Technical Review
Number 12

Criteria for Distinguishing Effectiveness From Efficacy Trials in Systematic Reviews

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. 290-02-0016

Prepared by:
RTI-International–University of North Carolina Evidence-based Practice Center
Research Triangle Park, NC

Investigators
Gerald Gartlehner, M.D., M.P.H.
Richard A. Hansen, Ph.D.
Daniel Nissman, M.D., M.P.H.
Kathleen N. Lohr, Ph.D.
Timothy S. Carey, M.D., M.P.H.

AHRQ Publication No. 06-0046
April 2006

This report is based on research conducted by the RTI-International–University of North Carolina Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-02-0016). The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. No statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

Suggested Citation:
Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS. Criteria for Distinguishing Effectiveness From Efficacy Trials in Systematic Reviews. Technical Review 12 (Prepared by the RTI-International–University of North Carolina Evidence-based Practice Center under Contract No. 290-02-0016.) AHRQ Publication No. 06-0046. Rockville, MD: Agency for Healthcare Research and Quality. April 2006.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-Based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To bring the broadest range of experts into the development of evidence reports and health technology assessments, AHRQ encourages the EPCs to form partnerships and enter into collaborations with other medical and research organizations. The EPCs work with these partner organizations to ensure that the evidence reports and technology assessments they produce will become building blocks for health care quality improvement projects throughout the Nation. The reports undergo peer review prior to their release.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality.

We welcome comments on this evidence report. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to **epc@ahrq.gov.**

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Beth A. Collins Sharp, Ph.D., R.N.
Acting Director, EPC Program
Agency for Healthcare Research and Quality

Acknowledgments

We would like to thankfully acknowledge the directors of the U.S. and Canadian Evidence-based Practice Centers who selected the sample studies and provided insightful comments to the study protocol.

Structured Abstract

Objectives: To propose and test a simple instrument based on seven criteria of study design to distinguish effectiveness (pragmatic) from efficacy (explanatory) trials while conducting systematic reviews.

Design: Currently, no validated definition of effectiveness studies exists. We asked the directors of 12 Evidence-based Practice Centers (EPCs) to select six studies each: four that they considered to be examples of effectiveness trials and two considered efficacy studies. We then applied our proposed criteria to test the construct validity using the selected studies as if they had been identified by a gold standard.

Results: Based on the rationale to identify effectiveness studies reliably with minimal false positives (i.e., a high specificity), a cut-off of six criteria produced the most desirable balance between sensitivity and specificity. This setting produced a specificity of 0.83 and a sensitivity of 0.72.

Conclusions: When applied in a standardized manner, our proposed criteria can provide a valid and simple tool to distinguish effectiveness from efficacy studies. The applicability of systematic reviews can improve when analysts place more emphasis on the generalizability of included studies. In addition, clinicians can also use our criteria to determine the external validity of individual studies given an appropriate population of interest.

Contents

Tables

Technical Review

Chapter 1. Introduction

Randomized controlled trials (RCTs) are the gold standard in evaluating the effects of treatments. To be clinically meaningful, results must be relevant to specific patient populations in specific settings.[1] Multiple factors determine the external validity (i.e., generalizability or applicability) of RCTs: patient characteristics, condition under investigation, drug regimens, costs, compliance, co-morbidities, and concomitant treatments. For practical reasons, trials cannot always take these factors fully into consideration (e.g., costs, poor compliance). Also, certain aspects of study design—eligibility criteria, study duration, mode of intervention, outcomes, adverse events assessment, or type of statistical analysis greatly influence the degree of generalizability, given an appropriate population of interest.

Clinicians and policymakers often distinguish between the *efficacy* and the *effectiveness* of an intervention. Efficacy trials (explanatory trials) determine whether an intervention produces the expected result under ideal circumstances. Effectiveness trials (pragmatic trials) measure the degree of beneficial effect under "real world" clinical settings.[2] Hence, hypotheses and study designs of an effectiveness trial are formulated based on conditions of routine clinical practice and on outcomes essential for clinical decisions.

Efficacy and effectiveness exist on a continuum. Generalizability depends largely on the viewpoint of the observer and the condition under investigation. Baseline patient characteristics (e.g., sex, age, severity of the disease, racial groups) are primary factors in generalizability; thus, depending on the population of interest, generalizability of the same study can range from low to high. Geographic settings (urban versus rural) and health care systems can also be significant, factors,[1] although geography may have less influence on generalizability of drug trials than trials of other interventions (e.g., screening programs, behavioral therapy).

Ensuring generalizability may compromise internal validity. Under everyday clinical settings, factors such as patient or doctor preferences,[3,4] or patient-doctor relationships[5,6] can influence response and compliance. Random allocation, allocation concealment, and blinding negate these factors, thereby increasing internal validity on the one hand and decreasing external validity on the other. Therefore, to some extent, the operational definition of "effectiveness trial" delineates the necessary trade-offs with internal validity. An ideal definition would balance this equilibrium at a point at which satisfactory internal validity accompanies a high degree of generalizability.

Systematic reviews, including meta-analyses, have become an increasingly important source of information for clinical practice. If well conducted, they synthesize large amounts of information and provide estimated effect sizes that have greater precision and generalizability than individual studies.[7] Distinguishing between efficacy and effectiveness contributes an important aspect to analyzing any body of clinical evidence. Furthermore, greater emphasis on effectiveness studies may lead to changes in presentation in systematic reviews and policy initiatives.

In this article we propose and test seven hallmarks of study design to create a tool that can help researchers and those producing systematic reviews to distinguish more readily and more consistently between efficacy and effectiveness studies.

Chapter 2. Methods

Based on clinical and methodological considerations and the published literature, we selected seven domains of study design that, in our view, demonstrably influence the generalizability of trial results (Table 1). We searched MEDLINE® to identify published literature on instruments to distinguish effectiveness from efficacy studies. We found various definitions of effectiveness studies [8-12] but no validated rating instruments. Additional searches on Web sites of the U.S. Agency for Healthcare Research and Quality (AHRQ), U.S. Centers for Disease Control and Prevention (CDC), Cochrane Collaboration, and the U.K. National Institute for Clinical Effectiveness Web sites did not yield any results.

Table 1: Proposed criteria to distinguish effectiveness from efficacy trials

Item 1	Populations in primary care
Item 2	Less stringent eligibility criteria
Item 3	Health outcomes
Item 4	Long study duration; clinically relevant treatment modalities
Item 5	Assessment of adverse events
Item 6	Adequate sample size to assess a minimally important difference from a patient perspective
Item 7	Intention-to-treat analysis

Given the lack of a validated gold standard, we asked the directors of 12 Evidence-based Practice Centers (EPCs) in the United States and Canada to nominate six trials each. Four were to exemplify effectiveness studies and two, efficacy trials. The EPCs are programs that conduct systematic reviews for the AHRQ for a variety of audiences, including the CDC, U.S. National Institutes of Health, the U.S. Preventive Services Task Force (USPSTF), professional societies, and other health care groups. Any trial was eligible, regardless of design; observational studies or observational follow-ups of trials were ineligible. Our intent was to use the selected effectiveness and efficacy studies as if they had been identified by a gold standard method. The EPC directors possess many years of experience in systematic reviews, and we considered them the natural source of expertise for this effort. For masking purposes, we did not disclose our proposed criteria to EPC directors during their trial selection.

Once we had the nominated trials in hand, two independent raters applied our criteria to distinguish effectiveness from efficacy trials; they were blinded to which studies the EPC directors had identified as efficacy or effectiveness studies. Reviewers also assessed the internal validity (quality) of trials based on predefined criteria from the USPSTF (ratings: good-fair-poor)[13] and the National Health Service Centre for Reviews and Dissemination.[14]

We viewed diagnostic test parameters (i.e., sensitivity, specificity, likelihood ratios) as an intuitive and appropriate way to test the construct validity of our criteria. We determined diagnostic parameters for different cut-off points (i.e., seven criteria fulfilled, six criteria fulfilled, and so on) to assess how well our criteria identified effectiveness studies. To ensure reliability, we applied our criteria to seven trials that are frequently referred to as effectiveness studies in the published literature. We did all statistical analyses with StatsDirect 2.3.8.

Proposed Criteria

1. Populations in Primary Care

Efficacy studies are frequently conducted in large tertiary-care, referral settings, which tend to have more specialized clinicians and better technical equipment than primary care facilities. Subjects in such studies typically live in areas with ready access to such health centers and have accepted such referrals. They are often better educated and have better insurance coverage than the average primary care patient.

Primary care settings vary depending on health condition and available infrastructure. For people with most diseases, office-based locations, primary-care clinics, or community health centers are the initial setting for health care. Under specific circumstances, such as children or frail, elderly populations, schools or nursing homes may be the site of primary care.

For effectiveness trials, settings should reflect the initial care facilities available to a diverse population with the condition of interest. For persons with rare or severe diseases or those requiring high-risk interventions, such as organ transplantations, specialized secondary or tertiary care settings may provide initial care. Therefore, depending on the indication of interest, primary care settings may not always be an adequate criterion.

2. Less Stringent Eligibility Criteria

A common criticism of RCTs is that enrolled populations are highly selected and unrepresentative of the general population affected by the condition under consideration. Recruitment often employs stringent eligibility criteria to minimize adverse events and potential nonresponders. Some trials screen up to 68 people for each person enrolled.[15] Prerandomization run-in periods to exclude placebo-responders or poorly compliant patients additionally limit external validity.

For effectiveness trials, eligibility criteria must allow the source population to reflect the heterogeneity of external populations: the full spectrum of the human population, their co-morbidities, variable compliance rates, and use of other medications (or other therapies, such as psychotherapies, or complementary and alternative medications). Co-morbidities and other medications cannot be general exclusion criteria unless they contraindicate the use of the agent in ordinary practice.

Recruitment issues (e.g., volunteer bias, cultural barriers, language issues) may limit generalizability, perhaps severely for certain populations. For example, persons in minority ethnic groups are often underrepresented, and findings from trials among adults cannot be extrapolated to children. In North America, English language ability and literacy are prerequisites for participation in most trials, rendering immigrant groups ineligible for enrollment.

3. Health Outcomes

Clinical trials generally evaluate three types of outcome measures: subjective, objective, and health-related.[16] Efficacy studies, especially phase III clinical trials, commonly use objective or subjective outcomes (e.g., symptom scores, laboratory data, or time to disease recurrence) to determine intermediate (surrogate) outcomes. Assessments of health outcomes (e.g., functional capacity, quality of life, mortality) may be less commonly included as primary outcome measures. Short-term changes in symptom scores or laboratory data may provide valuable information regarding treatment mechanisms of improvement. Improvements in intermediate

outcome measures, however, cannot always be reliably extrapolated to improvements in health outcomes.[17]

Health outcomes, relevant to the condition of interest, should be the principal outcome measures in effectiveness studies. Intermediate outcomes are adequate only if empirical evidence verifies that the effect of the intervention on an intermediate endpoint predicts and fully captures the net effect on a health outcome.[17]

4. Long Study Duration, Clinically Relevant Treatment Modalities

External validity is limited if study protocols do not reflect clinical practice. Efficacy trials (of pharmaceuticals) are required for approval purposes, and investigators design study durations and treatment modalities to prove an effect and ensure safety. Such trials may not last as long as therapy would in everyday practice. Additionally, they may rely on strict diagnostic criteria that are usually not employed in primary care settings.

In effectiveness trials, study durations should mimic a minimum length of treatment in a clinical setting to allow the assessment of health outcomes. Treatment modalities should reflect clinical relevance (e.g., no fixed-dose designs; equivalent dosages for head-to-head comparisons). Diagnosis should rely on diagnostic standards that practicing physicians use.

In efficacy trials, investigators need to ensure (or measure) compliance to determine whether an intervention works. In clinical settings, however, adherence to therapy is often low;[18,19] it may depend on dosage regimens,[20] side effects profiles, and demographic or socioeconomic circumstances of the patients. In effectiveness trials, therefore, investigators should define compliance as an outcome measure, because unpredictable or "poor" compliance can render an efficacious treatment ineffective.[2]

5. Assessment of adverse events

Objective assessment of adverse events over an appropriate period of time is crucial to evaluate the balance of benefits and risks of any treatment. Reporting adverse events in RCTs is often limited; the methods of adverse events assessments are frequently poor. Rarely do investigators employ objective scales of adverse events (e.g., the World Health Organization scale of adverse events). Patient self-reporting often excludes "embarrassing" adverse events such as sexual side effects. To some extent, discontinuation rates and compliance, if assessed as outcomes, reflect adverse events.

Ideally, effectiveness studies use objective scales with predefined adverse events to determine adverse events rates. However, using an extensive objective adverse events scale is often not feasible in daily clinical practice because of time constraints and practical considerations. Therefore, adverse events assessments in effectiveness trials could be limited to critical issues based on experiences from prior trials.

6. Adequate Sample Size To Assess a Minimally Important Difference From a Patient Perspective

The power of a study to detect a statistically significant difference depends primarily on sample size. Large, simple trials with few levels of analysis provide the ideal study design to detect small but clinically significant treatment effects.[12] Small studies, specifically noninferiority drug trials may lack the statistical power to detect clinically significant differences between two treatments. Norman et al. propose one-half of a standard deviation as a good guideline to determine the effect size of a minimally important difference on health-related

quality of life instruments from a patient's perspective.[21] In a normally distributed sample this equates to a sample size of n = 64 per arm for a two-tailed test, not considering any loss to follow up. This sample size calculation, however, cannot be applied to studies with dichotomous outcomes or skewed continuous data.

The sample size of an effectiveness trial should be sufficient to detect at least a minimally important difference on a health-related quality of life scale. Therefore, we propose a minimum starting sample size of n = 75 participants per treatment arm to factor in a possible attrition of 15 percent. For conditions where rare but significant outcomes such as mortality or hospitalizations are of main interest, sample sizes must be greater and based on adequate power calculations. For example, differences in the rates of mortality (e.g., antiplatelet drugs for acute myocardial infarction) or hospitalization (e.g., inhaled corticosteroids for chronic obstructive pulmonary disease) may be of greater interest than health-related quality of life scores.

7. Intention-to-treat (ITT) analysis

ITT analysis maintains treatment groups that are similar except for random variation.[22] Given sound randomization and allocation concealment, ITT distributes known and unknown confounders equally across treatment groups. To some extent ITT analysis takes the effects of lack of adherence and varying reasons for treatment discontinuations into consideration when estimating effect sizes. The primary goal of efficacy trials is to determine if a treatment works under ideal circumstances. Ideal circumstances, however, require minimization of factors that can alter a treatment effect. Therefore, statistical analyses in efficacy trials frequently exclude patients with protocol deviations. In clinical practice, however, factors such as compliance, adverse events, drug regimens, co-morbidities, concomitant treatments, or costs all can alter efficacy. A "completers only" analysis would not take these factors adequately into account.

Internal Validity

Apart from designating a trial as an effectiveness or efficacy study, internal validity should be assessed; various rating scales and methods are available for such assessments.[23] Good or fair internal validity are prerequisites of external validity. To maintain internal validity, adequate randomization and allocation concealment are critical elements. Effectiveness trials often require cluster randomization to deal with contamination issues.[2]

A triple-blinded design (investigators, patients, and outcomes assessors or data analysts, if different from investigators) is a very difficult design to implement in effectiveness studies; even a double-blinded design (investigators/assessors and patients) is often not possible. For certain outcomes, such as all-cause mortality, blinding is not necessary because subjectivity plays no role. For investigations of surgical interventions masking patients may be extraordinarily difficult for practical and ethical reasons; masking observers measuring outcomes may be more feasible. Insofar as masking is crucial to avoid measurement bias, however, even when a double-blinded design is not achievable, outcomes assessment or data analysis must be blinded whenever possible.

Chapter 3. Results

The EPC directors identified 26 studies. Of these, six were intended to illustrate efficacy trials and 20, effectiveness studies. We excluded two studies from the latter group because they did not meet eligibility criteria:[24,25] one was an observational follow-up of three RCTs [24] and the other a pooled analysis of clinical trials.[25] Of the remaining 24 studies (Table 2, alphabetical by author), 22 were RCTs, three with an open-label design[26-28]; the other two were a nonrandomized, controlled trial[29] and an uncontrolled trial.[30]

Table 2: Overview of studies identified by the EPC directors

Author, Year	Title	Study design	Sample Size	Funding
Bridges et al. 2000 [31]	Effectiveness and cost-benefit of influenza vaccination of healthy working adults: a randomized controlled trial	RCT, double-blind, placebo controlled	2,375	National Center for Infectious Diseases, CDC
Conley et al. 2001[32]	A randomized double-blind study of risperidone and olanzapine in the treatment of schizophrenia or schizoaffective disorder	RCT, double-blind, head-to-head	377	Janssen Research Foundation
Farlow et al. 1992 [33]	A controlled trial of tacrine in Alzheimer's disease	RCT, double-blind, placebo-controlled	468	Parke-Davis
Follath et al. 2002 [34]	Efficacy and safety of intravenous levosimendan compared with dobutamine in severe low-output heart failure (the LIDO study): a randomised double-blind trial	RCT, double-blind, head-to-head	203	Orion Pharma, Espoo, Finland
Gane et al. 1997[35]	Randomised trial of efficacy and safety of oral ganciclovir in the prevention of cytomegalovirus disease in liver-transplant recipients	RCT, double-blind, placebo-controlled	304	Roche Global Development Palo Alto, CA
Geldmacher et al. 2003[24]	Donepezil is associated with delayed nursing home placement in patients with Alzheimer's Disease	Observational follow-up of RCTs	1,115	Eisai, Inc. and Pfizer, Inc.
Jerrell et al. 2002 [26]	Cost-effectiveness of risperidone, olanzapine, and conventional antipsychotic medications	RCT, open-label, head-to-head	108	South Carolina Dept of Mental Health
Kawai et al. 2005 [29]	Factors influencing the effectiveness of oseltamivir and amantadine for the treatment of influenza: a multicenter study from Japan of the 2002-2003 influenza season	non-randomized, open-label trial	2,163	Not stated
Klassen et al. 1996[36]	The efficacy of nebulized budesonide in dexamethasone-treated outpatients with croup	Randomized, double-blind, placebo-controlled	50	Ontario Ministry of Health grant

Table 2: Overview of studies identified by the EPC directors (cont'd)

Author, Year	Title	Study design	Sample Size	Funding
Knapp et al. 1994[37]	A 30-week trial of high-dose tacrine in patients with Alzheimer's disease	RCT, double-blind, placebo-controlled	653	Parke-Davis
Kroenke et al. 2001 [27]	Similar effectiveness of paroxetine, fluoxetine, and sertraline in primary care: a randomized trial	RCT, open-label, head-to-head	573	Ely Lilli
Little et al. 2001 [28]	Pragmatic randomized controlled trial of two prescribing strategies for childhood acute otitis media	RCT, open-label, head-to-head	315	NHS Research & Development
Maskell et al. 2005[38]	U.K. controlled trial of intrapleural streptokinase for pleural infection	RCT, double-blind, placebo-controlled	427	UK Medical Research Council
McFalls et al. 2004[39]	Coronary-artery revascularization before elective major vascular surgery	RCT, open-label	510	Cooperative Studies Program, Dept. VA Office of R&D
Meltzer et al. 2003[40]	Clozapine treatment for suicidality in schizophrenia	RCT, open-label, head-to-head	980	Novartis; William K. Warren Research Foundation; Donald Test Foundation
Mendelmann et al. 2001[25]	Safety, efficacy and effectiveness of the influenza virus vaccine, trivalent, types A and B, live, cold-adapted (CAIV-T) in healthy children and adults	Pooled analysis of RCTs	10,443	Aviron , NIH
Physicians Health Study 1989[41]	Final report on the aspirin component of the ongoing physicians' health study	RCT, double-blind, placebo-controlled	22,071	NIH
Plint et al. 2000[42]	The efficacy of nebulized racemic epinephrine in children with acute asthma: a randomized, double-blind trial	RCT, double-blind, head-to-head	121	Children's Hospital of Eastern Ontario Research Institute Grant;
Purdon et al. 2000[43]	Neuropsychological change in the early phase of schizophrenia during 12 months of treatment with olanzapine, risperidone, or haloperidol	RCT, double-blind, head-to-head	65	Eli Lilly
Robles et al. 2005[30]	Effectiveness and safety of eprosartan on pulse pressure for the treatment of hypertensive patients	Uncontrolled, open-label trial	566	None listed

Table 2: Overview of studies identified by the EPC directors (cont'd)

Author, Year	Title	Study design	Sample Size	Funding
Rosenheck et al. 1999[44]	Cost-effectiveness of clozapine in patients with high and low levels of hospital use	RCT, double-blind, head-to-head	423	Dept. of VA Health Services; Sandoz
Rosenheck et al. 2003[45]	Effectiveness and cost of olanzapine and haloperidol in the treatment of schizophrenia	RCT, double-blind, head-to-head	309	Eli Lilly; VA Cooperative Studies Program
Schmid et al. 2005[46]	Effectiveness of a 10-day melarsoprol schedule for the treatment of late-stage human African trypanosomiasis: confirmation from a multinational study (Impamel II)	Uncontrolled, open-label trial	2,020	Swiss Agency for Development and Cooperation
Stiell et al. 2004[47]	Advanced cardiac life support in out-of-hospital cardiac arrest	Open-label, controlled trial	5,638	Ontario Ministry of Health and Long-term Care
The Food Trial 2005[48]	Effect of timing and method of enteral tube feeding for dysphagic stroke patients (FOOD): a multicenter randomized controlled trial	RCT, open-label	859 321	Multiple government assoc. in the UK, Singapore, & New Zealand
Wassef et al. 2005[49]	Lower effectiveness of divalproex versus valproic acid in a prospective, quasi-experimental clinical trial involving 9,260 psychiatric admissions	Quasi-experimental, controlled trial	5,228	Not stated

NIH: National Institute of Health

Table 3 indicates whether each study met (Y) or did not meet (N) each criterion and gives the quality grade (good, fair, or poor); studies thought to be efficacy trials are given first, then studies thought to be effectiveness trials, and both sets are ranked by quality grade. For example, Follath et al. (a good efficacy trial) met only three of the seven criteria; by contrast, Bridges et al. (a good effectiveness trial) met all seven.

Table 3: Results of applying effectiveness criteria to studies selected by the EPC directors

Study	Item 1: Populations in primary care setting	Item 2: Less stringent eligibility criteria	Item 3: Health outcomes QOL	Item 4: Long study durations, clinically relevant study modalities	Item 5: Assessment of adverse events	Item 6: Adequate sample size	Item 7: ITT analysis	Quality Rating
Studies submitted as efficacy studies								
Conley et al. 2001[32]	N	N	N	N	Y	Y	Y	Fair
Follath et al. 2002[34]	N	N	N	Y	N	Y	Y	Good
Gane et al. 1997[35]	Y	Y	Y	Y	Y	Y	Y	Fair
Klassen et al. 1996[36]	N	N	N	N	N	N	N	Fair
Plint et al. 2000[42]	N	N	N	Y	N	Y	N	Fair
Purdon et al. 2000[43]	Y	Y	N	Y	Y	N	N	Poor*
Studies submitted as effectiveness studies								
Bridges et al. 2000[31]	Y	Y	Y	Y	Y	Y	Y	Good
Farlow et al. 1992[33]	Y	N	N	N	Y	Y	N	Poor†
Jerrell et al. 2002[26]	N	N	Y	Y	Y	Y	N	Poor*
Kawai et al. 2005[29]	Y	Y	Y	Y	N	Y	N	Fair
Knapp et al. 1994[37]	Y	Y	N	Y	Y	Y	Y	Poor †
Kroenke et al. 2001[27]	Y	Y	Y	Y	Y	Y	Y	Fair
Little et al. 2001[28]	Y	Y	Y	Y	Y	Y	Y	Fair
Maskell et al. 2005[38]	Y	N	Y	Y	Y	Y	Y	Good
McFalls et al. 2004[39]	N	Y	Y	Y	Y	Y	Y	Good
Meltzer et al. 2003[40]	Y	Y	Y	Y	Y	Y	Y	Fair
Physicians Health Study 1989[41]	Y	Y	Y	Y	Y	Y	Y	Good
Robles et al. 2005[30]	Y	Y	N	N	Y	Y	N	Fair
Rosenheck et al. 1999[44]	N	Y	Y	Y	Y	Y	Y	Poor†
Rosenheck et al. 2003[45]	N	Y	Y	Y	Y	Y	Y	Fair
Schmid et al. 2005[46]	Y	Y	Y	Y	Y	N	Y	Poor† ‡
Stiell et al. 2004[45]	Y	Y	Y	Y	N	Y	Y	Fair
The Food Trial 2005[47]	Y	Y	Y	Y	N	Y	Y	Fair
Wassef et al. 2005[49]	N	Y	Y	Y	Y	Y	N	Fair

*High number of post-randomization exclusions
† High attrition
‡ Completers-only analysis

Table 4 summarizes the diagnostic parameters at different cut-off points. Using a cut-off point of seven criteria (i.e., a trial meets *all* criteria to consider it an effectiveness study), we identified five of the 18 suggested effectiveness studies. This yielded a sensitivity of 0.28 and a specificity of 0.83. Employing six criteria as a cut-off raised the number of identified effectiveness trials to 13 with a corresponding sensitivity of 0.72 and a specificity of 0.83. A cut-off of five criteria led to a sensitivity of 0.89 and a specificity of 0.67.

Table 4: Summary of diagnostic parameters for different cut-off points

Diagnostic Parameters	Estimate	95% Confidence Interval
Cut off: 7 (all) criteria fulfilled		
Sensitivity (%):	0.28	0.10 to 0.53
Specificity (%):	0.83	0.36 to 1.00
+ Likelihood ratio	1.7	0.4 to 10.1
- Likelihood ratio	0.9	0.6 to 1.7
Cut off: 6 criteria fulfilled		
Sensitivity (%):	0.72	0.46 to 0.90
Specificity (%):	0.83	0.36 to 1.00
+ Likelihood ratio	4.3	1.2 to 24.4
- Likelihood ratio	0.3	0.1 to 0.8
Cut off: 5 criteria fulfilled		
Sensitivity (%):	0.89	0.65 to 0.99
Specificity (%):	0.67	0.22 to 0.96
+ Likelihood ratio	2.7	1.2 to 9.3
- Likelihood ratio	0.2	0.0 to 0.6

Based on the rationale that we want to identify effectiveness studies reliably with minimal false positives (i.e., high specificity), a cut-off of six criteria produced the most desirable balance between sensitivity and specificity. At this point, the positive likelihood ratio was 4.3, the negative likelihood ratio 0.3. In other words, a true effectiveness study is 4.3 times more likely to be identified as such than an efficacy trial.

We did not reach a higher specificity because of one study that was nominated as an efficacy trial.[35] However, three reviewers independently classified this trial as an effectiveness study. This trial assessed the efficacy and safety of oral ganciclovir in preventing cytomegalovirus disease in liver-transplant recipients. The settings were multiple university clinics in Europe and the United States. In our view, liver transplantations will never be conducted in primary care facilities; tertiary care facilities participating in this trial, although highly specialized, will always be the setting of initial care for patients undergoing organ transplantations. Because all the other criteria were also fulfilled, we deemed this study to be an effectiveness trial. Removing this study from our calculations raises the positive likelihood ratio to 8.5 with a specificity of 0.92 and a sensitivity of 0.71.

The methodological quality of the studies was mixed. Of the 18 effectiveness studies, five (28 percent) received a “poor” quality rating. High attrition and a high number of post-randomization exclusion were the principal reasons for poor internal validity. Nevertheless, most of these studies still provide valuable information. For example, two trials assessing the effectiveness of tacrine in patients with Alzheimer’s disease were rated “poor” because of high attrition (>50%).[33,37] However, the high attrition rate was attributable primarily to the hepatotoxicity of tacrine. Therefore, although the internal validity to determine the effectiveness

of tacrine is compromised, these studies still provide valuable information about the risk-benefit ratio of tacrine. An efficacy study, with an active run-in period before randomization, might have concealed such findings.

In addition to the studies identified by the EPC directors we applied our criteria to seven trials that are frequently used as examples of effectiveness studies in the published literature.[50-56] Using a cut-off point of six criteria we identified all seven trials as effectiveness studies.

Chapter 4. Discussion

Our objective was to identify drug effectiveness studies reliably based on seven proposed criteria. We focused on studies of medications because they are common, but many of the same principles can be applied to other types of interventions. Because we attribute greater value to effectiveness studies than to efficacy studies, the specificity of this process had to be high. That is, we wanted to ensure that efficacy studies are not falsely rated as effectiveness studies. Erring on the side of exclusions appeared to be better than erring on the side of inclusions, given that most analysts will give greater emphasis to inferences drawn from effectiveness studies than from efficacy trials. Thus, on the one hand, trials identified as effectiveness studies must reliably have great external validity. On the other hand, criteria must not be so stringent as to exclude a large proportion of effectiveness studies. We found a cut-off of six criteria, which produced a specificity of 0.83 and a sensitivity of 0.72, most suitable for this rationale.

Our results should be interpreted cautiously for several reasons. First, the sample size of articles was small, which produces limitations in terms of using diagnostic parameters to test construct validity. The estimate of specificity is based on six efficacy studies only and has a great degree of uncertainty, with a "true" estimate between 0.36 and 1.00. Second, no validated definition of "effectiveness study" exists, and, EPC directors' views on the matter differed greatly. Nonetheless, the different EPCs represent a broad, valid spectrum of the current thinking in evidence-based medicine about effectiveness studies. Third, we excluded observational studies and included only trials (although not limited to RCTs). We strongly believe that observational studies have an important role in determining the effectiveness of drug interventions, especially with respect to long-term health outcomes and rare but severe adverse events. However, we limited our eligibility criteria to trials because they (especially RCTs) can best establish causation and minimize bias. Other designs, such as observational studies, may better reflect "real world" settings but can mainly indicate associations and only to a lesser degree determine causal effects.[57] Completely ruling out confounding in observational studies is not possible.

The applicability of our criteria for standard use in systematic reviews remains to be seen. Depending on the topic, some criteria could be predefined as "effectiveness eligibility criteria" at the start of a systematic review. For example, for a systematic review on Alzheimer's medications minimum length of follow-up (e.g., 6 months), typical primary care settings (e.g., nursing homes, office-based settings), and relevant health outcomes (e.g., nursing home placements, activities of daily living) could be defined a priori and reduce inter-rater differences.

Overall, our proposed criteria provide a simple, valid tool to distinguish effectiveness from efficacy studies. When the generalizability of included studies is given greater emphasis, through inclusion of effectiveness trials, systematic reviews can be made more generally applicable. Clinicians can apply our criteria to determine the external validity of individual studies given their particular populations of interest.

References

1. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?" Lancet 2005;365:82-93.
2. Godwin M, Ruhland L, Casson I, et al. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. BMC Med Res Methodol 2003;3:28.
3. Benson J, Britten N. Patients' decisions about whether or not to take antihypertensive drugs: qualitative study. BMJ 2002;325:873.
4. Redelmeier DA, Rozin P, Kahneman D. Understanding patients' decisions. Cognitive and emotional perspectives. JAMA 1993;270:72-6.
5. Thomas KB. General practice consultations: is there any point in being positive? Br Med J (Clin Res Ed) 1987;294:1200-2.
6. Di Blasi Z, Harkness E, Ernst E,et al. Influence of context effects on health outcomes: a systematic review. Lancet 2001;357:757-62.
7. Mulrow CD. Rationale for systematic reviews. BMJ 1994;309:597-9.
8. Brook RH, Lohr KN. Efficacy, effectiveness, variations, and quality. Boundary-crossing research. Med Care 1985;23:710-22.
9. Hoagwood K, Hibbs E, Brent D, et al. Introduction to the special section: efficacy and effectiveness in studies of child and adolescent psychotherapy. J Consult Clin Psychol 1995;63:683-7.
10. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. J Chronic Dis 1967;20:637-48.
11. Califf RM, DeMets DL. Principles from clinical trials relevant to clinical practice: Part I. Circulation 2002;106:1015-21.
12. Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol 1995;48:23-40.
13. Harris RP, Helfand M, Woolf SH, et al. Current methods of the U.S. Preventive Services Task Force: a review of the process. Am J Prev Med 2001;20(3 Suppl):21-35.
14. Centre for Reviews and Dissemination. Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews. CRD Report No. 4, 2nd ed. York, UK: NHS Centre for Reviews and Dissemination; 2001.
15. Gross CP, Mallory R, Heiat A, et al. Reporting the recruitment process in clinical trials: who are these patients and how did they get there? Ann Intern Med 2002;137:10-6.
16. Spahn J. Clinical trial efficacy: what does it really tell you? J Allergy Clin Immunol 2003;112:S102-6.
17. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med 1996;125:605-13.
18. Coutts JA, Gibson NA, Paton JY. Measuring compliance with inhaled medication in asthma. Arch Dis Child 1992;67:332-3.
19. Gibson NA, Ferguson AE, Aitchison TC, et al. Compliance with inhaled asthma medication in preschool children. Thorax 1995;50:1274-9.
20. Kelloway JS, Wyatt RA, Adlis SA. Comparison of patients' compliance with prescribed oral and inhaled asthma medications. Arch Intern Med 1994;154:1349-52.
21. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003;41:582-92.
22. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 1999;319:670-4.
23. West S, King V, Carey TS, et al. Systems to rate the strength of scientific evidence. Evid Rep Technol Assess (Summ) 2002 Mar;(47):1-11.
24. Geldmacher DS, Provenzano G, McRae T, et al. Donepezil is associated with delayed nursing home placement in patients with

Alzheimer's disease. J Am Geriatr Soc 2003;51:937-44.

25. Mendelman PM, Cordova J, Cho I. Safety, efficacy and effectiveness of the influenza virus vaccine, trivalent, types A and B, live, cold-adapted (CAIV-T) in healthy children and healthy adults. Vaccine 2001;19:2221-6.

26. Jerrell JM. Cost-effectiveness of risperidone, olanzapine, and conventional antipsychotic medications. Schizophr Bull 2002;28:589-605.

27. Kroenke K, West SL, Swindle R, et al. Similar effectiveness of paroxetine, fluoxetine, and sertraline in primary care: a randomized trial. JAMA 2001;286:2947-55.

28. Little P, Gould C, Williamson I, et al. Pragmatic randomised controlled trial of two prescribing strategies for childhood acute otitis media. BMJ 2001;322:336-42.

29. Kawai N, Ikematsu H, Iwaki N, et al. Factors influencing the effectiveness of oseltamivir and amantadine for the treatment of influenza: a multicenter study from Japan of the 2002-2003 influenza season. Clin Infect Dis 2005;40:1309-16.

30. Robles NR, Martin-Agueda B, Lopez-Munoz F, et al. Effectiveness and safety of eprosartan on pulse pressure for the treatment of hypertensive patients. Int J Clin Pract 2005;59:478-84.

31. Bridges CB, Thompson WW, Meltzer MI, et al. Effectiveness and cost-benefit of influenza vaccination of healthy working adults: a randomized controlled trial. JAMA 2000;284:1655-63.

32. Conley RR, Mahmoud R. A randomized double-blind study of risperidone and olanzapine in the treatment of schizophrenia or schizoaffective disorder. Am J Psychiatry 2001;158:765-74.

33. Farlow M, Gracon SI, Hershey LA, et al. A controlled trial of tacrine in Alzheimer's disease. The Tacrine Study Group. JAMA 1992;268:2523-9.

34. Follath F, Cleland JG, Just H, et al. Efficacy and safety of intravenous levosimendan compared with dobutamine in severe low-output heart failure (the LIDO study): a randomised double-blind trial. Lancet 2002;360:196-202.

35. Gane E, Saliba F, Valdecasas GJ, et al. Randomised trial of efficacy and safety of oral ganciclovir in the prevention of cytomegalovirus disease in liver-transplant recipients. The Oral Ganciclovir International Transplantation Study Group. Lancet 1997;350:1729-33.

36. Klassen TP, Watters LK, Feldman ME, et al. The efficacy of nebulized budesonide in dexamethasone-treated outpatients with croup. Pediatrics 1996;97:463-6.

37. Knapp MJ, Knopman DS, Solomon PR, et al. A 30-week randomized controlled trial of high-dose tacrine in patients with Alzheimer's disease. The Tacrine Study Group. JAMA 1994;271:985-91.

38. Maskell NA, Davies CW, Nunn AJ, et al. U.K. Controlled trial of intrapleural streptokinase for pleural infection. N Engl J Med 2005;352:865-74.

39. McFalls EO, Ward HB, Moritz TE, et al. Coronary-artery revascularization before elective major vascular surgery. N Engl J Med 2004;351:2795-804.

40. Meltzer HY, Alphs L, Green AI, et al. Clozapine treatment for suicidality in schizophrenia: International Suicide Prevention Trial (InterSePT). Arch Gen Psychiatry 2003;60:82-91.

41. Final report on the aspirin component of the ongoing Physicians' Health Study. Steering Committee of the Physicians' Health Study Research Group. N Engl J Med 1989;321:129-35.

42. Plint AC, Osmond MH, Klassen TP. The efficacy of nebulized racemic epinephrine in children with acute asthma: a randomized, double-blind trial. Acad Emerg Med 2000;7:1097-103.

43. Purdon SE, Jones BD, Stip E, et al. Neuropsychological change in early phase schizophrenia during 12 months of treatment with olanzapine, risperidone, or haloperidol. The Canadian Collaborative Group for research in schizophrenia. Arch Gen Psychiatry 2000;57:249-58.

44. Rosenheck R, Cramer J, Allan E, et al. Cost-effectiveness of clozapine in patients with high and low levels of hospital use. Department of Veterans Affairs Cooperative Study Group on Clozapine in Refractory Schizophrenia. Arch Gen Psychiatry 1999; 56:565-72.

45. Rosenheck R, Perlick D, Bingham S, et al. Effectiveness and cost of olanzapine and haloperidol in the treatment of schizophrenia: a randomized controlled trial. JAMA 2003;290:2693-702.

46. Schmid C, Richer M, Bilenge CM, et al. Effectiveness of a 10-day Melarsoprol schedule for the treatment of late-stage human African Trypanosomiasis: confirmation from a multinational study (Impamel II). J Infect Dis 2005;191:1922-31.

47. Stiell IG, Wells GA, Field B, et al. Advanced cardiac life support in out-of-hospital cardiac arrest. N Engl J Med 2004;351:647-56.

48. Dennis MS, Lewis SC, Warlow C. Effect of timing and method of enteral tube feeding for dysphagic stroke patients (FOOD): a multicentre randomised controlled trial. Lancet 2005;365:764-72.

49. Wassef AA, Winkler DE, Roache AL, et al. Lower effectiveness of divalproex versus valproic acid in a prospective, quasi-experimental clinical trial involving 9,260 psychiatric admissions. Am J Psychiatry 2005;162:330-9.

50. Davis BR, Cutler JA, Gordon DJ, et al. Rationale and design for the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). ALLHAT Research Group. Am J Hypertens 1996;9:342-60.

51. Yusuf S, Mehta SR, Xie C, et al. Effects of reviparin, a low-molecular-weight heparin, on mortality, reinfarction, and strokes in patients with acute myocardial infarction presenting with ST-segment elevation. JAMA 2005;293:427-35.

52. Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI). Lancet 1986;1:397-402.

53. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet 2002;360:7-22.

54. National Emphysema Treatment Trial Research Group. Patients at high risk of death after lung-volume-reduction surgery. N Engl J Med 2001;345:1075-83.

55. Schein OD, Katz J, Bass EB, et al. The value of routine preoperative medical testing before cataract surgery. Study of Medical Testing for Cataract Surgery. N Engl J Med 2000;342:168-75.

56. Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA 2002;288:321-33.

57. Jenicek M, Hitchcock DL. Logic and critical thinking in medicine. Chicago: AMA Press; 2005.