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 élargissant 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.

**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. 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. 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. 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. 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. 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. 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. 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. A previous paper in this section provides a review of these and other measures of association. 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. 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). The risk reduction was 0.13, resulting in a NNT of 8 (95% CI range 5–13) for any difference in asthma admission rates.

| Comparison: 01 Any steroid (po, IM, IV, inhaled) v. placebo |
| Outcome: 01 Admitted to hospital (all times) |
| Study, year | CS n / N | Placebo n / N | Relative risk (95% CI random) | Weight, % | Relative risk (95% CI random) |
| Connett, 1994 (A) | 13 / 19 | 15 / 18 | 14.2 | 0.82 (0.57, 1.19) |
| Connett, 1994 (B) | 7 / 18 | 12 / 15 | 6.1 | 0.49 (0.26, 0.91) |
| Lin, 1997 | 7 / 23 | 5 / 22 | 2.7 | 1.34 (0.50, 3.60) |
| Lin, 1999 | 8 / 30 | 11 / 26 | 4.6 | 0.63 (0.30, 1.33) |
| Littenberg, 1986 | 9 / 48 | 23 / 49 | 5.7 | 0.40 (0.21, 0.77) |
| Rodrigo, 1994 | 4 / 49 | 5 / 49 | 1.7 | 0.80 (0.23, 2.80) |
| Scarfone, 1993 | 11 / 36 | 19 / 39 | 6.9 | 0.63 (0.35, 1.13) |
| Schneider, 1988 | 5 / 27 | 12 / 27 | 3.3 | 0.42 (0.17, 1.02) |
| Stein, 1990 | 21 / 44 | 23 / 47 | 11.6 | 0.98 (0.64, 1.49) |
| Storr, 1987 | 53 / 73 | 65 / 67 | 33.6 | 0.75 (0.65, 0.87) |
| Tal, 1990 | 4 / 17 | 4 / 13 | 1.9 | 0.76 (0.23, 2.50) |
| Wolfson, 1994 | 17 / 42 | 15 / 46 | 7.6 | 1.24 (0.71, 2.16) |
| Total (95% CI) | 159 / 426 | 209 / 418 | 100.00 | 0.75 (0.63, 0.88) |

Test for heterogeneity chi-squared = 13.85; df = 11; p = 0.24
Test for overall effect z = 3.39, p = 0.0007

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, with permission from John Wiley and Sons Limited.
of cases.2
The pooled result should be viewed with caution and rea-
sons for the variability should be explored. Statistical het-
erogeneity 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 deter-
mate and clinically important effect sizes, a limited num-
bers of subgroups analyzed, indirect supporting evidence,
within- versus between-study differences, and consistency
across studies are all factors to be considered when decid-
ing whether subgrouping is valid.14

Sensitivity and subgroup analyses
Subgroup analyses, based on patient characteristics like
age, gender, ethnicity, comorbidity or severity of presenta-
tion, are often performed to determine whether key out-
comes or findings differ in patient groups. Sensitivity
analyses, based on non-patient characteristics like medica-
tion dose, methodological quality and type of statistical
analysis, are primarily utilized to explore heterogeneity
and determine the robustness of the pooled results. For ex-
ample, if the pooled results of an SR are robust, the sensi-
tivity analysis will show qualitatively similar outcomes re-
gardless of the study design or statistical analysis. If results
are not robust, the sensitivity analysis may show that out-
comes 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. Bio-
logical plausibility, a priori subgroup identification, statisti-
cally and clinically important effect sizes, a limited num-
er of subgroups analyzed, indirect supporting evidence,
within- versus between-study differences, and consistency
across studies are all factors to be considered when decid-
ing 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 con-
cerns led to the creation of the QUOROM (Quality of Re-
porting of Meta-analyses) statement, a guideline for report-
ing the methods and results of SRs.7 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 simi-
lar subjective statements. A pooled estimate is, by conven-
tion, only statistically significant if the 95% confidence in-
terval 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

---

### Comparison: 01 Homogeneous primary outcome

#### Outcome: 01 Treatment v. Control

<table>
<thead>
<tr>
<th>Study</th>
<th>Treatment n / N</th>
<th>Control n / N</th>
<th>Relative risk (95% CI random)</th>
<th>Weight, %</th>
<th>Relative risk (95% CI random)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>10 / 100</td>
<td>20 / 100</td>
<td>17.9 0.50 (0.25, 1.01)</td>
<td>100.0</td>
<td>0.48 (0.36, 0.65)</td>
</tr>
<tr>
<td>Study 2</td>
<td>5 / 50</td>
<td>10 / 50</td>
<td>8.9 0.50 (0.18, 1.36)</td>
<td>100.0</td>
<td>0.48 (0.36, 0.65)</td>
</tr>
<tr>
<td>Study 3</td>
<td>8 / 75</td>
<td>15 / 75</td>
<td>14.1 0.53 (0.24, 1.18)</td>
<td>100.0</td>
<td>0.48 (0.36, 0.65)</td>
</tr>
<tr>
<td>Study 4</td>
<td>11 / 120</td>
<td>24 / 150</td>
<td>19.7 0.57 (0.29, 1.12)</td>
<td>100.0</td>
<td>0.48 (0.36, 0.65)</td>
</tr>
<tr>
<td>Study 5</td>
<td>1 / 25</td>
<td>3 / 30</td>
<td>1.8 0.40 (0.04, 3.61)</td>
<td>100.0</td>
<td>0.48 (0.36, 0.65)</td>
</tr>
<tr>
<td>Study 6</td>
<td>20 / 200</td>
<td>45 / 185</td>
<td>37.6 0.41 (0.25, 0.67)</td>
<td>100.0</td>
<td>0.48 (0.36, 0.65)</td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td>55 / 570</td>
<td>117 / 590</td>
<td>100.0 0.48 (0.36, 0.65)</td>
<td>100.0</td>
<td>0.48 (0.36, 0.65)</td>
</tr>
</tbody>
</table>

Test for heterogeneity chi-squared = 0.77; $df = 5; p = 0.98$

Test for overall effect $z = 4.82, p = 0.00001$

---

### Comparison: 02 Heterogeneous primary outcome

#### Outcome: 01 Treatment v. Control

<table>
<thead>
<tr>
<th>Study</th>
<th>Treatment n / N</th>
<th>Control n / N</th>
<th>Relative risk (95% CI random)</th>
<th>Weight, %</th>
<th>Relative risk (95% CI random)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>10 / 100</td>
<td>20 / 100</td>
<td>19.0 0.50 (0.25, 1.01)</td>
<td>100.0</td>
<td>0.57 (0.36, 0.80)</td>
</tr>
<tr>
<td>Study 2</td>
<td>10 / 50</td>
<td>5 / 50</td>
<td>15.7 2.00 (0.74, 5.43)</td>
<td>100.0</td>
<td>0.57 (0.36, 0.80)</td>
</tr>
<tr>
<td>Study 3</td>
<td>8 / 75</td>
<td>15 / 75</td>
<td>18.0 0.53 (0.24, 1.18)</td>
<td>100.0</td>
<td>0.57 (0.36, 0.80)</td>
</tr>
<tr>
<td>Study 4</td>
<td>24 / 150</td>
<td>11 / 120</td>
<td>19.4 1.75 (0.89, 3.42)</td>
<td>100.0</td>
<td>0.57 (0.36, 0.80)</td>
</tr>
<tr>
<td>Study 5</td>
<td>1 / 25</td>
<td>3 / 30</td>
<td>6.7 0.40 (0.04, 3.61)</td>
<td>100.0</td>
<td>0.57 (0.36, 0.80)</td>
</tr>
<tr>
<td>Study 6</td>
<td>45 / 185</td>
<td>20 / 200</td>
<td>21.4 2.43 (1.49, 3.96)</td>
<td>100.0</td>
<td>0.57 (0.36, 0.80)</td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td>98 / 585</td>
<td>74 / 575</td>
<td>100.0 1.11 (0.57, 2.16)</td>
<td>100.0</td>
<td>0.57 (0.36, 0.80)</td>
</tr>
</tbody>
</table>

Test for heterogeneity chi-squared = 20.73; $df = 5; p = 0.0009$

Test for overall effect $z = 0.30, p = 0.8$
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


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