Assessment of clinical trial quality and its impact on meta-analyses

Carlos Rodrigues da Silva Filho[a], Humberto Saconato[b], Lucieni Oliveira Conterno[a], Iara Marques[b] and Álvaro Nagib Atallah[c]

[a]Faculdade Estadual de Medicina de Marília. Marília, SP, Brasil. [b]Universidade Federal do Rio Grande do Norte. Natal, RN, Brasil. [c]Escola Paulista de Medicina. Universidade Federal de São Paulo. São Paulo, SP, Brasil

Keywords
Randomized controlled trials. Quality assurance, healthcare. Questionnaires. Outcome assessment, healthcare. Quality control.

Abstract

Objective
To evaluate whether different quality assessment tools applied to a group of clinical trials could be correlated, and what would be their impact on meta-analysis results.
Methods
Thirty-eight randomized controlled clinical trials were analyzed. These had been selected for a systematic review of the therapeutic efficacy of alpha interferon for treating chronic hepatitis B. The following tools were utilized: Maastricht (M), Delphi (D), Jadad (J) and the Cochrane Collaboration (CC) method (gold standard). The Spearman correlation coefficient was used to compare the results from the three methods. The Kappa test was used to assess the concordance between the reviewers in applying the tools, and the weighted Kappa test was applied to compare the quality ranking determined by the tools. The outcomes assessed in the meta-analyses were clearance of HBV-DNA and HBeAg.
Results
The studies presented regular to low quality. The concordance between reviewers varied according to the instrument utilized: D=0.12; J=0.29; M=0.33; and CC=0.53. The correlation was moderate and homogeneous (D/J=0.51; D/M=0.53; and J/M=0.52). The meta-analysis result relating to HBV-DNA ranged from RR=0.71 (95% CI: 0.66-0.77) to RR=0.67 (95% CI: 0.58-0.79). For HBeAg, the results ranged from RR=0.85 (95% CI: 0.80-0.90) to RR=0.85 (95% CI: 0.77-0.93). These results depended on the quality of the studies included.
Conclusions
The quality assessment tools presented good correlation. In systematic reviews with the same direction of effect, the quality assessment may not significantly change the results. The Cochrane Collaboration method was the most reproducible method and easiest to apply.

INTRODUCTION

The adoption of the paradigm of scientific grounding for medical practice took place slowly and gradually over the course of the twentieth century, with an acceleration towards the end of the century. This gave rise to increasing quantities of articles published in the medical literature. To enable such information to be progressively assimilated, there was intensive production of reviews, so-called narratives, grounded in this concentrated experience.[13]

This process was dependent on the authors' subjectivity in relation to selecting and interpreting the information on related clinical topics, from both published and unpublished data.[15] In most cases, no explicit criteria for critically analyzing the information were utilized. This meant that the conclusions and sug-

Correspondence:
Carlos Rodrigues da Silva Filho
Departamento de Medicina
Faculdade Estadual de Medicina de Marília
Av. Monte Carmelo, 800
17519-030 Marília, SP, Brasil
E-mail: silvacr@famema.br / crodrig2@uottawa.ca

Based on Doctoral thesis presented to Universidade Federal de São Paulo, in 2003.
Collaborative study carried out between the Postgraduate Course on Internal Medicine and Therapeutics, Discipline of Internal Medicine, Faculdade Estadual de Medicina de Marília; Centro Cochrane do Brasil (affiliated to the Cochrane Collaboration), Escola Paulista de Medicina, Universidade Federal de São Paulo; and the Department of Medicine, Universidade Federal do Rio Grande do Norte.
Presented at the 19th and 20th Global Meetings of the International Clinical Epidemiology Network (INCLEN), held respectively in Kunming, China (2003), and in Agra, India (2004); and at the 12th Cochrane Colloquium, in Ottawa, Canada (2004).
Received on 13/12/2004. Reviewed on 11/7/2005. Approved on 14/7/2005.

gestions coming from such reviews were subject to bias and errors from various sources. In extreme cases, this model for medical practice led to loss of effectiveness and efficiency within the healthcare services.[18]

In attempts to weigh up the problem of the quality of the information in relation to the pathways adopted, the application of the epidemiological method to clinical practice has shown the greatest consistency. This method is known as clinical epidemiology.[6] Within this methodological approach, the strongest evidence for a given prophylactic or therapeutic intervention is derived from studies in which several randomized controlled clinical trials are developed with a focus on the same treatment for the same condition. This no longer follows the precepts of a narrative review, but rather those of a systematic review.[18]

Such reviews form a modern method for evaluating a set of simultaneous data,[1] through making a summary of the medical literature (primary research). They make use of explicit and reproducible methods for systematically searching for, critically analyzing and producing syntheses of individual studies in what can now be called a systematic review (secondary research).[4]

With regard to therapeutic intervention, evidence level 1a, i.e. the best evidence that exists in relation to a given therapeutic effect, comes from systematic reviews of randomized controlled studies that may or may not be accompanied by a statistical synthesis called meta-analysis.[1]

Naturally, the quality of individual studies has obvious relevance to systematic reviews and has an influence on the magnitude of the results. If the quality of the primary material is inadequate, this may falsify the conclusions of the review.[11]

Clinical studies may become corrupted accidentally or deliberately, in various ways: in the randomization process, in the masking to the allocated treatment, in the random generation of number sequences, in the analysis, or even when the double-blind type of masking is not implemented.

The need for analysis of the quality of these studies has become obvious. Differences in quality ranking, resulting from the use of different lists of criteria, may cause serious problems in conducting systematic reviews. Thus, empirical evidence is needed for establishing whether the quality assessment methods are valid and reproducible.[20]

The objective of the present study was to compare the results obtained from applying different quality assessment tools to clinical trials included in a systematic review on the efficacy of alpha interferon in the treatment of chronic hepatitis B.

METHODS

An observational and analytical study was carried out within the context of a systematic review on the efficacy of treating chronic hepatitis B using alpha interferon. Search strategies with different sensitivities and specificities were utilized, as suggested by the liver and bile duct research group of the Cochrane Collaboration, including all the clinical trials identified in the MedLine, EmBase and Lilacs databases and in the Library of the Cochrane Collaboration. The period covered was from 1966 to 2001.

The relevant clinical trials were identified in all the databanks cited, by using the specific terms *hepatitis B* and *Interferon* and the following strategy: *randomized controlled trial [publication Type] OR controlled clinical trial [Publication Type] OR randomized controlled trial [Mesh Terms] OR random allocation [Mesh Terms] OR double blind methods [Mesh Terms] or single blind methods [Mesh Terms]*.

In all, 2,838 trials were identified, of which 46 were randomized controlled clinical trials. From these, the trials that referred only to comparisons between different doses of alpha interferon were removed (eight studies). Thus, 38 clinical trials were selected.

Four distinct assessment methods for clinical trials were utilized: the Maastricht, Delphi, Jadad and Cochrane Collaboration methods.

Two pairs of reviewers took part in the study, who worked independently. The Cochrane Collaboration method was applied by two infectologists with training in clinical epidemiology in the same place and year, and it served as a comparison parameter for the other three tools. This method classifies trials as A, B or C, according to whether there is a low, moderate or high chance of bias in the primary studies, respectively. This classification is mainly based on the internal validity of the study, its randomization method and how it bypassed or minimized the bias. Thus, it differs from the other assessment tools in that it is a scrutinizing tool rather than a structured tool.[17]

The Delphi system is so called because the 206 items associated with study quality that were initially listed were reduced to nine by means of the Delphi consensus technique. It seeks to assess three dimensions of the quality of studies: internal validity, external validity and statistical analysis.[19]

The Jadad system consists of three topics that are directly related to reducing bias (centered on the internal validity). All the questions have yes/no choices. There are five points possible for its quality score: three single points for yes responses and two additional points for appropriate methods of randomization and ensuring blindness of allocation.[9]

The Maastricht system consists of 15 main items based on methodological criteria for quality assessment. These are divided into 47 subitems totaling 100 points, and these evaluate three dimensions of the quality of a clinical trial: internal validity, external validity and statistical method. The list gives four response choices and weighting is given to the items that reflect relative importance.[20]

It was decided to use these three methods because they are tools that provide different types of assessment, with very different numbers of items, scoring systems with and without weighting, and sometimes apparently complementary assessments. Thus, it was sought to include assessment of the greatest possible number of characteristics within the methodologies of the clinical trials.

The quality scores for each system were calculated in accordance with their original weighted values (Jadad and Maastricht). For the Delphi list, equal weighting was used for each item.

To make comparisons between the different systems, the quality scores obtained were transformed into percentages, in relation to the maximum score for each system. The percentages obtained for each system were categorized into five classes: from 0 to 20% =1; from 21 to 40% =2; from 41 to 60% =3; from 61 to 80% =4; and from 81 to 100% =5.

The Spearman correlation coefficient was calculated between the quality percentages obtained by the reviewers.

The Kappa test was performed to evaluate the concordance between the reviewers regarding their classifications of the studies into categories, when the utilized the different methods. Subsequently, consensus meetings were held to reach definitive scoring for each randomized controlled clinical trial.

After obtaining consensual scores, the studies were ranked in decreasing order in terms of quality, taking the ranking obtained after applying the Maastricht system as the reference.

The weighted Kappa test was utilized for comparing the quality rankings defined by applying each of the evaluation systems.

To analyze the effect of intervention versus non-intervention, it was determined that the studies that obtained scores of greater than or equal to 50% of the possible score using each tool, for at least two of the quality tools utilized, would be considered to be the best ones. For the clinical epidemiologists, the best studies were the ones classified as A or B.

The statistical package Review Manager 4.1 was utilized for performing the meta-analysis. Since only dichotomous variables were utilized, the relative risks were calculated with a 95% confidence interval.

To assess the estimated magnitude and/or direction of effect resulting from the use of alpha interferon for treating chronic hepatitis B, studies that reported the primary outcomes of clearance of the hepatitis B virus DNA (HBV-DNA) and/or the antigen "e" of the hepatitis B virus (HBeAg) within six months of ending the treatment were included in the meta-analysis.

Given the heterogeneity identified among the clinical trials included, the random effect model was utilized for performing the meta-analysis.

RESULTS

Among the tools utilized, Delphi and Jadad were considered easy to apply. The Maastricht method presented greater problems and required adaptations. Because the Cochrane Collaboration method is unstructured, it depends on previous familiarity with the methods of clinical epidemiology.

Out of the 38 trials selected, only six (15.7%) were published more recently than 1996, the year in which CONSORT[3] was published: a consensus that sought to regulate the undertaking and publication of clinical trials. All these trials were defined as randomized and had in fact undergone some randomization method. Twelve of them (31.5%) gave descriptions of the sequences and methodologies utilized. Only one of the articles, by Tchrveniakova/Radev in 1999, described the use of placebo control (multivitamin), although without describing whether or not the placebo was utilized in a way that was analogous to the treatment.

Only seven of the articles (18.4%) provided a report of how the sample size was calculated.

In 22 of the articles (57.8%), losses and withdrawals from the trial and their causes were identified. Detailed analysis of many trials that did not report

the occurrence of losses found that such losses did in fact occur and sometimes reached a proportion of more than 10%, thereby introducing a potentially significant bias in the conclusions from these trials.

In 23 of the articles (60.5%), the analysis took into account the "intention to treat", i.e. it took into consideration all the patients that actually began the trials in the intervention or control group, and not just those who completed the study period.

In 27 of the articles (71%), the losses reported were less than or equal to 10% of the randomized patients, a quantity that is generally considered acceptable for avoiding erroneous conclusions.

Table 1 presents quality scores for the tools utilized, in decreasing order after reaching a consensus between the reviewers.

Table 2 shows the concordance coefficient (Kappa) relating to the application of the tools.

The Spearman correlation coefficient between the tools showed that the correlation obtained between the Maastricht and Jadad methods was 0.52 ($p<0.0005$), between Maastricht and Delphi was 0.53 ($p<0.0007$), and between Delphi and Jadad was 0.51 ($p=0.0001$). Even though these were only moderate correlations, they were very homogeneous and much greater that would be expected just by chance, and thus statistically significant.

Comparison of the trial quality rankings obtained through applying structured systems gave combined concordance of 0.16, using the weighted Kappa test. This was considered to be statistically significant ($p=0.02$).

Figure 1 shows the meta-analysis of the effect of intervention versus non-intervention, in relation to the outcome of clearance of HBV-DNA, including all the studies, regardless of quality assessment. Figure 2 presents the grouped analysis with the inclusion only of the best-quality trials. The relative risk (RR) and 95% confidence interval (95% CI) when only

Table 1 - Quality ranking of the trials on the treatment of chronic hepatitis B using alpha interferon, after reaching a consensus, laid out in decreasing order.

N	Author*	Year	Maastricht	(rank)	Delphi	(rank)	Jadad	(rank)	CE (rank)
1	Sokal / Conjeevaran	1998	64	(1)	5	(1)	3	(1)	A (1)
2	Jansen / Gerken	1999	62	(2)	4	(2)	3	(1)	A (1)
3	Brook / Macdonald	1989	60	(3)	3	(3)	3	(1)	A (1)
4	Fattovich / Farci	1992	58	(4)	4	(2)	3	(2)	B (2)
5	Porres / Carreño	1988	57	(5)	4	(2)	2	(2)	B (2)
6	Di-Biscegli / Fong	1993	56	(6)	4	(2)	3	(2)	B (2)
7	Hoofnagle / Peters	1988	56	(6)	4	(2)	3	(2)	B (2)
8	Sarin / Guptan	1996	56	(6)	4	(2)	3	(1)	A (1)
9	Lok / Wu	1988	55	(7)	4	(2)	3	(3)	C (3)
10	Pastore / Santantonio	1988	55	(7)	4	(2)	2	(2)	B (2)
11	Saracco / Mazzela	1989	55	(7)	4	(2)	3	(2)	B (2)
12	Carreño / Marcelin	1999	54	(8)	4	(2)	2	(2)	B (2)
13	Muller / Baungarten	1990	54	(8)	3	(3)	2	(2)	B (2)
14	Ruiz-Moreno / Rua	1991	54	(8)	3	(3)	2	(2)	B (2)
15	Willian / Graig	1990	54	(8)	4	(2)	2	(3)	C (3)
16	Fattovich / Brollo	1989	52	(9)	3	(3)	3	(1)	A (1)
17	Ruiz-Moreno / Jimen	1990	52	(9)	3	(3)	3	(2)	B (2)
18	Wong / Yim	1995	52	(9)	4	(2)	2	(1)	A (1)
19	Carreño / Porres	1987	51	(10)	3	(3)	3	(2)	B (2)
20	Carreño / Porres	1991	51	(10)	3	(3)	3	(3)	B (2)
21	Pastore / Millella	1992	51	(10)	4	(2)	2	(2)	B (2)
22	Realdi / Fattovich	1990	51	(10)	2	(4)	4	(2)	B (2)
23	Alexander / Brahn	1986	50	(11)	3	(3)	3	(2)	C (3)
24	Barbera / Bortolotti	1994	50	(11)	3	(3)	3	(2)	A (1)
25	Brunetto / Olivieri	1989	50	(11)	3	(3)	3	(2)	A (1)
26	Dusheiko / Paterson	1986	50	(11)	3	(3)	3	(1)	C (3)
27	Perrilo / Schiff	1990	50	(11)	4	(2)	2	(1)	A (1)
28	Brook / Chan / Yap	1989	48	(12)	5	(1)	1	(2)	C (3)
29	Rumi / Romeu	1993	47	(13)	3	(3)	3	(2)	B (2)
30	Tchrveniakova / Radev	1999	47	(13)	3	(3)	3	(3)	C (3)
31	Utilli / Sagnelli	1991	45	(14)	2	(4)	4	(3)	C (3)
32	Thomas / Lok	1994	44	(15)	4	(2)	2	(2)	C (3)
33	Lampertico / Delninn	1997	43	(16)	4	(2)	2	(1)	C (3)
34	Waked / Amin	1990	39	(17)	3	(3)	3	(2)	C (3)
35	Hadzyianes / Bramou	1990	36	(18)	3	(3)	3	(2)	C (3)
36	Mazzela / Villanova	1988	34	(19)	3	(3)	3	(2)	C (3)
37	Lok / Lai	1986	32	(20)	2	(4)	4	(2)	C (3)
38	Barbara / Mazzela	1986	30	(21)	2	(4)	4	(2)	C (3)
	Minimum / Maximum		30/	64	2/	5	1/	3	
	Mean score (SD)		50.13	(7.7)	3.42	(0.75)	2.28	(0.61)	

*The list of articles analyzed can be obtained upon request, from the first author of the present article
SD: standard deviation; CE: clinical epidemiologists

the worst trials (n=25) were included were RR=0.73 and 95% CI: 0.68-0.78.

The meta-analysis performed utilizing the tool suggested by the Cochrane Collaboration, on the trials that the infectologists considered to be in the A and B categories, found RR=0.71 and 95% CI: 0.65-0.78, which was little different from the previous findings.

Figure 3 presents the effect of the treatment in relation to the clearance of HBeAg with all the trials included in the meta-analysis (n=23). When only the best ones were included (n=11), RR=0.85 and 95% CI: 0.77-0.93 were obtained, which thus only differed in relation to the confidence interval.

DISCUSSION

The best articles were those that reported on the randomization procedures, sample size calculation, losses and their causes, and analysis of intention to treat. They were especially the articles that very uniformly reported losses of less than 10% of the total number of randomized patients, thereby decreasing the chances of bias in the analysis and erroneous conclusions.

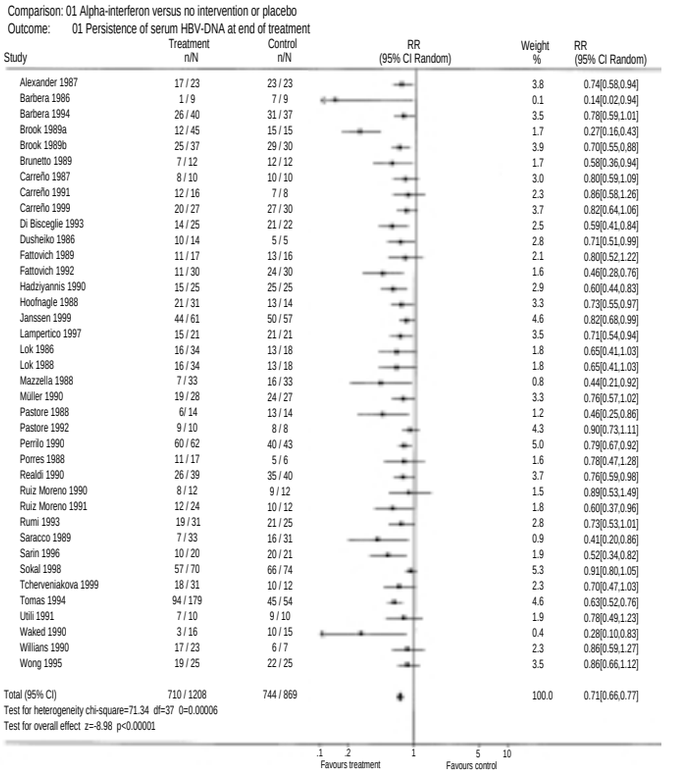

Figure 1 - Clearance of HBV-DNA at the end of the treatment using alpha interferon, including all the trials (N=38).

Naturally, the worst trials had these characteristics at much lower percentages, or did not present them.

Similar moderate statistically significant correlations of between 0.51 and 0.53 were observed between the tools utilized for assessing the quality of the studies. This was doubtlessly consequential to the presence of a common core of items that sought to analyze the internal validity of the studies. All three of the lists contained items that verified the randomization procedures and the presence/absence of secrecy of allocation, for example.

In comparing the reproducibility of the application of these tools between the reviewers, as measured using Kappa, statistical significance was not reached for the Delphi system, even though the reproducibility was greater than expected. In a way, this was an unexpected result, since this is a system that considers the assessment of the dimensions involved in the quality of the clinical trials, and it is easily understood and fast in its application. Three of its items (the ones relating to the presence or absence of placebo control) simply were not applicable, given that only one of the studies incorporated in the review included placebo control. This in itself would have decreased the chance of discordance between the reviewers.

The numbers show that, between the epidemiologists and the method adopted by them for evaluating the quality of the studies, the concordance coefficient (Kappa) was moderate (0.53) and statistically significant (p=0.0001).

The difference observed between the reviewers using the quality assessment tools and the epidemiologists using the method suggested by the Cochrane Collaboration can be explained by the lower number of items and practicality of the latter. Another explanation might

Table 2 - Concordance coefficient (Kappa) for the trials on the treatment of chronic hepatitis B using alpha interferon.

Method	Observed concordance	Expected concordance	Kappa	Standard deviation	p
Delphi	52.63	45.98	0.12	0.12	0.15
Jadad	55.26	36.91	0.29	0.10	0.002
Maastricht	78.95	68.35	0.33	0.12	0.004
Cochrane	68.42	31.72	0.53	0.10	0.0001

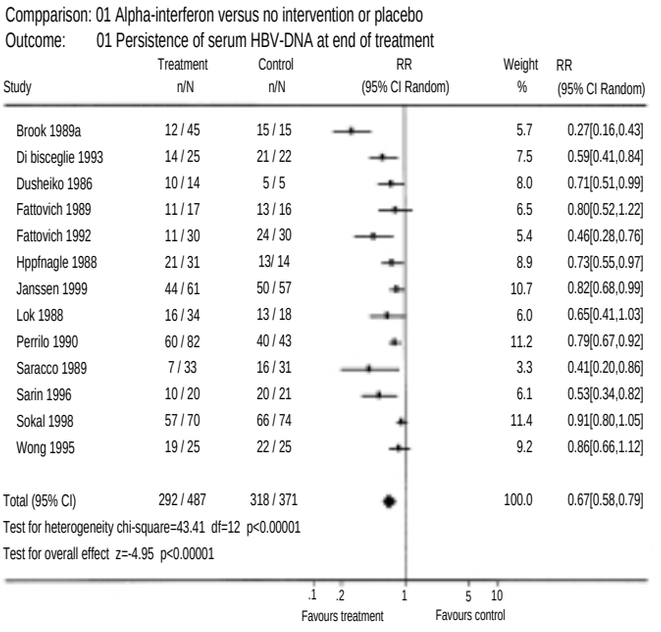

Figure 2 - Clearance of HBV-DNA at the end of the treatment using alpha interferon, including the trials that obtained scores of ≥50% from the quality assessment tools (N=13).

that, in the latter, the score assessment is given on the basis of whether a characteristic was reported (how the patients were allocated, for example), rather than whether this was done adequately in the study.[17]

The present study has shown that, depending on the type of scale utilized for evaluating the quality of the publications, different quality ranks will be defined among the studies incorporated into the systematic review. This has already been observed by other investigators.[10]

The statistically significant result from the weighted Kappa test (p=0.02) allows the conclusion that, despite the low concordance obtained (Kappa=0.16), the quality ranking for the studies was similar.

When all the studies were included and the effect on the clearance of the HBV-DNA was measured, the value obtained was RR=0.71; 95% CI: 0.66-0.77. When only the studies considered to be the best ones were included, the value obtained was RR=0.67; 95% CI: 0.58-0.79, and thus a difference in risk of 33%. When only the studies that did not reach the quality threshold were tested, the RR was 0.73 (95% CI: 0.68-0.76), which was very close to the preceding values and in the direction of lower effect from the intervention (27%).

The literature regularly indicates an overestimate of the effect when lesser studies are introduced into the meta-analysis, because these are usually developed with lower methodological rigor.[4]

The same would occur if studies with lower quality scores were to be introduced.[12] On the other hand, when it was sought to compare the effect of the treatment between the low and high-quality studies, no correlation was found between the quality measurement and the magnitude of the effect of the treatment, and in particular in relation to the variables of utilization of double-blind procedure and secrecy of allocation. However, the possibility that certain quality measurements might be associated with the effect of the treatment cannot be discarded, with regard to clinical situations or questions of specific interest.[2]

The result from the meta-analysis utilizing the best articles (A+B) defined by the epidemiologists while measuring the same outcome (RR=0.72; 95% CI: 0.64-0.80) was very similar to what was obtained by means of scrutiny using the other three quality assessment tools. This reinforces the efficacy of the method proposed by the Cochrane Collaboration for assessing the quality of clinical studies.

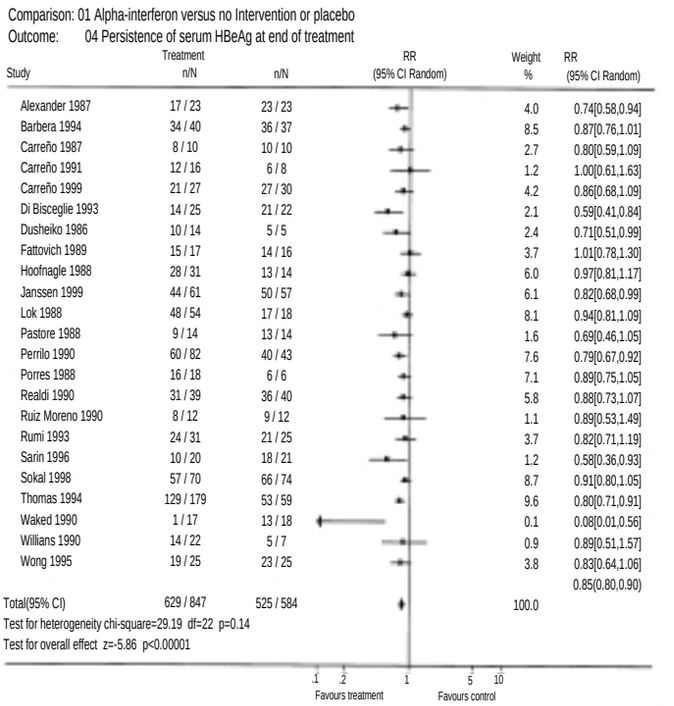

Figure 3 - Clearance of HBeAg at the end of the treatment using alpha interferon, including all the studies that assessed this outcome (N=23).

From analysis of the magnitude of the effect in relation to the second outcome examined (clearance of HBeAg), the results were found to differ only in relation to the confidence interval. The magnitude of the effect measured was smaller than for the previous outcome.

All the articles analyzed determined an outcome favoring intervention. Thus, comparison of the best versus the worst did not alter the direction of the effect, which was maintained with different magnitudes but always in favor of intervention.

A hypothesis can be put forward that, in performing a systematic review and meta-analysis using studies that are discordant in relation to the direction of the effect, there will be greater likelihood of observing variation in the effect of the intervention, depending on the quality of the studies included.

The inclusion of articles of both good and bad quality in the systematic reviews may increase their heterogeneity (the extent to which the results from "similar" randomized controlled clinical trials diverge from each other). This may not necessarily be a bad thing, given that heterogeneity is a reality in clinical practice.

This point has been raised previously, with emphasis on the fact that if all randomized controlled clinical trials were filled with patients of uniform age, sex, location and therapeutic intervention, the estimates from grouped analyses ought to be very precise, but their conclusions would only apply to similar groups and locations. Their external validity would be reduced.[16]

In developing secondary studies, such as systematic reviews, and assessing the quality of primary studies, certain characteristics that produce consequences for modern reviews must be taken into account. One of these is that the results from older clinical trials may be less reliable, since they were conducted at a time when the methodology was less rigorous and reproducible. Therefore, they may not be comparable with more modern studies, with regard to outcomes and how to measure them, for example.[7]

On the other hand, among studies that offer an inadequate description of the tests performed or the population studied, even if it is plausible to believe that this shows design deficiencies and hides bias, it is not always what has in fact occurred. It is often impossible to distinguish between poor design quality and research conduct and poor quality of the description. This latter may have occurred, for example, because of the meager space available for publication in top journals.[4]

The procedures required for performing such trials demonstrate that it is not simply the quality assessment for the randomized controlled clinical trials that should be included in a systematic review and meta-analysis. Unless a consensus for defining the quality measurement is robustly constructed and validated, the concordance when applying it may be found to be unacceptably low.

It should be considered that trained epidemiologists making use of simpler methodology from the Cochrane Collaboration, for classifying the quality of the primary studies, arrived at practically the same result.

It is clear that, until there is a consensus regarding the development and publication of randomized and controlled clinical trials that is adopted without restriction,[3] and also regarding how to report them, assessments of their methodological quality must be considered to be a routine procedure when performing systematic reviews and meta-analyses.[14]

Assessment of the quality of randomized clinical trials offers an estimate of the likelihood that their results will express the truth. This does not rule out the need to perform procedures for meticulous individual evaluation, especially in relation to the generation of randomization and conducting of allocation procedures for the intervention, the secrecy of the distribution and assessment of the effect of the intervention, and also the degree of adherence to the treatment and how losses are handled in the analysis.

Researchers of little ethical commitment or who lack capability, and even those in the pharmaceutical industry, may cause their randomized controlled clinical trials to reach the maximum scores possible by utilizing a wide variety of quality scales. They may in this way defraud premises, objectives, analyses and the most appropriate moment for disseminating the information necessary for good clinical practice.[8]

Perhaps the best way of overcoming the problem will be through increasingly explicit statement of potential conflicts of interest in the development and publication of studies. It will furthermore come from progressive public support for well-trained researchers who do not have links to objectives other than seeking the wellbeing of the majority of the population, and who patiently develop detailed and discerning analysis of the design and conduct of clinical trials, in order to build and strengthen Evidence-Based Medicine.

REFERENCES

1. Atallah AN. Revisão sistemática da literatura médica e metanálise. *Diagn Tratamento* 1997;2(2):12-5.

2. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JPA, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* 2002; 287(22):2973-82.

3. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996;276(8):637-9.

4. Berlin JA, Rennie D. Measuring the quality of trials: the quality of quality scales. *JAMA* 1999;282(11):1083-5.

5. Chalmers I, Haynes B. Reporting, updating, and correcting systematic reviews of the effects of health care. *BMJ* 1994;309(6958):862-5.

6. Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology: the essentials. 3rd ed. Baltimore, MD: Willians & Wilkins; 1996.

7. Hopayian K. The need for caution in interpreting high quality systematic reviews. *BMJ* 2001;323(7314):681-4.

8. Ioannidis JPA, Lau J. Can quality of clinical trials and meta-analyses be quantified? *Lancet* 1998(9128);352:590-1.

9. 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):1-12.

10. Jüni 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.

11. Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001;323(7303):42-6.

12. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analysis. *Ann Intern Med* 2001;135(11):982-9.

13. Leite AJM. Medicina baseada em evidências: um exemplo no campo da pediatria. *J Pediatr* (Rio de J) 1999;75(4):215-26.

14. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Ann Intern Med* 2001;134(8):657-62.

15. Mulrow CD. The medical review article: state of science. *Ann Intern Med* 1987;106(3):485-8.

16. Naylor CD. Two cheers for meta-analysis: problems and opportunities in aggregating results of clinical trials. *CMAJ* 1988;138(10):891-5.

17. Oxman A, Clarke M, editors. Cochrane reviewers' handbook 4.1.1: updated december 2000. In: Assessment of study quality. Oxford: The Cochrane Library; 2001. p. 39-50.

18. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: how to practice and teach EBM. 2nd ed. London: Churchill Livingstone; 2000.

19. Verhagen AP, Vet HC, Bie RA, Kessels AG, Boers M, Bouter LM, et al. The Delphi list: a criteria list for quality assessment of randomized clinical for conducting sistematic reviews developed by Delphi consensus. *J Clin Epidemiol* 1998;51(12):1235-41.

20. Verhagen AP. Quality assessment of randomized clinical trials. In: Quality assessment of trials. Maastricht: Database; 1999. p. 9-11.

21. Verhagen AP, Bie RA, Lenssen AF, Vet HC, Kessels AGH, Boers M, et al. Quality assessment of trial: a comparison of three criteria lists. *Phys Ther Rev* 2000;5:49-58.