Chapter 2. Methods

Original Proposed Key Questions

The American College of Physicians-American Society of Internal Medicine (ACP/ASIM) and the American Academy of Pediatrics (AAP) nominated the topic of this report. Other interested groups include the American Academy of Family Physicians (AAFP) and the National Heart, Lung, and Blood Institute (NHLBI). These groups provided us this initial list of questions.

1. *What is the evidence that pharmacotherapy is effective in weight loss and maintenance of weight loss?*

2. *Are certain agents more effective than others?*

3. *Do certain populations (e.g., gender- or age-related, racial/ethnic populations) benefit more from different agents?*

4. *What is the optimum amount of time to treat, and what is the optimum level of weight loss to target? Do optimum amount of time to treat and optimum levels of weight loss differ according to a patient's age? Gender? Racial/ethnic population?*

5. *What are the most effective non-pharmacological, non-surgical treatment approaches (e.g., individual vs. group; specific dietary regimens in conjunction with other therapies; alternative medicine)?*

6. *What is the safety and efficacy of surgical therapies, such as stomach stapling and bypass surgeries, as interventions for children and adolescents with morbid obesity?*

Technical Expert Panel

Each AHRQ evidence report is guided by a Technical Expert Panel (TEP). We invited a distinguished group of scientists and clinicians, including individuals with expertise in obesity, human nutrition, surgery, pediatrics, and pharmacology, to participate in the TEP for this report. A list of members is included in Appendix D. Our TEP conference call was held on January 28, 2003. The subjects discussed were the following: identification of pharmacological agents to include in review; limitation of the scope of question 5 (regarding nonpharmacologic and alternative treatments); and broadening the scope of question 6 to include adults.

TEP members suggested that our assessment of pharmacological agents include FDA-approved weight loss medications and other medications for which reports have begun to appear regarding their use as weight loss agents. The FDA-approved weight loss drugs are phentermine, sibutramine, orlistat, diethylpropion, and mazindol; however, our TEP advised us to disregard mazindol because it is no longer used. Drugs for which reports have begun to appear regarding weight loss, but that are not currently FDA approved for weight loss, included the antidepressants fluoxetine and bupropion, and topiramate, a drug indicated for seizure control.

Note: Appendixes and Evidence Tables cited in this report are provided electronically at http://www.ahrq.gov/clinic/epcindex.htm.

Our TEP also instructed us to include only studies that had treatment durations of six months or longer, considering shorter durations of treatment to be less informative about effectiveness.

RAND staff suggested that the panel narrow the focus of question 5, above, because the list of nonpharmacological and alternative medicine therapies is potentially infinite. The TEP agreed that diet was the most important therapy to focus on. At a subsequent conference call with AHRQ and ACP/ASIM, the topic of diet was dropped from further review because the number of studies and their heterogeneity made it impossible to synthesize the data satisfactorily within the resources of the project. On this phone call, we agreed that the subject of this report would be the efficacy of drug therapy and surgical therapy.

Finally, regarding question 6, RAND staff expressed curiosity as to why the scope of this question was restricted to adolescent and pediatric populations, since most of the data on these procedures would derive from studies of adult populations only. The TEP agreed that we should examine data on bariatric surgery with patients of all ages in mind. In fact, we broadened our search of all interventions to include adolescent and pediatric populations.

Literature Search

Our search for controlled human studies of pharmacological and surgical treatments of obesity began with an electronic search of MEDLINE® on October 16, 2002. Appendix A shows our specific search strategies. MEDLINE®, which is maintained by the U.S. National Library of Medicine, is widely recognized as the premier source for bibliographic coverage of biomedical literature. It encompasses information from *Index Medicus*, the Index to Dental Literature, and the Cumulative Index to Nursing and Allied Health Literature ([CINAHL]; allied health includes occupational therapy, speech therapy, and rehabilitation), as well as other sources of coverage in the areas of health care organization, biological and physical sciences, humanities, and information science as they relate to medicine and health care. Subsequently, our librarian conducted "current awareness" search updates on May 22, June 2, June 12, and July 3, 2003.

We also searched the Cochrane Controlled Clinical Trials Register Database. The Cochrane Collaboration is an international organization that helps people make well-informed decisions about health care by preparing, maintaining, and promoting the accessibility of systematic reviews on the effects of heath care interventions.

Additional Sources of Evidence

U.K. University of Aberdeen report. Broom and colleagues prepared a systematic review of the long-term outcomes of treatments for obesity, the implications for health improvement, and the economic consequences, for the National Health Service of the United Kingdom.[34] Members of the project team kindly shared with us the draft report October 2002. We ordered all studies referenced therein; using their search terms, we conducted an update of their library search. Search terms are provided in Appendix A. The final report has since been released.

Orlistat review. In 2001, O'Meara and colleagues published a rapid and systematic review of the clinical effectiveness and cost-effectiveness of orlistat in the management of obesity.[35] The review was published as a Health Technology Assessment by the United Kingdom National Health Service (NHS). We ordered all studies referenced therein.

Note: Appendixes and Evidence Tables cited in this report are provided electronically at http://www.ahrq.gov/clinic/epcindex.htm.

Sibutramine review. We identified a high-quality meta-analysis on the effects of sibutramine that was "in press" at the time of our search.[36] The first author of this review agreed to allow us to incorporate their results into our evidence report. The authors of that review did not restrict their literature search by language and made extensive efforts to identify unpublished and ongoing trials, including contacting representatives of the pharmaceutical industry.

NICE surgery review. In July 2002, the United Kingdom's National Institute for Clinical Excellence (NICE) published an assessment report[37] on the clinical and cost effectiveness of surgery for morbidly obese patients, authored by Clegg and colleagues at the University of Southampton. We ordered all studies referenced therein.

Cochrane surgery review. Our literature search also identified a Cochrane review[38] on surgery for morbid obesity, published by the team that authored the NICE surgery review. The Cochrane review was updated in February 2003 during our literature search process. The authors included both randomized controlled trials (RCTs) and non-randomized controlled trials comparing surgery with nonsurgical management for morbid obesity and assessed RCTs comparing different surgical procedures. The review was restricted to adults age 18 years or older with morbid obesity defined as a BMI greater than 40 or a BMI index greater than 35 with serious comorbid disease. In addition to summarizing their results, we obtained copies of studies referenced therein and conducted an update of their library search.

Article Review

We reviewed the articles retrieved from the various sources against our exclusion criteria to determine whether to include them in the evidence synthesis. A one-page screening review form (checklist) that contains a series of categorization questions was created to track the articles (see Appendix B). After being evaluated against this checklist, each article was either accepted for further review or rejected. Three reviewers, each trained in the critical analysis of scientific literature, independently reviewed the studies, abstracted data, and resolved disagreements by consensus (each study was reviewed by two reviewers: One reviewer assessed all the studies and worked as a team with the other two reviewers, each of whom reviewed half the studies.) The principal investigator resolved any disagreements that remained unresolved after discussions among the reviewers. Project staff entered data from the forms into an electronic database that was used to track all studies through the screening process.

To be accepted for analysis, studies of drug therapy had to be controlled clinical trials according to the following definitions:

Randomized controlled trial (RCT). A trial in which the participants (or other units) are definitely assigned prospectively to one of two (or more) alternative forms of health care, using a process of random allocation (e.g., random number generation, coin flips).

Controlled clinical trial (CCT). A trial in which participants (or other units) are either

(a) definitely assigned prospectively to one of two (or more) alternative forms of health care using a quasi-random allocation method (e.g., alternation, date of birth, patient identifier)

OR

Note: Appendixes and Evidence Tables cited in this report are provided electronically at http://www.ahrq.gov/clinic/epcindex.htm.

(b) possibly assigned prospectively to one of two (or more) alternative forms of health care using a process of random or quasi-random allocation.

For the analysis of surgical studies, we broadened these inclusion criteria to encompass case series, since our TEP suggested, and a brief scan of the literature confirmed, that RCTs would be few in number. While acknowledging that inferences about efficacy cannot easily be made from case series, we did judge that such studies provided useful information in the absence of RCT data and, furthermore, would be useful to assess complications and adverse events of surgery. To avoid reviewing potentially numerous case reports, we set a threshold of ten or more patients per series for inclusion in our review.

Extraction of Study-Level Variables and Results

We abstracted data from the articles that passed our screening criteria onto a specialized Quality Review Form (See Appendix B). The form contains questions about the study design, the number of patients and comorbidities, dosage, adverse events, the types of outcome measures, and the time from intervention until outcome measurement. With input from the project's TEP, we selected the variables for abstraction. Three reviewers, working in groups of two, extracted data from the same articles and resolved disagreements by consensus. A senior physician resolved any disagreements not resolved by consensus.

To evaluate the quality of the studies, we collected information on the study design, withdrawal/dropout rate, method of random assignment (and blinding), and method for concealment of allocation (the attempt to prevent selection bias by concealing the assignment sequence prior to allocation). We also calculated the percentage of attrition by dividing the number of persons who dropped out of the trial (i.e., the number of people who entered the trial minus the number who completed the trial) by the number of persons entering the trial. The elements of design and execution (randomization, blinding, and withdrawals) have been aggregated into a summary score developed by Jadad.[39] The Jadad score rates studies on a 0 to 5 scale, based on the answer to three questions:

1. Was the study randomized?

2. Was the study described as double-blind?

3. Was there a description of withdrawals and dropouts?

One point is awarded for each "yes" answer, and no points are given for a "no" answer. Additional points are awarded if the randomization method and method of blinding were described and were appropriate. A point is deducted if the method is described but is not appropriate. Empirical evidence has shown that studies scoring 2 or fewer points show larger apparent differences between treatment groups than do studies scoring 3 or more.[40]

Choice of Outcomes

The outcome of interest specified by our sponsor was weight loss. Excess weight is associated with other negative health outcomes, such as diabetes mellitus, hypertension, sleep

Note: Appendixes and Evidence Tables cited in this report are provided electronically at http://www.ahrq.gov/clinic/epcindex.htm.

apnea, and osteoarthritis. A loss of 5 percent to 10 percent of body weight by an obese person, followed by long-term weight maintenance, is associated with improved health outcomes.[16]

Weight loss can be measured in several ways, most commonly kilograms of weight lost, "excess" weight loss, or percentage of excess weight loss. Excess weight is defined as that above a patient's "ideal body weight" based on height and weight tables. Among these, the most commonly reported outcome, by far, is kilograms (or pounds) of weight lost. Among 111 surgical studies reporting weight loss that we reviewed, 43 reported weight loss only in terms of kilograms or pounds, 17 reported only excess weight loss or some variant, 46 reported both of these outcomes, and 5 reported neither of these. Among the 76 pharmaceutical studies that we assessed, none reported excess weight loss. Therefore, we chose weight loss (in kilograms or pounds) as the principal outcome measure, since this choice allowed us to include the maximum number of studies in our analysis and afforded us the only way of comparing the effectiveness of surgical and pharmaceutical therapies across studies.

Because weight loss achieves health benefits primarily by reducing the incidence or severity of weight-related comorbidities like diabetes, we also endeavored to assess treatments by comparing their effects on these outcomes. Very few of the pharmaceutical studies reported these outcomes; thus, it was not feasible to make cross-study comparisons to address the control of comorbidities. We did search for studies that made within-study comparisons of the control of comorbidities, and we summarize their findings in this report. We also assessed the case series reports of obesity surgery for the control of selected comorbidities and compared these results to those reported in studies containing within-group comparisons.

Meta-Analyses of Weight Loss Medications

Of the medications we assessed, three had up-to-date existing meta-analyses (sibutramine, phentermine, and diethylpropion) and three others had a sufficient number of new studies to justify a new meta-analysis (orlistat, topiramate, and fluoxetine). Our meta-analytic methods are the same for the orlistat and fluoxetine weight loss meta-analyses, so we discuss the approach for both medications simultaneously. We conducted all analyses and drew all graphs using the statistical package Stata.[41]

Selection of Trials for Meta-Analysis

The outcome of interest was weight loss between baseline and followup. Based on clinical considerations, we focused on weight loss measured at six months or later. To make our analyses comparable, we stratified the analysis in the same manner as the "in press" meta-analysis on sibutramine.[36] Thus, we defined data collected at "six months" to be data collected at any point between 16 and 24 weeks; likewise, "one year" followup data were collected at any point between 44 and 54 weeks. If a study presented data for two or more time points in an interval, e.g., at 16 and 18 weeks, we chose the longest followup measurement for our analysis.

For some trials, several publications presented the same outcome data. In these cases, we picked the more informative of the duplicates; for example, if one publication was a conference abstract with preliminary data and the second was a full journal article, we chose the latter. The publications dropped for duplicate data do not appear in the evidence table but are noted in the text. We note that multiple citations of the same article were removed at the title screening stage of the project.

For a trial to be included for further analysis, the associated publication(s) had to report on weight loss, provide data prior to the crossover point if the trial was a crossover design, and contain sufficient statistical information for the calculation of a mean difference at six months and/or one year followup as defined below.

Mean Difference

Each trial contained one control or placebo group, which is referred to below simply as the control group. Some trials contained more than one medication group, e.g., at different dose levels. In order not to double-count patients for each trial, we chose the most clinically relevant medication group for our analysis, or in some cases, we combined medication groups. In the discussion below, we assume that a single medication group per study has been determined by choice or defined by combination.

For each trial, we extracted the followup mean weight loss for the control group, the followup mean weight loss for the medication group, and the standard deviation for each group. For studies that included measures for both a six-month and a 12-month followup, we collected those measures separately. If a study did not report a followup mean or if a followup mean could not be calculated from the given data, the study was excluded from analysis. We extracted weight loss as a positive quantity—i.e., greater than zero. For studies that did not report a standard deviation or for which a standard deviation could not be calculated from the given data, we imputed the standard deviation by using those studies and groups that did report a standard deviation and weighting all groups equally.

We converted all means and standard deviations to kilograms. The vast majority of studies measured weight loss in pounds or kilograms. For the few that measured weight loss in terms of BMI, we converted BMI to kilograms by assuming an average height of 5 feet 4 inches. (Past experience has shown that our results—efficacy of weight loss therapies—vary little if any height between 5 feet and 6 feet is chosen to convert BMI to weight.) We then calculated a mean difference for each study, which was the difference between the control group followup mean weight loss and the medication group followup mean weight loss:

mean difference = control followup mean – medication followup mean

We also estimated the standard deviation for that mean difference.[42] A negative mean difference indicates that the medication group experienced more weight loss at followup than did the control group. For example, if the medication group lost 3 kg on average and the control group lost 1 kg, the mean difference is –2 kg. The mean difference is readily interpretable because it is measured in kilograms.

Sensitivity Analyses

We conducted sensitivity analyses on four study dimensions: Jadad quality score (less than or equal to 2 versus greater than or equal to 3); year of publication (1998 or earlier versus 1999 or later); completion rate (less than 80 percent versus 80 percent or greater; and less than 70 percent versus 70 percent or greater); and dosage (for the study[43] of bupropion, the 300 mg arm was used rather than the 400 mg arm; for the fluoxetine analysis, the study[44] with the 20 mg dose was

excluded because all other studies used a 60 mg dose; for the topiramate analysis, the 96 mg study arms were analyzed separately from the 192 mg study arms). We tested for differences between subgroups—e.g., the high-quality versus lower-quality studies—by conducting a meta-regression analysis using a single dichotomous variable to indicate subgroup membership.

Performance of Meta-Analysis

For the 6-month and 12-month analyses, respectively, we estimated a pooled random-effects estimate[45] of the overall mean difference. The individual trial mean differences are weighted by both within-study variation and between-study variation in this synthesis. We also report the chi-squared test of heterogeneity p-value based on Cochran's Q.[46] We constructed a forest plot in which each individual trial mean difference is shown as a box whose area is inversely proportional to the estimated variance of the mean difference in that trial. The trial's confidence interval is shown as a horizontal line through the box. The pooled "weighted mean difference" and its confidence interval are shown as a diamond at the bottom of the plot with a dotted vertical line indicating the pooled estimate value. A vertical solid line at zero indicates no effect of medication on weight loss.

Publication Bias

We assessed the possibility of publication bias by evaluating a funnel plot of the trial mean differences for asymmetry, which can result from the nonpublication of small trials with negative results. These funnel plots include a horizontal line at the pooled fixed-effects estimate and pseudo-95% confidence limits.[47] If bias due to nonpublication exists, the distribution is asymmetric or skewed. Because graphical evaluation can be subjective, we also conducted an adjusted rank correlation test[48] and a regression asymmetry test[47] as formal statistical tests for publication bias. The correlation approach tests whether the correlation between the mean differences and their variances is significant, and the regression approach tests whether the intercept of a regression of the mean differences on their precision differs from zero; that is, both approaches formally test for asymmetry in the funnel plot. We acknowledge that other factors, such as differences in trial quality or true study heterogeneity, could produce asymmetry in funnel plots.

Medication Adverse-Events Meta-Analysis

As we described for the meta-analysis of weight loss, our methods for the adverse-event analysis are the same for all medications.

Extraction of Adverse-Event Data

Each trial included in the weight loss analysis was examined to determine whether it reported data on adverse events. Adverse events were recorded onto a spreadsheet that identified each trial group, the description of the adverse event as listed in the original article, and the number of subjects in each group. We then abstracted either the number of events or the number of people, depending on how the trial chose to report events. The majority of trials recorded the number of

events, rather than the number of unique people who experienced the event. Each event was counted as if it represented a unique individual. Because a single individual might have experienced more than one event, this assumption may have overestimated the number of people having an adverse event. In fact, for some adverse events in a few trials, the number of events reported in a group was greater than the number of individuals in that group. In those cases, we assumed that all individuals in the group experienced the adverse event when we calculated the risk of the event in that group, as described below.

If a report of a trial mentioned a particular type of adverse event in the discussion but did not report data on that adverse event, we did not include that trial in that particular event's analysis. In other words, we did not assume zero events occurred unless the trial report specifically stated that zero events were observed. By taking this approach, we may have overestimated the number of patients for whom a particular adverse event was observed.

After abstracting the data, we identified mutually exclusive subgroups of similar events, based on clinical expertise. For example, one subgroup was "gallbladder problems," consisting of all adverse events concerning this body system. When we subgrouped events, we again treated all observed events as having occurred in unique individuals. For example, we considered bloating, abdominal pain, and dyspepsia as a single subgroup: For a trial that reported abdominal pain events and dyspepsia events separately, we assumed the events that occurred in each category were unique and occurred in different individuals. The number of individuals who were at risk of being affected is the total number of patients in the trial's relevant group (medication or control). Tables 1 and 2 present the categories of adverse events for which we considered a pooled analysis, and the exact descriptions used in the original articles that we aggregated into each category.

For each adverse-event subgroup, we report the number of trials that provided data for any event in the subgroup. We also report the total number of individuals in the medication groups in the relevant trials who were observed to have experienced the event and the total number of patients in the medication groups in those trials. We then report the analogous counts for the control groups in the relevant trials.

Meta-Analysis

For subgroups of events that occurred in two or more trials and occurred at least once in the medication group and at least once in the control group, we performed a meta-analysis to estimate the pooled odds ratio and its associated 95% confidence interval. Given that many of the events were rare, we used exact conditional inference to perform the pooling rather than applying the usual asymptotic methods that assume normality. Asymptotic methods require corrections if zero events are observed; generally, half an event is added to all cells in the outcome-by-treatment (two-by-two) table in order to allow estimation, because these methods are based on assuming continuity. Such corrections can have a major impact on the results when the outcome event is rare. Exact methods do not require such corrections. We conducted the meta-analysis using the statistical software package StatXact.[49]

For interpretability, for any significant pooled odds ratio greater than one (which indicates that the odds of the adverse event being associated with medication is larger than the odds of the event associated with being in the control group), we calculated the RR and number needed to harm (NNH). To perform these calculations, we assume that the expected rate of the adverse

event in an untreated population was equal to the observed crude rate among all control patients (the total number of adverse events observed across all control groups divided by the total number of control patients). The number needed to harm is the number of patients who would have to be treated with medication to produce one adverse event, on average, and is clinically more interpretable than the odds ratio. We calculated the risk ratio by making the same assumption about the expected rate among untreated patients.

Power Calculation

We also conducted a power calculation to determine the lowest adverse-event rate that the included medication trials had at least 80 percent power to detect. First, we assumed a sample size equal to the total number of patients receiving each medication across all trials for which we had sample size data, either reported directly in trials we reviewed or from sample size data reported in other systematic reviews. We then determined the lowest detectable adverse-event rate based on a two-sided test of level 0.05. This calculation was performed to assess the statistical power we actually had available to detect adverse events if few or none were observed. Even if no adverse events were observed, we cannot necessarily conclude that the rate is zero, because the available sample size may have been too small to detect a rare event.

Surgery Analyses

Because the surgery studies included both controlled trials and case series (which do not include controls), we conducted several types of analysis. The way that studies were handled depended on their design and the type of analysis to which they could be subjected. Duplicate publications on surgery studies were treated as described above for the medication publications.

The vast number of types of surgery reported on in the literature required that we aggregate clinically similar surgeries and also identify the key comparisons between different types of surgeries that were of most interest to the clinical audience. Based on discussions with three expert bariatric surgeons, we categorized obesity surgery procedures by (1) procedure type, (2) whether the procedure was performed laparoscopically or open, and (3) more specific surgical details such as length of Roux limb or whether the band was adjustable or nonadjustable. To allow for comparisons in our analysis, we found it necessary to combine certain procedures that were judged clinically similar. Table 3 shows the categories used for the analysis. Vertical banded gastroplasty (VBG) is the only “gastroplasty” procedure that is performed currently. Thus, the other “gastroplasty” procedures were placed into a separate, single category, to be used for historic comparisons as needed.

Analysis of the Efficacy of Surgical Weight Loss

We first summarized the Cochrane review[38] on surgical studies of weight loss. For all surgical studies that we found, we also extracted the mean weight loss and its standard deviation for each study group, generally defined by surgery procedure, at 12 postoperative months and at the maximum followup time greater than or equal to 36 months, as available. We followed the same procedures regarding imputation, conversion to kilograms, and conversion of BMI measurements as we did for the medication meta-analysis.

For randomized controlled trials that reported a within-study comparison of two procedures of interest, a mean difference was calculated (mean weight loss in procedure "1" group minus mean weight loss in procedure "2" group). A positive mean difference indicates that patients in the procedure 1 group lost more weight on average than patients in the procedure 2 group. A negative mean difference indicates that patients in the procedure 1 group lost less weight on average than patients in the procedure 2 group. These mean differences were pooled using a random effects model, and a 95% confidence interval was estimated. For all studies, randomized or not, a pooled mean weight loss for each procedure group was estimated using a random effects model, and an associated 95% confidence interval was constructed.

Analysis of Surgery Mortality

For each group in each study, we recorded the number of deaths observed and the total number of patients in the group. If the study self-identified the deaths as “early” or “postoperative” or if it identified the deaths as within 30 days of the surgery, we termed these “early deaths.” If the deaths were self-identified as “late” or if they were identified as after 30 days, we termed these “late deaths.” If the study was unclear as to the timing of the recorded deaths, we termed these “unclear deaths.” If a study did not report data on death for a group, we recorded zero unclear deaths for that group. That is, we imputed zero for missing data, under the assumption that the authors would have reported a death if there had been one.

For each group of similar procedures (described above in our discussion of the surgery categorization), we conducted analysis for four separate combinations of death definition and type of study: late deaths for RCTs/CCTs; late deaths for CSs; early or unclear deaths for RCTs/CCTs; and early or unclear deaths for CSs. In each setting, we calculated the crude death rate. That is, we divided the total number of deaths observed by the total number of patients in the relevant study groups. This calculation treats all patients from all studies equally and does not take into account any variation across studies in mortality rates, but given the small number of observed deaths, this statistic is simple and easily interpretable.

We were also interested in examining the variation in mortality rates across studies. Thus, for any study with one or more observed deaths, we determined the probability that we would observe as many deaths or more in a sample the size of the study sample, if the population mortality rate is equal to the observed crude death rate. We identified in our results any studies that have a probability less than or equal to 0.05 and, in this way, identified studies with outlying mortality rates compared with the those of the majority of studies. We made this identification by conducting an exact binomial test.

Analysis of Surgery Comorbidity

Three clinicians extracted comorbidity data for all case series studies. Data for diabetes, hypertension, sleep apnea, and lipids were extracted. For each condition, we collected data on the number of people who had the condition at the start of the study. We then collected data on the number of those people whose condition resolved, improved, or was unchanged. A crude proportion was calculated across studies for those who resolved or improved (e.g., the number of people who resolved or improved divided by the number of people with the condition at baseline).

Analysis of Surgery Adverse Events

Each surgery study (RCT/CCT or CS) was examined to determine whether it reported data on adverse events (other than death). The extraction of data for the surgery adverse-event analysis was the same as that described above for the medication trials. After abstracting the data, we identified mutually exclusive subgroups of similar events based on clinical expertise. The actual terms used in the articles that were aggregated into the adverse-event categories are listed in Table 3.

For selected surgery comparisons (one type of surgery versus another type of surgery) for which RCT/CCT data were available, we estimated a pooled odds ratio and its associated 95% confidence interval using exact methods as described above for the medication adverse-events meta-analysis. We also report the crude adverse-event rate for each RCT/CCT surgery group (total number of affected patients divided by total number of patients at risk). In addition, we report the crude adverse-event rate for each surgery group across all studies (RCT/CCT/CS) combined.

Rating the Body of Evidence

We rated the body of evidence on each topic in terms of quality, quantity, and consistency. The quality of the evidence is the aggregate of the quality of the individual studies and is influenced by study design, execution, and analysis of the results. In terms of study design, our minimum criterion for conclusive evidence for the efficacy of a treatment for a chronic disease is that a study have concurrent comparison groups, i.e., conclusive evidence cannot derive solely from case series or studies with historical controls. Empirical evidence shows that studies without a concurrent comparison group report estimates of effect that are inflated when compared with estimates from studies with a concurrent comparison group.[50] Numerous examples of interventions—both medical and surgical—can be cited that reported benefits in case series, only to be shown to be less effective or even ineffective when subjected to a study with a concurrent comparison group.[51,52] Consequently, in order to judge a treatment as having conclusive evidence of effect, we required statistically and clinically significant within-study comparisons of outcomes. In order to have confidence in making cause-and-effect conclusions between the treatment and outcome when assessing a difference in outcome between two groups receiving different treatments, we need to be confident that the groups were sufficiently similar before the treatment. Random assignment of a large number of patients is the best way to obtain similar groups, but that does not mean that conclusive evidence can come only from randomized trials. Rather, when drawing conclusions, we judge the size of the difference in outcome compared to the possibility that there were pretreatment differences between groups that might explain the outcome differences. If a large number of patients are randomly assigned to groups, the two groups are very likely to be similar at baseline, and we are more confident in drawing conclusions about cause-and-effect relationships, even if the difference in outcome is small. However, even when patients are not assigned randomly to groups, we can still draw conclusions about cause-and-effect if the differences in outcome are very large, so large that we judge it unlikely that measured or unmeasured differences in groups could account for the outcome differences.

The quantity of evidence refers to the number of studies, the sample sizes of the studies, and the sizes of the study effects. In the absence of a single very large study, then the number of studies, their sizes, and the sizes of the effects are taken into consideration when assessing the quantity of evidence.

Peer Review

A draft of this report was prepared in October 2003 and sent to the peer reviewers listed in Appendix D. The comments we received from peer review were used to prepare the final report. Service as a reviewer of this report should not in any way be construed as agreeing with or endorsing the contect of the report.

Note: Appendixes and Evidence Tables cited in this report are provided electronically at http://www.ahrq.gov/clinic/epcindex.htm.