Intern Emerg Med (2012) 7:563–568
DOI 10.1007/s11739-012-0846-1

CE - COCHRANE CLINICAL EVIDENCE AND HEALTH TECHNOLOGY ASSESSMENT

Multiple systematic reviews: methods for assessing discordances of results

Lorenzo Moja · M. Pilar Fernandez del Rio · Rita Banzi · Cristina Cusi · Roberto D'Amico · Alessandro Liberati · Giovanni Lodi · Ersilia Lucenteforte · Silvia Minozzi · Valentina Pecoraro · Gianni Virgili · Elena Parmelli

Received: 3 July 2012 / Accepted: 8 August 2012 / Published online: 2 September 2012

Abtract

Background The process of systematically reviewing research evidence is useful for collecting, assessing and summarizing results from multiple studies planned to answer the same clinical question. The term "systematic" implies that the process, besides being organized and complete, is transparent and fully reported to allow other independent researchers to replicate the results, and therefore come to the same conclusions. Hundreds of new systematic reviews are indexed every year. The growing number increases the likelihood of finding multiple and discordant results.

Objectives To clarify the impact of multiple and discordant systematic reviews, we designed a program aimed at finding out: (a) how often different systematic reviews are done on the same subject; (b) how often different systematic reviews on the same topic give different results or conclusions; (c) which methods or interpretation

L. Moja
Dipartimento di Scienze Biomediche per la Salute,
Università degli Studi di Milano, Milan, Italy
e-mail: lorenzo.moja@unimi.it

L. Moja
Unità di Epidemiologia Clinica,
IRCCS Istituto Ortopedico Galeazzi, Milan, Italy

L. Moja · R. Banzi · V. Pecoraro
Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy
e-mail: rita.banzi@marionegri.it

V. Pecoraro
e-mail: valentina.pecoraro@marionegri.it

M. P. Fernandez del Rio
Fundacion para la Formacion e Investigacion Sanitaria
de la Region de Murcia (FFIS), Murcia, Spain
e-mail: mariapilar.fern@gmail.com

C. Cusi
Cochrane Neurological Field, Direzione
Sanità e Servizi Sociali, Regione Umbria, Perugia, Italy
e-mail: cusi.cris@gmail.com

R. D'Amico · A. Liberati · E. Parmelli
Dipartimento di Medicina Diagnostica,
Clinica e di Sanità Pubblica, Centro Cochrane Italiano,
Università di Modena e Reggio Emilia, Modena, Italy
e-mail: roberto.damico@unimore.it

E. Parmelli
e-mail: elena.parmelli@unimore.it

G. Lodi
Dipartimento di Scienze Biomediche, Chirurgiche e
Odontoiatriche, Università degli Studi di Milano, Milan, Italy
e-mail: giovanni.lodi@unimi.it

E. Lucenteforte (✉)
Dipartimento di Farmacologia Preclinica e Clinica,
Università degli Studi di Firenze, viale Gaetano Pieraccini,
6-50139 Florence, Italy
e-mail: ersilia.lucenteforte@unifi.it

S. Minozzi
Dipartimento di Epidemiologia, Regione Lazio,
Rome, Italy
e-mail: minozzi.silvia@gmail.com

G. Virgili
Dipartimento di Scienze Chirurgiche Specialistiche,
Università degli Studi di Firenze, Florence, Italy
e-mail: gianni.virgili@unifi.it

characteristics can explain the differences in results or conclusions.

Methods This paper outlines the method used to explore the frequency and the causes of discordance among multiple systematic reviews on the same topic. These methods were then applied to a few medical fields as case studies.

Conclusion This aim is particularly relevant for both clinicians and policy makers. Judgments about evidence and recommendation in health care are complex, and often rely on discordant results, especially when there are no empirical results to help serve as a guideline.

Introduction

The number of systematic reviews published has risen dramatically, and it has been estimated that approximately 2,500 new publications are indexed annually on Medline [1]. This increases the likelihood of finding multiple and discordant results.

The process of systematically reviewing research evidence is useful for collecting, assessing and summarizing results from multiple studies planned to answer the same clinical question. The term "systematic" implies that the process, besides being organized and complete, is transparent and well reported, so that other independent researchers following the same methods can replicate the results and therefore come to similar conclusions [2, 3]. But what can be done when there is more than one systematic review, published by independent researchers, aimed at answering the same clinical question, but ends up with discordant results?

In 1997, the question of interpreting discordant results from similar systematic reviews was first addressed by Jadad et al. [4]. Results can differ, or the interpretation and inferences made by the review authors can be discordant. Discordance can arise between systematic reviews regarding quantitative results (direction, magnitude or significance) or their interpretation. The following potential sources of discordance are identified: clinical question, including the dimensions considered by the popular acronym—patients, intervention, comparison and outcomes (PICO); study selection and inclusion; data extraction; assessment of study risk of bias; assessment of the ability to combine studies; and the statistical methods used for data synthesis.

In this paper, we present an approach to assess the scientific validity and reproducibility of the results of multiple systematic reviews. We targeted multiple systematic reviews on the same topic to see how often they agreed or came to different conclusions/interpretations. We examined: (a) how often different systematic reviews are done on the same subject; (b) how often different systematic reviews try to answer the same question, yet reach different results or conclusions; (c) which methods or interpretation characteristics can explain the differences in results or conclusions. We developed and applied these methods to select few medical fields and their relative systematic reviews. The context for this methodological development is a research program of work in systematic review science. The team of researchers working on this program comprised of specialty clinicians, clinical epidemiologists and biostatisticians, whom we have called: the 'Discordance Team'.

Methods

Basic methods basis for all phases

We divided our methods into phases (Fig. 1) providing transparent and accurate reporting of what has been done in each phase. To generate the multiple systematic reviews dataset, we started an iterative screening of large numbers of eligible records concerning selected medical conditions. At each step, eligibility was assessed independently by two reviewers, following standard rules and using ad hoc forms. Operational guidelines were reported in a background document available to all reviewers. Disagreements were resolved by consensus while arbitration with a third reviewer was possible when necessary.

Phase 1: information source

To identify overlapping systematic reviews, we used the clinical evidence search strategy process and outputs [5]. Briefly, clinical evidence is an authoritative decision-support resource, which summarises the current state of knowledge and uncertainty about the prevention and treatment of clinical conditions, based on thorough searches and appraisal of the literature. For each topic of interest, clinical evidence's information specialists search for systematic reviews in several databases—Cochrane database of systematic reviews, Medline, Embase, and other databases (e.g., PsycInfo) as appropriate. Additional resources are: Centre for Reviews and Dissemination (CRD) website, database of abstracts of reviews of effects (DARE) online database, Health Technology Assessment (HTA) online database, and National Institute for Health and Clinical Excellence (NICE) website. Clinical evidence search strategies and relative filters, all created for the OVID online search interface, are based on strategies developed by in-house BMJ evidence centre information specialists and others (e.g., the Haynes team [6], the Cochrane collaboration [7, 8]). Clinical evidence search strategies are regularly updated.

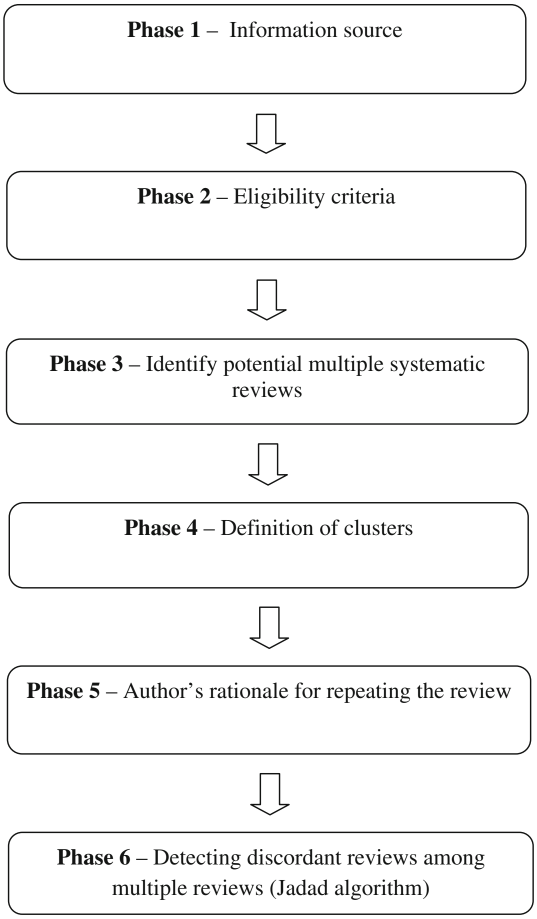

Phase 1 – Information source
Phase 2 – Eligibility criteria
Phase 3 – Identify potential multiple systematic reviews
Phase 4 – Definition of clusters
Phase 5 – Author's rationale for repeating the review
Phase 6 – Detecting discordant reviews among multiple reviews (Jadad algorithm)

Fig. 1 Methodology phase

For any topic of interest, the clinical evidence search strategy outputs were collated, and passed to the Discordance Team.

In addition to searching these databases, through clinical evidence strategies, we used two supplementary approaches: (a) we checked relevant chapter in the clinical evidence database to identify any additional relevant systematic reviews introduced by clinical evidence contributors following the search process; (b) we checked the reference lists of all selected systematic reviews to identify other relevant ones. We only considered documents published in scientific journals (i.e., we excluded meeting proceedings, theses or pharmacological datasets).

Phase 2: eligibility criteria

In phase 2, the reviewers screened the retrieved records (typically title or title and abstract) to identify the proportion of truly systematic reviews on any one selected condition (e.g., myocardial infarction, colorectal cancer). Operational items were:

Inclusion we only included systematic reviews of efficacy or safety that mentioned the terms "systematic review" or "meta-analysis" in the title or abstract, or reported that there had been a search in at least one bibliographic database (i.e., Medline). We included systematic reviews irrespective of their qualitative or quantitative nature. Systematic reviews had to have been published from January 1997 to December 2007. Languages of publication were: English, Italian, Spanish, or French.

Exclusion we excluded systematic reviews embedded in editorials, correspondence/letters, or reports of randomised controlled trials, those published only as abstracts, recommendations from consensus conferences, primary studies, systematic reviews of other diseases, systematic reviews of diagnostic accuracy, prognosis and economic assessments of treatments and systematic reviews published before 1997 and after 2007.

Phase 3: identify potential multiple systematic reviews

Reviewers extracted information from the title and abstract to identify potentially multiple systematic reviews. In this study, potential multiplicity has been defined as at least two independent systematic reviews sharing the same population, condition/pathology and intervention, irrespective of the sources of clinical heterogeneity for outcomes and controls. Reviewers extracted the following details from the titles and abstract: ID, author, title, complete reference, type of systematic review (safety, efficacy or both), population, condition/pathology and intervention. Given the iterative nature of the selection, exclusion was still possible. Reasons for exclusion, such as a duplicate publication or a narrative nature of the review, were recorded (Box 1).

Phase 4: definition of clusters

In this study, a cluster is made up of at least two independent systematic reviews with the same objective, population, condition/disease, intervention, control, and at least one outcome, reaching a status of clinical homogeneity (multiple systematic reviews). Phase 4 differed from phase 3 because researchers had access to the full text of reviews, analysing the overlap of control intervention and outcomes. If overlapping was complete, we signed the systematic reviews off as 'truly' multiple. From the full texts, reviewers extracted extensive details to generate the final list of clusters: ID, author, title, reference, objective, population, condition/pathology, experimental intervention, control intervention, outcome measures, information sources and search, study design, number of studies included, quantitative results (only if meta-analyses were presented), and interpretation of results. We did double entry of all details to ensure data quality. To batch systematic reviews within clusters, we sequentially and manually filtered the objective, population, condition/disease,

Box 1 Main criteria for exclusion

Narrative review [16–20]

- Broad objective (not addressable in experiments)
- Highly incomplete entry for PICOS (participants, interventions, comparators/controls, outcomes, and study design)
- Complete lack of details about information sources and search

Duplicate/double publications [21–24]

- The same authors, although the order of names could be reversed or different
- Substantial amount of the same methods (i.e., eligibility criteria, information sources, planned methods of analysis)
- Duplication of outcomes and studies included
- Same or very similar results and conclusions

experimental and control interventions, and outcomes, searching for match dimensions.

There was no one-to-one relationship between individual systematic reviews and clusters: an individual article could serve more than one purpose, and be part of more than one cluster.

After extensive attempts, some of the reviews could not be matched to others, and thus became a single systematic review on a specific topic.

Phase 5: author's rationale for repeating the review

In each systematic review within clusters, we searched for references to previous overlapping systematic reviews. We investigated whether one review had cited previous reviews on the basis of the date of the last literature search. If this date was not available, we used its acceptance for publication date, publication date or the date of the most recent citation in the references. When a previous review was cited, we abstracted the authors' rationale for repeating the review, if reported [9].

Phase 6: detecting discordant reviews among multiple reviews (Jadad algorithm)

Meta-analysis methods and results

For each cluster of multiple systematic reviews, we examined concordance or discordance for direction and for the statistical significance of meta-analysis results, if available. The meta-analysis was classified according to [10]: outcome type (efficacy or safety), type of effect size [based on means (standardized or unstandardized mean difference or response ratios), binary (risk ratio, odds ratio and risk difference) or time-to-event (hazard ratio)], primary or secondary outcomes, statistical significance, completeness of estimate reporting (Fig. 2), statistical methods (e.g., Mantel–Haenszel for binary data), models (e.g., fixed effect), and measures of heterogeneity (e.g., Q statistic).

Overall estimate reporting was defined on four primary levels based on the completeness of data presented in the results section of the publications (Fig. 2) [11]. The nature and amount of data required to compare results from meta-analysis are shown in Box 2. A fully reported overall estimate had all the details necessary to compare the results of the meta-analysis. Partially reported overall estimate had some but not all of the data necessary for meta-analysis, and a qualitatively reported overall estimate had no reliable data except for a statement on effect size and its precision. Unreported overall estimates were those in which the publication provided no data even though the outcome was specified in either the Methods or Discussion. Finally, meta-analyses not done were those that provided reasons for not calculating a summary estimate.

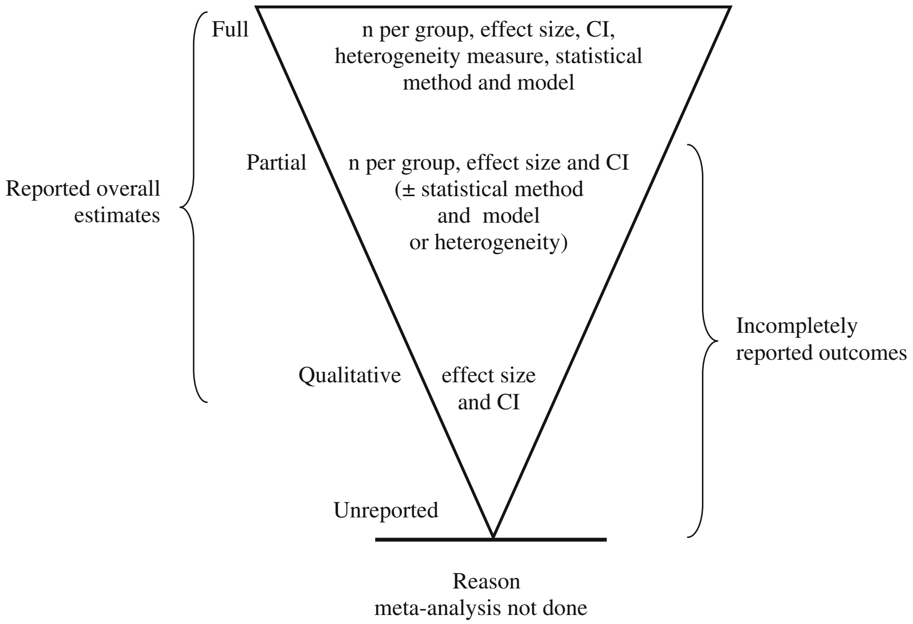
To assess discordances on direction, we moved from a model in which measures of association lower than 1.0 meant that outcomes had a favourable profile for the intervention group (e.g., an HR of 0.80 for overall survival means a reduction in mortality).

To assess discordances on statistical significance, we moved from a model in which a p value <0.05 was statistically significant. This was chosen based on standard practice in the research community where p values <0.05 are often reported as "statistically significant", and interpreted as being small enough to justify rejection of the null hypothesis.

A priori we reasoned that two meta-analyses of the same outcome (e.g., response rate) were discordant if they presented an effect size based on the same data type (i.e., binary) but: (a) the results were in the opposite direction to the no-difference value (e.g., one for relative risk), i.e., a 'qualitative interaction'; (b) the two effect sizes were in the same direction, but just one reached statistical significance, which we called 'significance cliché'; (c) there was a statistical criterion, i.e., when the heterogeneity within the two effect sizes was significant (p significant for values less than 0.1), i.e., a 'quantitative interaction' [12].

Although comparing effect estimates in different groups by considering the meta-analysis results from each effect size separately (criteria a and b) is considered at best naïve approach, it is commonly done by review authors. These were considered discordances since they may cause difference in the interpretation of results. In our study, we considered 'qualitative interaction' and 'quantitative interaction' following the definitions given by Yusuf 1991 [13]. Qualitative interaction exists if the direction of effect is reversed, meaning an intervention is beneficial in one meta-analysis, but harmful in another. Quantitative interaction exists when the magnitude of the effect varies but not the direction, so an intervention is beneficial to

Fig. 2 Levels of overall estimate reporting (*n* number of participants per group, *CI* confidence intervals)


Full
n per group, effect size, CI, heterogeneity measure, statistical method and model
Partial
n per group, effect size and CI (± statistical method and model or heterogeneity)
Reported overall estimates
Qualitative
effect size and CI
Unreported
Incompletely reported outcomes
Reason meta-analysis not done

Box 2 Amount of data required for comparing meta-analyses

Data required for meta-analysis
The total number of participants in the experimental and control groups
A meta-analysis (point estimate and CI) using the chosen effect measure, method (i.e., Mantel–Haenszel for binary data) and model (fixed or random effects), both graphically (block or diamond) and as text
Heterogeneity statistics (tau-squared, Chi-squared test or the I^2 statistic)

different degrees in different subgroups. Overall effect estimates were compared following our classification irrespective of the total number of participants, association measures within data type (odds or risk ratio for binary outcomes), statistical model or method and heterogeneity. Each discordancy could be classified in more than one category.

Assessment of conclusions

We reviewed all multiple systematic reviews, and recorded all concluding statements in the abstract and the main text addressing the efficacy of the intervention in modifying outcomes. A concluding statement had to explicitly include the intervention and remarks about the causal relationship, with or without the outcome [14]. We excluded comments about implications for practice.

To assess the efficacy of intervention, we assessed each quote as reported in the original paper according to these categories [14]: efficacious (the tone of statements is assertive, implying that the intervention modifies primary outcome/s); detrimental (the tone is negative, control is better); mixed results (the tone is partially positive implying weak efficacy, with a small effect in the primary outcome measure or with some outcomes positive and others not); no effect (no difference between intervention and control); no conclusion quote. Categories were then compared.

Analysis of the reviews included using the Jadad decision algorithm

We used the guide to interpret discordant systematic reviews proposed by Jadad et al. [4] to assess multiple systematic reviews. The algorithm helps readers to evaluate possible sources of discordance among reviews [9]. The algorithm first step involves overlapping of (a) the clinical question (population of patients, interventions, control and outcomes), which characterizes our clusters. The other steps include: (b) primary study selection and inclusion (selection criteria, search strategies), (c) data extraction from the primary studies [methods used to measure outcomes, end points, human error (random or systematic)], (d) assessment of primary study quality (methods used to assess quality, interpretation of quality assessments, methods used to incorporate quality assessments in review), (e) assessment of the ability to combine primary studies (statistical methods, clinical criteria used to judge the ability to combine studies), and (f) statistical methods for data synthesis.

Conclusion

Pre-specified methods are an important requirement for any type of research. The research methods presented here underline a series of studies conducted in different fields of medicine by our group. Their detailed presentation might be useful for other researchers assessing these important and potentially problematic issues, as well as to interested readers. Systematic reviews are used to inform clinical practice and public health policy, and therefore it is important to foster knowledge about potential sources of discordance. Actions to limit the impact of discordant systematic reviews are important to avoid differences in practice being based on unreliable evidence.

As the risk of finding conflicting information in the scientific literature is increasing, and those who produce practice guidelines rely on systematic reviews [15], it is important to understand how disagreements can modify the transfer of evidence from the literature to recommendations, and influence clinical and policy deliberations. A systematic and explicit approach to explore the process through which recommendations are produced when there is discordant literature might help prevent errors, facilitate critical appraisal of these judgments, assign value to minority opinions and foster communication of this information, providing guides for decision-making for clinicians, researchers, peer reviewers and journal editors. This is particularly important since structured approaches evaluating evidence, such as the GRADE system, are increasingly used.

Conflict of interest None.

References

1. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG (2007) Epidemiology and reporting characteristics of systematic reviews. PLoS Med 4(3):e78
2. Oxman AD, Guyatt GH (1993) The science of reviewing research. Ann NY Acad Sci 703:125–133 discussion 133–124
3. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC (1992) A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA 268(2):240–248
4. Jadad AR, Cook DJ, Browman GP (1997) A guide to interpreting discordant systematic reviews. CMAJ 156(10):1411–1416
5. Clinical Evidence Study design search filters. Available at http://clinicalevidence.bmj.com/x/set/static/ebm/learn/665076.html
6. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC (1994) Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc 1(6):447–458
7. Dickersin K, Scherer R, Lefebvre C (1994) Identifying relevant studies for systematic reviews. BMJ 309(6964):1286–1291
8. Chalmers I (1993) The Cochrane Collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. Ann N Y Acad Sci 703:156–165
9. Poolman RW, Abouali JA, Conter HJ, Bhandari M (2007) Overlapping systematic reviews of anterior cruciate ligament reconstruction comparing hamstring autograft with bone-patellar tendon-bone autograft: why are they different? J Bone Joint Surg Am 89(7):1542–1552
10. Borenstein M, Hedges L, Higgins J, Rothstein H (2009) Introduction to meta-analysis. Wiley, West Sussex
11. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG (2004) Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 291(20):2457–2465
12. Higgins JPT, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. BMJ 327(7414):557–560
13. Yusuf S, Wittes J, Probstfield J, Tyroler HA (1991) Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 266(1):93–98
14. Li LC, Moja L, Romero A, Sayre EC, Grimshaw JM (2009) Nonrandomized quality improvement intervention trials might overstate the strength of causal inference of their findings. J Clin Epidemiol 62(9):959–966
15. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schunemann HJ (2008) GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 336(7650):924–926
16. Bhandari M, Devereaux PJ, Montori V, Cina C, Tandan V, Guyatt GH (2004) Users' guide to the surgical literature: how to use a systematic literature review and meta-analysis. Can J Surg 47(1):60–67
17. Collins JA, Fauser BCJM (2005) Balancing the strengths of systematic and narrative reviews. Hum Reprod Updat 11(2):103–104
18. Halligan S, Altman DG (2007) Evidence-based practice in radiology: steps 3 and 4—appraise and apply systematic reviews and meta-analyses. Radiology 243(1):13–27
19. Mulrow CD (1987) The medical review article: state of the science. Ann Intern Med 106(3):485–488
20. Hillier TLB, Jadad AR (1996) The measurement of clinical pain: an appraisal of published reviews. In: Abstracts of the eighth world congress on pain. The international association for the study of pain press, Seattle, p 304
21. Errami M, Sun Z, Long TC, George AC, Garner HR (2009) Deja vu: a database of highly similar citations in the scientific literature. Nucl Acids Res 37(suppl_1):D921–D924
22. von Elm E, Poglia G, Walder B, Tramer MR (2004) Different patterns of duplicate publication: an analysis of articles used in systematic reviews. JAMA 291(8):974–980
23. International Committee of Medical Journal E (1997) Uniform requirements for manuscripts submitted to biomedical journals. N Engl J Med 336(4):309–316
24. National Institutes of health, National Library of Medicine. Fact sheet: errata, retractions, partial retractions, corrected and republished articles, duplicate publications, comments (including author replies), updates, patient summaries, and republished (reprinted) articles policy for MEDLINE®. Available at http://www.nlm.nih.gov/pubs/factsheets/errata.html