Conflict of Evidence: Resolving Discrepancies When Findings from Randomized Controlled Trials and Meta-analyses Disagree

Richard J. Sylvester a,*, Steven E. Canfield b, Thomas B.L. Lam c, Lorenzo Marconi d, Steven MacLennan e, Yuhong Yuan e, Graeme MacLennan f, John Norrie f, Muhammad Imran Omar c, Harman M. Bruins g, Virginia Hernández h, Karin Plass i, Hendrik Van Poppel j, James N’Dow c

a EAU Guidelines Office, Brussels, Belgium; b Division of Urology, University of Texas McGovern Medical School, Houston, TX, USA; c Academic Urology Unit, University of Aberdeen, Aberdeen, UK; d Department of Urology, Coimbra University Hospital, Coimbra, Portugal; e Department of Medicine, McMaster University, Hamilton, ON, Canada; f Health Services Research Unit, University of Aberdeen, Aberdeen, UK; g Department of Urology, Radboud University Medical Center, Nijmegen, The Netherlands; h Department of Urology, Hospital Universitario Fundacion Alcorcon, Madrid, Spain; i EAU Central Office, Guidelines Office, Arnhem, The Netherlands; j Department of Urology, University Hospital Gasthuisberg, Katholieke Universiteit Leuven, Leuven, Belgium

Article info

Article history:
Accepted November 16, 2016

Associate Editor:
James Catto

Keywords:
Conflict of evidence
Meta-analyses
Randomized controlled trials
Systematic reviews
Treatment guidelines

Abstract

Context: Clinicians and treatment guideline developers are faced with a dilemma when the results of a new, large, well-conducted randomized controlled trial (RCT) are in direct conflict with the results of a previous systematic review (SR) and meta-analysis (MA).

Objective: To explore and discuss possible reasons for disagreement in results from SRs/MAs and RCTs and to provide guidance to clinicians and guideline developers for making well-informed treatment decisions and recommendations in the face of conflicting data.

Evidence acquisition: The advantages and limitations of RCTs and SRs/MAs are reviewed. Two practical examples that have a direct bearing on European Association of Urology guidelines on treatment recommendations are discussed in detail to illustrate the points to be considered when conflicts exist between the results of large RCTs and SRs/MAs.

Evidence synthesis: RCTs are the gold standard for providing evidence of the effectiveness of interventions. However, concerns regarding the internal and external validity of an RCT may limit its applicability to clinical practice. SRs/MAs synthesize all evidence related to a given research question, but two urologic examples show that the validity of the results depends on the quality of the individual studies, the clinical and methodological heterogeneity of the studies, and publication bias.

Conclusions: Although SRs/MAs can provide a higher level of evidence than RCTs, the quality of the evidence from both RCTs and SRs/MAs should be investigated when their results conflict to determine which source provides the better evidence. Guideline developers should have a well-defined and robust process to assess the evidence from MAs and RCTs when such conflicts exist.

Patient summary: We discuss the advantages and limitations of using data from randomized controlled trials and systematic reviews/meta-analyses in informing clinical practice when there are conflicting results. We provide guidance on how such conflicts should be dealt with by guideline organizations.

© 2016 European Association of Urology. Published by Elsevier B.V. All rights reserved.

* Corresponding author. EAU Guidelines Office, 500 avenue Moliere, Brussels 1050, Belgium.
Tel. +322 346 4916.
E-mail address: richard.sylvester@skynet.be (R.J. Sylvester).

http://dx.doi.org/10.1016/j.eururo.2016.11.023
0302-2838/© 2016 European Association of Urology. Published by Elsevier B.V. All rights reserved.
1. Introduction

The practice of evidence-based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research [1].

Treatment recommendations in European Association of Urology (EAU) guidelines are underpinned, whenever possible, by the results of systematic reviews (SRs)/meta-analyses (MAs) and large randomized controlled trials (RCTs). According to the 2009 Oxford Centre for Evidence-based Medicine, SRs of RCTs (with or without a meta-analysis) that are free of worrisome variations (heterogeneity) in results between individual studies provide the highest level of evidence (LE), 1a, whereas individual RCTs with a narrow confidence interval provide the next highest LE, 1b [2]. As SRs can provide a higher LE than RCTs, the results of SRs are generally considered to take precedence when developing treatment recommendations.

The quality of the results of an SR/MA depends on the quality of the studies included. Kjaergard et al [3] found a correlation between methodologic quality and discrepancies in the results of large and small RCTs included in MAs. Intervention effects were exaggerated in small trials with inadequate allocation sequence generation, inadequate allocation concealment, and no double blinding.

Discrepancies have also been noted between large RCTs and previously published MAs on the same subject [4–6]. For 12 large RCTs carried out subsequent to 19 MAs addressing the same question, LeLorier et al [7] found that the results of subsequent RCTs disagreed with those of earlier MAs 35% of the time.

To illustrate these points and provide guidance to guideline developers in dealing with conflicting data from different sources, two examples that have a direct bearing on EAU Guidelines treatment recommendations are presented. In the first example, the EAU Guidelines Office has recently been confronted with the results of a large RCT that found no beneficial effect of medical expulsive therapy (MET) on stone passage, contrary to results of previous meta-analyses that formed the basis for treatment recommendations [8]. In the second example, which compares the efficacy of partial and radical nephrectomy for localized renal tumors, discordance between the results of the meta-analysis and the only available RCT are investigated [9,10].

2. Advantages and limitations of RCTs

As summarized in Table 1, RCTs have a number of advantages and limitations.

### Table 1 – Advantages and limitations of randomized controlled trials (RCTs)

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomization minimizes the influence of both known and unknown prognostic variables on treatment outcome</td>
<td>It may be difficult to recruit and follow up patients</td>
</tr>
<tr>
<td>RCTs can demonstrate causality</td>
<td>Ethical considerations may make randomization difficult</td>
</tr>
<tr>
<td>Patients are treated according to a common protocol</td>
<td>Required study power might not be met</td>
</tr>
<tr>
<td>Quality control of treatment and outcome assessment</td>
<td>Generalizability may be low</td>
</tr>
<tr>
<td>RCTs provide the strongest empirical evidence of treatment efficacy</td>
<td>RCTs are expensive and resource-intensive</td>
</tr>
</tbody>
</table>

2.1. Advantages of RCTs

RCTs are the gold standard for providing evidence on the effectiveness of interventions [11,12]. Randomization balances, on average, the distribution of both known and unknown prognostic factors at baseline in the intervention groups, thereby minimizing selection bias when assigning patients to treatments. Although adjusting for baseline covariates used in the randomization process can improve statistical power, complex adjustment procedures such as propensity score weighting are not usually required when comparing outcomes.

Patients are selected, treated, followed, and assessed according to a common protocol testing a specific hypothesis. Blinding of participants and physicians to the allocated intervention may be possible to minimize performance bias, and is especially important when assessing outcomes [13]. Quality control measures and external review of key parameters maximize study quality.

2.2. Limitations of RCTs

RCTs can be challenging to design (randomization and blinding), conduct (poor recruitment, loss to follow-up), analyze (missing data), and report (patient exclusions).

RCTs require an adequate sample size and follow-up to have sufficient power to detect clinically relevant differences between interventions [14]. In practice, many clinical trials do not meet their prespecified power requirements, so a conclusion of “no significant difference” in outcome should not be interpreted as meaning that two or more treatments are equivalent in effect. Sample size estimation requires data about expected differences and variability of the primary outcome. Often these data are unknown or only available from observational studies prone to bias.

Although analyses using the intention-to-treat principle can provide an unbiased estimate of the treatment effect, this assumes that there are no differences in follow-up or missing outcome data that may bias the treatment comparison [15]. In some RCTs, not all participants receive their randomized intervention; they may, for example, cross over to the other randomized treatment, in which case a per-protocol analysis may also provide useful information. Various analysis strategies exist, depending on whether the objective is to estimate treatment efficacy (the intervention effect under perfect conditions, in which case intent to treat can dilute the size of the treatment effect) or effectiveness (the real-world intervention effect with “imperfect” compliance).
They should only be combining data from two or more separate RCTs asking sources of bias. Studies, draw attention to their differences, and identify an intervention. SRs explore the findings of individual and critical qualitative appraisal of the evidence related to replicable form of literature review that provide a rigorous study inclusion criteria. They are the only transparent and comparator, outcome (PICO)–based protocol outlining the research evidence relevant to that question. SRs and MAs are vital for guideline developers, healthcare providers, patients, researchers, and policymakers in order to guide clinical practice, research, and healthcare policies.

3. Advantages and limitations of SRs and MAs

Table 2 outlines the advantages and limitations of SR/MAs.

3.1. Advantages of SR/MAs

An SR is a literature review focused on a research question that tries to identify, appraise, select, and synthesize all research evidence relevant to that question. SRs are a priori defined in a participant, intervention, comparator, outcome (PICO)–based protocol outlining the study inclusion criteria. They are the only transparent and replicable form of literature review that provide a rigorous and critical qualitative appraisal of the evidence related to an intervention. SRs explore the findings of individual studies, draw attention to their differences, and identify sources of bias.

An MA is a statistical technique for quantitatively combining data from two or more separate RCTs asking the same or a similar question. They should only be performed as part of an SR, otherwise the MA is a combined analysis that is susceptible to study selection bias. Two different types of MAs exist: literature-based or aggregate data MAs, and individual patient data (IPD) MAs.

MAs provide an overall estimate of the size of the treatment effect, giving due weight to the size of the individual RCTs. They are useful when individual studies are underpowered or yield inconclusive or conflicting results, or when an overall, more precise estimate of the size of the treatment effect is required. MAs increase the power to detect moderate but clinically meaningful differences in treatment outcome and assess if the treatment effect is similar across different studies or types of patients. They are useful in exploring the effects of an intervention in subgroups of patients, especially in IPD MAs.

SRs and MAs are vital for guideline developers, healthcare providers, patients, researchers, and policymakers in order to guide clinical practice, research, and healthcare policies.

3.2. Limitations of SR/MAs

The validity of an MA depends on the quality of the SR on which it is based. SRs and MAs have a number of potential limitations, including poor quality of the studies included, heterogeneity, and publication bias.

The literature summary provided in an SR and the results of an MA are only as reliable as the quality of the studies included. Although IPD MAs and multicenter RCTs can be analyzed using the same statistical techniques for clustered data, where the clusters are studies and centers, respectively, there may be important clinical and methodological heterogeneity between the studies in an MA since they are not carried out based on a common protocol. The studies may be heterogeneous regarding patients included, the intervention, or the assessment of treatment outcome. Although heterogeneity in treatment effect can be better investigated in IPD MAs, the primary studies should be similar enough to be combined, otherwise genuine differences in effects may be obscured. Since institutions participating in a multicenter study are supposed to treat, follow up, and assess patients according to a common protocol, there is potentially a greater degree of standardization and higher quality data in multicenter clinical trials compared to studies included in MAs.

If bias is present in the individual studies included in an MA, MAs will compound these errors and produce a biased

---

### Table 2 – Advantages and limitations of systematic reviews and meta-analyses

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focused, well-defined clinical question with a clear objective and explicit, predefined study eligibility criteria</td>
<td>Depends on the quality of the studies included</td>
</tr>
<tr>
<td>Comprehensive literature search strategy to guarantee the identification of all potentially eligible studies</td>
<td>Susceptible to the effects of heterogeneity of the studies included</td>
</tr>
<tr>
<td>Critical appraisal of all the included studies that is used to guide the analysis and conclusions</td>
<td>• Clinical heterogeneity</td>
</tr>
<tr>
<td>Increases the power to detect differences between interventions</td>
<td>– Participants (eg, age, gender, disease severity and subtype, study eligibility criteria)</td>
</tr>
<tr>
<td>Increases the precision of the estimate of the treatment effect</td>
<td>– Interventions (eg, drug doses, duration/intensity of treatment, delivery, co-interventions, surgeon experience)</td>
</tr>
<tr>
<td>Allows the comparison of treatment effects across different studies or subgroups of patients, interventions and outcomes</td>
<td>– Outcomes (eg, definition of outcome, outcomes reported, timing and method of measurements, follow-up duration, cutoff points)</td>
</tr>
<tr>
<td>Time- and resource-consuming</td>
<td>• Methodological heterogeneity (eg, different study designs, reporting bias across studies)</td>
</tr>
<tr>
<td>• Statistical heterogeneity</td>
<td>Publication bias</td>
</tr>
<tr>
<td>• Comprehensive literature search strategy to guarantee the identification of all potentially eligible studies</td>
<td></td>
</tr>
</tbody>
</table>
result. The risk of bias (RoB) for the outcomes in each study should be systematically assessed and sensitivity analyses performed to examine the effect of RoB on the conclusions. Observational and nonrandomized comparative studies in SRs of interventions should not be included in MAs because the MA may provide very precise but spurious results because of confounding and patient selection bias.

Only a nonrandom proportion of research projects ultimately reach publication in an indexed journal and become readily identifiable for SRs. Statistically significant “positive” results favoring an intervention are more likely to be published, published more quickly, and published in higher impact journals, leading to publication bias [26]. When these trials are pooled together in an MA, this may lead to exaggeration of the treatment effect. Begg and Egger tests, along with funnel graphs and plots, can be applied for detection of publication bias, but these have limited power in small MAs, such as those including fewer than ten studies [27]. To minimize publication bias, authors should not only perform a comprehensive systematic literature search looking for published trials in various electronic databases, but should also search trial registries for unpublished studies and conference abstracts or proceedings [18].

4. When the results of an RCT are in conflict with the results of an SR/MA

It is not uncommon for the results of a large RCT to appear to be inconsistent with evidence from SRs/MAs. The most extreme case is when an intervention thought to be beneficial is demonstrated to be harmful in a large RCT [9,10]. More commonly, an RCT may show that a treatment is ineffective, or less effective than found in a previous MA, [9,10]. More commonly, an RCT may show that a treatment is ineffective, or less effective than found in a previous MA, [9,10]. Assuming the conflicting RCT was of high quality, a number of issues should be explored to try to explain the discrepancies.

4.1. Quality of the SR

The starting point is the methodological quality of the SR. Assessment of Multiple Systematic Reviews (AMSTAR) and Documentation and Appraisal Review Tool (DART) checklists [28–30] allow readers to judge a review’s quality by focusing on the essential components of a well-conducted SR. Items include the comprehensiveness of the search strategy, a description of the characteristics of studies included and an assessment of their scientific quality. A poor quality SR/MA may produce biased results that conflict with a large RCT.

4.2. Small study effects and publication bias

Small study effects and publication bias can individually and jointly produce results in an SR/MA that conflict with a large RCT. Studies have shown that small RCTs can exaggerate intervention effects owing to shortcomings in methodological rigor, which may then introduce bias [3]. Small studies that find statistically significant (but unrealistically large) treatment effects are more likely to be published than negative studies, and then included in an SR and MA, leading to publication bias. Both of these phenomena can be investigated using funnel plots [31].

4.3. Heterogeneity

Heterogeneity within an SR/MA can arise from many sources, including the population recruited (age, sex, disease severity, etc), the intervention(s) and control treatments, and the definition and timing of outcome measurements. If studies included in an SR/MA differ substantially from a subsequent large RCT, then judgment is required on whether similar findings should be expected.

Another source of heterogeneity is differences in the methodological quality of the studies included. Deficiencies in the generation and concealment of the allocation sequence, adherence to treatment, handling of missing data, and outcome assessment can all introduce bias in the outcomes reported in the studies included [18]. Bias may then be propagated in MAs through the pooling of biased study effects, thus contributing to different estimates of effectiveness between an SR/MA and subsequent large RCTs. Nevertheless, since an MA is generally seen to have a higher LE than a single RCT, the results of a poor-quality MA may have more impact than a well-conducted RCT.

Heterogeneity should be assessed using both clinical knowledge and statistical methods. If substantial heterogeneity from any source is suspected, random effects models are recommended; however, the pooling of data and estimation of an overall treatment effect may be inappropriate with any statistical model in the presence of heterogeneity. Meta-regression is a useful tool for exploring the relationship between RCT effect sizes and characteristics at a study level [32]; however, IPD MAs are required for assessment at a patient level [21,33]. Appropriate statistical modeling may show that after correcting for sources of bias and heterogeneity, discrepancies between SRs/MAs and definitive RCTs are reduced. Whatever the approach, interpretation of results is less straightforward when heterogeneity is present.

To provide guidance to clinicians and guideline developers when there is a conflict in results between a large RCT and an SR/MA, a practical checklist of points to consider is provided in Table 3.

5. Examples of discrepancies between findings from MAs and large RCTs

5.1. Medical expulsive therapy

Five SRs and MAs on the management of uncomplicated symptomatic ureteric stones using MET were published in the past 10 yr [34–38]. All five suggested that alpha blockers and nifedipine were more effective in increasing spontaneous passage of ureteric stones compared to control (risk ratios ranging from 1.45 to 1.59). The reviews identified numerous sources of potential bias that limited the strength
<table>
<thead>
<tr>
<th>Criteria to consider</th>
<th>Questions to ask</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selection bias</strong></td>
<td>Were the sequence generation and allocation concealment adequate in both the studies included in the SR/MA and the subsequent RCT?</td>
<td>If the sequence generation was not truly random or the allocation was not effectively concealed, this can lead to exaggerated estimates in individual studies, and these may be amplified in MAs.</td>
</tr>
<tr>
<td><strong>Confounding bias</strong></td>
<td>Were the groups balanced for known prognostic factors at baseline and were any imbalances controlled for in the analysis?</td>
<td>Imbalances in known and unknown prognostic factors are possible even in well-designed RCTs. Baseline imbalances may explain differences in estimates of effect if not controlled for in the analysis.</td>
</tr>
<tr>
<td><strong>Performance and detection bias</strong></td>
<td>Where possible, in all the studies included in the SR/MA and the new RCT, was there blinding of study participants, clinicians administering the treatment, ancillary care-givers, and outcomes assessors? When blinding is not possible, could knowledge of the treatment received affect interpretation of any of the outcomes?</td>
<td>Some objective outcomes are unlikely to be affected by knowledge of the intervention arm, but failure to blind (particularly for subjective outcomes) may lead to an exaggeration of effect sizes in individual studies, and these may be amplified in MAs.</td>
</tr>
<tr>
<td><strong>Attrition bias</strong></td>
<td>Were all dropouts documented and unlikely to be related to the treatment outcome in the studies included in the SR/MA and in the new RCT?</td>
<td>If dropout rates differ between the treatment arms, then the reasons may be related to the outcome of interest and may hide important outcome effects.</td>
</tr>
<tr>
<td><strong>Reporting bias</strong></td>
<td>Were all outcomes that were stated in the methods and/or protocol for all the studies included in the SR/MA and in the new RCT documented in the trial report? Were all the outcomes measured appropriately (as defined in the protocol) or were deviations reasonably explained?</td>
<td>Selective reporting of outcomes, or selective methods of reporting, may lead to exaggerated estimates of effect.</td>
</tr>
<tr>
<td><strong>Publication bias</strong></td>
<td>Were funnel plots used to investigate publication bias in the SR/MA? Is the funnel plot symmetric or is there reason to believe there is a systematic difference between published and unpublished studies? Note: This is difficult to assess when there are &lt;10 RCTs contributing to an MA.</td>
<td>Asymmetric funnel plots raise suspicion that there are systematic differences between published and unpublished studies and that some positive or negative trials may be unpublished. This may lead to exaggerated effect sizes in an MA.</td>
</tr>
<tr>
<td><strong>Consistency and heterogeneity of outcome</strong></td>
<td>Did the studies included in the SR/MA have overlapping 95% CIs for the outcome? Was variation more than would be expected by chance alone? Was the I² statistic &lt;40%? (Cochrane GRADE rule of thumb) Were subgroups used to explain any observed heterogeneity? Were event rates in the control group similar in the different studies? Note: Subgroups of the population, the intervention/control types, or the outcome measurement may explain heterogeneity.</td>
<td>If it can be shown that the outcomes are more effective in certain subgroups, or with variations of an intervention (eg, a higher dose), then this explained heterogeneity may indicate a key difference that may justify the results in the new RCT. Where unexplained heterogeneity exists, then the estimate of effect is likely to be uncertain, even if precise.</td>
</tr>
<tr>
<td><strong>Directness</strong></td>
<td>Do the studies included in the SR/MA and the new RCT both directly assess the research question about the population, interventions, and outcomes?</td>
<td>Indirect populations, interventions, surrogate outcome measures, or indirect comparisons may conceal or exaggerate important differences within and between studies, and may impact the estimate of effect.</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>Were the sample sizes for the studies included in the SR/MA and the new RCT powered to address the outcomes of interest? Does the 95% CI in the MA include clinically judged appreciable benefit and harm?</td>
<td>If any of the trials in the SR/MA or the new RCT were not powered to detect a clinically meaningful difference in the effect estimate, this may reduce confidence in the estimate of effect. If the lower and upper 95% CI thresholds indicate that the intervention may be beneficial at one end, but harmful at the other, this will probably reduce confidence in the estimate of effect.</td>
</tr>
<tr>
<td><strong>Sensitivity analyses</strong></td>
<td>When some studies included in an SR/MA are judged to be at high RoB and others at low RoB, or extreme variations in the populations or interventions in the studies are apparent, did the authors conduct a sensitivity analysis to ascertain the estimates of effect for only studies judged to be at low RoB?</td>
<td>Sensitivity analyses are different from subgroup analyses. Some studies are actively omitted as we are only interested in the results when the biased or “different” studies are omitted.</td>
</tr>
</tbody>
</table>

SR = systematic review; MA = meta-analysis; RCT = randomized controlled trial; CI = confidence interval; RoB = risk of bias.
of the evidence, and the authors concluded that there was an urgent need to conduct a large, robust, multicenter RCT to address these shortcomings. Pickard et al [8] published the results of such an RCT in 1167 patients and found no evidence that either tamsulosin or nifedipine increased the rate of spontaneous stone passage compared to placebo. The results were consistent across subgroup and sensitivity analyses.

We compare the RCT by Pickard et al [8] to the MA with the most studies, by Seitz et al [36], to explore and discuss discordant findings. Most RCTs included in the Seitz MA were small and recruited from a single center; only six of 35 (17%) recruited more than 100 patients. The majority had low internal validity and only one RCT reported allocation concealment. As small RCTs may report larger effect sizes compared to larger RCTs, an MA of small RCTs can lead to biased estimates of treatment effects [39]. Seitz et al also found evidence of publication bias, which can lead to overestimation of treatment effects and compromise the validity of the MA findings [40].

Seitz et al [36] found evidence of clinical heterogeneity concerning the patient inclusion criteria, stone characteristics, intervention, treatment in the control group, and outcome measurement. In the MA, the primary outcome of being stone-free was inconsistently defined, assessed using different imaging modalities, and measured at a variety of time points. In the RCT of Pickard et al [8], the primary outcome was any need for further intervention within 4 wk of randomization, which is compared here to being stone-free. In the control group of the Pickard RCT, 80% of patients were stone-free, whereas in the Seitz review, stone-free rates ranged from 4% to 78%, which highlights the potential impact of heterogeneity in the studies included.

With contrasting primary outcomes and different baseline event rates in the control groups, it is not surprising that the RCT and MA reported discordant findings. The choice of primary outcome is clearly of paramount importance in any trial. Heterogeneity in the conduct, design, and reporting of trials in this MA makes pooled treatment effects difficult, if not impossible, to interpret.

5.2. Partial versus radical nephrectomy

In a European Organization for Research and Treatment of Cancer (EORTC) RCT involving 541 patients with a solitary T1–2N0M0 renal tumor of ≤5 cm, 21 patients progressed, nine after radical nephrectomy (RN) and 12 after partial nephrectomy (PN). An intent-to-treat analysis found an overall survival (OS) advantage in favor of RN (hazard ratio [HR] 1.5, p = 0.03); however, only 12 of the 117 deaths were due to kidney cancer, four after RN and eight after PN [10]. Subsequently, Kim et al [9] published an SR and MA including some 41 000 patients and found statistically significant improvements in both OS (HR 0.81, p < 0.001) and disease-specific survival (HR 0.71, p < 0.001), but this time in favor of PN [9]. How can this discordance be explained?

The Kim MA has a number of limitations. First, the 38 trials included were mostly retrospective, single-center studies. The only RCT was the EORTC study. No information was provided about the distribution of follow-up or patient characteristics by treatment group (T category when >T1, tumor size, tumor grade, cell type, or renal function). Consequently, the differences in survival observed may not be directly due to differences in treatment efficacy. In addition, it is not clear to which patients the results can be generalized. Lastly, there was significant heterogeneity in the size of the treatment effect across the studies, so the overall estimate of the HR is not meaningful. Nevertheless, the EORTC RCT also had limitations and should be interpreted with caution: 55 patients crossed over to the other randomized treatment, 140 patients were clinically or pathologically ineligible, and there were few cancer-related events.

The MA found that PN was associated with a lower risk of severe chronic kidney disease (CKD); however, the EORTC study only found lower incidence of at least moderate renal dysfunction, not of advanced kidney disease or renal failure, and this was not associated with a corresponding difference in survival [41]. The studies in the MA did not always specify the status of the contralateral kidney, whereas in the EORTC study the contralateral kidney had to be normal.

Critical information regarding the biases of the studies included in the SR were not made explicit, since a Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach [42] was not used to assess the quality of the evidence. The quality of the studies in the SR and the heterogeneity of the results call into question the validity of the conclusions of the MA, which should thus be viewed with skepticism. In the same year, another SR suggested that localized renal cell carcinomas are best managed by PN where technically feasible. However, the evidence base had significant limitations owing to studies of low methodological quality and high RoB [43].

Further nonrandomized studies have found improved survival after PN [44,45] and a reduction in the risk of cardiovascular events [46] relative to RN; however, patients chosen for PN had a higher baseline likelihood of long-term survival [47,48]. In another study, only patients with stage II CKD had a lower risk of developing significant renal impairment after PN [49]. More recently, an SR and MA of 21 nonrandomized comparative studies in patients with clinical T1b and T2 renal tumors found better tumor control and survival for PN compared to RN [50], but the SR is subject to the same biases as the Kim MA.

Taking into account all the available efficacy data and a perceived advantage in renal function, the 2016 EAU Guidelines recommend, with several exceptions, that localized renal cancers are better managed by PN than RN.

6. Discussion

It is generally accepted that a high-quality SR of RCTs and an associated MA can provide a higher LE than a single RCT addressing the same question [2]. This can be problematic, however, when the results of the MA are in direct conflict with the RCT, making it difficult for guideline organizations to interpret the evidence and issue recommendations.
Guideline groups should follow well-defined methodological rules to assess the studies in these situations. RCTs should be appraised for their internal and external validity using established tools [51]. The conflicting SR/MA should be appraised in the same fashion to determine the methodological quality of the review, the quality of the studies included, inconsistency within the studies, unexplained heterogeneity, and the likelihood of publication bias using tools such as AMSTAR [28,29] and DART [30]. In some cases, the discrepancy may be due to errors in the MA in applying study eligibility criteria or even data extraction [52], so there is a need for an SR/MA protocol and strict quality control.

When MAs include many small underpowered studies, especially combined with likely presence of publication bias, there is immediate concern for overinflation of, or completely erroneous, effect size measurement. In addition, when a great degree of heterogeneity exists in the MA that cannot be easily accounted for, the results may be highly unreliable. In this regard, IPD MAs provide a better platform for assessing and explaining heterogeneity than aggregate data MAs do.

Two examples were discussed in this manuscript to illustrate the assessment process. In the case of MET for ureteric stones, a large, high-quality RCT [8] contradicted many well-established MAs that pointed to a benefit of this therapy. Analysis of a representative MA [36] revealed the inclusion of many small RCTs, poor internal validity, significant study heterogeneity and likely publication bias. When such MA concerns are present, a single high-quality RCT may be considered as having the higher LE. For guideline organizations, this process can be used to justify a change in recommendations based on methodologically sound principles.

Radical versus partial nephrectomy provides a more complex example. The MA [9] included only a single RCT, which was the study in conflict with its own results. The other studies included were all retrospective, which in general provide a lower LE. Risk of bias was poorly assessed, and significant study heterogeneity was present. It is important to reiterate that combining observational studies in general, and even comparative nonrandomized studies with RCTs in an intervention MA, may produce unreliable results and is not considered valid. In light of all this, the single RCT [10] in this circumstance might provide more guidance than the MA if it was of significantly high quality. However, this RCT also had some methodology concerns, so the comparison is not so simple.

Instead of automatically assigning a higher LE to SR/MA that conflict with RCTs, these examples have shown that the quality of the evidence and the RoB of studies included in SRs/MA should be assessed to determine which source provides the better evidence.

Although non-RCTs can be included in SRs, we have emphasized that only RCTs should be included in intervention MAs. RCTs are not required for MAs of prognostic factors and the accuracy of diagnostic tests, however, the studies included in these MAs should preferably be prospective in nature and based on a protocol to minimize RoB.

Despite the availability of MAs and RCTs, and in cases where high LE does not exist, we may still not know what the best treatment is. The GRADE system, which takes into account the quality of evidence (high, moderate, low, very low) for critical outcomes, provides strengths of recommendations (strong, weak) for or against a treatment to aid clinicians in their practice when consensus is not possible [42,53]. A decision curve approach, which takes into account a patient’s values and preferences, may also be used to help choose between the different treatment options.

7. Conclusions

New or existing RCT data can lead to conflicts with MA data. In this paper, we present examples of and explore reasons for such conflicts. Guidance is provided to guideline developers on how to interpret conflicting data in such circumstances to help assess which source is more reliable.

For guideline organizations both within and outside urology, having a well-defined and robust process to deal with such conflicts is essential to improve guideline quality.

Author contributions: Richard J. Sylvester had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Sylvester, N’Dow.

Acquisition of data: Sylvester, Lam, Marconi, S. MacLennan, Yuan, Van Poppel, N’Dow.

Analysis and interpretation of data: Sylvester, Canfield, Lam, Marconi, S. MacLennan, Yuan, G. MacLennan, Norrie, Omar, Bruins, Hernandez, Plass, Van Poppel, N’Dow.

Drafting of the manuscript: Sylvester, Canfield, Lam, Marconi, S. MacLennan, Yuan, G. MacLennan, Norrie, Omar, Bruins, Hernandez, Plass, Van Poppel, N’Dow.

Critical revision of the manuscript for important intellectual content: Sylvester, Canfield, Lam, Marconi, S. MacLennan, Yuan, G. MacLennan, Norrie, Omar, Bruins, Hernandez, Plass, Van Poppel, N’Dow.

Statistical analysis: None.

Obtaining funding: None.

Administrative, technical, or material support: None.

Supervision: Sylvester, N’Dow.

Other: None.

Financial disclosures: Richard J. Sylvester certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (eg, employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: None.

Funding/Support and role of the sponsor: None.

References


[38] EAU/AUA Nephrolithiasis Guideline Panel. 2007 guideline for the management of ureteral calculi. www.auanet.org/education/guidelines/ureteral-calculi.cfm


Would you like to receive all the benefits of EAU membership, but have no time for tedious paperwork?

**Becoming a member is now fast and easy!**

Go to [uroweb.org/membership](http://uroweb.org/membership) to apply online. It will only take you a couple of minutes to submit your application, the rest is for you to enjoy!