Chapter 2. Methodology

Defining Treatment-Resistant Epilepsy

There is no generally accepted definition of treatment-resistant epilepsy. However, for the purposes of retrieving articles for this report, an operational definition of this term was needed. Accordingly, we defined treatment resistance as failure of one or more AEDs at a maximum tolerable dose to provide complete seizure relief.

This definition was based on consensus obtained during a 1-day meeting with an Expert Panel and subsequent discussions with Technical Experts, during discussions with the two agencies that requested this report, the Centers for Disease Control and Prevention (CDC) and the Social Security Administration (SSA), and in consultation with the Agency for Healthcare Research and Quality (AHRQ).

Many articles did not provide sufficient information to allow a determination of the definition employed (see our conclusions to Question 1). Consequently, and after consultation with the Technical Experts, CDC and SSA, we included articles even if they only stated that the enrolled patients were "treatment-resistant", or used some other synonym (see Question 1 for examples of such synonyms).

Expert Panel

At the beginning of this project, we worked with an Expert Panel that assisted in defining the scope of this Evidence Report, developing its questions, defining the outcomes of interest, and developing the criteria for retrieving and including articles. The involvement of this panel consisted of their participation in a 1-day meeting with ECRI, AHRQ, and representatives of SSA and CDC.

To establish the Expert Panel, we solicited nine organizations to nominate individuals who could serve as its members. All solicitations were preapproved by AHRQ, and all nine organizations nominated an individual. Thus, the Expert Panel was comprised of individuals from the following organizations:

- American Academy of Neurology
- American Academy of Pediatrics
- American College of Occupational and Environmental Medicine
- American Epilepsy Society
- Child Neurology Society
- Citizens United for Research in Epilepsy
- Epilepsy Foundation/Penn Epilepsy Center
- National Association of Epilepsy Centers
- Society for Behavioral and Cognitive Neurology

The participation of these individuals and organizations in this project does not imply their endorsement of the findings of this Evidence Report.

Technical Experts

Subsequent to the 1-day meeting, the Expert Panel was disbanded, and a group of Technical Experts was formed. We collaborated with this group to further refine this project's scope, questions, outcomes of interest, and criteria for retrieving and including articles. The Technical Experts also served as a source of information throughout the project. Collaboration with these Experts was accomplished through telephone conversations and e-mail.

The Technical Experts were comprised of all of the members of the Expert Panel and representatives from:

- Columbia Presbyterian Medical Center
- Harborview Medical Center
- Johns Hopkins School of Medicine
- Strategic Health Institute

As with the Expert Panel, the participation of these individuals and organizations in this project does not imply their endorsement of the findings of this Evidence Report.

Key Questions

This report addresses nine Questions arrived at through the discussions with the Expert Panel, Technical Experts, and representatives from AHRQ, SSA and CDC. These are:

Question #1: What are the definitions of treatment-resistant epilepsy used in the literature?

Question #2: Which methods of rediagnosing or re-evaluating treatment-resistant epilepsy lead to, or can be expected to lead to improved patient outcomes?

Question #3: Is there evidence that patients with treatment-resistant epilepsy are not optimized at their current level of treatment?

Question #4: Which drug treatment strategy, 1) sequential monotherapy, 2) polytherapy, or 3) optimized current therapy leads to improved outcomes for patients with treatment-resistant epilepsy, and what are the relative improvements obtained with each strategy?

Question #5: Which methods of nondrug treatment for epilepsy after initial treatment failure lead to improved outcomes for patients with treatment-resistant epilepsy?

Question #6: Which social, psychological or psychiatric services for treatment-resistant epilepsy lead to, or can be expected to lead to improved patient outcomes?

Question #7: What characteristics of treatment-resistant epilepsy interfere with ability to obtain and maintain employment, or attend and perform well in school?

Question #8: What is the mortality rate of patients with treatment-resistant epilepsy?

Question #9: Is there a correlation between the number and/or type of seizure and sudden death?

Causal Pathway

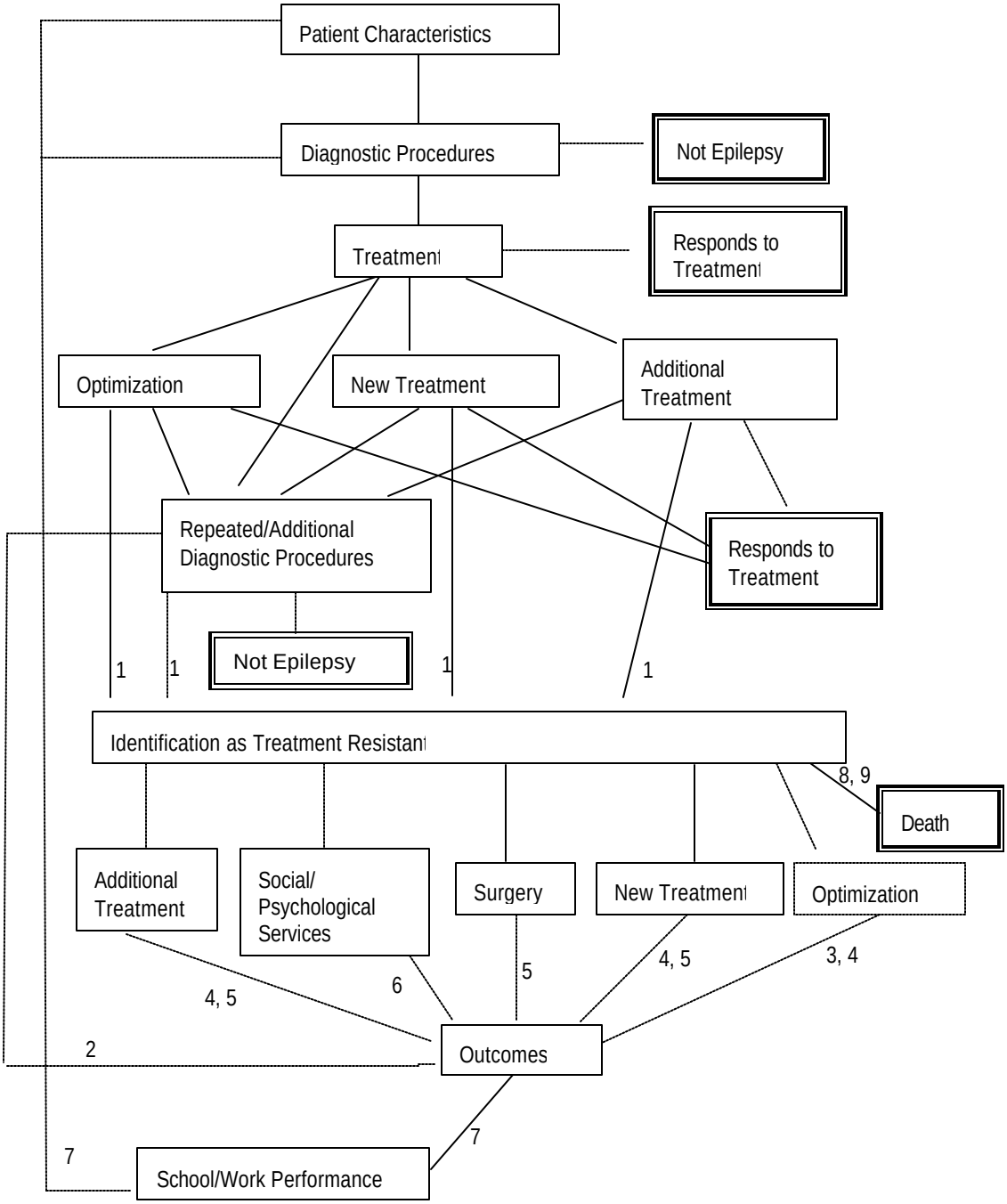
The scope of this report can be illustrated by a causal pathway. More specifically, this pathway illustrates the Key Questions and the relationships among them. It also illustrates items that are beyond the scope of this Evidence Report. This pathway is shown in Figure 1. The rectangles in this figure depict the primary clinical "events," from presentation of a patient (who has certain symptoms that may be at least partly diagnostic and/or prognostic) to the outcomes that the patient experiences (e.g., improves/does not improve). This pathway proceeds in an approximate chronological order that is depicted by solid arrows that connect the rectangles in Figure 1. Because these arrows connect two rectangles, they are termed "links." The numbers next to each link represent the number of the question that addresses that link.

Several boxes represent endpoints in the causal pathway. These are identified by double borders. Patients reaching these endpoints do not go on to additional treatments or diagnostic procedures. Although boxes with no arrows emerging from them represent end points in terms of reporting in published studies, the patients themselves may go on to receive additional treatments and experience further outcomes. The outcomes examined in this Evidence Report were determined by the Expert Panel, Technical Experts, CDC, and SSA. Two outcomes, death and performance in school or work, are broken out from other outcomes because they are specifically addressed by their own questions.

The dashed lines in the figure "overarch" several rectangles. We have drawn these lines as dashed because they do not depict the sequence of events in the clinical pathway. In general, these lines portray questions about how patient characteristics (including clinical findings) may influence outcomes.

Theoretically, a question can be derived by drawing a line between any two rectangles in Figure 1. Therefore, rectangles not connected by solid or dashed lines are beyond the scope of this Evidence Report.

Figure 1. Causal pathway



Literature Searches

Electronic Database Searches

To obtain information for this report, we systematically searched 23 electronic databases. These were:

- Center for International Rehabilitation Research Information and Exchange (CIRRIE) (searched November 30, 2001)
- Cochrane Database of Systematic Reviews (through 2001, Issue 4)
- Cochrane Registry of Clinical Trials (through 2001, Issue 4)
- Cochrane Review Methodology Database (through 2001, Issue 4)
- Cumulative Index to Nursing and Allied Health (CINAHL) (1988 through January 11, 2002)
- Database of Reviews of Effectiveness (Cochrane Library) (through 2001, Issue 4)
- ECRI Health Devices Alerts (1977 through January 2002)
- ECRI Health Devices Sourcebase (through January 2002)
- ECRI Healthcare Standards (1975 through January 2002)
- ECRI International Health Technology Assessment (IHTA) (1990 through January 2002)
- ECRI Library Catalog (through January 2002)
- ECRI TARGET (through January 2002)
- Embase (Excerpta Medica) (1975 through January 2002)
- ERIC (Educational Resources Information Center) (searched January 8, 2002)
- Health and Psychosocial Instruments (HAPI) (through April 27, 2001)
- LocatorPlus (through January 2002)
- NDA Pipeline (searched November 1, 2001)
- PsycINFO (1975 through January 31, 2002)
- PubMed® (MEDLINE®, PreMEDLINE®, HealthSTAR) (1975 through January 2002)
- Rehabdata (searched April 24, 2001)
- U.K. National Health Service (NHS) Economic Evaluation Database (EED) (through January 2002)
- U.S. Centers for Medicare & Medicaid Services (CMS) (formerly HCFA) (through January 2002)
- U.S. National Guidelines Clearinghouse™ (NGC) (through January 2002)

Search Strategies

We employed different searches for different sections of the report, including different searches for different questions. The strategies for these different searches, given in PubMed® /MEDLINE® syntax, are provided in Appendix A.

Other Sources

In addition to the above searches, we also reviewed the bibliographies and reference lists of all studies included in this Evidence Report, and searched Current Contents—Clinical Medicine® on a weekly basis.

Article Inclusion Criteria

To be included in this Evidence Report, an article had to meet specific *a priori* criteria. Some of these criteria were specific to each question, and these are listed at the beginning of the discussion of each question in the Results section of this Evidence Report. Some criteria were common to all questions except Questions #1 and #9. These common inclusion criteria[a] were:

1. The article described a study that enrolled (unless otherwise noted in certain questions) only patients with treatment-resistant epilepsy or, if other patients were enrolled, data from treatment-resistant patients were separately presented.
2. Only full-length articles were included. We did not include meeting abstracts because they are often preliminary reports of results, and they seldom contain sufficient detail to allow evaluation of study design.
3. The article described a study that must (unless otherwise noted in certain questions) have been published in 1985 or later. We adopted this criterion in accordance with the wishes of the Expert Panel, which stated that treatments for epilepsy, the technologies associated with the diagnosis of the disease, and the classification of its seizure types have substantially changed from what they were prior to 1985.
4. Articles had to be English-language. We adopted this criterion out of consideration of the time and budget allotted for this project.
5. The article described a study that enrolled 10 or more patients. Smaller studies may be of unusual methods or patients. Therefore, their findings may not be applicable to other patients or to settings outside the ones in which the study was conducted. Further, small surgical series may represent studies conducted by physicians who have comparatively little experience with the procedure.
6. The article described a study that quantitatively reported an outcome or diagnostic test result of interest. Qualitative expressions of results do not allow conclusions to be drawn about how well a treatment works (i.e., about effect sizes) and provide little assurance that results were rigorously evaluated by the investigators.
7. As per the desires of the Expert Panel and Technical Experts, a study of an intervention must have reported data on one or more of the following outcomes:
   a. *Outcomes related to seizure frequency:*
      - Absolute seizure frequency
      - Percentage change in seizure frequency from baseline
      - Proportion of patients seizure-free
      - Proportion of patients with >50 percent reduction in seizure frequency from baseline
      - Engel Classification

[a] Throughout the remainder of this Evidence Report, we refer to these inclusion criteria as "general" inclusion criteria.

- Rundown time to seizure-free
- Seizure-free period
- Proportion of patients with any reduction in seizures
- Proportion of patients with any increase in seizures
- Proportion of patients exiting a trial due to harmful seizure increases

b. *Nonseizure frequency outcomes:*

- Quality of life
- Mood (we used this as a general term to describe a range of outcomes, from depression to psychosis)
- Functional status/ability
- Cognitive Function
- Ability to stay in or return to work
- Ability to stay in or return to school
- Ability to hold a driver's license
- Adverse events
- Mortality

8. Articles that present data pertaining to quality of life, mood, or cognitive function must have used a validated psychometric instrument. For the purposes of this Evidence Report, a validated psychometric instrument is an instrument for which there is evidence in the peer-reviewed literature to demonstrate that it has construct validity (it measures what it purports to measure) and good reliability (e.g. test-retest reliability; inter-rater reliability). Ideally, the instrument would have been validated using patients with epilepsy. However, given the scarcity of such data on quality of life and psychological status, we decided *a priori* not to require that all psychometric instruments be validated in a population of patients with epilepsy.
9. If there were fewer than five studies of a given intervention or diagnostic, and none of these studies was a randomized controlled trial (RCT) that enrolled 50 or more patients in the treatment group, we did not include any studies of the diagnostic or intervention. Where an RCT with 50 or more patients did exist, we included that RCT even if there were fewer than five studies. We adopted the criterion partly out of consideration of the time and budget allotted for this project. However, this criterion also reduces the potential that publication bias will influence our conclusions. Conclusions drawn from small literature bases could be overturned by the results of only one or two unpublished studies. Therefore, requiring that five studies be available before analyzing a given intervention or diagnostic helps to reduce publication bias.
10. When five or more controlled studies addressed a given intervention or diagnostic, we included only controlled studies. Otherwise, uncontrolled studies were included.
11. Except for surgical topics (where the great majority of studies were retrospective) when five or more prospective studies addressed a given intervention or diagnostic, we included only the prospective studies. Otherwise, retrospective studies were included.
12. When there were several publications describing the same trial, only the largest and most recent publication was included. This avoids double counting patients and the consequent distortion of measurements of effect size. We included earlier studies that reported data not in later publications, and earlier publications that contained data from more patients than later publications.

All surgery trials meeting the inclusion criteria were further examined for patient overlap by cross-matching years of patient enrollment within articles published by the same surgery center. When articles presented overlapping patient populations, the article containing the most recent patient enrollment periods was included. When patient populations did not overlap, all articles from a single center were included.

13. The seizure types examined in a study were classified according to the International League Against Epilepsy's International Classification of Epileptic Seizures, published in 1981,[5,14] or the article used terminology that was consistent with this classification.

As stated above, these criteria do not apply to Key Questions 1 and 9. We provide the criteria relevant to these questions in the text associated with them.

We reviewed the abstracts of articles identified by our searches against the inclusion criteria to determine whether we would retrieve it. Five research analysts independently performed this task, and each analyst worked on different questions. We retrieved an article whenever there was uncertainty about whether it met the inclusion criteria. We also retrieved articles when an abstract was not present in the search results, but the title of the article suggested that it might be relevant.

Once an article was retrieved, it was examined to determine whether it met the appropriate inclusion criteria. Articles that met these criteria were first examined for "fatal flaws" that precluded interpreting their results. Such articles were excluded. When an article was excluded for design flaws, we presented the reason(s) for its exclusion in Evidence Tables associated with each of the nine Key Questions in this report.

We then examined the remaining articles for design flaws that could potentially bias their results. Wherever possible, we empirically evaluated studies for the presence of bias. When data-driven evaluations of study quality were not possible, we documented and explained a study's potential for bias.

Articles Identified

We identified 11,111 articles with our searches. We retrieved 2,356 of these for Questions 2-9 according to *a priori* criteria. Three hundred fifty-seven articles remained after evaluating whether the full article met these criteria. After evaluating these latter articles for design flaws so severe that their results could not be interpreted, and after determining whether there were too few articles to permit a firm evidence-based conclusion (see above for Article Inclusion Criteria), 305 articles remained, and were included in this Evidence Report for Questions 2-9. One hundred eighty-five articles for Question #1 (on definitions of treatment-resistant epilepsy) were randomly sampled from these 305 articles.

The number of articles retrieved, that met the inclusion criteria for each question, and were included in the Evidence Report are shown in Table 2.

Table 2. Number of articles in the Evidence Report

Key Question Number	Retrieved	Met Criteria	Included
1	185	185	185
2	470	34	10
3	201	20	20
4	654	57	50
5	804	206	206
6	86	12	0
7	48	5	0
8	61	14	10
9	32	9	9
Totals =	**2541**	**542**	**490**

Note: Articles for Question 1 were also used to address other questions. Because these articles were evaluated twice, each for a different purpose, they are double-counted in the totals shown in the table. The article count in the text does not incorporate this double-counting

Statistical Methods

Meta-Analyses

We performed meta-analyses of data from RCTs and uncontrolled trials. We performed meta-analyses of RCTs to estimate an average effect of treatment. We performed meta-analytic threshold analyses of uncontrolled studies to determine whether an intervention was *plausibly* effective. All meta-analyses, except meta-regressions, were performed with software programs developed by ECRI. This software has been extensively validated using published examples and hand calculations. Meta-regressions were performed using SPSS (version 10.1) Statistical Software (Copyright © SPSS Inc., 1989-2000).

Meta-analyses of Randomized Controlled Trials

Our meta-analyses of RCTs are exclusively comprised of random effects models. These analyses appear only in Question 4, and only in regard to the polytherapy drug strategy. We employed random effects models because of the types of trials that addressed this question. These trials were each of a single AED, and not all trials studied the same drug. Therefore, no assumption is made that these trials were all drawn from a single population.

An important aspect of these meta-analyses is that each of the trials is an *instance* of polytherapy, rather than a study of polytherapy, *per se.* This is because a planned study of polytherapy would investigate more than one drug and means that each trial in each meta-analysis in Question 4 represents only one way polytherapy might be tested. However, combining these trials into a single meta-analysis of polytherapy means that this analysis approximates a single trial that directly studied this strategy of drug administration.

Another important aspect of the RCTs in our meta-analyses is that some of the trials consisted of more than two groups and therefore have more than two effect sizes. These effects

are not independent of each other. Ideally, sophisticated statistical techniques would be used (e.g., general linear modeling or hierarchical regression) to account for this dependence. However, in the present case, there was too little published information to permit such analyses. Therefore, we conducted two analyses on the data from any given set of trials. In the first analysis, we evaluated the effects of the intervention by comparing outcomes in the group that received a study's highest drug dose to outcomes in the placebo group. In the second analysis, we compared outcomes in the group that received a study's lowest drug dose to outcomes in the placebo group. Additional details about the specific analyses we conducted are provided in Question 4.

Because the results of these analyses are not independent, if a meta-analysis of a low dose of drug is statistically significant, a subsequent finding that the effects of high doses are also statistically significant does not provide "twice as much" evidence that the treatment is effective. Further, because the data allow computation of only a meta-analytic summary statistic, and do not permit attempts to explain the cause(s) of any between-studies heterogeneity, some information loss accompanies all of these analyses.

Our random effects analyses were performed as described by DerSimonian and Laird.[15] As the measure of the effectiveness of treatment, we employed Cohen's h, the difference between the arcsine transform of two proportions divided by the pooled standard error.

We did not compare the effectiveness of different drugs. This conforms to the wishes of the Expert Panel and Technical Experts, who advised that such comparisons were of secondary importance.

Meta-analyses of Uncontrolled Studies

As mentioned above, our meta-analytic threshold analyses of uncontrolled studies are not intended to produce a summary statistic. Rather, we performed these analyses to assist readers in determining whether an intervention is *plausibly* effective. In these threshold analyses, we meta-analytically compared the improvement rate in treated patients to increasing rates of improvement in a hypothetical control group. Starting at 0 percent, we increased the rate of improvement in the "control" patients until the difference in improvement between the treated and "control" groups was no longer statistically significant. This value is the threshold. Thus, the threshold is the proportion of untreated patients who would have to improve to render the effect of treatment statistically nonsignificant. Except for analyses of sequential monotherapy, all of these analyses employed fixed effects models. In analyses of sequential monotherapy, we employed random effects models due to the use of different drugs as monotherapy. Where possible, we provide context for these thresholds by supplementing them with historical data on "control" patients obtained from published articles.

We also report the percentage of patients who improved after the intervention, but note that this percentage is not the difference between improvement in a treated and a control group and, therefore, is not the net effectiveness of the intervention. Nevertheless, this percentage is informative because it represents the proportion of patients likely to improve, regardless of the cause of their improvement. We estimated this percentage by back-transforming the summary statistic from the meta-analysis into a percentage.

We conducted these analyses using Cohen's h as the test statistic. In all analyses, we assumed that the number of patients in the "control" group equaled the number of patients in the treated group. We chose Cohen's h because, under these conditions, the Q statistic and each

study's standardized residual remain constant as the proportion of improved patients in the "control" group increases.

Meta-regression

Whenever statistically significant heterogeneity among the study results was detected in our threshold analyses, we attempted to "explain" the heterogeneity using meta-regression. We used the Q statistic to determine whether an analysis was heterogeneous. Because this statistic is conservative,[16,17] we adopted a p-value of 0.10 (as opposed to the traditional significance level of 0.05) as the critical value for statistical significance.[18]

Typically, there was no strong *a priori* hypothesis to "explain" this heterogeneity. Therefore, we generated a set of regression models for any meta-analysis in which we found statistically significant heterogeneity. We constructed this set by first computing all possible models containing one predictor variable. The number of available predictor variables was often limited by incomplete reporting. This is because we required that at least 90 percent of the studies report the value of a given variable before we entered it into a meta-regression. When more than 90 percent of studies, but less than 100 percent of studies reported the value of a given variable, we assumed the mean value of the variable for the missing data.

After generating all possible one-predictor models, we generated all possible two-predictor models except those containing the coefficients that were not significant in the one-predictor models. Finally, we constructed all three-predictor models except those containing a pair of coefficients that were nonsignificant in the two-predictor models. We only constructed three-predictor models when $(Q_{E1}-Q_{E2})/Q_{E0} > 0.25$, where Q_{E0} is the value of Q_E when there were no predictors in the regression model, Q_{E1} is the value of Q_E when there was one predictor in the regression model, and Q_{E2} is the value of Q_E when there were two predictors in the model. We employed this rule to avoid over fitting data from small numbers of studies. Constructing multiple models also assisted in detecting multicolinearity.

For the purposes of constructing models, we set the alpha level required for significance of the regression coefficients at 0.10. This is anticonservative, but allows for examination of a broader range of models than would an alpha of 0.05.

In the text of the Evidence Report, we consider a model to be a plausible "explanation" of variability only if: (1) it was the only model in a set to produce a statistically nonsignificant ($p > 0.10$) Q_E, (2) all coefficients in the model were statistically significant and, (3) adding another predictor variable to the model caused the value of Q_E to decrease by less than 25 percent with respect to the value of Q_E with no predictors in the model. For interpreting models, we used the traditional alpha level of 0.05 for the regression coefficients.

Other Meta-analyses

To ensure there were no systematic biases in the enrollment of patients in RCTs, we conducted, wherever possible, meta-analyses of the characteristics of patients enrolled in them. In these analyses, we compared the characteristics of patients in the control groups to the characteristics of patients who received the intervention. For example, in one such meta-analysis, we sought to determine whether females tended to be enrolled more in the control groups than in the experimental groups of studies of AEDs. These analyses employed fixed effects models, and we used Cohen's h or Hedges' d, as appropriate, to estimate the between-group differences.

Recognizing that meta-analysis has low statistical power to detect influences of patient characteristics on outcomes,[19,20] we extended our fixed effects analysis on surgical outcomes by performing appropriate meta-analyses of data from nested case-control studies. These studies reported the proportion of patients with a given characteristic who had successful or unsuccessful surgery, or they separately reported a continuous variable for patients who had successful or unsuccessful surgery. We performed meta-analyses on proportions using Cohen's h as the test statistic. For studies reporting continuous variables, we computed each study's appropriate point-biserial correlation coefficient and then meta-analytically evaluated these coefficients.

We performed analyses of data from nested case-control studies using only patient-level data. Although meta-analysis of such data using more sophisticated modeling techniques is preferable (e.g., hierarchical models), this was beyond the scope of the present project.

Sensitivity Analyses

We used sensitivity analysis to test whether our meta-analytic summary statistics were robust. We employed four such analyses for each summary statistic. These were recalculations of the meta-analytic summary statistic with: (1) the largest study removed, (2) the smallest study removed, (3) the study with the largest effect removed, and (4) the study with the smallest effect removed from the meta-analysis.

Other Computations

In addition to computing the above-described meta-analytic statistics, we performed numerous other statistical computations. We note each of these computations in the text of this Evidence Report and/or in footnotes to the in-text tables and Evidence Tables. Briefly, the computations we performed included:

1. Statistical power analyses. Studies that do not contain a sufficient number of patients cannot detect statistically significant differences between groups, even when these differences are clinically meaningful. Therefore, when appropriate, we computed the smallest between-group difference that any given controlled study had the power to detect.
2. Determinations of whether there were statistically significant differences between the characteristics of patients in the groups of any given study. This is particularly important for studies that are not randomized, because the patients in the different groups of such studies may not be comparable. Further, although other studies may report that they were randomized, the randomization protocol may not have been adequately followed or the study may not have been truly randomized (i.e., randomization may have been nonstochastic). These departures from randomization can manifest themselves in pretreatment between-group differences in patient characteristics. We recognize that in a properly randomized trial, such differences can arise from chance. However, searching for such differences in the context of a systematic review is justifiable because there is no other way to audit whether the randomization was, indeed, accomplished.
3. Computation of pretreatment effect sizes. Departures from randomization can also manifest themselves as a statistically significant difference in the outcome between groups prior to the administration of treatment. For example, if the seizure frequencies

experienced by patients in different groups were significantly different before treatment, the study may not have been truly randomized.

4. Verification of 2 x 2 tables reported in studies of diagnostic tests. Because peer-reviewed published articles often contain errors in reported results, we attempted to verify the calculations in each article. If an error was found, we corrected the data and included it in our analysis.
5. Computations of t-tests, chi-square tests, Fisher's exact text, odds ratios (OR), and their 95 percent confidence intervals (CI). Some studies included in this Evidence Report did not report the results of statistical tests that were important for answering the questions. We computed these statistics when such studies reported sufficient data.
6. Computations of crude (CMRs) and standardized mortality ratios (SMRs). These quantities, which are useful for comparing the mortality rates among persons with epilepsy and those who do not have epilepsy, were not reported in all studies but some studies reported sufficient data to allow us to compute them. CMRs and SMRs are calculated as the number of observed deaths divided by the number of expected deaths. However, SMRs are standardized according to the age distribution of the study population and the age-specific death rates in the country of interest. If a study reports only a mean age or age range for their patient group, then only a crude estimate can be made as to the number of expected deaths. Therefore, caution is required in interpreting crude mortality ratios. We discuss the reasons required for this caution in detail in Question 8.
7. Numerous other calculations of descriptive statistics for patient characteristics (e.g., mean age) and outcomes (e.g., seizure frequency) were performed when patient-level data were reported.

Methods of Evaluating Literature Quality

Studies of Interventions

Our evaluation of the quality of interventional studies employed three tools. The first was an evidence hierarchy we used to determine which studies to retrieve, include and, in certain questions, to determine whether there was greater potential for bias in some included studies than in others. The second tool we used was a checklist for evaluating each study's internal validity. Finally, we considered the difficulties inherent in certain outcome measurements.

Evidence Hierarchies

We did not restrict the studies included in this Evidence Report to RCTs. Rather, ours is a "best evidence" synthesis in which we accept the best available evidence, not the best possible evidence. Performing such an analysis on studies of surgery for epilepsy is particularly important. This is because withholding treatment to perform such an RCT may be unethical.

To determine the best available evidence, we used an evidence hierarchy. This hierarchy served, in part, to determine which studies we would include and retrieve. In some cases, we also used this hierarchy to evaluate a study's potential for bias. This hierarchy, shown in order of study designs with the least potential for bias to those with the greatest potential for bias was:

- Randomized controlled trials
- Controlled clinical trials
- Studies that made measurements before and after treatment (pre/post studies)
- Uncontrolled studies

Within each level of the hierarchy, we assumed blinded studies to have lower potential for bias compared to nonblinded studies, and prospective studies to have lower potential for bias compared to retrospective studies. Wherever possible, we empirically evaluated these assumptions, and the assumption that studies lower in the hierarchy had results that were different (i.e. were biased) from studies higher in the hierarchy.

Internal Validity Checklist

The internal validity of an interventional study represents the degree of confidence one can have in whether the intervention caused a change in the outcome of interest. The confidence in this causal relationship can be weakened by a number of biases. Because of this, we employed a second tool, geared to evaluate the potential difficulties with each included study's internal validity. This tool was a checklist of such potential difficulties, and was a modification of the scheme of Cook and Campbell.[21] Thus, we evaluated each study to determine whether any of the potential biases listed below was present.

We stress that these are *potential* biases. The existence of a potential bias does not necessarily mean that a study's results were affected by the bias. We view this question as one that can be empirically determined.

Selection bias

Selection bias is relevant only to studies with control groups. This bias occurs when there are differences between the patients in the different arms of the study at the start of the study. These differences may lead to posttreatment differences in outcome that are not due to treatment. Random assignment of patients to the study arms protects against this bias, but the fact that a study states that assignment was random does not guarantee that randomization protocols were adequately followed. This is a concern when the method of randomization is not reported. In such instances, the method of randomization may not be truly stochastic.

Investigator bias

This bias can occur in studies that are not blinded. In such nonblinded studies, investigators are aware of who is receiving a particular treatment and who is not. This knowledge may influence the measurement of patient outcomes, especially when these outcomes rely on a degree of subjectivity. This bias can affect nonblinded studies of any design.

Patient bias

This bias can occur in open studies or in blinded studies in which blinding has been broken. As a result, patients are aware that they are, or are not, receiving a treatment. This knowledge may influence the way they report an outcome of interest. Given that many seizure frequency and quality of life outcomes rely heavily on patient reports (e.g., seizure diaries), this bias is particularly relevant to the studies considered in this Evidence Report. This bias affects nonblinded studies of any design.

Attrition bias

Attrition refers to the loss of patients before outcome measurements can be recorded. Patients may no longer return to the clinic because they have moved away, have improved to the extent that they believe they no longer need to see a physician, or have died. Because those who completed the study may not be representative of the entire group of patients who entered the study, analyses based only on study "completers" may be biased. This bias can affect studies of any design. In this report, we did not set any limit for attrition beyond which we would not consider the study in our assessment. Exiting a trial before completion was considered an important outcome in our evaluation of studies of drug treatment. In our evaluation of vagal nerve stimulation, we specifically looked for any influence of attrition on outcomes.

Measurement bias

Measurement bias occurs when the method used to measure a particular outcome systematically over- or underestimates the true effect of treatment on that outcome. For example, general health status instruments (e.g., SF-36) may be less sensitive than disease-specific instruments for detecting small changes in health status that are important to patients.[22] This bias can affect studies of any design. In the epilepsy literature, the use of seizure diaries to measure seizure frequency may be a source of potential measurement bias. We discuss why in the section entitled "Validity of Seizure-related Outcomes" (below).

Regression bias

This bias, also known as regression to the mean, can occur when there are patients who, upon entry into a study, have relatively good (or relatively poor) performance on an outcome. For example, patients may enter a study when their condition is at its worst. When the disease is not progressive, these patients are unlikely to be so ill upon subsequent measurement, even in the absence of treatment. Patients with extremely high pretreatment seizure frequencies may experience reductions in seizure frequency, even without treatment.[23] This bias affects studies of all designs except well-designed RCTs.

Extraneous event bias

This bias occurs when events other than the intervention of interest cause improvements in health outcomes. For example, in an uncontrolled longitudinal study, treatment may incorrectly appear to cause an improvement in health outcomes if patients are given new, effective methods of patient management. This bias can affect studies of all designs including RCTs. RCTs will be affected if the new methods are not uniformly applied to all patients.

Sampling bias

Sampling bias occurs when a study either does not include all enrolled patients who received the treatment of interest, or does not include a random sample of the enrolled patients who received the treatment of interest.

Maturation bias

Maturation bias occurs if individuals improved because of developmental maturation, and not because of treatment. In the present Evidence Report, we only evaluate studies for potential maturation bias if they used followup periods longer than 1 year.

Sample specification bias

In the context of the present report, sample specification bias occurs if a study enrolled some patients who were not treatment-resistant. For example, studies that did not explicitly state that all enrolled patients were experiencing seizures despite prior treatment with at least one AED given at maximum tolerable dose may not have exclusively enrolled patients who met the definition of treatment resistance that was suggested by the Expert Panel and Technical Experts (see the section entitled "Defining Treatment Resistant Epilepsy"). Sample specification bias is a lesser issue for studies of surgical interventions for epilepsy than for the other interventions discussed in this Evidence Report. This is because of the relatively extensive presurgical evaluations that surgical candidates receive.

Statistical power

Studies with low statistical power do not have the ability to find statistically significant differences between groups or between one test and a subsequent test. As such, the failure of a low power study to find a statistically significant difference does not always imply that an intervention is ineffective. This is because the study may not have had the power to detect clinically important differences.

Determining whether a study has sufficient power to detect clinically important differences is subjective. This is because determination of what a clinically important difference is ultimately requires the opinions of patients. Often, these opinions are not well studied. Consequently, we have refrained from making such judgments and, instead, provided the reader with sufficient information to make their own. We accomplish this by computing the smallest between-group percentage difference that a statistically nonsignificant study could have detected. The reader then needs to compare this percentage to the percentage change deemed clinically important. Assume, for example, that we computed that a study only had the power to detect a 30 percent decline in seizures as statistically significant. The reader may decide that, in fact, a 10 percent decline in seizure rates is clinically significant, and then note that the study did not have the statistical power to detect this difference. This would mean that the results of the study were not informative.

Because our consideration of a study's power involves *de novo* calculations, we consider power in the "synthesis of study results" section of each question, and not in the section devoted to evaluating a study's internal validity.

Validity of Seizure-related Outcomes

There are several commonly reported ways to measure seizures, including mean and median frequencies, the proportion of patients who experience a reduction in seizures greater than a certain percent (e.g., the proportion of patients who experience a greater than 50 percent reduction in frequency), and the percentage of patients who exit a trial due to seizure increases.

All of these outcomes depend upon patient reports, often in the form of seizure diaries. One difficulty with these diaries is that they rely on the objectivity, and memory of the individual responsible for keeping the diary. This affects the accuracy of records of all seizures (see "Measurement bias" above), and accurate recording of auras may be particularly problematic.

Another problem with the way in which seizure frequencies are reported is that not all outcome measurements capture what happens to *all* patients. For example, a common way to report seizure frequency is to report the percentage of patients who had a 50 percent (or some

other percentage) or greater reduction in seizure frequency after treatment. This type of outcome only captures information about patients whose seizure rates decreased. It does not capture information about patients who experienced increases in seizure frequency. In fact, expressing results in this way can be quite misleading. For example, assume a study that reported that seizure frequency reduced by 50 percent or more in 40 percent of their patients. Seizure rates may have actually increased in the remaining 60 percent of patients and, if this were true, the treatment might actually be harmful.

In contrast, measures of absolute seizure frequency do capture information about all patients because the results of patients who became worse are combined with the results of those who improved. However, absolute seizure frequencies are not normally distributed, so a simple average is not an appropriate summary of the data. As a result, many studies report median seizure frequencies. While medians are an appropriate measure of the central tendency of such data, they pose technical difficulties. In particular, methods for combining medians in a meta-analysis are not well developed. One way around this difficulty is to transform seizure frequencies by using a natural log transform. This would render the data normally distributed and allow for computation of meaningful averages and measures of dispersion. Published studies rarely report such results.

Another way of reporting results is to provide the number or proportion of patients who exited a trial due to changes in seizure type and/or frequency. Although trial exit *per se* can be accurately recorded, it is also based on seizure frequencies and, therefore, typically depends on the accuracy of seizure diaries. Another difficulty with this outcome is that it is relatively insensitive to small or moderate increases in seizure frequency. This is particularly true because a common criterion that investigators set for exiting a monotherapy drug trial is a doubling of seizure frequency.

Studies of Diagnostics

The biases that can affect studies of diagnostics are different from those that can affect studies of interventions. The checklist we employ for determining whether a diagnostic study was potentially affected by a bias incorporated items suggested by Lijmer, Mol, Heisterkamp, et al.,[24] Irwig, Tosteson, Gastsonis, et al.,[25] Gann,[26] Begg and Greenes,[27] and Ransohoff and Feinstein.[28] These biases are:

Spectrum bias

This bias occurs when there are differences between populations in the spectrum of disease presentation and severity. In the present report, it manifests itself in diagnostic case-control studies in which “cases” (in this instance, patients with epileptic seizures alone) and “controls” (patients with nonepileptic seizures) were selected for inclusion because they were known to have epileptic or nonepileptic seizures prior to the study. Such studies therefore enrolled cases that are relatively easy to diagnosis, and did not enroll cases that are more difficult to diagnose. The effects of spectrum bias have recently been demonstrated empirically by Lijmer, Mol, Heisterkamp et al.[24] who found that the diagnostic odds ratio was approximately three times greater in diagnostic case-control studies compared to studies of the same diagnostic carried out using unbiased populations.

Imperfect reference standard bias

This bias occurs when a reference standard against which the diagnostic performance of the diagnostic of interest was measured is not perfect (not a true "gold standard").

Differential reference standard bias

In the context of this Evidence Report, this bias occurs when patients allocated to the epileptic and nonepileptic seizure groups were not diagnosed using the same reference standard. For example, patients with epileptic seizures may have been diagnosed in a neurology department using a diagnostic such as video-EEG, but patients with syncopal seizures may have been diagnosed in a cardiac department using a diagnostic such as a tilt-table.

Prevalence bias

Prevalence bias occurs when the numbers of cases and controls in a case-control study are artificially chosen to be equal. This artificial prevalence introduces a bias that influences the positive predictive value (PPV) and negative predictive value (NPV) in a manner described by Bayes' theorem.[29]

Interpretation bias

This bias occurs when the results of the test of interest are subjective and can be influenced by factors that are unrelated to the disease of interest.

Patient bias

This bias may occur in diagnostic studies when patients are aware of their diagnostic group allocation. This bias is a particular problem when the diagnostic of interest involves patient input. For example, the Minnesota Multiphasic Personality Inventory (MMPI) has been proposed as a means of differentiating patients with epileptic seizures from patients with psychogenic seizures. This instrument requires patient input. Patients' awareness of their diagnostic group allocation may influence their input.

Investigator bias

This bias may occur in diagnostic studies when investigators interpreting the results of the diagnostic of interest are not blinded to the diagnostic group allocation of the patients in the study. This is a particular problem when the investigator is required to "interpret" the findings of a diagnostic test. For example, the interpretation of a CT scan requires that an investigator interpret the image. If the investigator is aware of the diagnostic categorization of the patient, his interpretation of the CT image may be influenced.

Verification bias

This bias is only relevant to studies that used followup to confirm the accuracy of the diagnostic of interest and occurs when only one group of patients is followed. This group typically consists of only those with a positive diagnosis. For example, only those diagnosed by the test of interest might be followed up.

Diagnostic yield bias

This bias may occur when only a subset of patients enrolled in a study is reassessed. For example, some patients do not experience a seizure during re-evaluation, so diagnostic data cannot be collected from them. If these patients are somehow different from patients in whom a

diagnostic reassessment was possible (for example, the subgroup contained a higher proportion of patients with nonepileptic seizures), then this may lead to a biased estimate of prevalence.

Studies of Mortality

Two questions in this Evidence Report, Questions #8 and #9, concern epilepsy-related mortality rates. Studies examining mortality rates may be biased by factors that are different from those that bias the results of other kinds of studies. Therefore, we examined studies of mortality for the following potential biases:

Sample specification bias

See the above definition of this bias.

Sampling bias

See the above definition of this bias.

Cause validation bias

This bias may occur in studies that did not determine the cause of death by autopsy. For example, investigators may assume that epilepsy is the cause of death in patients with epilepsy who died suddenly. This bias will artificially inflate death rates due to epilepsy.

Mortality ratio bias

This bias occurs in studies that did not present standardized mortality ratios or in studies that did not present sufficient information to allow us to calculate these ratios. Other methods of computing mortality do not allow mortality rates to be standardized by age, which could bias mortality differences in either direction.

Control selection bias

This bias affects only Question #9 regarding sudden unexplained death. It occurs when studies of mortality use an inappropriate control group. For example, in a case-control study of sudden unexpected death where all of the cases were children with epilepsy, the control group should not consist of adults with epilepsy. This would increase the likelihood of finding a spurious relationship between sudden death and a variable that may be unrelated to sudden death (e.g. childhood epilepsies are likely to differ from epilepsies that afflict adults in ways that may be unrelated to the risk of sudden death). An appropriate control group would be living children with epilepsy.

Statistical control bias

This bias also affects only Question #9. It may occur if studies evaluating a relationship between two variables did not use a statistical method that adjusts for the possible effects of other variables. For example, regression techniques are often useful for determining the influence of a variable on an outcome. When such techniques are not used, the magnitude of the relationship between a variable and the outcome may be misestimated.

External Validity

We evaluated each study's external validity (generalizability) according to patient characteristics appropriate to each question. These characteristics are provided as we address each question. We did not evaluate external validity when evidence-based conclusions could not be reached.

Presentation of Results

Evidence Tables vs. Tables

The results of our analyses of internal and external validity and of our meta-analyses are presented in two types of tables. Evidence Tables contain detailed information on each of the studies used in an assessment and the results of meta-analyses of these studies. The Evidence Tables tend to be large and are therefore contained in a separate volume of this report. They are organized according to the Key Questions addressed in this report. Other tables appear in the Results chapter following the discussion of each intervention being assessed. These tables are intended to provide a brief listing and description of the studies and the outcomes reported in these studies. Tables addressing the internal validity of studies used in an assessment are presented in Appendix B.

Figures

Figures are also presented in the Results chapter after the discussion of each intervention assessed in this report. Figures are designed to present summary information in the form of a forest plot (array of study effect sizes usually with a summary estimate), graphs of threshold analyses and meta-regressions, or other appropriate graphical presentations.

Peer Review

Internal Review

Throughout the preparation of this report, the five analysts and the Project Manager held numerous meetings to determine the strategy and methods of analysis. The Project Manager then individually reviewed each completed section of the report, and suggested changes. Upon completion of these changes, the individual sections were assembled into an initial draft report that was again reviewed by the Project Manager. Subsequent to changes made in response to this draft, it was distributed to the five analysts in the project team for review. Suggested changes were reviewed by the Project Manager and discussions were held among the project team to determine which suggestions would be incorporated. Upon incorporation of the appropriate changes, the draft report was sent for external review.

External Review

To select peer-reviewers for the draft Evidence Report, ECRI prepared a list of 27 potential reviewers. This list was submitted to AHRQ, which approved all reviewers. Letters inviting these

individuals to review the draft report were then mailed. Twenty-five individuals responded to these letters, 19 agreed to review the draft Evidence Report, and nine individuals returned reviews.

Upon receipt of reviews, ECRI revised the draft report accordingly. ECRI also prepared a document describing the disposition of all substantive reviewer comments and supplied this document to AHRQ for review and approval.