Technical Review

Number 13

Empirical Evaluation of the Association Between Methodological Shortcomings and Estimates of Adverse Events

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. 290-02-0024

Prepared by:
Oregon Evidence-based Practice Center, Portland, OR

Investigators
Roger Chou, M.D.
Rongwei Fu, Ph.D.
Susan Carson, M.P.H.
Somnath Saha, M.D., M.P.H.
Mark Helfand, M.D., M.P.H.

AHRQ Publication No. 07-0003
October 2006

This report is based on research conducted by the Oregon Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-02-0024). The findings and conclusions in this document are those of the authors, who are responsible for its content, and do not necessarily represent the views of AHRQ. No statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help clinicians, employers, policymakers, and others make informed decisions about the provision of health care services. This report is intended as a reference and not as a substitute for clinical judgment.

This report may be used, in whole or in part, as the basis for the development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

Suggested Citation:
Chou R, Fu R, Carson S, Saha S, Helfand M. Empirical Evaluation of the Association Between Methodological Shortcomings and Estimates of Adverse Events. Technical Review No. 13. (Prepared by the Oregon Evidence-based Practice Center under Contract No. 290-02-0024.**)** AHRQ Publication No. 07-0003. Rockville, MD: Agency for Healthcare Research and Quality. October 2006.

Dr. Saha was supported by a Research Career Development award from the Health Services Research and Development Service of the Department of Veterans Affairs and by a Generalist Physician Faculty Scholar award from the Robert Wood Johnson Foundation.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To bring the broadest range of experts into the development of evidence reports and health technology assessments, AHRQ encourages the EPCs to form partnerships and enter into collaborations with other medical and research organizations. The EPCs work with these partner organizations to ensure that the evidence reports and technology assessments they produce will become building blocks for health care quality improvement projects throughout the Nation. The reports undergo peer review prior to their release.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality.

We welcome comments on this evidence report. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to **epc@ahrq.gov.**

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Beth A. Collins Sharp, Ph.D., R.N.
Director, EPC Program
Agency for Healthcare Research and Quality

Gurvaneet Randhawa, M.D., M.P.H.
EPC Program Task Order Officer
Agency for Healthcare Research and Quality

Acknowledgments

The authors would like to acknowledge Peter Juni, M.D., for providing data on rates of myocardial infarction in clinical trials of rofecoxib; David Atkins, M.D., M.P.H., and Gurvaneet Randhawa, M.D., M.P.H., at the Agency for Healthcare Research and Quality for reviewing and providing insightful comments for earlier versions of this manuscript; and Jayne Schablaske for administrative support with this manuscript.

Structured Abstract

Objective: Accurate harms data are necessary to appropriately assess the balance between benefits and harms of interventions. Numerous deficiencies in the quality and reporting of harms associated with clinical interventions have been reported. Little is known, however, about whether such perceived methodological shortcomings are associated with lower estimates of harms.

Study Design and Setting: Studies reporting harms associated with (1) carotid endarterectomy (CEA) for symptomatic stenosis, (2) rofecoxib for arthritis, and (3) CEA for asymptomatic stenosis were identified using published systematic reviews. A standardized abstraction form, including eight pre-defined criteria for assessing the quality of harms reporting, was used to extract data. Univariate and multivariate analyses were performed to empirically evaluate the association between quality criteria and estimates of harms. A quality-rating instrument for studies reporting harms was developed based on the results of the analyses of studies of CEA for symptomatic stenosis. The quality-rating instrument was tested on the other data sets.

Results: In 111 studies of CEA for symptomatic stenosis, meeting five of the eight quality criteria was associated with significantly higher rates of stroke or death. A quality-rating instrument with four of the five criteria predicted adverse events (5.7 percent in studies rated “adequate”, compared to 3.9 percent in studies rated “inadequate” [p=0.0003]). In multivariate analyses, the quality rating assignment remained significant when controlling for other clinical and study-related variables. Different quality criteria, however, predicted lower estimates of risk for myocardial infarction in 16 trials of rofecoxib, and none of the quality criteria predicted lower estimates of stroke of death in 18 studies of CEA for asymptomatic stenosis. Evaluation of the latter two data sets was limited by small numbers of trials and low rates of evaluated adverse events.

Conclusions: The presence of methodological shortcomings can predict lower estimates of serious harms. Clinicians and researchers should carefully consider the potential effects of individual methodological shortcomings when evaluating estimates of harms associated with clinical interventions. However, we were unable to develop a generic summary quality-rating instrument for studies of harms because specific methodological shortcomings were not consistently predictive across data sets.

Contents

Tables

Technical Review

Chapter 1: Introduction

Every health care intervention is associated with a risk of harmful or adverse events that must be balanced against the potential favorable outcomes.[1] Recent highly publicized examples of harms associated with medical interventions emphasize the importance of accurate measurement and reporting of harms.[2-4] In studies that report harms, estimates of adverse events may vary because of differences in the underlying risk of the populations studied,[5, 6] differences in the delivery of the intervention (such as the skill of the surgeon[7] or dosing of medication), and random or systematic errors in adverse event assessment due to deficiencies in the design or execution of the study. Systematic reviews should therefore assess the generalizability of study findings to other clinical settings and populations, and evaluate for potential biases in assessment and reporting of harms.

Studies have consistently demonstrated that assessment and reporting of harms in clinical trials is suboptimal, with adverse events infrequently defined, unclear or inadequate methods for identifying adverse events, poor description of severity of harms, and little space devoted to adverse event reporting.[8-12] Guidelines for improving harms reporting have been published to help address perceived shortcomings in measurement, analysis, and reporting of harms data.[13] The Cochrane Collaboration has also proposed draft guidelines for evaluating adverse events in systematic reviews.[14] However, little is known about the empirical associations between such perceived shortcomings and estimates of harms.[15] Empiric evaluations have been published on the association between methodological shortcomings and estimates of efficacy[16-19] and diagnostic test characteristics.[20] Similar information is necessary to help readers appropriately judge the validity of studies reporting harms data.

The purpose of our study was to empirically assess the association between perceived methodological shortcomings and estimates of serious complications from surgical and medical interventions. We compared estimates of harms from studies meeting eight pre-defined criteria designed to assess the quality of harms reporting to studies not meeting the criteria. We hypothesized that estimates of harms would be understated in studies with methodological shortcomings as measured by the quality criteria, even after controlling for other clinical and study design-related variables that could affect adverse event rates.

Our main analyses were performed on a large data set of studies of carotid endarterectomy (CEA) for symptomatic stenosis. Based on the analyses of this data set, we developed a generic summary quality-rating instrument for studies reporting harms. In order to assess the reproducibility of our results, we also analyzed a smaller data set of studies of the same intervention (CEA) in patients with asymptomatic stenosis. In addition, because the association between methodological shortcomings and estimates of harms may differ across interventions, we also evaluated a third data set consisting of studies of the drug rofecoxib, a cyclooxygenase-2 selector inhibitor recently withdrawn from the market because of concerns over increased cardiovascular risk.[21]

Chapter 2: Methods

Selection of Relevant Studies

Studies of CEA for symptomatic stenosis were identified from our earlier evidence review of diagnostic strategies for stroke[22] and augmented with studies from the reference lists of three other systematic reviews.[23-25] We included all randomized controlled trials, cohort studies, and surgical series reporting complication rates. We retrieved all studies of CEA for asymptomatic stenosis from the reference lists of two systematic reviews.[25, 26] We also retrieved all of the randomized controlled trials of rofecoxib for arthritis included in a recent meta-analysis.[2] Abstracts were excluded because they provide insufficient information to adequately judge quality.

Assessment of Study Quality

Study quality was assessed using eight criteria (Table 1). Each criterion was chosen on the basis of: (1) face validity as a marker of rigorous adverse event assessment, or (2) results from an earlier meta-analysis[24] of CEA suggesting predictability for adverse event rates. For instance, studies with independent assessment of complications (criterion 6) were associated with higher complication rates than studies in which the surgeon performing the procedure assessed complications.[24] Prospective or retrospective study design, on the other hand, was not included in our criteria list because it was not a significant predictor in the earlier meta-analysis. Examples of criteria that have not been empirically studied but had face validity include pre-defining of adverse events (criterion 4) and non-biased selection of patients for assessment of harms (criterion 1).

Studies were scored either 0 (inadequate) or 1 (adequate) for each criterion. Studies that did not report enough information to accurately assess a criterion were rated inadequate.

Table 1. Quality assessment tool for studies reporting adverse events

Criterion	Score
Quality criterion 1: Non-biased selection	1: Study is a properly randomized controlled trial (according to Jadad criteria[17]), or an observational study with a clear pre-defined inception cohort (that attempted to evaluate all patients in the inception cohort) 0: Study does not meet above criteria (e.g., convenience samples)
Quality criterion 2: Adequate description of population	1: Study reports 2 or more demographic characteristics, presenting symptoms/syndrome and at least 1 important risk factor for complications 0: Study does not meet above criteria
Quality criterion 3: Low loss to follow-up	1: Study reports number lost to follow-up, and the overall number lost to follow-up is low (threshold set at 5% for studies of carotid endarterectomy and 10% for studies of rofecoxib) 0: Study does not meet above criteria
Quality criterion 4: Adverse events pre-specified and defined	1: Study reports explicit definitions for major complications that allow for reproducible ascertainment (what adverse events were being investigated and what constituted an event) 0: Study does not meet above criteria
Quality criterion 5: Ascertainment technique adequately described	1: Study reports methods used to ascertain complications, including who ascertained, timing, and methods used 0: Study does not meet above criteria
Quality criterion 6: Non-biased ascertainment of adverse events	1: Independent or masked assessment or complications (for studies of carotid endarterectomy, someone other than the surgeon who performed the procedure; for studies of rofecoxib, presence of an external endpoint committee blinded to treatment allocation) 0: Study does not meet above criteria
Quality criterion 7: Adequate statistical analysis of potential confounders	1: Study examines 1 or more relevant confounders/risk factors (in addition to the comparison group in controlled studies), using acceptable statistical techniques such as stratification or adjustment 0: Study does not meet above criteria
Quality criterion 8: Adequate duration of follow-up	1: Study reports duration of follow-up and duration of follow-up adequate to identify expected adverse events (threshold set at 30 days for studies of carotid endarterectomy and 6 months for studies of rofecoxib) 0: Study does not meet above criteria
Total quality score=sum of scores (0-8)	>6: Good 4-6: Fair <4: Poor

Adapted from Meenan et al. Effectiveness and Cost-Effectiveness of Echocardiography and Carotid Imaging in the Management of Stroke. Evidence Report/Technology Assessment Number 49. (Prepared by Oregon Health & Science University Evidence-based Practice Center under Contract No. 290-97-0018.) AHRQ Publication No. 02-E022. Rockville, MD: Agency for Healthcare Research and Quality. July 2002.

Several criteria deserve comment regarding their application. "Non-biased selection" (criterion 1) was judged adequate for randomized controlled trials if they were properly randomized according to the widely-used Jadad definition.[17] For observational studies, non-biased selection was judged adequate if the study defined an inception cohort and attempted to evaluate all patients meeting the definition. Examples of studies that would not meet this criterion are studies that evaluated convenience samples or studies in which an inception cohort was not clearly defined. “Adverse events pre-specified” (criterion 4) was judged adequate if the study reported the adverse events being investigated and how an “event” was defined.

Exact application of several criteria varied according to the intervention being studied. Specifically, we defined “adequate duration of follow-up” (criterion 8) as at least 30 days for studies of CEA and at least 6 months for studies of rofecoxib, and “low loss to follow-up” (criterion 3) as <5 percent for studies of CEA and <10 percent for studies of rofecoxib. For studies of CEA, “non-biased ascertainment” (criterion 6) was defined as ascertainment performed by someone other than the surgeon who did the procedure. For trials of rofecoxib, we defined “non-biased ascertainment” as assessment by a blinded, external endpoint committee.[2] For “adequate statistical analysis of potential confounders” (criterion 7), studies of CEA had to assess at least one confounder to meet the criterion. Trials of rofecoxib had to analyze at least one confounder in addition to the treatment group (rofecoxib versus control).

Data Abstraction

Two investigators (R.C. and S.C.) independently applied the quality criteria and abstracted data for all studies. When quality assessments differed, consensus was reached before assigning final ratings. All quality assessments were made before complication rates were abstracted from each study.

We also abstracted the proportion of patients with cardiovascular comorbidities, study-level demographic characteristics (gender, race, and mean age), year of publication, funding source, setting, use of blinding (for the rofecoxib studies), and other study design features (such as randomized controlled trial or observational study, prospective or retrospective design, population-based or not population-based) in order to assess their impact on estimates of adverse events. We also abstracted variables that might be proxies for study quality such as author category (based on reported departmental and institutional affiliations),[24] whether the study reported severity of adverse events, the amount of text devoted to adverse events methods and reporting (recorded as the proportion of the total amount of text, excluding abstract and discussion),[12] and high (>7) Journal Impact Factor (based on 2003 data). Journal Impact Factor is a general measurement of journal quality[27] and is based on citation counts.[28] We selected a modest uniform threshold for classifying a journal as high Journal Impact Factor, though the interpretation of a specific value can vary depending on the journal’s target audience (for example, a general medical audience versus a subspecialty journal). Because most CEA studies either did not report volume of surgeries or surgeon experience, we simply rated studies according to whether they did or did not analyze any surgeon-related variable. Variables such as use of shunting or type of anesthesia were not analyzed because they were not associated with differential complication rates in previous studies.[29, 30] For studies of rofecoxib, the authors of a recent meta-analysis provided published and unpublished data on rates of myocardial infarction.[2] Because that meta-analysis found that the dose of rofecoxib, the type of control (placebo, non-naproxen nonsteroidal anti-inflammatories, or naproxen), and adequate allocation concealment were not associated with differences in risks for myocardial infarction, we did not analyze these variables.

Data Analysis

Chi-square tests or Fisher's exact tests (when cell values were less than five) were used to evaluate associations between individual quality criteria and other independent variables for each data set. Standard tests for heterogeneity were performed for all meta-analyses, with p values <0.10 considered significant.

Univariate analyses were performed to determine the association between individual methodological and clinical variables and adverse event rates. For each data set, random effects meta-regression models (in which between-study variance is incorporated into the analysis) were used to examine the effects of quality criteria and other study design and clinical or demographic factors on rates of stroke or death (studies of CEA) or odds ratios (OR) for myocardial infarction (trials of rofecoxib). The OR was evaluated as the dependent variable for trials of rofecoxib because of low absolute event rates (mean 0.3 percent). Data analyses for the rofecoxib trials were similar to analyses performed in a recently published meta-analysis.[2] Specifically, trials of rofecoxib with no events in both the rofecoxib and control group were excluded; control groups were combined; for any trial in which there were no events in either the rofecoxib or control group the OR was calculated by adding 0.5 to all cells;[31] and for trials with extensions, appropriate weighting was performed to avoid duplicate counting of data.

We calculated an initial summary quality score by adding up the scores of the eight criteria and assigned an initial overall quality rating based on the following arbitrary cutoffs: good=>6, fair=4-6, poor=<4. The univariate analyses and tests of association were used to develop more parsimonious quality rating instruments by removing non-predictive quality criteria, or those that did not improve the performance of quality rating instruments. Quality rating instruments were analyzed using different cutoffs for good, fair, or poor (three categories); adequate or inadequate (two categories); and the raw quality score as predictors of adverse events.

Multivariate meta-regression models using individual quality rating criteria or summary quality rating instruments and other clinical or study design-related variables were compared for goodness-of-fit using Akaike's Information Criterion (AIC),[32] corrected AIC (AICC),[33] and the Schwarz-Bayesian Information Criterion (BIC).[34] A smaller value of these criteria indicates a better fit, with a difference of 3 to 7 indicating an important difference between models.[35] Variables initially entered into the multivariate models were chosen based on the univariate analyses. Additional model selection was performed using backwards elimination methods. Variables that were highly correlated with quality ratings according to tests of association were preferentially removed during the model selection process, but we assessed their effects by adding them back to the final models.

When possible, subgroup analyses were performed to better control for the effects of clinical factors on complications. Additional analyses were also performed to assess the association between the raw quality score and complication rates using the baseline quality rating instrument and more parsimonious quality rating instruments.

All analyses were performed with SAS 9.1 (SAS Institute Inc., Cary, NC, U.S.A.).

Chapter 3: Results

Characteristics of Studies of CEA for Symptomatic Stenosis

We included 111 studies of CEA for symptomatic stenosis. Nine were randomized trials. Of 102 observational studies, 16 were population-based (attempted to assess all patients undergoing CEA in a pre-defined population). Eight of the 111 studies were published in journals with a high (>7) Journal Impact Factor. Eighty-two studies were performed in North America and 23 in Europe. Ten had a single surgeon author, 53 had multiple surgeon authors, and 48 had at least one non-surgeon author. Many studies did not report cardiovascular risk factors. For example, the proportion of patients with diabetes, the most frequently reported co-morbidity, was reported in 57 percent of studies.

Tests of Association

Three quality criteria were very highly ($p<0.001$) associated with one another (Table 2): criterion 4 (adverse events pre-defined), criterion 5 (ascertainment technique adequate described), and criterion 6 (non-biased ascertainment). Overall quality rating (good, fair, or poor) was highly associated ($p<0.0001$) with author category (single surgeon, multiple surgeons, or at least one non-surgeon), funding source (mostly government, not reported, or other), and high (>7) Journal Impact Factor.

Table 2. Tests of association (p values shown*) for discrete independent variables in studies of carotid endarterectomy for symptomatic stenosis

	Quality criterion 1: Non-biased selection (yes or no)	**Quality criterion 2: Adequate description of population (yes or no)**	**Quality criterion 3: Low (<5%) loss to follow-up (yes or no)**	**Quality criterion 4: Adverse events pre-defined (yes or no)**	**Quality criterion 5: Ascertainment technique adequately described (yes or no)**	**Quality criterion 6: Non-biased ascertainment (yes or no)**	**Quality criterion 7: Adequate duration (>30 days) of follow-up (yes or no)**	**Quality criterion 8: Statistical analysis of confounders (yes or no)**	**Quality rating (poor, fair, or good)**
Quality criterion 1: Non-biased selection	—	0.6146	**0.5536**	**0.0015**	0.0008	0.0031	**0.1387**	0.0024	<.0001
Quality criterion 2: Adequate description of population	0.6146	—	1.000	0.2392	0.9580	0.4705	0.2358	0.8882	0.0128
Quality criterion 3: Low (<5%) loss to follow-up	**0.5536**	1.000	—	**0.5584**	1.00	1.00	**0.5774**	0.0518	0.4260
Quality criterion 4: Adverse events pre-specified and defined	**0.0015**	0.2392	**0.5584**	—	<.0001	<.0001	**0.0832**	0.0342	<.0001
Quality criterion 5: Ascertainment technique adequately described	0.0008	0.9580	1.00	<.0001	—	<.0001	0.0012	0.0085	<.0001
Quality criterion 6: Non-biased ascertainment	0.0031	0.4705	1.00	<.0001	<.0001	—	0.0092	0.0273	<.0001
Quality criterion 7: Statistical analysis of confounders	0.0024	0.8882	0.0518	0.0342	0.0085	0.0273	0.2116	—	0.0001
Quality criterion 8: Adequate duration (>30 days) of follow-up	**0.1387**	0.2358	**0.5774**	**0.0832**	0.0012	0.0092	—	0.2116	<.0001
Quality rating (poor, fair, or good)	<.0001	0.0128	0.4260	<.0001	<.0001	<.0001	<.0001	<.0001	—

* p values reported for Fisher's exact test when there were less than 5 events in 1 or more cells; otherwise p values for chi-square tests.
Bold values indicate associations between quality criteria in final 4-criteria quality rating instrument.

Table 2. Tests of association (p values shown*) for discrete independent variables in studies of carotid endarterectomy for symptomatic stenosis (continued)

	Author category (single surgeon author, multiple surgeons, or at least one non-surgeon)	Setting (North America, Europe, or other)	Funding source (pharmaceutical, mostly government, other, or unclear)	Prospective study (yes or no)	Population-based study (yes or no)	Randomized controlled trial (yes or no)	High (>7) Journal Impact Factor (yes or no)	Adverse events discussed by severity (yes or no)	Discussion of hospital or surgical volume or experience (yes or no)
Quality criterion 1: Non-biased selection	0.0723	0.9582	0.0024	0.0011	0.2849	0.0250	0.0252	0.6173	0.0592
Quality criterion 2: Adequate description of population	0.1170	0.7687	0.2090	0.5774	0.7164	0.2651	0.0173	0.8906	0.7378
Quality criterion 3: Low (<5%) loss to follow-up	1.00	1.00	0.3762	0.5584	1.00	1.00	1.00	0.5984	1.0000
Quality criterion 4: Adverse events pre-specified and defined	0.0023	0.6042	0.0149	0.6669	0.0077	0.0868	0.0002	0.6602	0.0035
Quality criterion 5: Ascertainment technique adequately described	<.0001	0.3734	<.0001	0.5769	<.0001	0.0267	0.0012	0.2807	0.0009
Quality criterion 6: Non-biased ascertainment	<.0001	0.7551	<.0001	0.1525	0.0004	0.0022	<.0001	1.00	<.0001
Quality criterion 7: Statistical analysis of confounders	0.0236	0.0358	0.1087	0.9061	0.0239	0.7277	0.7078	0.0974	0.0135
Quality criterion 8: Adequate duration (>30 days) of follow-up	0.1202	0.4790	0.003	0.0498	0.5554	0.0050	0.0114	0.8342	0.9684
Quality rating (poor, fair, or good)	<.0001	0.7060	<.0001	0.0413	0.0006	0.0501	<.0001	0.3064	0.0026

* p values reported for Fisher's exact test when there were less than 5 events in 1 or more cells; otherwise p values for chi-square tests

Bold values indicate associations between quality criteria in final 4-criteria quality rating instrument

Univariate Analyses

For five of the eight individual quality items (non-biased selection, low loss to follow-up, adverse events pre-specified and defined, ascertainment technique adequately described, and non-biased and accurate ascertainment of adverse events), meeting the criterion was significantly (p<0.05) associated with differences in rates of stroke or death (Table 3). Criteria that were not associated with significant differences in estimates of stroke or death were criterion 2 (adequate description of population), criterion 7 (statistical analysis of confounders), and criterion 8 (adequate duration of follow-up).

Other variables associated with significantly higher rates of stroke or death were author category (studies with multiple authors and at least one non-surgeon having the highest rates), prospective study, population-based study, randomized controlled trial, high Journal Impact Factor, analysis of surgeon-related variables, and the patient population variables mean age and the proportion with coronary artery disease (Tables 3 and 4).

Table 3. Univariate analyses of rates of stroke or death associated with carotid endarterectomy for symptomatic stenosis—discrete variables

Independent variable	Study category	Number of studies	Pooled rate of stroke or death (95% confidence interval) in univariate analyses	p value for difference in complication rates
Quality criterion 1: Non-biased selection	1: adequate	70	5.1% (4.4% to 5.8%)	p=0.0101
	0: inadequate	41	3.7% (2.9% to 4.5%)	
Quality criterion 2: Adequate description of population	1: adequate	67	4.6% (3.9% to 5.3%)	p=0.9024
	0: inadequate	44	4.5% (3.7% to 5.4%)	
Quality criterion 3: Low loss to follow-up, and patients lost to follow-up analyzed for adverse events	1: adequate	108	4.6% (4.1% to 5.2%)	p=0.0187
	0: inadequate	3	2.2% (0.3% to 4.2%)	
Quality criterion 4: Adverse events pre-specified and defined	1: adequate	43	5.9% (4.9% to 6.8%)	p=0.0006
	0: inadequate	68	3.8% (3.2% to 4.4%)	
Quality criterion 5: Ascertainment technique adequately described	1: adequate	35	5.7% (4.6% to 6.8%)	p=0.0082
	0: inadequate	76	4.0% (3.4% to 4.6%)	
Quality criterion 6: Non-biased and accurate ascertainment of adverse events	1: adequate	32	5.7% (4.6% to 6.9%)	p=0.0136
	0: inadequate	79	4.1% (3.5% to 4.7%)	
Quality criterion 7: Adequate statistical analysis of potential confounders	1: adequate	69	4.8% (4.1% to 5.5%)	p=0.1509
	0: inadequate	42	4.0% (3.2% to 4.9%)	
Quality criterion 8: Adequate duration of follow-up	1: adequate	48	5.0% (4.1% to 5.9%)	p=0.1877
	0: inadequate	63	4.2% (3.5% to 4.9%)	
Author category	Single author and surgeon	10	2.8% (1.7% to 3.8%)	p=0.0327 (single versus multiple surgeons) p<0.0001 (single surgeon versus non-surgeon)
	Multiple authors, all surgeons	53	4.2% (3.5% to 4.9%)	P=0.0181 (multiple surgeons versus non-surgeon)
	At least one non-surgeon author	48	5.6% (4.6% to 6.5%)	

Table 3. Univariate analyses of rates of stroke or death associated with carotid endarterectomy for symptomatic stenosis—discrete variables (continued)

Methodological variable	Study category	Number of studies	Pooled rate of stroke or death (95% confidence interval) in univariate analyses	p value for difference in complication rates

Setting	North America or Canada	82	4.2% (3.6% to 4.8%)	p=0.0771 (North America versus Europe);
				p=0.2395 (North America versus other)

	Europe	23	5.7% (4.2% to 7.3%)	p=0.8560 (Europe versus other)
	Other, not reported or unclear, or international	6	6.0% (3.1% to 9.0%)	
Funding source	Mostly government	14	5.3% (3.6% to 6.9%)	p=0.6221 (government versus other); p=0.2479 (government versus not reported)
	Other	15	5.9% (4.1% to 7.6%)	p=0.0810 (other versus not reported)
	Not reported or unclear	82	4.2% (3.6% to 4.8%)	
Prospective study	Yes	31	5.2% (4.0% to 6.4%)	p=0.1851
	No	80	4.3% (3.7% to 5.0%)	
Population-based study	Yes	16	6.3% (4.6% to 8.0%)	p=0.0259
	No	95	4.3% (3.7% to 4.8%)	
Randomized controlled trial	Yes	9	7.4% (4.5% to 10.2%)	p=0.0444
	No	102	4.4% (3.8% to 4.9%)	
Published in high Journal Impact Factor journal	Journal Impact Factor >7	8	7.8% (4.8% to 10.8%)	p=0.0262
	Journal Impact Factor <=7	103	4.3% (3.8% to 4.9%)	
Reports adverse events by severity	Yes	58	4.8% (4.0% to 5.6%)	p=0.5281
	No	51	4.4% (3.7% to 5.2%)	
Analysis of hospital or surgeon volume or surgeon experience	Yes	21	6.2% (4.7% to 7.7%)	p=0.0150
	No	90	4.2% (3.6% to 4.7%)	

Table 4. Univariate analyses of rates of stroke or death in studies associated with carotid endarterectomy for symptomatic stenosis—continuous variables

Demographic or risk factor variable	Number of studies with available data	Odds ratio for risk of stroke or death per unit change in independent variable (95% confidence interval)	p value
Proportion of patients who smoke	45	0.6137 (0.2087 to 1.8049)	0.3667
Mean age of patients in study	83	0.9671 (0.9356 to 0.9997)	0.0479
Proportion of patients who are male	87	1.0077 (0.9905 to 1.0252)	0.3769
Proportion of patients who are white	16	0.6741 (0.3626 to 1.2533)	0.1953
Proportion of patients with diabetes	63	0.7514 (0.1147 to 4.9244)	0.7622
Proportion of patients with hypertension	61	0.8433 (0.8433 to 3.1001)	0.7943
Proportion of patients with coronary artery disease	50	0.1645 (0.04864 to 0.5563)	0.0045
Proportion of space devoted to discussion of adverse event assessment methods (cm of text/total text, excluding discussion section)	111	1.0369 (1.0144 to 1.0598)	0.0014
Proportion of space devoted to discussion of adverse event assessment results (cm of text/total text, excluding discussion section)	111	1.0124 (0.9991 to 1.0258)	0.0684

Developing a Quality Rating Instrument

Using all eight quality criteria, reported rates of stroke or death were highest for studies rated good-quality (pooled rate 6.3 percent, 95% CI 4.6 percent to 8.0 percent), intermediate for fair-quality studies (5.3 percent, 95% CI 4.1 percent to 6.5 percent), and lowest for poor-quality studies (3.8 percent, 95% CI 3.2 percent to 4.4 percent); the differences were statistically significant for good versus poor (p=0.0076) and fair versus poor (p=0.0289), though not for good versus fair studies (p=0.3557).

To develop a more parsimonious quality rating instrument, we removed criterion 2 (adequate description of population), which was not associated with differential adverse event rates, and criteria 5 and 6, which were highly associated with (but not as predictive as) criterion 4 (adverse events pre-defined). After comparing all possible 4-item instruments using the remaining 5 criteria, we found that a quality rating instrument with criteria 1 (non-biased selection), 3 (low loss to follow-up), 4 (adverse events pre-defined), and 8 (adequate duration of follow-up) performed similarly to the 5-criteria instrument (Table 5). We also found that quality rating instruments using a single cutoff (adequate or inadequate) performed similarly to instruments using multiple cutoffs (good, fair, or poor). Using the 4-criteria instrument, studies rated "adequate" (score >=3 out of 4) had a significantly (p=0.003) higher pooled rate of stroke or death of 5.7 percent (95% CI 4.8 percent to 6.6 percent) compared to 3.7 percent (95% CI 3.1 percent to 4.3 percent) for studies rated "inadequate" (score <3). The summary quality rating also predicted rates of surgical complications better than any individual quality criterion.

For the four criteria instrument, a steady increase in the pooled complication rate was observed with increasing quality scores (scores of 1, 2, 3, and 4 associated with rates of 3.1 percent [95% CI 2.2 percent to 3.9 percent], 4.0 percent [95% CI 3.2 percent to 4.8 percent], 5.6 percent [4.4 percent to 6.7 percent], and 6.0 percent [4.5 percent to 7.4 percent], respectively). For instruments with more criteria, higher scores were associated with higher pooled complication rates until a threshold was reached at scores >=4 (Table 6).

Table 5. Analysis of different quality rating instruments on pooled rates of stroke or death associated with carotid endarterectomy for symptomatic stenosis

Quality rating instrument	Score	Number of studies	Pooled rate of stroke or death (95% confidence interval) in univariate analyses	Significance of differences in pooled complication rates	Goodness-of-fit
Quality instrument using 8 criteria and baseline cutoffs for good, fair, or poor	>6: good	16	6.3% (4.6% to 8.0%)	p=0.0074 (good vs. poor), p=0.3530 (good vs. fair)	AIC=755.0, AICC=755.4, BIC=765.9
	>4 and <=6: fair	27	5.3% (4.2% to 6.5%)	p=0.0290 (fair vs. poor)	
	<=4: poor	68	3.9% (3.3% to 4.5%)		
Quality instrument using 8 criteria and 'best' cutoffs for good, fair, or poor	>5: good	27	6.1% (4.8% to 7.4%)	p=0.0002 (good vs. poor), p=0.0187 (good vs. fair)	AIC=753.0, AICC=753.3, BIC=763.8
	>2 and <=5: fair	64	4.4% (3.7% to 5.1%)	p=0.0207 (fair vs. poor)	
	<=2: poor	20	3.0% (2.1% to 4.0%)		
Quality instrument using 7 criteria (excludes criterion 2, 'adequate description of population') and 'best' cutoffs for good, fair, or poor	>3: good	54	5.6% (4.8% to 6.5%)	p<0.0001 (good vs. poor), p=0.0024 (good vs. fair)	AIC=747.8, AICC=748.2, BIC=758.7
	>1 and <=3: fair	43	3.9% (3.2% to 4.6%)	p=0.0249 (fair vs. poor)	
	<=1: poor	14	2.5% (1.5% to 3.5%)		
Quality instrument using 7 criteria (excludes criterion 2) and 'best' cutoffs for adequate or inadequate	>=4: adequate	54	5.6% (4.8% to 6.5%)	p=0.0002	AIC=750.0, AICC=750.3, BIC=758.2
	<4: inadequate	57	3.6% (3.0% to 4.2%)		
Quality instrument using 5 criteria (1, 3, 4, 7, and 8) and 'best' cutoffs for adequate or inadequate	>=4: adequate	39	5.9% (4.8% to 6.9%)	p=0.0005	AIC=752.8, AICC=753.1, BIC=761.0
	<4: inadequate	72	3.9% (3.3% to 4.4%)		
Quality instrument using 4 criteria (1, 3, 4, and 8) and 'best' cutoffs for adequate or inadequate	>=3: adequate	50	5.7% (4.8% to 6.6%)	p=0.0003	AIC=750.4, AICC=750.7, BIC=758.6
	<3: inadequate	61	3.7% (3.1% to 4.3%)		

Table 6. Pooled rate of stroke or death in studies of carotid endarterectomy for symptomatic stenosis according to number of quality criteria met

Quality score	Number of studies	Pooled rate of stroke or death (95% confidence interval)
7 criteria quality rating instrument (criteria 1, 3-8)		
1	14	2.5% (1.5% to 3.5%)
2	21	4.1% (2.9% to 5.2%)
3	22	3.8% (2.8% to 4.7%)
4	21	5.3% (4.0% to 6.6%)
5	11	5.2% (3.9% to 6.6%)
6	10	5.3% (3.9% to 6.6%)
7	12	5.2% (3.8% to 6.6%)
5 criteria quality rating instrument (criteria 1, 3, 4, 7, 8)		
1	15	2.6% (1.6% to 3.5%)
2	22	4.1% (3.0% to 5.3%)
3	35	4.2% (3.4% to 5.0%)
4	21	6.0% (4.6% to 7.4%)
5	18	6.0% (4.5% to 7.5%)
4 criteria quality rating instrument (criteria 1, 3, 4, 8)		
1	23	3.1% (2.2% to 3.9%)
2	38	4.0% (3.2% to 4.8%)
3	30	5.6% (4.4% to 6.7%)
4	20	6.0% (4.5% to 7.4%)
3 criteria quality rating instrument (criteria 3, 4, 8)		
1	45	3.4% (2.7% to 4.1%)
2	44	5.1% (4.2% to 6.0%)
3	22	5.9% (4.6% to 7.3%)

Inter Rater Reliability

The overall quality rating (adequate or inadequate) using the four-criteria instrument was in agreement between two investigators for 19/20 of a random selection of studies (Kappa = 0.90).

Multivariate Regression Analyses

A model with four individual quality criteria (1, 3, 4, and 8) performed similarly to a model using the composite four-criteria quality rating instrument that categorized studies as “adequate” or “inadequate” (AICC goodness-of-fit test 751.4 vs. 750.7). An “adequate” rating using the four-item instrument was an independent predictor of the reported adverse event rate after adjustment for North American setting (associated with lower complication rates), randomized controlled trial or population-based study design (each associated with higher complication rates), and the proportion of total text devoted to reporting adverse event results (higher proportion associated with higher complication rates) (Table 7). The four-criteria instrument performed similarly as a binary (adequate versus inadequate) or continuous (quality score 0-4) variable (AICC goodness of fit statistic 738.1 versus 738.6).

Table 7. Final multivariate model for rates of stroke or death in studies of carotid endarterectomy for symptomatic stenosis

Model	Independent variables initially entered into model	Independent variables in final model	Effect size (ln odds ratio)	p value	Goodness of fit statistics for model
Multivariate model using 4 criteria quality rating instrument	study rated 'adequate' (score >=3)	study rated 'adequate' (score >=3)	0.3038	p=0.0130	AIC=737.1, AICC=738.1, BIC=756.0
	setting: North America	setting: North America	-0.4011	p=0.0034	
	setting: Europe				
	randomized controlled trial	randomized controlled trial	0.5361	p=0.0140	
	population-based study	population-based study	0.4192	p=0.0058	
	proportion of text devoted to reporting adverse event results (continuous variable)	proportion of text devoted to reporting adverse event results (continuous variable)	0.01505	p=0.0167	
	proportion of text devoted to reporting adverse event methods (continuous variable)				
	prospective study				
	Journal Impact Factor				
	adverse events reported by severity				
	surgical factors discussed (hospital or surgeon volume or experience)				
	funding source not reported				
	funding source governmental				

AIC=Akaike's Information Criterion
AICC=corrected Akaike's Information Criterion
BIC=Schwarz-Bayesian Information Criterion

Adding high (>7) Journal Impact Factor or the variable “single surgeon author” to the multivariate model did not improve the model. No other variables were significant in multivariate models, though incomplete reporting in the primary studies limited analysis of demographic factors and comorbidities.

Subgroup Analyses

An “adequate” quality rating (using the four-criteria instrument) was associated with higher reported complication rates across most clinical subgroups of patients, including patients with transient ischemic attacks (5.3 percent vs. 3.6 percent, p=0.0783), patients undergoing early (<3 to 6 weeks after symptoms) CEA (9.7 percent vs. 3.1 percent, p=0.0053), asymptomatic patients (reported separately from patients with symptoms) (3.2 percent vs. 1.4 percent, p=0.0021), and stroke patients (6.8 percent vs. 5.9 percent, p=0.3681). In multivariate analyses limited to the studies that reported mean age or proportion of patients with coronary artery disease, an “adequate” quality rating remained significant when controlling for those factors (Table 8). An “adequate” rating also predicted higher complication rates in univariate analyses of population-based observational studies (7.1 percent vs. 4.4 percent, p=0.0424) and non-population based observational studies (5.0 percent vs. 3.4 percent, p=0.0071), as well as in a multivariate model using data from all observational studies (excluding randomized trials).

Table 8. Multivariate models for rates of stroke or death in subgroups of studies of carotid endarterectomy for symptomatic stenosis

Model	**Number of studies**	**Independent variables in final model**	**Effect size (in odds ratio)**	**p value**
Multivariate model limited to studies reporting mean age	83	study rated ‘adequate’ (score >=3)	0.4151	p=0.0022
		setting: North America	-0.3708	p=0.0137
		randomized controlled trial	0.4568	p=0.0562
		population-based study	0.3195	p=0.0685
		proportion of text devoted to reporting adverse event results (continuous variable)	0.01444	p=0.00776
		mean age in years (continuous variable)	-0.02681	p=0.1004
Multivariate model limited to studies reporting proportion of patients with coronary artery disease	50	study rated ‘adequate’ (score >=3)	0.3395	p=0.0135
		setting: North America	-0.3618	p=0.0258
		randomized controlled trial	dropped from model (p>0.10)	
		population-based study	dropped from model (p>0.10)	
		proportion of text devoted to reporting adverse event results (continuous variable)	0.01441	p=0.0278
		proportion of patients with coronary artery disease (continuous variable)	-0.01380	p=0.0182

Studies of CEA for Asymptomatic Stenosis

We included 18 studies of CEA for asymptomatic stenosis. Six were randomized trials, 15 were set in North American, and none were population-based studies. Rates of stroke and death were low in these studies, ranging from 0 percent to 4.6 percent.

None of the eight quality criteria (Table 9), quality rating instruments, or other clinical or study-design related variables were associated with significant differences in rates of stroke or death. Using the final four-criteria instrument developed based on the analyses of studies of CEA for symptomatic stenosis, rates of stroke or death were 2.4 percent and 2.5 percent in studies rated adequate and inadequate, respectively. In fact, little correlation was observed between quality scores and complication rates. For example, 7seven studies that only met one or two of the eight quality criteria reported complication rates ranging from 1.4 percent to 4.6 percent. Two studies that met seven or eight of the quality criteria reported complication rates of 1.3 percent and 3.0 percent.

Because no variables were significant in univariate analyses, we did not attempt multivariate regression analyses.

Table 9. Univariate analyses of association between quality criteria and rates of stroke or death in studies of carotid endarterectomy for asymptomatic stenosis

Independent variable	Score	Number of studies	Pooled rate of stroke or death (95% confidence interval)	p value for difference in complication rates
Quality criterion 1: Non-biased selection	1: adequate	10	2.4% (1.8% to 3.1%)	0.9763
	0: inadequate	8	2.4% (1.4% to 3.5%)	
Quality criterion 2: Adequate description of population	1: adequate	13	2.4% (1.8% to 3.0%)	0.7304
	0: inadequate	5	2.6% (1.2% to 4.1%)	
Quality criterion 3: Low loss to follow-up and patients lost to follow-up analyzed for adverse events	1: adequate	18	2.4% (1.8% to 3.0%)	--
	0: inadequate	0	--	
Quality criterion 4: Adverse events pre-specified and defined	1: adequate	6	2.2% (1.4% to 3.0%)	0.3551
	0: inadequate	12	2.7% (1.8% to 3.6%)	
Quality criterion 5: Ascertainment technique adequately described	1: adequate	6	2.6% (1.8% to 3.4%)	0.4132
	0: inadequate	12	2.2% (1.4% to 3.0%)	
Quality criterion 6: Non-biased and accurate ascertainment of adverse events	1: adequate	6	2.6% (1.8% to 3.4%)	0.4132
	0: inadequate	12	2.2% (1.4% to 3.0%)	
Quality criterion 7: Adequate statistical analysis of potential confounders	1: adequate	5	2.4% (1.7% to 3.2%)	0.9710
	0: inadequate	13	2.4% (1.6% to 3.3%)	
Quality criterion 8: Adequate duration of follow-up	1: adequate	7	2.4% (1.7% to 3.2%)	0.9463
	0: inadequate	11	2.4% (1.6% to 3.2%)	

Studies of Rofecoxib

We included 16 randomized controlled trials of rofecoxib in patients with arthritis. Two studies published only as abstracts were excluded.[36, 37] Two trials were published in journals with high (>7) Journal Impact Factor, four were published since 2001, and two were written by authors from a single department and institution. Pharmaceutical companies funded all of the trials.

In univariate analyses of the eight quality criteria, only criteria 6 (blinded external endpoint committee; pooled OR 3.69 versus 0.68, p<0.0001) and 7 (adequate statistical analysis of confounders; pooled OR 4.99 vs. 1.39, p=0.0164) were significantly associated with a higher risk for myocardial infarction (Table 10). Studies published in high Journal Impact Factor journals were the same as the studies that met criterion 7. Presence of an external endpoint committee blinded to treatment allocation was a stronger predictor of differences in risk for myocardial infarction when its presence was assessed using published or unpublished data than when relying only on published information (pooled OR 3.83 vs. 1.37, p=0.0468). Blinded assessment of complications (defined as assessor not aware of treatment allocation, but not necessarily a review committee independent from the study) was not associated with a significantly lower risk of myocardial infarction (pooled OR 1.68 vs. 3.94, p=0.2134).

Table 10. Univariate analyses of risk for myocardial infarction in randomized controlled trials of rofecoxib

Independent variable	Score	Number of studies	Pooled odds ratio for myocardial infarction, rofecoxib versus control	p value for difference in odds ratios
Quality criterion 1: Non-biased selection	1: adequate	6	2.19 (0.93 to 5.15)	0.7212
	0: inadequate	10	1.80 (0.96 to 3.35)	
Quality criterion 2: Adequate description of population	1: adequate	8	1.97 (1.02 to 3.78)	0.9118
	0: inadequate	8	1.85 (0.83 to 4.13)	
Quality criterion 3: Low loss to follow-up and patients lost to follow-up analyzed for adverse events	1: adequate	8	2.19 (1.16 to 4.12)	0.5313
	0: inadequate	8	1.56 (0.69 to 3.51)	
Quality criterion 4: Adverse events pre-specified and defined	1: adequate	4	2.64 (1.19 to 5.84)	0.3417
	0: inadequate	12	1.59 (0.85 to 2.96)	
Quality criterion 5: Ascertainment technique adequately described	1: adequate	10	2.10 (0.82 to 5.38)	0.8282
	0: inadequate	6	1.85 (1.02 to 3.38)	
Quality criterion 6: Non-biased and accurate ascertainment of adverse events	1: adequate	9	3.69 (2.61 to 5.20)	<0.0001
	0: inadequate	7	0.68 (0.42 to 1.09)	
Quality criterion 7: Adequate statistical analysis of potential confounders	1: adequate	2	4.99 (2.28 to 10.93)	0.0164
	0: inadequate	14	1.39 (0.86 to 2.25)	
Quality criterion 8: Adequate duration of follow-up	1: adequate	7	1.69 (0.85 to 3.33)	0.5874
	0: inadequate	9	2.24 (1.07 to 4.71)	
Publication year	After 2001	4	2.71 (0.99 to 7.40)	0.4563
	Before or during 2001	12	1.72 (0.97 to 3.06)	
Setting	North America	8	1.41 (0.63 to 3.15)	0.4306
	Other	6	2.22 (1.05 to 4.70)	

Table 10. Univariate analyses of risk for myocardial infarction in randomized controlled trials of rofecoxib (continued)

Independent variable	Score	Number of studies	Pooled odds ratio for myocardial infarction, rofecoxib versus control	p value for difference in odds ratios
Blinded outcomes assessment	Study had blinded assessment of complications.	13	1.68 (0.99 to 2.83)	0.2134
	Study not adequately blinded or unclear.	3	3.94 (1.22 to 12.72)	
External endpoint committee (published data only)	Study had an external endpoint committee according to published data.	3	3.83 (1.81 to 8.07)	0.0468
	Study does not meet above criteria or unclear.	13	1.37 (0.80 to 2.36)	
Published in high Journal Impact Factor journal	Journal Impact Factor >7	2	4.99 (2.28 to 10.93)	0.0164
	Journal Impact Factor <=7	14	1.39 (0.86 to 2.25)	
Author category	Multiple institutions or specialties	14	2.11 (1.28 to 3.49)	0.2260
	Single specialty and institution	2	0.66 (0.12 to 3.72)	
Reports adverse events by severity	Yes	6	1.43 (0.63 to 3.27)	0.3947
	No	10	2.27 (1.23 to 4.21)	

Using the four criteria quality rating instrument from the analyses of studies of CEA for symptomatic stenosis, quality rating (adequate or inadequate) was not associated with significant differences in odds ratios for myocardial infarction with rofecoxib relative to control interventions (2.59 versus 1.49, p=0.28) or with absolute rates of myocardial infarction on rofecoxib (0.34 percent versus 0.31 percent).

Analyses of this data set were limited by the small number of studies, and by the fact that two studies[38, 39] with high odds ratios (4.98 and 5.00) for myocardial infarction contributed over half of the patients (13633 of 23725) in the included trials. Both of these trials also met seven out of the eight quality criteria.

Chapter 4: Discussion

We empirically evaluated the association between perceived methodological shortcomings and estimates of serious adverse events associated with clinical interventions. Our main results are based on a large set of studies of CEA for symptomatic stenosis. Although we also analyzed studies of CEA for asymptomatic stenosis and randomized controlled trials of rofecoxib for arthritis, those results are mainly hypothesis generating because of the low rate of serious adverse events and the small number of studies included in those data sets.

We found that certain pre-defined quality criteria predicted differences in pooled rates of stroke or death in randomized controlled trials, cohort studies, and uncontrolled surgical series of CEA for symptomatic stenosis, and remained predictive after controlling for other methodological and clinical variables through multiple regression or subgroup analyses. We are not aware of any other studies that have empirically tested a large number of quality criteria designed to measure shortcomings in the measurement or reporting of adverse events against actual estimates of harms.

We also found that it may be feasible to develop empirically validated quality rating instruments for assessing the validity of studies reporting harms. To our knowledge, this is the first quality rating instrument that has been developed using a data set that included both randomized trials and observational studies. We found that a summary quality rating instrument with four methodological criteria predicted adverse events associated with CEA for symptomatic stenosis as well as instruments with more criteria. The summary quality rating instrument predicted adverse events better than any individual criterion. We also found a dose-response relationship: the more criteria met, the higher the estimate of adverse events.

Each of the four criteria included in the final quality rating instrument may measure different aspects of adverse event assessment. Biased selection (criterion 1), for example, could systematically affect adverse event rates if patients who underwent the procedure but were excluded from analysis were more likely to have an adverse event. Patients lost to follow-up (criterion 3) would not be assessed for adverse events for the full duration of the study, and could also be at higher risk for adverse events, which could lead to attrition bias. Pre-specifying adverse events (criterion 4) suggests increased attention paid to adverse event assessment, and was highly associated with (but slightly more predictive than) two other criteria that may measure a similar characteristic: adequate description of ascertainment technique (criterion 5) and independent assessment (criterion 6). Adequate duration of follow-up (criterion 8) is important because studies that only evaluated patients until discharge from the hospital could miss complications that occurred within 30 days but after discharge.

Some of the pre-defined criteria that were excluded from the final quality rating instrument may not be associated with predictable biases in rates of adverse events, but could still reflect important aspects of adverse event assessment. For example, although inadequate description of population (criterion 2) would make it difficult to assess external validity, it could be associated with patients at either higher or lower risk for complications. Similarly, inadequate description of ascertainment technique (criterion 5) could be associated with systematic over- or under-reporting of adverse events, depending on the ascertainment technique used.

Three other findings from analyses of this set of studies deserve mention. First, high Journal Impact Factor,[27], author category,[24] and proportion of text devoted to reporting adverse event results appeared to be proxies for quality of adverse event assessment. Reporting bias could confound the association between author category and lower complication rates, as surgeons may be more apt to report good results. Second, adverse events were more frequent in randomized, controlled trials compared to observational studies. This could be because the effects of patient and surgeon selection[40] in randomized trials are offset by generally better adverse event assessment, or because observational studies are more likely than randomized controlled trials to go unpublished if study findings are unfavorable. Among observational studies, higher quality ratings predicted higher rates of adverse events—a factor that should be taken into account when comparing the results of randomized controlled trials and observational studies.[41-44] Third, population-based studies were associated with higher rates of complications than non-population-based studies, even when quality criteria were also considered. This could be because population-based studies are more representative of the entire population and different surgeons than other observational studies. Alternatively, population-based studies could be more effective in obtaining complete outcomes data through large databases.[45]

We were unable to replicate the associations between quality ratings and estimates of harms in a smaller data set of studies of CEA for asymptomatic stenosis. One possible explanation for these results is that these analyses had less power to detect differences related to quality because of the lower rate of complications, less variance in rates of complications between studies, and substantially fewer studies to analyze. In addition, factors pertaining to external validity (such as patient selection or factors related to the delivery of the intervention) could be an important source of variation in this set of studies but more difficult to adequately control for because of the smaller data set. An important finding is that even for the same intervention (CEA), the same quality criteria may not consistently predict adverse events when applied to studies evaluating different populations or settings.

For studies of rofecoxib, our findings were generally similar to a recent meta-analysis by Juni et al—namely, that the presence of an independent, external endpoint committee blinded to treatment allocation was the strongest predictor for a higher risk of myocardial infarction.[2] On the other hand, blinded outcomes assessment (not necessarily as part of an independent review committee) was not associated with higher estimates of risk. Appropriate allocation concealment, another factor commonly used to assess internal validity of clinical trials, did not predict reported risk of myocardial infarction in an earlier meta-analysis of rofecoxib trials.[2] These findings support the hypothesis that unique considerations in adverse event reporting may require a distinct set of criteria separate from those used to assess internal validity.

Our findings have several limitations. First, the only outcomes assessed were major adverse drug events and post-surgical complications. Applicability of the results to assessment of minor side effects and complications is unknown. Second, as in other studies, our assessment of methodological shortcomings primarily relied on information available in published reports. However, even though poor reporting and poor quality are often associated, they are not synonymous**.**[46] Inadequate reporting of adverse events methods can lead to misclassification, or assumptions that studies are methodologically deficient even when they were designed, conducted, and analyzed properly.[17] This is illustrated by the fact that the use of unpublished information to determine the presence of an external endpoint committee in trials of rofecoxib resulted in better predictions of myocardial infarction risk than determinations based on published reports alone. Appropriate methods for detecting *unsuspected* adverse events (such as myocardial infarction in earlier trials of rofecoxib) could be particularly susceptible to poor reporting. Third, important aspects for rating the quality of adverse event assessment may be difficult to measure using quality rating criteria. The use of accurate and precise ascertainment techniques, for example, is likely to be an important factor,[9] but difficult to define objectively. One possibility could be to distinguish between studies that used active techniques to identify adverse events versus those that used more passive methods. One recent study found that using a checklist to identify 53 possible adverse events resulted in identification of 20-fold more events compared to using open-ended questions.[47] However, the validity of using different methods for assessing adverse events had not yet been assessed. Fourth, publication bias could have distorted our conclusions, if either high-quality studies with lower estimates of harms or low-quality studies with higher estimates of harms were less likely to be published.

Quality criteria could also have differential predictive ability depending on whether relative or absolute measures are used to quantify harms. Relative measures such as odds ratios or relative risks are particularly important when comparing harms from different interventions. Absolute rates of adverse events, on the other hand, are helpful for quantifying the balance of harms and benefits associated with a particular intervention.[48] Methodological factors that affect estimates of one measure of harms may not affect the other. A factor that leads to systematic under-counting of adverse events and therefore affects the absolute rate, for example, might not significantly change the odds ratio if the bias affects both treatment groups similarly.

Most importantly, analyses need to be performed to determine whether methodological shortcomings are associated with lower estimates of harms for other surgical or drug interventions and outcomes. Like studies evaluating the effects of methodological shortcomings on estimates of efficacy from clinical trials,[49, 50] we found that developing a generic summary instrument for rating quality of adverse event assessment is problematic because different aspects of quality were more important for one set of studies compared to another. Specifically, major complications associated with CEA for symptomatic stenosis and rofecoxib for arthritis were predicted by different quality rating criteria, and no quality rating criteria predicted adverse events in studies of CEA for asymptomatic stenosis. It is therefore important for systematic reviewers to evaluate individual quality criteria when judging the quality of adverse event estimates, rather than relying on generic summary scales.

A key lesson from analyzing studies of harms is that in addition to doing a better job of looking for adverse events and measuring them reliably, it is also important for researchers to adequately evaluate and report factors that may influence complication rates.[13] Readers should carefully assess for potential sources of bias as well as other sources of variation (such as differences in populations and interventions[51]) when interpreting results of studies reporting harms. Future studies of this area are needed and should investigate data sets large enough to detect differences in adverse event rates, include studies utilizing both randomized and non-randomized designs, evaluate associations using absolute as well as relative event rates, and carefully examine the association between individual and summarized quality criteria and differential estimates of harms.

References

1. Cuervo LG, Aronson JK. The road to health care: balancing benefits and harms of interventions is essential. BMJ 2004;329(7456):1-2.

2. Juni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger PM. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. Lancet 2004;364(9450):2021-9.

3. 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(21):2622-31.

4. 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(9418):1341-5.

5. Alamowitch S, Eliasziw M, Barnett HJM. The risk and benefit of endarterectomy in women with symptomatic internal carotid artery disease. Stroke 2005;36(1):27-31.

6. Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?" Lancet 2005;365(9453):82-93.

7. Birkmeyer JD, Stukel TA, Siewers AE, Goodney PP, Wennberg DE, Lucas FL. Surgeon volume and operative mortality in the United States. N Engl J Med 2003;349(22):2117-27.

8. Derry S, Loke YK, Aronson JK. Incomplete evidence: the inadequacy of databases in tracing published adverse drug reactions in clinical trials. BMC Medical Research Methodology Vol 1: BMC Medical Research Methodology; 2001:Available at: http://www.biomedcentral.com/1471-2288/1/7.

9. 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(6):427-37.

10. Ethgen M, Boutron I, Baron G, Giraudeau B, Sibilia J, Ravaud P. Reporting of harm in randomized, controlled trials of nonpharmacologic treatment for rheumatic disease. Ann Intern Med 2005;143(1):20-25.

11. Loke YK, Derry S. Reporting of adverse drug reactions in randomised controlled trials—a systematic survey. BMC Clinical Pharmacology. Vol 1; 2001:Available at http://www.biomedcentral.com/1472-6904/1471/1473.

12. Ioannidis JPA, Lau J. Completeness of safety reporting in randomized trials. JAMA 2001;285(4):437-43.

13. Ioannidis JPA, Evans SJW, Gotzsche PC, et al. Better reporting of harms in randomized trials: an extension of the consort statement. Ann Intern Med 2004;141(10):781-8.

14. Loke Y, Price D, Herxheimer A. Including adverse effects in your review. Paper presented at Cochrane Colloquium: 2003; Barcelona.

15. Chou R, Helfand M. Challenges in systematic reviews that assess treatment harms. Ann Intern Med 2005;142(12 Pt 2):1090-9.

16. Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess. 2003;7:1-76.

17. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17(1):1-12.

18. Moher D, Jones A, Cook DJ, et al. Does quality of randomised trials affect estimates of intervention efficacy reported in meta-analyses. Lancet 1998;352:609-13.

19. Schulz KF, Chalmers I, Hayes RJ. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273(5):408.

20. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-6.

21. Horton R. Vioxx, the implosion of Merck, and aftershocks at the FDA. Lancet. 2004;364(9450):1995-6.

22. Meenan RT, Saha S, Chou R, et al. Effectiveness and cost-effectiveness of echocardiography and carotid imaging in the management of stroke. Rockville, MD: Agency for Healthcare Research and Quality; 2002.

23. 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(9):2290-303.

24. 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(2):260-5.

25. Rothwell PM, Slattery J, Warlow CP. A systematic comparison of the risks of stroke and death due to carotid endarterectomy for symptomatic and asymptomatic stenosis. Stroke 1996;27(2):266-9.

26. Benavente O, Moher D, Pham B. Carotid endarterectomy for asymptomatic carotid stenosis: a meta-analysis. BMJ. 1998;317:1477-80.

27. Lee KP, Schotland M, Bacchetti P, Bero LA. Association of journal quality indicators with methodological quality of clinical research articles. JAMA 2002;287(21):2805-8.

28. Garfield E. The history and meaning of the journal impact factor. JAMA 2006;295(1):90-3.

29. Bond R, Rerkasem K, Rothwell PM. Routine or selective carotid artery shunting for carotid endarterectomy (and different methods of monitoring in selective shunting). Stroke 2003;34(3):824-5.

30. Rerkasem K, Bond R, Rothwell PM. Local versus general anaesthesia for carotid endarterectomy. Cochrane Database Syst Rev 2004;2.

31. Pagano M, Gauvreau K. Principles of biostatistics. 2nd ed. Pacific Grove CA: Duxbury Press; 2000.

32. Akaike H. A new look at the statistical model identification. IEEE Transaction on Automatic Control 1974;AC-19:716-23.

33. Hurvich CM, Tsai CL. Regression and time series model selection in small samples. Biometrika 1989;76(2):297-307.

34. Schwarz G. Estimating the dimension of a model. Annals of Statistics 1978;6(2):461-4.

35. Burnham KP, Anderson DR. Model selection and inference: a practical information-theoretic approach. New York: Springer; 1998.

36. Geba GP, Polis AB, Najarian DK, Dixon ME, Storms WW, Weaver AL. Onset of efficacy and patient assessment of clinical response in osteoarthritis (OA): comparison of rofecoxib to nabumetone. J Am Geriatr Soc 2001;49:S126.

37. Truitt KE, Lee M, DeTora LM, Anderson M, Zhao Rahway PL. Results of a pivotal (phase iii) placebo and active comparator controlled efficacy trial of rofecoxib 12.5 and 25 mg in adult patients with rheumatoid arthritis (ra). Arthritis Rheum. 2001;44(S9):S369.

38. Bombardier C, Laine L, Reicin A, 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(21):1520-8.

39. Lisse JR, Perlman M, Johansson G, et al. Gastrointestinal tolerability and effectiveness of rofecoxib versus naproxen in the treatment of osteoarthritis: a randomized, controlled trial. Ann Intern Med 2003;139(7):539-46.

40. Barnett HJM. Carotid endarterectomy. Lancet. 2004;363(9420):1486-7.

41. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342(25):1878-86.

42. 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(13).

43. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342(25):1887-92.

44. Ioannidis JPA, Haidich A-B, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001;286(7):821-30.

45. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005;58:323-37.

46. Huwiler-Muntener K, Juni P, Junker C, Egger M. Quality of reporting of randomized trials as a

measure of methodologic quality. JAMA 2002;287(21):2801-4.

47. Bent S, Padula A, Avins AL. Better ways to question patients about adverse medical events. A randomized, controlled trial. Ann Intern Med 2006;144:257-61.

48. Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. From subgroups to individuals: general principles and the example of carotid endarterectomy. Lancet. 2005;365(9455):256-65.

49. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999;282(11):1054-60.

50. Balk EM, Bonis PAL, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002;287:2973-82.

51. Finlayson EVA, Birkmeyer JD, Stukel TA, Siewers AE, Lucas FL, Wennberg DE. Adjusting surgical mortality rates for patient comorbidities: More harm than good? Surgery. 2002;132(5):787-94.