ORIGINAL PAPER

V. Lent · A. Langenbach

# A retrospective quality analysis of 102 randomized trials in four leading urological journals from 1984–1989

Received: 21 August 1995 / Accepted: 7 November 1995

**Abstract** The objective of the present study was to analyse critically the quality of the reporting in 102 randomized trials from four leading urological journals from 1984 to 1989 on the basis of an evaluation system we have developed. This comprises 21 principal parameters selected in terms of their significance for the validity of the studies. These parameters were evaluated by two readers independently of each other as to whether they were specified, not specified, could not be evaluated or were not applicable. The study score of each paper resulted from the sum of all specified criteria. In the 102 studies, out of 21 criteria 69.1% and 69.8% (investigators A and B, respectively) were reported; 29.8% and 29.4%, respectively were not reported, 0.4% and 0.1%, respectively, could not be evaluated and 0.7% did not apply. Such important principal parameters as the sample size (6.9% and 7.8%, respectively), statistical power (11.8%), method of randomization (22.5% and 23.5%, respectively), patient blinding (30.4%), investigator blinding (33.3%), loss to follow-up (34.8% and 35.3%, respectively) and rate of discontinuation (36.0% and 37.7%, respectively) were mentioned least often. The study score of all investigations ranged from 20.5 (97.6%) to 9.0 (42.9%) points. Most (60/59% and 62/61%, respectively) attained values between 16 (76.2%) and 13 (61.9%). Accordingly, randomized trials in urological journals show similar deficits to those in internal medicine, surgery and intensive care medicine. A particular problem is that they concern the most important techniques for systematic reduction of inadvertent errors (bias), and thus doubt is cast upon the hardcore of controlled studies. If it is possible for many authors to mention individual criteria completely, this should also apply (and in particular) to the most critical parameters. In our opinion, the 21 criteria selected for an evaluation system constitute a practical compromise between the 3 and 38 criteria alternatively suggested by other authors. Moreover, use of a comprehensive check list should be the precondition for acceptance of papers for publication.

**Key words** Randomized trials · Urological journals · Quality assessment · Quality criteria · Reporting check list

Randomized trials provide the most stringent test of unclear differences between alternative therapies. By methodological reduction of inadvertent errors (bias), they contribute to the clarification of these differences. However, they are only of maximum use when planning, implementation and presentation are optimal.

In fundamental investigations, quality criteria for randomized studies were suggested by Mosteller et al., 1980 [11] and Chalmers et al., 1981 [3]. In 1982, DerSimonian et al. [4] reported their method for systematically checking controlled studies. Of 11 selected criteria in 67 clinical studies from four medical journals, they found that 56% were clear, 10% were unclear and 34% were not mentioned at all. In 1984, Emerson et al. [5] arrived at similar results in general medicine investigations, as did Kelen et al. [9] in intensive care medicine studies in 1985. In 1988, Rohde [12] again found similar results in general surgery studies. All these findings were based on the 11 criteria used by DerSimonian et al. [4]

These criteria were chosen particularly for their fundamental significance in various fields of medicine [4]. Since then, further study analyses have been published using 3–38 principal parameters [1, 2, 6–8, 10, 13, 14]. They were evaluated as single [3–10] criteria [1, 8, 13], by an internal and external (24 points) validity score [10] or by a systematic checkup of study design/organization, conduction/protocol, analysis and presentation using 26–33 points [2, 6, 7, 14]. There is no doubt that

V. Lent (✉) · A. Langenbach
Department of Urology, St. Nikolaus Stiftshospital,
Academic Teaching Hospital, University of Bonn,
Hindenburgwall 1, D-56626 Andernach, Germany

all these (at least 70 validated) criteria are of significance. However, some seemed essential for a general evaluation. We therefore supplemented the 11 criteria of DerSimonian with 10 further criteria which in our opinion must be considered without fail. We used this 21-point score to check 102 randomized trials which were published in four urological journals from 1984 to 1989. The objective was to establish how precisely these studies are presented in urology and to identify their specific weaknesses.

## Materials and methods

Randomized trials published in four urological journals from 1984 to 1989 were evaluated in a retrospective analysis. The following leading English language and German-language journals were chosen: *Journal of Urology*, *European Urology*, *Aktuelle Urologie* and *Urologe A*. Only original papers which fulfilled the conditions of a randomized study were included. All other papers were excluded.

Twenty-one criteria were defined and listed in a study protocol for evaluation of the studies (Table 1). According to this list, all papers were evaluated to establish whether the individual criteria were specified or not specified, were not applicable or could not be evaluated. Only what had been described with sufficient clarity was considered to have been reported. A criterion was regarded as not specified when sufficient pertinent information was not provided. Descriptions which were not categorized because they were unclear could not be evaluated. Criteria which were not capable of being assessed a priori, e.g. patient and/or investigator blinding in surgical operations, remained not applicable. From the sum of the specified criteria, each study was given an overall evaluation as the "study score".

The studies were evaluated by a qualified woman doctor and a university lecturer, who determined the results completely independently of each other. After a training phase in 20 studies not included in this study, problems which had arisen were cleared up and binding guidelines were laid down. After the separate evaluation of all studies, the results were exchanged and misunderstandings were cleared up in 12 arbitration sessions. Different categorizations were only corrected when there was agreement that they were based on a wrong appraisal. Otherwise, they remained unchanged.

**Table 1** Criteria tested in the present analysis

| |
|---|
| Study goals |
| End point definition |
| Sample size calculation |
| Diagnostic methods |
| Therapeutic regimen |
| Inclusion criteria |
| Exclusion criteria |
| Follow-up schedule |
| Allocation before selection |
| Method of randomization |
| Patient blinding |
| Investigator blinding |
| Patient data |
| Side effects |
| Withdrawals |
| Loss to follow-up |
| Statistical analysis |
| Statistical methods |
| Statistical power |
| Conclusion justification |
| Results presentation |

The studies were analysed in two directions. In the horizontal analysis, it was checked how often the individual criteria were reported, not reported, could not be evaluated or were not applicable. In the vertical analysis, the study score of each individual study was determined. If a criterion was not applicable, the highest score was reduced by one point.

In respect of the evaluation criterion "rate of discontinuation", quarter points for the description of the withdrawal and the pertinent reasons as well as the description of a dropout and the pertinent reasons were used for better differentiation. Similarly, the criterion "lost to follow-up" was subdivided into half a point each for the rate of loss during the follow-up observation and the specification of reasons for this.

## Results

In the four urological journals from 1985 to 1989, only 102 (5%) out of 2051 original papers met the criteria which had been laid down: 78 (5.7%) out of 1368 papers in the *Journal of Urology*, 15 (4.6%) out of 323 papers in *European Urology*, 6 (2.6%) out of 231 in *Aktuelle Urologie* and 3 (2.3%) out of 126 papers in *Urologe A*.

As the result of horizontal analysis of all 102 studies, study goals, end point definition, diagnostic methods, therapeutic regimen, inclusion criteria, patient data, justification of conclusion and results presentation were almost always stated (98%–100%). Follow-up schedule, allocation before randomization, complications, side effects and statistical analysis were also frequently mentioned (81%–94%). The statistical method, exclusion criteria, rate of discontinuation, loss to follow-up, patient and investigator blinding and method of randomization were specified in 22.5%–72% of studies. On the other hand, statistical power (11.8%) and sample size (6.9% and 7.8%, respectively) were least often specified.

Very few criteria in individual studies proved to be not available (could not be evaluated) or were not applicable. This mainly involved patient or investigator blinding in the case of "not applicable".

Comparing the results from the individual journals, a large degree of agreement but with some differences was found. Of the studies included in this analysis, the exclusion criteria were not present in 67% of those in *Journal of Urology*, in 60% of those in *European Urology*, in 33% of those in *Aktuelle Urologie* and not at all in *Urologe A*. Data on patient and investigator blinding were found in 36% and 40% of studies, respectively, in *Journal of Urology*, in 13% of those in *European Urology*, in 17% of those in *Aktuelle Urologie* and in none in *Urologe A*. Losses to follow-up were included in 34% of studies in *Journal of Urology*, in 50% of those in *European Urology*, in 25% of those in *Aktuelle Urologie* and in 0% of those in *Urologe A*. Statistical analysis and the statistical methods were described in 94% and 80%, respectively, of studies in *Journal of Urology*, in 60% and 53%, respectively, of those in *European Urology*, in 66.7% of those in *Urologe A* and none in *Aktuelle Urologie*. However, these data are not significant because of the different and in some cases very small number of studies in the individual journals.

The results of the vertical analysis of all 102 studies and their differentiation according to the individual journals showed that none of the studies fulfilled all of the criteria stipulated. One study each was given the highest score of 20.5 and 19.5 (97.6% and 92.9%, respectively) by both investigators. Most studies [60, 62] scored between 16 and 13 points (76.2–61.9%). Four studies each showed a score of 10 (47.6%) and one only 9 (42.9%) points. Comparing the individual journals, the highest scores for *Journal of Urology* were 20.5 and 19.5 points, respectively, *European Urology* 17.5 points, *Urologe A* 13.5 and *Aktuelle Urologie* 13 points. Most studies attained a score between 15.5 and 13 points in *Journal of Urology* and between 13 and 14 points in *European Urology*. The papers with the best appraisal were only accorded scores of between 13.5 and 13 points in *Urologe A* and *Aktuelle Urologie*. On the other hand, the lowest scores were between 9 and 11 points in all studies.

Of the 2142 appraisals which were possible from 102 articles and using the 21 criteria, initially 169.5 (7.9%) differed between the two investigators. This applied to rate of discontinuation, 31.5 times (30.9%), loss to follow-up, 22 times (21.6%), exclusion criteria, 16 times (15.7%); complications and side effects or allocation before randomization, 14 times each (13.7%); and method of randomization or statistical power, 11 times (10.8%) each.

In the various arbitration sessions, 150.5 (7%) of the divergent evaluations were clarified. The experts were unable to reach final agreement on 19 (0.9%) points. These concerned exclusion criteria and allocation before randomization 5 times, rate of discontinuation twice and complications and side effects once each, inclusion criteria, sample size, method of randomization, patient blinding, follow-up schedule and loss to follow-up once each. These discrepancies were insignificant for the overall appraisal, so that they were not elucidated.

## Discussion

Randomized trials do not of themselves ensure validity. Preconditions for reliability are optimal use of all the necessary criteria in their planning, conduct and presentation. It is an open question as to what criteria must be met under all circumstances in the checking of studies.

Between 0 and 10 categories were tested with a score of 3–38 criteria in the study by DerSimonian or according to the suggestions of other authors in earlier analyses [1, 2, 4–10, 12–14]. These add up to at least 70 criteria which are distributed almost evenly in five categories (design, conduct, results, analysis, presentation). However, each of these criteria has its specific significance, and every study can be checked only with difficulty for all criteria.

These earlier analyses reveal specific priorities, which were subsumed in accordance with their general importance in a score of our own which comprises 21 criteria. In our opinion, this is the necessary minimum for the appraisal of studies. On the basis of this score, we found that 69.1% and 69.8%, respectively, of the criteria were met in the analysis of 102 randomized studies in four urological journals, whereas 29.8% and 29.4%, respectively, were not. This corresponds to the results of other authors with in some cases the same principal parameters (DerSimonian 56% + 10% and 34%, respectively, Emerson 59% + 5% and 36%, respectively) [4, 5] and also with different criteria [1, 2, 6–10, 12–14].

Compared with other specialties, randomized studies in urological journals show deficiencies similar to those in internal medicine, general surgery and intensive care medicine [4, 5, 9, 12]. Here, the deficits of the criteria on which the results of the studies are most strongly based are particularly problematical. If data on sample size are lacking, it is doubtful whether the study design is appropriate at all. Studies with groups which are too small tend to provide false-negative as well as false-positive results [8, 11]. On the other hand, if data on statistical power are lacking, it is doubtful whether the results are relevant at all. If data on method of randomization and on blinding are also lacking, it is questionable whether it is a randomized study. Finally, if data on loss to follow-up and on withdrawal are lacking, one must also doubt the validity of the results.

As compared to the time periods of previous analyses (before 1984), our results for 1984–1989 indicate that the deficiencies in the presentation of studies have on the whole not decreased substantially. This also corresponds to the results of Solomon and McLeod [13], who were unable to find any significant difference between the standard of studies in 1980 and 1990. Consequently, efforts to bring about improvements have so far been without success.

One limitation of our results might be that we have only evaluated criteria as reported, not reported, not applicable or not available without making quantitative or qualitative differentiations as have some other authors [4–7, 10, 12, 14]. Some points would be more appropriately rated as partial or weak mention. However, on the whole we were not concerned with making a qualitative appraisal of the criteria specified, but with recording the specified criteria quantitatively as indirect evidence for the quality of the studies. Liberati et al. [10] determined on inquiry by telephone that at least some deficiencies arose because the authors forgot to mention them, rather than because the criteria were actually ignored. However, the fact that the opposite is also possible further reduces the relative significance of quantitative recording.

Finally, in particular the choice of criteria determines the feasibility of analysing studies. Out of at least 70 specified criteria, we chose 21 which appear to us to be

**Table 2** Check list of criteria for randomized trials

*A. Design*
Main/major end points
Secondary/minor end points
Retrospective analysis
Sample size
Assessment mechanism
Experimental regimen/therapy
Control regimen/therapy
Inclusion criteria
Exclusion criteria
Measurement criteria
Assessor qualification
Risk factors
Criteria for stopping the trial
Follow-up schedule

*B. Conduct*
Mechanism of selection
Patient awareness
Patient consent
Method of allocation
Time of randomization
Randomization blinding
Method of randomization
Test of validity of randomization
Blinding of patients
Blinding of physician/investigator
Blinding of statistician
Testing of blinding
Testing compliance
Handling of withdrawals

*C. Results*
Population source
Population characteristics
Prerandomization data
Postrandomization data
Quality control of data
Differences of appearance in placebo/treatment
Differences of taste in placebo/treatment
Side effects
Timing of events
Concomitant treatment
Compliance with treatment
Withdrawals after selection
Withdrawals after consent
Withdrawals after allocation
Withdrawals after treatment assignment
Loss to follow-up

*D. Analysis*
Statement of null hypothesis
Statistical analysis of numbers
Statistical analysis of proportions
Confidence interval of treatment effect ($P < 0.05$)
Posterior estimate of power ($P > 0.05$)
Consideration of type II error in negative trials
Values of test statistic and probability
Comparability table
Analysis of imbalance in prerandomization
Analysis of imbalance in randomization
Multivariate analysis
Stratification of risk factors
Life-table/time-series analysis
Regression analysis correlation
Handling of withdrawals
Handling of loss to follow-up

*E. Presentation*
Assessability of raw data
Discernment of raw data
Results concerning major end points
Results concerning minor end points
Discussion of side effects
Consideration of concomitant treatment
Reproducibility of methods
Credibility of results
Justification of conclusions
Accuracy of title, abstract, etc.

essential for the general appraisal of controlled studies. The most frequent weaknesses according to the results presented were detected with this score. However, if we had concentrated only on the critical criteria, other fundamental data from which the general standard of studies can be determined would have been neglected.

It is concluded that randomized studies lacking essential information on basic criteria of study design and organization, conduct and protocol as well as analysis and presentation do not fulfill their objective. However, more effort is now required to overcome the stagnation in the presentation of studies. It is necessary for authors to check a comprehensive list of criteria (as in Table 2) before "starting and landing". Editors should make this a condition for acceptance of studies.

**Dedication** This study is dedicated to Bucknam McPeek, MD, Professor of Anesthesia at Massachusetts General Hospital (Harvard Medical School), in gratitude for his encouragement and suggestions.

## References

1. Altmann DG, Dorè CJ (1990) Randomisation and baseline comparisons in clinical trials. Lancet 335:149
2. Brown CG, Kelen GD, Moser M, Moeschberger ML, Rund DA (1985) Methodology reporting in three acute care journals: replication and reliability. Ann Emerg Med 14:986
3. Chalmers TC, Smith H Jr, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A (1981) A method for assessing the quality of a randomized control trial. Controlled Clin Trials 2:31
4. DerSimonian R, Charette J, McPeek B, Mosteller F (1982) Exporting on methods in clinical trials. N Engl J Med 306:1332
5. Emerson JD, McPeek B, Mosteller F (1984) Reporting clinical trials in general surgical journals. Surgery 95:572
6. Evans M, Pollock AV (1985) A score system for evaluating random control clinical trials of prophylaxis of abdominal surgical wound infection. Br J Surg 72:256
7. Evans M, Pollock AV (1987) The inadequacy of published random control trials of antibacterial prophylaxis in colorectal surgery. Dis Col Rect 30:743
8. Jacobsen BS, Meininger JC (1985) Randomized experiments in nursing: the quality of reporting. Methodology Corner 35:379
9. Kelen GD, Brown CG, Moser M, Ashton J, Rund DA (1985) Reporting methodology protocols in three acute care journals. Ann Emerg Med 9:880
10. Liberati A, Himel HN, Chalmers TC (1986) A quality assessment of randomized control trials of primary treatment of breast cancer. J Clin Oncol 4:942
11. Mosteller F, Gilbert JP, McPeek B (1980) Reporting standards and research strategies for controlled trials. Controlled Clin Trials 1:37
12. Rohde H, Otterbach C, Pütz T, Mizrahi M (1988) Reporting on randomised surgical trials from German as compared with English and American surgeons. Theor Surg 3:118
13. Solomon MJ, McLeod RS (1993) Clinical studies in surgical journals – have we improved? Dis Colon Rectum 36:43
14. Valleron AJ, Grimfeld A (1992) Evaluation of clinical trials of immunomodulators for prevention of recurrent respiratory infections in children. Develop Biol Stand 77:149