Low Methodological and Reporting Quality of Randomized, Controlled Trials of Devices to Treat Urolithiasis

Peter J. Zavitsanos, Vincent G. Bird,* Kathryn A. Mince, Molly M. Neuberger and Philipp Dahm†

From the Department of Urology, University of Florida and Malcom Randall Veterans Affairs Medical Center, Gainesville (PD), Florida

Abbreviations and Acronyms

CENTRAL = Cochrane Central Register of Controlled Trials

MD = mean difference

PCNL = percutaneous nephrolithotomy

RCT = randomized, controlled trial

SWL = shock wave lithotripsy

Accepted for publication October 17, 2013.

* Financial interest and/or other relationship with Boston Scientific.

† Correspondence: Department of Urology, University of Florida College of Medicine, Health Science Center, Box 100247, Room N2-15, Gainesville, Florida 32610-0247 (telephone: 352-273-8634; FAX: 352-273-7515; e-mail: p.dahm@urology.ufl.edu).

Purpose: We assessed the methodological and reporting quality of randomized, controlled trials of stone disease management and determined whether the reporting quality of randomized, controlled trials improved with time.

Materials and Methods: We systematically searched the literature for randomized, controlled trials of urolithiasis treatment. We developed and pilot tested a data extraction checklist based on CONSORT (Consolidated Standards of Reporting Trials) criteria as well as a clinical checklist relevant to urolithiasis, each scored as 0 to 25. Our primary outcome measures were the mean differences in CONSORT and clinical summary scores with time. We performed statistical hypothesis testing using the Student t-test with 2-sided $\alpha = 0.05$ to compare scores between 2002 to 2006 and 2007 to 2011.

Results: A total of 104 randomized, controlled trials met study inclusion criteria. The most common procedure types studied were percutaneous nephrolithotomy (41.3%), ureteral stenting (28.8%) and shock wave lithotripsy (25.0%). Mean $\pm$ SE CONSORT summary scores were 11.4 ± 0.4 and 12.1 ± 0.3 in 2002 to 2006 and 2007 to 2011, respectively, with a mean difference of 0.7 (95% CI −0.3−1.6, $p = 0.167$). Mean clinical summary scores were 7.4 ± 0.5 and 9.3 ± 0.4 in 2002 to 2006 and 2007 to 2011, respectively, with a mean difference of 1.8 (95% CI 0.6−3.1, $p = 0.004$).

Conclusions: While the number of randomized, controlled trials of urological devices used to treat stone disease substantially increased with time, methodological and clinical reporting quality remains suboptimal. This compromises their credibility and warrants efforts to promote appropriate performance of future endourological studies.

Key Words: kidney, ureter, urolithiasis, equipment and supplies, randomized controlled trials as topic

Well designed RCTs have the potential to provide the highest quality evidence for questions of therapeutic effectiveness, assuming that they are appropriately performed and analyzed, and transparently reported. Urologists use RCT results to guide clinical decision making in individuals and they are also being increasingly used in systematic reviews and clinical practice guidelines that define standards of care and shape health policy.[1] In the widely endorsed paradigm of evidence-based clinical practice high quality evidence supporting a given intervention provides a strong impetus for its clinical application while low quality evidence is the

0022-5347/14/1914-0988/0
THE JOURNAL OF UROLOGY®

http://dx.doi.org/10.1016/j.juro.2013.10.067
Vol. 191, 988-993, April 2014
Printed in U.S.A.

source of uncertainty and undesirable practice variation.

Treatment of urinary stones is central to the practice of urology and also notable for its heavy reliance on surgical devices. For various historical and practical reasons these devices have not been assessed by the same evidentiary standards as drugs before regulatory approval and widespread implementation. Although unsystematic observations suggest that the number of RCTs of surgical treatment of urolithiasis has increased, little is known about the quality of these studies.

Therefore, we formally assessed trends in the methodological and reporting quality of RCTs of surgical devices used to treat patients with stone disease over time. We sought to better define evidence gaps and suggest strategies for improvement.

MATERIALS AND METHODS

We defined a surgical device RCT as a prospective study comparing surgical interventions that included a device in at least 1 arm of the trial and had therapeutic intent in human participants randomly allocated to study groups.[2] Surgical procedures and devices included SWL, ureteroscopy, PCNL, lithotripters, lithotrites, endoscopes, ureteral stents and antiretropulsion devices. We systematically searched the literature using a defined search strategy in 2 databases (MEDLINE® and CENTRAL) with date restrictions (2002 to 2011) and publication type (RCT) to identify RCTs potentially eligible for study inclusion. We also assessed individual studies for eligibility that were referenced in systematic reviews and meta-analyses identified in the MEDLINE search. Two investigators (PJZ and KAM) independently screened all search results for eligibility. Consensus was achieved through discussion between the 2 investigators with arbitration by a third investigator (PD).

Two independent investigators with formal methodology training reviewed and scored each included article using a standardized, pilot tested data extraction form incorporating the 2010 CONSORT statement criteria and an evidence-based checklist used to standardize RCT reporting and prevent the introduction of bias into studies (supplementary file 1, http://jurology.com/).[3] We also included 25 clinical variables on the same checklist to evaluate baseline data and end points relevant to the treatment of patients with urinary stones (supplementary file 1, http://jurology.com/). Items were scored as met, not met or nonapplicable. Discrepancies were settled by discussion among the reviewers and in select cases by the third party arbiter.

Our primary end points were the MDs in CONSORT and clinical criteria summary scores, each on a scale of 0 to 25. As an a priori null hypothesis, we considered that the reporting quality of urolithiasis surgical device RCTs published within the 10-year study period in 2007 to 2011 was no different than in 2002 to 2006. Consistent with prior studies we assigned quarter, third and half points for multicomponent criteria to maintain weighting. Thus, if a study only mentioned 2 of 4 subcriteria for 1 of the 25 CONSORT criteria, the study received a half point for that criterion.[4] Clinical scores were calculated based on the reporting of 25 predefined baseline and end point criteria. We assessed interobserver agreement beyond chance using the κ statistic.[5]

Descriptive summary statistics for individual CONSORT and clinical criteria are shown as proportions and summary scores are shown as the mean $\pm$ SE and median. We calculated MDs and the 95% CI of summary scores between periods. We performed predefined subgroup analysis by continent of origin and post hoc subgroup analysis by journal of publication. Statistical hypothesis testing was done with SPSS® 20.0 using the chi-square and Student t-tests with 2-sided $\alpha = 0.05$. We did not adjust for multiple comparisons.

RESULTS

A total of 104 RCTs were included in our study (supplementary file 2, http://jurology.com/ and fig. 1). The MEDLINE search identified 209 records, of which 99 ultimately met inclusion criteria. Another 1 and 4 studies were identified for inclusion through CENTRAL and reference lists of systematic reviews, respectively. Figure 1 shows the reasons for study exclusion.

Table 1 lists the characteristics of included trials. The number of trials meeting inclusion criteria that were published in each 5-year period increased from 39 (2002 to 2006) to 65 (2007 to 2011). The most common types of procedures studied were PCNL, ureteral stenting and SWL. There was a marked increase from 17.9% (2002 to 2006) to 55.4% (2007 to 2011) in the proportion of studies of PCNL. Of the series 86.5% were parallel 1-arm studies. Mean sample size was 106.8 patients (range 18 to 903), including 82.2 in 2002 to 2006 and 122.6 in 2007 to 2011 ($p = 0.545$). The percent of studies with a sample size of greater than 100 patients increased during the periods. Although more than half of the trials did not mention the number of study sites, the number of multicenter trials performed between the periods increased (2.6% to 10.8%). In regard to study origin the absolute number of studies from North America remained similar but the relative percent of North American studies decreased since more studies originated from other continents, most notably Asia.

When comparing the 2 intervals, mean $\pm$ SE CONSORT summary scores were 11.4 ± 0.4 (2002 to 2006) vs 12.1 ± 0.3 (2007 to 2011) with a MD of 0.7 (95% CI -0.3–1.6, $p = 0.167$). Mean $\pm$ SE clinical summary scores were 7.4 ± 0.5 (2002 to 2006) and 9.3 ± 0.4 (2007 to 2011) with a MD of 1.8 (95% CI 0.6–3.1, $p = 0.004$). Median CONSORT summary scores were 12.0 (2002 to 2006) vs 12.1 (2007 to 2011) and median clinical summary scores were 8.0 (2002 to 2006) vs 10.0 (2007 to 2011).