - For permutation or randomization-based analysis approaches, simulated power is typically the best approach
  - For large-sample based analytic approaches, simply use your favorite power calculation package
Cluster-Randomized Trials: Sample Size

- For trials in which the patient or subject is the unit of analysis, sample size can be generated using either simulation techniques or standard calculation formulae.
- Standard formulas are presented for a variety of outcome measures in the Donner and Klar reference (2000), Design and Analysis of Cluster Randomization Trials in Health Research.
- Generally, the approaches in this reference simply adapt standard approaches to account for the intraclass correlation.

Bayesian Approaches: Why consider it for CER?

- **Frequentist** (or classical) methods assume that unknown parameters (e.g., risk of mortality in the population) are fixed constants; hence, you cannot make probability statements about parameters because they are fixed.
- **Bayesian methods** treat parameters as random variables and define probability based on our understanding of how likely an event is to be true, as such we can make probability statements about parameters.
- By making probability statements about the current state of knowledge (prior probabilities), *Bayesian statistics* provides an approach and method of analysis which combine prior knowledge and accumulated experience with current information (likelihood) in order to make inferences about a quantity of interest (posterior probabilities).

Bayesian Approaches: What are some possible advantages for CER research?

- If good prior information is available related to a question, Bayesian approaches provide a formal calculus for integrating that information with new data.
- Specifying prior probability distributions requires collaborators to think carefully and make explicit decisions about how to include prior information.
- Bayesian approaches have some advantages for interim monitoring of clinical trials in the sense of handling accumulating information in a formal way and for dealing with subgroup analysis without having to deal with the issue of multiplicity.
Bayesian Approaches: What issues will need to be addressed in designing CER RCTs?

- Both the design of the trial and the conclusions of the trial depend upon the prior distributions. Consequently, statisticians and clinicians will need to work together closely to reach a consensus of how priors will be defined.
- Because the conclusions of the trial rely on prior information and decisions about that information can be somewhat subjective, standard approaches for presenting results have not been established; during the design phase, the team will need to determine how study results are to be presented.
- Software for the methods is less developed and less user friendly than for classic methods, so added expertise may be needed.
- Care must be taken to not be overly optimistic during study design.

Bayesian Approaches: Steps in Designing a Trial from a Bayesian Perspective

- As with the classic design
  - Define the key outcome measures for the study
  - Define key subgroups that will be included a priori in the analyses and establish the outcome measures for each subgroup
  - For each key outcome measure and each subpopulation of interest, define the prior distribution of the parameters.
  - Typically for CER, prior distributions will be based on strong evidence provided via systematic reviews.
  - Define both optimistic and pessimistic alternatives to the prior distributions.

- Establish the Bayesian evidentiary criteria that will be used to compare effectiveness of the alternatives at the end of the trial.
- Utilize either simulation procedures or standard calculation methods based on conjugate priors to estimate sample sizes based on the full range of priors established above.
RCTs for CER: Statistical Approaches

Questions
Case Study #1

Compare the effectiveness of traditional risk stratification for coronary heart disease (CHD) and noninvasive imaging (using coronary artery calcium, carotid intima media thickness, and other approaches) on CHD outcomes

Coronary heart disease is the leading cause of death in the United States. In nearly half of cases, the first clinical manifestation is death or a life-threatening heart attack. Decades of pathological and imaging research has revealed that the disease often goes through a prolonged asymptomatic phase, one that often lasts for decades.

During the past 10-15 years, it has become possible to use CT scans to take pictures of coronary arteries and look for deposits of calcium, which serve as indirect measures of disease. Several large-scale studies have shown that coronary artery calcium measures are strong predictors of future heart attacks and death in otherwise asymptomatic people. Some advocates call for routine screening of the population with coronary calcium scanning. Others, however, argue that we’ve seen previous cases where a strongly predictive test failed to improve outcomes when used for screening (e.g. PSA for prostate cancer, urinary catecholamines for neuroblastoma).

By today’s standard of care, adults at potential risk for coronary disease are assessed for risk based on “traditional risk factors,” namely age, gender, smoking, blood pressure, and cholesterol. The IOM calls for CER the compares the effectiveness of traditional risk stratification with noninvasive imaging (e.g. coronary calcium) as ways to prevent premature deaths and heart attacks.

What kind of study should be done?

A recent article illustrating the problem is attached (Bonow).

I am also attaching an article that discusses the appropriateness of randomized trials that focus on diagnostic tests (Lord).
**Case study #2:** Compare the effectiveness of primary prevention methods, such as exercise and balance training, versus clinical treatments in preventing falls in older adults at varying degrees of risk.

**Background:** In the U.S., over one-third of adults 65 years and older have had a fall. Falls are the leading cause of injury-related deaths, hospital admissions for trauma, and non-fatal injuries in older adults (CDC, 2010). In addition to injuries, falls are associated with disability, reduced mobility and physical fitness, and greater fears for falling. There are also significant costs associated with treating, caring for, and preventing falls and fall-related outcomes.

Several risk factors for falls have been identified. Having had a previous fall, balance impairment (as determined by gait and strength), visual impairment, and taking multiple medications are among the most commonly identified risk factors. Medical conditions associated with higher risks for falls include depression, diabetes, arthritis, pain, and urinary incontinence. Women are more likely to have a non-fatal injury from a fall, while men are more likely to die from a fall.

Efforts aimed at preventing falls have included both behavioral and pharmacological-based interventions. Exercise that combines gait, strength, balance, and endurance training is a commonly used approach. More recently, Tai Chi has become popular as a prevention method. It is not clear whether physical therapy or other structured therapies are beneficial.

Vitamin D has been found to help with fracture prevention. However, most other clinical interventions have focused on managing medications that are associated with risk factors such as antihypertensive agents, anticoagulants, and antidepressants. A recent meta-analysis (Woolcott et al., 2009) found that sedative, hypnotics, antidepressants, and benzodiazepines were associated with a higher risk for falls among the elderly.
Case Study #3: Compare the effectiveness of various screening, prophylaxis, and treatment interventions in eradicating methicillin resistant *Staphylococcus aureus* (MRSA) in communities, institutions, and hospitals.

Since its emergence in the early sixties, Methicillin-resistant *Staphylococcus aureus* (MRSA) was recognized to be associated with some predisposing factors. In 2005, around 94,000 invasive MRSA infections were estimated to occur in the USA, causing about 18,500 deaths. Traditionally confined to medical institutions, a growing number of community-acquired (CA-MRSA) cases are being diagnosed. Among the different forms of pathologic presentations, MRSA sepsis and pneumonia can be lethal. Only few costly antibiotics are currently efficient in treating these infections. It remains unclear, however, whether MRSA will expand the scope of their resistance thus causing further severe infections and posing growing challenges to the medical community.

Preventive measures
Infection control has multiple important elements, such as early screening, identification of MRSA carries for isolation, nasal and skin decontamination, staff education, enforcement of hand hygiene and decontamination of patients' wards. Besides nasal and cutaneous swabs, throat and rectal areas are considered for routine swabbing. While these aggressive measures and strict guidelines can improve the efficacy of hospital bed usage, they clearly cannot completely eradicate this resistant organism. It seems that MRSA is slowly gaining this battle by acquiring more territories and even using medical staff and their instruments, such as faucets, computer keyboards and stethoscopes, for further expansion.

What study would you propose to prevent MRSA? Please consider the elements described in the workshop.
Case Study #4. : Compare the effectiveness of pharmacologic and non-pharmacologic treatments in managing behavioral disorders in people with Alzheimer’s disease and other dementias in home and institutional settings.

Alzheimers disease is a major health burden and is expected to worsen as the population ages. Behavioral disturbances are common problems that affect the management of person with Alzheimers disease with estimation that 60-80% of persons with AD will have psychological or behavioral disturbances during the course of the disease, more so in the stages of severe dementia. Aggressive behavior is especially problematic as it can lead to self harm, physical, and more often psychic harm to caregiver, transfer to a more restrictive environment, social isolation, etc. It is sometimes associated with paranoid delusions and hallucinations. The use of antipsychotics, hypnotics, and anticonvulsants to decrease aggressive behavior in AD has been studied to some limited extent but use of these agents remains controversial. Behavioral based treatment protocols are also touted as helpful in cohort studies and small RCTs at single institutions. Larger studies show some benefits of atypical narcoleptics in decreasing aggression but significant side effects – tardive dyskinesia, cardiovascular morbidity and mortality, sedation-related side effects such as aspiration, dehydration, depression, etc. Some side effects are dose related. Choices include treatment with medium to high doses of drugs at time of an aggressive event. Others include daily administration of lower doses in attempt to prevent aggressive events.

Scales available include aggression scale, BehaveAD scale, counting aggressive events, SAEs, quality of life scales for AD filled out by caregivers.

Troubles to consider

- caregiver reluctance in withholding medications in aggressive patients because of fear that things will get worse and come to crisis.

Most patients who have already demonstrated aggressive behavior are on drugs so need to consider wash out in a drug trial with its associated potential behavioral consequences.

Ethics of randomization in frail, severely demented, elderly persons who are unable to give consent.

Placebo effect which may be related to the enhanced human-interactions with patients in a study as compared to persons with same stage of disease who are not in the study.

A) You are retained by a health organization which owns and runs 100 skilled nursing facilities across the country with 100 AD patients each and their data suggests half of these have intermittent aggressive behavior. They have a $2million research budget and would like to know whether a neuroleptic (risperidone) should be prescribed for their AD patients (subgroup unspecified) alone, in place of, or in addition to, a behavioral program that they already have in place. What type of trials would you consider proposing?
B) A Geriatrician consults you about an NIH grant idea. They wish to do a multi-center trial in the Alzheimers Disease Research Consortium. The Consortium has 10,000 AD patients in early and mid stages of the disease. Their data suggests that 10% of the patients have intermittent behavioral disorders. They wish to study whether daily use of an anticonvulsant, topiramate, is superior on some clinically meaningful metric, to counseling caregivers on behavior management. What type of trials would you consider proposing?

See attached Cochrane review for further background.
Practical Clinical Trials: Increasing the Value of Clinical Research for Decision Making in Clinical and Health Policy

Sean R. Tunis; Daniel B. Stryer; Carolyn M. Clancy

http://jama.ama-assn.org/cgi/content/full/290/12/1624

Correction
Contact me if this article is corrected.

Citations
This article has been cited 361 times.
Contact me when this article is cited.

Topic collections
Medical Practice; Health Policy; Quality of Care; Evidence-Based Medicine; Statistics and Research Methods; Randomized Controlled Trial
Contact me when new articles are published in these topic areas.

Related Letters
Realizing the Benefits of Practical Clinical Trials

In Reply:
The need for careful scientific evaluation of clinical practice became a prominent focus during the second half of the 20th century. Demonstration of pervasive and persistent unexplained variability in clinical practice and high rates of inappropriate care, combined with increased expenditures, have fueled a steadily increasing demand for evidence of clinical effectiveness. The limited amount of high-quality evidence is recognized to be partly responsible for geographic variation, inappropriate care, and the limited success of quality improvement efforts. As a result, increased attention is being directed to the development of methods that can provide valid and reliable information about what works best in health care.

The need for high-quality evidence

Among the primary audiences for higher-quality evidence are clinical and health policy decision makers, including patients, physicians, payers, purchasers, health care administrators, and public health policymakers. Patients and physicians increasingly seek to combine their personal beliefs about health care choices with attention to high-quality evidence in making individual decisions about care. Medical professional societies produce guidelines to assist physicians and patients in making medical decisions. Health insurers and managed care organizations increasingly depend on systematic reviews and technology assessments to support quality improvement efforts and to develop coverage and payment policy. Hospitals and health systems in turn depend on high-quality evidence to support clinical and health policy choices; however, the quality of available scientific evidence is often found to be inadequate. Reliable evidence is essential to improve health care quality and to support efficient use of limited resources. The widespread gaps in evidence-based knowledge suggest that systematic flaws exist in the production of scientific evidence, in part because there is no consistent effort to conduct clinical trials designed to meet the needs of decision makers. Clinical trials for which the hypothesis and study design are developed specifically to answer the questions faced by decision makers are called pragmatic or practical clinical trials (PCTs). The characteristic features of PCTs are that they (1) select clinically relevant alternative interventions to compare, (2) include a diverse population of study participants, (3) recruit participants from heterogeneous practice settings, and (4) collect data on a broad range of health outcomes. The supply of PCTs is limited primarily because the major funders of clinical research, the National Institutes of Health and the medical products industry, do not focus on supporting such trials. Increasing the supply of PCTs will depend on the development of a mechanism to establish priorities for these studies, significant expansion of an infrastructure to conduct clinical research within the health care delivery system, more reliance on high-quality evidence by health care decision makers, and a substantial increase in public and private funding for these studies. For these changes to occur, clinical and health policy decision makers will need to become more involved in all aspects of clinical research, including priority setting, infrastructure development, and funding.
tems use these documents to make decisions about capital investments such as equipment purchasing and construction (BOX).

The current clinical research enterprise in the United States is not consistently producing an adequate supply of information to meet the needs of clinical and health policy decision makers. The inability to address many common, important clinical questions, despite a significant increase in public and private funding for clinical research, suggests a systemic problem in the production of clinical research. This article explains the impact of knowledge gaps on health care decision makers, describes the features of clinical trials that would more reliably answer the practical questions they face, and discusses why the current clinical research enterprise fails to address many important practical questions. This article also proposes strategies to address the current shortage of clinical trials to meet these needs.

PREVALENCE AND IMPACT OF KNOWLEDGE GAPS

The prevalence and significance of gaps in knowledge about clinical effectiveness are most readily appreciated by reviewing the results of most systematic literature reviews, technology assessments, and clinical practice guidelines. These reports are generally produced to provide comprehensive reliable information for decision makers and usually address common conditions with large aggregate cost, morbidity, and public health importance. A consistent finding of these reviews is that the quality of evidence available to answer the critical questions identified by experts is suboptimal. For example, a systematic review of newer pharmacologic agents for depression concludes that few studies provided data on the long-term effectiveness of treatment, the functional status of patients, or the outcomes of patients treated in typical practice settings. Furthermore, few studies compared the older inexpensive agents with newer agents in terms of adverse effects and clinical efficacy. Most well-done systematic reviews and clinical guidelines reach similar conclusions about the quality of evidence associated with common clinical problems.

These gaps in evidence undermine efforts to improve the scientific basis of health care decisions in several ways. Organizations that develop evidence-based clinical practice guidelines may not be able to develop clear, specific recommendations. For example, the background report for clinical guidelines on outpatient management of exacerbations of chronic obstructive pulmonary disease found that although numerous industry-sponsored clinical trials reported minor differences in the antimicrobial activity of alternative broad-spectrum antibiotics, no trials had been performed to determine whether any of the newer broad-spectrum antibiotics were better than older generic antibiotics or even placebo (for mild exacerbation). As a result, the guideline could not provide definitive recommendations on the appropriate choice of antibiotics for chronic obstructive pulmonary disease exacerbations.

The limited quantity and quality of available scientific information also impede the efforts of public and private health insurers in developing evidence-based coverage policies for many new and existing technologies. Poor-quality studies of new technologies can lead to millions of dollars being allocated for new technologies for which the long-term benefits and risks have not been determined. Minimally invasive technologies for treatment of benign prostatic hyperplasia (BPH) are in widespread use, yet no clinical trials have been performed to compare the risks and benefits of these treatments with standard surgical interventions. The Medicare program has spent millions of dollars per year for home use of special beds for patients with pressure ulcers, despite the fact that no well-designed study demonstrates that they improve healing of these ulcers. The limited production of this body of research becomes increasingly problematic as major increases in public and private funding for clinical research appear.

Box. Uses of Evidence in Decision Making

**Physician/Patient Decision Making**

Of existing diagnostic or treatment alternatives, which makes the most sense for an individual patient?

**Choosing Plans or Physicians**

Which plan or physician is likely to provide high-quality care?

**Practice Guidelines**

What is the best approach for patients with selected conditions?

**Quality Measurement and Improvement**

How can evidence-based clinical performance be assessed? Do improvement programs result in enhanced clinical care?

**Product Purchasing and Formulary Selection**

How does this product compare with existing alternatives?

**Benefit and Coverage Decisions**

Should a new service be reimbursed and for which patients?

**Organizational and Management Decisions**

Does a hospitalist program decrease costs and improve outcomes?

**Program Financing and Priority Setting**

Which services represent the best value for additional investments?

**Product Approval**

Should this product be approved and, if so, for which indications?

©2003 American Medical Association. All rights reserved.
Clinical trials designed to assist health care decision makers, referred to as pragmatic clinical trials or practical clinical trials (PCTs), are defined as trials for which the hypothesis and study design are formulated based on information needed to make a decision. They are distinguished from explanatory clinical trials, for which the goal is to better understand how and why an intervention works. Explanatory trials are designed to maximize the chance that some biological effect of a new treatment will be revealed by the study. The PCTs address practical questions about the risks, benefits, and costs of an intervention as they would occur in routine clinical practice. The most distinctive features of PCTs are that they select clinically relevant interventions to compare, include a diverse population of study participants, recruit participants from a variety of practice settings, and collect data on a broad range of health outcomes.

**Compare Clinically Relevant Alternatives**

Often PCTs are designed as head-to-head comparisons of viable alternative clinical strategies. These comparative studies have the potential to alter clinical decisions profoundly, because they are derived from practical choices facing patients and their physicians. In contrast, many of the clinical studies comparing individual agents with placebo may not provide a sound basis for choosing from among acceptable alternatives. For example, a PCT conducted by the Department of Veterans Affairs (VA) compared the benefits of pharmacologic therapy with terazosin hydrochloride, finasteride, or both for treatment of symptoms of benign prostatic hyperplasia (BPH). Both drugs are approved by the Food and Drug Administration (FDA) for this use based on trials comparing each drug with a placebo. There is no incentive or requirement for manufacturers to initiate studies to compare the products. The VA study randomized 1229 men with BPH and demonstrated that terazosin was more effective than finasteride for reducing BPH symptoms at 1 year. Once initiated, both manufacturers contributed to the design and funding for the study. The National Institutes of Health (NIH) funded a study that compared doxazosin mesylate, finasteride, and both drugs in 3047 patients with BPH for a mean of 4.5 years. This study revealed that the combination of the 2 drugs was significantly more effective at delaying progression of symptoms than either drug alone.

In addition, PCTs can address nonpharmacologic alternatives. Despite back pain being among the most prevalent complaints in primary care, few high-quality studies have been performed to compare results of the numerous existing treatment alternatives. A PCT of therapy for recent-onset low back pain randomized 323 patients to 1 of 3 widely used alternative adjunct treatment strategies: physical therapy, chiropractic care, and self-care (the principles of which are described in an educational booklet). The study showed that physical therapy and chiropractic care increased patient satisfaction and marginally reduced symptoms compared with the self-care principles outlined in the booklet; however, there were no differences between the 3 study groups in function or rates of recurrence. The educational booklet that outlined self-care was substantially less expensive. A study of similar design demonstrated that therapeutic massage was more effective and less costly than acupuncture in treating low back pain.

**Enroll a Diverse Study Population**

Typically, PCTs include a more diverse study population by having broad inclusion criteria and fewer exclusion criteria when enrolling patients. The goal is to enroll patients in the trial with characteristics that reflect the range and distribution of patients observed in clinical practice for a particular problem. This approach addresses the common concern of decision makers about the applicability of results from studies with restricted eligibility criteria. It can also ensure that the higher-risk patients likely to have the greatest benefit from some treatment are not excluded from clinical trials.

Most clinical trials of antidepressants have excluded elderly patients and focus on patients with major depression rather than patients with less severe depression, which is a more prevalent problem. Williams et al enrolled a substantial number of elderly patients in a PCT of 415 patients with minor depression or dysthymia, comparing the effectiveness of oral antidepressants, problem-solving treatment (a form of cognitive therapy), and placebo. The study found that treatment with paroxetine showed greater improvement in depressive symptoms and mental health function compared with problem-solving treatment or placebo. Reliable and relevant information such as this is more likely to convince physicians to provide medication for their elderly patients with this common cause of significant reversible morbidity.

Because physicians must often treat patients based on the likely rather than confirmed diagnosis, studies that enroll patients based on presenting symptoms rather than definitive test results may be of great practical value. A PCT of patients with sinusitis compared the effectiveness of antibiotics with and...
without nasal corticosteroids.\textsuperscript{34} Because primary care physicians typically use sinus radiographs to evaluate patients with suspected sinusitis, radiographic evidence of sinusitis was used as a study entry criterion (rather than sinus puncture). This approach maximized the generalizability of the study results by enrolling a patient population that physicians would recognize as similar to patients observed in daily practice. Another PCT randomized 478 patients to 2 alternative strategies for managing patients who presented with dyspepsia: (1) blood testing for \textit{Helicobacter pylori} with patients with positive test results referred for endoscopy or (2) empirical treatment with acid-suppressive drugs.\textsuperscript{35} Neither study group involved routine referral for endoscopy, replicating the standard of care for this common primary care problem.

\textbf{Recruit From a Variety of Practice Settings}

Also, PCTs may improve external validity by including a wider range of physicians and settings to which the study will be applicable. Pragmatic clinical trials have often been conducted in community-based settings, generally clinics and physician offices at which the primary activity is clinical care. Enrolling patients from a more diverse group of practice settings also allows for some variability in how the study intervention and associated clinical care will be provided.\textsuperscript{36} The ancillary care that these patients receive is more likely to reflect the average care patients would receive outside the research context. In the study of nasal corticosteroids for sinusitis, 18 of the 22 study sites were community-based primary care or otolaryngology clinics. Dolor et al\textsuperscript{37} noted that this selection of study settings would ensure that the study sample represented the patient population typically observed in general practice. Outcomes of risky or technically demanding procedures are particularly likely to vary across sites, making it especially valuable to have results that reflect outcomes across a range of institutions.

The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) compared treatment outcomes for various drug treatments for hypertension in more than 42,000 patients enrolled from 623 primary care clinics.\textsuperscript{38} Many of the study sites had not previously been sites for clinical research. The study demonstrated that use of inexpensive diuretic medication resulted in a lower incidence of major cardiovascular events compared with the newer and more expensive classes of antihypertensive agents, including calcium channel blockers, angiotensin-converting enzyme inhibitors, and \textit{\&}-blockers.\textsuperscript{37,38} Physicians and policymakers can be confident that the reported outcomes in this study are likely to predict results that will be observed across a wide range of practice settings. This eliminates a common obstacle to physician implementation of clinical research findings.

\textbf{Measure a Broad Range of Relevant Health Outcomes}

Selection of the outcomes to be measured in PCTs is based on the most important anticipated effects of the intervention, taking into account those outcomes of greatest relevance to decision makers. As a result, the study endpoints collected in PCTs include a broad range of functional outcomes, including quality of life, symptom severity, satisfaction, and costs, as well as more traditional end points, such as mortality and major morbidity.\textsuperscript{39,40} The low back pain study by Cherkin et al\textsuperscript{41} reported disability days, satisfaction, and ability to function. Many large trials of cardiovascular interventions now report general and disease-specific quality of life rather than focusing only on mortality or major nonfatal outcomes such as stroke or myocardial infarction.

In addition, the period of follow-up for PCTs is often longer than in traditional clinical trials to better reflect more of the natural history of a disease. Many therapies have different results in the short term than in the long term, a phenomenon commonly observed in surgical trials that usually demonstrate high short-term risks associated with surgery.\textsuperscript{42} Two PCTs that compared the results of immediate surgical repair with surveillance for small abdominal aortic aneurysms followed up patients for 4.9 years in 1 trial and 8.0 years in the other trial.\textsuperscript{41,42} Previous studies on this question had been much shorter and had left substantial uncertainty about the safety of deferring surgical repair of small aneurysms. Both studies demonstrated that survival was not improved by elective surgical repair of these aneurysms, a result that could lead to the avoidance of many unnecessary surgical procedures.

Inclusion of economic outcomes in clinical trials provides information that has a significant impact on clinical and policy decisions.\textsuperscript{43} A randomized study evaluated the results of treatment with fluoxetine compared with tricyclic antidepressants for the initial management of mild depression.\textsuperscript{44} Primary outcomes of the study included the type and frequency of adverse effects, the number of patient visits associated with these adverse effects, and the rate at which patients were switched from their originally assigned medication to another agent. The study concluded that, despite the higher cost of fluoxetine, total costs of care were no different between the 2 groups because the incidence of adverse effects and frequency of physician visits were lower for patients initially treated with fluoxetine. The findings provided important input for formulary decision makers and were also given substantial weight in the development of evidence-based clinical guidelines on the management of depression.\textsuperscript{45}

\textbf{WHY IS THE SUPPLY OF PCTs INADEQUATE?}

The supply of PCTs is currently inadequate in large measure because trials often require substantial resources, and the current level of public- and private-sector funding for these studies is inadequate. Neither of the major sources of funding for clinical research in the United States—the NIH and the medical products industry—has as a primary mission the goal of ensuring that
CLINICAL RESEARCH FOR DECISION MAKING

...studies are performed to address clinical questions important to decision makers. The funding entities that are focused on supporting these studies have relatively limited resources.

Practical clinical trials may be extremely costly to conduct because they can require large sample sizes and may follow up patients for a long period. The 7-year, $42,000-patient ALLHAT study is estimated to have cost approximately $120 million to complete (C. Furberg, MD, PhD, written communication, January 9, 2003). The National Emphysema Treatment Trial enrolled 1200 patients in the surgical and medical therapy groups, took approximately 6 years to complete, and cost more than $35 million for the research, with substantially more than this amount paid by Medicare for the clinical care of patients in the trial (G. Weinman, MD, written communication, April 16, 2003). Even studies of relatively modest size and simple design can have substantial costs. An NIH study of St John’s wort and citralopram will enroll 300 patients with mild depression and follow them up for 20 weeks, with an estimated cost of $4 million over 4 years.

Although PCTs are not a major focus for the NIH, they have funded a number of important PCTs, most of which have provided crucial information for clinical and health policy decision makers. The NIH does not, however, have any organized systematic mechanism in place to identify the high-priority questions of decision makers that might need to be resolved through a PCT. Estimates of the distribution of NIH extramural grants indicate that approximately 70% of NIH grants are devoted to basic biomedical research, with the remainder going to clinical research of all types. The NIH scientific review committees, as currently constituted, are more familiar with traditional explanatory trials and have relatively limited experience with PCTs. Consistent with its primary mission and funding priorities, the NIH has been successful at biomedical discovery but has been less consistent in supporting clinical research that ensures that these discoveries will be effectively translated into improved patient care and public health.

Some public research funding organizations have an explicit commitment to supporting research that addresses practical questions of health care decision makers. The Agency for Healthcare Research and Quality (AHRQ) has funded several important PCTs, but devotes most of its resources to health services research, quality measurement, and patient safety. According to AHRQ (L. Levine, oral communication, February 24, 2002), total spending on PCTs was approximately $30 million in fiscal year 2001, and the current agency budget of approximately $300 million would not allow for a significant expansion for such studies while maintaining its other core activities. The AHRQ and FDA jointly oversee the Centers for Education and Research in Therapeutics, which were established for the purpose of improving the effective use of pharmaceuticals. The Centers for Education and Research in Therapeutics are directed to accomplish this by conducting “essential research studies on drugs and therapies that are not being performed by the pharmaceutical industry or the NIH” and communicating these findings to practicing physicians. The annual funding by the Centers for Education and Research in Therapeutics of $7 million is adequate to identify, but not to support, important PCTs related to pharmaceutical therapy.

The Cooperative Studies Program of the VA has produced numerous studies designed to support clinical and policy decisions. This focus of activity emerges directly from the purpose of that research program: to support a research agenda designed to improve the effectiveness of care for veterans. This program funded some of the earliest trials clearly demonstrating the value of coronary artery bypass graft surgery. More recently, the VA funded an evaluation of arthroscopic surgery for osteoarthritis of the knee, which showed that the results of placebo surgery were no different than results from actual lavage and debridement. Although the clinical implications of the study are still being debated, the Department of Veterans Affairs issued an advisory to its physicians, recommending that they not perform this surgery in patients similar to those in the clinical trial. The study also prompted the Medicare program to consider specific limitations for use of this procedure, guided primarily by results of the VA study. With approximately $55 million to support such trials, the VA is an important source of PCTs in the United States, but the number of important questions is far beyond this level of funding.

Industry funding for clinical trials is several times greater than NIH spending ($4.1 billion vs $850 million in 2000), however, most of the clinical research supported by the pharmaceutical and medical device industries is designed primarily to meet the regulatory requirements for marketing approval by the FDA. More than 90% of industry spending on clinical trials supports phase 1, 2, and 3 studies required for regulatory approval, with the remaining 10% spent on phase 4 clinical trials. Most PCTs would be classified as phase 4 trials.

Most devices reviewed by the FDA are cleared for marketing based on the similarity of those devices to previously approved devices, and there is no requirement for clinical studies that directly evaluate the clinical effectiveness of most devices. Once FDA approved, devices may be widely used off label for clinical purposes that have never been carefully evaluated. Most of the pulmonary artery catheters in modern use have been approved by the FDA without clinical trial data based on the similarity of these catheters to those in use before 1976 (the year that the FDA began to regulate medical devices). Large randomized trials have recently shown that the pulmonary artery catheter does not provide clinical benefit compared with standard care in elderly high-risk surgical patients who require intensive care. Most of the therapies in common use for treatment of pressure ulcers and other chronic wounds, such as electrical stimulation and hyperbaric oxygen, have never been...
the subject of high-quality trials or received FDA approval for treatment of wounds.

For pharmaceuticals, the FDA requires placebo-controlled studies for approval and generally does not require trials that compare a new drug with existing alternatives, unless a company is specifically seeking FDA approval to make a claim of superiority to another marketed product. These comparative studies are also unattractive to drug manufacturers if there is any chance that the results might reflect poorly on their product. Therefore, industry-sponsored drug studies will usually not provide the comparative information about risks and benefits sought by most healthcare decision makers. Furthermore, because the FDA requires information about efficacy rather than effectiveness, the study populations enrolled in FDA-approval studies are highly selected, in contrast with the more diverse study populations that enhance the generalizability of results from PCTs.

Although medical product companies increasingly support additional clinical studies after FDA approval, postmarket research priorities for industry are understandably guided by the objective of maximizing market share rather than effectiveness, the study populations enrolled in FDA-approval studies are highly selected, in contrast with the more diverse study populations that enhance the generalizability of results from PCTs.

We propose a set of strategies that we believe would promote more consistent production of clinical trials that meet the needs of healthcare decision makers. These strategies would call for sustained effort and substantial resources from all affected stakeholders.

**STRATEGIES FOR IMPROVEMENT**

**Systematic Identification and Prioritization of Knowledge Gaps**

Currently, there is no institution with the responsibility for identifying and prioritizing important unanswered clinical research questions that would be useful to clinical and health policy decision makers. It is essential that this function be established. Ideally, the responsible entity would have majority representation from the decision makers in need of this information, including representatives of payers, purchasers, medical professionals, and patients. One possibility would be for the Institute of Medicine to serve this function, providing staff and supportive infrastructure within a neutral scientific environment. The list generated by this group would be available to public and private funders as they develop new initiatives and make their funding decisions. The degree to which funders supported studies from this list would highlight whether research funders were meeting important public health needs. To the extent that other health care stakeholders begin to devote additional resources to clinical research, the priority list will also serve as a guide to research most likely to generate reliable usable information.

There are a number of sources from which high-priority questions for PCTs could be identified. Virtually every clinical guideline, technology assessment, systematic review, and consensus report includes a section that lists specific clinical research priorities. These priorities deserve special attention because of the systematic and comprehensive method by which they have been generated.

**Decision Makers Insist on High-Quality Evidence in Making Decisions**

The production of high-quality clinical trials will increase significantly when health care decision makers decide to consistently base their decisions on high-quality evidence. Research sponsors (public and private) will be motivated to provide the type of clinical research required by decision makers. Payers and purchasers can clearly indicate to the drug and device industry that favorable coverage and payment decisions will be expedited by reliable evidence from PCTs. In particular, manufacturers will be motivated to perform head-to-head comparative trials if these are required to justify payments higher than the existing less expensive alternatives. Physicians and medical professional organizations can also increase the degree to which care of individual patients and professional society clinical policy are guided by attention to reliable evidence.

Patients and their advocacy organizations have a critical stake in the quality of clinical research. The fundamental premise of improving health care through informed patient choice cannot be realized if unreliable evidence is used to inform patients. Patient advocacy organizations should become more insistent on the production of high-quality evidence and work with research funders to increase support for PCTs. These groups could also facilitate the research by encouraging patients, physicians, and health care organizations to participate in clinical trials in those situations for which current evidence is uncertain. Many years and lives could have been saved had advocates worked to ensure the rapid completion of clinical trials on bone marrow transplantation for breast cancer, instead of putting pressure on state and federal policymakers to force payers to cover this procedure when supportive high-quality data were lacking.

©2003 American Medical Association. All rights reserved.
Create Operational Infrastructure

To conduct PCTs efficiently, a durable infrastructure within the primary care setting is needed. Considerable effort and resources will be required to develop additional networks of physicians willing and able to participate in clinical research. The HMO Research Network, a consortium of large health maintenance organizations, including Kaiser Northern California, Group Health of Puget Sound, and Harvard Vanguard Health System, may serve as a model for its capacity to implement studies and capture data within the context of usual care.\(^5^9\) Other examples include AHRQ's 55 practice-based research networks, the VA Cooperative Studies Program, and a number of specialty-specific research networks sponsored by various NIH institutes. The NIH also supports the General Clinical Research Centers, a robust infrastructure for clinical research within academic medical centers. Additional attention to expanding this infrastructure beyond teaching hospitals and clinics and to extending the capacity for clinical research into real-world clinical care settings would be extremely helpful in facilitating the conduct of clinical trials of value to decision makers.

It will also be necessary to significantly expand the number of physicians capable of conducting PCTs. The NIH and AHRQ should expand their programs to provide special career awards, scholarship, and loan-repayment programs aimed at expanding the clinical research workforce. Although some individuals will need extensive training to become principal investigators and devote their professional careers to these studies, a minimum level of skills in clinical investigation should become part of the basic training of physicians. Ultimately, with the expansion of electronic medical record systems, most clinical encounters could provide useful data and most physicians could function in part as clinical investigators.

Address Methodological and Ethical Issues

Practical clinical trials pose a number of methodological and ethical challenges that will need to be identified and addressed. Clinical trialists have developed some strategies to reduce the cost of conducting large lengthy trials. Large simple trials feature simple protocols, enrollment from a large number of research sites, limited patient exclusion criteria, and data collection limited to the smallest possible number of elements.\(^6^0\) This approach can make it possible to study thousands of patients at relatively low cost and was first shown to be feasible by Richard Peto and his colleagues, who conducted several large landmark trials of drug treatment for acute myocardial infarction.\(^6^0\) Broader application of these strategies to PCTs will be needed to minimize cost and participant burden. Because ancillary patient management in a PCT may not be standardized, there are likely to be clustering effects at the level of physicians and organizations that will require specialized methods to facilitate analysis or randomization at the organizational level. Such study designs pose some unusual methodological challenges that will require additional attention to refine.

For questions of comparative effectiveness, the expected outcome differences between 2 active therapies are likely to be considerably smaller than the differences between a treatment and placebo. Trials designed to demonstrate such differences will require large sample sizes and may be expensive. In some circumstances, fewer patients may be necessary to compare 2 active treatments in an equivalence trial: a study designed to demonstrate that there are no important differences in effectiveness between 2 active treatments. However, equivalence trials raise a number of challenging scientific and ethical issues and are not yet widely accepted, so additional conceptual development of this approach will be necessary.\(^6^2\)

Several important ethical issues may be encountered in the design and conduct of some PCTs. Because a number of these trials may involve comparison of new vs old technology, expensive vs inexpensive drugs, or high-intensity vs low-intensity services, there may be concerns about the economic motivations for the trials and whether some patients enrolled in the trials are being denied optimal care (ie, whether true clinical equipoise exists).\(^6^3\) Informed consent and confidentiality issues may be raised by more extensive use of clinical and administrative records to identify potential study participants and to augment the data collected specifically for research purposes. Consent procedures that are both adequate and efficient will need further exploration, because the number of PCTs that can be performed may be limited by the operational burden they impose, particularly in settings not primarily focused on research. In addition, efficient strategies for obtaining institutional review board approval from multiple institutions will be necessary, perhaps through approaches that involve coordination of central and local institutional review board reviews.

Funding Options for PCTs

A significant amount of funding will be necessary to support PCTs, given the large number and high cost of such trials. Furthermore, a substantial amount of support will be needed to develop the infrastructure necessary to efficiently conduct these studies. Resources to fund these trials will necessarily come either from redistribution of current research funds to important PCTs or by finding additional sources of new funds. Several possible public and private sources of additional or redistributed funding should be considered.

Although increasing the portion of NIH resources flowing to PCTs may have merit, given the NIH mission to improve public health through research, there is likely to be concern that the money is being diverted from important basic research. However, the NIH has recently acknowledged the need for improvements in the clinical research enterprise\(^6^4\) and may be increasingly willing to collaborate with decision makers once they have more clearly identified their critical clinical research priorities. Funds from the NIH are also needed to support further development of the PCT infrastructure, including support for research networks, expansion of the Gen-
CLINICAL RESEARCH FOR DECISION MAKING

Several Clinical Research Centers beyond academic centers, and training physician-investigators who have skills to design and conduct PCTs.

Industry funding could also support PCTs. This is most likely to occur in response to the increased use of evidence-based policy development by decision makers. Industry will be particularly responsive to coverage and payment policy firmly linked to evidence of improved health outcomes and increased payment to evidence of substantially improved outcomes. The regulatory framework of the FDA is designed to ensure that medical technologies approved for marketing are safe and effective for their intended use and is not structured to ensure that research is conducted that will inform optimal clinical use of these technologies. Additional requirements for comparative studies before FDA approval could significantly increase the time and expense of bringing new technologies to market. Recent policy discussions have been focused on increasing the speed of FDA approvals by developing strategies for gathering additional data in the postmarketing period. Increased diffusion of health information technology will facilitate more rapid and efficient conduct of requisite comparative studies, in both the premarket and postmarket phases.

Because payers and purchasers increasingly demand high-quality evidence to support decisions, they are often proposed as the most likely source of additional funding for PCTs. The benefit of an enhanced evidence base for making clinical and health policy decisions ultimately accrues to those who use health care services. For that reason, financial support and priority setting for this body of research would most sensibly be associated with the funding streams that support the provision of health care services. Furthermore, the amount of funding for this body of applied research should be linked to the level of resources being consumed in supplying health care services.

Several health care delivery systems allocate dedicated resources for clinical research. The VA spends approximately $24 billion on the delivery of health care services and maintains a budget of $3.5 billion for research, including approximately $600 million for clinical trials (virtually all of which are PCTs). The National Health Services in England and Canada also have provided funding for a large number of PCTs. These are examples of health systems where health care services are delivered within a fixed budget and clinical research is understood to be a critical mechanism to ensure optimal use of existing resources. Such an approach should be considered in the broader US health care systems and could provide substantial resources for high-priority applied clinical research if even a small fraction of total health care spending were reserved for this purpose.

This proposal, although conceptually simple, may be challenging to implement. Most private insurance contracts exclude payment for experimental and investigational services, although coverage of clinical trials is increasingly popular. At present, the Medicare program does not have broad legal authority to fund clinical research, because the statute does not allow payment for care unless medically necessary. If payers and purchasers were to identify funds for clinical research, it would be essential to have a functioning system for priority setting free of political interference and robust mechanisms for oversight of sponsored research. Finally, costs for supporting clinical research will, at least initially, increase overall health care spending or replace spending now assigned to other technologies and services. Although PCTs might be expected to increase the efficiency of spending and the health benefit derived from given dollars, any increased costs for clinical research will eventually be passed on to the payers, purchaser, employers, public programs, and eventually the general public. Health care consumers will ultimately need to decide whether this is a desirable use of their money.

CONCLUSION

The United States is now seeing the result of its heavy investment in biomedicalex: numerous promising medical products have been developed and many more are on the way to initial clinical trials. With this success comes an equally important additional need: to develop a systematic approach to efficiently evaluate the risks and benefits of these new technologies in the context of existing alternatives. Currently, decision makers in health care are not provided with information of adequate quality to make well-informed decisions regarding alternative strategies for diagnosis and treatment of most common clinical conditions. Improving the quality of clinical research will depend on more active involvement of clinical and health policy decision makers in all aspects of clinical research, including priority setting, study design, study implementation, and funding.

Disclaimer: The opinions contained in this article represent the views of the individual authors and do not necessarily reflect the views or policies of the Centers for Medicare & Medicaid Services or AHRQ.

Acknowledgment: We thank Joe Selby, MD, Allan Korn, MD, and Robert Califf, MD, for comments on earlier drafts of the manuscript. Alex Omya, PhD, and the members of the Institute of Medicine Clinical Research Roundtable contributed greatly through many insightful discussions of these issues. We also acknowledge the wisdom and support of Peter Mazzone, MD, in the earliest stages of this work.

REFERENCES

Clinical Research for Decision Making

...
Should Coronary Calcium Screening Be Used in Cardiovascular Prevention Strategies?

Robert O. Bonow, M.D.

A 52-year-old man requests a coronary-artery calcium (CAC) scan for assessment of his risk of coronary events after seeing an advertisement from a local facility that offers the test. He has no symptoms of cardiac disease, has never smoked, and is not overweight, but he does not exercise regularly. His father, who was a heavy smoker, had a fatal myocardial infarction at 45 years of age. The patient’s blood pressure is 130/85 mm Hg. His total cholesterol level is 220 mg per deciliter (5.7 mmol per liter), and his low-density lipoprotein (LDL) and high-density lipoprotein (HDL) cholesterol levels are 160 mg per deciliter (4.1 mmol per liter) and 38 mg per deciliter (1.0 mmol per liter), respectively. His fasting blood glucose level is 92 mg per deciliter (5.1 mmol per liter). What would you advise regarding CAC scanning?

During the past four decades, there has been a dramatic decline in the age-adjusted rate of death from cardiac disease in the United States and many other developed countries. This reduction is attributed in large part to primary and secondary prevention strategies that target modifiable risk factors. Despite these advances, cardiovascular disease remains the leading cause of death in developed countries as well as in most developing countries, and there is concern that the growing prevalence of obesity and type 2 diabetes will reverse the gains of the past 40 years. Furthermore, risk factors for cardiovascular disease are present and poorly controlled in the majority of persons who have no symptoms.

Adult Treatment Panel III (ATP III) of the National Cholesterol Education Program based its 2001 recommendations for the treatment of hypercholesterolemia for primary prevention of cardiovascular disease on the assessment of individual risk factors and global risk as indicated by the Framingham risk score, which is determined on the basis of age, blood pressure, levels of total cholesterol and HDL cholesterol, and smoking status. With the use of these criteria, persons without established coronary heart disease are considered to be at high risk if they have either a condition that is equivalent to coronary heart disease with respect to the risk of a coronary event (e.g., peripheral-artery atherosclerotic disease or diabetes) or more than a 20% estimated likelihood of a coronary event over the next 10 years; they are considered to be at low risk if the estimate is less than 10%. Although at the time of their first acute coronary event more than 85% of patients have at least one of the major established risk factors for heart disease, the presence of only a single risk factor would place most people in the low-risk category according to the ATP III criteria — meaning that intensive measures to reduce risk factors would not be considered necessary. The 2004 ATP III update also defined patients at “moderate
risk” for a coronary event as those having two or more risk factors and a 10-year risk of less than 10% and patients at “moderately high” risk (previously termed “intermediate risk”) as those with a 10 to 20% likelihood of having a coronary event during the next 10 years. Neither patients at moderate risk nor those at moderately high risk would be considered candidates for the aggressive treatment of risk factors that is recommended for patients at high risk.

The observation that a person classified as being at low, moderate, or moderately high risk for a coronary event according to the updated ATP III criteria10 may nonetheless have a coronary event has spurred interest in new markers of such risk that might identify those who would benefit from more rigorous attention to risk-factor modification.11-13 In this regard, the potential role for imaging of subclinical atherosclerosis — in particular, the use of electron-beam computed tomography (EBCT) or multidetector computed tomography (MDCT) to detect CAC — has generated considerable attention and debate.

**STRAATEGIES AND EVIDENCE**

EBCT is distinguished from standard computed tomography (CT) by the use of a scanning electron beam that is steered by an electromagnetic deflection system across a series of fixed tungsten target rings; x-rays generated by the tungsten rings fan across the patient and are captured by detectors placed opposite to the tungsten rings. In contrast, MDCT uses the traditional x-ray tube and physically rotates the x-rays around the patient. The absence of mechanical motion in EBCT results in rapid image acquisition, with exposure times of just 50 to 100 msec for each tomogram. MDCT is also rapid, the exposure time is more than 220 msec. Both EBCT and MDCT are sensitive methods for detecting coronary-artery calcification and provide highly reproducible data.14-17 Although MDCT is more widely available than EBCT and appears to yield results of similar usefulness in terms of risk assessment,18 most of the information relating CAC scores to cardiovascular risk comes from studies using EBCT.

With very rare exception, coronary calcification is a marker of coronary atherosclerosis. Hence, the presence of coronary calcium signifies underlying coronary-artery disease, with essentially no false positive findings, although calcification is not synonymous with luminal stenosis or obstruction. The CAC score, developed initially with the use of EBCT to quantify the extent of coronary calcification,19 correlates closely with the volume

---

**Figure 1. CT Scans of the Left Coronary Artery in Two Asymptomatic Men.**

Two asymptomatic men, 51 years of age and 81 years of age, underwent coronary-artery calcium (CAC) imaging with multidetector CT. The resulting transaxial cardiac tomograms, obtained at the level of the origin of the left coronary artery, show calcification of the left main and proximal left anterior descending coronary arteries in both the younger patient (Panel A) and the older patient (Panel B). The CAC score for the younger man, although relatively low at 80.1, places him in the 85th percentile for severity of CAC for men in his age group. The older man’s CAC score is higher, at 1054, but the severity of his CAC relative to that for men in his age group is lower — in the 70th percentile. (Images courtesy of James Carr, M.D., Northwestern University, Chicago.)
The risk of a major coronary event (death or nonfatal myocardial infarction) with increasing coronary-artery calcium (CAC) scores, adjusted for standard risk factors, was determined in 6722 initially asymptomatic healthy adults who were followed for a median of 3.8 years. Higher CAC scores were associated with higher event rates, but even with high CAC scores, the event rate was only approximately 1% per year. A CAC score of 0 indicates no detectable calcium; higher CAC scores indicate more severe degrees of coronary calcification, with the extent of severity influenced by age, ethnicity, and sex. CI denotes confidence interval. Data are from Detrano et al.\textsuperscript{18}

Population studies indicate that CAC scores increase with advancing age, reflecting the natural progression of atherosclerosis, and men generally have higher CAC scores than women of similar age. Women tend to have calcium scores similar to those of men who are 15 years younger. CAC scores are often reported as percentiles of calcification in a reference population according to age and sex. Younger people (men younger than 45 years and women younger than 55 years) may have underlying atherosclerosis with little or no CAC, yet any degree of CAC would place a man or woman in these age groups in a very high percentile for the risk of coronary events as compared with the risk among controls matched for age and sex (Fig. 1). Similarly, the Framingham risk score does not classify most young adults as being at high risk, since in this age group, coronary events are unlikely to occur within 10 years, even in young adults with multiple risk factors. Estimates of the lifetime risk of coronary events, rather than the 10-year risk, may be more appropriate for this age group.\textsuperscript{23} Distributions of CAC scores adjusted for age and sex also vary according to race and ethnic group.\textsuperscript{22,24}

Since the CAC score indicates the presence or absence and measures the extent of coronary atherosclerosis, it is not surprising that multiple studies have shown that a high CAC score is a marker for an increased risk of coronary events. Thus, a CAC score of zero is associated with a very low risk of subsequent coronary events,\textsuperscript{25,26} whereas increasing CAC scores are associated with a stepwise increase in the risk of events. Numerous studies indicate that the CAC score correlates with the number and severity of conventional risk factors but is also an independent marker for the risk of coronary events, after adjustment for these conventional risk factors.\textsuperscript{18,24,27-38} For example, the Multi-Ethnic Study of Atherosclerosis (MESA) (ClinicalTrials.gov number, NCT00005487), a cohort study of 6722 initially asymptomatic adults (38.6% white, 27.6% black, 21.9% Hispanic, and 11.9% Chinese),\textsuperscript{18} has provided strong evidence of the incremental prognostic information provided by the CAC score (Fig. 2). No major differences in the predictive value of CAC scores were noted among racial or ethnic groups. Although MESA showed a clear gradient of risk as CAC scores increased, the absolute risk of events was low (roughly 1% per year, even among participants with high scores).\textsuperscript{18}

The association between CAC scores and cardiovascular events, which has been well documented, is not unexpected, since the presence and extent of CAC reflect the actual presence and severity of atherosclerosis, whereas risk factors, risk scores, and biomarkers are merely markers for the likelihood of the disease. However, it remains unclear whether CAC scanning has a favorable effect on clinical outcomes, and there are concerns regarding the associated radiation exposure.\textsuperscript{39-42} The usual radiation burden associated with CAC scanning is small but real (generally, 0.6 to 1.0 mSv for EBCT and 0.9 to 2.0 mSv for MDCT),\textsuperscript{44} and some MDCT imaging protocols are associated with estimated radiation doses higher than 10
mSv. In comparison, the standard chest radiograph yields a radiation dose of 0.01 to 0.02 mSv. An effective dose of 2.3 mSv is estimated to result in a small measurable increase in the risk of cancer, and this estimate would need to be considered if CAC testing (and repeated testing) were used for widespread population screening.

There are also uncertainties with regard to how, when, and in whom the test should be performed, what CAC-score threshold should trigger more aggressive treatment of risk factors, and what the most appropriate treatment is once that threshold is crossed. In the absence of data on outcomes, the most appropriate treatment is once that threshold is crossed. In the absence of data on outcomes, the most appropriate treatment is once that threshold is crossed. In the absence of data on outcomes, the most appropriate treatment is once that threshold is crossed. In the absence of data on outcomes, the most appropriate treatment is once that threshold is crossed. In the absence of data on outcomes, the most appropriate treatment is once that threshold is crossed.

In persons categorized as having a low risk of coronary events according to the ATP III criteria, the presence of coronary calcium increases the risk of future events, but even among those in this group with a high CAC score, the likelihood of coronary events remains low (Fig. 3). Similarly, for those who are in the high-risk category according to the ATP III criteria, a low CAC score does not eliminate the risk of an unfavorable outcome associated with an ominous risk-factor profile, and aggressive attention to risk reduction is warranted regardless of the CAC score. If any group might benefit from the additive risk stratification provided by CAC scoring, it might be persons initially identified by the Framingham criteria as having an intermediate risk of coronary events — that is, those with a 10-year risk of 10 to 20% (Fig. 3). An absence of coronary calcium in these patients might shift them to a low-risk category, whereas a high CAC score might indicate a higher level of risk that calls for more intensive treatment of modifiable risk factors. Furthermore, women in general tend to be categorized as having a low risk according to the ATP III criteria, and a high CAC score appears to identify a group with a substantively higher risk of coronary events than their low-risk profile in the Framingham study would suggest.

The goal of CAC scanning in asymptomatic persons is to refine the risk assessment in order to determine whether preventive strategies should be intensified, not to identify persons with asymptomatic coronary-artery stenoses. However, high CAC scores often lead to additional downstream testing and procedures, including stress imaging, coronary angiography, and even percutaneous coronary interventions. Although the highest CAC scores represent substantial plaque burdens with the potential to encroach on the coronary lumen, there is no evidence that interventions other than risk-factor modification will lead to a better outcome in patients without symptoms. Widespread CAC screening runs the risk of increasing the number of unnecessary tests and procedures downstream — and of escalating health care costs. A corollary to this view is that CAC assessment is unlikely to alter the treatment plan for patients whose blood pressure and lipid levels are already well controlled.

**AREAS OF UNCERTAINTY**

The major unresolved question is whether routine, widespread CAC screening or determination of the CAC score in an individual patient will lead to an overall improvement in quality of care and clinical outcomes. The identification of patients at increased risk is useful only if it leads to a successful strategy that helps avert future coronary events. Since no studies have been designed to demon-
strate this effect, there are no convincing data to suggest that this desired result can be achieved. Studies evaluating surrogate endpoints related to patient and physician behavior have had mixed results. Two observational studies reported that knowledge of a positive CAC scan was associated with greater use of statin and aspirin therapy and risk-reducing changes in lifestyle, but these studies cannot prove cause and effect. Moreover, in a randomized trial, patients who were informed of the results of CAC testing did not have a favorable improvement in their risk-factor profile after 1 year; the results for these patients were no different from the results for patients who were not informed of the results of testing. Knowledge of the CAC score does, however, lead to more anxiety, more hospitalizations, and more revascularization procedures.

The final quandary concerning CAC tests is typical of screening tests in general. Should the subgroup with the highest CAC scores, identified by screening, become the target population for intensive risk-reduction measures? This group with the highest relative risk will represent only a small segment of the total population. The much larger group with low CAC scores will have a much lower relative risk, but because of its sheer magnitude, this group is likely to include persons who are destined to have at least as many coronary events as those in the high-scoring group. This is borne out in most of the longitudinal studies examining the relation between CAC scores and outcome (Table 1). One would have to treat virtually all patients with even minimal CAC to achieve the goal of screening, thereby negating the value of the screening program. Twenty years ago, my colleagues and I pointed out this inevitability of cardiovascular screening in an article that focused on stress testing, and it is demonstrated once again with CAC scoring.

**Table 1. Coronary-Artery Calcium Scores and Outcomes for Asymptomatic Subjects in Eight Studies.**

<table>
<thead>
<tr>
<th>Study, Outcome Reported, and CAC Score †</th>
<th>Event Rate</th>
<th>Total No. of Patients</th>
<th>No. of Patients with Event</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Shaw et al., death at 5 yr</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAC score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1.0</td>
<td>5,915</td>
<td>59</td>
</tr>
<tr>
<td>1–100</td>
<td>1.4</td>
<td>7,990</td>
<td>111</td>
</tr>
<tr>
<td>&gt;1000</td>
<td>5.0</td>
<td>311</td>
<td>16</td>
</tr>
<tr>
<td><strong>Budoff et al., death at 10 yr</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAC score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–100</td>
<td>1.1</td>
<td>19,643</td>
<td>226</td>
</tr>
<tr>
<td>&gt;1000</td>
<td>12.2</td>
<td>964</td>
<td>118</td>
</tr>
<tr>
<td><strong>Detrano et al., death or MI at 3.8 yr</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAC score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–100</td>
<td>0.6</td>
<td>5,137</td>
<td>33</td>
</tr>
<tr>
<td>&gt;300</td>
<td>3.8</td>
<td>833</td>
<td>32</td>
</tr>
<tr>
<td><strong>Nasir et al., death at 8 yr</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAC score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–100</td>
<td>2.8</td>
<td>10,519</td>
<td>296</td>
</tr>
<tr>
<td>&gt;1000</td>
<td>25.9</td>
<td>819</td>
<td>212</td>
</tr>
<tr>
<td><strong>Raggi et al., death or MI at 32 mo</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAC score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–100</td>
<td>2.5</td>
<td>511</td>
<td>13</td>
</tr>
<tr>
<td>&gt;400</td>
<td>12.8</td>
<td>47</td>
<td>6</td>
</tr>
<tr>
<td><strong>Greenland et al., death or MI at 7 yr</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAC score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–100</td>
<td>5.5</td>
<td>637</td>
<td>35</td>
</tr>
<tr>
<td>&gt;300</td>
<td>15.4</td>
<td>221</td>
<td>34</td>
</tr>
<tr>
<td><strong>Vliegenthart et al., death at 3.3 yr</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAC score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–100</td>
<td>3.2</td>
<td>905</td>
<td>29</td>
</tr>
<tr>
<td>&gt;1000</td>
<td>12.2</td>
<td>196</td>
<td>24</td>
</tr>
<tr>
<td><strong>Lakoski et al., CHD event at 3.8 yr‡</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAC score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–99</td>
<td>0.6</td>
<td>2403</td>
<td>14</td>
</tr>
<tr>
<td>&gt;300</td>
<td>6.7</td>
<td>105</td>
<td>7</td>
</tr>
</tbody>
</table>

*CAC denotes coronary-artery calcium, CHD coronary heart disease, and MI myocardial infarction.

† A CAC score of 0 indicates no detectable calcium; higher CAC scores indicate more severe degrees of coronary calcification, with the extent of severity influenced by age, ethnicity, and sex.

‡ The participants in this study were women with low Framingham risk scores.

§ A CHD event was defined as death from coronary disease, nonfatal myocardial infarction, development of angina, coronary revascularization for chest pain, or resuscitation after cardiac arrest.

Widespread CAC screening of the asymptomatic adult population is not currently recommended because of the lack of prospective data showing that such screening ultimately results in improved outcomes and reduced coronary events, as well as the inherent limitations of screening. The U.S. Preventive Services Task Force recommends that adults at low risk for coronary events not under-
go routine screening and has found that there is insufficient evidence to make a recommendation for or against routine screening of those at high risk for events. The expert-panel statement from Prevention Conference V, sponsored by the American Heart Association in 2000, does not advocate routine CAC assessment, although it does note that such an assessment appears to offer the greatest potential value for the group that the Framingham criteria place at intermediate risk. The ATP III guidelines that were published a year later concur with this recommendation against the use of indiscriminate screening, noting that CAC scoring might be an option for advanced risk assessment in appropriately selected patients with multiple risk factors, provided the test is ordered by a physician who is familiar with the strengths and weaknesses of noninvasive testing. Whether updated recommendations will be forthcoming from the newly constituted Adult Treatment Panel IV is uncertain at this time.

The scientific statement from the American Heart Association and the expert-consensus document from the American College of Cardiology–American Heart Association conclude that it may be reasonable to consider CAC screening in asymptomatic persons identified as having an intermediate risk of coronary events on the basis of an assessment of multiple risk factors; this view is based on the possibility that such patients might be reclassified in a higher risk group on the basis of the CAC score and that the management of risk factors might then be intensified. As noted previously, no studies have shown that improved outcomes are associated with this approach. The lack of supporting data is reflected in the report on the appropriate use of cardiac CT developed by the American College of Cardiology and partner organizations, which concluded that CAC screening is inappropriate in asymptomatic patients who are at low risk for coronary events according to the ATP III criteria; the authors of the report were uncertain about the appropriateness of screening for those at intermediate or high risk.

**Conclusions and Recommendations**

A broad population-based strategy of CAC screening does not appear to be warranted. It is not clear whether it is reasonable to consider CAC scanning in persons whose global risk assessment places them in the intermediate-risk category or whether the findings from such testing will lead to a beneficial increase in the intensity of treatment. This issue needs to be addressed in future trials focusing on clinical outcomes and cost-effectiveness.

The patient presented in the vignette has a low Framingham risk score (less than a 10% risk of a coronary event over the next 10 years). I would not recommend CAC scanning but would discuss with him the absence of evidence that the use of this test improves outcomes, as well as the downside of testing, including the radiation burden and the possibility that the results will lead to further unnecessary testing. I would start therapy with aspirin and a statin and counsel him to initiate a program of regular exercise several times per week.

No potential conflict of interest relevant to this article was reported.

**References**

8. Greenland P, Knoll MD, Stamler J, et al. Major risk factors as antecedents of
fetal and nonfetal coronary heart disease events. JAMA 2003;290:891-7.


44. He ZX, Hedrick TD, Pratt CM, et al. Severity of coronary artery calcification by electron beam computed tomography
predicts silent myocardial ischemia. Circulation 2000;101:244-51.


47. O'Malley PG, Feuerstein IM, Taylor AJ. Impact of electron beam tomography, with or without case management, on motivation, behavioral change, and cardiovascular risk profile: a randomized controlled trial. JAMA 2003;289:2215-23.


Clinical trials bureaucracy: unintended consequences of well-intentioned policy
Robert M Califf
Clin Trials 2006; 3: 496
DOI: 10.1177/1740774506073173

The online version of this article can be found at:
http://ctj.sagepub.com/cgi/content/abstract/3/6/496

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
SCT
The Society for Clinical Trials

Additional services and information for Clinical Trials can be found at:
Email Alerts: http://ctj.sagepub.com/cgi/alerts
Subscriptions: http://ctj.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.co.uk/journalsPermissions.nav

Citations (this article cites 30 articles hosted on the SAGE Journals Online and HighWire Press platforms):
http://ctj.sagepub.com/cgi/content/refs/3/6/496
Clinical trials bureaucracy: unintended consequences of well-intentioned policy

Robert M Califf

Background: As randomized controlled trials have become the ‘gold standard’ for medical research, a complex regulatory structure for the conduct of clinical trials has emerged. However, this structure has not been adequately assessed to ensure that regulations governing human subjects research actually produce the desired effects.

Purpose: Our purpose is to identify some of the major shortcomings in the current regulatory system of human clinical trials oversight, and to propose some potential solutions to these problems.

Methods: We discuss the evolution of the current US regulatory environment and its application in the context of several widely-used drug therapies.

Results: Despite numerous randomized controlled trials, performed within a structure of extensive documentation and data collection, serious shortcomings in a number of pharmaceutical therapies were not detected until after the drugs were approved and widely adopted by clinicians.

Conclusion: The current system of regulatory bureaucracy in clinical trials has led to an extremely expensive research paradigm that, in spite of complex systems of oversight and exhaustive data collection, cannot be shown to adequately ensure the integrity of the research process and the protection of human research subjects. Some parts of the system, including Research Ethics Review Boards, may not be well-suited to carrying out their core mission of overseeing research conduct, and other aspects of clinical trials regulatory structure, such as monitoring/auditing review and adverse event reporting, may constitute a waste of money and resources. Misdirected data collection and adverse events reporting divert valuable resources and hamper development of large, simple clinical trials powered to definitively answer important research questions. Careful scrutiny of the utility of current or proposed regulatory schemes is required to ensure the integrity of human subjects research and to enhance the effectiveness of research dollars.


Introduction

The tool of randomization represents a critical advance that has enabled modern medicine to distinguish modest but important effects, both favorable and detrimental, of medical technologies on patients’ well-being. Although the first modern multicenter trial was reported in 1947 in the British Medical Journal [1], it was only in the 1970s that the clinical trial became a widely-accepted tool for the biomedical enterprise. Over the past 30 years, layer after layer of regulation has been added to the conduct of clinical trials, but there has been little accountability for whether this bureaucracy is having the net effect of improving the human condition, or whether it is unnecessarily impeding the conduct of complex endeavors critical to guiding the appropriate use of technology [2].
A major milestone in the history of this scientific tool was the decision by the US Food and Drug Administration (FDA) to review the evidence for all pharmaceuticals on the market in 1962 [3]. At that point, it was determined that for most drugs on the market, the evidence to determine the balance of risk and benefit was not available. Many products were withdrawn from the market and the randomized controlled trial (RCT) was established as the standard for determining the efficacy of therapies from the FDA perspective.

The adoption of this standard dramatically increased the diligence with which pharmaceutical companies pursued randomized trials but also led to an understandable desire for a regulatory structure under which the interpretation of regulatory trials could be accepted and interpreted. When problems were found with the conduct of RCTs, new rules were developed to prevent recurrence of the issues [4], but the complex maze of rules was not reviewed to eliminate those without benefit. Indeed, this situation led to multiple poorly-coordinated efforts to codify standard operating procedures (SOPs) for the conduct of trials and ultimately created an entirely new industry: the contract research organization (CRO). CROs have become finely attuned to producing trials that satisfy the stated needs of FDA officials and regulatory departments within the medical technology industry; these entities (FDA, regulatory departments and CROs) wield tremendous power in determining the study designs that ultimately define which products are marketed and how those products are labeled.

A second series of demands for regulation, and therefore creation of bureaucracy, resulted from the increasingly widespread realization that clinical trials are, in essence, human experiments. As recognition of the potential for abuse of volunteers in human studies mounted, regulations intended to protect human subjects proliferated [5]. A series of controversial failures in the system [6,7] recently have led to an even more stringent set of rules intended to prevent investigators and corporations from exposing subjects to undue conflicts of interest or unnecessary risks as a result of participating in a clinical trial.

The end result of this expression of regulatory and ethical concerns about clinical trials has been the development of a series of well-intentioned regulations and business standards. Due at least in part to the nature of the regulations and the risk of penalty and public humiliation if regulators decide that a breach of the rules has occurred, few have challenged each new layer of bureaucracy. This situation has culminated in a feeling among many investigators in the United States that the regulatory burden has become so great that the risks of participating as an investigator exceed the potential benefits. Furthermore, the cost of regulatory compliance has become large enough to create a major incentive for corporations to move research to countries with a much lower cost structure for doing the same work. Fundamentally, the transactional cost of the bureaucracy originally intended to improve the process of clinical research threatens its viability.

**Definitions**

The definition of bureaucracy provides insight into issues that need to be addressed in ‘fixing’ the perceived problem of excessive bureaucracy. Merriam-Webster’s dictionary defines bureaucracy as a body of nonelective government officials or an administrative policy-making group. More specifically, a bureaucracy can be described as a government (or governing entity) characterized by specialization of functions, adherence to fixed rules, and a hierarchy of authority. Finally and of most concern to clinical researchers is the definition: ‘a system of administration marked by officialism, red tape, and proliferation.’ A tenet of this manuscript is that while a bureaucracy is needed, that bureaucracy should be one that actively avoids this latter definition. Instead, the bureaucracy should be characterized by an unceasing effort to optimize the function of clinical research, and clinical trials as components of clinical research, to answer important questions about medical technologies and practice while also protecting human subjects from harm due to poor research conduct [8].

**Clinical trials bureaucracy: the general landscape**

The overall ‘megatrends’ in health care and biomedical research cry out for a reassessment of the bureaucracy that regulates research efforts. The aging of the population in the economically developed world and the rapid population growth and rapidly-emerging health systems in developing countries are leading to significant financial pressure on health care systems. This societal financial pressure, combined with the rapid evolution of consumerism at the level of the individual, has produced substantial demand for an understanding of the relationships between value and cost in medical technologies. The knowledge base for medical therapeutics has evolved so that clinical trials are recognized as the most reliable evidence upon which to base our decisions about appropriate medical therapies [9]. Most governments and professional societies are vested in the development of clinical practice guidelines and technology assessments in which the randomized clinical trial is viewed as the highest level of evidence [10,11].
The post-genomic revolution will increase the pressure on the system to produce irrefutable evidence about which technologies are effective and which are ineffective or even dangerous. Knowledge of genes, proteins, and metabolites, along with functional imaging, will divide populations into multiple 'phenotypes', creating both an unprecedented opportunity to improve clinical outcomes as well as an enormous subgroup issue that will require many more trials in larger numbers of human subjects. As the market becomes flooded with inadequately-studied tests used to stratify treatment, the demand for adequate comparative studies will understandably increase. With current regulations, however, the cost of the required increase in clinical trials enrollment will be prohibitive.

A broad system for optimizing medical therapeutics has now been envisioned. The conduct of a larger number of pragmatic clinical trials is absolutely essential to the generation of the level of evidence needed to guide practice. When the right trials are done, definitive evidence is developed that can be used to create clinical practice guidelines. From clinical practice guidelines, the context of appropriate practice can be described and performance metrics for practice can be generated [12]. We now know that these metrics can be used to improve practice, thereby reducing death and disability [13]. We also know that practice can be examined to continuously improve delivery. However, this 'cycle of quality' is completely dependent upon the foundation of well-designed pragmatic clinical trials [14]. The combination of increasing societal awareness of the critical nature of clinical trials in defining appropriate uses of technology, coupled with the implications of the genomic revolution, requires a radical restructuring of bureaucracy to permit many more subjects to participate in trials at a much lower cost.

The general problem

Given the proliferation of regulations and standards governing clinical trials, one would hope that the effort is yielding increasingly efficient results with fewer problems, including reductions in violations of human research subject protections [15]. Instead, we have ample evidence that the research system has become increasingly expensive, and the few studies that are available indicate that the research dollar is producing fewer results [16]. As clinical research expenses have skyrocketed, the number of new molecular entities reaching the market has declined in recent years. Furthermore, when the recent histories of late drug failures are reviewed [17], the problem that emerged was not failure to adhere to regulations, but expenditure of huge amounts of money on the collection of vast amounts of irrelevant data. These failures have occurred in many different fields of medicine.

The failure of the Type I ‘anti-arrrhythmic’ drugs provides the prototype example. Multiple clinical trials were completed to demonstrate that these drugs reduced the frequency of ventricular premature beats in patients thought to be at risk of sudden death. None of the trials was large enough to determine the actual effect on death. Thousands of pages of data were collected and adverse events were dutifully reported to regulatory authorities. Unfortunately, these trials did not ask the fundamentally simpler but more important question: did these drugs reduce or increase the risk of sudden death? Unfortunately, the answer came after the drugs had been on the market for years: they did indeed increase rather than decrease the risk of sudden death [18]. Thus, the bureaucracy diverted attention from the critical question and created a cost structure that delayed the crucial trials for years while the drugs were being used in practice in a lethal manner.

The field of women’s health was rocked by the demonstration in two consecutive trials that the use of hormone replacement therapy (HRT) in postmenopausal women increased rather than decreased cardiovascular events [19,20]. At the time these trials were performed, HRT was the most commonly prescribed drug therapy globally. Many trials had been done with small numbers of women with the intention of providing mechanistic insight into the ‘benefits’; however, little consideration was given to the potential for harm. Additionally, years of adverse event reporting and mountains of detailed data collection in these small samples failed to uncover the problem. The cost of the Women’s Health Initiative project (over $200 million) has been used as a reason to recommend decreasing the public funding of clinical trials [21], yet without the trial this treatment would still be leading to thousands of unnecessary deaths, strokes and myocardial infarctions in the hands of well-intentioned doctors and postmenopausal women.

The exposure of large numbers of patients to COX-2 inhibitors occurred before their effect on cardiovascular events was discovered. Again, the pattern of failure involved multiple small clinical trials, with huge amounts of data collected on each patient at a tremendous cost. The studies were heavily monitored: fortunes were spent on flying highly-paid monitors as often as once per month to investigational sites, where all data items were checked to ensure that the data submitted to the coordinating center matched the clinical source documentation. As in the cases of Type I anti-arrhythmics and HRT, one cannot help but wonder whether spending the same amount of money on
large studies with comparisons of clinical outcomes, while at the same time dispensing with all the monitoring and source documentation, could have achieved a reliable estimate of the balance of risk and benefit.

We are currently experiencing a period in which multiple pragmatic questions are plaguing drugs commonly used in the treatment of mental health problems. Selective serotonin reuptake inhibitor (SSRI) antidepressants seem to cause an increase in suicidal ideation [22], and 10% of American boys are said to be taking drugs to treat attention deficit disorder that are known to raise heart rate and blood pressure, and for which anecdotal cases of sudden death have been reported [23]. Perhaps no field other than mental health drug development has conducted so many small trials with mountains of data with so little understanding of the balance of risks and benefits of taking the drugs.

All of these failures have in common a tremendous investment in detailed data collection, extensive auditing of the data collected, and scrutiny that the hypotheses tested in the studies were confirmed. These expensive methods would not have allowed early detection of these problems, because much larger size trials would have been needed and the costs would have been untenable. Given the propensity among those who do not fully comprehend the risks involved in skipping evaluative clinical trials to advocate the use of unproven putative surrogate endpoints as a cost-saving measure [24], it is especially critical to eschew unnecessary costs so that these vital pragmatic trials can be done.

**Research ethics review boards, institutional review boards and ethics committees**

Few would quibble with the concept that human studies should have oversight by an unbiased group of people representing diverse interests. Yet the system of research ethics review boards (RERBs), also known as ethics committees or institutional review boards (IRBs), has taken on a life of its own. Bureaucracy proliferates as new rules are added while outdated or even useless regulations often remain in place [25–27]. Indeed, almost no empirical studies have been done to determine which aspects of human subjects protection oversight are effective and which are useless or even counterproductive.

In May 2000 a multidisciplinary meeting involving investigators, regulators, RERB chairs, and industry representatives was held to address this problem [28]. But while progress has been made since 2000 in some areas, it would be difficult to conclude that the bureaucracy has been ‘right-sized’ for the function of human subjects protection oversight.

At this meeting, numerous gaps in the functioning of the system were identified. RERBs were thought to be incapable of making safety assessments based on the data received from multicenter trials, except in the case of rare, idiosyncratic adverse events. An enormous duplication of effort occurs in multicenter trials, with each RERB reviewing consent forms and looking at the same useless reports of adverse events accruing from all over the world. Empirical evidence indicates that tampering with consent forms by local RERBs often renders them harder to understand, and despite all of our efforts subjects frequently cannot demonstrate that they truly understand the nature of the study in which they have agreed to participate [29]. Recent empirical data [30] indicate that the present fervor over informing subjects about details of potential conflicts of interest may have been overblown from the perspective of what subjects actually want to know.

Most of the misplaced and very expensive efforts of local RERBs would be better performed by well-functioning data monitoring committees (DMCs). Only in 2002 was the first textbook for DMCs written, and the ‘rules of engagement’ are still in evolution [31]. However, it is obvious that in a multicenter trial, each local RERB is not in a position to interpret individual adverse event reports, particularly when the RERB is neither unblinded nor aware of the denominator that forms the essential context for interpreting adverse events in trials that are still enrolling subjects. Despite dozens of meetings and statements decrying this absurd generation of mountains of expensive paper with almost nothing to show for it, the policies have yet to change. Furthermore, DMCs are now under the gun of the legal system, causing an increase in the need for expensive documentation and indemnification for those who fill this vital societal role [32].

While RERBs remain bogged down reviewing protocols and consent forms, there is little to no oversight of the actual conduct of clinical trials. Such an effort would require real-time, sample-based observation of the functioning of the research team as it interacts with human subjects, together with interviews of subjects to determine the degree to which they understood the issues raised by participation in the trial.

**Adverse events reporting and collection**

When medical products (drugs, devices or biologics) are made available to prevent or treat illness, a system for reporting and tracking adverse events related to the use of the technology makes sense.
Indeed, despite its shortcomings [33], the spontaneous adverse events reporting system currently in place has resulted in the identification of a number of adverse effects of drugs and devices that have led to modifications in use or withdrawal from the market. However, partial success should not justify continued proliferation of useless components. In particular, many companies have interpreted regulations as requiring that all adverse events be reported to all RERBs participating in clinical trials with the drug or device, a practice that costs hundreds of millions of dollars. When RERBs receive these useless documents, they feel compelled to assign a clinical expert to assess the report. The mailing and filing costs alone are enormous, but added to this is the incalculable cost of professional time, much of it provided on a voluntary basis. The FDA recently developed a tentative guidance that would dramatically reduce the proliferation of adverse event reports, but this guidance has not been finalized, and the practice continues.

Auditing and monitoring

Well-publicized examples of untruthful behavior by clinical investigators have fueled an understandable interest in the creation of systems to ensure that investigators are both truthful and accurate when they report clinical research data. The response to this interest has been to develop an elaborate system of data auditing in clinical trials that are submitted for regulatory review. For many industry-sponsored studies, the SOPs call for every case report form to be reviewed in person by a monitor who physically visits the enrolling site. This is a very expensive process, as study monitors are well-paid and incur significant travel and lodging expenses.

This labor-intensive approach has been defended from two perspectives. First, a company hoping to gain regulatory approval of its product for marketing can ill afford to fail an audit from regulators, with the result that extra up-front expense is justified. Second, there is deep concern about the veracity of investigators as they report data in the course of clinical trials. On-site auditing to detect fraud and correct sloppiness, and the extensive efforts to verify the accuracy of the data, meet ‘common sense’ views of how to produce reliable research results. This rationale is reinforced by the FDA policy of on-site auditing, often performed by individuals with little knowledge of the clinical material, which relegates them to checking process-related issues and examining concordance of different data sources. Unfortunately errors, and even on occasion frank fraud, are found often enough to lend credence to the concept of on-site auditing.

One effect of extensive auditing is the generation of ‘data queries’, a process by which sites are solicited to correct or verify possibly discrepant or incorrect data. An entire industry has developed around creating, communicating, and archiving the documentation necessary for verifying the nature of, and individuals responsible for, changes made to study records. It is estimated that the cost of generating and resolving a single query is over $150. Yet only a few empirical studies have been done to examine the merits of this practice, and these studies indicate that it may be a colossal waste of money.

An empirical approach is sorely needed to develop quality standards in clinical trials that lead to reliable results at the lowest cost. This approach will involve statistical sampling in the same manner as utilized in other industries. By examining patterns in the reported data, answers that appear either impossible or too ‘perfect’ can be detected. Indeed, studies of fraud indicate that statistical process control will more often detect fraud than on-site auditing, since someone clever enough to attempt to fabricate data will be inclined to prepare for an audit [34–37].

Dollars spent

The implications of the issues enumerated are profound in terms of the efficiency and effectiveness of clinical research. One view of this problem was generated from an exercise recently published in the American Heart Journal [38]. A group of academics, industry experts, and FDA officials convened to discuss the cost of conducting clinical trials. Two hypothetical clinical trials were developed and a group of companies applied their operational models to determine the time frames and resources needed to conduct the trials. The average total cost for a trial in acute coronary syndromes was $83 million, but estimates ranged from $57 million to $158 million. A heart failure trial was priced at an average of $135 million, with a range of $102 to $207 million. The cost of managing research sites (monitoring and data auditing) and payments to the investigators accounted for >65% of total costs, with the main drivers of these costs being excessive data collection and monitoring visits, with the associated proliferation of travel and labor costs.

Possible solutions

The development of the internet has made it possible to automate much of what first developed as a tedious and expensive set of manual processes in the conduct of clinical trials. The current system is
driven by companies seeking guidance in regulatory and reimbursement issues, as well as seeking marketing advantages for their products, and also by officials appointed to prevent poorly-conducted research of the type that has led to public concerns. Investigators are ‘hired’ to conduct the research under a set of regulations as described above. A parallel system, funded by government agencies to answer questions driven by the academic scientific community, exists in some countries. These parallel universes overlap at the level of individual research sites, but neither functions optimally in answering the questions that would best inform decisions about medical care.

An alternative approach would be to create networks of investigators motivated to answer questions of clinical relevance for both the medical products industry and the health policy communities. This infrastructure could be funded by governments that maintain a continuous record of the quality of site performance. Industry could use these sanctioned networks without bearing the cost of creating expensive data collection instruments and data auditing systems. As electronic health records are phased in, this possibility will become even more substantial, creating the opportunity to maintain ‘interoperable’ networks that could evaluate technologies and behavioral interventions across multiple disease categories. Statistical process control could be used to prompt cause-specific audits based on data patterns, and fees to perform research activities could be reimbursed at a professional level directly to staff actually dealing with subjects, instead of to auditors [39]. Both the US National Institutes of Health and the British Medical Research Council are making initial efforts in this direction. The networks themselves could participate in the design of the studies to ensure that they are feasible and properly directed to answer important questions.

If these efforts succeed, the bureaucracy currently devoted to processes that have not been tuned to optimize the societal value of research could focus on driving the systems toward answering questions that would better inform medical decision-making. Such an approach could also reverse the worrisome trend of control of data by the medical products industry instead of by the investigators [40].

Conclusion

While bureaucracy is necessary in providing rules and guidance for the conduct of clinical research, the system has become unnecessarily complex and expensive. Many well-intentioned regulations have resulted in huge amounts of effort and paperwork with little or no measurable positive impact on the ethical conduct of clinical research. However, it is clear that these regulations and standards have dramatically increased the cost of clinical trials, thereby making the enterprise much less efficient in its primary mission of answering questions that inform decisions for patients and consumers, doctors and other health care providers and administrators. Correction of this enormous societal waste is critical not just to the clinical trials industry or the industries that develop and sell medical products. Rather, since the pragmatic clinical trial is the cornerstone of evidence upon which the quality of medical care is built, tuning the logistics to the most efficient delivery of meaningful results is critical to the future of medicine.

References


http://ctj.sagepub.com


Diagnostic tests are used to confirm, exclude, classify, or monitor disease to guide treatment. Their clinical value depends on whether the information they provide leads to improved patient outcomes; this can be assessed by randomized trials that compare patient outcomes from the new diagnostic test versus the old test strategy. However, randomized trials of test-and-treatment strategies are not routinely performed. They are not required for marketing approval, and they are not always feasible because they require large sample sizes. As a result, new diagnostic tests frequently enter clinical practice without evidence of improved patient outcomes.

Studies of diagnostic test accuracy can show how well diagnostic strategies that include a new test identify the presence or absence of disease compared with an old test strategy by comparing each with the results of a reference standard test. However, if clinicians use only information about test accuracy to decide whether to adopt a new diagnostic test, they sometimes may harm patients (for example, if subsequent treatments are unsafe or ineffective). It is therefore worthwhile to investigate how best to decide if clinicians can rely on evidence about test accuracy or if they need to wait for patient outcome results from randomized trials. Fryback and Thornbury (1) described a hierarchy of 6 levels of evidence for the assessment of a diagnostic test: 1) the technical quality of test information; 2) diagnostic accuracy; 3) change in the referring physician’s diagnostic thinking; 4) change in the patient management plan; 5) change in patient outcomes; and 6) societal costs and benefits. This framework does not provide guidelines about if and when lower levels of evidence are adequate to assess a test and always requires randomized trials for conclusions about improved patient outcomes. Some researchers have suggested that accuracy studies alone may sometimes suffice (2–6). We provide a practical framework to help clinicians decide whether a new diagnostic test can be adopted on the basis of evidence of test accuracy alone or if they need to wait for results from randomized trials.

A FRAMEWORK FOR DECIDING IF EVIDENCE OF TEST ACCURACY WILL SUFFICE

Studies of diagnostic test accuracy can suffice if clinicians already have evidence from randomized trials showing that treatment of the cases detected by the diagnostic test improved patient outcomes (Figure 1). This approach may seem straightforward; however, it requires a clear understanding of the proposed use and benefits of the new test, as well as careful consideration of whether the cases detected are representative of the patients included in treatment trials.

The benefits of a new diagnostic test will vary according to how it is used. Investigators of a new diagnostic test need to explicitly state who will be tested, where the new test will fit in the existing diagnostic pathway, and what tests it will supplement or replace. This information will allow them to identify the expected benefits of adopting the new test and, therefore, the most relevant questions to ask to assess its value. Test attributes generally fall into 3 categories: 1) The test is safer or is less costly; 2) the test is more specific (excludes more cases of nondisease) and thus avoids unnecessary treatment; and 3) the test is more sensitive (detects more cases of disease) and thus promotes more appropriate treatment.

We propose a simple sequence of questions to guide decisions about whether evidence of test accuracy will suffice for each of these 3 categories (Figure 2). The first step in assessing a new diagnostic test is to classify it according to whether it is more sensitive than the old test. We describe the key concepts of this framework using simple examples to describe situations where the new test offers better safety or specificity with similar sensitivity, followed by...
consideration of situations where the test is more sensitive. We also consider other, more complex scenarios. In each of these examples, we have some existing evidence of treatment efficacy for cases detected by an old test, and therefore, the rationale for testing has already been established.

**When a New Test Has Similar Sensitivity to an Old Test**

If a new diagnostic and old diagnostic test have similar sensitivity, it is generally reasonable to assume that they will detect the same true cases of disease. However, a new test may offer other positive attributes, such as better safety or more specificity than an old test. If studies of test accuracy show that the new test offers other positive attributes without a loss of sensitivity, it is logical to assume that cases detected by either test will show the same response to treatment. Therefore, new trials assessing treatment efficacy in the cases detected by the new test are not needed.

**When the New Test Is More Sensitive than the Old Test**

If a new test is more sensitive than an old test but has similar specificity, its value is directly related to the treatment response in the extra cases detected. If treatment response has already been assessed by treatment trials enrolling patients detected by the new test, decisions to use the test will be based on whether these trials showed that treatment improves patient outcomes. In this instance, evidence of test accuracy linked with evidence of treatment efficacy replaces the need for new randomized trials (Figure 1).

There will also be a good match between tested and treated populations if the results for patients identified by a new test are analyzed in a treatment trial as potential predictors of treatment response. For example, trials of adjuvant tamoxifen among women with early breast cancer have shown that the estrogen receptor status of the tumor determines its response to treatment (10). These trials demonstrate the clinical value of testing for estrogen receptor status, as well as the effectiveness of tamoxifen therapy. However, more commonly, treatment trials have only en-
rolled cases detected by an old test. In these situations, clinicians need to consider whether the results apply to cases detected by the new test.

Assessing Whether the Extra Cases Detected by a New, More Sensitive Test Respond to Treatment

Clinicians first need to ask whether the extra cases detected by a new diagnostic test represent the same spectrum or subtypes of disease as those included in treatment trials. If they do, then randomized trials may not be required (Figure 2). If they do not, or if clinicians cannot be certain that they do, then clinicians also need to ask whether treatment response is known to be similar across the range of disease spectrum or subtype. These considerations are important regardless of the magnitude of the difference in sensitivity between tests.

At one extreme, a good match between tested and treated populations is possible if the reference standard used to determine test sensitivity and specificity is also the starting point for trials conducted in similar populations. For example, computed tomography colonography is more sensitive for the detection of large colorectal polyps when scanning is performed with the patient in prone and supine positions versus supine positioning alone, without reduced specificity (11). Treatment trials showing improved survival following the early detection and treatment of colorectal polyps have been based on cases detected by colonoscopy, the reference standard for the computed tomography colonography accuracy studies. All of these tests are used to identify the same disease characteristic (adenomatous colorectal polyps) and spectrum of disease (classified by polyp size), so it is reasonable to conclude that computed tomography colonography with dual positioning will improve pa-

---

RCT = randomized, controlled trial. * New test = diagnostic strategies that include the new test; old test = standard diagnostic strategies that do not include the new test.
tent outcomes compared with supine positioning alone (Table, example 3).

At the other extreme, a good match between tested and treated populations will not be possible if a new test measures a different biological characteristic to define disease and leads to a different selection of cases for treatment. For example, a new nuclear magnetic resonance spectroscopy technique to measure lipoprotein particle size and concentration has been proposed as a more accurate test than plasma lipoprotein cholesterol levels to identify patients who will benefit from lipid-lowering therapy (12). However, the effectiveness of treatment for the extra patients detected by the new rather than the old test has not been assessed (Table, example 4).

Often we need to consider situations between these 2 extremes, including situations where the evidence is incomplete. Accuracy studies and treatment trials are usually done by different investigators at different times and in different settings. Variations in the clinical setting and spectrum of disease may lead to genuine differences (heterogeneity) in test accuracy and treatment effect (13, 14), limiting clinicians’ ability to link evidence between studies and draw conclusions about the value of a new test. Consider magnetic resonance angiography versus conventional arteriography for detecting lower-limb arterial disease. Although magnetic resonance angiography is a more sensitive test for patients with claudication, the effectiveness of surgical and percutaneous interventions has been well established only in patients with critical limb ischemia (15). Whether this evidence can be linked to infer that magnetic resonance angiography will improve patient outcomes depends on whether clinicians can assume that test accuracy and treatment response are similar for patients with different disease severity.

A new test may also produce a shift in the spectrum of disease detected. For example, breast magnetic resonance imaging in addition to mammography is more sensitive than mammography alone for the early detection of invasive breast cancer in young women at high risk. However, the extra cases detected may represent a different spectrum of disease that does not show improved treatment response. Therefore, the value of breast magnetic resonance imaging

<table>
<thead>
<tr>
<th>Test and Indication</th>
<th>Proposed Benefits of New Diagnostic Test</th>
<th>Does the Evidence of Effective Treatment Apply to the Cases Detected by the New Test?</th>
<th>Can Studies of Test Accuracy Suffice?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1: Doppler ultrasonography vs. venography for detection of deep venous thrombosis</td>
<td>Safer, less costly</td>
<td>Yes; no change to the definition or spectrum of disease.</td>
<td>Yes; the value of the new test may be inferred from an assessment of its relative safety and cost.</td>
</tr>
<tr>
<td>Example 2: new vs. standard FOBT for early detection of colorectal cancer</td>
<td>More specific</td>
<td>Yes; no change to the definition or spectrum of disease.</td>
<td>Yes; the value of the new test may be inferred from an assessment of its relative safety and cost and the benefits of avoiding a false-positive result.</td>
</tr>
<tr>
<td>Example 3: supine and prone positioning for CT colonography vs. supine positioning alone for detection of adenomatous colorectal polyps</td>
<td>More sensitive</td>
<td>Yes; no change to the definition or spectrum of disease because cases detected by both tests are subsequently confirmed by colonoscopy.</td>
<td>Yes; the value of the new test may be inferred from an assessment of its relative sensitivity, given no substantial loss in specificity and existing trial evidence that treatment improves patient outcomes.</td>
</tr>
<tr>
<td>Example 4: NMR spectroscopy vs. plasma lipoprotein levels for the detection of hyperlipidemia</td>
<td>More sensitive</td>
<td>Unknown; is it unknown whether NMR detects patients who would benefit more or less from lipid-lowering agents than those whose disorder was detected using the existing test.</td>
<td>No; trial evidence is required to determine the value of the new test.</td>
</tr>
<tr>
<td>Example 5: MRI vs. mammography for earlier detection of invasive breast cancer in women at high risk for the disease</td>
<td>More sensitive</td>
<td>No; it is uncertain if any benefits of treatment at the earlier stage of disease outweigh the harm of overdetection of cancer that would never have presented clinically.</td>
<td>No; trial evidence is required to determine the impact of testing on patient outcomes or, at least, the interval breast cancer rate.</td>
</tr>
<tr>
<td>Example 6: PET, MRI, and EEG vs. MRI and EEG to detect an epileptogenic focus in patients with medically refractory epilepsy who are being considered for surgery</td>
<td>More sensitive</td>
<td>Uncertain; it is uncertain whether patients with functional lesions detected by PET for whom existing standard imaging yields negative or inconclusive results will show the same treatment response to surgery as patients with structural lesions that can be detected with standard imaging alone.</td>
<td>Judgment is needed about whether clinicians require randomized trials to assess treatment response in the extra cases detected by PET or whether they can rely on existing trials conducted in patients with disease detected by MRI and EEG and observational evidence about treatment response in cases detected by PET.</td>
</tr>
</tbody>
</table>

* CT = computed tomography; EEG = electroencephalography; FOBT = fecal occult blood test; MRI = magnetic resonance imaging; NMR = nuclear medicine resonance; PET = positron emission tomography.
is uncertain without trials comparing patient outcomes after early versus standard detection (16) (Table, example 5). When the possibility of a spectrum effect and its impact on treatment efficacy are considered, it is useful to determine whether a new test performs consistently in different populations and whether treatment of the detected condition is effective across different patient subgroups. Previous studies that identify subgroups in which the test is less accurate or the treatment less effective may be available and may assist in making judgments about generalizability between the cases detected by a new test and patients included in treatment trials. In many cases, generalizability will depend on expert opinion. Sometimes, differences in the spectrum of disease detected may be obvious because the new test detects disease of a different size, stage, grade, or severity. In other cases, this judgment will be less clear.

Consider the use of positron emission tomography as an incremental test to detect an operable epileptogenic focus in patients with medically refractory epilepsy who have no structural lesion on magnetic resonance imaging. There is trial evidence that seizure control improves after surgery for patients with structural lesions (17). However, the extra cases detected by positron emission tomography may represent a different spectrum of disease. Thus, clinicians’ decisions whether evidence of test accuracy will suffice will depend on whether it is plausible that the response to treatment is similar in patients with structural lesions (the treated population in this example) and those with functional lesions detected by positron emission tomography (positive results on the new test in the tested population) (Table, example 6). The willingness of clinicians to accept assumptions about generalizability depends on the estimated costs of a new test and the severity of the consequences if the assumptions are proved false. If the costs or consequences are substantial, clinicians should wait for direct evidence from trials that include the extra cases detected by the new test.

Other Considerations

The examples discussed here do not cover all possible scenarios. For example, it is possible for a new test to have overall sensitivity and specificity that are similar to those of an old test but to detect different cases of disease if it identifies patients at a different disease spectrum. However, we think our approach provides a reasonable and efficient framework to deal with most situations. It helps clinicians to quickly identify more complex situations where a new test offers a tradeoff between positive and negative attributes (for example, if a new test is less invasive but also less specific than the existing test). Here, the benefits of better safety need to be assessed against the harms arising from additional false-positive results. Such tradeoffs can be assessed by using a decision analytic model to assess the benefits and harms of new and old tests, including the rates and the consequences of true-positive, false-positive, true-negative, and false-negative results, as well as test complications. Decision analysis also allows clinicians to compare the effects of testing in populations with a different prevalence of disease. If a new test offers better sensitivity but has other negative attributes, decision analysis may not be appropriate because it relies on assumptions that evidence of test accuracy can be linked to evidence of treatment efficacy. If this linkage is uncertain, clinicians will need to call for new randomized trials. In these situations, trials investigating the effect of treatment in patients who have positive results on the new test and negative results on the old test may be more efficient and more clinically relevant than trials conducted on all patients in whom the new test yields positive results (18).

Conclusion

Whenever clinicians use a new diagnostic test because it is more sensitive than an old test, they need to be clear about the assumptions linking this evidence to improved patient outcomes, such as evidence that the new test detects the same spectrum of disease as the old test or has similar treatment efficacy across the spectrum of disease. The selection of effective tests is just as essential as the selection of effective treatments; thus, whenever clinicians’ assumptions about a new diagnostic test are in doubt, they should wait for evidence from randomized trials.

From The University of Sydney, Sydney, Australia.

Grant Support: The authors have received funding from the Australian Medical Services Advisory Committee for the development of guidelines for the assessment of diagnostic technologies and National Health and Medical Research Council Program Grants (253602, 402764)

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Sarah J. Lord, MBBS, MS, National Health and Medical Research Council Clinical Trials Centre, University of Sydney, 88 Mallett Street, Camperdown NSW, 2050 Australia; e-mail, sally.lord@ctc.usyd.edu.au.

Current author addresses are available at www.annals.org.

References

12. Otvos JD, Jeyarajah EJ, Cromwell WC. Measurement issues related to lipoprotein.
A pragmatic–explanatory continuum indicator summary (PRECIS): a tool to help trial designers

Kevin E. Thorpe MMath, Merrick Zwarenstein MD MSc, Andrew D. Oxman MD, Shaun Treweek BSc PhD, Curt D. Furberg MD PhD, Douglas G. Altman DSc, Sean Tunis MD MSc, Eduardo Bergel PhD, Ian Harvey MB PhD, David J. Magid MD MPH, Kalipso Chalkidou MD PhD

Published at www.cmaj.ca on Apr. 16, 2009. An abridged version of this article appeared in the May 12 issue of CMAJ. This article was published simultaneously in the May 2009 issue of the Journal of Clinical Epidemiology (www.jclinepi.com).

Randomized trials have traditionally been broadly categorized as either an effectiveness trial or an efficacy trial, although we prefer the terms “pragmatic” and “explanatory.” Schwartz and Lellouch described these 2 approaches toward clinical trials in 1967. These authors coined the term “pragmatic” to describe trials that help users choose between options for care, and “explanatory” to describe trials that test causal research hypotheses (i.e., that a given intervention causes a particular benefit).

We take the view that, in general, pragmatic trials are primarily designed to determine the effects of an intervention under the usual conditions in which it will be applied, whereas explanatory trials are primarily designed to determine the effects of an intervention under ideal circumstances. Thus, these terms refer to a trial’s purpose and, in turn, structure. The degree to which this purpose is met depends on decisions about how the trial is designed and, ultimately, conducted.

Very few trials are purely pragmatic or explanatory. For example, in an otherwise explanatory trial, there may be some aspect of the intervention that is beyond the investigator’s control. Similarly, the act of conducting an otherwise pragmatic trial may impose some control resulting in the setting being not quite usual. For example, the very act of collecting data required for a trial that would not otherwise be collected in usual practice could be a sufficient trigger to modify participant behaviour in unanticipated ways. Further, several aspects of a trial are relevant, relating to choices of trial participants, health care practitioners, interventions, adherence to protocol and analysis. Thus, we are left with a multidimensional continuum rather than a dichotomy, and a particular trial may display varying levels of pragmatism across these dimensions.

In this article, we describe an effort to develop a tool to assess and display the position of any given trial within the pragmatic–explanatory continuum. The primary aim of this tool is to help trialists assess the degree to which design decisions align with the trial’s stated purpose (decision-making v. explanation). Our tool differs, therefore, from that of Gartlehner and associates in that it is intended to inform trial design rather than provide a method of classifying trials for the purpose of systematic reviews. It can, however, also be used by research funders, ethics committees, trial registers and journal editors to make the same assessment, provided trialists declare their intended purpose and adequately report their design decisions. Hence, reporting of pragmatic trials is addressed elsewhere.

Ten ways in which pragmatic and explanatory trials can differ

Trialists need to make design decisions in 10 domains that determine the extent to which a trial is pragmatic or explanatory. Explanatory randomized trials that seek to answer the question “Can this intervention work under ideal conditions?” address these 10 domains with a view to maximizing whatever favourable effects an intervention might possess. Table 1 illustrates how an explanatory trial, in its most extreme form, might approach these 10 domains.

Pragmatic randomized trials that seek to answer the question “Does this intervention work under usual conditions?” address these 10 domains in different ways when there are important differences between usual and ideal conditions. Table 1 illustrates the most extreme pragmatic response to these domains.

From the Dalla Lana School of Public Health (Thorpe) and the Department of Health Policy, Management and Evaluation (Zwarenstein), University of Toronto, Toronto, Ont.; the Centre for Health Services Sciences, Sunnybrook Research Institute, Sunnybrook Health Sciences Centre, and the Institute for Clinical Evaluative Sciences (Zwarenstein), Toronto, Ont.; the Preventive and International Health Care Unit, Norwegian Knowledge Centre for the Health Services (Oxman, Treweek), Oslo, Norway; the Division of Clinical and Population Sciences and Education, University of Dundee (Treweek), Dundee, UK; the Division of Public Health Sciences, Wake Forest University School of Medicine (Furberg), Winston-Salem, USA; the Centre for Statistics in Medicine, University of Oxford (Altman), Oxford, UK; the Center for Medical Technology Policy (Tunis), Baltimore, USA; the UNDP/UNFPA/WHO/World Bank Special Programme of Research, Development and Research Training in Human Reproduction, Department of Reproductive Health and Research, World Health Organization (Bergel), Geneva, Switzerland; the Faculty of Health, University of East Anglia (Harvey), Norwich, UK; the Institute for Health Research, Kaiser Permanente Colorado, and the Departments of Preventive Medicine and Biometrics and of Emergency Medicine, University of Colorado Health Sciences Center (Magid), Denver, USA; and the National Institute for Health and Clinical Excellence (Chalkidou), London, UK
Table 1: PRECIS domains illustrating the extremes of explanatory and pragmatic approaches to each domain

<table>
<thead>
<tr>
<th>Domain</th>
<th>Pragmatic trial</th>
<th>Explanatory trial</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Participants</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant eligibility criteria</td>
<td>All participants who have the condition of interest are enrolled, regardless of their anticipated risk, responsiveness, comorbidities or past compliance.</td>
<td>Stepwise selection criteria are applied that (a) restrict study individuals to those previously shown to be at highest risk of unfavourable outcomes, (b) further restrict these high-risk individuals to those who are thought likely to be highly responsive to the experimental intervention and (c) include just those high-risk, highly responsive study individuals who demonstrate high compliance with pretrial appointment-keeping and mock intervention.</td>
</tr>
<tr>
<td><strong>Interventions and expertise</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental intervention — flexibility</td>
<td>Instructions on how to apply the experimental intervention are highly flexible, offering practitioners considerable leeway in deciding how to formulate and apply it.</td>
<td>Inflexible experimental intervention, with strict instructions for every element.</td>
</tr>
<tr>
<td>Experimental intervention — practitioner expertise</td>
<td>The experimental intervention typically is applied by the full range of practitioners and in the full range of clinical settings, regardless of their expertise, with only ordinary attention to dose setting and side effects.</td>
<td>The experimental intervention is applied only by seasoned practitioners previously documented to have applied that intervention with high rates of success and low rates of complications, and in practice settings where the care delivery system and providers are highly experienced in managing the types of patients enrolled in the trial. The intervention often is closely monitored so that its “dose” can be optimized and its side effects treated; co-interventions against other disorders often are applied.</td>
</tr>
<tr>
<td><strong>Follow-up and outcomes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow-up intensity</td>
<td>No formal follow-up visits of study individuals. Instead, administrative databases (e.g., mortality registries) are searched for the detection of outcomes.</td>
<td>Study individuals are followed with many more frequent visits and more extensive data collection than would occur in routine practice, regardless of whether patients experienced any events.</td>
</tr>
<tr>
<td>Primary trial outcome</td>
<td>The primary outcome is an objectively measured, clinically meaningful outcome to the study participants. The outcome does not rely on central adjudication and is one that can be assessed under usual conditions (e.g., special tests or training are not required).</td>
<td>The outcome is known to be a direct and immediate consequence of the intervention. The outcome is often clinically meaningful but may sometimes (e.g., early dose-finding trials) be a surrogate marker of another downstream outcome of interest. It may also require specialized training or testing not normally used to determine outcome status or central adjudication.</td>
</tr>
<tr>
<td><strong>Compliance/adherence</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant compliance with “prescribed” intervention</td>
<td>There is unobtrusive (or no) measurement of participant compliance. No special strategies to maintain or improve compliance are used.</td>
<td>Study participants’ compliance with the intervention is monitored closely and may be a prerequisite for study entry. Both prophylactic strategies (to maintain) and “rescue” strategies (to regain) high compliance are used.</td>
</tr>
<tr>
<td>Practitioner adherence to study protocol</td>
<td>There is unobtrusive (or no) measurement of practitioner adherence. No special strategies to maintain or improve adherence are used.</td>
<td>There is close monitoring of how well the participating clinicians and centres are adhering to even the minute details in the trial protocol and “manual of procedures.”</td>
</tr>
<tr>
<td><strong>Analysis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis of primary outcome</td>
<td>The analysis includes all patients regardless of compliance, eligibility, and others (intention-to-treat analysis). In other words, the analysis attempts to see if the treatment works under the usual conditions, with all the noise inherent therein.</td>
<td>An intention-to-treat analysis is usually performed. However, this may be supplemented by a per-protocol analysis or an analysis restricted to “compliers” or other subgroups in order to estimate maximum achievable treatment effect. Analyses are conducted that attempt to answer the narrowest, “mechanistic” question (whether biological, educational or organizational).</td>
</tr>
</tbody>
</table>

Note: PRECIS = pragmatic-explanatory continuum indicator summary.
The design choices for a trial intended to inform a research decision about the benefit of a new drug are likely to be more explanatory (reflecting ideal conditions). Those for a later trial of the same drug intended to inform practical decisions by clinicians or policy-makers are likely to be more pragmatic (reflecting usual conditions). When planning their trial, trialists should consider whether a trial’s design matches the needs of those who will use the results. A tool to locate trial design choices within the pragmatic–explanatory continuum could facilitate these design decisions, help to ensure that the choices that are made reflect the intended purpose of the trial, and help others to appraise the extent to which a trial is appropriately designed for its intended purpose.

Such a tool could, for example, expose potential inconsistencies, such as the use of intensive adherence monitoring and intervention (explanatory tactics) in a trial being designed to answer a more pragmatic question. Alternatively, a trial might include a wide range of participants and meaningfully assess the impact (pragmatic tactics) but evaluate an intervention that is enforced or tightly monitored (explanatory tactics) and thus not widely feasible. By supporting the identification of potential inconsistencies such as these, a pragmatic–explanatory indicator could improve the extent to which trial designs are fit for purpose by highlighting design choices that do not support the needs of the intended users of the trial’s results. In this article we introduce such a tool.

The pragmatic–explanatory distinction comprises a continuous spectrum, not an either/or dichotomy of the extremes, as illustrated in Table 1. Moreover, it is probably impossible ever to perform a “purely” explanatory or “purely” pragmatic trial. For example, no patients are perpetually compliant, and the hand of the most skilled surgeon occasionally slips, so there can never be a “pure” explanatory trial. Similarly, a “pure” pragmatic trial loses its purity as soon as its first eligible patient refuses to be randomized.

Development of the PRECIS tool

The proposal for the pragmatic–explanatory continuum indicator summary (PRECIS) was developed by an international group of interested trialists at 2 meetings in Toronto (2005 and 2008) and in the time between. The initiative grew from the Pragmatic Randomized Controlled Trials in HealthCare (Practhc) project (www.practhc.org), an initiative funded by Canada and the European Union to promote pragmatic trials in low- and middle-income countries.

The development of the PRECIS indicator began with the identification of key domains that distinguish pragmatic from explanatory trials. As illustrated in Table 1, they comprise:

- The eligibility criteria for trial participants.
- The flexibility with which the experimental intervention is applied.
- The degree of practitioner expertise in applying and monitoring the experimental intervention.
- The flexibility with which the comparison intervention is applied.
- The degree of practitioner expertise in applying and monitoring the comparison intervention.
- The intensity of follow-up of trial participants.
- The nature of the trial’s primary outcome.
- The intensity of measuring participants’ compliance with the prescribed intervention, and whether compliance-improving strategies are used.
- The intensity of measuring practitioners’ adherence to the study protocol, and whether adherence-improving strategies are used.
- The specification and scope of the analysis of the primary outcome.

During the 2005 meeting, 8 domains emerged during a brainstorming session. Furthermore, 5 mutually exclusive definitions were used to assign the level of pragmatism in each domain. Attempts to use the initial tool on a number of published trials revealed some difficulties. The mutually exclusive categories were technically difficult to understand and use, and in some cases contradictory among domains. The current approach, for the most part, is to consider a number of design tactics or restrictions consistent with an explanatory trial in each domain. The more tactics that are present, the more explanatory the trial is. However, these design tactics and restrictions (see “The domains in detail” section for some examples) are not equally important, so it is not a simple matter of adding up tactics. Where exactly to place a trial on the pragmatic–explanatory continuum is, therefore, a judgment best made by trialists discussing these issues at the design stage of their trial and reaching consensus. Initially, the domains for intervention flexibility and practitioner expertise addressed both the experimental and comparison interventions. Discussions at the 2008 meeting led to the separation of experimental and comparison interventions into their own domains and the replacement of a domain regarding trial duration with the domain related to the nature of the primary outcome.

At this point, a brief explanation of our use of some terminology may be helpful. In this paper, we view a trial participant as the recipient of the intervention. In many trials, the participants are patients. However, in a trial of a continuing education intervention, for example, the participants may be physicians. By practitioner we mean the person delivering the intervention. Again, for many trials the practitioners are physicians. For a continuing education intervention, the practitioners may be trained instructors.

We defined the purpose of a pragmatic trial as answering the question “Does an intervention work under usual conditions?,” where we take “usual conditions” to mean the same as, or very similar to, the usual-care setting. Characterizing the pragmatic extreme of each domain is less straightforward, since what is considered “usual care” may depend on context. For some interventions, what is usual for each domain may vary across different settings. For example, the responsiveness and compliance of patients, adherence of practitioners to guidelines, and the training and experience of practitioners may be different in different settings. Thus, characterizing the pragmatic extreme requires specifying the settings for which a trial is intended to provide an answer. Occasionally a pragmatic trial addresses a question in a single specific setting. For example, a randomized trial of interventions to improve the use of active sick leave was designed to answer a pragmatic question under
usual conditions specific to the Norwegian context, where active sick leave was being promoted as a public sickness benefit scheme offered to promote early return to modified work for temporarily disabled workers. More often pragmatic trials will address questions across specific types of settings or across a wide range of settings. Examples of specific types of settings include settings where chloroquine-resistant falciparum malaria is endemic, where hospital facilities are in close proximity, or where trained specialists are available.

Conversely, we defined the purpose of an explanatory trial as answering the question “Can an intervention work under ideal conditions?” Given this definition, characterizing the explanatory extreme of each domain is relatively straightforward and intuitive. It simply requires considering the design decisions one would make in order to maximize the chances of success. Thus, for example, one would select patients that are most likely to comply and respond to the intervention, ensure that the intervention is delivered in a way that optimizes its potential for beneficial effects, and ensure that it is delivered by well-trained and experienced practitioners.

Thus, we recommend that trialists or others assessing whether design decisions are fit for purpose do this in 4 steps:

1. Declare whether the purpose of the trial is pragmatic or explanatory.
2. Specify the settings or conditions for which the trial is intended to be applicable.
3. Specify the design options at the pragmatic and explanatory extremes of each domain.
4. Decide how pragmatic or explanatory a trial is in relation to those extremes for each domain.

For some trials, there may not be any important difference between the pragmatic and explanatory extremes for some dimensions. For example, delivering an intervention, such as acetylsalicylic acid (ASA) therapy to someone with an acute myocardial infarction, does not require practitioner expertise. As mentioned earlier, for domains where the extremes are clear, it should not be difficult to decide whether a design decision is at one extreme or the other. For design decisions that are somewhere in between the extremes, it can be more challenging to determine how pragmatic or explanatory a trial will be. For this reason we recommend that all the members of the trial design team rate each domain and compare.

To facilitate steps 3 and 4, we have identified a number of design tactics that either add restrictions typical of explanatory trials or remove restrictions in the fashion of pragmatism. The tactics that we describe here are not intended to be prescriptive, exhaustive or even ordered in a particular way, but rather illustrative. They are to aid trialists or others in assessing where, within the pragmatic–explanatory continuum, a domain is, allowing them to put a “tick” on a line representing the continuum. To display the “results” of this assessment, the lines for each domain are arranged like spokes of a wheel, with the explanatory pole near the hub and the pragmatic pole on the rim (Figure 1). The display is completed by joining the locations of all 10 indicators as we progress around the wheel.

The proposed scales seem to make sense intuitively and can be used without special training. Although we recognize that alternative graphical displays are possible, we feel that the proposed wheel plot is an appealing summary and is informative in at least 3 ways.

First, it depicts whether a trial is tending to take a broad view (as in a pragmatic trial asking whether an intervention does work, under usual conditions) or tending to be narrowly “focused” near the hub (as for an explanatory trial asking whether an intervention can work, under ideal conditions).

Second, the wheel plot highlights inconsistencies in how the 10 domains will be managed in a trial. For example, if a trial is to admit all patients and practitioners (extremely pragmatic) yet will intensely monitor compliance and intervene when it falters (extremely explanatory), a single glance at the wheel will immediately identify this inconsistency. This allows the researcher to make adjustments in the design, if possible and appropriate, to obtain greater consistency with their objective in conducting the trial.

Third, the wheel plot can help trialists better report any limitations in interpretation or generalization resulting from design inconsistencies. This could help users of the trial results make better decisions.

The domains in detail

Participant eligibility criteria

The most extremely pragmatic approach to eligibility would seek only to identify study participants with the condition of interest from as many sources (e.g., institutions) as possible. As one moves toward a more explanatory attitude, additional restrictions will be placed on the study population. These restrictions include the following:

- Excluding participants not known or shown to be highly compliant to the interventions under study.
- Excluding participants not known or shown to be at high risk for the primary trial outcome.
• Excluding participants not expected to be highly responsive to the experimental intervention.
• Using a small number of sources (or even 1) for participants. The first 3 restrictions noted above are typically achieved by applying various exclusion criteria to filter out participants thought least likely to respond to the intervention. So, explanatory trials tend to have more exclusion criteria than pragmatic trials. Exclusion criteria for known safety issues would not necessarily count against a pragmatic trial, since such individuals would not be expected to get the intervention under usual practice.

Flexibility of experimental intervention
The pragmatic approach leaves the details of how to implement the experimental intervention up to the practitioners. For example, the details of how to perform a surgical procedure are left entirely to the surgeon. How to deliver an educational program is left to the discretion of the educator. In addition, the pragmatic approach would not dictate which co-interventions were permitted or how to deliver them. Several restrictions on the intervention’s flexibility are possible:
• Specific direction could be given for administering the intervention (e.g., dose, dosing schedule, surgical tactics, educational material and delivery).
• Timing of the delivery of the intervention could be designed to maximize the intervention effect.
• The number and permitted types of co-interventions could be restricted, particularly if excluded co-interventions would dilute any intervention effect.
• Specific direction could be given for applying permitted co-interventions.
• Specific direction could be given for managing complications or side effects from the primary intervention.

Experimental intervention — practitioner expertise
A pragmatic approach would put the experimental intervention into the hands of all practitioners treating (educating, and others) the study participants. The choice of practitioner can be restricted in a number of ways:
• Practitioners could be required to have some experience, defined by length of time, in working with the participants like the ones to be enrolled in the trial.
• Some specialty certification appropriate to the intervention could be required.
• For an intervention that has been in use (e.g., surgery) without a trial evaluation, experience with the intervention itself could be required.
• Only practitioners who are deemed to have sufficient experience in the subjective opinion of the trial investigator would be invited to participate.

Flexibility of the comparison intervention
Specification of the flexibility of the comparison intervention complements that of the flexibility of the experimental intervention. A pragmatic trial would typically compare an intervention to “usual practice” or the best alternative management strategy available, whereas an explanatory trial would restrict the flexibility of the comparison intervention and might, in the case of early-phase drug development trials, use a placebo rather than the best alternative management strategy as the comparator.

Comparison intervention — practitioner expertise
Similar comments apply as for the specification of the flexibility of the comparison intervention. In both cases, the explanatory extreme would maximize the chances of detecting whatever benefits an intervention might have, whereas the pragmatic extreme would aim to find out the benefits and harms of the intervention in comparison with usual practice in the settings of interest.

Follow-up intensity
The pragmatic position would be not to seek follow-up contact with the study participants in excess of the usual practice for the practitioner. The most extreme position is to have no contact with study participants and instead obtain outcome data by other means (e.g., administrative databases to determine mortality). Various adjustments to follow-up intensity are possible. The extent to which these adjustments could lead to increased compliance or improved intervention response will determine whether follow-up intensity moves toward the explanatory end.
• Follow-up visits (timing and frequency) are prespecified in the protocol.
• Follow-up visits are more frequent than typically would occur outside the trial (i.e., under usual care).
• Unscheduled follow-up visits are triggered by a primary outcome event.
• Unscheduled follow-up visits are triggered by an intervening event that is likely to lead to the primary outcome event.
• Participants are contacted if they fail to keep trial appointments.
• More extensive data are collected, particularly intervention-related data, than would be typical outside the trial.

Often the required trial outcomes may be obtained only through contact with the participants. Even in the “no follow-up” approach, assessment of outcomes may be achieved with a single “end of study” follow-up. The end of study would need to be defined so that there is sufficient time for the desired study outcomes (see “Primary trial outcome” section) to be observed. When the follow-up is done in this way, it is unlikely to have an impact on compliance or responsiveness. However, there may often be considerable tension between unobtrusive follow-up and the ability to collect the necessary outcomes. Often, although not always, explanatory trials are interested in the effect of an intervention during the intervention period, or shortly afterward. On the other hand, pragmatic trials may follow patients well beyond the intervention period in their quest to answer the “does this work?” question. Such longer term follow-up may well require more patient contact than usual care. However, it is not necessarily inconsistent with a pragmatic approach if it does not result in patient management that differs from the usual conditions, which may in turn increase the chance of detecting an intervention effect beyond what would be expected under usual conditions.
Table 2: A PRECIS assessment of 4 trials (part 1)

<table>
<thead>
<tr>
<th>Domain; trial Assessment of domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Participant eligibility criteria</strong></td>
</tr>
<tr>
<td>DOT&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>NASCET&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>CLASP&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Caritis et al.</td>
</tr>
<tr>
<td><strong>Experimental intervention — flexibility</strong></td>
</tr>
<tr>
<td>DOT</td>
</tr>
<tr>
<td>NASCET</td>
</tr>
<tr>
<td>CLASP</td>
</tr>
<tr>
<td>Caritis et al.</td>
</tr>
<tr>
<td><strong>Experimental intervention — practitioner expertise</strong></td>
</tr>
<tr>
<td>DOT</td>
</tr>
<tr>
<td>NASCET</td>
</tr>
<tr>
<td>CLASP</td>
</tr>
<tr>
<td>Caritis et al.</td>
</tr>
<tr>
<td><strong>Comparison intervention(s) — flexibility</strong></td>
</tr>
<tr>
<td>DOT</td>
</tr>
<tr>
<td>NASCET</td>
</tr>
<tr>
<td>CLASP</td>
</tr>
<tr>
<td>Caritis et al.</td>
</tr>
</tbody>
</table>
Table 2: A PRECIS assessment of 4 trials (part 2)

<table>
<thead>
<tr>
<th>Comparison intervention(s) — practitioner expertise</th>
<th>Assessment of domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOT</td>
<td>All clinic nurses were involved, with no particular specialization or additional training, which was extremely pragmatic.</td>
</tr>
<tr>
<td>NASCET</td>
<td>Like the surgical patients, the patients in the medical arm were managed and followed by board-certified neurologists or their senior subspecialty trainees.</td>
</tr>
<tr>
<td>CLASP</td>
<td>Since there was no difference in care provider with respect to treatment, this domain was treated the same as for the experimental arm.</td>
</tr>
<tr>
<td>Caritis et al.</td>
<td>Since there was no difference in care provider with respect to treatment, this domain was treated the same as for the experimental arm.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Follow-up intensity</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DOT</td>
<td>No extra clinic visits were scheduled. In fact, in the experimental arm, no visits whatsoever were required, since even the weekly drug collection could be delegated to a family member. This was the most extreme pragmatic approach.</td>
</tr>
<tr>
<td>NASCET</td>
<td>NASCET patients had prescheduled appointments at 1, 3, 6, 9, 12, 16, 20 and 24 months (and every 4 months thereafter). Each consisted of a medical, neurologic and functional-status assessment. All blood pressure records were reviewed centrally, and elevated readings triggered reminder letters. None of the 659 patients were lost to follow-up. A highly explanatory approach was taken here.</td>
</tr>
<tr>
<td>CLASP</td>
<td>There was a single scheduled follow-up, which happened after delivery of the infant and any of the primary study outcomes. Infant deaths up to 1 year were also recorded. This was very pragmatic.</td>
</tr>
<tr>
<td>Caritis et al.</td>
<td>Study follow-ups were scheduled to occur with the standard patient care schedules at each centre. Usually the patients were seen every 4 weeks up to 28 weeks’ gestation, then every 2 weeks up to 36 weeks’ gestation and then weekly thereafter until delivery. Although the visit schedule was no more intense than it would have been at these centres outside the trial, there would have been trial-related data collected that may not normally have been collected, which may have altered patient management from standard care. This was very explanatory but not extreme.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Primary trial outcome</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DOT</td>
<td>The primary outcome was “successful treatment,” which included all patients who were cured and all patients who completed the treatment. All patients were followed up for a year, until they completed their treatment, died, were classified as “incompletely treated” or were lost to follow-up; very pragmatic.</td>
</tr>
<tr>
<td>NASCET</td>
<td>The primary time was to ipsilateral stroke, the clinically relevant, explanatory outcome most likely to be affected by carotid endarterectomy. Other outcomes were more pragmatic: all strokes, major strokes and mortality were secondary outcomes.</td>
</tr>
<tr>
<td>CLASP</td>
<td>The primary outcome of pre-eclampsia was defined in a clinically relevant way that required only investigations common to standard care. Deaths up to 1 year post-delivery were recorded and adjudicated for cause. This was very pragmatic, but not the most extreme position.</td>
</tr>
<tr>
<td>Caritis et al.</td>
<td>The primary outcome of pre-eclampsia was defined in a clinically relevant way that required only investigations common to standard care. There was blinded adjudication of the primary outcome. There were a number of other short-term outcomes. Although the primary outcome itself was consistent with a pragmatic approach, the adjudication and focus on short-term outcomes moved this some way toward an explanatory approach.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Participant compliance with “prescribed” intervention</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DOT</td>
<td>Compliance was an element of the outcomes, and so was measured for this purpose, but was not used to improve patient compliance. This was pragmatic, but not at the most extreme end.</td>
</tr>
<tr>
<td>NASCET</td>
<td>The experimental intervention in NASCET was offering a one-time operation. Because the 50% probability of operation was clearly stated in the original consent documents, patients who did not want surgery were unlikely to enter the trial (only 0.3% of admitted patients randomized to the operation refused it). This was a prophylactic strategy for achieving compliance and was thus an explanatory approach.</td>
</tr>
<tr>
<td>CLASP</td>
<td>Compliance was asked about at the follow-up visit. Since this was after the completion of treatment, it could in no way affect compliance in the trial. Thus, it was extremely pragmatic.</td>
</tr>
<tr>
<td>Caritis et al.</td>
<td>Compliance was measured by pill count and direct questioning during follow-up. A research nurse periodically contacted women to “survey and reinforce compliance.” This was an extremely explanatory approach.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Practitioner adherence to study protocol</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DOT</td>
<td>There were no measurements of protocol adherence, and no adherence-improving strategies were used. This was the most pragmatic approach possible.</td>
</tr>
<tr>
<td>NASCET</td>
<td>The completeness, timeliness and accuracy of clinical data forms generated at admission, follow-up and for events were monitored centrally. Both at regular intervals, and more frequently when they were deficient, the NASCET principal investigator made a personal visit to that centre. In addition, blood pressure reports from each visit were scrutinized centrally, with letters pestering clinical collaborators when readings were elevated. An extremely explanatory approach was evident here.</td>
</tr>
<tr>
<td>CLASP</td>
<td>Not specified; assumed not extreme in either direction.</td>
</tr>
<tr>
<td>Caritis et al.</td>
<td>Not specified; assumed not extreme in either direction.</td>
</tr>
</tbody>
</table>
### Table 2: A PRECIS assessment of 4 trials (part 3)

<table>
<thead>
<tr>
<th>Domain; trial</th>
<th>Analysis of primary outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOT</td>
<td>All randomized patients were included in the primary analysis. Patients who failed to meet the criteria for “successful treatment” (including those who died, were lost to follow-up or were transferred to another clinic) were classified “failures.” This was an extremely pragmatic approach.</td>
</tr>
<tr>
<td>NASCET</td>
<td>The primary analysis was restricted to fatal and nonfatal strokes affecting the operated side of the cerebral circulation. In addition, blind adjudicators removed 3 NASCET patients after they were randomized because a review of their data before randomization revealed that they had other explanations for their symptoms (glaucoma, symptoms not arising from a carotid territory of the brain) or were inoperable (total occlusion of their carotid artery). However, patients were not excluded if they did not have a carotid endarterectomy or had uncontrolled blood pressure. This leaned toward an explanatory approach.</td>
</tr>
<tr>
<td>CLASP</td>
<td>An intention-to-treat analysis was conducted on patients who completed the follow-up. Some subgroups, notably high-risk subgroups, were considered a priori. This was a fairly pragmatic approach.</td>
</tr>
<tr>
<td>Caritis et al.</td>
<td>An intention-to-treat analysis was conducted on women with outcome data. An analysis, adjusted for compliance, was also performed. A number of additional “explanatory” analyses were conducted. This was fairly explanatory in its approach.</td>
</tr>
</tbody>
</table>

Note: PRE = pragmatic-explanatory continuum indicator summary, DOT = directly observed treatment, NASCET = North American Symptomatic Carotid Endarterectomy Trial, CLASP = Collaborative Low-dose Aspirin Study in Pregnancy.

### Primary trial outcome

For primary trial outcome, it is more intuitive to begin from the explanatory pole and describe the progression to the pragmatic pole. The most explanatory approach would consider a primary outcome (possibly surrogate, as in dose-finding trials intended to demonstrate a biological response) that the experimental intervention is expected to have a direct effect on. Phase 3 and 4 trials often have patient-important outcomes and thus may be more pragmatic in this domain. There may well be central adjudication of the outcome, or assessment of the outcome may require special training or tests not normally used to apply outcome definition criteria. Two obvious relaxations of the strict outcome assessment present in explanatory trials are the absence of central outcome adjudication and the reliance on usual training and measurement to determine the outcome status. For some interventions, the issue may be whether to measure outcomes only during the intervention period or up to a “reasonable” time after the intervention is complete. For example, stroke could be a primary outcome for explanatory and pragmatic trials. However, time horizons may vary from short term following a one-time intervention (more explanatory) to long term (more pragmatic).

### Participant compliance with “prescribed” intervention

The pragmatic approach recognizes that noncompliance with any intervention is a reality in routine medical practice. Because measurement of compliance may possibly alter subsequent compliance, the pragmatic approach in a trial would be not to measure or use compliance information in any way. The more rigorous a trial is in measuring and responding to noncompliance of the study participants, the more explanatory it becomes:

- Compliance is measured (indirectly) purely for descriptive purposes at the conclusion of the trial.
- Compliance data are measured and fed back to providers or participants during follow-up.
- Uniform compliance-improving strategies are applied to all participants.
- Compliance-improving strategies are applied to participants with documented poor compliance.

For some trials, the goal of an intervention may be to improve compliance with a treatment guideline. Provided the compliance measurement is not used, directly or indirectly, to influence subsequent compliance, a trial could still be “very pragmatic” in this domain. On the other hand, if measuring compliance is part of the intervention (e.g., audit and feedback), this domain would, appropriately, move toward a more explanatory approach if audit and feedback could not be similarly applied as part of the intervention under usual circumstances.

### Practitioner adherence to study protocol

The pragmatic approach takes account of the fact that providers will vary in how they implement an intervention. A purely pragmatic approach, therefore, would not be concerned with how practitioners vary or “customize” a trial protocol to suit their setting. By monitoring and (especially) acting on protocol nonadherence, a trial shifts toward being more explanatory:

- Adherence is measured (indirectly) purely for descriptive purposes at the conclusion of the trial.
- Adherence data are measured and fed back to practitioners.
- Uniform adherence-improving strategies are applied to all practitioners.
- Adherence-improving strategies are applied to practitioners with documented poor adherence.

### Analysis of the primary outcome

Recall that the pragmatic trial is concerned with the question “Does the intervention work under usual conditions?” Assuming other aspects of a trial have been treated in a pragmatic fashion, an analysis that makes no special allowance for noncompliance, nonadherence or practice variability, for example, is most appropriate for this question. So, the pragmatic approach to the primary analysis would typically be an intention-to-treat analysis of an outcome of direct relevance to the study.

---

**Analysis**
The PRECIS tool is an initial attempt to identify and quantify trial characteristics that distinguish between pragmatic and explanatory trials to assist researchers in designing trials. As such, we welcome suggestions for its further development. For example, the tool is applicable to individually randomized trials. It would probably apply to cluster randomized trials as well, but we have not tested it for those designs.

It is not hard to imagine that a judgment call is required to position the dots on the wheel diagram, especially for domains that are not at an extreme. Because trials are typically designed by a team of researchers, the PRECIS tool should be used by all involved in the design of the trial, leading to a consensus view on where the trial is situated within the pragmatic–explanatory continuum. The possible subjectiveness of dot placement should help focus the researcher’s attention on those domains that are not as pragmatic or explanatory as they would like. Clearly, domains where consensus is difficult to achieve warrant more attention.

There are other characteristics that may more often be present in pragmatic trials but, because they can also be found in explanatory trials, are not immediately helpful for discrimination. An appreciation of these characteristics helps round out the picture somewhat and assists with the interpretation of a given trial. For example, in a pragmatic trial, the comparison intervention is, by definition, standard care. So, one would be unlikely to use a placebo group in a pragmatic trial. Therefore, although the presence of a placebo group suggests an explanatory trial, absence of a placebo group does not necessarily suggest a pragmatic trial. Another example of this is blinding, whether it be blinded intervention delivery or outcome assessment blinded to treatment assignment. Blinding is desirable in all trials to the extent possible. Blinding may be less practical to achieve in some pragmatic trials, but that does not imply that blinding is inconsistent with a pragmatic trial.

Understanding the context for the applicability of the trial results is essential for all trials. For example, the intervention
studied in a pragmatic trial should be one that is feasible to implement in the “real world” after the completion of the trial. However, feasibility is often context specific. For example, an intervention could be easy to implement in Ontario, Canada, but all but impossible to implement in a low-income country because of cost, different health care delivery systems and many other reasons.

Our initial experiences developing the PRECIS tool suggest that it has the potential to be useful for trial design, although we anticipate that some refinement of the scales will be required. The reporting of pragmatic trials is addressed elsewhere.4 The simple graphical summary is a particularly appealing feature of this tool. We believe it has value for the planning of trials and the assessment of whether the design of a trial is fit for purpose. The tool can help ensure the right balance is struck to achieve the primary purpose of a trial, which
may be to answer an “explanatory” question about whether an intervention can work under ideal conditions or to answer a “pragmatic” question about whether an intervention does work under usual conditions. The PRECIS tool highlights the multidimensional nature of the pragmatic–explanatory continuum. This multidimensional structure should be borne in mind by trial designers and end-users alike so that overly simplistic labelling of trials can be avoided.

We would also like to caution readers to not confound the structure of a trial with its usefulness to potential users. Schwartz and Lellouch clearly linked the ability of a trial to meet its purpose with decisions about how the trial is designed and that, taken together, these decisions affect where the trial is placed on the explanatory–pragmatic continuum. However, how useful a trial is depends not only on design but on the similarity between the user’s context and that of the trial. Although it is unreasonable to expect the results of a trial to apply in all contexts, trials should be designed and reported in such a way that users of the results can make meaningful judgments about applicability to their own context.

Finally, we stress that this article, building on earlier work from multiple investigators, describes a “work in progress.” We welcome suggestions from all who read it, especially those who wish to join us in its further development. The words with which Schwartz and Lellouch closed their 1967 paper continue to apply: “This article makes no pretension to originality, nor to applicability to their own context.”

Competing interests: None declared.

Contributors: All of the authors made significant contributions to the intellectual content of this paper, reviewed multiple drafts for important omissions and have approved the final manuscript.

Acknowledgements: We are especially indebted to Dr. David L. Sackett for his encouragement and advice during the development of the tool and preparation of this manuscript. We would also like to acknowledge the contributions made by the numerous attendees at the Toronto workshops in 2005 and 2008.

The Practhe group was supported by the European Commission’s 5th Framework INCO program (contract ICA4-CT-2001-10019). The 2005 Toronto meeting was supported by a Canadian Institutes for Health Research grant (no. FRN 63095). The 2008 Toronto meeting was supported by the UK Medical Research Council, the Centre for Health Services Sciences at Sunnybrook Health Sciences Centre, Toronto, Canada, the Center for Medical Technology Policy, Baltimore, USA, and the National Institute for Health and Clinical Excellence, London, UK.

REFERENCES

12. Rothwell PM. External validity of randomised controlled trials: “To whom do the results of this trial apply?” Lancet 2005;365:82-93.

Correspondence to: Prof. Kevin E. Thorpe, Dalla Lana School of Public Health, University of Toronto, 6th floor, Health Sciences Building, 155 College St., Toronto ON M5T 3M7; fax 416 864-6057; kevin.thorpe@utoronto.ca
Atypical antipsychotics for aggression and psychosis in Alzheimer’s disease (Review)

Ballard CG, Waite J, Birks J

This is a reprint of a Cochrane review, prepared and maintained by The Cochrane Collaboration and published in The Cochrane Library 2008, Issue 4

http://www.thecochranelibrary.com
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEADER</td>
<td>1</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>1</td>
</tr>
<tr>
<td>PLAIN LANGUAGE SUMMARY</td>
<td>2</td>
</tr>
<tr>
<td>BACKGROUND</td>
<td>2</td>
</tr>
<tr>
<td>OBJECTIVES</td>
<td>3</td>
</tr>
<tr>
<td>METHODS</td>
<td>3</td>
</tr>
<tr>
<td>RESULTS</td>
<td>5</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>11</td>
</tr>
<tr>
<td>AUTHORS’ CONCLUSIONS</td>
<td>12</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>13</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>13</td>
</tr>
<tr>
<td>CHARACTERISTICS OF STUDIES</td>
<td>21</td>
</tr>
<tr>
<td>DATA AND ANALYSES</td>
<td>34</td>
</tr>
</tbody>
</table>

Analysis 1.1. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 1 BEHAVE-AD or NPI TOTAL (change from baseline at 10-13 weeks) ITT.  
Analysis 1.2. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 2 BEHAVE-AD or NPI PSYCHOSIS (change from baseline at 8-13 weeks) ITT.  
Analysis 1.3. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 3 BEHAVE-AD AGGRESSIVENESS (change from baseline at 12-13 weeks) ITT.  
Analysis 1.4. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 4 CMAI TOTAL AGGRESSIVENESS (change from baseline at 10-13 weeks) ITT.  
Analysis 1.5. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 5 Number of dropouts before end of treatment at 8-13 weeks.  
Analysis 1.6. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 6 Number of dropouts due to adverse events before end of treatment at 8-13 weeks.  
Analysis 1.7. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 7 Number of adverse events before end of treatment at 8-13 weeks.  
Analysis 1.8. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 8 Number of adverse events of somnolence before end of treatment at 8-13 weeks.  
Analysis 1.9. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 9 Number of adverse events of injury before end of treatment at 8-13 weeks.  
Analysis 1.10. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 10 Number of adverse events of purpura before end of treatment at 12-13 weeks.  
Analysis 1.11. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 11 Number of adverse events of pain before end of treatment at 12-13 weeks.  
Analysis 1.12. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 12 Number of adverse events of coughing before end of treatment at 12-13 weeks.  
Analysis 1.13. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 13 Number of adverse events of rhinitis before end of treatment at 12-13 weeks.  
Analysis 1.14. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 14 Number of adverse events of upper respiratory tract infection before end of treatment at 8-13 weeks.  
Analysis 1.15. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 15 Number of adverse events of a fall before end of treatment at 8-13 weeks.  
Analysis 1.16. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 16 Number of adverse events of extrapyramidal disorder before end of treatment at 8-13 weeks.  
Analysis 1.17. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 17 Number of adverse events of urinary tract infection before end of treatment at 12-13 weeks.  
Analysis 1.18. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 18 Number of adverse events of peripheral oedema before end of treatment at 12-13 weeks.  
Analysis 1.19. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 19 Number of adverse events of fever before end of treatment at 12-13 weeks.
Analysis 1.20. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 20 Number of adverse events of agitation before end of treatment at 8-13 weeks.

Analysis 1.21. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 21 Number of adverse events of tremor before end of treatment at 8-13 weeks.

Analysis 1.22. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 22 Number of adverse events of abnormal gait before end of treatment at 8-13 weeks.

Analysis 1.23. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 23 Number of adverse events of confusion before end of treatment at 8-13 weeks.

Analysis 1.24. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 24 Number of adverse events of hallucination before end of treatment at 8-13 weeks.

Analysis 1.25. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 25 Number of adverse events of urinary incontinence before end of treatment at 8-13 weeks.

Analysis 1.26. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 26 Number of adverse events of asthenia before end of treatment at 8-13 weeks.

Analysis 1.27. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 27 Number of adverse events of hostility before end of treatment at 8-13 weeks.

Analysis 1.28. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 28 Number of adverse events of flu syndrome before end of treatment at 8-13 weeks.

Analysis 1.29. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 29 Number of adverse events of dyspnea before end of treatment at 8-13 weeks.

Analysis 1.30. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 30 Number of adverse events of angina pectoris before end of treatment at 8-13 weeks.

Analysis 1.31. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 31 Number of cerebrovascular adverse events before end of treatment at 8-13 weeks.

Analysis 1.32. Comparison 1 RISPERIDONE VS PLACEBO, Outcome 32 Number of deaths before end of treatment at 8-13 weeks.

Analysis 2.1. Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 1 BEHAVE-AD (change from baseline at 12 weeks) ITT.

Analysis 2.2. Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 2 BEHAVE-AD PSYCHOSIS (change from baseline at 12 weeks) ITT.

Analysis 2.3. Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 3 BEHAVE-AD AGGRESSION (change from baseline at 12 weeks) ITT.

Analysis 2.4. Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 4 Number of dropouts before end of treatment at 12 weeks.

Analysis 2.5. Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 5 Number of dropouts due to adverse events before end of treatment at 12 weeks.

Analysis 2.6. Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 6 Number of adverse events of somnolence before end of treatment at 12 weeks.

Analysis 2.7. Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 7 Number of adverse events of injury before end of treatment at 12 weeks.

Analysis 2.8. Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 8 Number of adverse events of fall before end of treatment at 12 weeks.

Analysis 2.9. Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 9 Number of adverse events of extrapyramidal disorder before end of treatment at 12 weeks.

Analysis 2.10. Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 10 Number of adverse events of urinary tract infection before end of treatment at 12 weeks.

Analysis 2.11. Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 11 Number of adverse events of peripheral oedema before end of treatment at 12 weeks.

Analysis 2.12. Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 12 Number of adverse events of purpura before end of treatment at 12 weeks.

Analysis 2.13. Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 13 Number of adverse events of fever before end of treatment at 12 weeks.

Analysis 2.14. Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 14 Number of adverse events of depression before end of treatment at 12 weeks.
| Analysis 2.14. | Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 14 Number of adverse events of pain before end of treatment at 12 weeks. | 77 |
| Analysis 2.15. | Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 15 Number of adverse events of coughing before end of treatment at 12 weeks. | 78 |
| Analysis 2.16. | Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 16 Number of adverse events of agitation before end of treatment at 12 weeks. | 78 |
| Analysis 2.17. | Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 17 Number of adverse events of rhinitis before end of treatment at 12 weeks. | 79 |
| Analysis 2.18. | Comparison 2 RISPERIDONE 2 mg/day vs 1 mg/day, Outcome 18 Number of adverse events of upper respiratory tract infection before end of treatment at 12 weeks. | 79 |
| Analysis 3.1. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 1 NPI-NH TOTAL (change from baseline at 6-10 weeks) ITT. | 80 |
| Analysis 3.2. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 2 NPI-NH PSYCHOSIS (change from baseline at 6-10 weeks) ITT. | 81 |
| Analysis 3.3. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 3 NPI-NH AGITATION/AGGRESSION (change from baseline at 6-10 weeks) ITT. | 82 |
| Analysis 3.4. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 4 NPI-NH ANXIETY (change from baseline at 6-10 weeks) ITT. | 83 |
| Analysis 3.5. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 5 NPI-NH APATHY/INDIFFERENCE (change from baseline at 6-10 weeks) ITT. | 84 |
| Analysis 3.6. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 6 NPI-NH EUPHORIA/ELATION (change from baseline at 6-10 weeks) ITT. | 85 |
| Analysis 3.7. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 7 NPI-NH IRRITABILITY/LABILITY (change from baseline at 6-10 weeks) ITT. | 86 |
| Analysis 3.8. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 8 NPI-NH ABERRANT MOTOR (change from baseline at 6-10 weeks) ITT. | 87 |
| Analysis 3.9. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 9 NPI-NH DEPRESSION (change from baseline at 6-10 weeks) ITT. | 88 |
| Analysis 3.10. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 10 NPI-NH CARER DISTRESS (change from baseline at 6-10 weeks) ITT. | 89 |
| Analysis 3.11. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 11 BPRS TOTAL (change from baseline at 6-10 weeks) ITT. | 90 |
| Analysis 3.12. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 12 Number of dropouts before end of treatment at 6-10 weeks. | 91 |
| Analysis 3.13. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 13 Number of dropouts due to adverse events before end of treatment at 6-10 weeks. | 92 |
| Analysis 3.14. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 14 Number of deaths before end of treatment at 6-10 weeks. | 93 |
| Analysis 3.15. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 15 Number of adverse events due to accidental injury before end of treatment at 6-10 weeks. | 94 |
| Analysis 3.16. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 16 Number of adverse events due to somnolence before end of treatment at 6-10 weeks. | 95 |
| Analysis 3.17. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 17 Number of adverse events due to pain before end of treatment at 6-10 weeks. | 96 |
| Analysis 3.18. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 18 Number of adverse events due to abnormal gait before end of treatment at 6-10 weeks. | 97 |
| Analysis 3.19. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 19 Number of adverse events due to anorexia before end of treatment at 6-10 weeks. | 98 |
| Analysis 3.20. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 20 Number of adverse events due to ecchymosis before end of treatment at 6-10 weeks. | 99 |
| Analysis 3.21. | Comparison 3 OLANZAPINE VS PLACEBO, Outcome 21 Number of adverse events due to fever before end of treatment at 6-10 weeks. | 100 |
Analysis 3.22. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 22 Number of adverse events due to agitation before end of treatment at 6-10 weeks.

Analysis 3.23. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 23 Number of adverse events due to weight loss before end of treatment at 6-10 weeks.

Analysis 3.24. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 24 Number of adverse events due to increased cough before end of treatment at 6-10 weeks.

Analysis 3.25. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 25 Number of adverse events due to peripheral oedema before end of treatment at 6-10 weeks.

Analysis 3.26. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 26 Number of adverse events due to nervousness before end of treatment at 6-10 weeks.

Analysis 3.27. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 27 Number of adverse events due to hostility before end of treatment at 6-10 weeks.

Analysis 3.28. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 28 Number of adverse events due to confusion before end of treatment at 6-10 weeks.

Analysis 3.29. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 29 Number of adverse events due to anxiety before end of treatment at 6-10 weeks.

Analysis 3.30. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 30 Number of adverse events due to dry mouth before end of treatment at 6-10 weeks.

Analysis 3.31. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 31 Number of adverse events due to akathisia before end of treatment at 6-10 weeks.

Analysis 3.32. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 32 Number of adverse events due to falls before end of treatment at 6-10 weeks.

Analysis 3.33. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 33 Number of adverse events due to weight increase before end of treatment at 6-10 weeks.

Analysis 3.34. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 34 Number of adverse events due to hallucination before end of treatment at 6-10 weeks.

Analysis 3.35. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 35 Number of adverse events due to urinary incontinence before end of treatment at 6-10 weeks.

Analysis 3.36. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 36 Number of adverse events due to asthenia before end of treatment at 6-10 weeks.

Analysis 3.37. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 37 Number of adverse events due to flu syndrome before end of treatment at 6-10 weeks.

Analysis 3.38. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 38 Number of adverse events due to dyspnea before end of treatment at 6-10 weeks.

Analysis 3.39. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 39 Number of adverse events due to angina pectoris before end of treatment at 6-10 weeks.

Analysis 3.40. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 40 Number of cerebrovascular adverse events before end of treatment at 6-10 weeks.

Analysis 3.41. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 41 CGI (change from baseline at 6-10 weeks) ITT.

Analysis 3.42. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 42 MMSE (change from baseline at 6-10 weeks) ITT.

Analysis 3.43. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 43 BEHAVE-AD TOTAL (change from baseline at 6-10 weeks) ITT.

Analysis 3.44. Comparison 3 OLANZAPINE VS PLACEBO, Outcome 44 CGI-Severity (change from baseline at 6-10 weeks) ITT.

Analysis 4.1. Comparison 4 ARIPIPRAZOLE VS PLACEBO, Outcome 1 BEHAVE-AD PSYCHOSIS (change from baseline at 10 weeks) ITT.

Analysis 4.2. Comparison 4 ARIPIPRAZOLE VS PLACEBO, Outcome 2 BPRS TOTAL (change from baseline at 10 weeks) ITT.

Analysis 4.3. Comparison 4 ARIPIPRAZOLE VS PLACEBO, Outcome 3 BPRS - PSYCHOSIS SUBSCALE (change from baseline at 10 weeks) ITT.
Analysis 4.4. Comparison 4 ARIPIPRAZOLE VS PLACEBO, Outcome 4 Number of dropouts before end of treatment at 10 weeks. ............................................. 120
Analysis 4.5. Comparison 4 ARIPIPRAZOLE VS PLACEBO, Outcome 5 Number of dropouts due to adverse events before end of treatment at 10 weeks. ............................................. 120
Analysis 4.6. Comparison 4 ARIPIPRAZOLE VS PLACEBO, Outcome 6 Number of deaths before end of treatment at 10 weeks. ............................................. 121
Analysis 4.7. Comparison 4 ARIPIPRAZOLE VS PLACEBO, Outcome 7 Number of adverse events due to urinary tract infection before end of treatment at 10 weeks. ............................................. 121
Analysis 4.8. Comparison 4 ARIPIPRAZOLE VS PLACEBO, Outcome 8 Number of adverse events due to accidental injury before end of treatment at 10 weeks. ............................................. 122
Analysis 4.9. Comparison 4 ARIPIPRAZOLE VS PLACEBO, Outcome 9 Number of adverse events due to somnolence before end of treatment at 10 weeks. ............................................. 122
Analysis 4.10. Comparison 4 ARIPIPRAZOLE VS PLACEBO, Outcome 10 Number of adverse events due to bronchitis before end of treatment at 10 weeks. ............................................. 123
Analysis 4.11. Comparison 4 ARIPIPRAZOLE VS PLACEBO, Outcome 11 Number of adverse events due to extrapyramidal symptoms before end of treatment at 10 weeks. ............................................. 123
Analysis 5.1. Comparison 5 QUETIAPINE VS PLACEBO, Outcome 1 CMAI TOTAL (change from baseline at 6 weeks) ITT. ............................................. 124
Analysis 5.2. Comparison 5 QUETIAPINE VS PLACEBO, Outcome 2 CMAI TOTAL (change from baseline at 26 weeks) ITT. ............................................. 124
Analysis 5.3. Comparison 5 QUETIAPINE VS PLACEBO, Outcome 3 SIB (change from baseline at 6 weeks) ITT. ............................................. 125
Analysis 5.4. Comparison 5 QUETIAPINE VS PLACEBO, Outcome 4 SIB (change from baseline at 26 weeks) ITT. ............................................. 125
Analysis 5.5. Comparison 5 QUETIAPINE VS PLACEBO, Outcome 5 Number of dropouts by end of treatment at 26 weeks. ............................................. 126
WHAT'S NEW ............................................................................. 126
HISTORY ............................................................................... 126
CONTRIBUTIONS OF AUTHORS ............................................... 127
DECLARATIONS OF INTEREST ................................................ 127
SOURCES OF SUPPORT .......................................................... 127
NOTES .................................................................................. 127
INDEX TERMS ........................................................................ 127
ABSTRACT

Background

Aggression, agitation or psychosis occur in the majority of people with dementia at some point in the illness. There have been a number of trials of atypical antipsychotics to treat these symptoms over the last five years, and a systematic review is needed to evaluate the evidence in a balanced way.

Objectives

To determine whether evidence supports the use of atypical antipsychotics for the treatment of aggression, agitation and psychosis in people with Alzheimer's disease.

Search strategy

The trials were identified from a last updated search of the Specialized Register of the Cochrane Dementia and Cognitive Improvement Group on 7 December 2004 using the terms olanzapine, quetiapine, risperidone, clozapine, amisulpride, sertindole, zotepine, aripiprazole, ziprasidone. This Register contains articles from all major healthcare databases and many ongoing trials databases and is updated regularly.

Selection criteria

Randomised, placebo-controlled trials, with concealed allocation, where dementia and psychosis and/or aggression were assessed.

Data collection and analysis

1. Three reviewers extracted data from included trials
2. Data were pooled where possible, and analysed using appropriate statistical methods
3. Analysis included patients treated with an atypical antipsychotic, compared with placebo
Main results

Sixteen placebo controlled trials have been completed with atypical antipsychotics although only nine had sufficient data to contribute to a meta-analysis and only five have been published in full in peer reviewed journals. No trials of amisulpiride, sertindole or zotepine were identified which met the criteria for inclusion.

The included trials led to the following results:

1. There was a significant improvement in aggression with risperidone and olanzapine treatment compared to placebo.
2. There was a significant improvement in psychosis amongst risperidone treated patients.
3. Risperidone and olanzapine treated patients had a significantly higher incidence of serious adverse cerebrovascular events (including stroke), extrapyramidal side effects and other important adverse outcomes.
4. There was a significant increase in drop-outs in risperidone (2 mg) and olanzapine (5-10 mg) treated patients.
5. The data were insufficient to examine impact upon cognitive function.

Authors’ conclusions

Evidence suggests that risperidone and olanzapine are useful in reducing aggression and risperidone reduces psychosis, but both are associated with serious adverse cerebrovascular events and extrapyramidal symptoms. Despite the modest efficacy, the significant increase in adverse events confirms that neither risperidone nor olanzapine should be used routinely to treat dementia patients with aggression or psychosis unless there is severe distress or risk of physical harm to those living and working with the patient. Although insufficient data were available from the considered trials, a meta-analysis of seventeen placebo controlled trials of atypical neuroleptics for the treatment of behavioural symptoms in people with dementia conducted by the Food and Drug Administration suggested a significant increase in mortality (OR 1.7). A peer-reviewed meta-analysis (Schneider 2005) of 15 placebo controlled studies (nine unpublished) found similarly increased risk in mortality (OR=1.54, 95% CI 0.004 to 0.02, p=0.01) for the atypical neuroleptics.

Plain Language Summary

Atypical antipsychotics benefit people with dementia but the risks of adverse events may outweigh the benefits, particularly with long term treatment

Atypical antipsychotics have become the pharmacological treatment of choice for many clinicians in the treatment of behavioural and psychiatric symptoms in people with dementia, and the largest evidence base for double blind placebo controlled trials in this area is for risperidone. Particularly in view of recent safety concerns, a meta-analysis of efficacy and adverse events to inform clinical practice is timely. Modest efficacy is evident, but the elevated risk of cerebrovascular adverse events, mortality, upper respiratory infections, oedema and extrapyramidal symptoms is a concern, particularly as selective reporting makes interpretation of other potential adverse outcomes impossible.

Background

Five percent of people aged over 65 and 20% of those over 80 have dementia, with 700,000 sufferers in the UK alone, a number that will continue to rise as the age of the population increases. More than 50% of people with dementia experience behavioural and psychological disturbances (BPSD). BPSD are distressing for the patients (Gilley 1991), problematic for their carers (Rabins 1982) in whom they are associated with clinically significant depression (Ballard 1996) and are frequently the trigger for placement in residential or nursing home care (Steele 1990).

Pharmacological treatment with antipsychotic agents (also known as neuroleptics) is often the first line treatment for these disorders, despite evidence of only modest efficacy. Up until the last decade there were only a few double-blind placebo-controlled trials pertaining to the treatment of BPSD, the majority of which were conducted using typical antipsychotics such as haloperidol. However, these agents have a particularly high risk of harmful side effects for people with dementia, and the Chief Medical Officer for England and Wales (CMO 2004) has recommended that they should be