Challenges in Using Nonrandomized Studies in Systematic Reviews of Treatment Interventions

Susan L. Norris, MD, MPH, and David Atkins, MD, MPH

Randomized, controlled trials (RCTs) are firmly established as the standard for determining which medical treatments are effective. In some areas of health care, however, among them surgery, public health, and the organization of health care delivery, most evidence addressing the effectiveness of clinical or policy interventions rests on nonrandomized studies. We examine the use of study designs other than RCTs in Evidence-based Practice Center reports addressing questions of the effectiveness of treatment interventions. These reports offer the opportunity to examine the approaches used and the challenges faced by reviewers when nonrandomized studies are included and their quality assessed. We then offer recommendations for using these studies in systematic reviews of treatment interventions.


Over the 50 years since the first randomized, controlled trial (RCT) in clinical medicine, RCTs have become firmly established as the standard for determining which medical treatments are effective (1). The limitations of nonrandomized study designs are well known to researchers. The Canadian Task Force on the Periodic Health Examination was the first organization to incorporate an explicit hierarchy of different study designs to assist in the evaluation of evidence to support clinical recommendations (2). Similar classifications developed by other organizations all rank well-designed RCTs or meta-analyses of RCTs as the highest-quality evidence of clinical effectiveness. Some have argued that agencies “should fund nonrandomized studies only when convinced that a randomized study is not feasible” (3).

Study designs other than RCTs remain critical for evaluating diagnostic (4) and prognostic strategies, as well as for assessing the harms of interventions (5). However, the debate continues on the role of nonrandomized studies in formulating recommendations about treatments (3, 6). Recent experiences with vitamin supplementation and hormone replacement therapy, in which large clinical trials failed to confirm benefits reported by multiple observational studies (7, 8), have focused renewed attention on the pitfalls of drawing conclusions from nonrandomized studies. Yet in some areas of health care—among them surgery, public health, and the organization of health care delivery—most evidence addressing the effectiveness of clinical or policy interventions rests on nonrandomized study designs. Systematic reviews therefore frequently need to include these studies in order to provide a more detailed picture of our current knowledge and its limitations for clinicians and policymakers.

A distinguishing feature of the Evidence-based Practice Center (EPC) program of the Agency for Healthcare Research and Quality (AHRQ) is the wide range of topics addressed in its systematic reviews (9) and the resulting broad array of study designs included in these reviews. In this paper, we examine the use of study designs other than RCTs in EPC reports addressing questions of the effectiveness of treatment interventions. These reports, completed by 15 different EPCs, offer the opportunity to examine the approaches used and the challenges faced by EPC reviewers when they included nonrandomized studies and assessed their quality. We then offer recommendations for the use of these studies in systematic reviews of treatment interventions.

Of the 107 EPC reports released between February 1999 and September 2004, 78 examined at least 1 question of efficacy or effectiveness of a clinical intervention (Figure 1). Twenty-seven of these reports restricted their review to RCTs (25 were of pharmacotherapy interventions, and 2 examined the effectiveness of medical devices). Forty-nine reports included evidence from study designs other than RCTs. We focus here on these reports, which examined pharmacotherapy, medical devices, surgery, complementary and alternative medicine, and behavioral interventions.

Challenge: Study Design Terminology

The terminology to describe different nonrandomized study designs is inconsistent in the clinical research literature (3), a pattern we also observed in the EPC reports. Herein we use the term nonrandomized study to mean a study with a design other than an RCT, including studies...
in which the investigator assigns treatment group on the basis of a nonrandom strategy (nonrandomized trial), observational studies in which the investigator does not assign treatments (such as case–control and cohort studies), and single-group studies (such as before–after studies and case series) without a comparison group.

CHALLENGE: WHEN TO INCORPORATE NONRANDOMIZED STUDIES INTO EPC REPORTS

No established guidelines address situations in which nonrandomized studies can or should be considered for inclusion in a systematic review or what study designs to consider. Study designs varied among the 49 EPC reports that included at least 1 nonrandomized study. Prospective cohort studies, single-group comparison studies, and nonrandomized trials were included in the largest number of reports (Figure 1).

Most EPC reports did not state the rationale for including nonrandomized studies. Of the 19 reports that provided an explicit rationale for including these studies, all but 1 cited a lack of sufficient evidence from RCTs. In a review of tocolytics and antibiotics for preterm labor (10), the authors explicitly sought observational data despite availability of many RCTs.

In most reports, the decision to include nonrandomized studies appears to have been made early in the review process, but several reviewers broadened or narrowed their inclusion criteria over the course of their review. A review of therapies for treatment-resistant epilepsy took an explicit “best-evidence” approach; the reviewers restricted the review to “controlled studies” when 5 or more such studies addressed a specific question and expanded the review to include case series of various surgical interventions when too few trials were identified (11). Other reviews expanded inclusion criteria on the basis of a more subjective assessment of the evidence from trials (12–16) or dropped plans to include nonrandomized studies after they identified numerous RCTs (17).

The inclusion of different study designs varied with the nature of the clinical question that the review was addressing. Reviews of pharmacotherapy (n = 29), complementary and alternative medicine (n = 8), and behavioral interventions (n = 4) usually combined RCTs with evidence from nonrandomized trials, cohort studies, or before–after studies (Figure 2). On the other hand, reviews of surgical interventions (n = 16) were less likely to find relevant RCT evidence and more likely to include case series and before–after studies.

In addition to examining the number of reports that used nonrandomized studies, we examined contributions of these studies to the total body of evidence within each report. Many of the 49 EPC reports that included nonrandomized studies nonetheless derived most of their evidence from RCTs. In 18 reports, RCTs accounted for most
(≈75%) of the studies used to address an effectiveness question. In 4 of these reviews, the reviewers did not explicitly use the nonrandomized studies to formulate conclusions (18–21).

We identified many discrete situations in which nonrandomized studies were helpful in addressing questions on treatment effectiveness. The following examples illustrate 6 distinct reasons for including nonrandomized studies in a systematic review of treatment interventions.

**Difficulty Conducting Randomized Trials**

Randomized trials are increasingly becoming the standard with which to evaluate surgical and technologic interventions, but certain conditions can make randomization problematic. When interventions are complicated or potentially hazardous and have long-term implications for treatment, patients and clinicians are often reluctant to have treatments assigned by chance. For example, a review of pancreatic islet-cell transplantation for type 1 diabetes mellitus found only case series to assess effectiveness (22). Despite the limitations of case series, the finding that 76% of transplant recipients remained insulin independent for 1 year after transplantation provides fairly compelling evidence of benefit on that important outcome. In most of the 16 reports on surgical interventions, most available studies used nonrandomized designs. Such data may not always be sufficient for clinical recommendations, but excluding them a priori will deny patients, providers, and policymakers potentially useful information.

Occasionally, results of nonrandomized studies may be so striking that it would be considered unethical to randomly assign patients. A careful examination of available nonrandomized studies may nonetheless provide a more accurate picture of the size of benefits, the factors that may modify effectiveness, and the harms of the intervention. A review of total knee replacement found no trials that compared arthroplasty to medical therapy or sham surgery and acknowledged that such studies would probably not be possible (13). Numerous studies, however, measured pain and function with validated instruments before and after surgery. When results were compared and pooled across different instruments using standardized mean differences (whereby the size of the treatment effect in each study is expressed as the difference in means between groups relative to the variability observed in the study), total knee replacement resulted in large and consistent improvements in function, corresponding to improvements of 2 to 4 standard deviations on these scales. The fact that such dramatic improvements are hardly ever seen spontaneously, were consistent across many studies, and persisted in follow-up beyond 5 years lends further credence to the findings despite limitations of the study design.

**Examining Long-Term Outcomes of Treatments Assessed in Trials**

Cohort studies and case series can more easily examine long-term outcomes that may be difficult to observe in RCTs. A review of treatment options for chronic hepatitis C (23)
found numerous RCTs to address efficacy and safety of treatment with interferon and other antiviral agents using biochemical or pathologic end points, but data from trials on long-term outcomes such as clinical cirrhosis and hepatocellular carcinoma were rare. Cohort studies and case series with outcomes beyond 5 years for treated and untreated patients demonstrated effects of treatment that were generally consistent with intermediate outcomes in RCTs. This supported a conclusion that standard interferon-based therapy moderately decreases the risk for hepatocellular carcinoma and cirrhosis in responders to treatment.

Exploring Applicability of Findings from RCTs

Even when well-conducted RCTs are available, questions may remain about the applicability of findings from carefully controlled studies to the typical patient. Observational data can test whether outcomes seen in RCTs can be obtained in more representative populations or settings. If outcomes in observational studies differ greatly from those in RCTs, greater attention to patient selection or provider training may be needed in implementation and evaluation of future interventions. For example, a review of β-mimetic tocolytic drugs for arresting uterine contractions in preterm labor (10) identified 35 RCTs. Reviewers decided to include an additional 12 nonrandomized studies because they felt that participants in these studies were more representative of patients and clinical settings than those in the available RCTs. The nonrandomized data generally supported those from RCTs, strengthening the conclusions.

Clarifying Outcomes for Patients

Before–after studies and uncontrolled case series estimate expected outcomes that are useful to patients. A review of outcomes of bariatric surgery (12) found many trials comparing different surgical techniques but only 2 trials comparing surgical with nonsurgical treatment. Reviewers therefore examined case series to estimate 3-year weight loss for specific procedures. These 20 studies represented 3000 participants, compared with fewer than 300 patients with 3-year follow-up in the RCTs. While case series do not directly compare a treatment with an alternative therapy (for example, diet) and are subject to bias for estimating efficacy, a review of case series may provide a more generalizable estimate of the long-term outcomes that patients care about than does a review restricted to a small number of trials.

Addressing Policy Issues Raised by Nonrandomized Studies

Evidence from nonrandomized studies may create pressure on clinicians and policymakers to adopt new treatments, especially in areas where effective treatments are lacking and trials are difficult to perform. Systematically addressing the strengths and weaknesses of different studies can help clarify controversial issues even if the review does not support definitive conclusions. A review of hyperbaric oxygen therapy examined both RCTs and nonrandomized studies to assess effects on morbidity and mortality and possible adverse effects in patients with brain injury, stroke, and cerebral palsy (24). The RCTs made up only 25% (9 of 39 studies) of the body of evidence and produced inconsistent findings. Observational studies tended to demonstrate more favorable effects of hyperbaric oxygen therapy, but the report outlined specific sources of bias that might account for the results, such as underlying imbalances between compared groups, lack of a stable baseline in before–after studies, and the psychological effects of enhanced care. The report also outlined the poor quality of data on possible harms of hyperbaric oxygen therapy and uncertainty about underlying mechanisms. While the review concluded that the evidence was insufficient for clinical decision making, it outlined how specific study designs might overcome the existing barriers to recruiting patients to RCTs or improve the information gained from high-quality nonrandomized studies.

Clarifying Research Priorities

Careful review of nonrandomized studies can help clarify the potential importance of RCTs and identify ways to improve the quality of nonrandomized studies pending such trials. A review of vaginal birth after cesarean section found no trials comparing outcomes in women offered a trial of labor compared with repeat cesarean (25). A careful review of case series and cohort studies to quantify the potential risks for either approach revealed important differences in risk in 2 large population-based studies. The review concluded that valid estimates of risk and possible benefits of a trial of labor would require an RCT or better cohort studies that matched women on other risk factors, used more reliable measures of outcomes, and collected better information on potential confounders and co-interventions.

Challenge: Incorporating Quality Assessment into Reviews

Assessing quality of nonrandomized studies poses a great challenge. Quality rating has traditionally emphasized identifying threats to internal validity. Quality has been defined as “the confidence that the trial design, conduct, and analysis has minimized or avoided biases in its treatment comparisons” (26). Although problems of selection bias and confounding are widely recognized as important weaknesses in nonrandomized study designs, the empirical basis for defining specific criteria for assessing the quality for nonrandomized studies is less developed than for RCTs. Problems have been demonstrated with quality scoring systems proposed for RCTs (27), and such problems are even greater for other study designs. A more promising approach may be to examine the importance of individual quality components (3).

A variety of approaches were used to assess the quality of individual nonrandomized studies in EPC reports. Of the 49 reports that included nonrandomized study designs, 12 (25%) did not assess study quality, 16% used a previ...
ously published checklist or scoring system, 10% adapted a previously published instrument, and the remainder (49%) used instruments that the reviewers had developed themselves. Quality assessment was most common among the reports that included nonrandomized trials and those including prospective cohort studies (95% and 89% rated quality, respectively). Quality assessment of individual studies was less common (60%) in reports that included case series or before–after designs.

The existing quality assessment instruments that were used or modified included those from the Guide to Community Preventive Services (28), Downs and Black (29), and the U.S. Preventive Services Task Force (30). The EPC reviews used a variety of approaches to develop de novo quality assessment instruments. One third used a score or checklist with 3 or fewer attributes. Some derived an overall quality score from domains that have been demonstrated to, or were assumed to, affect internal validity (for example, blinding of outcome measurement and reliability of outcome measures) (31). Other reviews classified studies into a limited number of categories based on an overall qualitative assessment of quality, similar to that used by the U.S. Preventive Services Task Force (14, 24, 25). Many EPC reports applied the Jadad scale (32) to nonrandomized trials (18, 19, 33), effectively creating a 3-point scale based on blinding and attrition.

The EPC reports incorporated findings of quality assessment into their conclusions in a variety of ways. Among reports that did not formally assess quality of individual nonrandomized studies, many nonetheless commented on how the limitations of existing study designs affected conclusions or recommendations for future research. Although 36 of the 37 reports that presented quality assessment did so in the evidence tables, fewer reports discussed the quality of available studies in the narrative results (69%), conclusions (81%), or recommendations for future research (81%).

Few reports explicitly examined the impact of quality of nonrandomized studies on overall findings (34, 35). One plotted the number of quality criteria met and the percentage of patients available at follow-up against the effect size (35). One review omitted “poor”-quality studies from summary results and conclusions while presenting all studies in the evidence tables and making recommendations for future research using both randomized and nonrandomized methods (24).

Deeks and colleagues identified 193 different quality assessment instruments that were used in publications assessing nonrandomized studies (3). Using epidemiologic and study design principles, they identified 2 key domains for assessing the internal validity of nonrandomized studies: the creation of treatment groups (how allocation occurred and attempts to balance the groups by design) and the comparability of groups at the analysis stage (including assessment of baseline comparability, identification of prognostic factors, and case-mix adjustment). In addition, the authors identified 4 other important domains of quality: blinding of participants, investigators, and outcome assessment; the soundness of information about the intervention and the outcomes; adequacy of follow-up; and appropriateness of the analysis. The authors then identified 14 “best tools” that encompassed at least 5 of the 6 domains; 6 tools were judged suitable for systematic reviews on the basis of their purpose and ease of use. Only 1 of these tools (28) was used in an EPC report (35).

Adapting the Jadad scale (32) for use in nonrandomized studies, as several reports did, does not appear to be adequate for these studies. This approach involves only 2 elements potentially relevant to nonrandomized studies (blinding and attrition), and blinding of patients or clinicians is uncommon in most nonrandomized studies. Moreover, the Jadad scale does not consider the role of important imbalances between comparison groups as a result of selection bias.

**CONCLUSION AND RECOMMENDATIONS**

Evidence from nonrandomized studies was included in most systematic reviews of treatment interventions conducted by the AHRQ EPCs, usually to address questions for which there was little evidence from randomized trials. These studies constituted the bulk of available information for some interventions, especially surgical procedures, and contributed to conclusions in a variety of ways. Reports were rarely explicit, however, about the rationale and tradeoffs involved in including nonrandomized evidence. The methods to assess quality of nonrandomized evidence varied and rarely used validated tools.

Our findings echo those of a comprehensive review on the use of nonrandomized studies in systematic reviews (3). In an overview of 1100 systematic reviews entered into the Database of Abstracts of Reviews of Effect (DARE) before 2000, Deeks and colleagues found that 511 (44%) included nonrandomized studies. Of these, only one third assessed the quality of those studies. In reviews that did assess quality, reviewers rarely used tools that had been validated or that assessed the domains that Deeks and colleagues felt were most important to internal validity.

Given the known limitations of nonrandomized studies, it is understandable that the lack of evidence from RCTs appears to be the major factor leading to consideration of other study designs. A “best-evidence” approach may be the most useful approach for determining what study designs to include and may prevent unnecessary review of nonrandomized studies when randomized data exist. In practice, however, this approach may require repeating initial search strategies because there are no sensitive and specific search strategies for nonrandomized studies as there are for RCTs (36). In addition, changing review protocols in the middle of a review could increase the potential for bias.

Our review describes the various approaches taken by
Using Nonrandomized Studies in Systematic Reviews of Treatment Interventions


Table. Recommendations for the Use of Nonrandomized Studies in Systematic Reviews

<table>
<thead>
<tr>
<th>For systematic reviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assess availability of randomized, controlled trials before determining final inclusion criteria.</td>
</tr>
<tr>
<td>Consider the pros and cons of nonrandomized study designs.</td>
</tr>
<tr>
<td>Provide explicit rationale for decisions to include or exclude specific study designs.</td>
</tr>
<tr>
<td>Assess important domains of quality of individual studies.</td>
</tr>
<tr>
<td>Discuss how the inclusion of various study designs may affect conclusions.</td>
</tr>
<tr>
<td>Discuss how quality of individual studies and the body of evidence may affect conclusions.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>For researchers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimize sources of bias in both randomized and nonrandomized studies.</td>
</tr>
<tr>
<td>Examine the effects of various sources of bias on measured outcomes.</td>
</tr>
<tr>
<td>Use consistent terminology for study design.</td>
</tr>
<tr>
<td>Devise sensitive and specific search strategies for nonrandomized study designs.</td>
</tr>
</tbody>
</table>

15 EPCs with expertise in systematic reviews. It was not designed to test the validity of including evidence from nonrandomized studies. We also did not try to quantify the exact impact of doing so on the report conclusions because we felt doing this retrospectively would require too much subjective judgment. Finally, we did not examine changes over time. Current methods used by EPCs have probably evolved from those used in the earliest reports.

We offer recommendations to reviewers who consider using study designs other than RCTs in their reviews (Table). First, reviewers should assess the availability of RCTs addressing their review question before determining final inclusion criteria. They should search the Cochrane Central Register of Controlled Trials and MEDLINE and contact content experts. Second, when evidence from RCTs is limited, consider the arguments for and against including nonrandomized studies. Reviewers should consider the potential sources of bias and whether they can be minimized with well-conducted nonrandomized studies. Although Deeks and colleagues (3) observed that it is not always possible to predict when observational studies will be biased, certain conditions increase concerns that particular study designs will give misleading results. Are there secular trends in health care delivery that necessitate concurrent comparison groups? Does the disease process fluctuate such that a before–after design might be misleading? Do outcomes involve subjective judgments of patients or clinicians such that blinding of patients or assessors is more important? Similarly, reviewers should consider whether nonrandomized studies are useful to augment information from RCTs (for example, determination of long-term outcomes or testing the generalizability of an intervention to other populations and settings). Third, reviewers must provide an explicit rationale for their decisions to include or exclude various study designs. If nonrandomized study designs are included, how will they inform the review question? Reviewers must consider which specific study designs can provide useful information while adequately protecting against major sources of bias. Any change in inclusion criteria from the original review protocol must be explained and justified. Fourth, the important domains of individual study quality must be considered. Some existing tools address these areas but may need to be modified to facilitate their use and to incorporate topic-specific issues that affect validity. Fifth, reviewers must consider how their inclusion criteria based on study design may affect their conclusions, and the reviewers should discuss the potential impact of possible biases on their conclusions. Sixth, the quality of execution of studies in a body of literature, given specific study designs, should be incorporated into the discussion and conclusions. Are there consistent flaws that might affect the direction or magnitude of benefits or harms?

Our review also suggests several priorities for those conducting nonrandomized studies. In addition to using approaches to minimize known sources of bias, such as selection bias and confounding, researchers should continue to examine the effect of specific quality elements on the magnitude and direction of bias in nonrandomized studies across different clinical areas. Given the frequent use of case series and before–after studies, more research is needed on methods to estimate or reduce bias in these study designs. Researchers and journal editors should work together to promote more consistent terminology for describing specific nonrandomized designs. Finally, efficient search strategies for locating such studies need to be developed and tested.

Systematic reviews of important treatment questions will frequently need to consider nonrandomized studies. Given the challenges inherent in using these study designs in evidence synthesis, following these recommendations will help ensure a more explicit, transparent, and reproducible process.

From the Agency for Healthcare Research and Quality, Rockville, Maryland.

Potential Financial Conflicts of Interest: Authors of this paper have received funding for Evidence-based Practice Center reports.

Requests for Single Reprints: Susan L. Norris, MD, MPH, Agency for Healthcare Research and Quality, Center for Outcomes and Evidence, Room 6325, 540 Gaither Road, Rockville, MD 20850; e-mail, snorris@ahrq.gov

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

References


Current Author Addresses: Drs. Norris and Atkins: Agency for Healthcare Research and Quality, Center for Outcomes and Evidence, Room 6325, 540 Gaither Road, Rockville, MD 20850.