Pediatric Nephrology (2023) 38:3509–3512
https://doi.org/10.1007/s00467-023-06118-2

EDITORIAL COMMENTARY

Meta-analyses: how can we ensure that the hole is not greater than the sum of the parts?

Kathleen Kieran[1,2]

Received: 29 May 2023 / Revised: 30 July 2023 / Accepted: 31 July 2023 / Published online: 9 August 2023

Buder and colleagues [1] undertook a systematic review and meta-analysis to evaluate whether non-surgical management was supported in children with congenital non-refluxing primary megaureter. They concluded—based on eight studies—that the high prevalence of spontaneous resolution and low pooled prevalence of surgical intervention suggest that non-surgical intervention appeared to be supported. While it is reassuring to find current practice supported by a review of available data, this study raises more questions than it answers, first among them: "Wait, what?".

To be clear, Buder and colleagues undertook a well-designed and well-executed study. They scoured multiple databases for peer-reviewed publications, reviewed conference proceedings, and even checked clinical trial registries. They registered their systematic review prospectively in PROSPERO, and the data were reported in accordance with current best practices (PRISMA statement and the Cochrane Handbook for Systematic Reviews) [2, 3]. They reviewed manuscripts, conference proceedings, and registries published between 1947 and 2022. Seventy-six records were reviewed in full-text format, and only eight studies were included in the final analysis. As over 28,000 records were identified through the initial search, ultimately those eight studies represented fewer than 0.3% of records.

In rare conditions for which large-scale, randomized controlled trials are not feasible or practical, meta-analyses offer the opportunity to combine the results of multiple smaller studies. This increased power associated with an effectively larger sample size can more accurately characterize the magnitude of an observed effect than any of the smaller studies alone [4]. Similarly, systematic reviews can be useful for proactively informing the direction of future research [5]. In both cases, the quality of the studies selected and the data analysis performed will determine the quality of the conclusions drawn. While both meta-analyses and systematic reviews have the potential to include a majority (if not all), of the published data on a given topic, the aggregate sample size may still be relatively small. Consequently, the included patients may be different in ways that may not be readily evident even through careful review of a manuscript or use of thoughtful selection criteria.

It is by no means unusual that systematic reviews and meta-analyses report on relatively few manuscripts: a review of recent publications in PubMed finds 103 studies included in one systematic review [6], 25 in another [7], and 19 in a third [8]. This fivefold range illustrates two points: that a meta-analysis including 103 studies is fairly large despite the low absolute number of manuscripts and that the number of publications on less common conditions is widely variable. Stringent inclusion criteria further decrease the number of eligible publications: Buder's group identified 28,000 manuscripts that were culled to the eight included in their manuscript. What is most noteworthy about Buder and colleagues' work is the quality of the data included in the meta-analysis: not only did almost half of the eight included studies have a high risk of bias, but the data reported in the eight studies—though ostensibly inclusive of the same variables and outcomes—had substantially different levels of quality and detail. Furthermore, all eight studies reported retrospectively reviewed data from a single institution. In other words, each of the eight studies provided a curated presentation of a single institution's data.

Why is this a problem? Meta-analyses and systematic reviews will, by their very nature, include studies that have different inclusion criteria, sample composition, and study design. Buder and colleagues should be commended for reviewing conference proceedings and clinical trial registries to minimize one common pitfall of systematic reviews: publication bias. However, the eight studies included in the

✉ Kathleen Kieran
kathleen.kieran@seattlechildrens.org

[1] Division of Urology, Seattle Children's Hospital, 4800 Sand Point Way NE, OA.9.220, Seattle, WA 98105, USA

[2] Department of Urology, University of Washington, Seattle, WA, USA

current systematic review, while often including the same endpoints and measurement tools, varied in how measurements were taken and reported. Differential kidney outcome and urinary drainage during follow-up were both reported inconsistently and often qualitatively in the included studies. "Resolution" of non-refluxing primary megaureter was reported in almost two-thirds of patients in seven (not eight) studies, though Buder et al. found that the criteria for what constituted "resolution" was not uniform across studies nor even documented in most of the included manuscripts. Simply put, the published studies tended to report interpretations of the primary data, rather than the primary data itself. Moreover, the lack of agreement on nomenclature and classification raises concern: if different groups of authors cannot even agree on an operational definition for a clinical outcome, are these studies too heterogeneous to analyze together?

The current meta-analysis highlights one of the most maddening truths of research: the findings are only as good as the underlying data. Peer-reviewed manuscripts reflect only those papers that are selected for publication, and presentation of an abstract at a conference is no guarantee that a manuscript will follow. Almost two-thirds of conference abstracts were not published within 2–5 years (ironically, these data are derived from meta-analyses) [9–11]. Studies in which the outcome is the decision to proceed with surgical intervention have been heavily criticized, since without clear prospective criteria to consider surgery, surgeon preference and/or patient-specific nuances may play a disproportionate role [12, 13]. There is no expectation that every study included in a meta-analysis would have exactly the same design, subject group, and endpoints. However, each added layer of heterogeneity among studies further decreases the generalizability of the results of the meta-analysis. At some point, often unknown to the authors, the groups in different studies become too dissimilar to analyze together, and the well-intentioned recommendation of the meta-analysis is unknowingly unsupported by data.

For rare diseases, the above impacts are magnified, since uncommon conditions are often the subject of meta-analyses. According to the National Organization for Rare Diseases (NORD), rare diseases are those that affect fewer than 200,000 Americans, and 25–30 million Americans are living with a rare disease at any given time [14]. While this is an enormous number of people living with a rare disease, all of these people are not, of course, living with the same rare disease. NORD notes over 7000 known rare diseases [14]. Rare conditions require collaborative analysis of data from patients at geographically different sites and at different periods in time, making it less likely that data will be collected in precisely the same way in different studies.

Wilms tumor is the paradigm for successful progress in the management of a rare disease. Dismal clinical outcomes prompted the creation of the National Wilms Tumor Study (later the Children's Oncology Group), the model for inter-institutional collaboration in data collection and analysis. This collaborative work has generated treatment protocols that share best practices for the diagnosis and management of Wilms tumor and have more importantly facilitated identification of prognostic factors that direct tailored medical and surgical intervention, as well as development of clinical trials to support further progress [15].

Critical to the success of the collaborative Wilms tumor research is the collection of the same primary data, in the same way, for every included patient. Clinical data (e.g., operative notes), imaging tests, and surgical specimens are evaluated by institutional and central reviewers to minimize errors and variability in how data are reported and recorded [16, 17]. One example of the importance of recording primary rather than interpreted data is kidney function. Kidney function is assessed by reviewing serum creatinine, which allows calculation of creatinine clearance or glomerular filtration rates using approved equations. Advances in medicine, such as the recent exclusion of race from the calculation of glomerular filtration rate [18, 19] at many institutions, still permit utilization of the primary data to assess kidney function using the new equation; had only glomerular filtration rate been recorded *a priori*, such recalibration would not be possible.

In contrast, meta-analyses are often composed of collections of manuscripts and other available data that were intended for use by the authors and readers only, not for intentional inclusion in a larger analysis [20]. The authors of the meta-analyses should not be faulted, as their intent was almost certainly to share their clinical findings in what they believed to be the most clear and succinct way. Similarly, journal reviewers and editors clearly valued the information and its presentation when recommending the manuscript for publication. However, intent and impact often differ, and the impact of eight different author groups each presenting data in their preferred way is that there are eight distinct studies rather than eight institutions cohesively assessing a similar clinical question.

Research is, at its core, the intent to gather and analyze data in a way that generates generalizable findings and allows those analyzing and interpreting the data to conclude with a recommendation based on their findings [21]. With this in mind, the extreme variation in how endpoints are defined and reported in the eight studies included in this meta-analysis raises concern that the data may not be generalizable. In reporting their single-institution, retrospective experiences, have the authors of those eight studies simply reported their own data points? Or are these eight institutions independently and consistently reporting findings that would unquestionably apply to larger cohorts of children across the globe? We will never know, because the lack of

consistency in how the data were collected, interpreted, and recorded (in particular the subjective, qualitative, and general assessment of differential kidney function and kidney unit excretion) makes the eight groups uncertainly comparable. While patient characteristics and study design may vary somewhat among studies, the absence of shared operational definitions for quantifying differential kidney function, urinary drainage, or resolution is surprising—and worrisome.

Inconsistency in what should presumably be readily consistent data has been reported for many clinical findings in medicine. Medical students and residents may hear a pulmonary wheeze or a cardiac murmur that has disappeared by the time the attending examines the patient, and repeated blood pressure measurements are typically similar, rather than identical. Within urology, there is significant inter- (and intra-) rater variability in the grading of vesicoureteral reflux on voiding cystourethrography, the assessment of bladder qualities and function on urodynamic studies, and even how a voiding cystourethrogram is performed and reported [22–25]. In any test that relies on collection as well as interpretation of data, there is the potential for error and variation at each step. For example, a voiding cystourethrogram or urodynamic study may show vesicoureteral reflux on one cycle but not on the next (variation in data generated by the test), and two clinicians reviewing the same images from the test may grade the reflux differently (variation in data interpretation) [26]. When data are presented only in a refined, interpreted form, nuances are lost. Some of these nuances may be important, but specific data can only be made general, not vice versa. The generalizability of results, as Kukull and Ganguli note [27], is predicated on the ability of the researcher to cull relevant from irrelevant information—easier said than done. When quantitative data are reported qualitatively and only on some participants, teasing apart data to identify which kidney units have altered differential function or drainage is akin to attempting to identify the top students in a class when the only provided data is whether they have passed a single test or not.

As authors, reviewers, and editors, we can, and we must, do better. No study is perfect: biases and confounders abound. However, it is incumbent upon every researcher to include the highest quality data possible. In some cases, this may mean including primary data rather than interpretations, which can feel clumsy. Reviewers and editors must hold authors accountable: are the data provided in a manuscript sufficiently granular? In reviewing a single manuscript, a lapse in quality or a large proportion of missing data may not be noticeable or may be easily rationalized. However, Buder and colleagues have nicely illustrated that, when multiple studies have missing data and/or high levels of bias, the ability to draw meaningful conclusions from an analysis of those studies together is extremely limited. Uncommon conditions with limited publications derived from retrospective, incomplete, and non-primary data are at particular risk of this happening. The consequence is that individual manuscripts—and not larger-scale analyses—drive clinical practice. Apparent "evidence-based practice" may in fact be an echo chamber of the published experiences of a rarefied few, rather than a considered and thoughtful analysis of aggregate data.

Declarations

Conflict of interest The author declares no competing interests.

References

1. Buder K, Opherk K, Mazzi S, Rohner K, Weitz M (2023) Non-surgical management in children with non-refluxing primary megaureter: a systematic review and meta-analysis. Pediatr Nephrol. https://doi.org/10.1007/s00467-023-05938-6
2. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, McKenzie JE (2021) PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 372:n160. https://doi.org/10.1136/bmj.n160
3. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (eds) (2022) Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022). Cochrane. www.training.cochrane.org/handbook
4. Ioannidis JPA, Cappelleri JC, Lau J (1998) Issues in comparisons between meta-analyses and large trials. JAMA 279:1089–1093. https://doi.org/10.1001/jama.279.14.1089
5. Yuan Y, Hunt RH (2009) Systematic reviews: the good, the bad, and the ugly. Am J Gastroenterol 104:1086–1092. https://doi.org/10.1038/ajg.2009.11
6. van Netten JJ, Raspovic A, Lavery LA, Monteiro-Soares M, Paton J, Rasmussen A, Sacco ICN, Bus SA (2023) Prevention of foot ulcers in persons with diabetes at risk of ulceration: a systematic review and meta-analysis. Diabetes Metab Res Rev. https://doi.org/10.1002/dmrr.3652
7. Kerling DA, Clarke SC, DeLuca JP, Evans MO, Kress AT, Nadeau RJ, Selig DJ (2023) Systematic review and meta-analysis of the effect of loop diuretics on antibiotic pharmacokinetics. Pharmaceutics 15:1411. https://doi.org/10.3390/pharmaceutics15051411
8. Greco T, Zangrillo A, Biondi-Zoccai G, Landoni G (2013) Meta-analysis: pitfalls and hints. Heart Lung Vessel 5:219–225
9. Scherer RW, Meerpohl JJ, Pfeifer N, Schmucker C, Schwarzer G, von Elm E (2018) Full publication of results initially presented in abstracts. Cochrane Database Syst Rev 11:MR000005. https://doi.org/10.1002/14651858.MR000005.pub4
10. Jian-Yu E, Ramulu PY, Fapohunda K, Li T, Scherer RW (2020) Frequency of abstracts presented at eye and vision conferences being developed into full-length publications. A systematic review and meta-analysis. JAMA Ophthalmol 138:689–697. https://doi.org/10.1001/jamaophthalmol.2020.1264
11. Chua KJ, Mikhail M, Patel HV, Tabakin AL, Doppalapudi SK, Ghodoussipour SB, Kim IY, Jang TL, Srivastava A, Singer EA (2022) Quantifying publication rates and time to publication for American Urological Association podium presentations. J Urol 207:684–691. https://doi.org/10.1097/JU.0000000000002258

12. Sabharwal S, Patel NK, Griffiths D, Athanasiou T, Gupte CM, Reilly P (2016) Trials based on specific fracture configuration and surgical procedures likely to be more relevant for decision making in the management of fractures of the proximal humerus. Findings of a meta-analysis. Bone Joint Res 5:470–480. https://doi.org/10.1302/2046-3758.510.2000638
13. Hopkins C, Hettige R, Soni-Jaiswal A, Lakhani R, Carrie S, Cervin A, Douglas R, Fokkens WJ, Harvey R, Hellings PW, Leunig A, Lund VJ, Philpott C, Smith T, Wang DY, Rudmik L (2018) CHronic Rhinosinusitis Outcome MEasures (CHROME) – developing a core outcome set for trials of interventions in chronic rhinosinusitis. Rhinology 56:22–32. https://doi.org/10.4193/Rhin17.247
14. National Organization for Rare Diseases (2023) https://www.rarediseases.org. Accessed 28 May 2023
15. Dome JS, Graf N, Geller JI, Fernandez CV, Mullen EA, Spreafico F, Van den Heuvel-Eibrink M, Pritchard-Jones K (2015) Advances in Wilms tumor treatment and biology: progress through international collaboration. J Clin Oncol 33:2999–3007. https://doi.org/10.1200/JCO.2015.62.1888
16. Ehrlich PF, Hamilton TE, Gow K, Ehrlich PF, Hamilton TE, Gow K, Barnhart D, Ferrer F, Kandel J, Glick R, Dasgupta R, Naranjo A, He Y, Perlman EJ, Kalapurakal JA, Khanna G, Dome JS, Geller J, Mullen E (2016) Surgical protocol violations in children with renal tumors provides an opportunity to improve pediatric cancer care: a report from the Children's Oncology Group. Pediatr Blood Cancer 63:1905–1910. https://doi.org/10.1002/pbc.26083
17. Ehrlich PF, Ritchey ML, Hamilton TE, Haase GM, Ou S, Breslow N, Grundy P, Green DM, Norkool P, Becker J, Shamberger RC (2005) Quality assessment for Wilms' tumor: a report from the National Wilms' Tumor Study-5. J Pediatr Surg 40:208–212. https://doi.org/10.1016/j.jpedsurg.2004.09.044
18. Shi J, Lindo EG, Baird GS, Young B, Ryan M, Jefferson JA, Mehrotra R, Mathias PC, Hoofnagle AN (2021) Calculating estimated glomerular filtration rate without the race correction factor: observations at a large academic medical system. Clin Chim Acta 520:16–22. https://doi.org/10.1016/j.cca.2021.05.022
19. Nkinsi NT, Young BA (2022) How the University of Washington implemented a change in eGFR reporting. Kidney360 3:557–560. https://doi.org/10.34067/KID.0006522021
20. Farrar JL, Childs L, Ouattara M, Akhter F, Britton A, Pilishvili T, Kobayashi M (2022) Systematic review and meta-analysis of the efficacy and effectiveness of pneumococcal vaccines in adults. Pathogens 12:732. https://doi.org/10.3390/pathogens12050732
21. US Department of Health and Human Services. Office of Research Integrity (2023) Module 1: introduction: what is research? https://ori.hhs.gov/module-1-introduction-what-research. Accessed 28 May 2023
22. Schaeffer AJ, Greenfield SP, Ivanova A, Cui G, Zerin JM, Chow JS, Hoberman A, Mathews RI, Mattoo TK, Carpenter MA, Moxey-Mims M, Chesney RW, Nelson CP (2017) Reliability of grading of vesicoureteral reflux and other findings on voiding cystourethrography. J Pediatr Urol 13:192–198. https://doi.org/10.1016/j.jpurol.2016.06.020
23. Smith PP, Hurtado EA, Appell RA (2009) Post hoc interpretation of urodynamic evaluation is qualitatively different than interpretation at the time of urodynamic study. Neurourol Urodyn 28:998–1002. https://doi.org/10.1002/nau.20730
24. Metcalfe CB, Macneily A, Afshar K (2012) Reliability assessment of international grading system for vesicoureteral reflux. J Urol 188:1490–1492. https://doi.org/10.1016/j.juro.2012.02.015
25. Swanton AR, Arlen AM, Alexander SE, Kieran K, Storm DW, Cooper CS (2017) Inter-rater reliability of distal ureteral diameter ratio compared to grade of VUR. J Pediatr Urol 13:207e1–5. https://doi.org/10.1016/j.jpurol.2016.10.021
26. Chou FH, Ho CH, Chir MB, Linsenmeyer TA (2006) Normal ranges of variability for urodynamic studies of neurogenic bladders in spinal cord injury. J Spinal Cord Med 29:26–31. https://doi.org/10.1080/10790268.2006.11753853
27. Kukull WA, Ganguli M (2012) Generalizability: the trees, the forest, and the low-hanging fruit. Neurology 78:1886–1891. https://doi.org/10.1212/WNL.0b013e318258f812