Chapter 2. Methods

Original Proposed Key Questions

The topic of this report was nominated by the National Institutes of Health (NIH) Office of Dietary Supplements (ODS). The following questions were originally proposed:

Weight Loss

1. *What is the evidence for efficacy of ephedra-containing dietary supplement products for weight loss, over a sustained period of time?*
2. *Can efficacy for weight loss be attributed to ephedra alone, or ephedra in combination with other ingredients (e.g., caffeine)?*
3. *Does ephedra have additive effects with other agents?*
4. *What dosage levels of ephedra are necessary to achieve weight loss?*

Athletic Performance

1. *What is the evidence for efficacy of ephedra-containing dietary supplement products in terms of energy enhancement and enhancement of athletic performance, over a sustained period of time?*
2. *Can efficacy for energy enhancement and enhancement of athletic performance be attributed to ephedra alone, or ephedra in combination with other ingredients (e.g., caffeine) that produces energy enhancement and/or enhancement of athletic performance?*
3. *Does ephedra have additive effects with other agents?*
4. *What dosage levels of ephedra are necessary to achieve energy enhancement and enhancement of athletic performance?*

Safety Assessment

1. *Does use of ephedra-containing dietary supplement products over a sustained period of time increase the risk of cardiovascular disease (CVD) or other serious and life-threatening events in specific populations?*
2. *What populations are at risk of CVD and other life-threatening events through use of ephedra over a sustained period of time?*
3. *Can the risk for adverse events in these populations be attributed to ephedra alone, or in combination with other ingredients (e.g., caffeine)?*
4. *Does ephedra have additive effects with other agents?*
5. *What dosage levels of ephedra produce risk of CVD or other life-threatening events?*
6. *Do ephedra-containing dietary supplement products alter physiologic markers of cardiovascular function?*
7. *What are the metabolic actions of ephedra, so as to explain its beneficial and adverse effects?*

In addition to the questions related to ephedra-containing dietary supplement products, the sponsor also requested a review of the scientific literature on ephedrine (the purified alkaloid) regarding its efficacy and safety. A brief review of the mechanism of action of ephedra was also requested.

We were also asked about the gaps in knowledge about the effects of ephedra, alone or in combination with other agents, on weight loss, energy enhancement, and enhancement of athletic performance. We were asked to focus on the following categories of potential consumers: children, adolescents, young athletes (male and female), and adults (male and female).

Technical Expert Panel

Each AHRQ evidence report is guided by a Technical Expert Panel (TEP). We invited a distinguished group of basic scientists and clinicians, including individuals with expertise in cardiac electrophysiology, exercise, herbs, obesity and human nutrition, pharmacognosy (the study of developing drugs from plant and animal sources), pharmacology, and toxicology. Panel members are listed in Table 2.

Our expert panel meeting was held at RAND's Arlington, Virginia, office on Wednesday, November 28, 2001. Margaret Coopey, the Task Order Officer, represented AHRQ. Dr. Paul Coates, head of the NIH ODS, also attended. At the meeting, we discussed the focus of the report. The TEP agreed that we should review articles that discuss either ephedra or ephedrine. Studies or case reports on pseudoephedrine were not to be reviewed, except in the context of ephedra/ephedrine. We agreed to include a brief description of the other alkaloids (pseudoephedrine, norephedrine, etc.) in the introduction to our report.

The TEP also provided a number of suggestions regarding data collection. These suggestions are shown in Table 3.

Assessment of Adverse Events

With regard to adverse events, EPC staff and the TEP recognized that, even in aggregate, the number of patients included in randomized trials was likely to be too few to allow adequate statistical power to assess the rate of serious adverse events (such as death, myocardial infarction, stroke, or seizure) due to ephedra. Because of this likelihood, the EPC staff recognized the necessity of relying on case reports to help inform the sponsor regarding the key questions concerning serious adverse events. A long discussion occurred at the TEP meeting about criteria for assessing causality based on case reports. The framework for this discussion was based on an unpublished article by Cynthia Mulrow, MD (C. Mulrow, personal communication). This paper summarized the criteria used in all of the major published algorithms for establishing different levels of causality in case reports of adverse events from drugs (see Table 4). Our TEP judged that, to establish definite causality from case reports, a "de-challenge/re-challenge" test needed to be performed (that is, it had to be documented that the adverse event in question went away when the offending drug was withdrawn and reoccurred when the offending drug was reinstated). Clearly, such a de-challenge/re-challenge was not possible or feasible in the case of serious adverse events such as death or myocardial infarction.

Consequently, our TEP judged that case reports alone would be insufficient to establish definite causality between ephedra use and serious adverse events. The TEP discussed the key characteristics of a case report that would signal the need for additional study. Such characteristics would include the following:

- Documentation (preferably medical) that the adverse event occurred.

- Documentation that the patient took ephedra and that the dose and timing were consistent with the known pharmacology of ephedrine (for cases of death, myocardial infarction, stroke, or seizure). (The TEP later quantified this characteristic for acute events such as stroke or myocardial infarction to mean a dose preferably within six hours of the adverse event and in no cases greater than 24 hours before the adverse event.)

- Performance of an evaluation sufficient to rule out other potential causes for the adverse event.

The TEP and EPC staff discussed extensively the types of information necessary to satisfy this last criterion. The TEP agreed that the absence of data could not be construed as a negative result. For example, the absence of information about prior cardiac disease could not be construed as an absence of cardiac disease. Furthermore, the TEP emphasized that verbal histories indicating no prior history of serious conditions were not sufficient to rule out alternative explanations for the most serious adverse events, since unrecognized preexisting cardiac disease, congenital abnormalities, berry aneurysms in the cerebral circulation, and other such conditions occur with some frequency and are known to cause death, myocardial infarction, or stroke without warning in otherwise "healthy" individuals. Realizing that it would be very difficult to attempt to define all of the possible evaluations and interpretation of results in the abstract, the TEP left it to EPC staff to resolve these issues, guided by the three characteristics listed above.

Literature Search

Our search for controlled human studies of the effects of ephedra and ephedrine began with an electronic search of library databases in April 2001. Tables 5 and 6 show our specific search strategies. We started with Medline, which is maintained by the U.S. National Library of Medicine and 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 (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. We also searched EMBASE, the Excerpta Medica database produced by Elsevier Science, which is a major biomedical and pharmaceutical database indexing over 3,800 international journals. EMBASE currently contains over six million records, with more than 400,000 citations and abstracts added annually. We also searched BIOSIS, the most complete database for the life sciences; the Allied & Complementary Medicine Database (AMED); the Manual Alternative and Natural Therapy Index System

(MANTIS), which is the largest index of peer-reviewed articles in the area of complementary and alternative forms of therapy; and the Cochrane Controlled Clinical Trials Register Database. AMED is produced by the Health Care Information Service library in the United Kingdom. It covers journals in allied health professions as well as complementary and alternative medicine. Similarly, MANTIS covers manual, alternative, and natural therapy. 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. The Cochrane Register of Controlled Trials is available on CD-ROM by subscription.

Our TEP then suggested that we search three additional databases: the International Pharmaceutical Abstracts; Pascal (produced by the Institut de l'Information Scientifique et Technique (INIST) of the French National Research Council (CNRS), whose subject areas include physics, chemistry, biology, medicine, psychology, applied sciences, technology, earth sciences, and information sciences); and SciSearch. SciSearch contains all records published in Science Citation Index and additional records from about 1,000 journals whose table of contents pages are listed and indexed in the weekly Current Contents publications. Every subject area within the broad fields of science, technology, and biomedicine is included. Mary Hardy, MD, and Margaret Maglione, MPP, reviewed a total of 1,780 retrieved titles. Of those, 452 articles were deemed relevant to our undertaking and were ordered. Thirty-four additional articles were found through mining reference lists, and 64 were contributed by the TEP or AHRQ. We reviewed the reference list of every retrieved article for additional literature we might have missed and ordered any we found. Literature was tracked using ProCite and Access software.

Additional Sources of Evidence

We obtained the report "Safety Assessment and Determination of a Tolerable Upper Limit for Ephedra," published in December 2000 by CANTOX Health Sciences and funded by the Council for Responsible Nutrition, an association of dietary supplement manufacturers. We ordered copies of all literature cited in this report. We also obtained transcripts of a public meeting, held in Washington, DC on August 8 and 9, 2000 and sponsored by the HHS Office on Women's Health, on the safety of dietary supplements containing ephedrine alkaloids. We contracted with physicians proficient in Japanese and Chinese to search for scientific literature in their native languages. These searches identified little on the use of ephedra for weight loss and exercise enhancement because ephedra is not used in that manner in Eastern cultures. In addition, we found nothing about ephedra on Phytonet, a European database. We also contacted Baptist University, Hong Kong, which has a database on herbal medicine, as well as the Taiwan Poison Control Center, but did not receive any data from either.

On January 31, 2002, we spoke to Dr. Phillip Waddington, Director of the Natural Health Products Directorate for Health Canada. He agreed to send us 60 adverse event reports regarding ephedra/ephedrine products. However, at the time of this report, we had not received anything.

In January 2002, we created an announcement regarding our project's need for any unpublished studies on the use of ephedra/ephedrine for weight loss or exercise enhancement. The announcement was submitted to both the journal *Phytomedicine* and the *Herbalgram* newsletter. The intent was to reach individuals who might know of small studies being done on

ephedra or ephedrine of which the TEP were not aware. We receive no responses to this announcement.

In March 2002, we obtained a recent monograph on ephedra, written by Dennis McKenna, from the Institute for Natural Products Research, a nonprofit research and education foundation.

Finally, Wes Seigner, an attorney for the Ephedra Education Council in Washington, D.C., agreed to send us unpublished industry studies. We developed a confidentiality agreement, and Mr. Seigner sent us several reports on then-unpublished controlled trials conducted by members of the council.

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 yes/no questions was created to track the articles (Figure 1). After being evaluated against this checklist, each article was either accepted for further review or rejected. Two physicians and a policy analyst, each trained in the critical analysis of scientific literature, independently reviewed each study, abstracted data, and resolved disagreements by consensus. The principal investigator resolved any disagreements that remained unresolved after discussions among the reviewers. Project staff entered data from the checklists into an electronic database that was used to track all studies through the screening process.

To be accepted for analysis, studies 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
> (b) Possibly assigned prospectively to one of two (or more) alternative forms of health care using a process of random or quasi-random allocation.

Extraction of Study-Level Variables and Results

We abstracted data from the articles that passed our screening criteria onto a specialized Quality Review Form (QRF—see Figure 2). 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. We selected the variables for abstraction with input from the project's TEP. Two physicians, working independently, each

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.[82] The Jadad score rates studies on a 0 to 5 scale, based on the answer to three questions:

- Was the study randomized?
- Was the study described as double-blind?
- 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 less show larger apparent differences between treatment groups than do studies scoring 3 or more.[83]

Meta-Analysis

Selection of Trials for Meta-Analysis

In selecting trials for the meta-analysis of weight loss, we considered all weight loss trials that included a treatment duration of at least eight weeks. Our TEP suggested that shorter treatment durations were insufficient to assess long-term weight loss. Trials on athletic performance encompassed a wide variety of interventions. Because of this heterogeneity, we compared and contrasted athletic performance studies in a narrative review and did not perform a meta-analysis. This section focuses on methods used for the meta-analysis of the weight-loss trials.

Trial Inclusion

The available weight loss trials were judged to be sufficiently clinically homogeneous to support a pooled analysis. For some trials, several publications presented the same outcome data. In these cases, we picked the most 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 of Chapter 3, Results. We note that multiple citations of the same article were removed at the title screening stage of the project.

Based on input from our TEP, we chose weight loss as the most clinically relevant outcome for the included trials. In order for a trial to be included in the analysis, the associated publication

had to report on weight loss as an outcome, provide data prior to the crossover point if the trial was a crossover design, and contain sufficient statistical information for the calculation of an effect size. We calculated an effect size for every comparison of interest, e.g., ephedra versus placebo, at each relevant follow-up time-point, as described below. The effect size is calculated by dividing the difference between the weight loss in the treatment group and the weight loss in the placebo group by its standard deviation. The effect size is a unitless measure that is useful when comparing trials assessing outcomes that are similar (such as weight loss) but are measured in different ways (pounds versus body mass index). We synthesized effect sizes within comparison and follow-up subgroups. The percentage of weight lost, compared to pretreatment weight, is another clinically relevant outcome. However, we did not choose this outcome for our primary analysis for two reasons. First, pooling percentage of weight loss within a treatment group (e.g., an ephedra group) eliminates the placebo comparison from the trial and therefore does not make use of the strength of the randomized controlled design. Comparison of the treatment group to the placebo group within a trial utilizes the full strength of a randomized controlled trial, as patients who are similar in all aspects except treatment assignment are compared to each other. Thus, if one wanted to perform an analysis of weight loss percentage, we would advise pooling the difference in weight loss percentage between the treatment and placebo groups. The second, and more important, reason for not performing an analysis of weight loss percentage, regardless of whether the internal placebo comparison is made, is lack of data. The vast majority of trials did not report percentage of weight loss as an outcome. As a result, we would have had to make two assumptions in our calculations. First, to estimate mean percentage weight loss for a group in a trial, we took the ratio of mean weight loss between baseline and follow-up divided by mean baseline weight. The mean of a set of ratios does not equal the ratio of the means, but this would have been the best estimate we could obtain. Second, to estimate the standard deviation of our ratio, we would have had to use the delta method to approximate the standard deviation and furthermore would have had to estimate the correlation between the baseline and follow-up weights to be 0.5. We are unable to check either of these assumptions. In contrast, the vast majority of trials did report weight loss as an outcome, and also presented the standard deviation of this statistic. Hence, weight loss became our primary outcome for analysis.

Stratification of Interventions

The literature included 6 different types of comparisons: (1) ephedrine versus placebo; (2) ephedrine plus caffeine versus placebo; (3) ephedrine plus caffeine versus ephedrine; (4) ephedrine versus other active treatment; (5) ephedra versus placebo; and (6) ephedra plus herbs containing caffeine versus placebo. Only one trial compared the effect of ephedra alone versus placebo. If a trial had other treatment arms such as caffeine only, we dropped those arms from our analysis. Effect sizes were pooled separately within each comparison subgroup. In addition, a cross-subgroup synthesis using meta-regression was conducted on the ephedrine versus placebo; ephedrine plus caffeine versus placebo; and ephedra plus herbs containing caffeine versus placebo effect sizes as well as a direct within-study comparison for those few studies that presented data for more than one comparison, as described below.

Weight Loss Effect Size

For each trial, we calculated effect sizes for any of the six comparisons of interest for which the study provided data. The majority of trials included only one comparison—between a single treatment (e.g., ephedrine) arm and placebo. One trial[84] included both an ephedrine plus caffeine plus aspirin arm and an ephedrine plus caffeine arm. However, we combined these arms into a single ephedrine plus caffeine arm, based on the clinical reasoning that aspirin has relatively little effect on weight loss.

Nevertheless, a small number of trials contained more than one relevant comparison between arms and thus contributed more than one effect size to be considered for analysis. Double-counting patients is a concern if a trial contributed more than one effect size to an analysis, and patients were included more than once in calculating those effect sizes. For example, if a trial had one placebo arm, an ephedrine arm, and an ephedrine plus caffeine arm, it contributed two effect sizes, both based on the same placebo patients. Fortunately we encountered relatively few instances of double-counting of patients within the analyses. One trial[85] included two ephedrine doses and a placebo arm and thus contributed two ephedrine versus placebo effect sizes, that is, two effect sizes within a single comparison group.

Four trials[84, 86-88] contributed effect sizes in more than one of the six comparison groups. Since we conducted the comparison group analyses separately, the four latter trials do not double-count patients within comparison group analysis. We discuss the possible influence of multiple effect sizes per study on the meta-regression analysis below.

For each trial, we extracted the means and standard deviations of weight loss between baseline and the relevant follow-up times for each arm, if available. For example, if a trial with placebo and ephedra arms reported follow-up data at two months, we extracted the means and standard deviations of weight loss at two months for the ephedra and placebo arms. If trials did not report a weight loss mean for any arm, or this mean could not be calculated from the given data, the trial was excluded from the meta-analysis.

We initially considered four separate treatment duration measurement times: two months, three months, four months, and six months. However, only one ephedra trial[89] and two ephedrine plus caffeine trials[89, 90] reported an outcome measure for a treatment duration of six months. These numbers are too small to perform a separate pooled analysis on six-month outcomes. Thus, we considered three treatment durations: The two-month duration of treatment included only outcomes for 8 weeks of treatment. However, for the analysis of three-month treatment durations, we included data collected anytime between 12 and 15 weeks, and for the four-month analysis, we included data collected between 18 and 24 weeks. We also analyzed the rate of monthly weight loss, as described below.

The large majority of included trials reported weight loss in kilograms; some trials reported weight loss in pounds. Since an effect size is unitless, data expressed in either unit of measure could be extracted for analysis. One trial[91] reported weight loss only in terms of body mass index (BMI). Because this measure involves both height and weight, we first transformed the study data to kilograms by assuming an average height of 68 inches (within a range of reasonable values, the height that was chosen made little difference in the results).

As mentioned above, for each arm in each included trial, we also calculated the mean monthly weight loss by dividing by the number of months of treatment. Thus, using our previous example, we calculated the mean monthly weight loss for the placebo and ephedra arms respectively by dividing the associated two-month mean weight loss by two. For those trials that had more than one treatment duration time, we used the longest treatment duration time data to calculate the monthly weight loss. We extracted both weight loss at specific time points (e.g., two, three, and four months) and monthly weight loss to compare the results for both types of outcomes. This comparison allows us to check trends in weight loss, for example, whether weight loss is linear or dampens over time. Using meta-regression, we verified that weight loss was linear over the range of time for which data were available by comparing pooled monthly weight loss rates based on the two-month, three-month, and four-month data separately in each comparison group. Thus, our primary analysis focuses on monthly weight loss. We note that the included trials had relatively short-term follow-up; thus, our results address only short-term weight loss and should not be extrapolated beyond four months.

If a trial reported a standard deviation of weight loss at a relevant follow-up time, we extracted those data and used them to calculate the standard deviation of the monthly weight loss. Eight trials[84, 87, 88, 92-96] failed to report a standard deviation for weight loss at a given follow-up time, or a standard deviation could not be calculated from the given data. For these trials, we imputed the standard deviation of the monthly weight loss by using those trials and arms that did report a standard deviation. We averaged the monthly weight loss standard deviations by weighting all arms equally in the imputed value calculation. For those trials missing standard deviations, we then used the imputed monthly weight loss standard deviation to calculate the standard deviation for weight loss at the relevant follow-up time.

For each pair of arms, an unbiased estimate[97] of Hedges' g effect size[98] and a 95 percent confidence interval were calculated. A negative effect size indicates that the treatment arm (ephedrine or ephedrine plus caffeine, or ephedra plus herbs containing caffeine) is associated with a larger weight loss at follow-up (or a larger monthly weight loss) than is the comparison arm, e.g., the placebo.

Performance of Meta-Analysis

We estimated a pooled random-effects estimate[99] by combining effect sizes for comparison subgroups that contained three or more effect sizes. We also report the chi-squared test of heterogeneity p-value.[97]

Forest plots were constructed for each comparison subgroup. Each individual trial effect size is shown with confidence intervals as a box whose area is inversely proportional to the estimated variance of the effect in that trial. The pooled estimate 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 treatment effect.

For each trial, we calculated the monthly weight loss percentage for each treatment group and the placebo group. Monthly weight loss percentage is defined as the mean monthly weight loss divided by the mean baseline weight in that group. Unfortunately, we were not able to calculate monthly weight loss percentage on an individual level. To determine the standard

deviation of the monthly weight loss percentage, we used the delta method[100] and assumed a correlation of 0.5 between the baseline and follow-up weights.[101] For each comparison subgroup, we pooled monthly weight loss percentages in the treatment groups and placebo groups separately using a random effects model[99] and produced associated 95% confidence intervals. We acknowledge that combining estimates within treatment groups only, or placebo groups only, does not take advantage of the randomization and pairing of treatment and control within a trial. This lack of pairing, and the fact that the monthly weight loss percentage in the treatment group must be compared to the associated monthly weight loss percentage in the placebo group, should be kept in mind when interpreting the results of this analysis.

Sensitivity Analyses

When relevant, we conducted sensitivity analyses on subgroups of trials to determine the robustness of our conclusions. In order to assess the possible impact of attrition, we divided the trials into two groups: (1) those with less than 20 percent attrition in all arms and (2) all others. Twenty percent attrition is a commonly accepted threshold above which concerns about bias increase, due to loss to follow-up. For trials in which attrition was unknown, we assumed it was not less than 20 percent. We conducted the main analyses for the two attrition strata separately.

We also conducted further analyses on the attrition rates. To determine whether the attrition rate varied between treatment and placebo groups within a trial, we first collapsed all the treatment groups together within a trial and estimated a single attrition rate for treatment. We then conducted a paired t-test that assessed whether the difference between the treatment and placebo attrition rates within a trial was significantly different from zero. All studies were weighted equally in this analysis. We also categorized each trial as significant or not significant based on its effect size. Trials that had more than one effect size agreed in terms of significance (in other words, the trial reported consistent result with respect to significance at multiple time points). We then categorized each trial as to whether the attrition rate for the treatment group was higher than, lower than, or the same as that of the placebo group. We examined the bivariate distribution of studies into these six categories, (three relationships between group attrition rates categories, and whether each of these relationships was significant or nonsignificant), and conducted a chi-squared test of the association between significance and the relationship between group attrition rates.

When relevant, we also performed our calculations a second time, excluding the trial by Moheb and colleagues.[84] This trial was presented only in abstract form and provided only the total sample size, not the sample sizes for each arm; thus, we had to assumed equal sample sizes across arms.

For the ephedrine plus caffeine versus placebo trials, we performed two sensitivity analyses. In the first, we dropped one trial[102] that had synephrine in the ephedrine plus caffeine arm. In the second, we dropped one arm of one trial[103] in which aspirin was combined with ephedrine plus caffeine; the sensitivity analysis was performed with the ephedrine plus caffeine arm alone.

A final sensitivity analysis concerned the choice of summary statistics to pool. Instead of pooling effect sizes or "standardized mean differences," we applied a "weighted mean difference" approach. In the latter, we pooled the absolute differences in weight loss between the

treatment and placebo groups, inversely weighted by the trial variances of the differences. That is, we did not first divide the differences by their standard deviations to produce effect sizes and then weight by the inverse variances of the effect sizes. If the variances are not homogeneous and/or the variances are not well estimated, these two methods may not produce the same results. The weighted mean difference approach has the appeal of being conducted entirely in the clinical units of interest—in this case, pounds.

Analysis of Dose

We tested for a dose effect using a random-effects meta-regression model.[104] A separate model was fitted within each comparison subgroup. We defined a low dose of ephedrine as 10–20 mg; a medium dose of ephedrine as 40–90 mg; and a high dose of ephedrine as 100–150 mg. We characterized each dose level as an indicator variable in a main-effects model and chose the medium-dose group as the level to exclude. The meta-regression approach allowed us to test directly the efficacy of low and high doses versus the excluded medium dose group, as well as to estimate the effect size for each dose level.

Publication Bias

We assessed the possibility of publication bias by evaluating a funnel plot of effect sizes for asymmetry, which can result from the nonpublication of small trials with negative results. These funnel plots include a horizontal line at the fixed-effects pooled estimate and pseudo–95% confidence limits.[105] 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[105] and a regression asymmetry test[106] as formal statistical tests for publication bias. The correlation approach tests whether the correlation between the effect sizes and their variances is significant, and the regression approach tests whether the intercept of a regression of the effects sizes on their precision differs from zero; that is, both 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.

Meta-Regression

As described above, in order to compare monthly weight loss effect sizes across comparisons, we conducted a random-effects meta-regression.[104] The observations in this meta-regression were all monthly weight loss effect sizes across the ephedrine, ephedrine plus caffeine, and ephedra plus caffeine-containing herbs comparisons. The variables are indicator flags, one for each comparison. Only one trial[85] had multiple effect sizes in the regression, and we did not account for the correlation between these two effect sizes in our model.

Three trials[84, 86, 88] contained both ephedrine and ephedrine plus caffeine arms. For these trials, we were able to conduct a direct, or "head-to-head," comparison of these treatments by pooling the effect sizes for each trial together. In the estimation of an effect size in this situation, the comparison group is that group of individuals who received ephedrine alone. Thus, a negative effect size means that ephedrine plus caffeine is associated with a larger monthly weight loss than is ephedrine alone. This direct comparison is more robust than the cross-group meta-regression described above, because the former compares groups only within a trial. However, due to the small number of trials that provided more than one treatment arm and the lack of any

direct comparisons of ephedrine alone or ephedrine plus caffeine versus ephedra, we conducted both analyses.

Interpretation of the Results

To aid in interpreting our results, we back-transformed all pooled estimates to weight loss in pounds. In order to do this, we multiplied each pooled estimate by the average standard deviation across trials, and then further multiplied by 2.2 to transform kilograms to pounds. In this way, we were able to equate our unitless pooled effect size with weight loss in pounds. However, we note this back-transformation requires assuming a particular underlying standard deviation. Readers may wish to apply their own standard deviation, based on the particular patient population to which they wish to apply the results.

We conducted all analyses and drew all graphs using the statistical package Stata.[107]

Safety Assessment

Controlled Trial Adverse Events

Data Collection

Each trial that we identified was examined to determine whether it reported data on adverse events. Adverse events were recorded onto a spreadsheet that identified each study arm, the description of the adverse event as listed in the original article, the number of adverse events in each category, and the number of subjects in each arm.

Meta-Analysis

The strongest level of evidence for attributing an adverse event to an exposure comes from placebo-controlled randomized trials of the exposure in question. In this evidence report, such evidence would come from placebo-controlled trials of ephedra or ephedrine. We therefore searched all such trials that we identified and extracted from each trial the adverse events that were reported associated with it, as described above. Because each event was counted as if it represented a unique individual, and because a single individual might have experienced more than one adverse event, this method may have overestimated the number of people having an adverse event. We then compared event rates in the people who received ephedra or ephedrine with those in people who received placebo. We performed a meta-analysis on those adverse events for which there was an appreciable number of reports in the randomized trials.

We collected data on adverse events for the randomized controlled trials. For each adverse event, e.g., vomiting, and for each treatment group and for the placebo group, we 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. We treated all events as if they occurred in unique individuals, which, as we stated, may overestimate the number of people apparently affected in a particular event category.

We note that some trials recorded zero events in a particular event category, and these data were thus recorded. However, some trials recorded no data for a certain event category or

recorded no adverse events at all. These trials did not enter the adverse event meta-analysis, in that we did not assume zero observed events if a trial did not mention a particular type of event. By excluding these trials, we may have underestimated the number of patients for whom a particular adverse event was *not* observed. We note that, for the power calculation (described below) for serious adverse events (deaths, myocardial infarctions, strokes, seizures, and serious psychiatric symptoms), the sample sizes of all trials were taken into account, regardless of whether they mentioned these serious events. We assumed that such serious events would have been recorded had they been observed, so that a record of zero or no mention of a serious event could both be taken to mean that no such events were observed.

After abstracting the data, we identified mutually exclusive subgroups of similar events, based on clinical expertise. When we subgrouped events, we again treated all observed events as having occurred in unique individuals. For example, we considered nausea and vomiting as a single subgroup. For a trial that reported nausea events and vomiting 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 (placebo or treatment).

For each 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 placebo groups in the relevant trials who were observed to have experienced the event (calculated as described above) and the total number of patients in the placebo groups in those trials. We then report the analogous counts for all applicable treatment groups (ephedrine, ephedrine plus caffeine, ephedra) in the relevant trials. We specifically do not provide crude placebo and treatment rates (total number of affected patients divided by total number of patients at risk). Such crude rates do not weight trials appropriately.

Based on clinical importance and the availability of data, we chose a limited number of event subgroups for meta-analysis. For each chosen event subgroup, we estimated the pooled odds ratio across the trials that reported on any events in the subgroup, as well as a 95% confidence interval for the pooled odds ratio. Given that many of the events were rare, we utilized 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, since these methods are based on assuming underlying 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.[108]

We also conducted a power calculation to determine the lowest adverse-event rate that the clinical trials we identified had at least 80 percent power to detect. That is, we assumed a sample size equal to all the trials combined, and assuming a two-sided test of level 0.05, we determined the lowest detectable adverse-event rate. 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 are observed, we cannot necessarily conclude that the rate is zero, because the sample size available may have been too small to detect a rare event.

Case Report Adverse Events

Data Collection

Because the clinical trial data had low statistical power to detect a rate of serious adverse events, we therefore assessed case reports of adverse events associated with ephedrine or ephedra-containing dietary supplement use in order to inform the sponsors regarding the Safety Assessment key questions concerning serious adverse events. We reviewed case reports from three sources: the FDA MedWatch file, published case reports, and a file kept by the ephedra supplement manufacturer, Metabolife. Published case reports were identified through our literature search process previously described.

FDA Medwatch Data

In September 2001, the FDA's Office of Nutritional Products, Labeling, and Dietary Supplements produced an Excel spreadsheet with a master list of adverse-event report case numbers and summary information, in response to our request for all herbal ephedra–related adverse-event reports from their database. After several discussions and several months of work, the dataset construction algorithm was reproduced and limited to only herbal ephedra–related adverse-events reports, because some ephedrine adverse events had mistakenly been entered into the initial Excel file. We also received several sets of compact disks containing portable document format (PDF) files of events reported in the FDA Adverse Reaction Monitoring System (ARMS) for the dates specified. Documents retrieved included MedWatch Reports (FDA form 3500); Consumer Complaint Injury Reports (FDA Form 2516); Complaint/Injury Follow-up Forms (FDA Form 2516a); Adverse Reaction Questionnaires (Form A); letters from family members, health care professionals, or lawyers; affidavits collected from witnesses during FDA-held investigations; police reports; medical records, including physician notes (both inpatient and outpatient), emergency department reports, nurse notes, and laboratory reports; product labeling and related information; and product analysis results.

The second master list of only ephedra-related adverse-event reports was created at the product level, so that adverse-event report identification numbers (IDs) were repeated if multiple products appeared in one report. Because our analysis was at the adverse-event report level, where there were multiple products per single ID, we joined those into one record. We established a cutoff date of September 30, 2001, the production date for the CDs that contained actual reports. Our analysis does not include case reports filed after that cutoff date, since these files had not been redacted of identifying information.

The data were analyzed in a series of steps. First, we coded each unique report according to type of adverse event listed in the summary information on the Excel spreadsheet. The categories into which we grouped the reports are listed in Table 7. Then, we separated those reports with events coded as most serious (death, stroke, myocardial infarction (MI), seizure, and certain psychiatric symptoms) from those considered moderately serious. Reports that contained events considered most serious were analyzed using specialized data-collection instruments called Adverse Events Analysis Forms (AEA Forms—see Figures 3a–3c). We developed these instruments to collect information from the corresponding PDF file or published case report on whether the report was actually on ephedra or ephedrine, whether the data were adequate to analyze the report, and whether or not the adverse event qualified as a "sentinel event" (see

below). When a case report dealt with more than one individual, an AEA form was completed for each individual.

To understand the other potentially serious adverse events, we reviewed all case reports that had been grouped into the categories of "other serious cardiovascular," "other serious neurological," and "psychiatric" in our initial review of the master Excel file. For this review, we used a brief data abstraction form (Figure 4). This brief form was developed to assess the evidence supporting the prior use of ephedra and to define the adverse event more precisely. Again, when a case report contained more than one subject, a brief form was completed for each subject. Then, the data collected in the brief review were used to justify including certain more-serious events into the more-detailed review described above. These more-serious adverse events included ventricular tachycardia/fibrillation, cardiac arrest, pulmonary arrest, transient ischemic attack, and brain hemorrhage. Select adverse-event reports were then reviewed a second or third time by project staff physicians to reach an implicit judgment about whether an adequate investigation of other potential causes had been performed. Internists performed the initial reviews of cases of death, myocardial infarction, stroke, and seizure, and were assisted (as appropriate) by a cardiologist, rheumatologist, or neurologist. Psychiatric events were reviewed by two experienced professionals: a psychiatrist specializing in addictions and a psychologist who leads RAND's Drug Policy Research Center. All cases were reviewed by two individuals, with differences resolved by consensus.

As part of additional work we were requested to perform, we received hard copies of MedWatch data on ephedrine, organized in the same manner as the data on ephedra. We first reviewed all these events with our short form to identify the serious adverse events. These events were then reviewed using the same methods we developed for the ephedra database. Two types of adverse events associated with ephedrine were not associated with ephedra. The first involved the intravenous use of ephedrine given during surgery; several such reports were filed by medical personnel. The second involved attempted suicide. We note these two types of case reports in our analysis.

Literature Cases

During our literature search, we identified published case reports of adverse events associated with ephedra use. These published case reports were then reviewed using the same criteria used for the MedWatch events.

Metabolife File

We received the following materials from the Food and Drug Administration (FDA), which had, in turn, received them from Metabolife:

- A CD-ROM labeled "MIPER" (described in more detail below).
- Photocopies of medical information pertaining to 43 cases (also described in more detail below).

- A two-page *Listing of Key Complaint for the Metabolife Medical Records Submitted*, which is a listing of the key complaints for 46 cases, with photocopied medical information. (Note: we received medical information for only 43 cases.)

- A two-page sheet entitled *Index of Redacted Consumer Medical Records with Corresponding MIPER Numbers,* which contained a listing of the 46 cases with additional medical record information and the file numbers for related information on the MIPER CD-ROM.

- Three reviews of the Metabolife adverse-event file, which Metabolife commissioned. Note that to prevent their assessment from biasing our own, we did not read any of these reviews prior to our assessment, but did review them briefly when our assessment was completed.

- A file entitled *77 'serious' AE's as identified by Metabolife,* which contains photocopies of reports of events that were selected by Metabolife as being the most serious in nature. Most, but not all, of these reports were contained on the MIPER CD-ROM. Again, in order to avoid bias, we did not examine this file until after our initial assessment was performed.

- Several journal articles, all of which were already in our possession.

Later, we also received a report entitled *Adverse Event Reports from Metabolife* that had been prepared for Sen. Richard J. Durbin, Rep. Henry A. Waxman, and Rep. Susan A. Davis by the Minority Staff Report Special Investigations Division, Committee on Government Reform, U.S. House of Representatives, and which consisted of an analysis of the MIPER CD-ROM.

The MIPER CD-ROM contains several thousand files of adverse event reports organized in 20 folders. The adverse-event files are numbered from 15111 to 35069, and are continuous, except for three gaps—between 21121 and 22035; 25535 and 27472; and 30627 and 35047. Each file is a TIF picture file, generally of a single sheet of paper, on which is recorded information regarding the potential adverse event or events. This information was recorded in many different ways, including an email record of a telephone conversation between a company representative and the consumer; typed or handwritten letters from the consumer to the company; handwritten notes of telephone conversations with consumers, written on either a rudimentary form or on whatever piece of paper seems to have been handy at the moment; and a form developed by Metabolife for systematically collecting information about possible adverse events. Examples of all of these types of files are presented in Figure 5. Personal identifiers had been redacted from the files we received.

Each consumer could experience one or more adverse events. We referred to a particular adverse event for a person as a "case," and our analysis was conducted at the case level, rather than at the person level. Thus, a person could contribute more than one case to our analysis. We use this terminology throughout the remainder of the report. Practically speaking, in most instances of serious adverse events such as death, heart attack, or stroke, a person contributed only a single case in this manner.

In general, each file on the MIPER CD-ROM contained only a single sheet of paper. We did identify some files that were exactly the same as other files, and we excluded these files from our analysis. The information on a case might reside in a single file or in more than one file. For example, if a letter from a consumer concerning one of the adverse events experienced by that consumer was three pages long, each page resided in a separate file. If possible, we tried to identify all files that pertained to the same case (which we called "duplicate" files), so that we would not count a case more than once. Whether we identified all such instances is unknown, since information in each file was insufficient to allow us to check for duplicate files by matching on key variables such as age, gender, and the type of adverse event. No other mechanism for checking was possible within the time and resources available to the project. Therefore, while we did our best to identify and exclude or in some other way resolve duplicate files, we cannot be certain that all such files were identified. An example of the difficulty in identifying duplicates is given in Figure 6. In this instance, Metabolife had identified in the *77 'serious' AE's* document that these two files belonged to the same case. We would not have been able to make this determination, because the files are separated by more than 7000 numbers on the MIPER CD-ROM (file 16897 and file 24209), and the notes in one file specify "seizure," whereas the other file states "no history of seizure."

In contrast to a duplicate file, a file might contain information on more than one case, either a set of adverse events all experienced by the same consumer or one or more adverse events experienced by several different consumers (see Figure 5, Example 5c). For this reason and because of duplicate files, the number of cases of possible adverse events does not equal the number of files.

In order to review this large CD-ROM dataset within the given time frame, we chose to have the initial data collected by a team of abstractors, each working on a portion of the MIPER CD-ROM. We retained six nurses, each with many years of experience in medical record abstraction. We developed a one-page data collection form to collect key variables related to age, gender, nature of the reported adverse event, and need for hospitalization, which is reproduced in Figure 7. After undergoing training by the principal investigator, each nurse abstractor completed a sample of 135 records, each of which was reviewed in a group meeting with the principal investigator to identify areas of possible misinterpretation and vague language. Based on this experience, we revised the form and developed a "codesheet" to define how certain complaints were to be coded. Formal inter-rater reliability testing was performed on a 1 percent systematic sample of the MIPER files. This sample was stratified into two parts, the larger (N = 114) portion containing only a single adverse event in each file and the smaller portion (N = 16) containing more than one case per file. Inter-rater reliability was assessed using both absolute percentage agreement among abstractors and the kappa statistic, which adjusts for agreement due to chance. Kappa varies between 0 and 1.0, with values of 0.4 to 0.6 usually indicating moderate agreement beyond chance, 0.6 to 0.8 indicating substantial agreement beyond chance, and greater than 0.8 indicating almost perfect agreement.[109] Inter-rater reliability testing demonstrated a kappa statistic of greater than 0.8 or absolute agreement of 95 percent or greater for all variables, indicating almost perfect agreement, for the "one case, one file" (N = 114) records. For the files with multiple cases, two produced disagreement over the number of multiple cases contained in the file. For the remaining 14 multiple-case files, this analysis

showed levels of reliability similar to the "one case, one file analysis." Based on these results, we concluded that the inter-rater reliability for the six nurse abstracters trained in this manner was acceptable for this project. Each nurse was then assigned approximately one-sixth of the MIPER file. Questions that arose during abstraction were posted by email or telephone to our EPC's lead physician abstracter (WAM), who answered their questions, reviewed files himself, and consulted with the principal investigator on decisions requiring nuanced judgment. He maintained the codesheet, keeping it up—to-date and redistributing it to the abstractors whenever changes or additions were made.

We reviewed the forty-three cases that included photocopies of medical information. Personal identifiers had been redacted. Some of these cases were related to cases contained on the MIPER CD-ROM. However, matching these cases was a challenge. As previously noted, we were sent a two-page *Index of Redacted Consumer Medical Records with Corresponding MIPER Numbers*, which indicated a number and the associated files on the MIPER CD-ROM. Unfortunately, the medical records we received were not numbered. Furthermore, a second table that we received entitled *Listing of Key Complaint for the Metabolife Medical Records Submitted* contained a list of main complaints, also numbered. However, the two numbering systems did not agree. We numbered the cases in the order in which we received them in the shipping box. Our numbering system and the two numbering systems we received start out in agreement, but discrepancies occur as we progress through and compare the three systems. We did our best to resolve them.

Analysis of Case Reports

In our draft report, we assigned a likelihood of causality to selected cases, based on our modification of published methods. Many of the peer review comments received for this report pertained to our attempts to assign causality. These comments varied widely, ranging from critiques of our method for being too conservative (meaning, in the opinion of some reviewers, we had excluded or assigned too low a level of causality to certain cases) to critiques for being too liberal (meaning, in the opinion of some reviewers, we had assigned too high a level of causality to certain cases). Often, these conflicting comments concerned the same cases. We believe these peer review comments demonstrate that case report reviews involve considerably more subjective interpretation than do reviews of randomized trials. Because our goal in this evidence report is to report the evidence as objectively as possible, we ceased to assign assessments of causality to the case reports. Rather, we tried to identify those cases that would be classified medically as "idiopathic" in etiology, meaning the cause is not known. For such cases, given the known pharmacology of ephedrine, if use of ephedra or ephedrine was documented, a potential role for ephedra or ephedrine in causing the event must be considered. We classified such cases as "sentinel events."

In order to be classified as a sentinel event, three criteria had to be met:

1. Documentation existed that an adverse event meeting our selection criteria occurred.
2. Documentation existed that the person having the adverse event took an ephedra-containing supplement within 24 hours prior to the event (for cases of death, myocardial infarction, stroke, or seizure).
3. Alternative explanations were investigated and excluded with reasonable certainty.

Within the time and resources available for this evidence report, we were able to do an in-depth review of FDA case reports only for those events classified as death, myocardial infarction (which included acute coronary syndromes), stroke (which included intracerebral hemorrhage), seizures, and severe psychiatric symptoms (see below). Cases that met all three criteria were classified as "sentinel events." Cases where another condition by itself could have caused the adverse event, but for which the known pharmacology of ephedrine made it possible that ephedra or ephedrine may have helped precipitate the event, were classified as "possible sentinel events." "Probably not related" was used for events that had other clear causes discovered on detailed investigation and to which the pharmacology of ephedrine was unlikely to have potentially contributed. We also classified many cases as having insufficient information because crucial information was missing, such as the presence of ephedrine or a metabolite in the blood or documentation that the patient took ephedra within 24 hours prior to the event (for cases of death, myocardial infarction, stroke, or seizure); or other possible causes were insufficiently investigated. (We also classified as "sentinel events" a few cases that, on detailed review, led us to question whether an event meeting our inclusion criteria had actually occurred.)

We translated the criteria for identifying sentinel events into the following set of procedures:

- We required medical record documentation that an adverse event had occurred.

- For adverse events described as seizure, cases described as generalized tonic-clonic seizures underwent further review.

- For psychiatric symptoms, we reviewed cases described as psychosis, mania or severe agitation, severe depression, hallucinations, confusion or delusion, suicide attempt, paranoia, or violence.

- We required (for all but psychiatric events) that there be documentation that the subject had consumed ephedra or ephedrine within 24 hours prior to the adverse event, or that a toxicological examination revealed ephedrine or one of its associated products in the blood or urine. Cases with no such documentation were not reviewed further. For the Metabolife cases, we assumed ephedra use to have been within the prior 24 hours for all but psychiatric events.

- For psychiatric cases, we did not require documentation that the product was taken within 24 hours prior to the event. Ephedrine psychosis (as with amphetamine psychosis in general) is associated with prolonged use, which may lead to neurotoxicity, resulting in depletion of dopamine and other brain monoamines.[110]

- To be eligible for detailed review to investigate other potential causes of death, a file required evidence that an autopsy had been performed, and the results had to be available.

- To be eligible for detailed review to investigate other potential causes for cases of myocardial infarction, coronary angiography had to have been performed and the results had to be available.

- All cases of stroke that met the criterion of having consumed ephedra or ephedrine within 24 hours were reviewed in more detail. To be classified as a "sentinel event," reports of thrombotic stroke needed to have an assessment for a hypercoagulable state and vasculitis, reports of embolic stroke needed to have an embolic evaluation performed, whereas reports of hemorrhagic stroke required an examination to assess structural problems with the circulatory system of the brain.

- Other potential causes of seizure were assessed by searching cases for the results of vital signs, brain imaging (CT or MRI), serum glucose and electrolytes, blood calcium and magnesium, an EEG, and prior history of a seizure disorder or substance abuse.

- For cases with psychiatric symptoms, cases in which patients had a history of psychiatric or severe psychological problems were excluded from further review as reports of possible sentinel events. Cases where the patient reported use of or tested positive for other substances known to cause psychiatric symptoms were also excluded as possible sentinel events. For patients with a prior psychiatric history or use of other substances, these cases were classified as "inconclusive."

One of the key questions we were asked to answer by the sponsoring agencies concerned the relationship between dose and the likelihood of serious adverse events. We do not believe such an analysis is justifiable based on the case report evidence presented here, for the following reasons. First, such an analysis assumes a cause-and-effect relationship that has not been proven by conventional standards of medical science. Second, it would rely to a great extent on patients' recall of dose after having suffered an adverse event, which increases the likelihood of recall bias. Third, and most important, for more than half the adverse-event cases, no dose data were available.

Peer Review

This report was subjected to a lengthy peer review process. An initial draft report was prepared in July 2002. We received comments from 37 reviewers, including representatives from the American Herbal Products Association; Centers for Disease Control and Prevention; Consumer Healthcare Products Association; Council for Responsible Nutrition; Food and Drug Administration; National Center for Complementary and Alternative Medicine; National Institute of Diabetes and Digestive and Kidney Diseases; National Institute of Neurological Disorders and Stroke; National Heart, Lung and Blood Institute; National Institute of Health Office of Dietary Supplements; National Institute of Health Office of Research on Women's Health; National Nutritional Foods Association; Public Citizen Health Research Group; Center for Science in the Public Interest; Utah Natural Products Alliance; and members of the U.S. House of Representatives and U.S. Senate. Additional work requested, involving case report assessments, was performed during Autumn 2002. The "safety" section of the revised report, which contains the new material, was reviewed by additional experts in December 2002. A complete list of Reviewers is in Table 8.

We considered each peer review comment (more than 100 pages in total) and detail our responses in Appendix 3. Service as a reviewer of this report should not in any way be construed as agreeing with or endorsing the content of the report.