Rethinking the “Whodunnit” Approach to Assessing the Quality of Health Care Research – A Call to Focus on the Evidence in Evidence-Based Practice

Kathleen A. Fairman, MA, and Frederic R. Curtiss, PhD, RPh, CEBS

When we wrote our March, 2008 editorial in *JMCP* regarding the lack of evidence of value in value-based insurance design (VBID),¹ we correctly anticipated that those working in the “trenches” of managed care pharmacy would appreciate an objective evaluation of the evidence about prescription drug cost-sharing. What we did not anticipate was a query we received from a consultant about our real reasons for writing the editorial. He asked about our business interest in the topic and why we had not disclosed the financial relationships that must have prompted our work. It was apparently inconceivable that 2 journal editors would closely examine the research evidence underlying a health care management tool for no reason other than an interest in determining the quality of the evidence. Apparently equally unthinkable was the fact that we had no conflicts of interest to disclose.

We wonder if the cynical perspective evident in the question posed to us is unavoidable. We work in an industry in which most research is funded by parties with a commercial interest in the study outcomes. It is easy to conclude that a relationship between economic interest and research findings is inevitable. Adriane Fugh-Berman, author of numerous books and articles about complementary medicine, tells a similar tale: “When a university or medical society invites me to speak, I am often asked which company sponsors my talks. Not whether a company sponsors me, but which company.”²

Why So Cynical? Widely Publicized “Scientists Behaving Badly”³

Speculation about the relationship between business interest and the conduct of research is understandable given the media attention paid to financially-motivated unethical behavior – real or alleged – in the medical research enterprise in the past few years. Richard Smith, who edited the *British Medical Journal* from 1979 to 2004, described medical journals as “an extension of the marketing arm of pharmaceutical companies in publishing trials that [favor] their products.”⁴ In April, 2008, reports of excessive pharmaceutical manufacturer influence over publications in several major research journals prompted the editor of the *Journal of the American Medical Association (JAMA)* to describe the “manipulation” of the research literature as “disgusting.”⁵⁻⁷

Eyebrows were raised when Jeffrey Lisse, the first author of the Assessment of Differences between Vioxx and Naproxen to Ascertain Gastrointestinal Tolerability and Effectiveness

(ADVANTAGE) trial, (*Annals of Internal Medicine*, 2003), told a *New York Times* reporter in 2005 that he had not actually had any role in the trial but had merely lent his name to a manuscript written by the manufacturer.⁸ In response, an editorial published in *Annals* in 2005 declared that the frank admission of ghostwriting had “sent more than a few shivers up the spines of the editors.”⁸ Sismondo described such practices as “ghost management” (i.e., marketing masquerading as science) of research articles in the medical literature.⁹ However, the raised eyebrows were later singed when new allegations emerged from the rofecoxib product liability litigation. Paid consultants for the plaintiffs contended that the manufacturer of rofecoxib not only ghost-managed the research and the ensuing publication but did so in a business practice known as “seeding,” in which a drug manufacturer funds a clinical trial not only for research purposes but also to “jump-start” sales by raising awareness of a drug among likely prescribers.¹⁰ The manufacturer has denied the allegation.¹¹

Clashes between journal editors and authors over relationships between research findings and commercial interests are not uncommon. In one particularly high-profile example in 2005, a conflict developed between the editors of the *New England Journal of Medicine (NEJM)* and the authors of the Vioxx (rofecoxib) Gastrointestinal Outcomes Research (VIGOR) trial report (2000)¹² over what the editors described as “inaccuracies in data” about cardiovascular risk.¹³ The debate had been prompted by the court-ordered discovery of the manufacturer’s internal documents, which suggested that 3 additional myocardial infarctions in the rofecoxib treatment group were known to the manufacturer more than 4 months prior to publication but not reflected in the report’s data analysis.¹³ The VIGOR trial authors not employed by the manufacturer countered that because the infarctions had occurred after an a priori data collection and unblinding cutoff date, it would have been inappropriate to include them in the analysis.¹⁴ A subsequent (2008) analysis of the same manufacturer’s internal documents indicated that elevated mortality risk associated with rofecoxib treatment (pooled hazard ratio [HR] of 2.99; 95% confidence interval [CI] = 1.55-5.77) was known in 2001 but was not reported to the U.S. Food and Drug Administration (FDA) in timely fashion or clearly disclosed in study reports published in 2004 and 2005.⁶ The manufacturer has denounced the allegation that it failed to disclose known risks as “false and misleading.”⁷

The journal peer review process, widely considered to be a hallmark of scientific integrity and quality, has also recently

found itself the unwitting victim of bad behavior. In January, 2008, testimony before the U.S. Senate Finance Committee suggested that a prominent diabetes researcher, who had been reviewing a manuscript for *NEJM*, allegedly violated peer review confidentiality rules by leaking the manuscript to a pharmaceutical manufacturer that was likely to be negatively affected by its publication.¹⁵ A less-publicized but fascinating incident occurred at the *Journal of General Internal Medicine* in 2004 when editors sent a manuscript to peer review. Coincidentally, one of the assigned peer reviewers had been asked in the previous year to lend her name to it even though she had not done any of the work; the actual author was a medical education company employee.² When the reviewer realized that the medical education company had successfully recruited another named author for the ghost-written work, she reported the deception to the journal's editors, who rejected the paper, implemented new authorship disclosure policies, and publicly expressed concerns about the publication of literature that “injects bias and untruth into the scientific dialogue in order to enhance corporate profits.”¹⁶

Ghost Writing, Truth-Burying, and Other Bad Behavior: The Trees, Not the Forest

Although allegations of ethical lapses – such as inaccuracies in authorship lists, deliberate failure to disclose drug safety issues, or information leaks – are important and make for salacious reading, sole focus of our attention on these issues will draw our attention away from a broader and more important question: In assessing the credibility of research information, are decision-makers limited to the simplistic foregone conclusion that financial interest will dictate research outcomes irrespective of actual evidence? Are decision-makers' evaluations of *what* they can believe limited to a superficial assessment of *whom* they can believe? Or can they turn to more objective standards in assessing the quality of research information?

To address this broader question, we begin by examining current evidence of a relationship between financial interest and research outcomes that has caused justifiable concern among many observers. We explain why the approach taken in many proposed solutions to the problem is likely to be ineffective. We propose as an alternative a more direct examination of quality of evidence, using available standards for conducting and reporting research. After reviewing key research standards and discussing evidence of widespread nonadherence to these basic standards, we propose a way forward.

Financial Interest and Research Results: Empirical Evidence

The view that financial interest *is* *facto* equates to results is supported by evidence closely linking study sponsorship to reported research outcomes. In a meta-analysis of 324 RCTs of cardiovascular devices and drugs published between January, 2000 and July, 2005 in 3 prominent journals (*JAMA*, *Lancet*, and *NEJM*), Ridker and Torres identified a strong relationship between fund-

ing and positive outcomes for newer treatments. Of 205 drug trials, the proportions finding clinical superiority for the newer treatment were 39.5% for those with not-for-profit funding, 54.4% for those that were jointly funded, and 65.5% for those funded by for-profit companies ($P=0.002$).¹⁷ In an analysis of 504 studies of inhaled corticosteroids, Nieto et al. found a strong inverse relationship between pharmaceutical manufacturer funding and the likelihood of identifying adverse effects for a study drug; 34.5% of 275 studies funded by pharmaceutical manufacturers versus 65.1% of 229 studies with other funding reported that the study drug of interest had significantly higher rates of adverse effects than its comparators.¹⁸

Survey data provide direct evidence that scientists perceive, and sometimes respond to, pressure from their funding sources. In a 2002 survey of 3,427 scientists who had received financial support from the National Institutes of Health, 15.5% admitted to “changing the design, methodology, or results of a study in response to pressure from a funding source” in the previous 3 years. Although only 0.3% admitted to “falsifying or ‘cooking’ research data,” 13.5% said that they had used “inadequate or inappropriate research designs,” and 10.8% acknowledged “withholding details of methodology or results in papers or proposals.”¹⁹ The problem was, according to the scientists surveyed, not limited to their own work; 12.5% of the survey respondents admitted to “overlooking others’ use of flawed data or questionable interpretation of data.”²⁰

Proposed Solutions to the Problem of Money Driving the Evidence

Given this background, it is not surprising that many recent proposals to reform the manuscript publication system have focused primarily on methods to either detect or respond to violations of ethical improprieties, particularly authorship and sponsorship disclosure requirements. One observer has suggested that journal editors should share a “publicly available, regularly updated database of conflicts of interest and ethical transgressions” to enable medical editors to “punish transgressors.”²¹ (As editors, we choose not to think too carefully about how such punishment might be carried out.) A 2005 editorial in the *Annals of Internal Medicine* urged academicians to “insist” that articles “reflect their own interpretation of the evidence,” to demand full disclosure of all authors and contributors, and to ensure that all authors take responsibility for the manuscripts that bear their names. The *Annals* editorial further committed to reporting ethical breaches to academic institutions and encouraged those institutions to “mandate adherence” to ethical policies.⁸ A 2006 editorial in *Health Services Research (HSR)* emphasized disclosure not only of financial arrangements but also of evidence of bias, for example, public stands taken on policy issues. Somewhat in contrast to the positions taken by other journal editors, the *HSR* editors emphasized that both transparent disclosure and “peer reviewed evaluation of the appropriateness and quality of methods” were

Rethinking the “Whodunnit” Approach to Assessing the Quality of Health Care Research – A Call to Focus on the Evidence in Evidence-Based Practice

TABLE 1 Guidelines for Conducting and Reporting Research

GUIDELINE NAME	SOURCE OR ORGANIZATION	DESCRIPTION
GENERAL STANDARDS FOR ALL TYPES OF RESEARCH		
Systems to Rate the Strength of Scientific Evidence^{25,26}	AHRQ	Compilation and synthesis of 121 sources (scales, checklists or guidance documents) that “can be used in the production of systematic evidence reviews and technology assessments.” Systems were characterized with respect to “important domains and elements” that safeguard against bias or are “long-accepted” research practices.
Uniform Requirements for Manuscripts Submitted to Biomedical Journals ²⁷	ICMJE	Covers (a) the role of editors, authors and peer reviewers and good ethical practices for each; and (b) standards for content of study reports. Does not discuss methodology or analytic technique in detail, but refers readers to CONSORT and its offshoots (e.g., STROBE, STARD, QUORUM, and MOOSE) for detailed guidelines.
ETHICAL STANDARDS		
Guidelines on Good Publication Practice (GPP) ²⁸	Pharmaceutical company employees; working group	Devoted primarily to ethical issues of authorship and sponsorship. Refers readers to ICMJE and CONSORT for more complete standards.
ISPOR Code of Ethics ²⁹	ISPOR	Brief statement of ethical standards for design and conduct of research, sponsorship, publication, and relationships with students and employees.
OBSERVATIONAL RESEARCH AND CLAIMS DATABASE ANALYSES		
Checklist for Medication Compliance and Persistence Studies³⁰	ISPOR Medication Compliance and Persistence Special Interest Group	Checklist of items to be “included, or at least considered,” in retrospective database analyses of medication compliance or persistence. Includes definitions of commonly used terms. Refers to the ISPOR Retrospective Database Checklist for detailed standards to be used in assessing work derived from administrative claims databases.
Retrospective Database Checklist³¹	ISPOR working group	Checklist to assist decision makers in assessing quality of analyses of administrative databases; covers numerous areas including relevance, reliability and validity, methodological considerations unique to claims databases, statistical analysis and interpretation.
Meta-Analysis of Observational Studies in Epidemiology (MOOSE) ³²	EQUATOR network	Standards for reporting of meta-analyses of observational epidemiological studies, including background, search criteria, methods of selection and analysis, and presentation of findings and implications.
Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)³³	EQUATOR network	Best practices for reporting of observational research at every stage including study rationale, design, measurement, analysis, results, limitations, and discussion; includes citations to 237 references that can be accessed for additional information and examples.
Transparent Reporting of Evaluations with Nonrandomized Designs (TREND)³⁴	EQUATOR network	Consistent with CONSORT; explicitly addresses issues common in nonrandomized studies that examine the effects of interventions.
RANDOMIZED CONTROLLED TRIALS		
Consolidated Standards of Reporting Trials (CONSORT)³⁵	EQUATOR network	Standards for the reporting of randomized controlled trials including study rationale, design, measurement, analysis, results, limitations, and discussion; includes citations to 204 references that can be accessed for additional information and examples. Main focus is validity (internal and external).
Good Research Practices: Randomized Clinical Trials -- Cost-Effectiveness Analysis (RCT-CEA)³⁶	ISPOR working group	Standards for conducting and reporting analyses of economic end points incorporated into clinical trials.
STATISTICAL ANALYSES AND MODELS		
Chicago Guides^{37,38}	Chicago University Press	Comprehensive textbook standards for data analysis and reporting.
Cost-Effectiveness in Health and Medicine³⁹	U.S. Public Health Service Panel on Cost Effectiveness in Health and Medicine – expert panel convened in 1993	Guidelines for the conduct and reporting of cost-effectiveness analyses.
Principles of Good Practice for Decision Analytic Modeling⁴⁰	ISPOR Task Force	Comprehensive, detailed quality assessment document for decision analytic modeling. Emphasizes transparency, sensitivity analyses, empirically-based assumptions, and incorporation of new evidence as it becomes available.
Quality of Health Economic Studies (QHES) ⁴¹	Panel of 8 health economics experts; validated in a survey of 30 clinicians and 30 health economists	Quality scoring system for cost-minimization, cost-effectiveness and cost-utility analyses, based on 19 previous guidelines and checklists and validated against expert assessment of quality. Primarily focused on technical details specific to cost-effectiveness (e.g., time horizon, perspective, economic model structure).
OTHER RESEARCH TYPES		
Patient Reported Outcomes (PRO) Harmonization Group report⁴²	Representatives from ISOQOL, ISPOR, PhRMA-HOC, and ERIQA	Evolving standards for instruments and reporting of patient-reported data.
Quality of Reporting of Meta-analyses (QUORUM) ⁴³	EQUATOR network	Similar to CONSORT in concept, but addresses issues specific to meta-analyses of RCTs.
Standards for Reporting of Diagnostic Accuracy (STARD) ⁴⁴	EQUATOR network	Checklist of standards for reporting studies of diagnostic test accuracy.

AHRQ = Agency for Healthcare Research and Quality; EQUATOR = Enhancing the QUALity and Transparency Of health Research; ERIQA = European Regulatory Issues on Quality of Life Assessment; ICMJE = International Committee of Medical Journal Editors; ISOQOL = International Society for Quality of Life Research; ISPOR = International Society for Pharmacoeconomics and Outcomes Research; PhRMA-HOC = Pharmaceutical Manufacturer's Association Health Outcomes Committee; RCT = randomized controlled trial.

crucial to evaluation of the quality of research evidence.¹⁹ For our part, the *JMCP* disclosure forms for authors have been revised many times in response to instances of apparent ghost writing or ghost management. These *JMCP* disclosure forms require the corresponding author to certify the percentage contribution of every person that contributed in any way to the manuscript, including study concept and design, data collection, data interpretation, writing and revising the manuscript.²⁰

In 2001, the editors of *JAMA* supplemented the journal's enhanced disclosure policies with an additional response to the problems of “data misrepresentation ... and selective reporting.”²¹ Submissions of manuscripts reporting industry-sponsored studies must be accompanied by certification that the data analysis was conducted by “an academic statistician who is not employed by the sponsor and who is at an academic center, such as a medical school, or is an employee of a government research institute.” This approach was intended to provide “an additional layer of oversight for the integrity of the data analysis and reporting” and to rely on academic centers' usual “mechanism for investigation” of possible improprieties. *JAMA's* editors expressed considerable confidence in their policy change: “If all journals would have similar policies,” they opined, “the likelihood of manipulation of data, inappropriate data analysis and selective reporting of results could be substantially decreased.”²¹

Replacing the “Whodunnit” Approach: A Sharper Focus on Quality of Evidence

These proposed solutions focus primarily on “whodunnit” questions – who paid for the study, who conducted it, who interpreted the data, who wrote the report, and who enforced ethical standards. Although full disclosure of financial interests and authorship arrangements is clearly important, addressing the health care research literature's current problems through the “whodunnit” approach alone is unlikely to be effective.

The requirement that analyses be conducted by an academician instead of a for-profit company employee misses 3 key points: first, that high-quality work can be, and often is, produced by for-profit companies; second, that academicians, like all of us, are fallible human beings who can make errors or be influenced by bias; and third, that academicians are potentially just as vulnerable as other scientists to funding source pressures. As Ridker and Torres point out in their meta-analytic assessment of research with for-profit and nonprofit sponsorship, design problems in trials conducted in academic settings have been observed in the literature,¹⁷ such as in a 2006 investigation of high-dose statin therapy that was led by a medical school academician and published in a prominent journal despite having no control group.^{17,22} Additionally, Ridker and Torres observe that even academicians may “nevertheless be under pressure to present results in the best possible light, in particular emphasizing favorable subgroup analyses when an overall neutral effect was observed for the primary end point.”¹⁷

To their point, Mello et al. (2005) surveyed 107 medical school research administrators who had authority over contracts with industry research sponsors for the conduct of clinical trials. Respondents were presented with a list of potential contract provisions and asked which would be acceptable or unacceptable to them. Allowing sponsors to “alter the study design after the agreement is executed” was viewed as acceptable by 62% of respondents and unacceptable by 27%. Allowing sponsors the right to insert their own statistical analyses in manuscripts was seen as acceptable by 24% of respondents; 47% said that this provision would be unacceptable and 29% were not sure. A remarkably high 50% of the respondents said that they would permit the sponsor to “write up the results for publication and the investigators may review the manuscript and suggest revisions,” although 40% viewed such a provision as unacceptable.²³

Thus, an information consumer whose only approach for assessing quality of evidence is identifying *who* is “good” (e.g., who has pure motives, good training, etc.) is hopelessly relegated to making judgments with inadequate information, thereby inevitably rejecting some high-quality work and accepting some poor-quality work. We embrace a different and more effective approach. Instead of looking at the motives or quality of *people* – an exercise that is fraught with inherent uncertainty – we propose a more intense focus on the quality of *work output* using currently available standards for the conduct and reporting of research. Information consumers who rely on basic research quality standards will be in a strong position to distinguish poor-quality from high-quality work, no matter who performed or paid for it.

Research Standards: Are They Usable?

The report of a roundtable discussion of the use of real-world data in managed care decision-making, published in the April, 2008 issue of *JMCP*, observed that standards for the conduct and reporting of research are seldom used, perhaps because they are either not clearly available, not clearly understood, or both.²⁴ Confusion over the availability and utility of research standards is understandable, since a “Google” search on the term “research standards” yields dozens of entries ranging from vague admonitions to conduct scientifically valid and ethical work to detailed “how-to” publications.

Our analysis of the available tools to assess evidence quality examined documents that specifically address a question that is critically important to decision-makers: how to distinguish credible from less credible health care research results. Of those documents (Table 1),²⁵⁻⁴⁴ many address technical issues relevant to only a single type of work (e.g., decision analytic modeling, studies of diagnostic test sensitivity and specificity) or to specific problems (e.g., ghost authorship, sponsorship ethics) rather than principles broadly applicable to all types of health care research. The documents included in our analysis (shown in bold print in the table) met the following criteria:

(a) Address the question of how to conduct and/or report research.

Research reporting is especially critical because publications are typically a decision-maker’s only potentially valid and reliable source of information about the conclusions reached by study authors. No information consumer should rely solely on popular press accounts.

- (b) Provide specific guidance rather than vague admonitions. For example, the advice that study groups should be appropriate is accurate but not especially helpful to a decision-maker unfamiliar with research methods. In contrast, the guidance that study groups should be comparable with respect to demographic and clinical characteristics, including age, gender, payer type, industry type, benefit design, and comorbidities, is sufficiently specific to be usable.
- (c) Address quality measures that are applicable to many types of health care research rather than being limited to specific technical issues. For example, the Quality of Health Economic Studies instrument, a quality-assessment checklist for cost-effectiveness analyses, is not included in our detailed review because it addresses primarily technical issues (e.g., perspective of analysis, time horizon, discounting method) but not broader issues (e.g., how a literature review should be performed and reported or how to present implications of study findings for policy or future research.)⁴¹
- (d) Use is complementary with other guidelines (e.g., the AMCP dossier format for formulary submissions or the International Committee of Medical Journal Editors (ICMJE) standards for manuscript submissions).²⁷ For example, a managed care decision-maker assessing material for dossier inclusion or an author assessing studies for inclusion in a literature review may need to refer to detailed standards for quality of research evidence.

Although our review was not intended to be comprehensive of every research guideline available, it covers respected organizations and information sources frequently used by researchers in the managed health care field. These include:

- (a) University of Chicago Press reference guides to writing about numbers³⁷ and about multivariate analyses³⁸ (these 2 books provide virtually identical guidance);
- (b) International Society for Pharmacoeconomics and Outcomes Research (ISPOR) code of ethics and standards for conducting and reporting studies of medication compliance and persistence, cost-effectiveness analyses, decision analytic models, patient-reported outcomes, and retrospective analyses of administrative claims;^{29-31,36,40,42}
- (c) Key standards cited by the Enhancing the QUALity and Transparency Of health Research (EQUATOR) network funded by the National Library for Health and the National Institute for Health Research; these include reporting guidelines for RCTs (Consolidated Standards of Reporting Trials [CONSORT]),³⁵ observational research (Strengthening the Reporting of Observational Studies in Epidemiology

TABLE 2 Key Principles in Conducting and Reporting Research

Transparency
<ul style="list-style-type: none">• Study selection criteria are specified in detail• Quantitative effect of each sampling criterion on sample size is disclosed explicitly; a sample derivation flow chart or table is preferred.• Details of intervention and comparison conditions are described explicitly• Methods are described in sufficient detail to enable replication• Specific contribution of each author and other contributors is disclosed.^a
Hierarchy of evidence quality
<ul style="list-style-type: none">• To maximize internal validity (accurate causal inference) design minimizes the effects of confounding• Randomized designs are ideal; quasi-experimental is acceptable• Results of studies with simple pre-post or cross-sectional designs are given lesser weight
Quantitative exploration of possible bias
<ul style="list-style-type: none">• Report includes a demonstration (not just assertion) that study groups are comparable• Studies of claims databases address potential confounding factors specific to that data source• Possibility of bias is explored using sensitivity analyses on alternative methodological decisions or decision model inputs
Basic logic
<ul style="list-style-type: none">• Study report reflects a basic logical and evidence-based connection, or “plausible mechanism” between input and outcome• Interpretation does not confuse association with causation (e.g., white hair does not increase the risk of mortality)^b
Recognition of the meaning and limitations of statistical methodologies
<ul style="list-style-type: none">• Statistical significance is distinguished from practical or clinical significance• Descriptive analyses of study outcomes are presented• Multivariate analyses are presented only if necessary (e.g., if simpler bivariate analyses do not sufficiently address the research question because of confounding factors)• If multivariate (adjusted) analyses are used, descriptive (unadjusted) statistics are also presented• Degree of precision and amount of uncertainty around statistical estimates are presented
<i>^a Implementation of a contributorship policy is recommended by the International Committee of Medical Journal Editors.²⁷</i>
<i>^b From an example by Jane Miller in The Chicago Guide to Writing About Multivariate Analysis³⁸</i>

[STROBE]),³³ and intervention studies conducted in nonrandomized groups (Transparent Reporting of Evaluations with Nonrandomized Designs [TREND]);³⁴ and,

- (d) Guidelines promulgated by the U.S. Department of Health and Human Services and its agency arms, the U.S. Public Health Service (USPHS) and the Agency for Healthcare Research and Quality (AHRQ); these include a review of systems to rate the strength of scientific evidence^{25,26} and the report of the Panel on Cost-Effectiveness in Health and Medicine, which was charged in 1993 with making recommendations for the conduct of cost-effectiveness analyses “to improve their quality and encourage their comparability.”³⁹

Rethinking the “Whodunnit” Approach to Assessing the Quality of Health Care Research – A Call to Focus on the Evidence in Evidence-Based Practice

TABLE 3 Summary of Key Guideline Documents: Standards for Conducting and Reporting Research

	CHICAGO GUIDE ^a	ISPOR ^b	CONSORT/ STROBE/ TREND ^c	HHS ^d
INTRODUCTION AND BACKGROUND				
<i>Rationale for study is apparent and reasonable</i>				
Rationale for methodology (e.g., design, intervention, definitions used) is explained in the context of published evidence and theory; decision analytic model specifies an evidence-based “event pathway” ^{30,31,33,34,37-40}	√	√	√	√
How previously published or standard methods informed the present study is explained ^{26,30,31,33-35,37,38,40}	√	√	√	√
METHODS				
<i>Research is designed to minimize bias or confounding; method of minimizing bias is explained</i>				
Design reflects understanding of “hierarchy of evidence:” (a) randomization is ideal for internal validity, (b) well-done pseudo-randomized designs are acceptable, (c) a comparison group is nearly always essential, and (d) cross-sectional and pre-post designs are weak and may produce invalid results ^{26,31,35,37-39}	√	√	√	√
If design is nonrandomized, comparison group is as similar as possible to intervention group (in all respects except for exposure to intervention) ^{26,31,33,34,37-39}	√	√	√	√
Methods used to adjust, mitigate, or compensate for potential biases are described ^{26,29-31,33-39}	√	√	√	√
<i>Study design was a priori, not post hoc</i>				
A priori plan, and departures from that plan, are disclosed and explained ^{29,31,33-36,42}		√	√	
Method for determining sample size is specified; a priori power calculations are preferred ^{26,30,31,33,35,36}		√	√	√
<i>Explicit description of study inclusion and exclusion criteria</i>				
Study selection criteria are explicit (e.g., list used for selection, specific ages and diagnoses, location, type of clinic, etc.); description should be specific enough to enable reader to determine applicability to his/her population ^{26,30,31,33,35,37-39}	√	√	√	√
Quantitative effect of each criterion on sample size is reported for each stage of sample development ^{30,31,33-35,37,38}	√	√	√	
Sample derivation flow chart is provided ³³⁻³⁵			√	
Effect of subject refusal or withdrawal (e.g. survey response rate, refusal rate) on sample size is reported ^{26,33-35,37-39,42}	√	√	√	√
Reasons for dropouts are reported; at a minimum, dropouts for reasons relevant to outcomes (e.g., death) are reported separately from those irrelevant to outcomes (e.g., lost address) ^{33,35,37,38}	√		√	
<i>Intervention details are disclosed fully and specifically – treatment, settings, administration method^{26,31,33-35,37-40}</i>				
<ul style="list-style-type: none"> • In intervention trials, including drug trials, report includes explicit description of intervention (e.g., dose, administration schedule, blinding procedures, method for disguising placebo, qualifications of personnel administering intervention) • In drug benefit studies, formulary, step therapy, PA requirements, caps, deductibles, copayment structures, and other key details such as the proportion of mail order claims are described specifically • Carveouts, capitated arrangements, changes in reimbursement arrangements, or other policies that would affect findings, are disclosed if applicable • In decision analytic models, events and transitions associated with both interventions and alternatives are described explicitly 	√	√	√	√
<i>Outcome measures are valid and explicit</i>				
Methods section clearly and specifically defines outcome measures, including specific intermediate or end point outcomes, time horizon, measurement method, in sufficient detail to make replication possible ^{26,30,31,33,35,37-40}	√	√	√	√
<ul style="list-style-type: none"> • Decision analytic model specifies input values, data sources, and calculation methods explicitly • Claims data study report provides detailed definitions of all variables used (e.g., ICD-9-CM, GPI, CPT-4 codes) • Inclusion of sufficient detail may require technical appendix if journal space is scarce 				
Validated instruments, measures, and methods are used, or validity of new measures is assessed ^{26,30,36,39,40,42}		√	√	√
<i>Quantitative procedures are clear.</i> Calculation procedures and statistical methods are described in sufficient detail and specificity that a knowledgeable reader with access to the data could replicate the analysis ^{33,37-40}	√	√	√	√
<i>Studies involving analyses of multiple subgroups or multiple comparisons are performed according to technical standards</i>				
“Multiple comparison” problem is addressed using post hoc testing or P value adjustment (e.g., Bonferroni, Tukey’s Honestly Significant Difference Test) ^{26,30,31,33,35,36,42}		√	√	√
Subgroup effect assessments are made using interaction terms or other direct comparisons of groups, not by comparing P values across different groups ^{33,35-38}	√	√	√	
<i>Sensitivity analyses test effects of various methodological options</i>				
If more than 1 methodological decision rule was possible or appropriate, researcher examines effects of different decision rules (e.g., time periods, diagnoses, number of claims required for inclusion) for selecting study subjects ^{30,31,33,36-39}	√	√	√	√

Rethinking the “Whodunnit” Approach to Assessing the Quality of Health Care Research – A Call to Focus on the Evidence in Evidence-Based Practice

TABLE 3 Summary of Key Guideline Documents: Standards for Conducting and Reporting Research

↓	CHICAGO GUIDE ^a	ISPOR ^b	CONSORT/ STROBE/ TREND ^c	HHS ^d
If more than 1 method for measuring utilization would have been appropriate, researcher tests different methods for measuring utilization or cost (e.g., disease-related or treatment-related) ^{31,36,39}		√		√
Quantitative effects of potential bias are explored. Observational studies estimate potential effects of confounding bias, selection, bias, information bias. Decision analytic models assess degree of sensitivity to model assumptions ^{26,31,33,37-40}	√	√	√	√
RESULTS: ANALYSIS AND INTERPRETATION OF DATA				
<i>Key attributes of study population and setting are described transparently; readers are given enough information to assess appropriateness of study group(s)</i>				
Descriptive statistics on demographics, clinical profile, setting, potential confounders are presented ^{26,31,33-35,37-39}	√	√	√	√
In studies with multiple groups, descriptive information is presented for each group ^{26,33-35,37-39}	√		√	√
Length of follow-up is disclosed explicitly; in studies with variable length of follow-up, descriptive statistics about follow-up time (e.g., mean, median, range) are reported for each group to assess possibility of differential loss to follow-up ^{26,30,33,35,36,39}		√	√	√
Representativeness of the sample is explored: subjects included are compared with those not included in final analysis (e.g. refused, did not meet sampling criteria, lost to follow-up) or with population to which results will be generalized ^{26,31,33,34,37-39}	√	√	√	√
<i>Outcome measures are presented transparently</i>				
Report includes unadjusted descriptive statistics for study outcome measures (not a sole reliance on multivariate model results) ^{33,35,37,38}	√		√	
For continuous data, report mean and SD or SE (SD more commonly cited) ^{33,35,36-38}	√	√	√	
For skewed continuous data, present median and IQR ^{33,35,37,38}	√		√	
For continuous data, report range (minimum and maximum) ^{37,38}	√			
For grouped or categorical data, report category limits ^{33,37,38}	√		√	
Show counts of subjects or patients included in each analysis ^{33,35,37,38}	√		√	
<i>Sources of uncertainty, confounding, and bias are addressed</i>				
Measures of statistical precision (e.g., confidence intervals) are presented ^{26,30,33,35-40}	√	√	√	√
Potential for unmeasured confounding factors is recognized and assessed when using a multivariate model (e.g., measures of predictive accuracy or model fit) ^{30,31,33,37-39}	√	√	√	√
Quantitative details are presented (e.g., coefficients for statistical models, parameters for decision analytic models) ^{26,37-40}	√	√		√
Modeling process is described specifically (e.g., variable selection process, distribution assumed for statistical models; causal pathway for decision analytic models) ^{26,30,31,33,37-40}	√	√	√	√
<i>Handling of technical issues is explained</i>				
Techniques are appropriate for distribution (e.g. linearity, curvilinearity, skewness) ^{26,30,31,33,36-38}	√	√	√	√
Amount and handling of missing data, including dropouts, are reported ^{26,31,33-38,42}	√	√	√	√
DISCUSSION AND CONCLUSIONS				
<i>Association is distinguished from causation</i>				
The process that produces an association between variables is explored (e.g., a “plausible mechanism,” “causal pathway,” or “biologic credibility.”) Report recognizes that findings of association(s) between variables could be spurious ^{31,33-38}	√	√	√	
Alternative explanations for study findings, including bias, confounding, or reverse causation, are explored and reported specifically ^{26,31,33-35,37,38}	√	√	√	√
<i>Clinical or practical significance (not just statistical significance) is presented</i>				
Results are presented as outcome effect sizes (e.g., amount of reduction in cost, number of additional months of persistence, NNT, absolute risk reduction), not just P values ^{26,30,33-39,42}	√	√	√	√
Researcher recognizes phenomenon of statistical power; large sample sizes can produce findings that are statistically significant but without clinical or practical significance, and small sample sizes can produce clinically significant but statistically insignificant findings; both statistical and practical significance are taken into account in interpretation ^{26,30,31,33,35,37-39,42}	√	√	√	√
<i>Results are discussed in light of previous research</i>				
Report makes specific quantitative comparisons to findings of previously published studies ^{30,33,35,37-40}	√	√	√	√
Report explains reasons for similarities and differences compared with previously published work ^{30,33,37-40}	√	√	√	√
Limitations of methods and design are discussed (e.g., potential biases, confounding) ^{26,30,31,33,36-39}	√	√	√	√
ETHICS				
Rights of research subjects are recognized and protected (e.g., evidence of IRB, explanation of IRB exemption, informed consent by subjects affected by research) ^{29,30,36-38}	√	√	√	
Full disclosure of sponsorship or business relationships related to research ^{26,29,30,33,35,39}		√	√	√

Rethinking the “Whodunnit” Approach to Assessing the Quality of Health Care Research – A Call to Focus on the Evidence in Evidence-Based Practice

TABLE 3 Summary of Key Guideline Documents: Standards for Conducting and Reporting Research

	CHICAGO GUIDE ^a	ISPOR ^b	CONSORT/ STROBE/ TREND ^c	HHS ^d
Authorship list is accurate ²⁹		√		
SPECIAL CONSIDERATIONS WHEN USING DATABASES INITIALLY DESIGNED FOR BILLING AND PAYMENT PURPOSES				
Quality (accuracy and completeness) of data has been considered; in a claims database analysis, assurance checks and remedies for quality problems are performed, even when data have been pre-processed by a third party prior to delivery to researcher ^{30,31,37-39}	√	√		√
In an analysis of insured populations, eligibility for health care services is accounted for, either by requiring continuous eligibility for study inclusion or by using appropriate measures (e.g. per member per month, rate per person-year of follow-up) ^{30,31}		√		

CPT-4 = Current Procedural Terminology, Fourth Edition; GPI = Generic Product Identifier; ICD-9-CM = International Classification of Diseases, Ninth Revision, Clinical Modification; IQR = interquartile range; IRB = institutional review board; NNT = number needed to treat; PA = prior authorization

^a *The Chicago Guide to Writing About Numbers*³⁷ and *The Chicago Guide to Writing About Multivariate Analysis*³⁸

^b *International Society for Pharmacoeconomics and Outcomes Research. Includes:*

- *Code of Ethics*²⁹
- *Checklist for Medication Compliance and Persistence Studies*³⁰
- *Retrospective Database Checklist*³¹
- *Good Research Practices: Randomized Clinical Trials – Cost-Effectiveness Analysis*³⁶
- *Principles of Good Practice for Decision Analytic Modeling*⁴⁰
- *Patient Reported Outcomes Harmonization Group (in conjunction with European Regulatory Issues on Quality of Life Assessment, International Society for Quality of Life Research, and Pharmaceutical Manufacturer's Association Health Outcomes Committee)*⁴²

^c *From the EQUATOR (Enhancing the QUALity and Transparency Of health Research) network. Includes:*

- *Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)*³³
- *Consolidated Standards of Reporting Trials (CONSORT)*³⁵
- *Transparent Reporting of Evaluations with Nonrandomized Designs (TREND)*³⁴

^d *U.S. Department of Health and Human Services. Includes:*

- *Agency for Healthcare Research and Quality Systems to Rate the Strength of Scientific Evidence*^{25,26}
- *U.S. Public Health Service: Cost-Effectiveness in Health and Medicine*³⁹

Table 2 shows key principles reflected in the guideline documents. Table 3 contains a detailed quality checklist summarizing the guidance provided by each information source, sorted approximately in the order of sections in a typical research paper. Key principles include:

Key Principle: Transparency

All sources emphasize transparency in reporting every phase of a research project. First, study selection criteria should be specified in sufficient detail that readers can clearly determine the applicability of the study sample to their populations. For example, the results of a benefit design intervention study conducted in a hospitality industry sample consisting primarily of young adults paid at minimum wage probably do not inform the expected results of the same intervention implemented in an information technology company consisting of highly paid middle-aged white collar executives and computer programmers. Similarly, results obtained in primary care clinics serving low-income patients are unlikely to predict outcomes in a clinic that accepts only private insurance. For this reason, guidelines that discuss sampling procedures state that inclusion and exclusion criteria – such as age range, diagnoses, period of time for qualification into study, location, setting, or any factor that potentially could influence outcomes – should be specified explicitly.^{26,30,31,33,35,37-39}

Similarly, to enable the reader to assess how sampling procedures could have affected study validity, most guidelines emphasize that the quantitative effect of each sampling criterion on sample size should be disclosed explicitly.^{30,31,33-35,37,38} CONSORT and STROBE guidelines recommend a sample selection flowchart,³³⁻³⁵ a JMCP requirement (e.g., Figure 1 in Stockl et al.)⁴⁵ Some but not all guidelines recommend comparisons (e.g., on demographic and clinical characteristics) of cases included in the sample with either cases excluded from the sample or benchmark data (e.g., data for a sample of patients with diabetes could be compared with similar nationwide data from the American Diabetes Association).^{26,31,33,34,37-39} Guidelines for studies of patient-reported outcomes include reporting of the participation rate, such as a survey response rate.⁴²

Guideline documents also recommend complete disclosure of the details of interventions and outcomes, including time horizon, measurement method, and codes used (e.g., *International Classification of Diseases, Ninth Revision, Clinical Modification* [ICD-9-CM], *Current Procedural Terminology*, and *Health Care Common Procedural Coding System* codes). Although the specific content of recommendations varies by type of study, the principle of complete disclosure is remarkably consistent. For example, in RCTs of drug treatments, standards call for disclosure of dose, administration schedule, procedures used to blind participants

and investigators to placebo and active treatment, and qualifications of personnel administering the intervention.^{26,35} In drug benefit studies, guidelines call for disclosure of benefit design details, such as formularies, step therapy and prior authorization requirements, carveouts, mail order incentives, or any other arrangements that could potentially affect study findings.³¹ TREND guidelines for intervention evaluation studies employing nonrandomized designs describe as “critical” the reporting of “sufficient detail so that a reader has an understanding of the content and delivery of both the experimental intervention and the services in the comparison condition.”³⁴

Guideline sources, particularly those that address calculation methodologies in detail, are consistent with the ICMJE standard that statistical methods should be described “with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results.”²⁷ Both the USPHS recommendations for cost-effectiveness analyses and the ISPOR guidelines for decision analytic models include specification of all input values, data sources, and calculation methods.^{39,40}

CONSORT authors effectively summed up the guidance on transparency: “Readers should not have to speculate; the methods used should be transparent, so that readers can readily differentiate trials with unbiased results from those with questionable results. Sound science encompasses adequate reporting, and the conduct of ethical trials rests on the footing of sound science.”³⁵

Key Principle: Hierarchy of Evidence Quality

Guideline documents that address research design uniformly recognize that a key element in evidence quality is, as an AHRQ working group described it, “the extent to which all aspects of a study’s design and conduct can be shown to protect against systematic bias, nonsystematic bias, and inferential error.”²⁵ Key to this assessment is the minimization of confounding, defined as the degree to which subjects in intervention and control (or comparison) groups differ on a factor, other than the intervention, that might affect the study outcome(s). Randomization of study groups is widely recognized as the “gold standard” method to ensure the absence of confounding.³⁵ Quasi-experimental designs, such as those that compare pre-intervention versus post-intervention outcomes across 2 groups (i.e., a group that receives the intervention vs. a similar group that does not) are recognized as a reasonable alternative to randomization. Cross-sectional designs, which compare outcomes for nonrandomized groups during a single time period, and designs that compare pre-intervention versus post-intervention outcomes without a comparison group are recognized as more prone to producing invalid results because they are subject to confounding effects (e.g., unmeasured group differences or changes that would have occurred over time even without the intervention). This recognition is clear both in guideline documents and in methods textbooks.^{26,31,35,37-39,46}

For nonrandomized designs, a demonstration that study groups are comparable with respect to demographic and clinical

characteristics and other potentially confounding factors is recognized in guidelines as important.^{26,31,33-34,37-38} Potential confounders vary by study topic, but would include, for example: (a) baseline compliance in a study of whether a disease management program affects medication persistency; (b) comparability of drug formularies in a study of whether particular benefit design features affect patient outcomes; or (c) baseline drug expenditure in a study of the effects of an intervention on payer or out-of-pocket cost. In the AHRQ’s assessment of systems to rate the strength of scientific evidence, all 12 instruments that were used to grade quality of observational research included comparability of subjects as a standard.²⁵ Thus, guidelines specify the need for a descriptive table comparing study groups on relevant factors.^{26,31,33-34,37-38} For the author to merely state, rather than show, that study groups are comparable is recognized as inadequate.

Additionally, guidelines for analyses of claims databases identify possible confounding factors that are specific to that data source.^{30,31} First and most important is that eligibility for health insurance benefits must be addressed, either by limiting analyses to members continuously eligible for insurance benefits during the study period or by using outcomes measures that adjust for eligibility (e.g., cost per month of enrollment). Second is an assessment of whether features specific to particular health plans could affect data quality or capture. For example, in a study of cost and utilization of mental health services, the researchers must address whether a “carveout” arrangement with an external vendor for the provision of some or all mental health services could have affected the capture of claims in the dataset. When dealing with samples in which under-reporting of data is common, such as patients with mental illness or enrollees who are aged 65 or older and therefore Medicare-eligible, researchers should describe explicitly how potential concerns about data capture and completeness have been addressed.

Although all guidelines recognize the importance of validity, the authors of the CONSORT guidelines are particularly dismissive of the idea that departures from internal validity (i.e., accuracy in making causal inferences) are acceptable if they enhance external validity (i.e., the degree to which the data represent “real-world” conditions): “Internal validity is a prerequisite for external validity: the results of a flawed trial are invalid and the question of its external validity becomes irrelevant.”³⁵

Key Principle: Quantitative Exploration of Possible Bias

Closely related to the understanding of hierarchy of evidence is the admonition in guideline documents that researchers should explore in quantitative fashion the possibility of alternative explanations for study findings.^{26,31,33-35,37,38} For example, in an observational study that has found higher medication persistence rates for Drug X than for Drug Y, the researcher should explore the possibility that the patients treated with Drug X were more compliant at baseline (e.g., by examining history of compliance with other therapeutic classes in the pre-treatment year). In a

study in which ICD-9-CM codes of 296.2 (major depression, single episode) and 296.3 (major depression, recurrent episode) were used to select patients with depression, sensitivity analyses should explore the use of other diagnostic codes, such as 311 (depressive disorder not elsewhere classified), because of the possibility that diagnoses of major depression were deliberately or inadvertently miscoded in the bill submitted by the provider for payment. Readers should be especially alert for, and suspicious of, research in which multiple definitions or approaches could reasonably have been used, but the study authors have chosen just 1 of these and performed no sensitivity analyses to assess the effects of their choices.

The need for sensitivity analyses is featured prominently in guideline documents for decision analytic and cost-effectiveness modeling.^{39,40} ISPOR's recommendations for best practice in modeling studies indicate that “model results should never be presented as point estimates or as unconditional claims of effectiveness or cost,” but instead should be “represented as conditional upon the input data and assumptions, and they should include extensive sensitivity analysis to explore the effects of alternative data and assumptions on the results.”⁴⁰

Key Principle: Basic Logic

A common theme that emerges from review of the guideline documents is that of a logical and evidence-based connection or “plausible mechanism” between input and outcome; findings should have “biologic credibility.”^{31,33-38} Guidelines for decision analytic models emphasize the need for an evidence-based “event pathway,” explicitly showing the reader the relationships among study variables to make the rationale for the calculations clear.³⁹⁻⁴⁰ Weinstein et al., in writing about the ISPOR recommendations for modeling studies, note that a model's value “lies not only in the results it generates, but also in its ability to reveal the logical connection between inputs (i.e., data and assumptions) and outputs in the form of valued consequences and costs.” Thus, readers should be able to understand the “logic behind [the model's] results ... at an intuitive level.”⁴⁰

Guidelines for observational research emphasize that a plausible event pathway is especially critical to avoid confusion between association and causation. In a marvelous tongue-in-cheek example of how *not* to interpret observational data, statistician Jane Miller points out that a naïve researcher, noting that the mortality rate among people with white hair is higher than the rate for those whose hair has remained its original color, might write something like this: “White hair increased the risk of dying by 400%.”³⁸ Extending Miller's example to an error commonly committed in observational analyses in health care, one might see a naïve researcher report that, given the association between white hair and elevated risk of hospitalization and death, we could save millions of dollars and hundreds of thousands of lives each year if only everyone of a certain age would simply make a small investment in the cost of hair coloration products.

Presentation of an evidence-based logical pathway between input and outcome also helps the reader feel confident in another aspect of study design that is widely recognized as important – the use of a priori rather than post hoc methodologies.^{29,31,33-36,42} For example, if the research literature uniformly defines Disease X as a 3-month history of symptoms, a researcher who, without explanation, selects a study population based on a 6-month symptom history opens the analysis to criticism. Without a clear evidence-based explanation for the sample selection criterion, it appears to readers that methods were hand-picked to produce a desired answer rather than specified in advance.

Key Principle: Meaning and Limitations of Statistical Methodologies

Guideline documents, especially those that specifically address the topic of statistical methods, emphasize in several ways the dangers of over-interpreting or misinterpreting statistical analytic results. A first key example, cited by every guideline source, is the need to distinguish between statistical significance and clinical or practical significance.^{26,30,31,33,35,37-39,42} In analyses with very large sample sizes, which are common in retrospective analyses of claims databases, virtually any result will be statistically significant; however, practical significance must be assessed carefully using absolute, rather than relative, measures of effects on outcomes. For example, in a study comparing medication possession ratios for Drug X versus Drug Y with sample sizes of 30,000 in each group, very small differences of less than 0.5 percentage points, or about 1.5 days of pharmacotherapy per year, will be statistically significant but have no practical meaning whatsoever. Conversely, some studies may be underpowered by insufficient sample size; for example, a 1-year persistency rate of 67% for Drug X versus 50% for Drug Y, with sample sizes of 30 in each group, suggests a possibly clinically important difference that has not been tested adequately. Using a 2-tailed test and alpha (Type I error) of 0.05, the number necessary for 80% power to detect a difference of 67% versus 50% is 130 in each group. Thus, a finding of no significant difference between the groups is not interpretable.

Additional admonitions common to guidelines for statistical analyses address a common phenomenon in health care research, the use of multivariate analyses to adjust for baseline differences in studies with observational designs. Guidelines specify that: (a) descriptive analyses of study outcomes should always be presented, (b) multivariate analyses should be presented only if necessary, and (c) multivariate analyses should never be presented without accompanying descriptive statistics.^{33,35,37,38} While recommendations for the content of descriptive analyses vary slightly by source, *Chicago Guide* recommendations include mean and standard deviation, or median and interquartile range for skewed data, minimum and maximum (range), and counts of subjects or patients included in each analysis.^{37,38}

Finally, guidelines emphasize that researchers should report

degree of precision around statistical estimates, such as 95% confidence intervals in analyses of randomized trials or observational data, or “bootstrapping” estimates of uncertainty for decision analytic models.^{26,30,33,35–40} All sources also note the importance of recognizing the possibility of unmeasured confounding factors.^{26,31,33,36–39} Thus, reports of multivariate models should include estimates of model adequacy, such as goodness of fit measures or a measure of predictive accuracy (e.g., R square for linear regression).^{31,38}

Consistent But Not Consistently Used: Problems in the Published Literature

The remarkable consistency in the content of published guidelines is matched by an equally remarkable inconsistency in adherence to them. In 2002, a review of the quality of evidence in the medical literature described the problems of “poor methodology and reporting” as “widespread.”⁴⁷ The review documented that the prevalence rates of serious methodological errors in published randomized trials in multiple medical disciplines were high. For example, 25% of 364 reports in surgery journals failed to specify eligibility criteria for study inclusion. The percentages of articles failing to report the intervention allocation method were 89% of 196 reports in rheumatoid arthritis journals, 48% of 206 reports in obstetrics/gynecology journals, and 44% of 80 reports in general medical journals. Of 196 reports in rheumatoid arthritis journals, 63% analyzed multiple observations using inappropriate statistical techniques, and of 50 reports in general medical journals, 58% used incorrect techniques in comparing subgroups.⁴⁷

Similar problems have been identified in the observational research literature. The STROBE group reported that in an analysis of observational studies of stroke, 17 of 49 reports did not specify eligibility criteria for study inclusion.³³ In a review of studies that used multivariate statistical methodologies, only 93 of 169 articles (55%) stated clearly how the variables were entered into the model. A survey of published epidemiologic studies found that “some information regarding” participation rates was provided in only 59% of case-control and cross-sectional studies, and only 32% of cohort studies.³³ A study of 132 articles published in cancer journals found that only about one-half provided information about length of follow-up.³³ A recent study of abstracts of 222 articles published in leading medical journals found that when relative risks were reported, absolute risks were reported in 62% of RCT abstracts but only 21% of cohort study abstracts.³³ In Pocock et al.’s (2004) study of epidemiological publications in prominent journals, most exposures (e.g., dietary habits, lifestyle factors, or genetic markers) were grouped into categories without any stated rationale for the groupings, and most articles failed to explain the variables used to adjust for confounding; the authors described as “serious” the “risk that some epidemiological publications reach misleading conclusions.”⁴⁸

In some instances, the poor quality of the extant literature

has become an impetus for the development of guideline documents. In a particularly troubling example, the TREND group was formed in 2003 after a group of researchers tried unsuccessfully to synthesize multiple studies of behavioral interventions in patients with human immunodeficiency virus. So many study reports had “failed to include critical information (e.g., intervention timing and dosage, effect size data)” that the researchers were unable to complete their work.³⁴

Flaws of this type may have clinical consequences. A discussion of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system for ranking types of evidence notes that the recommendation of hormone replacement therapy to reduce cardiovascular risk in post-menopausal women, which had been “dutifully applied” by primary care physicians for about a decade, was based on observational research that ultimately was shown to have produced erroneous findings. “Had a rigorous system of rating the quality of evidence been applied at the time,” noted the GRADE observers in 2008, “it would have shown that because the data came from observational studies with inconsistent results, the evidence for a reduction in cardiovascular risk was of very low quality.”⁴⁹

In a similar sequence of events described by Jefferson and Di Pietrantonj in 2007, studies conducted in the decades prior to 2006, employing mostly observational designs, documented a widely accepted association between use of influenza vaccines in persons aged 65 and older and reduced all-cause death rate. When Jefferson and Di Pietrantonj pointed out that vaccine use was associated with only the all-cause death rate, not the death rate from pneumonia or influenza, making the association biologically implausible and probably due to confounding, they became the target of “much support and some scorn ... Eminent immunologists told readers that our interpretations were evidently false.”⁵⁰ Yet, in 2007, a review of the evidence concluded that the observational study results were attributable to selection bias, and sided with Jefferson and Di Pietrantonj.⁵⁰ These examples are unfortunately not unique; a 1999 examination of the methodological quality of clinical practice guidelines found that only 13% of the guideline development documents had graded their recommendations according to strength of the supporting evidence.⁵¹

Problems in Quality of Evidence: Preventable with Insistence on Better Reporting

Perhaps the best evidence that quality problems are preventable comes from research studies in which the authors directly reported data that either called the authors’ conclusions into question or entirely refuted them, yet the data were not reflected in the interpretation appearing in the study report.

For example, in a comparison of entecavir with lamivudine in the treatment of hepatitis B, Chang et al. reported that the study’s primary end point outcome, the rate of “histologic improvement,” was higher for entecavir than for lamivudine (72% vs. 62%;

$P=0.009$).⁵² However, the study’s primary data table showed that the percentages of patients with “no improvement” did not significantly differ by drug, and the counts of the “histologic improvement” and “no improvement” subgroups did not sum to the total number of study patients. A footnote to the data table explained the discrepancy; the denominators used in the calculations included patients who had no follow-up biopsy specimens, while the numerators excluded them. Excluding patients with missing follow-up specimens from both the denominator and the numerator, the percentages of cases with histologic improvement were 77% for entecavir and 72% for lamivudine, yielding a non-significant P value of 0.204 for the comparison (Fisher’s exact test). Nonetheless, the study abstract reported superiority for entecavir.⁵³

Yank et al. reported a similar problem in their systematic review of meta-analyses of antihypertensive medications that had been published from 1983-2004. When they compared studies funded by a single drug manufacturer with those funded by other sources (i.e., multiple funding sources, non-profit, and undisclosed), a financial tie to a drug manufacturer was not significantly associated with favorable *results* (i.e., objective findings, odds ratio [OR]=0.65, $P=0.25$). However, manufacturer funding was significantly associated with favorable *conclusions* (i.e., interpretation, OR=4.09, $P=0.016$). Yank et al. opined that their results reveal “a failure of peer review. Both editors and peer reviewers must have read manuscript versions of those meta-analyses containing discordant results and conclusions, yet they did not prevent publication of biased conclusions.”⁵⁴

Notably, as an example of the need for data analyses by academic statisticians, the *JAMA* editorial on pharmaceutical industry influence cited a secondary analysis of a 2005 trial of the use of rofecoxib in patients with mild cognitive impairment.^{6,21,55} As the secondary analysis noted, the 2005 report did not reflect the higher all-cause mortality rate for rofecoxib compared with placebo, even though this difference had been disclosed in internal manufacturer documents about 4 years previously.⁶ Yet, examination of the original study report suggests problems in reporting and calculation. The sample selection flow chart shows the numbers of patients in each group who experienced an “adverse event,” but does not show how many of those events were fatalities. Additionally, the report’s text discloses that in the “off-drug follow-up” period, for which data were available for “less than half of the patients,” the death rates were 4.8% (17/356) for rofecoxib and 1.6% (5/307) for placebo; the statistical significance of this difference (Fisher’s Exact Test $P=0.029$) was not reported.⁵⁵

Attention to Quality: The Best Solution to Quality Problems

It is tempting to think that if research is simply turned over to smart people who have no overt commercial interest in its outcome, concern over the quality of research evidence will become a thing of the past. Such a solution requires little real effort on the part of information consumers; simply ensure that

the right people are doing the work and settle back to enjoy the fruits of their labors. However, in an environment in which basic standards for quality of evidence are widely available and yet commonly ignored, this approach suggests abrogation of responsibility by information consumers, as well as peer reviewers and editors. We believe that the solution to the current problems with the quality of peer-reviewed health care literature does not lie solely, or even primarily, in a focus on “whodunnit” but rather in renewed attention to basic standards for quality of evidence. These standards represent nothing more complicated or difficult to understand than lessons repeated in undergraduate-level methodology classes that have been taught for decades, coupled with a little common sense. Using these standards does not take a degree in research methodology or statistics; it takes only the commitment to adhere to them.

As the editors of *Journal of General Internal Medicine* sagely observed 3 years ago in discussing the effect of pharmaceutical manufacturer funding on the research literature, “perfection is an asymptote, a goal that can be approximated, but never reached.”¹⁶ Yet, to strive for the goal of disseminating only high-quality evidence, it is necessary to assess distance from the goal. Comparisons of research standards with the characteristics of much of the published health care literature suggest that we have a long way to go.

A cynical perspective on health care research, although understandable, is neither inevitable nor wise. What is needed is critical analysis of the quality of the evidence, regardless of who did the research. It is more important to determine *what evidence* to trust than *whom* to trust. In managed care, the central focus should be on the customers – the patients, pharmacists, physicians and other care givers – who deserve accurate and transparent information about the expected clinical and economic effects of treatment options. All stakeholders, but especially patients, have a lot to lose when studies of treatment options are not validated against accepted standards for conducting and reporting research.

Authors

KATHLEEN A. FAIRMAN, MA, is Associate Editor and Senior Methodology Reviewer of the *Journal of Managed Care Pharmacy*.
FREDERIC R. CURTISS, PhD, RPh, CEBS, is Editor-in-Chief of the *Journal of Managed Care Pharmacy*.

AUTHOR CORRESPONDENCE: Kathleen A. Fairman, MA, Kathleen Fairman LTD, PO Box 31278, Phoenix, AZ 85046. Tel.: 602.867.1343; E-mail kathleenfairman@qwest.net.

DISCLOSURES

The authors report no conflicts of interest related to the subjects or products discussed in this article.

Rethinking the "Whodunnit" Approach to Assessing the Quality of Health Care Research – A Call to Focus on the Evidence in Evidence-Based Practice

REFERENCES

1. Fairman KA, Curtiss FR. Making the world safe for evidence-based policy: Let's slay the biases in research on value-based insurance design. *J Manag Care Pharm.* 2008;14(2):198-204. Available at: www.amcp.org/data/jmcp/JMCPMaga_March%2008_198-204.pdf.
2. Fugh-Berman A. The corporate coauthor. *J Gen Intern Med.* 2005;20:546-48.
3. Martinson BC, Anderson MS, de Vries R. Scientists behaving badly. *Nature.* 2005;435:737-38.
4. Smith R. Medical journals are an extension of the marketing arm of pharmaceutical companies. *PLoS Medicine.* 2005;2(5):0364-66. Available at: www.medicine.plosjournals.org/archive/1549-1676/2/5/pdf/10.1371_journal.pmed.0020138-L.pdf. Accessed August 27, 2008.
5. Ross JS, Hill KP, Egilman DS, Krumholz HM. Guest authorship and ghost-writing in publications related to rofecoxib. *JAMA.* 2008;299(15):1800-12.
6. Psaty BM, Kronmal RA. Reporting mortality findings in trials of rofecoxib for Alzheimer Disease or cognitive impairment. *JAMA.* 2008;299(15):1813-17.
7. Associated Press. Vioxx studies used drugmaker ghostwriters. MSNBC. April 15, 2008. Available at: www.msnbc.msn.com/id/24136335/print/1/displaymode/1098/.
8. Laine C, Mulrow CD, Exorcising ghosts and unwelcome guests. *Ann Intern Med.* 2005;143:611-12.
9. Sismondo S. Ghost management: how much of the medical literature is shaped behind the scenes by the pharmaceutical industry? *PLoS Medicine.* 2007;4(9):1429-33. Available at: www.medicine.plosjournals.org/archive/1549-1676/4/9/pdf/10.1371_journal.pmed.0040286-L.pdf. Accessed August 31, 2008.
10. Hill KP, Ross JS, Egilman DS, Krumholz HM. The ADVANTAGE seeding trial: a review of internal documents. *Ann Intern Med.* 2008;149:251-58.
11. Edelman JM. An open letter to the editors of *The Annals of Internal Medicine.* August 18, 2008. Available at: www.merck.com/newsroom/vioxx/pdf/Advantage_letter.pdf. Accessed August 27, 2008.
12. Bombardier C, Laine L, Reicin A, et al. for the VIGOR Study Group. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. *N Engl J Med.* 2000;343(21):1520-28.
13. Curfman GD, Morrissey S, Drazen JM. Expression of concern: Bombardier et al., "Comparison of Upper Gastrointestinal Toxicity of Rofecoxib and Naproxen in Patients with Rheumatoid Arthritis." (editorial) *N Engl J Med.* 2005;353:2813-14.
14. Bombardier C, Laine L, Burgos-Vargas R, et al. Response to expression of concern regarding VIGOR study. (letter) *N Engl J Med.* 2006;354:1196-98.
15. Saul S. Doctor accused of leak to drug maker. *New York Times.* January 31, 2008.
16. Tierney WM, Gerrity MS. Scientific discourse, corporate ghostwriting, journal policy, and public trust. *J Gen Intern Med.* 2005;20:550-51.
17. Ridker PM, Torres J. Reported outcomes in major cardiovascular clinical trials funded by for-profit and not-for-profit organizations: 2000-2005. *JAMA.* 2006;295(19):2270-74.
18. Nieto A, Mazon A, Pamies R, et al. Adverse effects of inhaled corticosteroids in funded and nonfunded studies. *Arch Int Med.* 2007;167(19):2047-53.
19. Luft HS, Flood AB, Escarce JJ. New policy on disclosures at Health Services Research. *Health Serv Res.* 2006;41(5):1721-32.
20. Editorial policy and author attestations. *J Manag Care Pharm.* Available at: www.amcp.org/amcp.ark?p=A44214B7.
21. DeAngelis CD, Fontanorosa PB. Impugning the integrity of medical science: the adverse effects of industry influence. *JAMA.* 2008;299(15):1833-35.
22. Nissen SE, Nicholls SJ, Sipahi I, et al. for the ASTEROID investigator group. Effect of very high-intensity statin therapy on regression of coronary atherosclerosis: The ASTEROID trial. *JAMA.* 2006;295:1556-65.
23. Mello MM, Clarridge BR, Studdert DM. Academic medical centers' standards for clinical-trial agreements with industry. *N Engl J Med.* 2005;352(21):2202-10.
24. Holtorf AP, Watkins JB, Mullins CD, Brixner D. Incorporating observational data into the formulary decision-making process – summary of a roundtable discussion. *J Manag Care Pharm.* 2008;14(3):302-08. Available at: www.amcp.org/data/jmcp/JMCPMaga_April08_302-308.pdf.
25. Agency for Healthcare Research and Quality. Systems to Rate the Strength of Scientific Evidence. Summary, Evidence Report/Technology Assessment: Number 47. AHRQ Publication No. 02-E015, March 2002. Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services. Rockville, MD. Available at: www.ahrq.gov/clinic/epcsums/strengthsum.htm. Accessed July 29, 2008.
26. West S, King V, Carey TS et al. Systems to Rate the Strength of Scientific Evidence: Evidence Report/Technology Assessment: Number 47. AHRQ Publication No. 02-E016, April 2002. Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services. Available at www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat1.chapter.70996. Accessed July 29, 2008.
27. International Committee of Medical Journal Editors. Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication. Updated October, 2007. Available at www.icmje.org/icmje.pdf. Accessed July 29, 2008.
28. Wager E, Field EA, Grossman L. Good publication practice for pharmaceutical companies. *Curr Med Res Opin.* 2003;19(3):149-54.
29. International Society for Pharmacoeconomics and Outcomes Research. ISPOR Code of Ethics. March, 2008. Available at www.ispor.org/workpaper/CodeOfEthics.asp. Accessed July 29, 2008.
30. Peterson AM, Nau DP, Cramer JA, Benner J, Gwadrý-Sridhar FG, Nichol M. A checklist for medication compliance and persistence studies using retrospective databases. *Value Health.* 2007;10(1):3-12. Available at: www.ispor.org/workpaper/research_practices/Peterson%20Scientific%20Report.pdf. Accessed August 2, 2008.
31. Moher BR, Brooks J, Clark MA et al. A checklist for retrospective database studies – report of the ISPOR Task Force on Retrospective Databases. *Value Health.* 2003;6(2):90-97. Available at: www.ispor.org/workpaper/research_practices/A%20Checklist%20for%20Retrospective%20Database%20Studies-Retrospective%20Database%20Studies.pdf. Accessed August 2, 2008.
32. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: A proposal for reporting. *JAMA.* 2000;283(15):2008-12.
33. Vandenberghe JP, von Elm E, Altman DG, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. *Ann Intern Med.* 2007;147:W-163-W-194. Available at: www.annals.org/cgi/reprint/147/8/W-163.pdf. Accessed August 2, 2008.
34. Des Jarlais DC, Lyles C, Crepaz N, and the TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *Am J Public Health.* 2004;94(3):361-66. Available at: www.ajph.org/cgi/reprint/94/3/361. Accessed August 2, 2008.
35. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med.* 2001;134(8):663-94. Available at: www.annals.org/cgi/reprint/134/8/663.pdf. Accessed July 29, 2008.
36. Ramsey S, Wilke R, Briggs A, et al. Good research practices for cost-effectiveness analysis alongside clinical trials: The ISPOR RCT-CEA Task Force report. *Value Health.* 2005;8(5):521-33. Available at: www.ispor.org/workpaper/research_practices/Good%20Research%20Practices-Cost%20Effectiveness%20Analysis%20with%20Clinical%20Trials.pdf. Accessed August 2, 2008.
37. Miller J. *The Chicago Guide to Writing About Numbers.* Chicago, IL: University of Chicago Press;2004.

Rethinking the “Whodunnit” Approach to Assessing the Quality of Health Care Research – A Call to Focus on the Evidence in Evidence-Based Practice

38. Miller J. *The Chicago Guide to Writing About Multivariate Analysis*. Chicago, IL: University of Chicago Press;2005.
39. Gold MR, Siegel JE, Russell LB, Weinstein MC. *Cost-Effectiveness in Health and Medicine*. New York, NY: Oxford University Press;1996.
40. Weinstein MC, O'Brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in health-care evaluation: Report of the ISPOR Task Force on Good Research Practices – Modeling Studies. *Value Health*. 2003;6(1):9-17. Available at: www.ispor.org/workpaper/research_practices/Principles%20of%20Good%20Practice%20for%20Decision%20Analytic%20Modeling-Modeling%20Studies.pdf. Accessed August 2, 2008.
41. Ofman JJ, Sullivan SD, Neumann PJ, et al. Examining the value and quality of health economic analyses: implications of utilizing the QHES. *J Manag Care Pharm*. 2003;9(1):53-61. Available at: www.amcp.org/data/jmcp/Formulary%20Management-53-61.pdf.
42. Acquadro C, Berzon R, Dubois D, et al. Incorporating the patient's perspective into drug development and communication: an ad hoc task force report of the Patient-Reported Outcomes (PRO) Harmonization Group meeting at the Food and Drug Administration, February 16, 2001. *Value Health*. 2003;6(5):522-31. Available at: www.ispor.org/workpaper/research_practices/PROConcepts_Definitions.pdf. Accessed August 31, 2008.
43. Moher D, Cook DJ, Eastwood S, et al. for the QUOROM Group. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet*. 1999;354:1896-900.
44. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Radiology*. 2003;226:24-28.
45. Stockl KM, Tijoe D, Gong S, Stroup J, Harada AS, Lew HC. Effect of an intervention to increase statin use in Medicare members who qualified for a medication therapy management program. *J Manag Care Pharm*. 2008;14(6):532-40. Available at: www.amcp.org/data/jmcp/JMCPMaga_532-540.pdf.
46. Cook TD, Campbell DT. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin Company; 1979.
47. Altman DG. Poor-quality medical research: what can journals do? *JAMA*. 2002;287(21):2765-67.
48. Pocock SJ, Collier TJ, Dandreo KJ et al. Issues in the reporting of epidemiological studies: a survey of recent practice. *BMJ*. 2004;329(7471):883. Available at: www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=15469946. Accessed August 28, 2008.
49. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336:924-26.
50. Jefferson T, Di Pietrantonj C. Inactivated influenza vaccines in the elderly – are you sure? *Lancet*. 2007;370(9594):1199-1200.
51. Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. *JAMA*. 1999;281(20):1900-05.
52. Chang TT, Gish RG, de Man R, et al., for the BEHoLD A1463022 Study Group. A comparison of entecavir and lamivudine for HBeAg-positive chronic hepatitis B. *N Engl J Med*. 2006;354(10):1001-10.
53. Fairman KA, Curtiss FR. It's only a pharmacoeconomic model - believe it or not. *J Manag Care Pharm*. 2008;14(1):83-85. Available at: www.amcp.org/data/jmcp/JMCPMaga_JanFeb%2008_083-085.pdf
54. Yank V, Rennie D, Bero LA. Financial ties and concordance between results and conclusions in meta-analyses: retrospective cohort study. *BMJ*. 2007;335(7631):1202-05.
55. Thal LJ, Ferris SH, Kirby L, et al. A randomized, double-blind, study of rofecoxib in patients with mild cognitive impairment. *Neuropsychopharmacology*. 2005;30:1204-15.