

Challenges in Systematic Reviews That Assess Treatment Harms

Roger Chou, MD, and Mark Helfand, MD, MPH

An evidence synthesis of a medical intervention should assess the balance of benefits and harms. Investigators performing systematic reviews of harms face challenges in finding data, rating the quality of harms reporting, and synthesizing and displaying data from different sources. Systematic reviews of harms often rely primarily on published clinical trials. Identifying important harms of treatment and quantifying the risk associated with them, however, often require a broader range of data sources, including unpublished trials, observational studies, and unpublished information on published trials submitted to the U.S. Food and Drug Administration. Each source of data has some potential for yielding important information. Criteria for judging the quality of

harms assessment and reporting are still in their early stages of development. Investigators conducting systematic reviews of harms should consider empirically validating the criteria they use to judge the validity of studies reporting harms. Synthesizing harms data from different sources requires careful consideration of internal validity, applicability, and sources of heterogeneity. This article highlights examples of approaches to methodologic issues associated with performing systematic reviews of harms from 96 Evidence-based Practice Center evidence reports.

Ann Intern Med. 2005;142:1090-1099.

www.annals.org

For author affiliations, see end of text.

To be useful to decision makers, an evidence synthesis of a medical intervention should assess the balance of benefits and harms (1, 2). Harms from medical interventions include adverse drug events (3) and complications following surgery or other procedures. An evidence synthesis that emphasizes only benefits is likely to lead to biased conclusions (4).

For most interventions, unfortunately, systematic reviews of harms are sparser than reviews of benefits. An analysis of more than 1000 systematic reviews or meta-analyses of health care interventions found that just 27% reviewed any harms data, and only 4% primarily focused on safety (5). Of 138 Cochrane systematic reviews of randomized trials with data from at least 4000 participants, only 18% included data on clearly defined harms (6). Difficulties in identifying studies of harms (7), poor quality of adverse event reporting in clinical trials (8), and uncertainty regarding whether systematic reviews of harms yield useful information (9) are some reasons systematic reviews have focused on evaluations of effectiveness.

Investigators performing systematic reviews of harms face challenges in finding and selecting data (1), rating the quality of harms reporting (9), and synthesizing (10) and displaying data from individual studies. In this article, we review these methodologic challenges and highlight examples of approaches to them from 96 Evidence-based Practice Center (EPC) evidence reports.

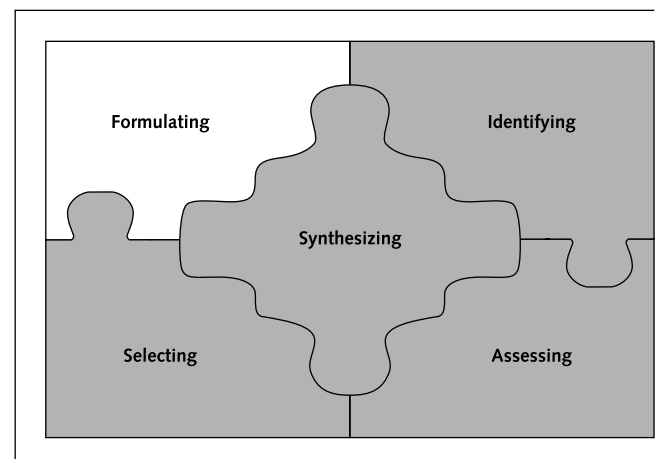
CHALLENGE: IDENTIFYING AND SELECTING INFORMATION ABOUT HARMS

Most systematic reviews rely on searches of electronic databases of published articles and hand searches of relevant journals. Identifying important harms of treatment and quantifying the risk associated with them, however, often require a broader range of data sources. In addition, the types of studies included in an evidence synthesis may influence the quality or amount of evidence regarding harms.

Reliance on Trials

In a sample of 60 meta-analyses that were published in 1995, 33 included only controlled trials (10). Properly designed and executed randomized, controlled trials are often considered the gold standard for evaluating efficacy because they minimize potential bias, but the quality and quantity of harms reporting in trials are often inadequate (11–14). Furthermore, surgical procedures and invasive diagnostic devices often become widely disseminated even though few or no randomized trial data are available (15, 16). In these cases, relying on controlled trials for information about harms is impossible.

By contrast, randomized trials of drugs are numerous. Premarketing trials, conducted according to the requirements of the U.S. Food and Drug Administration (FDA) (15), assess the effects of drugs under ideal circumstances. These “efficacy trials” have limited ability, however, to assess adverse events or applicability to everyday practice. They may exclude patients at high risk for harms or may not be applicable to clinical practice for other reasons (16), may be too short to identify long-term or delayed harms, or may have sample sizes too small to detect uncommon events (1, 17–20).



As required under the Freedom of Information Act, the FDA provides detailed information about the trials submitted in support of an application for approval of a new drug. Unfortunately, this information is often unavailable to the public at the time the drug is approved, and sometimes it is not disclosed until a year or more after the drug is marketed. When this information is available, the approval documents enable the reviewer to compare the results of published and unpublished trials and to compare the material published in journals to the material submitted to the FDA.

Unpublished clinical trials tend to report lower treatment effects than published trials (13, 21). The impact of unpublished trials on assessments of harms has not been well studied, but a recent meta-analysis of antidepressants in children found that addition of data from unpublished trials changed risk–benefit profiles from favorable to unfavorable for several drugs (22). In addition, scientists affiliated with pharmaceutical companies sometimes publish meta-analyses of premarketing trials, some of which have not been fully published individually. These meta-analyses may focus on a single adverse effect or a narrow range of related adverse effects. The narrow focus can obscure the tradeoffs of benefits and harms for the drug in comparison to its competitors. For example, a meta-analysis of studies of venlafaxine reported similar overall withdrawal rates compared with other antidepressants but did not note that the rate of nausea and vomiting was significantly higher for venlafaxine (23). These meta-analyses also may exclude studies not performed by the manufacturer and may not describe the quality-related design characteristics of the individual trials. In the venlafaxine example, the FDA has not released the details of the unpublished premarketing studies, even though this drug was approved in 1993. The lack of information about these trials—how many were done, and whether efficacy, discontinuations, and adverse events were similar to or different from those associated with comparison drugs—make this type of meta-analysis an unreliable and potentially misleading source of information.

Journal publications may omit important information from trials because of space limitations. Although most useful for efficacy analysis (24), drug approval information—especially the clinical and statistical reviews prepared by FDA staff—frequently provides details about harms not included in the journal article. For example, for a major trial (Vioxx Gastrointestinal Outcomes Research Study, or VIGOR) of rofecoxib, an FDA statistical review made available to the public in 2001 contains 6 pages of analysis on the issue of cardiovascular risk (25), compared with 3 lines in the *New England Journal of Medicine* report (26). In fact, before the publication of VIGOR, most trials of rofecoxib did not report the rate of myocardial infarctions. One FDA analysis of trial data showed that the additional risk for cardiovascular events associated with rofecoxib did not appear until after 6 months of follow-up (27). This

finding was of critical importance in evaluating the results of a major trial of celecoxib, the Celecoxib Long-term Arthritis Safety Study (CLASS). Findings from this study were published in the *Journal of the American Medical Association* as 6-month data; at 6 months, rates of serious gastrointestinal or cardiovascular events did not differ significantly between celecoxib and 2 comparators, diclofenac and ibuprofen (28). The article did not mention that some patients in the trial had been observed for longer than 6 months (29, 30). In contrast, the FDA review explains the changes in the protocol for CLASS, compared the 6-month results to longer-term results, and explained the rationale for early termination (27).

Including Observational Studies

Observational studies can be a useful supplement for systematically assessing harms, particularly when effectiveness trials are lacking (31, 32). The term *observational studies* refers to a broad range of study designs, including case reports; retrospective analyses of large claims databases; population-based, longitudinal cohort studies; uncontrolled surgical series of patients receiving an intervention (33); and others (34, 35). All have some potential for yielding useful information.

Well-controlled observational studies demonstrated the associations between maternal diethylstilbestrol use and vaginal adenocarcinoma in young women (36); angiotensin-converting enzyme inhibitors and scleroderma renal crisis; certain appetite suppressants and pulmonary hypertension; and aspirin use among children with influenza and the Reye syndrome (37). The traditional observational designs used in epidemiology—case–control and population-based cohort studies—are subject to confounding and biases that are encountered less commonly in randomized, controlled trials (10, 38, 39). However, those designs take stronger precautions against bias than other observational designs, and their strengths and weaknesses are well understood. Confounding by indication, for example, is usually not an issue with unexpected adverse drug events in these studies because such unpredictable outcomes are usually not associated with the indication for treatment (32, 40).

Of 96 EPC evidence reports published between December 1998 and April 2004, 71 addressed health care interventions associated with potential harms (38). Of these, 11 did not assess harms. Two of the 11 evaluated surgical interventions, and the other 9 evaluated pharmacologic interventions or approaches to diagnostic testing and disease management. Of the 60 remaining EPC reports, 34 included observational studies of adverse events.

Thirteen of 16 EPC reports evaluating surgical, obstetric, dental, or other invasive interventions included observational studies of harms, compared with 21 of 44 that evaluated noninvasive interventions. Most EPC reports evaluating alternative medicine interventions included observational studies of harms. Two EPC reports specifically included observational studies to identify uncommon or

rare adverse events (41, 42); 1 included observational studies when there was insufficient evidence from clinical trials (43); and 1 included observational studies because of concerns about applicability of clinical trials to community practice (44). In other EPC reports, the investigators did not explicitly state how they decided whether to include observational studies of harms, although it appeared that these studies were more likely to be included when clinical trial data were lacking.

Although several studies have found that well-designed controlled observational studies and randomized trials can report similar estimates of effects (45–48), no study has systematically evaluated how frequently observational studies result in different, yet valid, conclusions about harms. One meta-analysis of serious gastrointestinal complications associated with nonsteroidal anti-inflammatory drugs (49) and an EPC report on complications from carotid endarterectomy (44) found that clinical trials reported higher risks for adverse events than observational studies. This could be due to poorer assessment of harms in the observational studies or that observational studies are more likely to be published if they report good results. An EPC report on management of cancer pain found that a sample of observational studies did not change conclusions drawn from randomized trials (43). In other EPC reports, it was not clear whether including observational studies changed assessments about harms. Similarly, systematic reviews by non-EPC investigators either did not assess the effects of including observational studies (9, 50–53) or used observational studies only for areas where clinical trial data were lacking (54, 55).

Including Large Databases

Large databases may provide useful information about harms. For example, Fowler and Wennberg and colleagues used Medicare claims data to identify patients undergoing prostate surgery and surveyed them to estimate the frequency of side effects in practice (56, 57). They found that rates of sexual dysfunction, urinary incontinence, and urethral stricture were higher than reported in case series. Another study that compared Medicare data with data from randomized trials found a higher Medicare rate of acid-related upper gastrointestinal events in women receiving alendronate (58).

Pharmacoepidemiology is a developing science that uses large databases to study drug effects. Many pharmacoepidemiologic studies attempt to emulate 4 epidemiologic study designs: cohort studies, case-control studies, case-crossover studies, and case-time-control studies (34, 59). However, databases used in pharmacoepidemiologic research sometimes use fewer, and often weaker, precautions against bias than prospectively designed epidemiologic databases, such as the Framingham Study and the Nurses' Health Study. While pharmacoepidemiologic studies may be very valuable for examining the frequency of uncom-

mon adverse events, additional empirical research is needed to identify features that are associated with valid findings.

Including Data from Practice-Based Networks

Because medical research has traditionally been based in academic centers, findings may not apply to most patients who receive their care in community settings. Data collected by practice-based research networks may provide better information about benefits and harms of health care interventions in everyday clinical practice (60). Practice-based research data sets are often richer in clinical detail than administrative databases, making it possible to identify and measure likely confounders with more confidence.

One of the largest and most well-known practice-based research networks is the United Kingdom-based General Practice Research Database (GPRD) (61). A recently published study of suicide risk associated with selective serotonin reuptake inhibitors based on GPRD data reported no clear association with increased risk (62). Although this finding is similar to those of meta-analyses of randomized trials (63, 64), the GPRD analysis also suggests that the risks for suicide are not significantly higher in clinical practice, where patients may not be monitored as closely or be as highly selected as in the trials.

Compared with traditional observational studies, studies based on analyses of administrative or practice-based databases are often harder to find by using electronic searches because many such analyses are proprietary. In addition, administrative and practice-based databases often could address more questions than they do. The analyses that are published are determined largely by funders' interest in a particular question (for example, Is celecoxib safer than rofecoxib?) rather than the question that may be of interest to a systematic reviewer (for example, Is celecoxib safer than naproxen?). Because this can introduce a type of publication bias, systematic reviewers must be able to query databases (or their owners) to supplement published information on the questions a review addresses, rather than depending solely on questions posed by interested parties.

Including Case Reports

About 30% of the primary published literature on adverse drug events is in the form of case reports (65). Case reports can identify uncommon, unexpected, or long-term adverse drug events (18, 40) that are often different from those detected in clinical trials (66). Thirteen of 18 confirmed and important adverse drug reactions first reported in 1963, for example, were initially identified by case reports (67). Of 548 new chemical entities approved by the FDA between 1975 and 1999, 56 (10.2%) subsequently received one or more prominent black-box safety warnings ($n = 45$) or were withdrawn from the market ($n = 16$) (68). Although the proportion of withdrawals due to case report data in the United States is not known, case reports were the primary source used to withdraw 18 of 22 drugs from the Spanish market (69).

The FDA receives about 280 000 reports of postmar-

keting adverse events annually and collects them into a database (70). Although pharmaceutical companies may perform high-quality analyses of such data, these analyses are not always made public in a timely fashion, as in the case of the withdrawn lipid-lowering drug cerivastatin (71).

Including Pharmacokinetic and Pharmacodynamic Data

In conducting systematic reviews of adverse events, it is always important to consider whether specific drugs are more likely to be harmful for specific populations. Compared with white persons, for example, African Americans have an increased risk for angioedema from angiotensin-receptor inhibitors (72). Systematic reviews should bring attention to studies that deliberately look at the risks and benefits of specific drugs in subgroups. In many situations, however, the risk in different populations can be difficult to address systematically because controlled trials do not analyze harmful effects in subgroups or exclude patients at higher risk for harms. Often, a comprehensive search for trials and for observational clinical studies proves to be inadequate for evaluating clinical outcomes of drugs in specific populations.

When clinical data on subpopulations are lacking, investigators should consider including pharmacodynamic and pharmacokinetic studies, even though such data do not always correlate with clinical effects. In the case of the lipid-lowering agent rosuvastatin, for example, the FDA required labeling indicating that drug levels are higher in Asian persons (73), although a recently published meta-analysis of trials submitted to the FDA found no differences in clinical adverse events according to ethnicity, sex, or age (74). One role of systematic reviews in these cases is to help distinguish concerns based on clinical data from what is based on pharmacologic properties or on other considerations. The uncertainty regarding the safety of rosuvastatin in Asian persons illustrates how choosing among drugs for specific populations is often an imprecise use of evidence. Although it may not be clear whether the “best drug in general” is the “best drug for everyone,” overstating or overemphasizing differences in kinetics could lead to more harm than good if a decision is made to deprive people of what may prove to be the best drug; thus, such data should be interpreted with caution.

CHALLENGE: ASSESSING THE QUALITY OF HARMS REPORTING

Empirically validated criteria are available to judge the quality of randomized, controlled trials (13, 75, 76), although associations between quality measures and estimates of treatment effect (77, 78) are not straightforward. Unique considerations suggest the need for a distinct set of criteria for judging harms reporting. A recent meta-analysis of rofecoxib, for example, found that having an external end point committee was associated with higher risks for myocardial infarction, but allocation concealment was not associated with differences in risk (55). Widely used criteria

to assess the quality of observational studies (79, 80) and randomized trials, however, were not designed specifically to assess the quality of harms assessment (81). Guidelines for evaluating adverse events in systematic reviews (1) have been published, but they do not give firm recommendations on how to assess the quality of included studies.

Assessing Observational Studies

Because they lack randomization, observational studies should adhere to higher methodologic standards than randomized, controlled trials (1, 10, 37, 39). This point is often overlooked in debates about the integrity of pharmacoepidemiologic data. Randomized, controlled trials are expected to have outcomes recorded by blinded personnel, and to include all participants who were randomly assigned in the analysis of results. Using blinded outcome assessors and using an inception cohort are at least as important in observational studies.

In addition to not being designed for evaluating quality of harms assessment, systematic reviews have found that quality rating instruments for observational studies varied greatly in scope, the number and types of items used, and developmental rigor, and concluded that further study is needed to determine which methodologic characteristics are associated with bias (82–84). Several systematic reviews have evaluated specific methodologic characteristics for their effects on estimates of harms, although the generalizability of their findings is unknown. They found that prospective or retrospective design (84, 85), case–control compared with cohort studies (49, 55), and smaller compared with larger case series (84) had no clear effect on estimates of harms. On the other hand, a recent meta-analysis of observational studies of naproxen found that estimates of cardiovascular risk were lower in studies sponsored by the manufacturer of the competing drug rofecoxib (55).

Assessing Case Reports

Assessing the validity of case reports can be particularly difficult because of difficulties in establishing causality. Modeling, however, suggests that more than 1 to 3 spontaneously reported cases of uncommon or rare adverse events is very unlikely to be coincidental (86). Of 47 case reports published in 1963 in 4 major general medical journals, 35 subsequently proved to be clearly correct (87).

Four EPC reports (41, 42, 88, 89) used criteria to assess adverse drug events reported in case reports and series that included considerations of temporal relationship, lack of alternative causes, presence of toxic concentrations of the drug, response to discontinuation, dose–response relationship, and response to rechallenge. Other disease-specific (90, 91) and non–disease-specific (92, 93) methods for assessing the probability of adverse drug reactions from case reports have also been published. These methods have been validated by using expert opinion (90, 92, 94) or positivity on rechallenge (95) as the gold standard. Guidelines for improving the reporting of suspected adverse drug

Table 1. Quality Assessment Tool for Cohort Studies; Randomized, Controlled Trials; and Uncontrolled Surgical Series That Reported Complications from Carotid Endarterectomy*

Criterion	Score	Studies, <i>n</i>	Pooled Rate of Stroke or Death in Univariate Analyses (95% CI), %
Nonbiased selection	1: Study is a properly randomized, controlled trial or observational study with clear, predefined inception cohort	38	5.2 (4.9–5.6)
	0: Selection not clear or biased selection	9	3.2 (2.8–3.6)
Adequate description of population	1: Study reports ≥2 demographic characteristics, presenting symptoms/syndrome, and ≥1 important risk factor for complications	35	4.6 (4.4–4.9)
	0: Study does not meet above criteria	12	3.4 (2.8–3.6)
Low loss to follow-up, and patients lost to follow-up analyzed for adverse events	1: Study reports number lost to follow-up, analyzes patients lost to follow-up for adverse events, and has low overall number lost to follow-up (threshold set at 5% for studies of carotid endarterectomy)	47	4.5 (4.3–4.8)
	0: Study does not meet above criteria	0	No data
Adverse events prespecified and defined	1: Study reports explicit definitions for major complications that allow for reproducible ascertainment	17	6.7 (6.2–7.3)
	0: Study does not meet above criteria	30	3.5 (3.3–3.8)
Ascertainment technique adequately described	1: Study reports methods used to ascertain complications, including who ascertained, timing, and methods used	15	6.1 (5.7–6.6)
	0: Study does not meet above criteria	32	3.4 (3.1–3.8)
Nonbiased and accurate ascertainment of adverse events	1: Study provides independent assessment of complications (defined as assessment by someone other than the surgeon performing the procedure)	20	6.6 (6.1–7.2)
	0: Study does not meet above criteria	27	3.4 (3.1–3.7)
Adequate statistical analysis of potential confounders	1: Study examines ≥1 relevant confounders/risk factors using acceptable statistical techniques, such as stratification or adjustment	12	6.0 (5.5–6.6)
	0: Study does not meet above criteria	35	3.9 (3.6–4.2)
Adequate duration of follow-up	1: Study reports duration of follow-up and duration of follow-up is adequate to identify expected adverse events (threshold set at 30 days for studies of carotid endarterectomy)	27	5.5 (5.1–5.9)
	0: Study does not meet above criteria	20	3.4 (3.1–3.8)
Total quality score = sum of scores (0–8)	>6: Good	12	6.7 (6.2–7.4)
	4–6: Fair	14	5.1 (4.6–5.7)
	<4: Poor	21	2.9 (2.6–3.2)

* Adapted from reference 44: Meenan RT, Saha S, Swartztrauber K, Krages KP, O’Keefe-Rosetti M, McDonagh M, et al. Effectiveness and cost-effectiveness of echocardiography and carotid imaging in the management of stroke. Evidence Report/Technology Assessment No. 49. Rockville, MD: Agency for Healthcare Research and Quality; July 2002. AHRQ publication no. 02-E022.

events in case reports have recently been proposed (96). Of the 19 recommended items, the median number mentioned in 35 reports of 48 patients published in *BMJ* was only 9 (range, 5 to 12), although effects of missing information on the validity of case reports have not been studied.

Assessing Studies of Surgical Interventions

Studies of surgical interventions are often uncontrolled series and frequently do not meet standards for accurate and comprehensive reporting of complications (97). One EPC report created but did not empirically validate a 4-grade system to rate the quality of surgical series; this system incorporated several methodologic considerations, such as the number of centers, prospective or retrospective design, and use of intention-to-treat analysis (98). In our

EPC report on carotid endarterectomy, we developed and empirically tested an 8-criteria quality-rating instrument for assessing harms reporting from randomized, controlled trials; cohort studies; and uncontrolled surgical series (44). It incorporated factors potentially associated with more rigorous adverse event assessment and assigned an overall quality score and rating (Table 1). Several of these criteria are similar to those proposed in recent guidelines to improve reporting of harms in randomized trials (8).

The quality-rating instrument was pilot tested on 47 studies of carotid endarterectomy. Univariate analyses found pooled rates of stroke or death of 3.8% in poor-quality studies (95% CI, 2.7% to 5.2%), 6.4% in fair-quality studies (CI, 4.5% to 8.7%), and 6.8% in good-

quality studies (CI, 4.6% to 9.5%). As was found in an earlier systematic review (85), independent assessment was associated with higher complication rates. For 6 of the 7 other individual quality-ratings criteria in our quality-ratings tool (follow-up was reported as complete in all studies and could not be assessed), meeting the criterion adequately was also associated with higher complication rates.

Our quality-ratings instrument was tested only on a set of studies for 1 intervention, and analyses did not control for other patient or intervention factors (such as skill or experience of surgeon) that could affect complication rates. Nonetheless, we are aware of no other studies in which methodologic shortcomings identified by a quality-ratings tool for randomized trials and observational studies of surgical complications were associated with lower adverse event rates.

CHALLENGE: SYNTHESIZING AND DISPLAYING DATA FROM DIFFERENT TYPES OF STUDIES

Although the need to incorporate, synthesize, and weigh data from different types of studies for systematic reviews of harms has become widely recognized (31, 99–101), methods to combine data from different sources are

just starting to be developed (102). Because confounding and selection bias can distort findings from observational studies, it is especially important that researchers who include such studies avoid inappropriate statistical combination of data, carefully describe the characteristics and quality of included studies (103), and thoroughly explore potential sources of heterogeneity (10, 104, 105).

An approach proposed by the Grades of Recommendation Assessment, Development and Evaluation (GRADE) Working Group attempts to balance the need to be concise with full and transparent consideration of all important issues. This group recommends that investigators display important factors related to quality assessment (study design, quality, consistency, directness, and other modifying factors) and results (number of patients, effect size, quality, and importance) in a summary table (2). One EPC developed summary tables for evaluation of surgical complications from adrenalectomy that efficiently convey information about the number and types of studies, quality assessment, and results that could be used as a template for systematic reviews of surgical complications (Table 2) (98). Another EPC report developed summary tables to display the number and type of studies and estimate the magni-

Table 2. Summary Table of Evidence for Complications from Surgery for Clinically Inapparent Adrenal Mass*

Studies, n	Patients, n	Tumor Type, %†	Mean Tumor Size (Range), cm	Complications, %			Studies per Quality Grade‡, n			
				Death	Major	Minor	A	B	C	I
Case series of open transperitoneal adrenalectomy										
1	55	Pheochromocytoma, 38 Cancer, 0 Metastasis, ND	4.7 (0.7–8)	0	13	9	0	1	0	0
Case series of open retroperitoneal adrenalectomy										
8	470	Pheochromocytoma, 0–41 Cancer, 0–13 Metastasis, 0–6	1.5–4.3 (0.5–14)	0–3	2–24	0–14	0	4	4	0
Comparative studies of open adrenalectomy techniques										
4	Anterior adrenalectomy, 228	Pheochromocytoma, 0–24 Cancer, 0–48 Metastasis, 0	6.8	0–6.8	6–47	2–36	0	9	4	0
	Posterior adrenalectomy, 338	Pheochromocytoma, 0–17 Cancer, 0–6 Metastasis, 0	7.0	0–1.5	4–25	15–20				

* Adapted from reference 98. Lau J, Balk E, Rothberg M, Ioannidis JP, DeVine D, Chew P, et al. Management of clinically inapparent adrenal mass. Evidence Report/Technology Assessment No. 56. Rockville, MD: Agency for Healthcare Research and Quality; May 2002. AHRQ publication no. 02-E014. ND = not determined.

† Values expressed with a dash are the range.

‡ Grade A (least bias): a multicenter (or multisurgeon) prospective series with matched controls, applying the same exclusion criteria to all study groups; noncomparative case series of consecutive cases with data collected prospectively and no exclusions of difficult cases or bad outcomes (intention-to-treat analysis). Grade B (susceptible to some bias): a single-surgeon retrospective case series with matched controls and the same exclusion criteria for all study groups; retrospective noncomparative case series of consecutive cases without exclusions. Grade C (likely to have significant bias): a study with no matched controls or unequal treatment of study groups; cases are nonconsecutive because of exclusions. Grade I (indeterminate): a study with insufficient information to determine quality.

Table 3. Summary Table of Evidence for Serious Fetal and Neonatal Adverse Drug Reactions from Antihypertensive Drugs during Pregnancy*

Type of Drug	Consistency	Estimate of Magnitude
Angiotensin-converting enzyme inhibitors		
Fetal growth retardation	<10 case reports	Unknown
Pulmonary hypoplasia	<10 case reports	Unknown
Patent ductus arteriosus	<10 case reports	Unknown
Respiratory distress syndrome	>10 case reports	Unknown
Diuretics		
Thrombocytopenia	>10 case reports	Unknown
Neuroblastoma	2 case-control studies	Odds ratio, 4.1–5.75
Deafness	Case-control study	Not given
No serious adverse reactions	9 randomized, controlled trials (>5000 participants)	Not applicable

* Adapted from reference 106: Mulrow CD, Chiquette E, Ferrer RL, Sibai BM, Stevens KR, Harris M, et al. Management of chronic hypertension during pregnancy. Evidence Report/Technology Assessment No. 14. Rockville, MD: Agency for Healthcare Research and Quality; August 2000. AHRQ publication no. 00-E011.

tude of the association for serious adverse antihypertensive drug reactions during pregnancy (Table 3) (106). Other examples of methods for summarizing results from different data sources include Forrest plots of effect size (51) or pooled estimates of risk stratified by study design (49), and summary tables displaying assessments of the likelihood of bias and generalizability of results for each included study (53). A recently published meta-analysis juxtaposed clinical trials of rofecoxib and observational studies of naproxen (trials were not available) to evaluate and compare risks for myocardial infarction (55). These examples all illustrate potential complexities associated with interpreting and synthesizing results from systematic reviews that include different data sources. Until more is known about the methods for performing such evidence syntheses, oversimplification of findings could do more harm than good (10).

Table 4. Recommendations for Improving Systematic Reviews That Assess Harms

Use data from a broad range of sources, particularly when clinical trials are lacking; when generalizability is uncertain; and when investigating rare, long-term, or unexpected adverse events.
Remember that using comprehensive inclusion strategies may increase the detection of possible rare events but also may increase the risk for including biased and poor-quality data.
Use explicit eligibility criteria to determine the inclusion and exclusion of certain studies of harms.
Do not rely on nonvalidated quality-rating tools for studies of harms; rather, support development and validation of appropriate tools.
Consider analyzing the effects of individual study quality indicators, study design, study size, and funding source.
Clearly indicate the sources of data, assessments of study quality, and individual and summary estimates of results.
Avoid inappropriate combining of data, and thoroughly investigate heterogeneous results.
Clearly indicate the populations addressed by included studies, and carefully assess the applicability to other populations.

CONCLUSION AND RECOMMENDATIONS

Better data about harms are needed to conduct balanced systematic reviews. Systematic reviewers often focus on analyzing data from published clinical trials. Information from a broader range of sources, however, may help fill in gaps or provide a more comprehensive look at harms (19, 31, 40).

Additional research is needed to empirically determine the impact of including data from different sources on assessments of harms. Research on optimal strategies for identifying studies for systematic reviews has focused on clinical trials, and similar research on efficient identification of other published and unpublished studies of harms is needed (7, 107). Further development and testing of criteria for rating the quality of harms reporting will help facilitate judgments of the validity of included studies and conclusions of systematic reviews (9, 13, 108).

It is encouraging that attention to methodologic issues associated with conducting systematic evidence reviews on harms is increasing (99). For now, investigators should consider several key points when conducting systematic reviews of harms (Table 4): Investigators should explicitly state and explain the rationale for any decision to exclude specific data sources. Evidence syntheses evaluating associations between interventions and rare adverse drug events, for example, should strongly consider including observational studies, as randomized, controlled trials are unlikely to identify such events (3, 109). For assessments of adverse drug events, FDA documents may also provide important information not available in journal publications. In addition, investigators should specifically state what criteria are being used to evaluate the quality of studies of harms, and consider validating the quality ratings criteria used. This would help advance the science of performing systematic reviews of harms. Finally, investigators should clearly display their results (1, 2, 110), indicating the populations addressed by the included studies and the applicability of the results to other populations (45). They should avoid inappropriately combining data, and thoroughly investi-

gate heterogeneity to promote insight into areas of uncertainty and future research needs (10, 111).

From the Oregon Evidence-based Practice Center, Oregon Health & Science University, and Portland Veterans Affairs Medical Center, Portland, Oregon.

Acknowledgments: The authors thank Michele Freeman for her help in abstracting data.

Grant Support: This study was conducted by the Oregon Evidence-based Practice Center under contract to the Agency for Healthcare Research and Quality (contract 290-02-0024, task order 1).

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

Requests for Single Reprints: Roger Chou, MD, Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Mail Code BICC, Portland, OR; e-mail, chour@ohsu.edu.

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

References

1. Loke YK, Price D, Herxheimer A. Including adverse effects in your review. *Cochrane Adverse Effects Subgroup*. Accessed at www.dsr.org/wwwboard/latestdraft.pdf on 2 August 2004.
2. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. GRADE Working Group. *BMJ*. 2004;328:1490. [PMID: 15205295]
3. Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. *Lancet*. 2000;356:1255-9. [PMID: 11072960]
4. Cuervo LG, Clarke M. Balancing benefits and harms in health care [Editorial]. *BMJ*. 2003;327:65-6. [PMID: 12855496]
5. Ernst E, Pittler MH. Assessment of therapeutic safety in systematic reviews: literature review. *BMJ*. 2001;323:546. [PMID: 11546700]
6. Papanikolaou PN, Ioannidis JP. Availability of large-scale evidence on specific harms from systematic reviews of randomized trials. *Am J Med*. 2004;117:582-9. [PMID: 15465507]
7. Derry S, Kong Loke Y, Aronson JK. Incomplete evidence: the inadequacy of databases in tracing published adverse drug reactions in clinical trials. *BMC Med Res Method*. 2001;1:7. [PMID: 11591220]
8. Ioannidis JP, Evans SJ, Gotzsche PC, O'Neill RT, Altman DG, Schulz K, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med*. 2004;141:781-8. [PMID: 15545678]
9. McIntosh HM, Woolacott NF, Bagnall AM. Assessing harmful effects in systematic reviews. *BMC Med Res Method*. 2004;4:19. [PMID: 15260887]
10. Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. *BMJ*. 1998;316:140-4. [PMID: 9462324]
11. Edwards JE, McQuay HJ, Moore RA, Collins SL. Reporting of adverse effects in clinical trials should be improved: lessons from acute postoperative pain. *J Pain Symptom Manage*. 1999;18:427-37. [PMID: 10641469]
12. Papanikolaou PN, Churchill R, Wahlbeck K, Ioannidis JP. Safety reporting in randomized trials of mental health interventions. *Am J Psychiatry*. 2004;161:1692-7. [PMID: 15337661]
13. Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technol Assess*. 2003;7:1-76. [PMID: 12583822]
14. Ioannidis JP, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *JAMA*. 2001;285:437-43. [PMID: 11242428]
15. Moore SW. An overview of drug development in the United States and current challenges. *South Med J*. 2003;96:1244-55; quiz 1256. [PMID: 14696877]
16. Rothwell PM. External validity of randomised controlled trials: "to whom do

- the results of this trial apply?". *Lancet*. 2005;365:82-93. [PMID: 15639683]
17. Ray WA. Population-based studies of adverse drug effects. *N Engl J Med*. 2003;349:1592-4. [PMID: 14573730]
18. Kaufman DW, Shapiro S. Epidemiological assessment of drug-induced disease. *Lancet*. 2000;356:1339-43. [PMID: 11073036]
19. Vandembroucke JP. Benefits and harms of drug treatments [Editorial]. *BMJ*. 2004;329:2-3. [PMID: 15231587]
20. Dieppe P, Bartlett C, Davey P, Doyal L, Ebrahim S. Balancing benefits and harms: the example of non-steroidal anti-inflammatory drugs. *BMJ*. 2004;329:31-4. [PMID: 15231619]
21. McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet*. 2000;356:1228-31. [PMID: 11072941]
22. Whittington CJ, Kendall T, Fonagy P, Cottrell D, Cotgrove A, Boddington E. Selective serotonin reuptake inhibitors in childhood depression: systematic review of published versus unpublished data. *Lancet*. 2004;363:1341-5. [PMID: 15110490]
23. Thase ME, Entsuah AR, Rudolph RL. Remission rates during treatment with venlafaxine or selective serotonin reuptake inhibitors. *Br J Psychiatry*. 2001;178:234-41. [PMID: 11230034]
24. Santaguida PL, Helfand M, Raina P. Challenges in systematic reviews that evaluate drug efficacy or effectiveness. *Ann Intern Med*. 2005;142:1066-72.
25. Lee S. Statistical review. Center for Drug Evaluation and Research. Accessed at www.fda.gov/cder/foi/nda/2002/21-042S007_Vioxx_star.pdf on 28 February 2005.
26. Bombardier C, Laine L, Reicin A, Shapiro D, Burgos-Vargas R, Davis B, et al. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *N Engl J Med*. 2000;343:1520-8, 2 p following 1528. [PMID: 11087881]
27. Witter J. Medical review part 1. Center for Drug Evaluation and Research. Accessed at www.fda.gov/cder/foi/nda/2002/20-998S009_Celebrex_medr_P1.pdf on 28 February 2005.
28. Silverstein FE, Faich G, Goldstein JL, Simon LS, Pincus T, Whelton A, et al. Gastrointestinal toxicity with celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis: the CLASS study: a randomized controlled trial. Celecoxib Long-term Arthritis Safety Study. *JAMA*. 2000;284:1247-55. [PMID: 10979111]
29. Wright JM, Perry TL, Bassett KL, Chambers GK. Reporting of 6-month vs 12-month data in a clinical trial of celecoxib [Letter]. *JAMA*. 2001;286:2398-400. [PMID: 11712925]
30. Hrachovec JB, Mora M. Reporting of 6-month vs 12-month data in a clinical trial of celecoxib [Letter]. *JAMA*. 2001;286:2398; author reply 2399-400. [PMID: 11712924]
31. Glasziou P, Vandembroucke JP, Chalmers I. Assessing the quality of research. *BMJ*. 2004;328:39-41. [PMID: 14703546]
32. Vandembroucke JP. When are observational studies as credible as randomised trials? *Lancet*. 2004;363:1728-31. [PMID: 15158638]
33. Black N. Evidence-based surgery: a passing fad? *World J Surg*. 1999;23:789-93. [PMID: 10415204]
34. Etminan M, Samii A. Pharmacoepidemiology I: a review of pharmacoepidemiologic study designs. *Pharmacotherapy*. 2004;24:964-9. [PMID: 15338844]
35. Farrington CP. Control without separate controls: evaluation of vaccine safety using case-only methods. *Vaccine*. 2004;22:2064-70. [PMID: 15121324]
36. Herbst AL, Ulfelder H, Poskanzer DC. Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. *N Engl J Med*. 1971;284:878-81. [PMID: 5549830]
37. Ray JG. Evidence in upheaval: incorporating observational data into clinical practice. *Arch Intern Med*. 2002;162:249-54. [PMID: 11822916]
38. Psaty BM, Koepsell TD, Lin D, Weiss NS, Siscovick DS, Rosendaal FR, et al. Assessment and control for confounding by indication in observational studies. *J Am Geriatr Soc*. 1999;47:749-54. [PMID: 10366179]
39. Lawlor DA, Davey Smith G, Kundu D, Bruckdorfer KR, Ebrahim S. Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? *Lancet*. 2004;363:1724-7. [PMID: 15158637]
40. Stricker BH, Psaty BM. Detection, verification, and quantification of adverse drug reactions. *BMJ*. 2004;329:44-7. [PMID: 15231627]
41. Mulrow CD, Williams JW, Trivedi M, Chiquette E, Aguilar C, Cornell JE. Treatment of depression: newer pharmacotherapies. Evidence Report/Technology Assessment No. 7 (Prepared by San Antonio Evidence-based Practice Center

- based at the University of Texas Health Science Center at San Antonio under contract 290-97-0012). Rockville, MD: Agency for Health Care Policy and Research; February 1999. AHCPR publication no. 99-E014.
42. Shekelle P, Hardy M, Morton SC, Maglione M, Suttorp M, Roth E, et al. Ephedra and ephedrine for weight loss and athletic performance enhancement: clinical efficacy and side effects. Evidence Report/Technology Assessment No. 76 (Prepared by Southern California-RAND Evidence-based Practice Center under contract 290-97-0001). Rockville, MD: Agency for Healthcare Research and Quality; March 2003. AHRQ publication no. 03-E022.
 43. Goudas L, Carr DB, Bloch R, Balk E, Ioannidis JPA, Terrin N, et al. Management of cancer pain. Evidence Report/Technology Assessment No. 35 (Prepared by New England Medical Center Evidence-based Practice Center under contract 290-97-0019). Rockville, MD: Agency for Healthcare Research and Quality; October 2001. AHRQ publication no. 02-E002.
 44. Meenan RT, Saha S, Swartztrauber K, Krages KP, O'Keefe-Rosetti M, McDonagh M, et al. Effectiveness and cost-effectiveness of echocardiography and carotid imaging in the management of stroke. Evidence Report/Technology Assessment No. 49 (Prepared by Oregon Health Sciences University Evidence-based Practice Center under contract 290-97-0018). Rockville, MD: Agency for Healthcare Research and Quality; July 2002. AHRQ publication no. 02-E022.
 45. Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. *Health Technol Assess*. 1998;2:i-iv, 1-124. [PMID: 9793791]
 46. Concato J, Shah N, Horwitz RL. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med*. 2000;342:1887-92. [PMID: 10861325]
 47. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA*. 2001;286:821-30. [PMID: 11497536]
 48. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med*. 2000;342:1878-86. [PMID: 10861324]
 49. Ofman JJ, MacLean CH, Straus WL, Morton SC, Berger ML, Roth EA, et al. A metaanalysis of severe upper gastrointestinal complications of nonsteroidal antiinflammatory drugs. *J Rheumatol*. 2002;29:804-12. [PMID: 11950025]
 50. Bond R, Rerkasem K, Rothwell PM. Systematic review of the risks of carotid endarterectomy in relation to the clinical indication for and timing of surgery. *Stroke*. 2003;34:2290-301. [PMID: 12920260]
 51. Lawlor DA, Juni P, Ebrahim S, Egger M. Systematic review of the epidemiologic and trial evidence of an association between antidepressant medication and breast cancer. *J Clin Epidemiol*. 2003;56:155-63. [PMID: 12654410]
 52. Jefferson T, Rudin M, DiPietrantonj C. Systematic review of the effects of pertussis vaccines in children. *Vaccine*. 2003;21:2003-14. [PMID: 12706690]
 53. Jefferson T, Price D, Demicheli V, Bianco E, European Research for Improved Vaccine Safety Surveillance (EUSAFEVAC) Project. Unintended events following immunization with MMR: a systematic review. *Vaccine*. 2003;21:3954-60. [PMID: 12922131]
 54. Psaty BM, Smith NL, Siscovick DS, Koepsell TD, Weiss NS, Heckbert SR, et al. Health outcomes associated with antihypertensive therapies used as first-line agents. A systematic review and meta-analysis. *JAMA*. 1997;277:739-45. [PMID: 9042847]
 55. Juni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *Lancet*. 2004;364:2021-9. [PMID: 15582059]
 56. Fowler FJ Jr, Barry MJ, Lu-Yao G, Roman A, Wasson J, Wennberg JE. Patient-reported complications and follow-up treatment after radical prostatectomy. The National Medicare Experience: 1988-1990 (updated June 1993). *Urology*. 1993;42:622-9. [PMID: 8256394]
 57. Wennberg JE, Roos N, Sola L, Schori A, Jaffe R. Use of claims data systems to evaluate health care outcomes. Mortality and reoperation following prostatectomy. *JAMA*. 1987;257:933-6. [PMID: 3543419]
 58. Ettinger B, Pressman A, Schein J. Clinic visits and hospital admissions for care of acid-related upper gastrointestinal disorders in women using alendronate for osteoporosis. *Am J Manag Care*. 1998;4:1377-82. [PMID: 10338731]
 59. Etminan M. Pharmacoepidemiology II: the nested case-control study—a novel approach in pharmacoepidemiologic research. *Pharmacotherapy*. 2004;24:1105-9. [PMID: 15460170]
 60. Lindbloom EJ, Ewigman BG, Hickner JM. Practice-based research networks: the laboratories of primary care research. *Med Care*. 2004;42:III45-9. [PMID: 15026664]
 61. Wood L, Martinez C. The general practice research database: role in pharmacovigilance. *Drug Saf*. 2004;27:871-81. [PMID: 15366975]
 62. Martinez C, Rietbrock S, Wise L, Ashby D, Chick J, Moseley J, et al. Antidepressant treatment and the risk of fatal and non-fatal self harm in first episode depression: nested case-control study. *BMJ*. 2005;330:389. [PMID: 15718538]
 63. Gunnell D, Saperia J, Ashby D. Selective serotonin reuptake inhibitors (SSRIs) and suicide in adults: meta-analysis of drug company data from placebo controlled, randomised controlled trials submitted to the MHRA's safety review. *BMJ*. 2005;330:385. [PMID: 15718537]
 64. Fergusson D, Doucette S, Glass KC, Shapiro S, Healy D, Hebert P, et al. Association between suicide attempts and selective serotonin reuptake inhibitors: systematic review of randomised controlled trials. *BMJ*. 2005;330:396. [PMID: 15718539]
 65. Aronson JK, Derry S, Loke YK. Adverse drug reactions: keeping up to date. *Fundam Clin Pharmacol*. 2002;16:49-56. [PMID: 11903512]
 66. Loke YK, Derry S, Aronson JK. A comparison of three different sources of data in assessing the frequencies of adverse reactions to amiodarone. *Br J Clin Pharmacol*. 2004;57:616-21. [PMID: 15089815]
 67. Venning GR. Identification of adverse reactions to new drugs. III: Alerting processes and early warning systems. *Br Med J (Clin Res Ed)*. 1983;286:458-60. [PMID: 6401562]
 68. Lasser KE, Allen PD, Woolhandler SJ, Himmelstein DU, Wolfe SM, Bor DH. Timing of new black box warnings and withdrawals for prescription medications. *JAMA*. 2002;287:2215-20. [PMID: 11980521]
 69. Arnaiz JA, Carne X, Riba N, Codina C, Ribas J, Trilla A. The use of evidence in pharmacovigilance. Case reports as the reference source for drug withdrawals. *Eur J Clin Pharmacol*. 2001;57:89-91. [PMID: 11372600]
 70. Strom BL. Potential for conflict of interest in the evaluation of suspected adverse drug reactions: a counterpoint. *JAMA*. 2004;292:2643-6. [PMID: 15572722]
 71. Psaty BM, Furberg CD, Ray WA, Weiss NS. Potential for conflict of interest in the evaluation of suspected adverse drug reactions: use of cerivastatin and risk of rhabdomyolysis. *JAMA*. 2004;292:2622-31. [PMID: 15572720]
 72. Brown NJ, Ray WA, Snowden M, Griffin MR. Black Americans have an increased rate of angiotensin converting enzyme inhibitor-associated angioedema. *Clin Pharmacol Ther*. 1996;60:8-13. [PMID: 8689816]
 73. Food and Drug Administration Center for Drug Evaluation and Research. FDA public health advisory for Crestor (rosuvastatin). Accessed at www.fda.gov/cder/drug/advisory/crestor.htm on 1 March 2005.
 74. Shepherd J, Hunninghake DB, Stein EA, Kastelein JJ, Harris S, Pears J, et al. Safety of rosuvastatin. *Am J Cardiol*. 2004;94:882-8. [PMID: 15464670]
 75. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet*. 1998;352:609-13. [PMID: 9746022]
 76. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials*. 1996;17:1-12. [PMID: 8721797]
 77. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*. 1999;282:1054-60. [PMID: 10493204]
 78. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA*. 2002;287:2973-82. [PMID: 12052127]
 79. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health*. 1998;52:377-84. [PMID: 9764259]
 80. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (minors): development and validation of a new instrument. *ANZ J Surg*. 2003;73:712-6. [PMID: 12956787]
 81. Centre for Reviews and Dissemination. Undertaking Systematic Reviews of Research on Effectiveness. Khan SK, ter Riet G, Glanville J, Sowden AJ, Kleijnen J, eds. York, United Kingdom: Univ of York; 2001.
 82. Saunders LD, Soomro GM, Buckingham J, Jamtvedt G, Raina P. Assessing the methodological quality of nonrandomized intervention studies. *West J Nurs Res*. 2003;25:223-37. [PMID: 12666645]
 83. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47 (Prepared by Research Triangle Institute-University of North Carolina Evidence-based Practice Center under contract 290-97-0011). Rockville,

MD: Agency for Healthcare Research and Quality; April 2002. AHRQ publication no. 02-E016.

84. Dalziel K, Round A, Stein K, Garside R, Castelnuovo E, Payne L. Do the findings of case series studies vary significantly according to methodological characteristics? *Health Technol Assess*. 2005;9:iii-iv, 1-146. [PMID: 15588556]

85. Rothwell PM, Slattery J, Warlow CP. A systematic review of the risks of stroke and death due to endarterectomy for symptomatic carotid stenosis. *Stroke*. 1996;27:260-5. [PMID: 8571420]

86. Begaud B, Moride Y, Tubert-Bitter P, Chaslerie A, Haramburu F. False-positives in spontaneous reporting: should we worry about them? *Br J Clin Pharmacol*. 1994;38:401-4. [PMID: 7893579]

87. Venning GR. Validity of anecdotal reports of suspected adverse drug reactions: the problem of false alarms. *Br Med J (Clin Res Ed)*. 1982;284:249-52. [PMID: 6799125]

88. Mulrow C, Lawrence V, Ackermann R, Ramirez G, Morbidoni L, Aguilar C, et al. Garlic: effects on cardiovascular risks and disease, protective effects against cancer, and clinical adverse effects. Evidence Report/Technology Assessment No. 20 (Prepared by San Antonio Evidence-based Practice Center based at the University of Texas Health Science Center at San Antonio and the Veterans Evidence-based Research, Dissemination, and Implementation Center, a Veterans Affairs Health Services Research and Development Center of Excellence, under contract 290-97-0012). Rockville, MD: Agency for Healthcare Research and Quality; October 2000. AHRQ publication no. 01-E023.

89. Mulrow C, Lawrence V, Jacobs B, Dennehy C, Sapp J, Ramirez G, et al. Milk thistle: effects on liver disease and cirrhosis and clinical adverse effects. Evidence Report/Technology Assessment No. 21 (Prepared by San Antonio Evidence-based Practice Center based at the University of Texas Health Science Center at San Antonio and the Veterans Evidence-based Research, Dissemination, and Implementation Center, a Veterans Affairs Health Services Research and Development Center of Excellence, under contract 290-97-0012). Rockville, MD: Agency for Healthcare Research and Quality; October 2000. AHRQ publication no. 01-E025.

90. Maria VA, Victorino RM. Development and validation of a clinical scale for the diagnosis of drug-induced hepatitis. *Hepatology*. 1997;26:664-9. [PMID: 9303497]

91. Danan G, Benichou C. Causality assessment of adverse reactions to drugs—I. A novel method based on the conclusions of international consensus meetings: application to drug-induced liver injuries. *J Clin Epidemiol*. 1993;46:1323-30. [PMID: 8229110]

92. Naranjo CA, Busto U, Sellers EM, Sandor P, Ruiz I, Roberts EA, et al. A method for estimating the probability of adverse drug reactions. *Clin Pharmacol Ther*. 1981;30:239-45. [PMID: 7249508]

93. Michel DJ, Knodel LC. Comparison of three algorithms used to evaluate adverse drug reactions. *Am J Hosp Pharm*. 1986;43:1709-14. [PMID: 3752106]

94. Lucena MI, Camargo R, Andrade RJ, Perez-Sanchez CJ, Sanchez De La Cuesta F. Comparison of two clinical scales for causality assessment in hepatotoxicity. *Hepatology*. 2001;33:123-30. [PMID: 11124828]

95. Benichou C, Danan G, Flahault A. Causality assessment of adverse reactions

to drugs—II. An original model for validation of drug causality assessment methods: case reports with positive rechallenge. *J Clin Epidemiol*. 1993;46:1331-6. [PMID: 8229111]

96. Aronson JK. Anecdotes as evidence [Editorial]. *BMJ*. 2003;326:1346. [PMID: 12816800]

97. Martin RC 2nd, Brennan MF, Jaques DP. Quality of complication reporting in the surgical literature. *Ann Surg*. 2002;235:803-13. [PMID: 12035036]

98. Lau J, Balk E, Rothberg M, Ioannidis JP, DeVine D, Chew P, et al. Management of clinically inapparent adrenal mass. Evidence Report/Technology Assessment No. 56 (Prepared by New England Medical Center Evidence-based Practice Center under contract 290-97-0019). Rockville, MD: Agency for Healthcare Research and Quality; May 2002. AHRQ publication no. 02-E014.

99. Cuervo LG, Aronson JK. The road to health care [Editorial]. *BMJ*. 2004;329:1-2. [PMID: 15231586]

100. Smith LA, Moore RA, McQuay HJ, Gavaghan D. Using evidence from different sources: an example using paracetamol 1000 mg plus codeine 60 mg. *BMC Med Res Methodol*. 2001;1:1. [PMID: 11231885]

101. Rosendaal FR. Bridging case-control studies and randomized trials. *Curr Control Trials Cardiovasc Med*. 2001;2:109-110. [PMID: 11806781]

102. Wald NJ, Morris JK. Teleanalysis: combining data from different types of study. *BMJ*. 2003;327:616-8. [PMID: 12969936]

103. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA*. 2000;283:2008-12. [PMID: 10789670]

104. Berlin JA, Colditz GA. The role of meta-analysis in the regulatory process for foods, drugs, and devices. *JAMA*. 1999;281:830-4. [PMID: 10071005]

105. McPherson K, Hemminki E. Synthesising licensing data to assess drug safety. *BMJ*. 2004;328:518-20. [PMID: 14988197]

106. Mulrow CD, Chiquette E, Ferrer RL, Sibai BM, Stevens KR, Harris M, et al. Management of chronic hypertension during pregnancy. Evidence Report/Technology Assessment No. 14 (Prepared by San Antonio Evidence-based Practice Center based at the University of Texas Health Science Center at San Antonio under contract 290-97-0012). Rockville, MD: Agency for Healthcare Research and Quality; August 2000. AHRQ publication no. 00-E011.

107. Dickersin K. Systematic reviews in epidemiology: why are we so far behind? *Int J Epidemiol*. 2002;31:6-12. [PMID: 11914282]

108. Ross SD. Drug-related adverse events: a readers' guide to assessing literature reviews and meta-analyses. *Arch Intern Med*. 2001;161:1041-6. [PMID: 11322836]

109. Mann RD. Prescription-event monitoring—recent progress and future horizons. *Br J Clin Pharmacol*. 1998;46:195-201. [PMID: 9764958]

110. Schunemann HJ, Best D, Vist G, Oxman AD. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. *CMAJ*. 2003;169:677-80. [PMID: 14517128]

111. Petticrew M. Why certain systematic reviews reach uncertain conclusions. *BMJ*. 2003;326:756-8. [PMID: 12676848]

Current Author Addresses: Drs. Chou and Helfand: Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Mail Code BICC, Portland, OR 97239.