

UNITED STATES DISTRICT COURT
MIDDLE DISTRICT OF FLORIDA
TAMPA DIVISION

Robert L. Vazzo, LMFT, et al.,

Case No. 8:17-cv-02896-WFJ-AAS

Plaintiffs,

v.

City of Tampa, Florida,

Defendant.

DEFENDANT CITY OF TAMPA’S REPLY IN SUPPORT OF MOTION TO EXCLUDE EXPERT TESTIMONY OF BERNARDO O. HUDSON, M.D. AND CHRISTOPHER ROSIK, PH.D. AND INCORPORATED MEMORANDUM OF LAW

Defendant, City of Tampa, Florida (“City”), by and through counsel, hereby files its Reply to Plaintiffs’ Response to the City’s Motion to Exclude Expert Testimony of Bernardo O. Hudson, M.D. and Christopher Rosik, Ph.D. and Incorporated Memorandum of Law (“Motion to Exclude”). For the following reasons, this Court should exclude the expert testimony of Plaintiffs’ proposed experts, Bernardo O. Hudson, M.D. (“Dr. Hudson”) and Christopher Rosik, Ph. D. (“Dr. Rosik”).

INTRODUCTION

There is no dispute that over the past two decades, every major medical and mental health professional association in the United States has warned practitioners and the public that conversion therapy has no scientific foundation, is not supported by any scientifically reliable evidence of efficacy, and puts minors at risk of serious harms, including suicide. Plaintiffs’ experts reject that medical consensus and opine that conversion therapy should be practiced on minors because, in their opinions, there is no definitive proof that it causes harm.

Because those claims are not based on any scientifically valid principles or methods, as required by Federal Rule of Evidence 702, the Court should exclude their opinions.

ARGUMENT

I. PLAINTIFFS' EXPERTS FAILS TO SATISFY THE STANDARDS FOR ADMISSIBILITY UNDER FEDERAL RULE OF EVIDENCE 702.

Plaintiffs' experts claim that there is no "empirical" evidence that conversion therapy is harmful to minors and that, to be meaningful, any such evidence must be dispositive. Plaintiffs' experts are wrong on both counts. Plaintiffs' experts conflate the scientific standard for proving that a treatment works with the very different standard used to assess harm. While the former generally requires controlled randomized studies, the latter does not. To the contrary, when assessing whether a treatment is safe, *any* evidence of harm must be considered, including patient reports, and the more serious the harms the more weight such evidence must be given. Here, the evidence that conversion therapy puts minors at risk of serious harm is not only substantial, but compelling that it should not be used on minors under any standard of review.

A. There Is Substantial Evidence That Conversion Therapy Puts Minors at Risk of Serious Harms

Plaintiffs dismiss the medical consensus that conversion therapy is harmful to minors as mere "nose-counting." But that argument disregards the *reason* that such a strong medical consensus exists—namely, the substantial body of empirical evidence that conversion therapy is not just harmful to minors, it, in fact, puts them at risk of extremely serious harms, including death by suicide, suicide attempts, serious depression, and other lifelong negative health impacts. (Doc. 24-2, p. 377.) Plaintiffs' assertion that this evidence is not "empirical" rests on

a misunderstanding of that term, which simply means “[o]f, relating to, or based on experience, experiment, or observation.” Black’s Law Dictionary (11th ed. 2019). Here, the strong medical consensus that conversion therapy is harmful to minors is based on decades of observation and experience, including studies, clinical experience, and patient and provider reports. (Doc. 190, ¶ 21.)

The 2009 APA Task Force undertook a systematic review of then-existing scientific literature on conversion therapy. (Doc. 24-1, p. 394.) The Task Force “found no empirical evidence that providing any type of therapy in childhood can alter adult same-sex orientation,” and that “scientific evidence shows that SOCE is not likely to produce its intended outcomes and can produce harm for some of its participants.” (Doc. 24-2, p. 447.) It also found that patients who had undergone conversion therapy “reported negative social and emotional consequences,” including “self-reports of anger, anxiety, confusion, depression, grief, guilt, hopelessness, deteriorated relationships with family, loss of social support, loss of faith, poor self-image, social isolation, intimacy difficulties, intrusive imagery, suicidal ideation, self-hatred, and sexual dysfunction.” (Doc. 24-2, p. 377.) As a result, it concluded that therapists should not perform conversion therapy on minors. (Doc. 24-2, p. 447-48.)

In 2015, the Substance Abuse and Mental Health Services Administration of the U.S. Department of Health and Human Services conducted an updated survey of existing research on conversion therapy and published a report and recommendations based on “consensus statements developed by experts in the field after a careful review of existing research, professional health association reports and summaries, and expert clinical guidance.” (Doc. 24-4, p. 544.) The report found “none of the existing research supports the premise that mental health

or behavioral interventions can alter gender identity or sexual orientation.” *Id.* It concluded: “Interventions aimed at a fixed outcome, such as gender nonconformity or heterosexual orientation, including those aimed at changing gender identity, gender expression, and sexual orientation are coercive, can be harmful, and should not be part of behavioral health treatment.” *Id.*

Analyzing “the scientific evidence,” the American Academy of Child and Adolescent Psychiatry found that “[g]iven that there is no evidence that efforts to alter sexual orientation are effective, beneficial, or necessary, and the possibility that they carry the risk of significant harm, such interventions are contraindicated.” (Doc. 24-4, p. 521.) Similarly, the American Medical Association (AMA) found that conversion therapy “may cause significant psychological distress.” For example, the AMA explained that one study showed that 77 percent of ex-SOCE participants reported significant long-term harm, including the following symptoms: depression, anxiety, lowered self-esteem, internalized homophobia, self-blame, intrusive imagery, and sexual dysfunction. Participants also reported significant social and interpersonal harm such as alienation, loneliness, social isolation, interference with intimate relationships and loss of social supports. AMA Issue Brief, <https://www.ama-assn.org/system/files/2019-03/transgender-conversion-issue-brief.pdf>. (Exhibit A.) The AMA also found that:

SOCE may also increase suicidal behaviors in a population where suicide is prevalent. In young adults between 15 and 24 years old, suicide has been the second leading cause of death since 2011, and LGBTQ young adults are more than twice as likely to report a history of suicide attempts in comparison to their heterosexual peers. Similarly, LGB adults are three to five times more likely to have a suicidal attempt in comparison to their heterosexual counterparts. Young LGBTQ adults who report higher levels of parental and caregiver rejection are 8.4 times more likely to report having attempted suicide. One study found nearly 30 percent of individuals that underwent SOCE reported suicidal attempts.

Id. (internal citations omitted). In light of this evidence, “[a]ll leading medical and mental health associations reject ‘conversion therapy’ as a legitimate medical treatment.” *Id.*

Since the Ordinance was enacted, additional research has continued to document the serious harms to minors that are associated with conversion therapy. A 2018 study found that LGBT youth who were subjected to conversion therapy were nearly three times more likely to attempt suicide than LGBT youth who were not subjected to conversion therapy. (Doc. No. 190-3, p. 4093). Indeed, as recently as this month, a study published in the *Journal of American Medicine* found that transgender people who were subjected to conversion therapy as children were four times more likely to report lifetime suicide attempts than the general transgender population. (Doc. 198-1.) The fact that Plaintiffs’ experts simply disagree will not, and does not, change these alarming conclusions.

B. Medical Science Does Not Require Definitive Proof of Harm

Plaintiffs claim that the City must have definitive proof that conversion therapy causes harm before prohibiting its use on minors. But that is *not* the scientific standard for determining harm. While efficacy generally must be proved by the “gold standard” of controlled randomized studies, evidence that a treatment is harmful need not be—and rarely is—based on such research. *See* Chou R, Aronson N, Atkins D, et al. Assessing harms when comparing medical interventions. In: Agency for Healthcare Research and Quality. *Methods Reference Guide for Comparative Effectiveness Reviews* [posted November 2008]. Rockville, MD. Available at <http://effectivehealthcare.ahrq.gov/healthInfo.cfm?infotype=rr&ProcessID=60>. (Exhibit B.)

In order to protect patient safety, harm must be assessed based on “a broad range of data sources,” and *any* evidence of harm—including reports by patients—must be considered. Chou, R., Aronson, N., Atkins, D., Ismaila, A. S., Santaguida, P., Smith, D. H., ... Moher, D. (2010). AHRQ Series Paper 4: Assessing harms when comparing medical interventions: AHRQ and the Effective Health-Care Program. *Journal of Clinical Epidemiology*, 63(5), 502-512. <https://doi.org/10.1016/j.jclinepi.2008.06.007>. (Exhibit C.) That broad standard is especially important for “the most serious adverse events,” such as those at issue here. The more serious the harm, the more weight that must be given to patient or provider reports. *Id.*

Plaintiffs’ experts ignore this distinction between efficacy and safety entirely. But that distinction is critical to protecting patients. While rigorous evidence is required to prove that a treatment works, any significant evidence of serious or unexpected harm must be taken with the utmost seriousness in order to protect patient health and safety—including putting a stop to research that may be putting patients at undue risk. *See Abigail All. for Better Access to Developmental Drugs v. von Eschenbach*, 495 F.3d 695, 698 (D.C. Cir. 2007) (“At any time during the clinical trials, a drug sponsor is required to notify the FDA of “[a]ny adverse experience associated with the use of the drug that is both serious and unexpected,” *id.* § 312.32(c)(1)(A), and the FDA may order a “clinical hold” halting the trials if it determines that safety concerns so warrant, *id.* § 312.42.”).

Here, there is substantial evidence—including multiple studies involving patient reports—that conversion therapy puts minors at risk of serious harms. For this reason, Plaintiffs’ argument that the City must wait for further research is simply wrong. As the district court correctly noted in *Otto*, any such further research on minors would be unethical.

Otto v. City of Boca Raton, Fla., 353 F. Supp. 3d 1237, 1263 (S.D. Fla. 2019) (“Plaintiffs cannot demand a multiyear controlled study in which some minors, even those who would voluntarily seek SOCE, are subjected to SOCE and some are not.”).

Plaintiffs’ experts also ignore the fundamental scientific principle that a treatment “is unsafe if its potential for inflicting death or physical injury is not offset by the possibility of therapeutic benefit.” *U.S. v. Rutherford*, 442 US 542, 556 (1979). That principle is dispositive here, where “[n]o research has been published in the peer-reviewed literature that demonstrates the efficacy of conversion therapy efforts with gender minority youth, nor any benefits of such interventions to children and their families.” (SAMSHA Report, Doc. 24-5, p. 568). As the APA Task Force also found: “The positive experiences clients report in SOCE are not unique, and “the benefits reported by participants in SOCE may be achieved through treatment approaches that do not attempt to change sexual orientation.” (APA report 68).

In sum, conversion therapy provides no unique benefit, but it puts minors at risk of serious harms, up to and including death. This is why the medical community has issued such strong and urgent statements warning of the serious harms associated with conversion therapy and making clear that conversion therapies “should not be part of behavioral health treatments.” (Doc. 24-5, p. 554.) There is no scientific basis for permitting licensed professionals to continue subjecting minors to this dangerous treatment. Rather, doing so would be not only unethical, but it would be unconscionable.

The contrary opinions offered by Plaintiffs’ experts do not pass muster under Rule 702. An expert’s reliability “concerns whether the reasoning or methodology underlying the testimony is scientifically valid and. . . whether that reasoning or methodology properly can be

applied to the facts in issue.” *Seamon v. Remington Arms Co., LLC*, 813 F.3d 983, 988 (11th Cir. 2016) (quoting *Daubert*, 509 U.S. at 592-93). In ascertaining reliability, the court must consider whether an expert’s approach is generally accepted in the scientific community. *Quiet Tech. DC-8, Inc. v. Hurel-Dubois UK Ltd.*, 326 F.3d 1333, 1341 (11th Cir. 2003). While Plaintiffs seek to minimize that factor, it is central to *Daubert*’s test. “Rulings on admissibility under *Daubert* inherently require the trial court to conduct an exacting analysis of the proffered expert’s methodology,” *Quiet Tech. DC-8, Inc.*, 326 F.3d at 1341. Plaintiffs’ experts cannot survive that exacting analysis here. Their approach misapplies the scientific standard for assessing harm and disregards substantial evidence of harm, contrary to the evidence-based consensus of every leading medical and mental professional organization.

CONCLUSION

The testimony and opinions of Dr. Rosik and Dr. Hudson are neither relevant nor reliable. This Court, functioning in its gatekeeper role, is required to admit only reliable and relevant evidence under Rule 702. Because the testimony and opinions of Dr. Rosik and Dr. Hudson fail to meet the requirements under *Daubert* and 702, the City respectfully requests that this Court enter its order excluding Dr. Rosik and Dr. Hudson from testifying or otherwise offering any expert opinions in this case.

Respectfully submitted,

/s/ Robert V. Williams

Robert V. Williams

Florida Bar No. 144720

Dana L. Robbins

Florida Bar No. 106626

Primary: rwilliams@burr.com

Secondary: pturner@burr.com

BURR & FORMAN LLP

201 N. Franklin Street, Ste. 3200
Tampa, Florida 33602
Telephone: (813) 221-2626
Facsimile: (813) 221-7335

Attorneys for Defendants, City of Tampa

CERTIFICATE OF SERVICE

I HEREBY CERTIFY that on this 23rd day of September, 2019, I caused a true and correct copy of the foregoing to be served via electronic mail on counsel for Plaintiff, Horatio G. Mihet (hmihet@lc.org), Roger Gannam (rgannam@lc.org), and Daniel J. Schmid (dscmid@lc.org).

/s/ Robert V. Williams _____
Attorney

LGBTQ change efforts (“conversion therapy”)

Background

“Conversion therapy” refers to any form of interventions, such as individual or group, behavioral, cognitive or milieu/environmental operations, which attempt to change an individual’s sexual orientation or sexual behaviors (sexual orientation change efforts [SOCE]) or an individual’s gender identity (gender identify change efforts [GICE]).¹ Practitioners of change efforts may employ techniques including:

- Aversive conditioning (e.g., electric shock, deprivation of food and liquids, smelling salts and chemically induced nausea)
- Biofeedback
- Hypnosis
- Masturbation reconditioning
- Psychotherapy or systematic desensitization²

Underlying these techniques is the assumption that homosexuality and gender identity are mental disorders and that sexual orientation and gender identity can be changed. This assumption is not based on medical and scientific evidence. Professional consensus rejects pathologizing homosexuality and gender nonconformity, in addition, empirical evidence has demonstrated that homosexuality and variations in gender identity are normal variants of human expression not inherently linked to mental illness. However, the unfounded misconception of sexual orientation and gender identity “conversion” persists today in some health, spiritual and religious practitioners.³

According to the UCLA Williams Institute on Sexual Orientation and Gender Identity Law and Public Policy, as of 2018, almost 700,000 lesbian, gay, bisexual, transgender and queer/questioning (LGBTQ) adults in the U.S. had received “conversion therapy”; in addition, an estimated 57,000 youths will receive change efforts from religious or health care providers before they turn 18 years old.⁴

Health implications for LGBTQ individuals

Evidence does not support the purported “efficacy” of SOCE in changing sexual orientation.⁵ To the contrary, these practices may cause significant psychological distress.⁶ One study showed that 77 percent of ex-SOCE participants reported significant long-term harm, including the following symptoms:

- Depression
- Anxiety
- Lowered self-esteem

1. John Bancroft, et al., *Peer Commentaries on Spitzer*, 32 *Archives of Sexual Behavior* 5, 419-68 (Oct. 2003).

2. American Psychological Association, Report of the American Psychological Association Task Force on Appropriate Therapeutic Responses to Sexual Orientation (Aug. 2009).

3. Jack Drescher, *Ethical issues in treating gay and lesbian patients*, 25 *Psychiatric Clinics of North America* 3, 605-21 (Sep. 2002).

4. Christy Mallory, Taylor Brown & Kerith Conron, The Williams Institute on Sexual Orientation and Gender Identity Law, UCLA School of Law, Conversion therapy and LGBT youth (Jan. 2018)

5. American Psychological Association, *supra* note 2

6. *Id.*

- Internalized homophobia
- Self-blame
- Intrusive imagery
- Sexual dysfunction⁷

Participants also reported significant social and interpersonal harm such as alienation, loneliness, social isolation, interference with intimate relationships and loss of social supports.⁸

SOCE may also increase suicidal behaviors in a population where suicide is prevalent. In young adults between 15 and 24 years old, suicide has been the second leading cause of death since 2011, and LGBTQ young adults are more than twice as likely to report a history of suicide attempts in comparison to their heterosexual peers.⁹ Similarly, LGB adults are three to five times more likely to have a suicidal attempt in comparison to their heterosexual counterparts.¹⁰ Young LGBTQ adults who report higher levels of parental and caregiver rejection are 8.4 times more likely to report having attempted suicide.¹¹ One study found nearly 30 percent of individuals that underwent SOCE reported suicidal attempts.¹²

GICE may cause similar long-term harm as SOCE. According to the American Psychological Association Consensus on Efforts to Change Gender Identity, there is a lack of published research on efforts to change gender identity among children and adolescents. No existing research supports that mental health and behavioral interventions with children and adolescents alter gender identity.¹³

Ethical Concerns

All leading professional medical and mental health associations reject “conversion therapy” as a legitimate medical treatment. In addition to the clinical risks associated with the practice, the means through which providers or counselors administer change efforts violate many important ethical principles, the foremost of which: “First, do no harm.”

A health care provider’s nonjudgmental recognition of and respect for patients’ sexual orientations, sexual behaviors and gender identity are essential elements in rendering optimal patient care in health, as well as in illness. This recognition is especially important to address the specific health care needs of people who are or may be LGBTQ, as these patients often experience disparities in access to care. Yet administering change efforts is an inherently discriminatory practice often administered coercively and fraught with ethical problems, such as:

7. Ariel Shidlo & Michael Schroeder, *Changing Sexual Orientation: A Consumers’ Report*, 33 *Professional Psychology: Research and Practice* 3, 249-59 (2002).

8. *Id.*

9. National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, *10 Leading Causes of Death by Age Group, United States*, <https://www.cdc.gov/injury/wisqars/LeadingCauses.html>; Andrea Miranda-Mendizábal, et al., *Sexual orientation and suicidal behaviour in adolescents and young adults: systematic review and meta-analysis*, 211 *British Journal of Psychiatry* 2, 77-87 (Aug. 2017).

10. Travis Hottes, Laura Bogaert, Anne Rhodes, David Brennan & Dionne Gesink, *Lifetime Prevalence of Suicide Attempts Among Sexual Minority Adults by Study Sampling Strategies: A Systematic Review and Meta-Analysis*, 106 *Am J Public Health* 5, e1-e12 (May 2016).

11. Caitlin Ryan, David Huebner, Rafael Diaz, & Jorge Sanchez, *Family Rejection as a Predictor of Negative Health Outcomes in White and Latino Lesbian, Gay, and Bisexual Young Adults*, 123 *Pediatrics* 1, 346-52 (Jan. 2009).

12. Shidlo, *supra* note 7.

13. *Therapy Supporting and Affirming LGBTQ Youth: Statements of professional consensus regarding sexual orientation and gender identity and expression*, American Psychological Association, <https://www.apa.org/advocacy/civil-rights/sexualdiversity/lgbtq-therapy.aspx>.

- Uninformed consent: change efforts are often prescribed without full descriptions of risks and disclosure of lack of efficacy or evidence
- Breaches of confidentiality: content of treatment, sexual orientation and gender identity may be shared with family, school or religious leaders without proper consent
- Patient discrimination: change efforts reinforce bias, discrimination and stigma against LGBTQ individuals
- Indiscriminate and improper treatment: change efforts are recommended regardless of evidence
- Patient blaming: the failure of treatment may be blamed on the patient.¹⁴

It is clinically and ethically inappropriate for health care providers to direct mental or behavioral health interventions, including SOCE and GICE, with a prescriptive goal aimed at achieving a fixed developmental outcome of a child’s or adolescent’s sexual orientation, gender identity or gender expression.¹⁵

State laws

To date, 14 states (California, Connecticut, Delaware, Hawaii, Illinois, Maryland, Nevada, New Hampshire, New Jersey, New Mexico, Oregon, Rhode Island, Vermont and Washington) and the District of Columbia have enacted laws banning “conversion therapy” for minors. Importantly, these laws do not prohibit counseling and therapies that help patients struggling with sexual or gender identity to develop coping and self-acceptance skills.

Medical society and other healthcare association positions

The American Medical Association and GLMA: Health Professionals Advancing LGBTQ Equality (GLMA) oppose the use of reparative or conversion therapy for sexual orientation or gender identity. Other medical societies have policies or statements similarly opposing these policies, including the American Psychiatric Association, American Academy of Child and Adolescent Psychiatry, American College of Physicians and American Academy of Pediatrics.¹⁶ Other health care associations including the American Association for Marriage and Family Therapy, American Counseling Association, American Psychoanalytic Association, American Psychological Association, National Association of Social Workers, Pan American Health Organization: Regional Office of the World Health Organization have similar policies.¹⁷

AMA policy

H-160.991 Health Care Needs of Lesbian, Gay, Bisexual and Transgender Populations

1. Our AMA: (a) believes that the physician’s nonjudgmental recognition of patients’ sexual orientations, sexual behaviors, and gender identities enhances the ability to render optimal patient care in health as well as in illness. In the case of lesbian, gay, bisexual, transgender, queer/questioning, and other (LGBTQ) patients, this recognition is especially important to address the specific health care needs of people who are or may be LGBTQ; (b) is committed to taking a leadership role in: (i) educating physicians on the current state of research in and knowledge of LGBTQ Health and the need to elicit relevant gender and sexuality information from our patients;

14. *Id.*; Jack Drescher J, et al., *The Growing Regulation of Conversion Therapy*, 102 J Med Regulation 2, 7-12 (Jan 2016).

15. American Psychological Association, *supra* note 13.

16. See American Psychiatric Association, Commission on Psychotherapy by Psychiatrists, *Position Statement on Therapies Focused on Attempts to Change Sexual Orientation (Reparative or Conversion Therapies)*, 157 American Journal of Psychiatry 10, 1719-21 (Oct. 2000); American Academy of Child and Adolescent Psychiatry, The AACAP Policy on “Conversion Therapies (Feb. 2018); Hilary Daniel & Renee Butkus, American College of Physicians, *Lesbian, Gay, Bisexual, and Transgender Health Disparities: Executive Summary of a Policy Position Paper From the American College of Physicians*, 163 Ann Intern Med 2, 135-7 (July 2015); American Academy of Pediatrics, Committee on Adolescence, *Homosexuality and Adolescence*, 92 Pediatrics 4, 631-4 (1993).

17. Policy and Position Statements on Conversion Therapy, Human Rights Campaign, <http://www.hrc.org/resources/policy-and-position-statements-on-conversion-therapy>.



these efforts should start in medical school, but must also be a part of continuing medical education; (ii) educating physicians to recognize the physical and psychological needs of LGBTQ patients; (iii) encouraging the development of educational programs in LGBTQ Health; (iv) encouraging physicians to seek out local or national experts in the health care needs of LGBTQ people so that all physicians will achieve a better understanding of the medical needs of these populations; and (v) working with LGBTQ communities to offer physicians the opportunity to better understand the medical needs of LGBTQ patients; and **(c) opposes, the use of “reparative” or “conversion” therapy for sexual orientation or gender identity.**

2. Our AMA will collaborate with our partner organizations to educate physicians regarding: (i) the need for sexual and gender minority individuals to undergo regular cancer and sexually transmitted infection screenings based on anatomy due to their comparable or elevated risk for these conditions; and (ii) the need for comprehensive screening for sexually transmitted diseases in men who have sex with men; (iii) appropriate safe sex techniques to avoid the risk for sexually transmitted diseases; and (iv) that individuals who identify as a sexual and/or gender minority (lesbian, gay, bisexual, transgender, queer/questioning individuals) experience intimate partner violence, and how sexual and gender minorities present with intimate partner violence differs from their cisgender, heterosexual peers and may have unique complicating factors.
3. Our AMA will continue to work alongside our partner organizations, including GLMA, to increase physician competency on LGBTQ health issues.
4. Our AMA will continue to explore opportunities to collaborate with other organizations, focusing on issues of mutual concern in order to provide the most comprehensive and up-to-date education and information to enable the provision of high quality and culturally competent care to LGBTQ people. (CSA Rep. C, I-81; Reaffirmed: CLRPD Rep. F, I-91; CSA Rep. 8 - I-94; Appended: Res. 506, A-00; Modified and Reaffirmed: Res. 501, A-07; Modified: CSAPH Rep. 9, A-08; Reaffirmation A-12; Modified: Res. 08, A-16; Modified: Res. 903, I-17; Modified: Res. 904, I-17) (emphasis added)

GLMA policy

GLMA 099-97-114 Reparative or Conversion Therapy

GLMA: Health Professionals Advancing LGBTQ Equality condemns the behavioral and psychological interventions known as “reparative” or “conversion” therapies that attempt to change sexual orientation and gender identity. (Approved 1997; Amended & Reaffirmed pending final GLMA Board approval 2018)

For additional information or assistance with legislation to ban conversion therapy in your state, please visit www.ama-assn.org/go/arc or contact Annalia Michelman, JD, Senior Legislative Attorney, AMA Advocacy Resource Center at annalia.michelman@ama-assn.org or (312) 464-4788.



Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews

Assessing Harms When Comparing Medical Interventions



Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

Comparative Effectiveness Reviews are systematic reviews of existing research on the effectiveness, comparative effectiveness, and harms of different health care interventions. They provide syntheses of relevant evidence to inform real-world health care decisions for patients, providers, and policymakers. Strong methodologic approaches to systematic review improve the transparency, consistency, and scientific rigor of these reports. Through a collaborative effort of the Effective Health Care (EHC) Program, the Agency for Healthcare Research and Quality (AHRQ), the EHC Program Scientific Resource Center, and the AHRQ Evidence-based Practice Centers have developed a Methods Guide for Comparative Effectiveness Reviews. This Guide presents issues key to the development of Comparative Effectiveness Reviews and describes recommended approaches for addressing difficult, frequently encountered methodological issues.

The Methods Guide for Comparative Effectiveness Reviews is a living document, and will be updated as further empiric evidence develops and our understanding of better methods improves. Comments and suggestions on the Methods Guide for Comparative Effectiveness Reviews and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

This document was written with support from the Effective Health Care Program at AHRQ. Dr. Moher is supported by a University of Ottawa Research Chair. None of the authors has a financial interest in any of the products discussed in this document.

Suggested citation: Chou R, Aronson N, Atkins D, et al. Assessing harms when comparing medical interventions. In: Agency for Healthcare Research and Quality. Methods Reference Guide for Comparative Effectiveness Reviews [posted November 2008]. Rockville, MD. Available at: <http://effectivehealthcare.ahrq.gov/healthInfo.cfm?infotype=rr&ProcessID=60>.

This report has also been published in edited form: Chou R, Aronson N, Atkins D, et al. Assessing harms when comparing medical interventions: AHRQ and the Effective Health-Care Program. *J Clin Epidemiol* 2008 Sep 25 [Epub ahead of print].

Assessing Harms When Comparing Medical Interventions

Authors:

Roger Chou, M.D.^a

Naomi Aronson, Ph.D.^b

David Atkins, M.D., M.P.H.^c

Afisi S. Ismaila^d

Pasqualina Santaguida, Ph.D.^d

David H. Smith, Ph.D.^{a,e}

Evelyn Whitlock, M.D., M.P.H.^{a,e}

Timothy J. Wilt, M.D., M.P.H.^f

David Moher, Ph.D.^g

^aOregon Evidence-based Practice Center, Oregon Health & Science University, Portland, OR.

^bBlue Cross Blue Shield Evidence-based Practice Center, Blue Cross Blue Shield Association, Chicago, IL.

^cDepartment of Veterans Affairs, Washington, DC.

^dMcMaster Evidence-based Practice Center, McMaster University, Hamilton, ON.

^eOregon Evidence-based Practice Center, Kaiser Center for Health Research, Portland, OR.

^fMinnesota Evidence-based Practice Center, Minneapolis VA Center for Chronic Disease Outcomes Research, MN.

^gUniversity of Ottawa Evidence-based Practice Center, University of Ottawa, Ottawa, ON.

The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ, the Veterans Health Administration, or the Health Services Research and Development Service. Therefore, no statement in this report should be construed as an official position of these entities, the U.S. Department of Health and Human Services, or the U.S. Department of Veterans Affairs.

Assessing Harms When Comparing Medical Interventions

Key Points

- Assess all important harms, whenever possible.
- Use multiple sources of information, including clinical experts and stakeholders, to identify important harms.
- Use consistent and precise terminology when reporting data on harms, and avoid terms implying causality unless causality is reasonably certain.
- Gather evidence on harms from a broad range of sources, including observational studies, particularly when clinical trials are lacking; when generalizability is uncertain; or when investigating rare, long-term, or unexpected harms.
- Do not assume studies adequately assess harms because methods used to assess and report benefits are appropriate; rather, evaluate how well studies identify and analyze harms.
- Be cautious about drawing conclusions on harms when events are rare and estimates of risk are imprecise.
- Include placebo-controlled trials, particularly for assessing uncommon or rare harms, but be cautious about relying on indirect comparisons to judge comparative risks, and evaluate whether studies being considered for indirect comparisons meet assumptions for consistency of treatment effects.
- Avoid inappropriate combining of data on harms, and thoroughly investigate inconsistent results.

Introduction

Comparative Effectiveness Reviews (CERs) are systematic reviews that evaluate evidence on alternative interventions in order to help clinicians, policymakers, and patients make informed treatment choices.¹ To generate balanced results and conclusions, it is important for CERs to address both benefits and harms.² However, assessing harms can be difficult. Benefits have been accorded greater prominence when reporting trials, with little effort to balance assessments of benefits and harms. In addition, systematically reviewing evidence for all possible harms is often impractical, as interventions may be associated with dozens of potential adverse events. Furthermore, there are often important tradeoffs between increasing comprehensiveness and decreasing quality of harms data.³

Adequately assessing harms requires CER authors to consider a broad range of data sources. For that reason, they need to deal with important challenges, such as choosing which types of evidence to include, identifying studies of harms, assessing their quality, and summarizing and synthesizing data from different types of evidence.

Identifying Harms To Be Evaluated

CERs should always assess harms that are important to decisionmakers and users of the intervention under consideration.⁴ High-priority harms should include the most serious adverse events; they may also include common adverse events and other specific adverse events important to clinicians and patients. CER authors should examine previously published reviews, review publicly available safety reports from the U.S. Food and Drug Administration (FDA), and consult with technical experts and patients to set priorities for evaluating harms. Searches on postmarketing surveillance databases may also help identify important potential harms. The methods sections of the CER should specify the process used to identify harms of interest and list the specific harms for which evidence was sought.

Terminology

Terminology related to reporting of harms is poorly standardized.⁵ This can cause confusion or result in misleading conclusions. CER authors should strive for consistent and precise usage of terminology when reporting data on harms. For example, the term “harms” is generally preferred over the term “safety” because the latter sounds more reassuring and may obscure important concerns. “Harms” is also preferable to the term “unintended effects,” which could refer to either beneficial or harmful outcomes. Terms that do not imply causality (such as “adverse events”) should be the default term to describe harms, unless causality is reasonably certain.

Definitions for commonly used terms for harms reporting are summarized in Table 1, along with suggested usage.⁴⁻⁶

Sources of Evidence on Harms

Randomized Controlled Trials

Published trials. Properly designed and executed randomized controlled trials (RCTs) are considered the “gold standard” for evaluating efficacy because they minimize potential bias. However, relying solely on published RCTs to evaluate harms in CERs is problematic. First, most RCTs lack prespecified hypotheses for harms.⁵ Rather, hypotheses are usually designed to evaluate beneficial effects, with assessment of harms a secondary consideration. As such, the quality and quantity of harms reporting in clinical trials is frequently inadequate.^{7,8}

Second, few RCTs have large enough sample sizes or are long enough in duration to adequately assess uncommon or long-term harms.⁹

Third, most RCTs are explanatory, rather than pragmatic, in design—i.e., they assess benefits and harms in ideal, homogeneous populations and settings.¹⁰ Patients who are more susceptible to adverse events are often underrepresented in such “efficacy” trials. Even when harms are appropriately assessed and reported, the applicability of efficacy trials to general practice is limited.

Fourth, relatively few RCTs directly compare alternative treatment strategies. Although CER authors can evaluate benefits or harms of two competing interventions based on trials in which each is compared with a common third treatment (usually placebo), the results of indirect comparisons do not always agree with direct comparisons.^{11,12}

Fifth, publication and selective outcome(s) reporting bias can lead to distorted conclusions about harms when data are unpublished, partially reported, downplayed, or omitted.^{13,14}

Finally, in some cases, RCTs may not be available. For example, surgical procedures and medical devices often become widely disseminated with few or no randomized trial data. The same can be true for older therapeutic devices, such as hyperbaric oxygen chambers.¹⁵

Despite these limitations, RCTs are the gold standard for demonstrating efficacy, the basis for most regulatory approvals, and the source of most advertising and other claims made on behalf of drugs and other interventions. For this reason, CERs must address harms data from RCTs in detail when they are available.

“Head-to-head” RCTs provide the most direct evidence on comparative harms. However, placebo-controlled RCTs may also provide important information on absolute and relative risks and contribute to more precise estimates of harms. In addition, placebo-controlled trials can provide information about risks that may not be apparent from head-to-head trials. For example, a systematic review of nonsteroidal anti-inflammatory drugs (NSAIDs) found cyclo-oxygenase-2 selective NSAIDs associated with greater myocardial risk vs. placebo, but differences were not apparent vs. nonselective NSAIDs, which were also associated with increased risk.¹⁶ In general, CERs should routinely include placebo-controlled trials for assessment of harms, particularly for rare or uncommon adverse events. In lieu of examining individual placebo controlled trials, CERs may incorporate findings of well-conducted systematic reviews, provided they evaluate the specific harms of interest.

Unpublished supplemental trials data. In addition to evaluating results of published RCTs, CER authors should consider including results of completed or terminated but unpublished RCTs, as well as unpublished results from published trials. Such information has several potentially valuable uses:

- To assess the number of unpublished trials or frequency of unreported outcomes, which can help in evaluating risk for publication or outcomes reporting bias.
- To evaluate whether conclusions based on unpublished data are qualitatively different from those based on published RCTs.
- To conduct formal quantitative meta-analysis, including published and unpublished RCTs or outcomes.

Unpublished clinical trials tend to report lower estimates of treatment benefits than published trials (i.e., weaker intervention effects).^{17,18} The impact of unpublished trials on assessments of harms has not been extensively studied, but a systematic review of antidepressants in children found that addition of data from unpublished trials changed conclusions about the balance of risks and benefits from favorable to unfavorable for several drugs.¹⁹

Data from unpublished trials can be difficult to locate systematically. At a minimum, material from the FDA Web site should routinely be examined in order to assess what effect unpublished (completed or terminated) trials submitted for regulatory approval may have on conclusions regarding harms. In addition, starting in 2009, trial sponsors are required by the 2007 FDA reform bill to report results to a clinical trial results database (www.ClinicalTrials.gov).²⁰ Other resources for identifying unpublished trials include obtaining information from non-U.S. regulatory agencies and directly querying funding sources. Once unpublished trials are located, two caveats should also be considered. Frequently, there is insufficient information from unpublished trials to assess fully the risk of bias. Also, the results and conclusions of trials may change between initial presentation of data and publication in a peer-reviewed journal.²¹

Even when a trial is published, important information may be omitted because of space limitations or other reasons.^{22,23} For example, before the publication of the Vioxx Gastrointestinal Outcomes Research Study (VIGOR) in 2001,²⁴ information on myocardial infarctions was absent from most published reports of trials evaluating selective or nonselective NSAIDs because an association with cardiovascular events was not suspected. A systematic review that obtained unpublished myocardial infarction data from older trials found an increased risk with high doses of all evaluated NSAIDs (selective or nonselective) other than naproxen.¹⁶ An analysis of myocardial infarction risk based on only published information would have been seriously compromised by incomplete data.

Drug approval information—for example, the clinical and statistical reviews prepared by staff of the FDA—frequently provides details about harms not included in journal publications. For example, the Celecoxib Long-term Arthritis Safety Study (CLASS), a major trial of celecoxib, was published in the *Journal of the American Medical Association* as a 6-month study and reported fewer gastrointestinal adverse events for celecoxib than for two nonselective NSAID comparators.²⁵ The JAMA article did not mention that some patients in the trial had been observed for longer than 6 months.²⁶ In contrast, the FDA review reported all the outcomes data, including data that showed no difference in gastrointestinal adverse events at the end of followup.²⁷

Limited evidence suggests an inverse relationship between the proportion of included trials reporting a specific outcome and the estimates of treatment benefit for that outcome, possibly due to selective reporting of favorable outcomes.²⁸ How the proportion of included trials reporting outcomes affects estimates of harms has not been well studied. Nonetheless, when a significant proportion of published trials fail to report an important or critical adverse event, CER authors should report on this gap in the evidence and consider efforts to obtain unpublished data (e.g., by querying study authors, funding sources, or clinical trials results databases, or performing more detailed reviews of FDA documents).

Observational Studies

Observational studies are almost always necessary to assess harms adequately. The exception is when there are sufficient data from RCTs to reliably estimate harms. However, even though observational studies are more susceptible to bias than well-conducted RCTs, for some comparisons there may be few or no long-term, large, head-to-head, or effectiveness RCTs.²⁹ Observational studies may also provide the best (or only) evidence for evaluating harms in minority or vulnerable populations (such as pregnant women, children, elderly patients, or those with multiple comorbidities) who are underrepresented in clinical trials.

The term “observational studies” is commonly used to refer to cohort, case-control, and cross-sectional studies,³⁰ but can refer to a broad range of study designs, including case reports, uncontrolled series of patients receiving surgery or other interventions, and others.³¹ All can yield useful information as long as their specific limitations are understood.

The types of observational studies included in a CER will vary depending on the type or frequency of adverse events being evaluated. The choice of study designs also depends on whether investigators are seeking to determine what harms might be associated with a treatment (hypothesis generating) or whether certain harms are more likely (hypothesis testing). Different types of observational studies might be included or rendered irrelevant by availability of data from stronger study types.

Cohort and case-control studies. CER authors should routinely search for and include well-designed and reported case-control and population-based cohort studies.^{30,32} Such studies are well suited for testing hypotheses on whether one intervention is associated with a greater risk for an adverse event than is another and for quantifying the risk. They also take stronger precautions against bias than do other observational designs, and their strengths and weaknesses are well understood. For unexpected adverse events, for example, confounding by indication may not be as important an issue in case-control and cohort studies as when evaluating beneficial effects because their occurrence is usually not associated with the reasons for choosing a particular treatment.^{29,33} Although cross-sectional studies have features in common with cohort studies, it is difficult to establish causality because exposures, and outcomes are evaluated simultaneously. Indeed, associations in cross-sectional studies may sometimes be due to reverse causality.³⁴

A recent report found that large observational studies usually report smaller absolute risks of harm than do large randomized trials.³⁵ There was no clear tendency for randomized trials or observational studies to report larger relative risks. In more than one-half of the comparisons assessed, estimates of relative or absolute risk varied more than twofold. Discrepancies between randomized trials and observational studies may occur because of differences in populations, settings, or interventions; differences in study design, including criteria used to identify harms; differential effects of biases, or some combination of these factors.

Observational studies based on patient registries. Patient registries collect information on clinical outcomes in populations defined by a particular disease, condition, or exposure.³⁶ Clinical data are prospectively collected for specific research purposes using active methods to identify outcomes, although registry information can be supplemented by information from administrative databases and other sources. Registries can be designed as an active surveillance system for identifying harms and may be particularly useful for assessing long-term or uncommon adverse events.

Observational studies based on analyses of large databases. Pharmacoepidemiologic studies using large databases to identify exposures and outcomes may be valuable for comparing the risk of uncommon adverse events.³⁷ However, additional empirical research is needed to identify methods for collecting and analyzing data in pharmacoepidemiologic studies that are associated with valid findings.³⁸ Unlike studies based on patient registries, large administrative databases usually contain information routinely collected during clinic, hospital, laboratory, or pharmacy encounters, rather than for a specific research purpose. Such studies are probably most useful for evaluating serious harms that are more reliably reported and recorded (for example, death or acute myocardial infarction) than less serious harms that may not generate a specific clinic visit or diagnostic code (for example, sedation or nausea). In some cases, administrative data may be supplemented or verified by more detailed clinical information. Regardless of how data are obtained, all observational studies should employ appropriate methods for minimizing bias and misclassification of data.

Case reports and postmarketing surveillance. About 30 percent of the primary published literature on adverse drug events is in the form of case reports.³⁹ Case reports can be useful for identifying uncommon, unexpected, or long-term adverse events, particularly for new drugs or other interventions.⁴⁰ The adverse events identified by case reports often differ from those detected in clinical trials.⁴¹ However, case reports are usually considered to be hypothesis generating because it is difficult to calculate information from them about the frequency or comparative risk of adverse events.

In the United States, the FDA receives about 280,000 reports of postmarketing adverse events annually, collects them into a database,⁴² and issues information about adverse drug events on its MedWatch Web site (<http://www.fda.gov/medwatch/>). Although pharmaceutical companies and other investigators may also perform passive surveillance of harms on postmarketing data, such analyses are not always made public in a timely fashion.⁴³ Active, hypothesis-driven postmarketing surveillance systems have been developed recently for identifying and evaluating serious adverse drug events.⁴⁴

Case reports and other hypothesis-generating studies may be useful for CERs evaluating new drugs suspected of being associated with serious but uncommon adverse events. For other topics, CER authors may consider their inclusion on a case-by-case basis.

Other observational studies. Several other types of observational studies may also report data on harms. However, they are likely to be more prone to bias than RCTs or well-designed case-control or cohort studies, and their use needs to be considered cautiously. For example, studies reporting harms from surgical or other invasive interventions often consist of a series of patients who received the procedure. Data are often insufficient to assess the methods used to select participants.⁴⁵ In addition, because such studies lack control groups, evaluating effects of confounding is difficult, as is comparing risks of adverse events across interventions.

Other quasi-experimental study designs may not offer any advantage over RCTs in terms of their applicability to routine practice. For example, open-label extensions of clinical trials may follow patients for an extended period of time, but they usually enroll a more highly selected population (patients who completed the randomized trial, tolerated the medication, and agreed to participate in the extension), are unblinded, and often lack a comparison arm. Such studies can be excluded from CERs if more reliable long-term, comparative data are available. If they are included in CERs, their limitations should be described clearly.

Criteria to select observational studies for inclusion. In general, many more observational studies than randomized trials will be available for nearly all health care interventions. Evaluating a large number of observational studies can be impractical when conducting a CER, especially when a significant proportion either do not add useful information or carry a high risk of reporting biased results.

Several criteria have commonly been used in systematic reviews and CERs to screen observational studies of harms for inclusion. Empirical data are lacking on how use of different selection criteria affects estimates of harms. However, CERs should match inclusion criteria to the reasons for including observational studies. For example, inclusion criteria might specify minimum duration of followup if a priority is to identify evidence on long-term harms. If large, higher quality studies are available, it could be reasonable to specify a minimum sample size threshold in order to utilize resources efficiently. Methods sections should clearly describe selection criteria along with the rationale for choosing the criteria. Commonly used inclusion criteria for observational studies are shown in Table 2.

Assessing Risk of Bias (Quality) of Harms Reporting

Randomized Trials

A number of features of RCTs have been empirically tested and proposed as markers of higher quality (i.e., lower risk of bias). These include use of appropriate randomization generation and allocation concealment techniques; blinding of participants, health care providers, and outcomes assessors; and analysis according to intention-to-treat principles.⁴⁶ Whether these are equally important in protecting against bias in studies reporting harms is unclear. Moreover, because evaluating harms is often a secondary consideration in randomized trials, the quality of harms assessment and reporting can be inadequate even when assessment of the primary (beneficial) outcome is appropriate.

When evaluating the quality of harms assessment, CER authors should consider whether adequate methods were used to identify adverse events in the primary studies. Active methods, such as querying patients using a comprehensive checklist or standardized laboratory tests, are more likely to completely identify adverse events than passive methods, such as relying on patient self-report.⁴⁷ In addition, specific data on adverse events are likely to be more accurate and informative than generic statements, such as “no adverse events were noted” or “the interventions were well tolerated.” If a specific adverse event is not reported, it is generally safer for CER authors to assume that they were not ascertained or not recorded than to assume that the prevalence or incidence was zero.⁴

It is also important to assess how adverse events are assessed and categorized. Studies should predefine the qualifiers “serious” and “severe” to describe adverse events. Otherwise, it is impossible for readers to determine whether these labels were applied consistently within and across trials. Standardized criteria for grading severity of adverse events are available for certain conditions.^{48,49} CERs should note when grading severity or seriousness of adverse events is based on nonstandardized or poorly defined criteria, as such classifications may not be comparable across studies or may be poorly reproducible. Similarly, methods for classifying adverse events as “treatment related” are largely subjective, with unknown validity, and such data may be particularly unreliable.

It is not always necessary for trials to prespecify or define adverse events. For example, studies reporting unexpected outcomes can be very valuable for identifying previously unrecognized harms. However, when evaluating known harms, using validated or standardized criteria for adverse events may help reduce subjectivity or bias in their assessment and classification. In drug trials, use of an independent external endpoint committee may provide less biased estimates of harms than outcomes assessment performed by investigators connected to the study.⁵⁰

“Withdrawals due to adverse events” are commonly reported in trials, and they are often used in systematic reviews as a marker for intolerable or severe adverse events. However, the Cochrane Adverse Effects Methods Group suggests caution in interpreting withdrawals attributed to adverse events in this manner, for the following reasons:⁴

- Attribution of reasons for discontinuation is likely to be imprecise and to vary across trials.
- Pressure to keep dropouts low in trials may result in rates that do not reflect real-world practice.

- Unblinding often takes place before the decision to withdraw, which can lead to distortion of estimates of an intervention's effect on withdrawal (e.g., symptoms are less likely to lead to withdrawal if the patient is found to be on placebo).

Nonetheless, withdrawals due to adverse events are often reported even when serious or severe adverse events are not reported or are poorly defined, and they may provide some useful information.

Observational Studies

Because observational studies lack randomization, they should adhere to high methodological standards to be considered valid.^{30,32,51} RCTs are expected to have outcomes recorded by blinded personnel and to include all participants who were randomized in the analysis of results. Use of blinded outcome assessors and a clearly identified inception cohort (e.g., “new users”)⁵² is at least as important when assessing observational studies.

Instruments for assessing risk of bias in observational studies vary greatly in scope, number and types of items used, and developmental rigor.⁵³ Further study is needed to determine which methodological shortcomings in observational studies are consistently associated with bias in assessment and reporting of harms. However, some consensus exists on the major domains that should be considered when evaluating the overall validity of an observational study. For cohort studies, important factors include assembly of an inception cohort, complete followup, appropriate assessment of potential confounders, accurate determination of exposures and outcomes, and blinded assessment of outcomes.^{30,52-54}

Several studies have empirically evaluated effects of specific methodological characteristics on estimates of harms from observational studies. They found that prospective or retrospective design,^{55,56} case-control compared with cohort studies^{57,58} and smaller compared with larger case series⁵⁵ did not have consistent effects on estimates of harms. Two studies found that industry-funded studies tended to report more favorable outcomes than did studies with other funding sources.^{57,59} Because all of these studies evaluated fairly limited samples of studies, their wider applicability is uncertain.

Observational studies based on evaluations of large administrative databases should follow the same general principles to reduce bias as observational studies that directly collect data from patients. In these cases, reviewers should pay particular attention to the methods used for ascertaining exposures and outcomes and for measuring and analyzing potential confounders, as these issues are more likely to be problematic in studies relying on administrative claims (although not unique to them).³⁷

For all observational study designs, estimates of harms are less likely to be confounded when evaluating previously unsuspected adverse events than when evaluating a known harm or intended effects. For example, the finding that cyclo-oxygenase-2-selective NSAIDs were associated with an increased risk of myocardial infarction vs. nonselective NSAIDs was an unexpected finding from an RCT examining a different outcome.²⁴ This risk could be confirmed in observational studies, in part because the choice of type of NSAID in typical practice was unrelated to the patients' risk for myocardial infarction. In contrast, gastrointestinal bleeding was a known risk of nonselective NSAIDs, and clinicians were more likely to prescribe selective NSAIDs in patients at higher risk for gastrointestinal bleeding. Such “confounding by indication” led to the appearance of an apparent association between selective NSAID use and

bleeding in epidemiologic studies.⁶⁰ In some cases, such spurious associations may remain despite adjustment for known confounders (“residual confounding”).

Uncontrolled Studies

Studies of surgery, medical devices, and other nonpharmacologic interventions are often uncontrolled series of patients who received the therapy and then were followed over a period of time. Such studies can provide some information about rates of adverse events in clinical practice, and they may be most informative when the incidence of such events in untreated patients is low. Unfortunately, such studies frequently do not meet standards for accurate and comprehensive reporting of harms.⁶¹ Even when harms data are well described, an important limitation of uncontrolled studies is that it is difficult to evaluate confounding by indication. Authors are also more likely to submit for publication studies showing the best outcomes.

For some interventions, CER authors must consider including uncontrolled studies for assessing harms, as little or no other evidence may be available. Proposed criteria for evaluating case series are likely to promote improved reporting of results,⁶² but may provide only limited information about risk of bias. Important factors to consider when evaluating uncontrolled studies include whether the study enrolled or attempted to enroll all patients meeting prespecified inclusion criteria and whether the study clearly describes loss to followup.⁴⁵ When uncontrolled studies do not meet these criteria, determining the reliability and applicability of even well-described results may be impossible.

Instruments for Assessing Risk of Bias (Quality) in Studies on Harms

Development of instruments for assessing risk of bias specifically in studies of harms is still in an early stage of development. Two issues remain unclear: whether to use a specific rating instrument to evaluate harms assessment and reporting, or whether using instruments for rating the overall risk of bias of a study is sufficient, as long as particular attention is paid to how well adverse events are defined, ascertained, and reported.

Chou et al. empirically developed and tested an instrument for assessing quality of harms assessment and reporting in randomized trials and observational studies of carotid endarterectomy for symptomatic carotid artery stenosis.⁶³ This approach involved four criteria: nonbiased selection of subjects, low loss to followup, adverse events prespecified and defined, and adequate duration of followup. Studies meeting at least three of the four criteria reported a rate of postsurgical complications of 5.7 percent (95-percent confidence interval [CI], 4.8 percent to 6.6 percent), compared with 3.7 percent (95-percent CI, 3.1 percent to 4.3 percent) for studies meeting fewer than three of the criteria. However, the generalizability of this instrument to other datasets or interventions is unclear. When the authors applied these criteria to studies of rofecoxib, they were unable to show differences in estimates of risk of myocardial infarction. In addition, caution should be used when considering use of summary scores to assess risk of bias.⁶⁴ At a minimum, key methodological aspects should be assessed individually and their influence on estimates of harms explored.

Santaguida et al. have also developed a quality-rating instrument (McHarm) for evaluating studies reporting harms (Table 3).⁶⁵ The tool was developed from quality rating items generated by a review of the literature on harms and from previous quality assessment instruments. A formal Delphi consensus exercise was used to reduce the number of items. The subsequent list

of quality criteria specific to harms was tested for reliability and face, construct, and criterion validity. This quality-assessment tool is intended for use in conjunction with standardized quality-assessment tools for design-specific internal validity issues.

Case reports may provide valuable information about the possibility of rare or previously unrecognized adverse events. A 1982 study examined 47 case reports published in 1963 in four major general medical journals and judged that 35 of them were subsequently proved to be “clearly” correct.⁶⁶ However, the methods used to determine reliability of case reports in this study were subjective, and results have not been replicated. A recent study, in fact, found that only 18 percent of case reports of suspected adverse drug reactions have been subjected to rigorous evaluation in subsequent studies.⁶⁷ Nonetheless, statistical modeling study suggests that the likelihood of more than one to three spontaneously reported cases is very unlikely to be coincidental when the adverse event is rare or uncommon.⁶⁸ Case reports, however, cannot be used to estimate the rate of an adverse event, which may be critical to any decisions.

Several disease-specific⁶⁹ and non-disease-specific⁷⁰ methods for assessing the probability of causality from case reports of adverse events have been developed. These methods represent expert opinion and have not been validated empirically. Factors believed to increase the likelihood of causality are shown in Table 4.^{69,70}

Guidelines for improving the reporting of suspected adverse drug events in case reports have also recently been proposed.⁷¹ In 35 reports of 48 patients published in the *British Medical Journal*, the median number of recommended items that were reported was 9 of 19 (range 5-12), although effects of missing information on the validity of case reports have not been studied.

Synthesizing Evidence on Harms

CER authors should follow general principles for synthesizing evidence when evaluating data on harms. Such principles include: combining studies only when they are similar enough to warrant combining;⁷² adequately considering risk of bias, including publication and other related biases;⁷³ and exploring potential sources of heterogeneity.²³ Several other issues are especially relevant for synthesizing evidence on harms.

Uncommon or Rare Adverse Events

Evaluating comparative risks of uncommon or rare adverse events in CERs can be particularly challenging. A frequent problem in RCTs and systematic reviews is interpreting a nonsignificant probability value as indicating no difference in risk for rare adverse events, particularly when the confidence intervals are wide and encompass the possibility of clinically important risks.^{74,75} For example, one trial concluded that, in patients with meningitis, “treatment with dexamethasone did not result in an increased risk of adverse events” compared with placebo for treatment of hyperglycemia, herpes zoster, or fungal infection because P values for all three outcomes were more than 0.20.⁷⁶ However, the 95-percent confidence intervals for estimates of relative risks for these three adverse events encompassed clinically significant increases in risk (-13.5 percent to 77.6 percent, -60.4 percent to 377.7 percent, and -43.6 percent to 496.2 percent, respectively). In such cases, CERs should acknowledge the lack of statistical power to assess risk adequately and should interpret the confidence intervals, including the possibility or probability of excess harm.

Equivalence and Noninferiority

CER authors should draw conclusions about “equivalence” or “noninferiority” of interventions with regard to harms only when there are appropriate data to justify such statements.⁷⁷ Few CERs will have the statistical power to adequately assess noninferiority when the risk of an adverse event is on the order of 1 percent or lower. For example, about 100,000 patients would have been needed in the COBALT or GUSTOIII trials to rule out an excess relative death rate of 5 percent from alternative thrombolytic agents with 80-percent power.⁷⁸ Ruling out smaller event rates would require even higher sample sizes.

Indirect Analyses

Placebo-controlled trials can be helpful for evaluating absolute risks associated with an intervention. When head-to-head trials are sparse or unavailable, placebo-controlled trials may also be useful for indirectly evaluating comparative harms, particularly for rare or uncommon adverse events. However, for indirect analyses to be reliable, all studies should be comparable in terms of quality, factors related to applicability (population, dosing, co-interventions, and settings), measurement of outcomes, and incidence of adverse events in control groups.^{12,79}

For example, a meta-analysis found that rofecoxib was associated with an increased risk of arrhythmia compared with control treatments; celecoxib was not.⁸⁰ However, the rate of arrhythmia in the control arms was tenfold higher in trials of celecoxib (0.27 percent, or 18 of 6,568 subjects) than in trials of rofecoxib (0.02 percent, or 2 of 10,174 subjects). In this situation, indirect comparisons about the relative safety of celecoxib compared with rofecoxib are likely to be problematic. A more informative approach would be to explore reasons for the discrepancies in rates of arrhythmias in the control arms and how they may have affected comparisons.

More studies are needed to determine when indirect comparisons are most likely to be valid. In the meantime, CER authors considering indirect analyses to assess harms should carefully consider whether assumptions underlying valid indirect comparisons are likely to be met, compare results of indirect comparisons with head-to-head data if available, and draw conclusions from indirect comparisons cautiously.

Combining Data From Different Types Of Studies

Most CERs will include data on harms from different types of studies. Statistical combination of data from observational studies is often inappropriate and should be avoided unless there is a clear rationale to do so.⁸¹ If such analyses are undertaken, the justification should be clearly explained.

Discrepancies Between Randomized Trials and Observational Studies

A separate challenging situation occurs when results on harms from randomized trials and observational studies are discordant. Some reasons for discrepancies between randomized trials and observational studies are shown in Table 5. A reasoned analysis of potential sources of discrepancy is generally more helpful than simply presenting the different results.

Reporting Evidence on Harms

As when reporting evidence on benefits, CERs should emphasize the most reliable information for the most important adverse events. Summary tables should generally present data for the most important harms first, with more reliable evidence preceding less reliable evidence. Evidence on harms from each type of study should be clearly summarized in summary tables, narrative format, or both.² A critical role of CERs is to report clearly on the limitations of the evidence on harms and to analyze and interpret thoughtfully how these limitations may affect estimates of the balance of benefit and harm. Suggested elements to focus on when reporting harms are shown in Table 6.

Summary

A summary of the key points about assessment of harms discussed in this report is shown in Table 7.

Acknowledgments

The authors would like to acknowledge Gail R. Janes for participating in the workgroup calls.

References

- ¹ Lohr KN. Emerging methods in comparative effectiveness and safety: symposium overview and summary. *Med Care* 2007;45(10 Suppl 2):S5–S8.
- ² GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ* 2004;328:1490.
- ³ McIntosh HM, Woolacott NF, Bagnall A.-M. Assessing harmful effects in systematic reviews. *BMC Med Res Meth* 2004;4:19.
- ⁴ Loke YK, Price D, Herxheimer A. Systematic reviews of adverse effects: framework for a structured approach. *BMC Med Res Methodol* 2007;7:32.
- ⁵ Ioannidis JPA, Evans SJW, Gotzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Annals of Internal Medicine* 2004;141(10):781–788.
- ⁶ Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. *Lancet* 2000;356(9237):1255–1259.
- ⁷ Ioannidis JPA, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *JAMA* 2001;285(4):437–43.
- ⁸ Loke Y, Derry S. Reporting of adverse drug reactions in randomised controlled trials—a systematic survey. *BMC Clin Pharmacol* 2001;1:3.
- ⁹ Vandenbroucke JP. Benefits and harms of drug treatments. *BMJ* 2004;329(7456):2–3.
- ¹⁰ Rothwell PM. External validity of randomised controlled trials: “to whom do the results of this trial apply?” *Lancet* 2005;365(9453):82–93.

- 11 Chou R, Fu R, Huffman LH, Korthuis PT. Initial highly-active antiretroviral therapy with a protease inhibitor versus a non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses. *Lancet* 2006;368(9546):1503–15.
- 12 Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ* 2003;326(7387):472.
- 13 Chan A, Hrobjartsson A, Haahr M, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials. *JAMA* 2004;291(20):2457–65.
- 14 Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;337(8746):867–72.
- 15 McDonagh M, Helfand M, Carson S, Russman BS. Hyperbaric oxygen therapy for traumatic brain injury: a systematic review of the evidence. *Arch Phys Med Rehabil* 2004;85(7):1198–204.
- 16 Kearney PM, Baigent C, Godwin J, Halls H, Emberson JR, Patrono C. Do selective cyclo-oxygenase-2 inhibitors and traditional non-steroidal anti-inflammatory drugs increase the risk of atherothrombosis? Meta-analysis of randomized trials. *BMJ* 2006;332:1302–1308.
- 17 Egger M, Juni P, Bartlett C, Hohenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technology Assessment (Winchester, England)* 2003;7(1):1–76.
- 18 Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med* 2008;358:252–60.
- 19 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.
- 20 Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible research: moving toward research the public can really trust. *Ann Intern Med* 2007;146(6):450–3.
- 21 Toma M, McAlister FA, Bialy L, Adams D, Vandermeer B, Armstrong PW. Transition from meeting abstract to full-length journal article for randomized controlled trials. *JAMA* 2006;295(11):1281–7.
- 22 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):2270–2274.
- 23 Sterne JA, Egger M, Smith GD. Investigating and dealing with publication and other biases in meta-analysis. *BMJ* 2001;323(7304):101–5.
- 24 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 [see comment]. *New England Journal of Medicine* 2000;343(21):1520–8.
- 25 Silverstein FE, Faich G, Goldstein JL, et al. Gastrointestinal toxicity with celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis: the CLASS study: a randomized controlled trial. Celecoxib Long-term Arthritis Safety Study [see comment]. *JAMA* 2000;284(10):1247–55.
- 26 Hrachovec JB, Mora M. Reporting of 6-month vs 12-month data in a clinical trial of celecoxib. *JAMA* 2001;286(19):2398.
- 27 Witter J. Medical review part 1. Center for Drug Evaluation and Research. http://www.fda.gov/cder/foi/nda/2002/20-998S009_Celebrex_medr_P1.pdf. Accessed on April 3, 2008.

- 28 Furukawa TA, Watanabe N, Montori VM, Guyatt GH. Association between unreported outcomes and effect size estimates in Cochrane meta-analyses [letter]. *JAMA* 2007;297(5):468–70.
- 29 Vandembroucke JP. When are observational studies as credible as randomised trials? *Lancet* 2004;363(9422):1728–31.
- 30 von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandembroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med* 2007;147(8):573–7.
- 31 Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research. Principles and Quantitative Methods*. Belmont, CA: Wadsworth; 1982.
- 32 Vandembroucke 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(8):W163–94.
- 33 Psaty BM, Koepsell T, Lin D, et al. Assessment and control for confounding by indication in observational studies. *Journal of the American Geriatrics Society* 1999;47(6):749–54.
- 34 Rothman KJ, Greenland S. *Modern epidemiology*. 2nd ed. Philadelphia, PA: Lippincott-Raven; 1998.
- 35 Papanikolaou P, N, Christidi GD, Ioannidis J. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *CMAJ* 2006;174(5):635–41.
- 36 Gliklich R, Dreyer NA, eds. *Registries for evaluating patient outcomes: a user's guide*. AHRQ Publication NO. 07-EHC001-1. Rockville, MD: Agency for Healthcare Research and Quality; 2007.
- 37 Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of Clinical Epidemiology* 2005;58:323–337.
- 38 Sturmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration—a simulation study. *Amer J Epidemiol* 2007;165(10):1110–18.
- 39 Aronson JK, Derry S, Loke YK. Adverse drug reactions: keeping up to date. *Fundamental and Clinical Pharmacology* 2002;16:49–56.
- 40 Stricker BH, Psaty BM. Detection, verification, and quantification of adverse drug reactions. *BMJ* 2004;329(7456):44–7.
- 41 Loke YK, Derry S, Aronson JK. A comparison of three different sources of data in assessing the frequencies of adverse reactions to amiodarone. *British Journal of Clinical Pharmacology*. 2004;57(5):616–21.
- 42 Strom BL. Potential for Conflict of Interest in the Evaluation of Suspected Adverse Drug Reactions: A Counterpoint. *JAMA* 2004;292(21):2643–6.
- 43 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.
- 44 Bennett CL, Nebeker JR, Lyons EA, et al. The Research on Adverse Drug Events and Reports (RADAR) project. *JAMA* 2005;293(17):2131–40.
- 45 Oleson O. 2. Types of study design. The Cochrane Non-Randomised Studies Methods Group (NRSMG); 1999. <http://www.cochrane.dk/nrsmg/docs/chap2.pdf>. Accessed April 3, 2008.
- 46 Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001;323(7303):42–6.
- 47 Bent S, Padula A, Avins AL. Brief communication: better ways to question patients about adverse medical events: a randomized, controlled trial. *Annals of Internal Medicine* 2006;144(4):257–61.

- 48 NCI. Common Terminology Criteria for Adverse Events v3.0 (CTCAE); 2006. http://ctep.cancer.gov/reporting/ctc_v30.html. Accessed April 3, 2008.
- 49 NIAID. Division of AIDS table for grading the severity of adult and pediatric adverse events; 2004. <http://www3.niaid.nih.gov/research/resource/s/DAIDSclinRsrch/Safety/>. Accessed April 3, 2008.
- 50 Sydes MR, Spiegelhalter DJ, Altman DG, Babiker AB, Parmar MKB. Systematic qualitative review of the literature on data monitoring committees for randomized controlled trials. *Clinical Trials* 2004;1:60–79.
- 51 Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. 1974;66(5):688–701.
- 52 Rochon PA, Gurwitz JH, Sykora K, et al. Reader's guide to critical appraisal of cohort studies: 1. Role and design. *BMJ* 2005;330(7496):895–7.
- 53 Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies *Health Technology Assessment (Winchester, England)* 2003;7(27):iii–x, 1–173.
- 54 West S, King V, Carey TS. Systems to rate the strength of scientific evidence. Agency for Healthcare Research and Quality; Rockville, MD; 2002.
- 55 Dalziel K, Round A, Stein K, Garside R, Castelnuovo E, Payne L. Do the findings of case series studies vary significantly according to methodological characteristics? *Health Technol Assessment* 2005;9(2):1–146.
- 56 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):260–265.
- 57 Juni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *Lancet* 2004;364(9450):2021–9.
- 58 Ofman JJ, MacLean CH, Straus WL, et al. A metaanalysis of severe upper gastrointestinal complications of nonsteroidal antiinflammatory drugs [see comment]. *Journal of Rheumatology* 2002;29(4):804–12.
- 59 Shah RV, Albert TJ, Buegel-Sanchez V, Vaccaro AR, Hilibrand AS, Gauer JN. Industry support and correlation to study outcome for papers published in Spine. *Spine* 2005;30:1099–1104.
- 60 Laporte JR, Ibanez L, Vidal X, Vendrell L, Leone R. Upper gastrointestinal bleeding associated with the use of NSAIDs: new versus older agents. *Drug Safety* 2004;27(6):411–420.
- 61 Martin RCG, Brennan MF, Jacques DP. Quality of complication reporting in the surgical literature. *Annals of Surgery* 2002;235:803–813.
- 62 Carey TS, Boden SD. A critical guide to case series reports. *Spine* 2003;28:1631–1634.
- 63 Chou R, Fu R, Carson S, Saha S, Helfand M. Methodological shortcomings predicted lower harm estimates in one of two sets of studies of clinical interventions. *J Clin Epidemiol* 2006;60(1):18–28.
- 64 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.
- 65 Santaguida PL, Raina P. The Development of the McHarm Quality Assessment Scale for adverse events: Delphi Consensus on important criteria for evaluating harms. 2008. <http://hiru.mcmaster.ca/epc/mcharm.pdf>. Accessed May 14, 2008.
- 66 Venning GR. Validity of anecdotal reports of suspected adverse drug reactions: the problem of false alarms. *BMJ* 1982;284:249–52.

- 67 Loke YK, Price D, Derry S, Aronson JK. Case reports of suspected adverse drug reactions—systematic literature survey of follow-up. *BMJ* 2006;332(7537):335–9.
- 68 Begaud B, Moride Y, Tubert-Bitter P, Chaslerie A, Haramburu F. False-positives in spontaneous reporting: should we worry about them? *British Journal of Clinical Pharmacology* 1994;38(5):401–4.
- 69 Danan G, Benichou C. Causality assessment of adverse reactions to drugs—I. A novel method based on the conclusions of international consensus meetings: application to drug-induced liver injuries. *J Clin Epidemiol* 1993;46(11):1323–30.
- 70 Michel DJ, Knodel LC. Comparison of three algorithms used to evaluate adverse drug reactions. *American Journal of Hospital Pharmacy* 1986;43(7):1709–14.
- 71 Aronson JK. Anecdotes as evidence. *BMJ* 2003;326:1346.
- 72 Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. *Annals of Internal Medicine* 1997;127(9):820–6.
- 73 Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;352(9128):609–13.
- 74 Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuses of power when interpreting results. *Ann Intern Med* 1994;121:200–6.
- 75 Jonville-Bera AP, Giraudeau B, Autret-Leca E. Reporting of drug tolerance in randomized clinical trials: when data conflict with authors' conclusions. *Ann Intern Med* 2006;144:306–7.
- 76 de Gans J, van de Beek D. Dexamethasone in adults with bacterial meningitis. *N Engl J Med* 2002;347:1549–56.
- 77 Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ, CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. *JAMA* 2006;295(10):1152–60.
- 78 Ware JH, Antman EM. Equivalence trials. *N Engl J Med* 1997;337(16):1159–61.
- 79 Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology* 1997;50(6):683–91.
- 80 Zhang J, Ding EL, Song Y. Adverse Effects of Cyclooxygenase 2 Inhibitors on Renal and Arrhythmia Events: Meta-analysis of Randomized Trials. *JAMA* 2006;296:1619–32.
- 81 Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. *BMJ* 1998;316(7125):140–4.

Table 1. Terminology for reporting on harms

Active surveillance of harms.	Participants are asked in structured questionnaires or interviews about the occurrence of specific adverse events, or predefined laboratory or other diagnostic tests are performed at prespecified time intervals.
Adverse effect	A harmful or undesirable outcome that occurs during or after the use of a drug or intervention for which there is at least a reasonable possibility of a causal relation.
Adverse event	A harmful or undesirable outcome that occurs during or after the use of a drug or intervention but is not necessarily caused by it. When causality is uncertain or the purpose of the Comparative Effectiveness Review is to establish causality, “adverse event” should generally be the default term over “adverse effect” or “adverse reaction/adverse drug reaction.”
Adverse reaction/adverse drug reaction	An adverse effect specifically associated with a drug.
Complications	A term often used to describe adverse events following surgery or other invasive interventions.
Harms	The totality of all possible adverse consequences of an intervention.
Passive surveillance of harms	Participants are not specifically asked about or tested for the occurrence of adverse events. Rather, adverse events are identified based on patient reports made on their own initiative.
Risk-benefit ratio	A common expression for the comparison of overall harms and benefits. However, because benefits and harms of an intervention are usually very different in character and are measured on different scales, a true “risk-benefit ratio” is rarely calculable. In addition, there may be several distinct benefits and harms. A preferred term is “ <i>balance of benefits and harms.</i> ”
Safety	Substantive evidence of an absence of harm. Do not use this term (or the term “safe”) when evidence on harms is simply absent or insufficient.
Serious adverse event	Any adverse event with serious medical consequences, including death, hospital admission, prolonged hospitalization, and persistent or significant disability or incapacity.
Severe adverse event	An adverse event whose intensity is considered severe (including “nonserious” adverse events). For example, a rash could be “severe” but not “serious” (i.e., not resulting in death, hospital admission, prolonged hospitalization, or persistent or significant disability).
Side effects	Unintended drug effects (beneficial or harmful) given at doses normally used for therapeutic effects. Use of this term may tend to understate the importance of harms because the word “side” may be perceived to suggest secondary importance.
Tolerability	This term is often used imprecisely but should be used to refer to a patient’s or subject’s ability or willingness to tolerate or accept unpleasant drug-related adverse events without serious or permanent sequelae.
Toxicity	The term “toxicity” is used in pharmacology and microbiology to refer to the quality of being poisonous, especially the degree of virulence of a toxic microbe or of a poison. It is often measured in terms of the specific target affected (e.g., cytotoxicity or hepatotoxicity). In the context of systematic reviews, the term is often used to refer to laboratory-determined abnormalities, such as elevated liver function tests. However, the terms “abnormal laboratory measurements” and “laboratory abnormalities” are more specific and appropriate.

Table 2. Example criteria for selecting observational studies on harms for inclusion in a Comparative Effectiveness Review

<ul style="list-style-type: none"> • Studies meet certain study design definitions (e.g., cohort and case-control studies)
<ul style="list-style-type: none"> • Studies do not exceed a defined threshold for risk of bias (e.g., studies assessed as being at low risk of bias or meeting certain prespecified quality criteria)
<ul style="list-style-type: none"> • Studies meet a defined threshold for duration of followup
<ul style="list-style-type: none"> • Studies meet a sample size threshold
<ul style="list-style-type: none"> • Studies evaluate a specific population of interest (e.g., studies evaluating populations underrepresented in randomized trials, such as elderly, women, or minority populations)

Table 3. McMaster tool for assessing quality of harms assessment and reporting in study reports (McHarm)

1. Were the harms PRE-DEFINED using standardized or precise definitions?
2. Were SERIOUS events precisely defined?
3. Were SEVERE events precisely defined?
4. Were the number of DEATHS in each study group specified OR were the reason(s) for not specifying them given?
5. Was the mode of harms collection specified as ACTIVE?
6. Was the mode of harms collection specified as PASSIVE?
7. Did the study specify WHO collected the harms?
8. Did the study specify the TRAINING or BACKGROUND of who ascertained the harms?
9. Did the study specify the TIMING and FREQUENCY of collection of the harms?
10. Did the author(s) use STANDARD scale(s) or checklist(s) for harms collection?
11. Did the authors specify if the harms reported encompass ALL the events collected or a selected SAMPLE?
12. Was the NUMBER of participants that withdrew or were lost to follow-up specified for each study group?
13. Was the TOTAL NUMBER of participants affected by harms specified for each study arm?
14. Did the author(s) specify the NUMBER for each TYPE of harmful event for each study group?
15. Did the author(s) specify the type of analyses undertaken for harms data?

Source: Santaguida PL, Raina P. The development of the McHarm quality assessment scale for adverse events: Delphi consensus on important criteria for evaluating harms. <http://hiru.mcmaster.ca/epc/mcharm.pdf>. Accessed May 14, 2008.

Table 4. Criteria for evaluating the likelihood of a causal relationship in case reports

• Temporal relationship (exposure preceding adverse event and adverse event appearing at an appropriate time interval after exposure)
• Lack of alternative causes
• Drug levels in body fluids or tissues
• Resolution or improvement after discontinuation
• Dose-response relationship
• Recurrence following rechallenge (that is, restarting the drug to see whether the adverse reaction recurs)
• Confirmation of adverse event by objective information

Table 5. Sources of discrepancy between randomized controlled trials and observational studies

• Differences in risk of bias (study quality)
• Differences in applicability (study populations, interventions, or settings)
• Differences in methods used to define or measure outcomes
• Differential effects of publication or selective outcomes reporting bias
• Differential effects related to funding source (observational studies less likely to be funded by industry)

Table 6. Elements to report when describing results for harms in Comparative Effectiveness Reviews

Element	Factors
Risk of bias (quality)	Study design, number of studies, study quality, consistency of evidence, directness of evidence, other modifying factors
Applicability	Population characteristics, interventions, co-interventions, comparisons, outcomes, duration of followup for various harms
Results	Number of patients, absolute and relative estimates of risks
Publication bias or incomplete outcomes data	Graphic and/or statistical assessments for publication bias, known unpublished studies, number of studies not reporting key harms
Additional analyses	Sensitivity analyses, subgroup analyses, metaregression, etc.

Table 7. Summary of key points on assessment of harms in Comparative Effectiveness Reviews

• Assess all important harms, whenever possible.
• Use multiple sources of information, including clinical experts and stakeholders, to identify important harms.
• Use consistent and precise terminology when reporting data on harms, and avoid terms implying causality unless causality is reasonably certain.
• Gather evidence on harms from a broad range of sources, including observational studies, particularly when clinical trials are lacking; when generalizability is uncertain; or when investigating rare, long-term, or unexpected harms.
• Do not assume studies adequately assess harms because methods used to assess and report benefits are appropriate; rather, evaluate how well studies identify and analyze harms.
• Be cautious about drawing conclusions on harms when events are rare and estimates of risk are imprecise.
• Include placebo-controlled trials, particularly for assessing uncommon or rare harms, but be cautious about relying on indirect comparisons to judge comparative risks, and evaluate whether studies being considered for indirect comparisons meet assumptions for consistency of treatment effects.
• Avoid inappropriate combining of data on harms, and thoroughly investigate inconsistent results.



AHRQ Series Paper 4: Assessing harms when comparing medical interventions: AHRQ and the Effective Health-Care Program

Roger Chou^{a,*}, Naomi Aronson^b, David Atkins^{c,1}, Afisi S. Ismaila^d, Pasqualina Santaguida^d, David H. Smith^e, Evelyn Whitlock^e, Timothy J. Wilt^f, David Moher^g

^a*Oregon Evidence-based Practice Center, Oregon Health & Science University, Department of Medicine, 3181 SW Sam Jackson Park Road, Mail Code: BICC, Portland, OR, USA*

^b*Blue Cross Blue Shield Evidence-based Practice Center, Blue Cross Blue Shield Association, Chicago, IL, USA*

^c*Office of Research Development, Veteran's Affairs Administration, Washington, DC, USA*

^d*McMaster Evidence-based Practice Center, McMaster University, Hamilton, Ontario, Canada*

^e*Oregon Evidence-based Practice Center, Kaiser Center for Health Research, Portland, OR, USA*

^f*Minnesota Evidence-based Practice Center, Minneapolis VA Center for Chronic Disease Outcomes Research, MN, USA*

^g*University of Ottawa Evidence-based Practice Center, University of Ottawa, Ottawa, Ontario, Canada*

Accepted 21 June 2008

Abstract

Comparative effectiveness reviews (CERs) are systematic reviews that evaluate evidence on alternative interventions to help clinicians, policy makers, and patients make informed treatment choices. Reviews should assess harms and benefits to provide balanced assessments of alternative interventions. Identifying important harms of treatment and quantifying the magnitude of any risks require CER authors to consider a broad range of data sources, including randomized controlled trials (RCTs) and observational studies. This may require evaluation of unpublished data in addition to published reports. Appropriate synthesis of harms data must also consider issues related to evaluation of rare or uncommon events, assessments of equivalence or noninferiority, and use of indirect comparisons. This article presents guidance for evaluating harms when conducting and reporting CERs. We include suggestions for prioritizing harms to be evaluated, use of terminology related to reporting of harms, selection of sources of evidence on harms, assessment of risk of bias (quality) of harms reporting, synthesis of evidence on harms, and reporting of evidence on harms. © 2010 Elsevier Inc. All rights reserved.

Keywords: Evidence-based medicine; Research design; Review of literature; Meta-analysis; Postoperative complications; Therapeutics/ae (adverse effects)

1. Introduction

Comparative effectiveness reviews (CERs) are systematic reviews that evaluate evidence on alternative interventions to help clinicians, policy makers, and patients make informed treatment choices [1]. To generate balanced results and conclusions, it is important for CERs to address both benefits and harms [2]. However, assessing harms can be difficult. Benefits have been accorded greater prominence when reporting trials, with little effort to balance assessments of benefits and harms. In addition, systematically

reviewing evidence for all possible harms is often impractical, as interventions may be associated with dozens of potential adverse events. Furthermore, there are often important tradeoffs between increasing comprehensiveness and decreasing quality of harms data [3].

Adequately assessing harms requires CER authors to consider a broad range of data sources. For this reason, they need to deal with important challenges such as choosing which types of evidence to include, identifying studies of harms, assessing their quality, and summarizing and synthesizing data from different types of evidence.

2. Identifying harms to be evaluated

CERs should always assess harms that are important to decision makers and users of the intervention under consideration [4]. High-priority harms should include the most serious adverse events, and may also include common

Support: This article was written with support from the Effective Health Care Program at the US Agency for Healthcare Research and Quality. Dr. Moher is supported by a University of Ottawa Research Chair.

¹ Previously at the Agency for Healthcare Research and Quality, Rockville, MD, USA.

* Corresponding author. Tel.: +503-494-5367; fax: +503-494-4551.

E-mail address: chour@ohsu.edu (R. Chou).

adverse events and other specific adverse events important to clinicians and patients. CER authors should examine previously published reviews, review publicly available safety reports from the US Food and Drug Administration (FDA), and consult with technical experts and patients to set priorities for evaluating harms. Searches on postmarketing surveillance databases may also help identify important potential harms. The methods sections of the CER should specify the process used to identify harms of interest and list the specific harms for which evidence was sought.

3. Terminology

Terminology related to reporting of harms is poorly standardized [5]. This can cause confusion or result in misleading conclusions. CER authors should strive for consistent and precise usage of terminology when reporting data on harms. For example, the term “harms” is generally preferred over the term “safety” because the latter sounds more reassuring and may obscure important concerns. “Harms” is also preferable to the term “unintended effects,” which could refer to either beneficial or harmful outcomes. Terms that do not imply causality (such as “adverse events”) should be the default term to describe harms, unless causality is reasonably certain.

Definitions for commonly used terms for harms reporting are summarized in Table 1, along with suggested usage [4–6].

4. Sources of evidence on harms

4.1. Randomized controlled trials

4.1.1. Published trials

Properly designed and executed randomized controlled trials (RCTs) are considered the “gold standard” for evaluating efficacy because they minimize potential bias. However, relying solely on published RCTs to evaluate harms in CERs is problematic. First, most RCTs lack prespecified hypotheses for harms [5]. Rather, hypotheses are usually designed to evaluate beneficial effects, with assessment of harms a secondary consideration. As such, the quality and quantity of harms reporting in clinical trials is frequently inadequate [7,8].

Second, few RCTs have large enough sample sizes or are long enough in duration to adequately assess uncommon or long-term harms [9].

Third, most RCTs are explanatory, rather than pragmatic, in design. That is, they assess benefits and harms in ideal, homogeneous populations and settings [10]. Patients who are more susceptible to adverse events are often under-represented in such “efficacy” trials. Even when harms are appropriately assessed and reported, applicability of efficacy trials to general practice is limited.

Fourth, relatively few RCTs directly compare alternative treatment strategies. Although CER authors can evaluate benefits or harms of two competing interventions based on trials in which each is compared with a common third treatment (usually placebo), the results of indirect comparisons do not always agree with direct comparisons [11,12].

Fifth, publication and selective outcome(s) reporting bias can lead to distorted conclusions about harms when data are unpublished, partially reported, downplayed, or omitted [13,14].

Finally, in some cases, RCTs may not be available. For example, surgical procedures and medical devices often become widely disseminated with few or no randomized trial data. The same can be true for older therapeutic devices, such as hyperbaric oxygen chambers [15].

Despite these limitations, RCTs are the gold standard for demonstrating efficacy, the basis for most regulatory approvals, and the source of most advertising and other claims made on behalf of drugs and other interventions. For this reason, CERs must address harms data from RCTs in detail when they are available.

‘Head-to-head’ RCTs provide the most direct evidence on comparative harms. However, placebo-controlled RCTs may also provide important information on absolute and relative risks and contribute to more precise estimates of harms. In addition, placebo-controlled trials can provide information about risks that may not be apparent from head-to-head trials. For example, a systematic review of nonsteroidal anti-inflammatory drugs (NSAIDs) found cyclo-oxygenase-2-selective NSAIDs associated with greater myocardial risk vs. placebo, but differences were not apparent vs. nonselective NSAIDs that were also associated with increased risk [16]. In general, CERs should routinely include placebo-controlled trials for assessment of harms, particularly for rare or uncommon adverse events. In lieu of examining individual placebo-controlled trials, CERs may incorporate findings of well-conducted systematic reviews provided they evaluate the specific harms of interest.

4.1.2. Unpublished supplemental trials data

Besides evaluating results of published RCTs, CER authors should consider including results of completed or terminated but unpublished RCTs, and unpublished results from published trials. Such information has several potentially valuable uses:

- To assess the number of unpublished trials or frequency of unreported outcomes, which can help in evaluating risk for publication or outcomes reporting bias,
- To evaluate whether conclusions based on unpublished data are qualitatively different than those based on published RCTs, and
- To conduct formal quantitative meta-analysis, including published and unpublished RCTs or outcomes.

Table 1
Terminology for reporting on harms

Active surveillance of harms	Participants are asked about the occurrence of specific adverse events in structured questionnaires or interviews, or predefined laboratory or other diagnostic tests are performed at prespecified time intervals.
Adverse effect	A harmful or undesirable outcome that occurs during or after the use of a drug or intervention for which there is at least a reasonable possibility of a causal relation.
Adverse event	A harmful or undesirable outcome that occurs during or after the use of a drug or intervention but is not necessarily caused by it. When causality is uncertain or the purpose of the CER to establish causality, “adverse event” should generally be the default term over “adverse effect” or “adverse reaction/adverse drug reaction.”
Adverse reaction/adverse drug reaction	An adverse effect specifically associated with a drug.
Complications	A term often used to describe adverse events following surgery or other invasive interventions.
Harms	The totality of all possible adverse consequences of an intervention.
Passive surveillance of harms	Participants are not specifically asked about or tested for the occurrence of adverse events. Rather, adverse events are identified based on patient reports made on their own initiative.
Risk–benefit ratio	A common expression for the comparison of overall harms and benefits. However, because benefits and harms of an intervention are usually very different in character and are measured on different scales, a true ‘risk–benefit ratio’ is rarely calculable. In addition, there may be several distinct benefits and harms. A preferred term is <i>balance of benefits and harms</i> .
Safety	Substantive evidence of an absence of harm. Do not use this term (or the term “safe”) when evidence on harms is simply absent or is insufficient.
Serious adverse events	Any adverse event with serious medical consequences, including death, hospital admission, prolonged hospitalization, and persistent or significant disability or incapacity.
Severe adverse events	The term “severe” refers to the intensity of an adverse event (including “nonserious” adverse events). For example, a rash could be “severe” but not “serious” (i.e., not resulting in death, hospital admission, prolonged hospitalization, or persistent or significant disability).
Side effects	Unintended drug effects (beneficial or harmful) given at doses normally used for therapeutic effects. Use of this term may tend to understate the importance of harms because the word “side” may be perceived to suggest secondary importance.
Tolerability	This term is often used imprecisely but should be used to refer to a patient’s or subject’s ability or willingness to tolerate or accept unpleasant drug-related adverse events without serious or permanent sequelae.
Toxicity	The term “toxicity” is used in pharmacology and microbiology to mean “the quality of being poisonous, especially the degree of virulence of a toxic microbe or of a poison.” It is often measured in terms of the specific target affected (e.g., cytotoxicity, hepatotoxicity, and so on). In the context of systematic reviews, the term is often used to refer to laboratory-determined abnormalities, such as elevated liver-function tests. However, the terms “abnormal laboratory measurements” or “laboratory abnormalities” are more specific and appropriate.

Abbreviations: CERs, comparative effectiveness reviews.

Unpublished clinical trials tend to report lower estimates of treatment benefits than do published trials (i.e., weaker intervention effects) [17,18]. The impact of unpublished trials on assessments of harms has not been extensively studied, but a systematic review of antidepressants in children found that addition of data from unpublished trials changed conclusions about the balance of risks and benefits from favorable to unfavorable for several drugs [19].

Data from unpublished trials can be difficult to locate systematically. At a minimum, material from the FDA website should routinely be examined to assess what effect unpublished (completed or terminated) trials submitted for regulatory approval may have on conclusions regarding harms. In addition, starting in 2009, trial sponsors are required by the 2007 FDA reform bill to report results to a clinical trial results database (www.ClinicalTrials.gov) [20]. Other resources for identifying unpublished trials include non-US regulatory agencies and directly querying funding sources. Once unpublished trials are located, two

caveats should also be considered. Frequently, there is insufficient information from unpublished trials to fully assess the risk of bias. Also, the results and conclusions of trials may change between initial presentation of data and publication in a peer-reviewed journal [21].

Even when a trial is published, important information may be omitted because of space limitations or other reasons [22,23]. For example, before the publication of the Vioxx Gastrointestinal Outcomes Research Study (VIGOR) in 2001 [24], myocardial infarctions were absent from most published reports of trials evaluating selective or nonselective NSAIDs because an association with cardiovascular events was not suspected. A systematic review that obtained unpublished myocardial infarction data from older trials found an increased risk with high doses of all evaluated NSAIDs (selective or nonselective) other than naproxen [16]. An analysis of myocardial infarction risk based on only published information would have been seriously compromised by incomplete data.

Drug approval information—for example, the clinical and statistical reviews prepared by the staff of the FDA—frequently provides details about harms not included in journal publications. For example, the Celecoxib Long-term Arthritis Safety Study (CLASS), a major trial of celecoxib, was published in JAMA as a 6-month study and reported fewer gastrointestinal adverse events for celecoxib than for two nonselective NSAID comparators [25]. The JAMA article did not mention that some patients in the trial had been observed for more than 6 months [26]. In contrast, the FDA review reported all the outcome data, including no difference in gastrointestinal adverse events at the end of follow-up [27].

Limited evidence suggests an inverse relationship between the proportion of included trials reporting a specific outcome and the estimates of treatment benefit for that outcome, possibly because of selective reporting of favorable outcomes [28]. How the proportion of included trial reporting outcomes affects estimates of harms has not been well studied. Nonetheless, when a significant proportion of published trials fails to report an important or critical adverse event, CER authors should report on this gap in the evidence and consider efforts to obtain unpublished data (e.g., by querying study authors, funding sources, or clinical trial results databases, or by performing more detailed reviews of FDA documents).

4.2. Observational studies

Observational studies are almost always necessary to assess harms adequately. The exception is when there are sufficient data from RCTs to reliably estimate harms. However, even though observational studies are more susceptible to bias than well-conducted RCTs, for some comparisons there may be few or no long-term, large, head-to-head, or effectiveness RCTs [29]. Observational studies may also provide the best (or only) evidence for evaluating harms in minority or vulnerable populations (such as pregnant women, children, elderly patients, or those with multiple comorbidities) who are under-represented in clinical trials.

The term “observational studies” is commonly used to refer to cohort, case–control, and cross-sectional studies [30], but can refer to a broad range of study designs, including case reports; uncontrolled series of patients receiving surgery or other interventions; and others [31]. All can yield useful information as long as their specific limitations are understood.

The types of observational studies included in a CER will vary depending on the type or frequency of adverse events being evaluated. The choice of study designs also depends on whether investigators are seeking to determine what harms might be associated with a treatment (hypothesis generating) or whether certain harms are more likely (hypothesis testing). Different types of observational

studies might be included or rendered irrelevant by availability of data from stronger study types.

4.2.1. Cohort and case–control studies

CER authors should routinely search for and include well-designed and reported case–control and population-based cohort studies [30,32]. Such studies are well suited for testing hypotheses on whether one intervention is associated with a greater risk for an adverse event than is another and for quantifying the risk. They also take stronger precautions against bias than do other observational designs, and their strengths and weaknesses are well understood. For unexpected adverse events, for example, confounding by indication in case–control and cohort studies may not be as important an issue as when evaluating beneficial effects because their occurrence is usually not associated with the reasons for choosing a particular treatment [29,33]. Although cross-sectional studies have features in common with cohort studies, it is difficult to establish causality because exposures and outcomes are evaluated simultaneously. Indeed, associations in cross-sectional studies may sometimes be resulting from reverse causality [34].

A recent report found that large observational studies usually report smaller absolute risks of harm than do large randomized trials [35]. There was no clear tendency for randomized trials or observational studies to report larger relative risks. In more than one-half of the comparisons assessed, estimates of relative or absolute risk varied more than twofold. Discrepancies between randomized trials and observational studies may occur because of differences in populations, settings, or interventions; differences in study design, including criteria used to identify harms; differential effects of biases, or some combination of these factors.

4.2.2. Observational studies based on patient registries

Patient registries collect information on clinical outcomes in populations defined by a particular disease, condition, or exposure [36]. Clinical data are prospectively collected for specific research purposes using active methods to identify outcomes, though registries can be supplemented by information from administrative databases and other sources. Registries can be designed as an active surveillance system for identifying harms, and may be particularly useful for assessing long-term or uncommon adverse events.

4.2.3. Observational studies based on analyses of large databases

Pharmacoepidemiologic studies using large databases to identify exposures and outcomes may be valuable for comparing the risk of uncommon adverse events [37]. However, additional empirical research is needed to identify methods

for collecting and analyzing data in pharmacoepidemiologic studies that are associated with valid findings [38]. Unlike studies based on patient registries, large administrative databases usually contain information routinely collected during clinic, hospital, laboratory, or pharmacy encounters, rather than for a specific research purpose. Such studies are probably most useful for evaluating serious harms that are more reliably reported and recorded (e.g., death or acute myocardial infarction) than less serious harms that may not generate a specific clinic visit or diagnostic code (e.g., sedation or nausea). In some cases, administrative data may be supplemented or verified by more detailed clinical information. Regardless of how data are obtained, all observational studies should use appropriate methods for minimizing bias and misclassification of data.

4.2.4. Case reports and postmarketing surveillance

About 30% of the primary published literature on adverse drug events is in the form of case reports [39]. Case reports can be useful for identifying uncommon, unexpected, or long-term adverse events, particularly for new drugs or for other interventions [40]. The adverse events identified by case reports often differ from those detected in clinical trials [41]. However, case reports are usually considered hypothesis generating because calculating information from them about the frequency or comparative risk of adverse events is difficult.

In the US, the FDA receives about 280,000 reports of postmarketing adverse events annually, collects them into a database [42], and issues information about adverse drug events on its MedWatch website (<http://www.fda.gov/medwatch/>). Although pharmaceutical companies and other investigators may also perform passive surveillance of harms on postmarketing data, such analyses are not always made public in a timely fashion [43]. Active, hypothesis-driven postmarketing surveillance systems have been developed recently for identifying and evaluating serious adverse drug events [44].

Case reports and other hypothesis-generating studies may be useful for CERs evaluating new drugs suspected of being associated with serious but uncommon adverse events. For other topics, CER authors may consider their inclusion on a case-by-case basis.

4.2.5. Other observational studies

Several other types of observational studies may also report data on harms. However, they are likely to be more prone to bias than RCTs or well-designed case–control or cohort studies, and their use needs to be considered cautiously. For example, studies reporting harms from surgical or other invasive interventions often consist of a series of patients who received the procedure. Data are often insufficient to assess the methods used to select participants [45]. In addition, because such studies lack

control groups, evaluating effects of confounding is difficult, as is comparing risks of adverse events across interventions.

Other quasi-experimental study designs may not offer any advantage over RCTs in terms of their applicability to routine practice. For example, open-label extensions of clinical trials may follow up patients for an extended period of time, but they usually enroll a more highly selected population (patients who completed the randomized trial, tolerated the medication, and agreed to participate in the extension), are unblinded, and often lack a comparison arm. Such studies can be excluded from CERs if more reliable long-term, comparative data are available. If they are included in CERs, their limitations should be described clearly.

4.2.6. Criteria to select observational studies for inclusion

In general, many more observational studies than randomized trials will be available for nearly all health care interventions. Evaluating a large number of observational studies can be impractical when conducting a CER, especially when a significant proportion either does not add useful information or carry a high risk of reporting biased results.

Several criteria have commonly been used in systematic reviews and CERs to screen observational studies of harms for inclusion. Empirical data are lacking on how use of different selection criteria affects estimates of harms. However, CERs should match inclusion criteria to the reasons for including observational studies. For example, inclusion criteria might specify minimum duration of follow-up if a priority is to identify evidence on long-term harms. If large, higher-quality studies are available, it could be reasonable to specify a minimum sample size threshold to use resources efficiently. Methods sections should clearly describe selection criteria along with the rationale for choosing the criteria. Commonly used inclusion criteria for observational studies are shown in Table 2.

Table 2
Example criteria for selecting observational studies on harms for inclusion in a CER

-
- Studies meeting certain study design definitions (e.g., cohort and case–control studies)
 - Studies not exceeding a defined threshold for risk of bias (e.g., studies assessed as being at low risk of bias or meeting certain prespecified quality criteria)
 - Studies meeting a defined threshold for duration of follow-up
 - Studies meeting a sample size threshold
 - Studies evaluating a specific population of interest (e.g., studies evaluating populations under-represented in randomized trials such as the elderly, women, or minority populations)
-

Abbreviations: CERs, comparative effectiveness reviews.

5. Assessing risk of bias (quality) of harms reporting

5.1. Randomized trials

A number of features of RCTs have been empirically tested and proposed as markers of higher quality (i.e., lower risk of bias). These include use of appropriate randomization generation and allocation concealment techniques; blinding of participants, healthcare providers, and outcome assessors; and analysis according to intention-to-treat principles [46]. Whether these are equally important in protecting against bias in studies reporting harms is unclear. Moreover, because evaluating harms is often a secondary consideration in randomized trials, quality of harms assessment and reporting can be inadequate even when assessment of the primary (beneficial) outcome is appropriate.

When evaluating the quality of harms assessment, CER authors should consider whether adequate methods were used to identify adverse events in the primary studies. Active methods (such as querying patients using a comprehensive checklist or standardized laboratory tests) are more likely to completely identify adverse events than passive methods such as relying on patient self-report [47]. In addition, specific data on adverse events are likely to be more accurate and informative than generic statements, such as “no adverse events were noted” or “the interventions were well tolerated.” If a specific adverse event is not reported, it is generally safer for CER authors to assume that they were not ascertained or not recorded than to assume that the prevalence or incidence was zero [4].

It is also important to assess how adverse events are assessed and categorized. Studies should predefine the qualifiers “serious” and “severe” to describe adverse events. Otherwise, it is impossible for readers to determine whether these labels were applied consistently within and across trials. Standardized criteria for grading severity of adverse events are available for certain conditions [48,49]. CERs should note when grading severity or seriousness of adverse events is based on nonstandardized or poorly defined criteria, as such classifications may not be comparable across studies or may be poorly reproducible. Similarly, methods for classifying adverse events as “treatment-related” are largely subjective, with unknown validity, and such data may be particularly unreliable.

It is not always necessary for trials to prespecify or define adverse events. For example, studies reporting unexpected outcomes can be very valuable for identifying previously unrecognized harms. However, when evaluating known harms, using validated or standardized criteria for adverse events may help reduce subjectivity or bias in their assessment and classification. In drug trials, use of an independent external endpoint committee may provide less biased estimates of harms than outcome assessment performed by investigators connected to the study [50].

“Withdrawals due to adverse events” is commonly reported in trials and often used in systematic reviews as

a marker for intolerable or severe adverse events. However, the Cochrane Adverse Effects Methods Group suggests caution in interpreting withdrawals attributed to adverse events in this manner, for the following reasons [4]:

- Attribution of reasons for discontinuation is likely to be imprecise and to vary across trials.
- Pressure to keep drop-outs low in trials may result in rates that do not reflect real-world practice.
- Unblinding often takes place before the decision to withdraw, which can lead to distortion of estimates of an intervention’s effect on withdrawal (e.g., symptoms are less likely to lead to withdrawal if the patient is found to be on placebo).

Nonetheless, withdrawals because of adverse events are often reported even when serious or severe adverse events are not reported or are poorly defined, and they may provide some useful information.

5.2. Observational studies

Because observational studies lack randomization, they should adhere to high methodological standards to be considered valid [30,32,51]. RCTs are expected to have outcomes recorded by blinded personnel and to include all participants who were randomized in the analysis of results. The use of blinded outcome assessors and a clearly identified inception cohort (e.g., “new users”) [52] is at least as important when assessing observational studies.

Instruments for assessing risk of bias in observational studies vary greatly in scope, number and types of items used, and developmental rigor [53]. Further study is needed to determine which methodological shortcomings in observational studies are consistently associated with bias in assessment and reporting of harms. However, some consensus exists on the major domains that should be considered when evaluating the overall validity of an observational study. For cohort studies, important factors include assembly of an inception cohort, complete follow-up, appropriate assessment of potential confounders, accurate determination of exposures and outcomes, and blinded assessment of outcomes [30,52–54].

Several studies have empirically evaluated effects of specific methodological characteristics on estimates of harms from observational studies. They found that prospective or retrospective design [55,56], case–control compared with cohort studies [57,58], and smaller compared with larger case series [55] did not have consistent effects on estimates of harms. Two studies found that industry-funded studies tended to report more favorable outcomes than did studies with other funding sources [57,59]. Because all of these studies evaluated fairly limited samples of studies, their wider applicability is uncertain.

Observational studies based on evaluations of large administrative databases should follow the same general

principles to reduce bias as do observational studies that directly collect data from patients. In these cases, reviewers should pay particular attention to the methods used for ascertaining exposures and outcomes and for measuring and analyzing potential confounders, as these issues are more likely to be problematic in studies relying on administrative claims (although not unique to them) [37].

For all observational study designs, estimates of harms are less likely to be confounded when evaluating previously unsuspected adverse events than when evaluating a known harm or intended effects. For example, the finding that cyclo-oxygenase-2-selective NSAIDs were associated with an increased risk of myocardial infarction vs. nonselective NSAIDs was an unexpected finding from an RCT examining a different outcome [24]. This risk could be partly confirmed in observational studies because the choice of the type of NSAID in typical practice was unrelated to the patients' risk for myocardial infarction. In contrast, gastrointestinal bleeding was a known risk of nonselective NSAIDs and clinicians were more likely to prescribe selective NSAIDs in patients at higher risk for gastrointestinal bleeding. Such "confounding by indication" led to the appearance of an apparent association between selective NSAID use and bleeding in epidemiologic studies [60]. In some cases, such spurious associations may remain despite adjustment for known confounders ("residual confounding").

5.3. Uncontrolled studies

Studies of surgery, medical devices, and other nonpharmacologic interventions are often on uncontrolled series of patients who received the therapy and were then followed up over a period of time. Such studies can provide some information about rates of adverse events in clinical practice, and they may be most informative when the incidence of such events in untreated patients is low. Unfortunately, such studies frequently do not meet standards for accurate and comprehensive reporting of harms [61]. Even when harms data are well described, an important limitation of uncontrolled studies is that it is difficult to evaluate confounding by indication. Authors are also more likely to submit for publication studies showing the best outcomes.

For some interventions, CER authors must consider including uncontrolled studies for assessing harms, as little or no other evidence may be available. Proposed criteria for evaluating case series are likely to promote improved reporting of results [62], but may provide only limited information about risk of bias. Important factors to consider when evaluating uncontrolled studies include whether the study enrolled or attempted to enroll all patients meeting prespecified inclusion criteria and whether the study clearly describes loss to follow-up [45]. When uncontrolled studies do not meet these criteria, determining the reliability and applicability of even well-described results may be impossible.

5.4. Instruments for assessing risk of bias (quality) in studies on harms

Development of instruments for assessing risk of bias specifically in studies of harms is still in an early stage of development. Two issues remain unclear: whether to use a specific rating instrument to evaluate harms assessment and reporting, or whether using instruments for rating the overall risk of bias of a study is sufficient, as long as particular attention is paid to how well adverse events are defined, ascertained, and reported.

Chou et al. empirically developed and tested an instrument for assessing quality of harms assessment and reporting in randomized trials and observational studies of carotid endarterectomy for symptomatic carotid artery stenosis [63]. This approach involved four criteria: nonbiased selection of subjects, low loss to follow-up, adverse events prespecified and defined, and adequate duration of follow-up. Studies meeting at least three of the four criteria reported a rate of postsurgical complications of 5.7% (95% CI, 4.8–6.6%) compared with 3.7% (95% CI, 3.1–4.3%) for studies meeting fewer than three of the criteria. However, the generalizability of this instrument to other datasets or interventions is unclear. When the authors applied these criteria to studies of rofecoxib, they were unable to show differences in estimates of risk of myocardial infarction. In addition, caution should be used when considering use of summary scores to assess risk of bias [64]. At a minimum, key methodological aspects should be assessed individually and their influence on estimates of harms explored.

Santaguida and Raina have also developed a quality-rating instrument (McHarm) for evaluating studies reporting harms (Table 3) [65]. The tool was developed from quality-rating items generated by a review of the literature on harms and from previous quality-assessment instruments. A formal Delphi consensus exercise was used to reduce the number of items. The subsequent list of quality criteria specific to harms was tested for reliability and face, construct, and criterion validity. This quality-assessment tool is intended for use in conjunction with standardized quality-assessment tools for design-specific internal validity issues.

Case reports may provide valuable information about the possibility of rare or previously unrecognized adverse events. A 1982 study examined 47 case reports published in 1963 in four major general medical journals and judged that 35 of them were subsequently proved to be "clearly" correct [66]. However, the methods used to determine reliability of case reports in this study were subjective, and results have not been replicated. A recent study, in fact, found that only 18% of case reports of suspected adverse drug reactions have been subjected to rigorous evaluation in subsequent studies [67]. Nonetheless, a statistical modeling study suggests that the likelihood of more than one to three spontaneously reported cases is very unlikely to be coincidental when the adverse event is rare or uncommon [68]. Case

Table 3
McMaster tool for assessing quality of harms assessment and reporting in study reports (McHarm)

1. Were the harms PREDEFINED using standardized or precise definitions?
2. Were SERIOUS events precisely defined?
3. Were SEVERE events precisely defined?
4. Were the number of DEATHS in each study group specified OR were the reason(s) for not specifying them given?
5. Was the mode of harms collection specified as ACTIVE?
6. Was the mode of harms collection specified as PASSIVE?
7. Did the study specify WHO collected the harms?
8. Did the study specify the TRAINING or BACKGROUND of who ascertained the harms?
9. Did the study specify the TIMING and FREQUENCY of collection of the harms?
10. Did the author(s) use STANDARD scale(s) or checklist(s) for harms collection?
11. Did the authors specify if the harms reported encompass ALL the events collected or a selected SAMPLE?
12. Was the NUMBER of participants that withdrew or were lost to follow-up specified for each study group?
13. Was the TOTAL NUMBER of participants affected by harms specified for each study arm?
14. Did the author(s) specify the NUMBER for each TYPE of harmful event for each study group?
15. Did the author(s) specify the type of analyses undertaken for harms data?

Ref: Santaguida and Raina [65].

reports, however, cannot be used to estimate the rate of an adverse event, which may be critical to any decisions.

Several disease-specific [69] and nondisease-specific [70] methods for assessing the probability of causality from case reports of adverse events have been developed. These methods represent expert opinion and have not been validated empirically. Factors believed to increase the likelihood of causality are shown in Table 4 [69,70].

Guidelines for improving the reporting of suspected adverse drug events in case reports have also recently been proposed [71]. In 35 reports of 48 patients published in the BMJ, the median number of recommended items that were reported was nine (range 5–12) of 19, although effects of missing information on the validity of case reports have not been studied.

Table 4
Criteria for evaluating the likelihood of a causal relationship in case reports

- Temporal relationship (exposure preceding adverse event and adverse event appearing at an appropriate time interval after exposure)
- Lack of alternative causes
- Drug levels in body fluids or tissues
- Resolution or improvement after discontinuation
- Dose–response relationship
- Recurrence following rechallenge (i.e., restarting the drug to see whether the adverse reaction recurs)
- Confirmation of adverse event by objective information

6. Synthesizing evidence on harms

CER authors should follow general principles for synthesizing evidence when evaluating data on harms. Such principles include the following: combining studies only when they are similar enough to warrant combining [72]; adequately considering risk of bias, including publication and other related biases [73]; and exploring potential sources of heterogeneity [23]. Several other issues are especially relevant for synthesizing evidence on harms.

6.1. Uncommon or rare adverse events

Evaluating comparative risks of uncommon or rare adverse events in CERs can be particularly challenging. A frequent problem in RCTs and systematic reviews is interpreting a nonsignificant probability value as indicating no difference in risk for rare adverse events, particularly when the confidence intervals are wide and encompass the possibility of clinically important risks [74,75]. For example, one trial concluded that, in patients with meningitis, “treatment with dexamethasone did not result in an increased risk of adverse events” compared with placebo for treatment of hyperglycemia, herpes zoster, or fungal infection because *P* values for all three outcomes were >0.20 [76]. However, the 95% confidence intervals for estimates of relative risks for these three adverse events encompassed clinically significant increases in risk (–13.5% to 77.6%, –60.4% to 377.7%, and –43.6% to 496.2%, respectively). In such cases, CERs should acknowledge the lack of statistical power to assess risk adequately and interpret the confidence intervals, including the possibility or probability of excess harm.

6.2. Equivalence and noninferiority

CER authors should draw conclusions about “equivalence” or “noninferiority” of interventions with regard to harms only when there are appropriate data to justify such statements [77]. Few CERs will have the statistical power to adequately assess noninferiority when the risk of an adverse event is on the order of 1% or lower. For example, about 100,000 patients would have been needed in the CO-BALT or GUSTOIII trials to rule out an excess relative death rate of 5% from alternative thrombolytic agents with 80% power [78]. Ruling out smaller event rates would require even higher sample sizes.

6.3. Indirect analyses

Placebo-controlled trials can be helpful for evaluating absolute risks associated with an intervention. When head-to-head trials are sparse or unavailable, placebo-controlled trials may also be useful for indirectly evaluating comparative harms, particularly for rare or uncommon adverse events. However, for indirect analyses to be

reliable, all studies should be comparable in terms of quality, factors related to applicability (population, dosing, cointerventions, and settings), measurement of outcomes, and incidence of adverse events in control groups [12,79].

For example, a meta-analysis found rofecoxib associated with an increased risk of arrhythmia compared with control treatments; celecoxib was not [80]. However, the rate of arrhythmia in the control arms was 10-fold higher in trials of celecoxib (0.27% or 18 of 6,568 subjects) compared to trials of rofecoxib (0.02% or two of 10,174 subjects). In this situation, indirect comparisons about relative safety of celecoxib compared with rofecoxib are likely to be problematic. A more informative approach would be to explore reasons for the discrepancies in rates of arrhythmias in the control arms and how they may have affected comparisons.

More studies are needed to determine when indirect comparisons are most likely to be valid. In the meantime, CER authors considering indirect analyses to assess harms should carefully consider whether assumptions underlying valid indirect comparisons are likely to be met, compare results of indirect comparisons with head-to-head data if available, and draw conclusions from indirect comparisons cautiously.

6.4. Combining data from different types of studies

Most CERs will include data on harms from different types of studies. Statistical combination of data from observational studies is often inappropriate and should be avoided unless there is clear rationale to do so [81]. If such analyses are undertaken, the justification should be clearly explained.

6.5. Discrepancies between randomized trials and observational studies

A separate challenging situation is when results on harms from randomized trials and observational studies are discordant. Some reasons for discrepancies between randomized trials and observational studies are shown in Table 5. A reasoned analysis of potential sources of

Table 5
Sources of discrepancy between observational studies and RCTs

<ul style="list-style-type: none"> • Differences in risk of bias (study quality) • Differences in applicability (study populations, interventions, or settings) • Differences in methods used to define or measure outcomes • Differential effects of publication or selective outcomes reporting bias • Differential effects related to funding source (observational studies less likely to be funded by industry)

Abbreviations: RCTs, randomized control trials.

Table 6
Elements to report when describing results for harms in CERs

Element	Factors
Risk of bias (quality)	Study design, number of studies, study quality, consistency of evidence, directness of evidence, other modifying factors
Applicability	Population characteristics, interventions, cointerventions, comparisons, outcomes, duration of follow-up for various harms
Results	Number of patients, absolute and relative estimates of risks
Publication bias or incomplete outcome data	Graphic and/or statistical assessments for publication bias, known unpublished studies, number of studies not reporting key harms
Additional analyses	Sensitivity analyses, subgroup analyses, metaregression, and so on

Abbreviations: CERs, comparative effectiveness reviews.

discrepancy is generally more helpful than simply presenting the different results.

7. Reporting evidence on harms

As when reporting evidence on benefits, CERs should emphasize the most reliable information for the most important adverse events. Summary tables should generally present data for the most important harms first, with more reliable evidence preceding less reliable evidence. Evidence on harms from each type of study should be clearly summarized in summary tables, in narrative format, or in both [2]. A critical role of CERs is to report clearly on the limitations of the evidence on harms and to analyze and interpret thoughtfully how these limitations may affect estimates of the balance of benefit and harm. Suggested elements to focus on when reporting harms are shown in Table 6.

Acknowledgments

The authors would like to acknowledge Gail R. Janes for participating in the workgroup calls.

Disclaimer: The views expressed in this article are those of the authors and do not represent the official policies of the Agency for Healthcare Research and Quality, the Department of Health and Human Services, the Department of Veterans Affairs, the Veterans Health Administration, or the Health Services Research and Development Service.

Appendix

Summary of key points on assessment of harms in CERs

- Assess all important harms, whenever possible.
- Use multiple sources of information, including clinical experts and stakeholders, to identify important harms.
- Use consistent and precise terminology when reporting data on harms, and avoid terms implying causality unless causality is reasonably certain.
- Gather evidence on harms from a broad range of sources, including observational studies, particularly when clinical trials are lacking; when generalizability is uncertain; or when investigating rare, long-term, or unexpected harms.
- Do not assume studies adequately assess harms because methods used to assess and report benefits are appropriate; rather, evaluate how well studies identify and analyze harms.
- Be cautious about drawing conclusions on harms when events are rare and estimates of risk are imprecise.
- Include placebo-controlled trials, particularly for assessing uncommon or rare harms, but be cautious about relying on indirect comparisons to judge comparative risks, and evaluate whether studies being considered for indirect comparisons meet assumptions for consistency of treatment effects.
- Avoid inappropriate combining of data on harms, and thoroughly investigate inconsistent results.

Abbreviations: CERs, comparative effectiveness reviews.

References

- [1] Lohr KN. Emerging methods in comparative effectiveness and safety: symposium overview and summary. *Med Care* 2007;45:S5–8.
- [2] GRADE Working Group. Grading quality of evidence and strength of recommendations. *Br Med J* 2004;328:1490.
- [3] McIntosh HM, Woolacott NF, Bagnall A-M. Assessing harmful effects in systematic reviews. *BMC Med Res Methodol* 2004;4:19.
- [4] Loke YK, Price D, Herxheimer A. Systematic reviews of adverse effects: framework for a structured approach. *BMC Med Res Methodol* 2007;7:32.
- [5] Ioannidis JPA, Evans SJW, Gotzsche PC, O'Neill RT, Altman DG, Schulz K, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med* 2004;141:781–8.
- [6] Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. *Lancet* 2000;356:1255–9.
- [7] Ioannidis JPA, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *J Am Med Assoc* 2001;285:437–43.
- [8] Loke Y, Derry S. Reporting of adverse drug reactions in randomised controlled trials—a systematic survey. *BMC Clin Pharmacol* 2001;1:3.
- [9] Vandembroucke JP. Benefits and harms of drug treatments. *Br Med J* 2004;329:2–3.
- [10] Rothwell PM. External validity of randomised controlled trials: “to whom do the results of this trial apply?”. *Lancet* 2005;365:82–93.
- [11] Chou R, Fu R, Huffman LH, Korthuis PT. Initial highly-active anti-retroviral therapy with a protease inhibitor versus a non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses. *Lancet* 2006;368:1503–15.
- [12] Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *Br Med J* 2003;326:472.
- [13] Chan A, Hrobjartsson A, Haahr M, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials. *J Am Med Assoc* 2004;291:2457–65.
- [14] Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;337:867–72.
- [15] McDonagh M, Helfand M, Carson S, Russman BS. Hyperbaric oxygen therapy for traumatic brain injury: a systematic review of the evidence. *Arch Phys Med Rehabil* 2004;85:1198–204.
- [16] Kearney PM, Baigent C, Godwin J, Halls H, Emberson JR, Patrono C. Do selective cyclo-oxygenase-2 inhibitors and traditional non-steroidal anti-inflammatory drugs increase the risk of atherothrombosis? Meta-analysis of randomized trials. *Br Med J* 2006;332:1302–8.
- [17] Egger M, Juni P, Bartlett C, Hohenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technol Assess* 2003;7:1–76.
- [18] Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med* 2008;358:252–60.
- [19] Whittington CJ, Kendall T, Fonagy P, Cottrell D, Cotgrove A, Boddington E. Selective serotonin reuptake inhibitors in childhood depression: systematic review of published versus unpublished data. *Lancet* 2004;363:1341–5.
- [20] Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible research: moving toward research the public can really trust. *Ann Intern Med* 2007;146:450–3.
- [21] Toma M, McAlister FA, Bialy L, Adams D, Vandermeer B, Armstrong PW. Transition from meeting abstract to full-length journal article for randomized controlled trials. *J Am Med Assoc* 2006;295:1281–7.
- [22] Ridker PM, Torres J. Reported outcomes in major cardiovascular clinical trials funded by For-profit and Not-For-Profit organizations: 2000–2005. *J Am Med Assoc* 2006;295:2270–4.
- [23] Sterne JA, Egger M, Smith GD. Investigating and dealing with publication and other biases in meta-analysis. *Br Med J* 2001;323:101–5.
- [24] Bombardier C, Laine L, Reicin A, Shapiro D, Burgos-Vargas R, Davis B, et al. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. [see comment]. *N Engl J Med* 2000;343:1520–8.
- [25] Silverstein FE, Faich G, Goldstein JL, Simon LS, Pincus T, Whelton A, et al. Gastrointestinal toxicity with celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis: the CLASS study: a randomized controlled trial. Celecoxib Long-term Arthritis Safety Study. [see comment]. *J Am Med Assoc* 2000;284:1247–55.
- [26] Hrachovec JB, Mora M. Reporting of 6-month vs 12-month data in a clinical trial of celecoxib. *J Am Med Assoc* 2001;286:2398.
- [27] Witter J. Medical review part 1. Center for Drug Evaluation and Research. Available at: http://www.fda.gov/cder/foi/nda/2002/20-998S009_Celebrex_medr_P1.pdf. Accessed on April 3, 2008.
- [28] Furukawa TA, Watanabe N, Montori VM, Guyatt GH. Association between unreported outcomes and effect size estimates in Cochrane meta-analyses. [letter]. *J Am Med Assoc* 2007;297:468–70.
- [29] Vandembroucke JP. When are observational studies as credible as randomised trials? *Lancet* 2004;363:1728–31.
- [30] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandembroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med* 2007;147:573–7.
- [31] Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research. Principles and quantitative methods. Belmont, CA: Wadsworth; 1982.
- [32] Vandembroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *Ann Intern Med* 2007;147:W163–94.
- [33] Psaty BM, Koepsell T, Lin D, Weiss NS, Siscovick DS, Rosendaal FR, et al. Assessment and control for confounding by indication in observational studies. *J Am Geriatr Soc* 1999;47:749–54.
- [34] Rothman KJ, Greenland S. Modern epidemiology. 2nd edition. Philadelphia, PA: Lippincott-Raven; 1998.

- [35] Papanikolaou PN, Christidi GD, Ioannidis J. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *Can Med Assoc J* 2006;174:635–41.
- [36] AHRQ Publication NO. 07-EHC001-1. In: Gliklich R, Dreyer NA, editors. *Registries for evaluating patient outcomes: a user's guide*. Rockville, MD: Agency for Healthcare Research and Quality; 2007.
- [37] 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.
- [38] Sturmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration—a simulation study. *Am J Epidemiol* 2007;165:1110–8.
- [39] Aronson JK, Derry S, Loke YK. Adverse drug reactions: keeping up to date. *Fund Clin Pharmacol* 2002;16:49–56.
- [40] Stricker BH, Psaty BM. Detection, verification, and quantification of adverse drug reactions. *Br Med J* 2004;329:44–7.
- [41] Loke YK, Derry S, Aronson JK. A comparison of three different sources of data in assessing the frequencies of adverse reactions to amiodarone. *Br J Clin Pharmacol* 2004;57:616–21.
- [42] Strom BL. Potential for conflict of interest in the evaluation of suspected adverse drug reactions: a counterpoint. *J Am Med Assoc* 2004;292:2643–6.
- [43] 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. *J Am Med Assoc* 2004;292:2622–31.
- [44] Bennett CL, Nebeker JR, Lyons EA, Samore MH, Feldman MD, McKoy JM, et al. The research on adverse drug events and reports (RADAR) project. *J Am Med Assoc* 2005;293:2131–40.
- [45] Oleson O. The Cochrane Non-Randomised Studies Methods Group (NRSMG). Types of study design. Available at: <http://www.cochrane.dk/nrsmg/docs/chap2.pdf>. 1999; Accessed on April 3, 2008.
- [46] Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *Br Med J* 2001;323:42–6.
- [47] Bent S, Padula A, Avins AL. Brief communication: better ways to question patients about adverse medical events: a randomized, controlled trial. *Ann Intern Med* 2006;144:257–61.
- [48] NCI. Common terminology criteria for adverse events v3.0 (CTCAE). http://ctep.cancer.gov/reporting/ctc_v30.html. 2006; Available at Accessed on April 3, 2008.
- 49 NIAID. Division of AIDS table for grading the severity of adult and pediatric adverse events. Available at: <http://www3.niaid.nih.gov/research/resources/DAIDS/ClinRsrch/Safety/>. 2004; Accessed on April 3, 2008.
- [50] Sydes MR, Spiegelhalter DJ, Altman DG, Babiker AB, Parmar MKB. Systematic qualitative review of the literature on data monitoring committees for randomized controlled trials. *Clin Trials* 2004;1:60–79.
- [51] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974;66:688–701.
- [52] Rochon PA, Gurwitz JH, Sykora K, Mamdani M, Streiner DL, Garfinkel S, et al. Reader's guide to critical appraisal of cohort studies: 1. Role and design. *Br Med J* 2005;330:895–7.
- [53] Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovich C, Song F, et al. Evaluating non-randomised intervention studies. *Health Technol Assess* 2003;7:1–173.
- [54] West S, King V, Carey TS. *Systems to rate the strength of scientific evidence*. Rockville, MD: Agency for Healthcare Research and Quality; 2002.
- [55] Dalziel K, Round A, Stein K, Garside R, Castelnovo E, Payne L. Do the findings of case series studies vary significantly according to methodological characteristics? *Health Technol Assess* 2005;9:1–146.
- [56] Rothwell PM, Slattery J, Warlow CP. A systematic review of the risks of stroke and death due to endarterectomy for symptomatic carotid stenosis. *Stroke* 1996;27:260–5.
- [57] Juni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *Lancet* 2004;364:2021–9.
- [58] Ofman JJ, MacLean CH, Straus WL, Morton SC, Berger ML, Roth EA, et al. A metaanalysis of severe upper gastrointestinal complications of nonsteroidal antiinflammatory drugs. [see comment]. *J Rheumatol* 2002;29:804–12.
- [59] Shah RV, Albert TJ, Buegel-Sanchez V, Vaccaro AR, Hilibrand AS, Gauer JN. Industry support and correlation to study outcome for papers published in Spine. *Spine* 2005;30:1099–104.
- [60] Laporte JR, Ibanez L, Vidal X, Vendrell L, Leone R. Upper gastrointestinal bleeding associated with the use of NSAIDs: new versus older agents. *Drug Saf* 2004;27:411–20.
- [61] Martin RCG, Brennan MF, Jacques DP. Quality of complication reporting in the surgical literature. *Ann Surg* 2002;235:803–13.
- [62] Carey TS, Boden SD. A critical guide to case series reports. *Spine* 2003;28:1631–4.
- [63] Chou R, Fu R, Carson S, Saha S, Helfand M. Methodological shortcomings predicted lower harm estimates in one of two sets of studies of clinical interventions. *J Clin Epidemiol* 2006;60:18–28.
- [64] Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *J Am Med Assoc* 1999;282:1054–60.
- [65] Santaguida PL, Raina P. The development of the McHarm quality assessment scale for adverse events: Delphi consensus on important criteria for evaluating harms. Available at: <http://hiru.mcmaster.ca/epc/mcharm.pdf>. 2008; Accessed on May 14, 2008.
- [66] Venning GR. Validity of anecdotal reports of suspected adverse drug reactions: the problem of false alarms. *Br Med J* 1982;284:249–52.
- [67] Loke YK, Price D, Derry S, Aronson JK. Case reports of suspected adverse drug reactions—systematic literature survey of follow-up. *Br Med J* 2006;332:335–59.
- [68] Begaud B, Moride Y, Tubert-Bitter P, Chaslerie A, Haramburu F. False-positives in spontaneous reporting: should we worry about them? *Br J Clin Pharmacol* 1994;38:401–4.
- [69] Danan G, Benichou C. Causality assessment of adverse reactions to drugs—I. A novel method based on the conclusions of international consensus meetings: application to drug-induced liver injuries. *J Clin Epidemiol* 1993;46:1323–30.
- [70] Michel DJ, Knodel LC. Comparison of three algorithms used to evaluate adverse drug reactions. *Am J Hosp Pharm* 1986;43:1709–44.
- [71] Aronson JK. Anecdotes as evidence. *Br Med J* 2003;326:1346.
- [72] Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. *Ann Intern Med* 1997;127:820–6.
- [73] Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;352:609–13.
- [74] Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuses of power when interpreting results. *Ann Intern Med* 1994;121:200–6.
- [75] Jonville-Bera AP, Giraudeau B, Autret-Leca E. Reporting of drug tolerance in randomized clinical trials: when data conflict with authors' conclusions. *Ann Intern Med* 2006;144:306–7.
- [76] de Gans J, van de Beek D. Dexamethasone in adults with bacterial meningitis. *N Engl J Med* 2002;347:1549–56.
- [77] Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ, CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. *J Am Med Assoc* 2006;295:1152–60.
- [78] Ware JH, Antman EM. Equivalence trials. *N Engl J Med* 1997;337:1159–61.
- [79] Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol* 1997;50:683–91.
- [80] Zhang J, Ding EL, Song Y. Adverse effects of cyclooxygenase 2 inhibitors on renal and arrhythmia events: meta-analysis of randomized trials. *J Am Med Assoc* 2006;296:1619–32.
- [81] Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. *Br Med J* 1998;316:140–4.