Evaluating the Evidence: Statistical Methods in Randomized Controlled Trials in the Urological Literature

Charles D. Scales, Jr.,* Regina D. Norris,† Glenn M. Preminger,† Johannes Vieweg,† Bercedis L. Peterson,† and Philipp Dahm†,‡

From the Division of Urology, Department of Surgery (CDS, RDN, GMP), and Department of Biostatistics and Bioinformatics (BLP), Duke University Medical Center, Durham, North Carolina, and Department of Urology, College of Medicine, University of Florida, Gainesville, Florida (JV, PD)

Purpose: Randomized controlled trials potentially provide the highest level of evidence to inform clinical decision making. Appropriate use of statistical methods is a critical aspect of all clinical research, including randomized controlled trials. We report the first formal evaluation to our knowledge of the statistical methods of randomized controlled trials published in the urological literature in 1996 and 2004.

Materials and Methods: All human subjects randomized controlled trials published in 4 leading urology journals in 1996 and 2004 were identified for formal review. A standardized evaluation form was developed based on the Consolidated Standards of Reporting Trials statement. Each article was evaluated by 2 independent reviewers with formal training in research design and biostatistics who were blinded to study authors and institution. Discrepancies were settled by consensus.

Results: A total of 152 randomized controlled trials were reviewed (65 in 1996, 87 in 2004). The median sample size (IQR) per arm of parallel design randomized controlled trials published in 1996 and 2004 was 36 (11, 96) and 50 (26, 134) study subjects, respectively ($p = 0.157$). Sample size justifications were provided by 19% of studies in 1996 and 47% of studies in 2004 ($p = 0.001$). Of randomized controlled trials 16 (25%) vs 32 (37%) identified a single primary outcome variable ($p = 0.110$). Effect size estimates for primary or secondary outcome variables were provided by 5% vs 13% ($p = 0.090$) and the precision of the effect was detailed by 5% vs 10% of randomized controlled trials ($p = 0.195$).

Conclusions: This formal review suggests that statistical analysis in urological randomized controlled trials has improved. However, considerable deficiencies remain. Ongoing education in applied statistics may further improve urological randomized controlled trial reporting.

Key Words: statistics as topic, randomized controlled trials as topic, urology

Evidence-based medicine has been defined as the "conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients."[1] The central tenet of evidence-based clinical practice is the balanced integration of clinical expertise and judgment, patient and societal values, and the best available evidence.[2] The foundation for evidence-based clinical practice is clearly high quality evidence.

The highest level of evidence for evaluating the efficacy of health care interventions is provided by randomized controlled trials, if well designed and executed. High quality RCTs are characterized by trial design (eg randomization, blinding) as well as analytic methods (eg intent to treat analysis). Reporting of statistical analysis in RCTs is guided by the CONSORT statement which was published in 1996 and updated in 2001.[3,4] Key statistical elements identified by the CONSORT criteria include sample size calculations, intent to treat analysis, reporting of effect size and precision, and addressing the effects of multiple analyses on trial findings. Inadequate use or reporting of these methodological safeguards has been empirically associated with bias.[5–10]

Statistical hypothesis testing is the foundation of modern medical research. However, statistical methods in the medical literature are often suboptimal, undermining the validity of study conclusions.[11,12] A recent assessment suggests that statistical methods in the urological literature are not ideal, although RCTs comprised a small proportion (12%) of the designs in this investigation.[13] RCTs in the urological literature are often underpowered[14] and reporting of methodological criteria are lacking.[15,16] However, no dedicated analysis of the quality of statistical methods and reporting in RCTs in the urological literature has previously been published. Therefore, in a secondary analysis of a previously published assessment of

Submitted for publication February 11, 2008.

* Financial interest and/or other relationship with Tengion, Inc.

† Nothing to disclose.

‡ Correspondence: Department of Urology, University of Florida College of Medicine, Health Science Center, Box 100247, Gainesville, Florida 32610-0247 (telephone: 352-273-6815; FAX: 352-273-8846; e-mail: p.dahm@urology.ufl.edu).

***Editor's Note*: This article is the fifth of 5 published in this issue for which category 1 CME credits can be earned. Instructions for obtaining credits are given with the questions on pages 1578 and 1579.**

DOI:10.1016/j.juro.2008.06.026

RCT reporting we evaluate the statistical methods of RCTs published in the urological literature.[16]

METHODS

Selection of Studies

The selection criteria for the randomized controlled trials comprising the study sample have been previously published.[16] Randomized controlled trials published in The Journal of Urology®, Urology®, European Urology® and BJU International® in 1996 (before release of the CONSORT statement) and 2004 (after publication of the CONSORT statement) were identified using MEDLINE®. We examined only primary publications of RCTs. Secondary analyses and economic assessments were excluded from analysis (see figure).

Statistical Quality Assessment

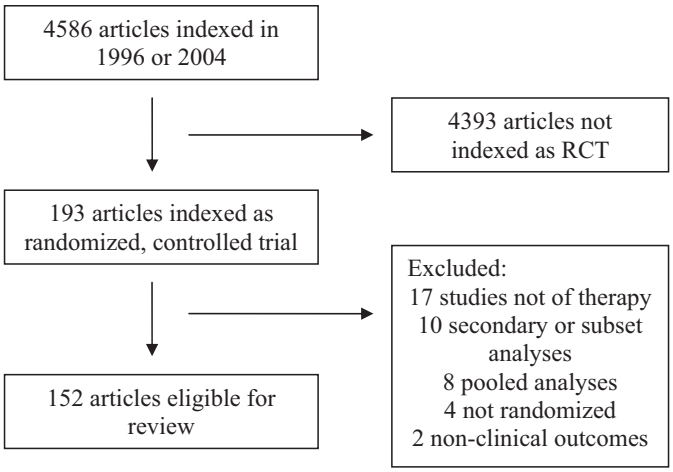
Two investigators (PD and CDS) with formal training in clinical research and statistics independently reviewed each study. The reviewers were blinded to the study authors, institution of origin and funding source by study personnel (RDN) who did not participate in the quality assessment review. The CONSORT criteria informed the development of the statistical evaluation form (see Appendix). In addition, an experienced biostatistician (BLP) assisted in adapting the CONSORT criteria. The criteria were each scored as either met or not met.

Data Collection

We created study databases using double data entry to record the reviewer assessments. Reviewer assessments were merged into a single database and discrepancies were resolved by consensus.

Analysis

The primary objective of this analysis was to compare key elements of statistical analysis reported in studies in 1996 and 2004. The chi-square or Fisher's exact test, as appropriate, was used to test differences between 1996 and 2004 in the proportion of articles that reported key statistical elements. The Mann-Whitney test was used to test continuous outcomes such as differences in RCT sample size between 1996 and 2004. Odds ratios and 95% confidence intervals are presented. Statistical testing was 2-sided with a Type I error threshold (α) of 0.05. Since this study represents a secondary analysis of an existing data set, no formal power calculations were performed. In addition, no adjustments were made for multiple testing. Kappa values were calculated as a measure of interrater agreement between reviewers for select criteria. Analysis was performed using SPSS® version 15.0 software.



Flow diagram of articles identified for review through MEDLINE search of 4 urology journals.

RESULTS

A total of 152 articles comprised the study sample (see figure). Voiding dysfunction (58, 38%) and oncology (40, 26%) were the most common clinical areas reported (table 1). The majority of trials were interventions with either a medication (102, 67%) or a device (29, 19%). Trials of surgical procedures (8, 5%) comprised only a small proportion of the RCTs reported in the study sample. Most studies (141, 93%) used a parallel group design with a median (IQR) sample size per arm of 40 (22, 105).

Reporting of some statistical criteria improved from 1996 to 2004 (table 2). The proportion of trials reporting sample size calculations improved from 19% to 47% (OR 2.36, 95% CI 1.39–4.02, $p < 0.001$). Similarly the proportion of trials that reported an α, the Type I error threshold, improved from 34% to 66% (OR 2.12, 95% CI 1.41–3.17, $p < 0.001$). Median sample size per arm of parallel design trials increased from 36 (11, 96) in 1996 to 50 (26, 134) in 2004, although this change was not statistically significant ($p = 0.157$).

Despite these improvements there was no statistically significant change in the reporting of the number of key statistical CONSORT criteria. Intent to treat analyses were reported in 34% of trials in 1996 but only 29% of trials in 2004 (OR 0.88, 95% CI 0.60–1.28, $p = 0.500$). Effect size reporting (eg OR) increased slightly from 5% in 1996 to 13% in 2004, although this change was not statistically significant (OR 2.10, 95% CI 0.76–5.81, $p = 0.090$). Similarly reporting of the precision of the effect size (eg 95% CI) remained low with 5% of RCTs in 1996 compared to 10% in 2004 (OR 1.77, 95% CI 0.65–4.80, $p = 0.195$). Addressing the effects of multiple testing remained low in both years (OR 0.96, 95% CI 0.45–2.04, $p = 0.916$).

DISCUSSION

To our knowledge this study represents the first dedicated assessment of statistical methods of RCTs in the urological literature. We found that the reporting of statistical elements of the CONSORT criteria improved from 1996 to 2004 but considerable deficiencies remain. These shortcomings pose a potential threat to the validity of study findings and may misguide readers' clinical decision making.

The CONSORT criteria were published in 1996 and revised in 2001 in an effort to improve reporting of RCTs, particularly with respect to items related to the internal and external validity of study findings.[17] The CONSORT statement addresses methodological and statistical elements, the lack of which are associated with exaggerated treatment effects and biased results.[5] Among these, several statistical elements are critical to the validity of study findings and are notably lacking in the urological literature as identified by

TABLE 1. *Study characteristics of 152 urological RCTs by publication year*

	1996		2004		Overall	
Overall No.	65		87		152	
No. study topic (%):						
Oncology	20	(31)	20	(23)	40	(26)
Stones/endourology/ laparoscopy	4	(6)	7	(8)	11	(7)
Trauma/reconstruction	2	(3)	4	(5)	6	(4)
Voiding dysfunction	24	(37)	34	(39)	58	(38)
Infection/inflammation	3	(5)	8	(9)	11	(7)
Infertility/erectile dysfunction	12	(19)	14	(16)	26	(17)
No. randomization (%):						
Drug	43	(66)	59	(68)	102	(67)
Chemotherapeutic agent	5	(8)	4	(5)	9	(6)
Surgical procedure	4	(6)	4	(5)	8	(5)
Device	12	(19)	17	(20)	29	(19)
Intervention	1	(2)	3	(3)	4	(3)
No. 2 or more institutions (%)	16	(25)	34	(39)	50	(33)
No. study population (%):						
Adult	65	(100)	81	(93)	146	(94)
Pediatric	0	(0)	6	(7)	6	(4)
Median overall sample size (IQR)	70	(42, 191)	103	(52, 265)	90	(49, 220)
Median sample size/arm (IQR)	32	(19, 93)	46	(25, 116)	40	(22, 105)
No. study arms (%):						
2	55	(85)	71	(82)	126	(83)
3	7	(11)	11	(13)	18	(12)
4 or More	3	(5)	5	(6)	8	(5)
No. design (%):						
Parallel groups	60	(94)	81	(93)	141	(93)
Crossover	5	(6)	6	(7)	11	(7)
No. placebo controlled (%)	18	(28)	37	(43)	55	(36)
No. funding (%):						
Industry	13	(20)	25	(29)	38	(25)
Institution	0	(0)	6	(7)	6	(4)
Government	7	(11)	9	(10)	16	(11)
No information	45	(69)	47	(54)	92	(61)

the results of the current study. These statistical elements include sample size, intent to treat analysis, reporting of effect size and precision for outcomes, and addressing the effects of multiple testing (eg multiple outcomes or subgroup analyses) on study findings.

Although calculations for trial sample size improved from 1996 to 2004 there was no increase in the median sample size per arm and more than half of RCTs in 2004 failed to report sample size calculations. Inadequate sample size can result in negative study findings (demonstrating no difference) when in fact a clinically meaningful difference exists.[18] This problem has been specifically identified in the urological literature where fewer than a third of negative clinical trials had sufficient power to detect a 25% difference.[14] Inaccurate conclusions from underpowered RCTs inhibit investigation of important clinical questions. Underpowered clinical trials are scientifically unsound, of questionable ethics and represent an inefficient use of health care research resources.[14] Therefore, targeted efforts should be focused on assuring that clinical trials in urology are appropriately powered to demonstrate clinically important differences if they exist.

Intent to treat analysis is another key statistical element identified by the CONSORT criteria. In our investigation less than 1 in 3 RCTs reported an intent to treat analysis, opting instead to report per protocol analyses. Intent to treat analysis protects the integrity of the random allocation and avoids the systematic error that may arise from the nonrandom loss of subjects, a form of selection bias.[19] Further analyses can accompany intent to treat results, which may reveal the effects of subject noncompliance for the intervention, but intent to treat results should always be included in trial reporting.[17] Reports of intent to treat analysis are often inadequate and may not be correctly applied.[9,10] Finally, there is empirical evidence that intent to treat analysis is associated with higher levels of overall methodological quality in clinical trials.[9,10] For these reasons the low use of intent to treat analysis in the urological literature is concerning, and should be a point of emphasis for investigators, reviewers and editors.

Reporting of results in a manner that is informative for the practicing urologist is also emphasized by the CONSORT criteria. Generally the outcome(s) of a study should include a measurement of the contrast between groups (eg intervention vs placebo) which is known as the effect size.[17] Examples of effect size measures include risk ratios or odds ratios for categorical outcomes, or differences between means for continuous outcomes. In addition, a measure of precision of the effect size should accompany the results, typically represented by the CI.[17] Solely reporting p values is less informative for the reader and is discouraged by the CONSORT guidelines. The low frequency (approximately 10%) of effect size and precision reporting in the urological literature indicates a need to raise awareness for this issue among clinical investigators, to facilitate interpretation and application of study findings for an evidence-based clinical practice.

One of the least commonly addressed statistical issues in this study refers to effects of multiple testing on study outcomes. This shortcoming is particularly concerning as multiple analyses (eg subgroup analyses or comparison of several outcomes) create a high risk of false-positive findings.[17] Analyses which are suggested by the data are not as sound as those which are prespecified by the study protocol.[17] Furthermore, it is frequently difficult to determine whether subgroup analyses were prespecified,[12] and there is empirical evidence that outcomes are selectively reported based on a comparison of protocols and published reports in the medical literature.[20] Subgroup analyses are most appropriately used for hypothesis generation but are all too often interpreted as confirmatory. Our findings suggest that the urological literature is susceptible to this

TABLE 2. *Differences in adherence to key statistical CONSORT criteria between 1996 and 2004*

Item Reported	No. (%) 1996	No. (%) 2004	OR (95% CI)	p Value
Sample size calculation	12 (19)	41 (47)	2.36 (1.39–4.02)	<0.001
Identification of α	22 (34)	57 (66)	2.12 (1.41–3.17)	<0.001
Identification of statistical tests	53 (82)	82 (94)	1.80 (1.24–2.61)	0.014
Identification of primary outcome	16 (25)	32 (37)	1.41 (0.90–2.21)	0.110
Intent to treat analysis	22 (34)	25 (29)	0.88 (0.60–1.28)	0.500
Reporting of nonsignificant p values	31 (52)	56 (69)	1.51 (1.04–2.19)	0.035
Reporting of effect size	3 (5)	11 (13)	2.10 (0.76–5.81)	0.090
Reporting of precision of effect size	3 (5)	9 (10)	1.77 (0.65–4.80)	0.195
Addressing multiple testing	4 (6)	5 (6)	0.96 (0.45–2.04)	0.916

danger and efforts to raise awareness of this issue seem indicated.

Our study does have limitations. We recognize that this study represents a secondary analysis of existing data and, therefore, is subject to the potential bias introduced by multiple testing.[16] The testing of multiple CONSORT criteria as outcomes increases the probability of false-positive findings. However, we noted few statistical differences between statistical criteria, and post hoc power calculations suggested that the sample size of this study was sufficient to demonstrate a methodologically important difference in quality of statistical reporting. Therefore, we believe that the secondary nature of this analysis does not draw the central conclusions of this study into question. We further recognize that select RCTs that are of interest to a broader audience are published in nonurological journals and, therefore, were not included in this review. However, the overall numbers of urological trials published in nonurological journals can be expected to be low. The reason to limit our analysis to 4 urological journals was further motivated by the fact that these journals likely represent the main source of primary, peer reviewed research evidence to urologists. Therefore, it appears important to raise awareness for these shortcomings among the readership, reviewers and editors of urological journals. Finally, our analysis is limited to 2 years of publication (1996 and 2004), which may not have been representative. However, there is no reason to suggest that the reporting quality may have been substantially better during other years from this period.

At the same time a number of design features strengthen the validity of our findings. The statistical criteria selected for this study have been empirically associated with bias in clinical research[5–10] and are grounded in the well established CONSORT criteria.[17] An experienced biostatistician assisted in development of the evaluation instrument. Both reviewers have formal training in clinical research methods and statistics. Ratings for key criteria achieved a substantial degree of concordance beyond chance with a kappa value of 0.70 (intent to treat analysis), 0.84 (identification of single primary outcome) and 0.91 (sample size considerations), thereby lending validity to our findings.

CONCLUSIONS

High quality evidence is critical to guide an evidence-based practice of urology. Our findings suggest that reporting of statistical methods in the urological literature has improved since the publication of the CONSORT criteria although substantial room for improvement remains. Authors, reviewers and editors should strive for higher standards of statistical methodology in publications. Efforts to raise the number and quality of RCTs published in the urology literature appear indicated.

ACKNOWLEDGMENTS

Ms. Susan Fesperman performed data entry and provided editorial assistance.

APPENDIX

The CONSORT Criteria (http://www.consort-statement.org)

SECTION Topic	Description
TITLE & ABSTRACT	How participants were allocated to interventions (eg "random allocation," "randomized" or "randomly assigned").
INTRODUCTION	
Background	Scientific background and explanation of rationale.
METHODS	
Participants	Eligibility criteria for participants and the settings and locations where the data were collected.
Interventions	Precise details of the interventions intended for each group and how and when they were actually administered.
Objectives	Specific objectives and hypotheses.
Outcomes	Clearly defined primary and secondary outcome measures and, when applicable, any methods used to enhance the quality of measurements (eg multiple observations, training of assessors).
Sample size	How sample size was determined and, when applicable, explanation of any interim analyses and stopping rules.
Randomization – Sequence generation	Method used to generate the random allocation sequence, including details of any restrictions (eg blocking, stratification).
Randomization – Allocation concealment	Method used to implement the random allocation sequence (eg numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned.
Randomization – Implementation	Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups.
Blinding (masking)	Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. When relevant, how the success of blinding was evaluated.
Statistical methods	Statistical methods used to compare groups for primary outcome(s); methods for additional analyses, such as subgroup analyses and adjusted analyses.
RESULTS	
Participant flow	Flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group report the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analyzed for the primary outcome. Describe protocol deviations from study as planned, together with reasons.
Recruitment	Dates defining the periods of recruitment and followup.
Baseline data	Baseline demographic and clinical characteristics of each group.
Numbers analyzed	Number of participants (denominator) in each group included in each analysis and whether the analysis was by "intention-to-treat." State the results in absolute numbers when feasible (eg 10/20, not 50%).
Outcomes and estimation	For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision (eg 95% confidence interval).
Ancillary analyses	Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those prespecified and those exploratory.
Adverse events	All important adverse events or side effects in each intervention group.
DISCUSSION	
Interpretation	Interpretation of the results, taking into account study hypotheses, sources of potential bias or imprecision and the dangers associated with multiplicity of analyses and outcomes.
Generalizability	Generalizability (external validity) of the trial findings.
Overall evidence	General interpretation of the results in the context of current evidence.

Abbreviations and Acronyms

CONSORT = Consolidated Standards of Reporting Trials
RCT = randomized controlled trial

REFERENCES

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB and Richardson WS: Evidence based medicine: what it is and what it isn't. BMJ 1996; **312:** 71.
2. Scales CD Jr, Preminger GM, Keitz SA and Dahm P: Evidence based clinical practice: a primer for urologists. J Urol 2007; **178:** 775.
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:** 637.
4. Moher D, Schulz KF and Altman DG: The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001; **357:** 1191.
5. Schulz KF, Chalmers I, Hayes RJ and Altman DG: Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995; **273:** 408.
6. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C et al: Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002; **287:** 2973.
7. Hutton JL and Williamson PR: Bias in meta-analysis due to outcome variable selection within studies. Appl Stat 2000; **49:** 359.
8. Gotzsche PC: Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. Control Clin Trials 1989; **10:** 31.
9. Ruiz-Canela M, Martinez-Gonzalez MA and de Irala-Estevez J: Intention to treat analysis is related to methodological quality. BMJ 2000; **320:** 1007.
10. Hollis S and Campbell F: What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 1999; **319:** 670.
11. Hall JC, Mills B, Nguyen H and Hall JL: Methodologic standards in surgical trials. Surgery 1996; **119:** 466.
12. Assmann SF, Pocock SJ, Enos LE and Kasten LE: Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; **355:** 1064.
13. Scales CD Jr, Norris RD, Peterson BL, Preminger GM and Dahm P: Clinical research and statistical methods in the urology literature. J Urol 2005; **174:** 1374.
14. Breau RH, Carnat TA and Gaboury I: Inadequate statistical power of negative clinical trials in urological literature. J Urol 2006; **176:** 263.
15. Welk B, Afshar K and MacNeily AE: Randomized controlled trials in pediatric urology: room for improvement. J Urol 2006; **176:** 306.
16. Scales CD Jr, Norris RD, Keitz SA, Peterson BL, Preminger GM, Vieweg J et al: A critical assessment of the quality of reporting of randomized, controlled trials in the urology literature. J Urol 2007; **177:** 1090.
17. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D et al: The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001; **134:** 663.
18. Moher D, Dulberg CS and Wells GA: Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994; **272:** 122.
19. Lachin JL: Statistical considerations in the intent-to-treat principle. Control Clin Trials 2000; **21:** 526.
20. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC and Altman DG: Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004; **291:** 2457.