Systematic reviews in emergency medicine: Part II. Critical appraisal of review quality, data synthesis and result interpretation

Peter J. Zed, BSc, BSc (Pharm), PharmD;[*†‡] Brian H. Rowe, MD, MSc;[§]
Peter S. Loewen, BSc (Pharm), PharmD;[*†] Riyad B. Abu-Laban, MD, MHSc[‡¶]

ABSTRACT

Reviews of the medical literature have always been an important resource for physicians. Increasingly, qualitative and quantitative systematic reviews (SRs) have replaced the traditional "narrative review" as a means of capturing and summarizing current evidence on a topic or, when possible, answering a specific clinical question. This paper is Part II of a 2-part series designed to provide emergency physicians with the background necessary to locate, critically evaluate and interpret SRs. The paper expands on the critical appraisal principles discussed in Part I by focusing on quality assessment, data synthesis and interpretation of results. To illustrate key points and facilitate readability, examples from the emergency medicine literature have been included and technical details have been kept to a minimum. The references, however, are comprehensive and provide a resource for readers seeking further information.

Key words: systematic reviews; emergency medicine; evidence-based medicine

RÉSUMÉ

Les revues de la littérature médicale ont toujours été une ressource importante pour les médecins. De plus en plus, les «revues systématiques» qualitatives et quantitatives ont remplacé les «revues narratives» comme moyen de saisir et résumer les données courantes sur un sujet précis ou, autant que possible, de répondre à une question clinique spécifique. Le présent article constitue la deuxième partie d'une série en deux parties conçue pour donner aux médecins d'urgence le contexte nécessaire pour trouver, faire une évaluation critique et interpréter les revues systématiques. Cet article approfondit davantage les principes de revue critique discutés dans la partie I en élaborant sur l'évaluation de la qualité, la synthèse des données et l'interprétation des résultats. Pour illustrer des points-clés et rendre la lecture plus facile, des exemples provenant de la littérature de médecine d'urgence sont inclus et les détails techniques sont maintenus à un strict minimum. Par contre, les références sont exhaustives et offrent des ressources pour les lecteurs à la recherche de plus d'information.

*Clinical Service Unit Pharmaceutical Sciences, Vancouver General Hospital, Vancouver, BC
†Faculty of Pharmaceutical Sciences, University of British Columbia, Vancouver, BC
‡Division of Emergency Medicine, Department of Surgery, Faculty of Medicine, University of British Columbia, Vancouver, BC
§Division of Emergency Medicine, Department of Public Health Sciences, University of Alberta and Capital Health Authority, Edmonton, Alta.
¶Department of Emergency Medicine and Center for Clinical Epidemiology and Evaluation, Vancouver General Hospital, Vancouver, BC

Received: Feb.1, 2003; final submission: July 9, 2003 ; accepted: July 18, 2003

This article has been peer reviewed.

Can J Emerg Med 2003;5(6):406-11

Introduction

This is the second of a 2-part series designed to provide emergency physicians with the background necessary to locate, critically evaluate and interpret systematic reviews (SRs). Part I provided a brief background on SRs and general principles of locating and critically appraising SRs.[1] Part II will broaden the discussion of critical appraisal principles by focusing on quality assessment, data synthesis and interpretation of results. To enhance readability, illustrative examples from the emergency medicine literature have been included and technical details have been kept to a minimum. The references, however, are comprehensive and provide a resource for readers seeking further information.

Appraising studies included in systematic reviews

After relevant studies have been identified for inclusion in an SR, authors should critically appraise these studies and describe the appraisal criteria so that readers can determine the validity of the studies, identify reasons for differences in study outcomes and judge the relevance of the review conclusions to their own clinical practice.[2,3] Critical appraisal provides an assessment of the quality of the primary research — quality referring to the extent that the study design, conduct and analysis minimize the potential for bias. This is important because biased studies are more likely to report misleading (usually positive) results that may substantially alter the SR conclusions.

Quality scores

The appraisal process involves an assessment of the patients, the study intervention, and outcome measurements in each study. Differences in any of these design features can lead to differences in study outcomes. In appraising primary research, reviewers may choose various tools, scales and approaches, but the most important methodologic criteria to assess are concealment of allocation, blinding, randomization and descriptions of patients lost to follow-up.[4–6] The most widely accepted tools for evaluating primary studies are the Cochrane and Jadad approaches, which have both been tested and validated. The Cochrane approach is based on the description of the concealment of allocation, and places studies in 1 of 3 categories: concealed allocation, unclear or clearly not concealed allocation. The Jadad criteria assess randomization, blinding and description of losses to follow-up using a 0–5 scale, with scores of 3–5 considered high quality. While these scales are widely used, they are applicable only to certain types of studies (i.e., randomized blinded pharmaceutical trials), and some researchers suggest they may be unreliable.[7] Although a full discussion of this controversy is beyond the scope of this article, it is clearly essential that some assessment of quality must be used, reported and applied to sensitivity analyses.

Quality scores may be applied in both qualitative and quantitative SRs but are generally not used to guide inclusion decisions. Instead they are utilized after inclusion to gauge the strength of evidence and to assist in the performance of sensitivity analyses. Reviewers sometimes use these scores to perform weighted analyses in which the relative weight of a selected study in a meta-analysis is determined by the magnitude of its methodological quality score.

Outcomes

High quality SRs identify explicit primary and secondary outcome measures. These measures should be clinically important and they should be specified a priori to avoid post-hoc analysis ("data dredging"). To illustrate, a recent SR of intravenous magnesium sulfate in acute asthma chose admission to hospital as the primary outcome, relegating other often used but less important measures (e.g., pulmonary function testing) to secondary status.[8] High quality SRs also report adverse events and economic outcomes. For example, an SR of the use of low molecular weight heparin (LMWH) in venous thromboembolism pooled data on thrombocytopenia and major bleeding as important secondary outcomes.[9] In this review, although the individual trials failed to detect differences between LMWH and unfractionated heparin with respect to these outcomes, pooled data confirmed statistically and clinically significant reductions in thrombocytopenia and major bleeding with LMWH.

Data synthesis

It is only reasonable to pool data from different studies if the studies are sufficiently similar, and the decision whether or not to pool data is one of the most important decisions a reviewer must make. High quality reviews painstakingly evaluate similarities among studies, considering patient population, intervention, control, outcomes and design. In many cases, included studies differ too much for pooling, and reviewers must limit themselves to a qualitative approach. If the studies are sufficiently similar, reviewers should employ and report the explicit and appropriate methods used for data synthesis. Unfortunately, this

is not always done, and only 48% of reviews published in leading emergency medicine journals reported the methods used to combine the findings of the relevant studies.[2]

For dichotomous outcomes (e.g., death, relapse), most reviewers report odds ratios (ORs) or relative risk (RR) with associated 95% confidence intervals (95% CI) for each individual trial and for the overall pooled estimate (illustrated by Figs. 1 and 2). The rationale for selecting OR versus RR is complicated and often based on tradition. Technically, RR estimates can only be generated from cohort studies, although ORs can be viewed as a practical approximation of RR.[11] A previous paper in this section provides a review of these and other measures of association.[12] Increasingly, reviews use RR as a method of reporting results, since RR is most appropriate to the randomized clinical trial designs combined in SRs. When displaying these data, the convention is that the effects favouring the experimental treatment are located to the left of the line of unity (1.0) while those favouring the control or comparison arm are located to the right of the line of unity. When the 95% CI crosses the line of unity, the result is considered non-significant (Fig. 1).

For continuous measures with standard units (e.g., height, blood pressure, airflow measurements), a weighted mean difference or effect size is calculated. The weight of each trial's contribution to the overall pooled result is based on the inverse of the trial's variance. For continuous outcomes, variance is largely a function of the standard deviation and sample size: the lower the standard deviation and the larger the sample size, the greater contribution the study makes to the pooled estimate. For continuous measures with variable units (e.g., quality of life or other functional scales), the use of a standardized mean difference is often used. For both the standardized mean difference and a weighted mean difference, the convention is the opposite of that for dichotomous variables, that is, effects favouring the experimental treatment are located to the **right** of the line of unity (0) while those favouring the control or comparison arm are plotted to the **left**. Once again, when the 95% CI crosses the line of unity, the result is considered non-significant.

Number needed to treat (NNT) is another way to express a measure of effect.[11] In the Cochrane Library reviews, the absolute risk reduction is represented by the risk reduction statistic, and the inverse of this (and its 95% CI) provides the NNT estimation. A less exact method is to examine the pooled percentages in each column. For example, in the meta-analysis on corticosteroid use in acute asthma to prevent admission the OR was 0.75 (95% CI, 0.63–0.86).[13] The risk reduction was 0.13, resulting in a NNT of 8 (95%

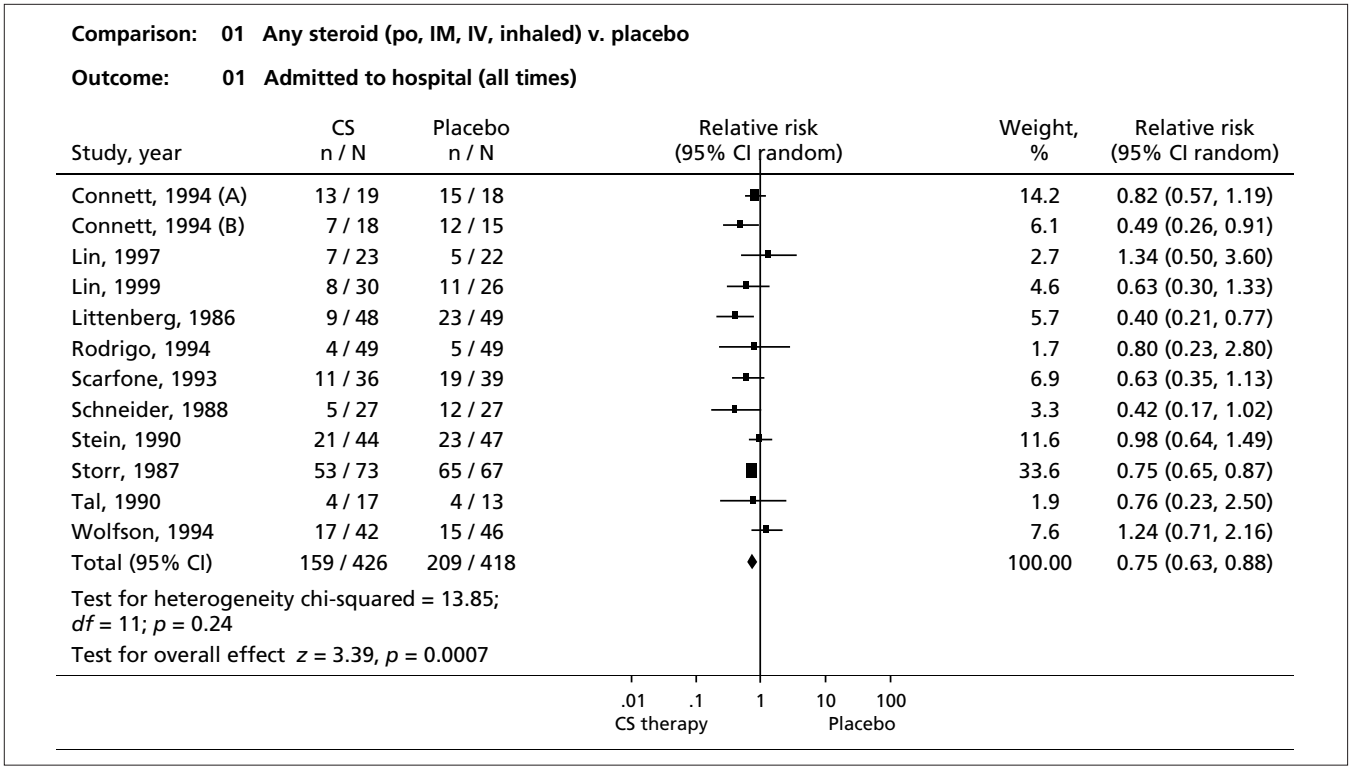

Fig. 1. Plot of odds ratio with 95% confidence intervals (CIs) from meta-analysis. po = by mouth; IM = intramuscular; IV = intravenous; CS = corticosteroid group. Reproduced from Rowe et al,[10] with permission from John Wiley and Sons Limited.

CI, 5–20). By subtracting the approximate admission rate in the control group (0.50) from that in the treatment group (0.37), one obtains an absolute risk reduction of 0.13 and a similar NNT of 8. Caution is advised, since this latter approach is an approximation.

Many clinicians are unfamiliar with the statistical methods applied in SRs. Although statistics can be used inappropriately, reviews that are of high quality based on the parameters discussed in this article are unlikely to fail because of improper statistical analysis. If readers are satisfied that the authors have pooled studies with similar designs, populations, interventions, controls and outcomes, the SR conclusions are probably valid. If these criteria are not satisfied, then no amount of statistical manipulation will salvage the review. Finally, when heterogeneity (discussed below) is visually or statistically apparent, this must be addressed.

Statistical and clinical heterogeneity

Different trials studying the same intervention rarely come to identical conclusions with respect to the estimated treatment effect. Heterogeneity refers to the differences, or variability, between studies in their estimated treatment effects. Statistical tests can be applied to determine if the degree of variability between studies is greater than that expected by chance. This type of variability is called statistical heterogeneity. If statistical heterogeneity exists, the pooled result should be viewed with caution and reasons for the variability should be explored. Statistical heterogeneity usually results from "clinical heterogeneity" (e.g., differences in patient populations, disease severity or interventions applied). In a recent study, over 30% of the SRs assessed provided insufficient information to determine whether the findings were combined appropriately. In the SRs that did provide this information, the authors had pooled statistically heterogeneous data in more than 25% of cases.[2]

Much has been written about handling clinical heterogeneity in quantitative SRs. At the same time, the importance of addressing heterogeneity in qualitative SRs has perhaps been underestimated. Formal rules for categorizing trials according to methodology, trial quality, and type of intervention or comparison (e.g., placebo v. active-controlled) are essential so that readers can understand how the authors evaluated heterogeneity in both qualitative and quantitative SRs. Readers can apply two crude assessments of heterogeneity. First, they can look to see if the studies all have similar populations, interventions, outcomes and designs. If so, there is theoretical homogeneity and it is probably rational to combine the data and calculate a summary statistic. Second, they can inspect the scatter plots, which will visually reveal heterogeneity if it exists (see Figs. 2a and 2b).

Sensitivity and subgroup analyses

Subgroup analyses, based on patient characteristics like age, gender, ethnicity, comorbidity or severity of presentation, are often performed to determine whether key outcomes or findings differ in patient groups. Sensitivity analyses, based on non-patient characteristics like medication dose, methodological quality and type of statistical analysis, are primarily utilized to explore heterogeneity and determine the robustness of the pooled results. For example, if the pooled results of an SR are robust, the sensitivity analysis will show qualitatively similar outcomes regardless of the study design or statistical analysis. If results are not robust, the sensitivity analysis may show that outcomes or treatment effects are qualitatively different when different methods are used. In a recent SR looking at the impact of magnesium sulfate on asthma admission rates, subgroup analysis revealed that patients with severe asthma benefited much more than patients with mild or moderate asthma.[8] Researchers have developed criteria to determine whether subgroup analysis is appropriate. Biological plausibility, a priori subgroup identification, statistically and clinically important effect sizes, a limited number of subgroups analyzed, indirect supporting evidence, within- versus between-study differences, and consistency across studies are all factors to be considered when deciding whether subgrouping is valid.[14]

Reporting and interpreting results

In the past, authors used widely diverse methods or no methods at all when writing review articles. These concerns led to the creation of the QUOROM (Quality of Reporting of Meta-analyses) statement, a guideline for reporting the methods and results of SRs.[5] The QUOROM statement is to SRs what the CONSORT statement was to randomized clinical trials, and many biomedical journals have since endorsed the QUOROM reporting style.[15]

Care is required at this final stage of the review, since, here, authors are not bound by explicit methods. Readers must be cautious of reviewers who use terms such as "trend towards significance," "almost significant" or similar subjective statements. A pooled estimate is, by convention, only statistically significant if the 95% confidence interval does not cross the line of unity (described above). If

it does cross, reviewers cannot claim a treatment is superior. Neither can they claim there is "no difference" between treatments when the CI includes unity, because in many cases the confidence range includes values that would represent clinically important effects favouring one, or even both, treatments. If the CI includes unity, it is usually best to say that "no evidence of a statistically significant difference between the treatments was detected."

Equivalence is particularly difficult to demonstrate, and reviewers should only conclude equivalence if the 95% CI is narrow and does not include any values that would represent a clinically important effect favouring either agent. When an SR fails to show a clear benefit or adverse impact of the treatment under study, the authors' discussion of the review implications should be viewed skeptically since, in reality, there are no clear implications. In addition, careful attention should be paid to whether all appropriate measures of benefit and harm were addressed. An SR that touts

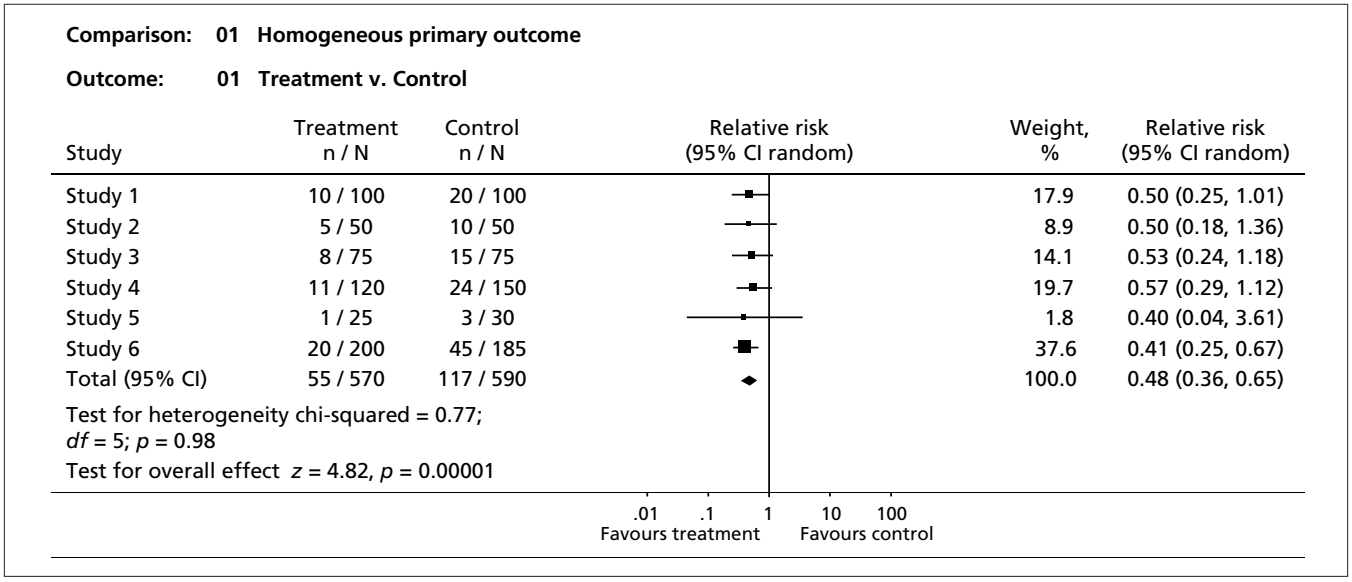

Fig. 2a. Homogeneity of study results from meta-analysis. Note: Results demonstrate visual (point estimates to the left of the neutral line, overlapping 95% CIs, same direction of effect) and statistical (*p* = 0.98) homogeneity. Reproduced from Rowe et al,[10] with permission from John Wiley and Sons Limited.

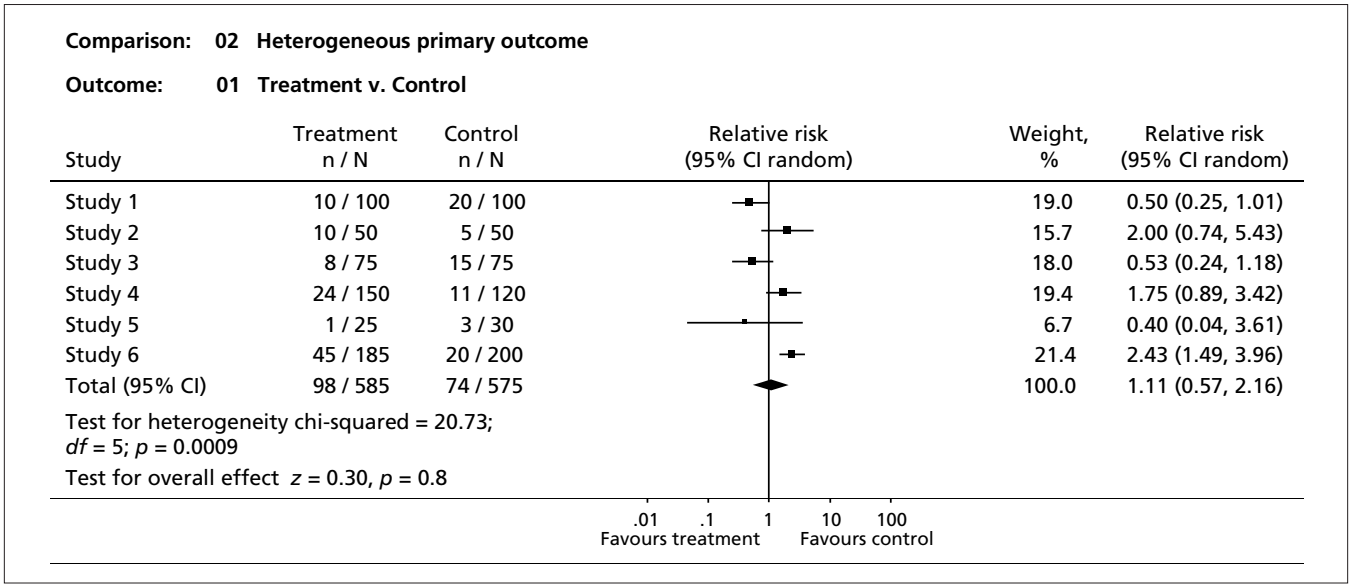

Fig. 2b. Heterogeneity of study results from meta-analysis. Note: Results demonstrate visual (point estimates to the left and right of the neutral line, non-overlapping 95% CIs, different direction of effect) and statistical (*p* = 0.0009) heterogeneity. Reproduced from Rowe et al,[10] with permission from John Wiley and Sons Limited.

the efficacy of an intervention without considering its toxicity or adverse effects may be misleading.

Summary

SRs are increasingly prevalent in the emergency medicine literature. Properly performed SRs can have far reaching implications for both patients and physicians, but poorly performed SRs can be meaningless. Clinicians should consider the clinical question posed, the methods used to identify studies, the assessment of their quality, the methods used to combine results and the appropriateness of the resulting conclusions. Without an understanding of the rigorous methodology required to produce valid SRs, readers are unlikely to correctly interpret and apply the results and conclusions presented.

Competing interests: None declared.

Acknowledgments: Dr. Brian Rowe is supported by the Canadian Institute of Health Research (CIHR) as a Canada Research Chair in Emergency Airway Diseases. Dr. Riyad B. Abu-Laban is supported by a Clinical Scholar award from the Michael Smith Foundation for Health Research.

References

1. Zed PJ, Rowe BH, Loewen PS, Abu-Laban RB. Systematic reviews in emergency medicine: Part I. Background and general principles for locating and critically appraising reviews. Can J Emerg Med 2003;5(5):331-5.
2. Kelly KD, Travers A, Dorgan M, Slater L, Rowe BH. Evaluating the quality of systematic reviews in the emergency medicine literature. Ann Emerg Med 2001;38:518-26.
3. Meade MO, Richardson S. Selecting and appraising studies for systematic review. Ann Intern Med 1997;127:531-7.
4. 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.
5. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet 1999;354:1896-900.
6. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17:1-12.
7. Clark HD, Wells GA, Huet C, McAlister FA, Salmi LR, Fergusson D, et al. Assessing the quality of randomized trials: reliability of the Jadad scale. Control Clin Trials 1999;20:448-52.
8. Rowe BH, Bretzlaff JA, Bourdon C, Bota GW, Camargo CA Jr. Intravenous magnesium sulfate treatment for acute asthma in the emergency department: a systematic review of the literature. Ann Emerg Med 2000;36:181-90.
9. van den Belt AGM, Prins MH, Lensing AWA, Castro AA, Clark OAC, Atallah AN, et al. Fixed dose subcutaneous low molecular weight heparins versus adjusted dose unfractionated heparin for venous thromboembolism [Cochrane review]. In: The Cochrane Library; Issue 2, 2001. Oxford: Update Software.
10. Rowe BH, Spooner CH, Ducharme FM, Bretzlaff JA, Bota GW. Early emergency department treatment of acute asthma with systemic corticosteroids [Cochrane review]. In: The Cochrane Library; Issue 2, 2000. Oxford: Update Software.
11. Clarke M, Oxman AD, editors. Cochrane reviewers' handbook 4.1.6 [updated Jan 2003]. In: The Cochrane Library; Issue 1, 2003. Oxford: Update Software. Updated quarterly.
12. Worster A, Rowe BH. Measures of association: an overview with examples from Canadian emergency medicine research. Can J Emerg Med 2001;3(3):219-23.
13. Rowe BH, Spooner CH, Ducharme FM, Bretzlaff JA, Bota GW. Corticosteroids for preventing relapse following acute exacerbations of asthma [Cochrane review]. In: The Cochrane Library; Issue 1, 2003. Oxford: Update Software.
14. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992;116:78-84.
15. Moher D, Schultz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001;357:1191-4.

Correspondence to: Dr. Peter J. Zed, CSU Pharmaceutical Sciences, Vancouver General Hospital, 855 West 12th Ave., Vancouver BC V5Z 1M9; 604 875-4077, fax 604 875-5267, zed@interchange.ubc.ca