BMJ Open

Frequency of equivocation in surgical meta-evidence: a review of systematic reviews within IBD literature

John D Delaney,[1] John T Holbrook,[2] Robert K Dewar,[1] Patrick J Laws,[3] Alexander F Engel[4]

To cite: Delaney JD, Holbrook JT, Dewar RK, *et al*. Frequency of equivocation in surgical meta-evidence: a review of systematic reviews within IBD literature. *BMJ Open* 2017;**7**:e018715. doi:10.1136/bmjopen-2017-018715

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2017-018715).

Received 19 July 2017
Revised 24 October 2017
Accepted 15 November 2017

[1]Colorectal Surgery, Northern Clinical School, University of Sydney, Sydney, New South Wales, Australia
[2]Royal Prince Alfred Hospital, Camperdown, New South Wales, Australia
[3]Prince of Wales Hospital, Sydney, New South Wales, Australia
[4]Department of Colorectal Surgery, Royal North Shore Hospital, Sydney, New South Wales, Australia

Correspondence to
Dr John D Delaney;
jdel2642@uni.sydney.edu.au

ABSTRACT

Objective To assess the level of equivocation among level 1 evidence in ulcerative colitis and Crohn's disease and determine whether any predisposing factors are present.

Method MEDLINE, Embase, CINHAL and Cochrane were searched from 2006 to 2017. Papers were scored using AMSTAR and categorised into surgical (S), medical (M) or medical and surgical (MS) groups. The ability of each paper to make a recommendation and conclusiveness in doing so was recorded.

Results 278 papers were assessed. 82% (n=227) could make a recommendation, 18% (n=51) could not. There was a significant difference in ability to provide a recommendation between S and M (P=0.003) but not MS and M (P=0.022) nor S and MS (P=0.79). Where a recommendation was made, S papers were more likely to be tempered than M papers (P=0.014) but not MS papers (P=0.987).

Conclusions Surgical meta-evidence within the inflammatory bowel disease domain is more than twice as likely as medical meta-evidence to be unable to provide a recommendation for clinical practice. Where a recommendation was made, surgical reviews were twice as likely to temper their conclusion.

Strengths and limitations of this study

- Large sample of papers, the use of multiple independent reviewers and the validity of AMSTAR as a quality assessment tool.
- The methods used in search and data-retrieval have been clearly outlined, with explicit inclusion and exclusion criteria.
- The inability of AMSTAR to discriminate between poor methodological quality of a study and poor reporting quality within the paper (internal validity).
- There are potential avenues for bias in this paper. The use of inflammatory bowel disease (IBD) as a framework may introduce selection bias, particularly given that surgical intervention typically represents a failure of medical therapy in IBD. The assessment of a paper's level of equivocation is subjective and open to bias. An author's bias towards a subject may also contribute to a paper's self-reported level of equivocation and the reasons for equivocation.
- The assessment of conclusion is subjective, and subtle changes in language may influence the perceived level of confidence and the rationale for uncertainty.

INTRODUCTION

Methods of aggregate literature review first emerged in the 17th century, developing in an ad hoc fashion until the modern era.[1] In the late 1980s, a need to synthesise and understand the increasing volume of medical research drove the development of more sophisticated and systematic techniques.[2] Since then, well-conducted systematic reviews and meta analyses have become the gold standard level of evidence in healthcare.[3] Such has been the success of these studies in medicine, the process has branched into disciplines as diverse as economics, the social sciences and environmental management.[3–5]

Meta-evidence is derivative in nature and as such is dependent on the validity of its input studies to be able to make useful recommendations for clinical practice. When original high-quality trials are combined, they yield more useful meta-evidence that mixed or low quality studies. Unfortunately, many difficulties have been identified that limit the production of high-quality clinical research in surgery[6] when compared with medicine, as surgical interventions are typically complex interventions involving the interaction of many independent variables. This creates significant obstacles to generating robust randomised control trials[7–9] on surgical topics, and consequently, evidence-based surgery relies heavily on observational studies.[10] Audits of methodological rigour within surgical observational studies have been critical.[6 10 11] Meta-evidence created from a lower quality selection of original studies has an unreliable foundation. Additionally, an increasing number of papers are being published that examine methodology with surgical meta-evidence. The results of

those studies suggest that, in general, meta-evidence within surgery is of poorer methodological quality.[12–14] We therefore have a situation where, despite best efforts, surgical meta-evidence is being created from studies of poorer methodological quality than their medical counterparts, and the systematic reviews and meta-analyses themselves are performed with less rigour.

The research question of this 'review-of-reviews' is: what are the factors that influence the ability meta-evidence to make recommendations for clinical practice? Of particular interest is the effect that intervention has; when compared with medical meta-evidence, do the known challenges of original surgical evidence, combined with the historical methodological inferiority of surgical reviews, produce meta-evidence that is more equivocal within the inflammatory bowel disease (IBD) domain?

METHOD

Literature search

We completed a thorough literature search across MEDLINE, Embase, CINAHL and the Cochrane Database of Systematic Reviews. In addition to the search terms identified in online supplementary appendix 1, a free search of MEDLINE, Embase and CINAHL was completed using the keywords 'surgery', 'meta-analysis' and either 'crohn's' or 'ulcerative colitis'. Validated filters for systematic reviews and meta-analyses, specific to each of the databases, were applied.[15]

Definitions

Papers to be analysed were systematic reviews or meta-analyses, as defined by the Cochrane Collaboration.[16] Ulcerative colitis (UC) and Crohn's disease (CD) were chosen as the framework for this study as they are relatively common, serious conditions,[17] with both medical and surgical therapy options.[18] The surgical therapies included were derived from International Classification of Diseases, 9th revision, Clinical Modification (ICD-9-CM) procedure codes, along with expert consultation and review of current surgical literature.[18–20] Use of ICD-9-CM codes has been previously validated.[21]

Retrieved meta-evidence was categorised into groups based on the type of intervention it assessed. Where a medical therapy was considered exclusively, the paper was included in the M group. Where a surgical therapy only was considered, the paper was included in the S group. Where a medical therapy was considered in the context of a surgical therapy, or vice versa, the paper was included in the MS group.

Papers were further classified as recommendation (R) or no recommendation (NR) based on whether they could provide a recommendation for clinical practice. Each conclusion was rated as either firm (F) or tempered (T) based on the definitiveness of the language used. The conclusion section of each paper was used to assess recommendation and definitiveness. Papers that were R–F were defined as ones that could make a clinical recommendation (positive or negative) using language that was definite and offered minimal or no caveats for the recommendation. Papers that said definitively that there was no difference between interventions, that is, they could confidently *not* recommend an intervention, were also classed as R–F. Papers that were R–T were those that made a recommendation for practice but offered significant caveats. NR–T papers were not able to offer a recommendation for practice but suggested a recommendation may be possible in the future based on an emerging trend or sound underlying theory. NR–F papers were completely uncertain and could not make a recommendation nor offer further advice due to lack of evidence. See table 1 for a reference list of definitions.

Table 1 Definitions for level of recommendation

Recommendation Category	Definition
Recommendation–Firm	A conclusion that makes a recommendation for practice (positive or negative), with minimal or no caveats.
Recommendation–Tempered	A conclusion that makes a recommendation for practice but places significant caveats on that recommendation.
No recommendation–Tempered	A conclusion that is unable to make a recommendation but suggests that a recommendation might be possible in the near future based on an emerging trend or underlying theory.
No recommendation–Firm	A conclusion that is unable to make a recommendation and is completely equivocal. There may be no evidence at all, or of too poor a quality, or the evidence may be contradictory.

The AMSTAR scoring system was used to assess methodological quality.[22] AMSTAR consists of 11 individual scoring criteria and is well established as a valid means of assessing meta-evidence.[23] AMSTAR is a 'checklist' style tool. A higher total AMSTAR score in a paper indicated a more reliable level of methodology.

Inclusion criteria

We included systematic reviews or meta-analyses printed between January 2006 and September 2017, inclusive, which assessed a surgical or medical intervention in adults with CD or UC. Review articles were excluded. Papers regarding other IBDs were excluded. The search was limited to full-length publications.

Data extraction

Three reviewers examined abstracts (JDD, RKD and PJL). Full text was obtained where abstracts were unable to provide enough information. Updated reviews were used preferentially. Papers that were deemed suitable for inclusion were placed into one of three groups depending on their interventional focus: S, M or MS. JDD, RKD and PJL

scored the methodology of the papers via AMSTAR. Any disagreements were resolved by discussion to arrive at a majority decision. Interobserver agreement was assessed using kappa (κ).

A paper's recommendation and level of conclusiveness was recorded by JTH and JDD according to the previously stated definitions. Data on the number of papers per review, number of patients included in each review and the 5-year impact factor of the journal in which the paper was published were also recorded. Impact factor was retrieved from the *Journal of Citation Reports*.[24] For papers that included meta-analyses, the number of trials, number of patients and heterogeneity scores (I^2 for each) were also extracted.

Additionally, financial information for each of the papers was extracted based on their description of funding sources or, where that was not available, the affiliations of the first and last authors. Our categories for sponsorship were corporate, government, academia, or those groups in combination, non-government organisations or unclear. An unclear source of funding was recorded where a paper did not offer a conflict of interest disclosure or where a conflict of interest disclosure was offered but the sponsorship of the paper was not clearly outlined.

Statistical analysis

All of the collected data were collated into a Microsoft Excel spreadsheet.[25] The means of continuous data were compared via analysis of variance (ANOVA). Categorical data were analysed via χ^2 test. In both formats, a two-tailed distribution with an alpha level of 0.05 was used. A multivariate ANOVA (MANOVA) assessment of the continuous data set was also performed. Statistical analysis was performed using SPSS V.24.[26]

RESULTS

We identified 739 meta-evidence papers from our initial search. Three hundred and eighty-nine (389) were excluded based on titles or abstracts or because they were duplicated results. Three hundred and fifty papers were reviewed in full. Seventy-two of these papers were excluded (online supplementary appendix 2) (κ=0.8). The 278 included papers were allocated into one of three categories, depending on their interventional focus: S (n=48), M (n=195) or MS (n=35). Descriptive statistics may be found in table 2. The trial flow diagram representing our inclusion and exclusion process is shown in figure 1. Details of the included papers may be found in online supplementary appendix 3.

Overall, 18% of papers (n=51) were unable to make a clinical recommendation based on the available evidence. Within the S group, NR papers made up 31% (n=15). Within MS, NR papers comprised 29% (n=10). Within M, NR papers made up 13% (n=26). A χ^2 test was performed, and a significant relationship was found between the intervention type and the likelihood of a paper to be able to make a recommendation (χ^2 (2, n=278)=11.049, P=0.004). Comparison of individual groups using χ^2 with a Bonferroni correction (α=0.017) revealed a significant difference between S and M (P=0.003) but not between S and MS (P=0.79) nor M and MS (P=0.022).

One-way ANOVA showed significant differences between S, M and MS groups when comparing the total number of patients (P=0.02) and heterogeneity via I^2 (P=0.008). No difference was found in total number of papers, impact factor of journal or AMSTAR rating. Planned contrasts found S papers to have a significantly higher number of patients per review than M papers or MS papers (P=0.001, P=0.009). Contrasts also showed significantly higher heterogeneity via I^2 in S when compared with M (P=0.002) and in S and MS combined when compared with M (P=0.016).

Comparison of R versus NR groups using one-way ANOVA showed no significant difference when comparing total number of patients, number of studies included, heterogeneity via I^2, impact factor or AMSTAR. MANOVA analysis of the same group revealed no difference.

Of papers that gave a recommendation (n=227), 64% were firm (R–F; n=145, 52% of papers overall) and 36% were tempered (R–T; n=82). Of papers that gave no recommendation (n=51), 31% were firm (NR–F; n=16) and 69% were tempered (NR–T; n=35). Within the M group, 58% were R–F (n=114), 29% were R–T (n=55), 9% NR–T (n=18) and 4% NR–F (n=8). Within S, 38% were R–F (n=18), 31% were R–T (n=15), 21% were NR–T (n=10) and 10% NR–F (n=5). For MS, 37% were R–F (n=13), 34% R–T (n=12), 20% NR–T (n=7) and 9% NR–F (n=3). A χ^2 test was performed, and a significant relationship was found between the intervention type and the level of conclusiveness of the paper (χ^2 (6, n=278)=14.493, P=0.025). Comparison of individual groups using χ^2 with a Bonferroni correction (α=0.017) revealed a significant difference between S and M (P=0.014) but not between S and MS (P=0.987) nor M and MS (P=0.065). The number of equivocal reviews (NR–T + NR–F) covered 355 papers and 104 160 patients in M, 503 papers and 385 898 patients in S and 124 papers and 15 371 patients in MS.

Financial support of the papers audited is detailed in table 3. Notably, government funding was identified as the major sponsor in 22% of M (n=42), 2% of S (n=1) and 11% of MS (n=4). Academia was the primary sponsor in 28% of M (n=55), 44% of S (n=21) and 45% of MS (n=16). The funding source was unclear in 22% of M (n=43), 37% of S (n=17) and 17% of MS (n=6). Comparison of individual groups using χ^2 with a Bonferroni correction (α=0.017) revealed a significant difference between S and M on government funding (P<0.001) but not within categories of corporate, academic, combination sponsorship or where the funding was unclear. The MS group was not significantly different from either group across all categories.

Table 2 Paper characteristics

	N	Mean	SD	Minimum	Maximum
Total number of papers					
Medical	195	15.65	21.71	0	255
Surgical	48	23.25	29.38	2	196
Medical and surgical	35	13.54	7.98	4	33
Total	278	16.70	22.22	0	255
Total number of patients					
Medical	195	4706.21	14655.34	0	133519
Surgical	48	54283.52	211 315.58	43	1 111 988
Medical and surgical	35	2317.89	2683.26	134	12586
Total	278	12 965.63	89923.3	0	1 111 988*
Heterogeneity (I^2)					
Medical	118	31.14%	28.4%	0%	88%
Surgical	28	51.10%	35.3%	0%	99%
Medical and surgical	23	37.79%	32%	0%	76%
Total	169	35.35%	30.8%	0%	99%
Impact factor					
Medical	192	4.94	2.47	0.674	17.469
Surgical	48	4.37	3.42	0.918	17.445
Medical and surgical	35	4.71	2.64	1.531	16.716
Total	275	4.81	2.67	0.674	17.469
AMSTAR					
Medical	195	6.64	2.22	0	10
Surgical	48	6.23	2.10	1	10
Medical and surgical	35	6.46	2.22	1	10
Total	278	6.54	2.20	0	10

*Outlying surgical study.

DISCUSSION

This paper has examined the differences in the level of equivocation between surgical and medical meta-evidence. To our knowledge, this is the first such comparison. We believe it is important to address this issue as meta-evidence continues to be produced in increasing numbers in both medicine and surgery.[27 28] While the utility of meta-evidence within medicine is widely acknowledged, surgical interventions are typically more complex and heterogeneous, making the generation of robust surgical meta-evidence difficult.[8 9 11] Although the justification for meta-evidence within surgery is weaker than in medicine, the academic cache is transferrable; that is, it maintains its premier position in the busy clinician's evidence heuristic.

Papers that could not make a recommendation for practice were more likely to involve a surgical therapy. Papers in the S group were 2.5 times more likely than M papers to be equivocal. MS papers were twice as likely. The only other comparator that was predictive on a paper's conclusiveness was the number of patients included. On metrics of methodology, number of included studies, heterogeneity and impact factor, there was no difference on univariate or multivariate analysis.

Surgical meta-evidence was also less likely than medical meta-evidence to be confident in its recommendations for clinical practice, by a factor of two, and more likely to be *completely* uncertain by a factor of three. In a combined medical and surgical paper, the ratios for these criteria were 1.6 and 2, respectively.

Previous studies have found that surgical meta-evidence is more likely to have poorer methodology,[12] though this paper did not find support for that claim (potentially demonstrating an improving methodology in surgical meta-evidence, a topic for further research). Despite parity on this and other metrics, our study has found that combined surgical evidence is more than twice as likely to be equivocal when compared with corresponding medical reviews. An important distinction to bear in mind here is that AMSTAR assesses the methodology of the meta-analysis or systematic review technique, as opposed to the quality of the original input papers. Audits of original research methodology have found surgical papers to be poorer than medical ones in that regard.[29] Reasons

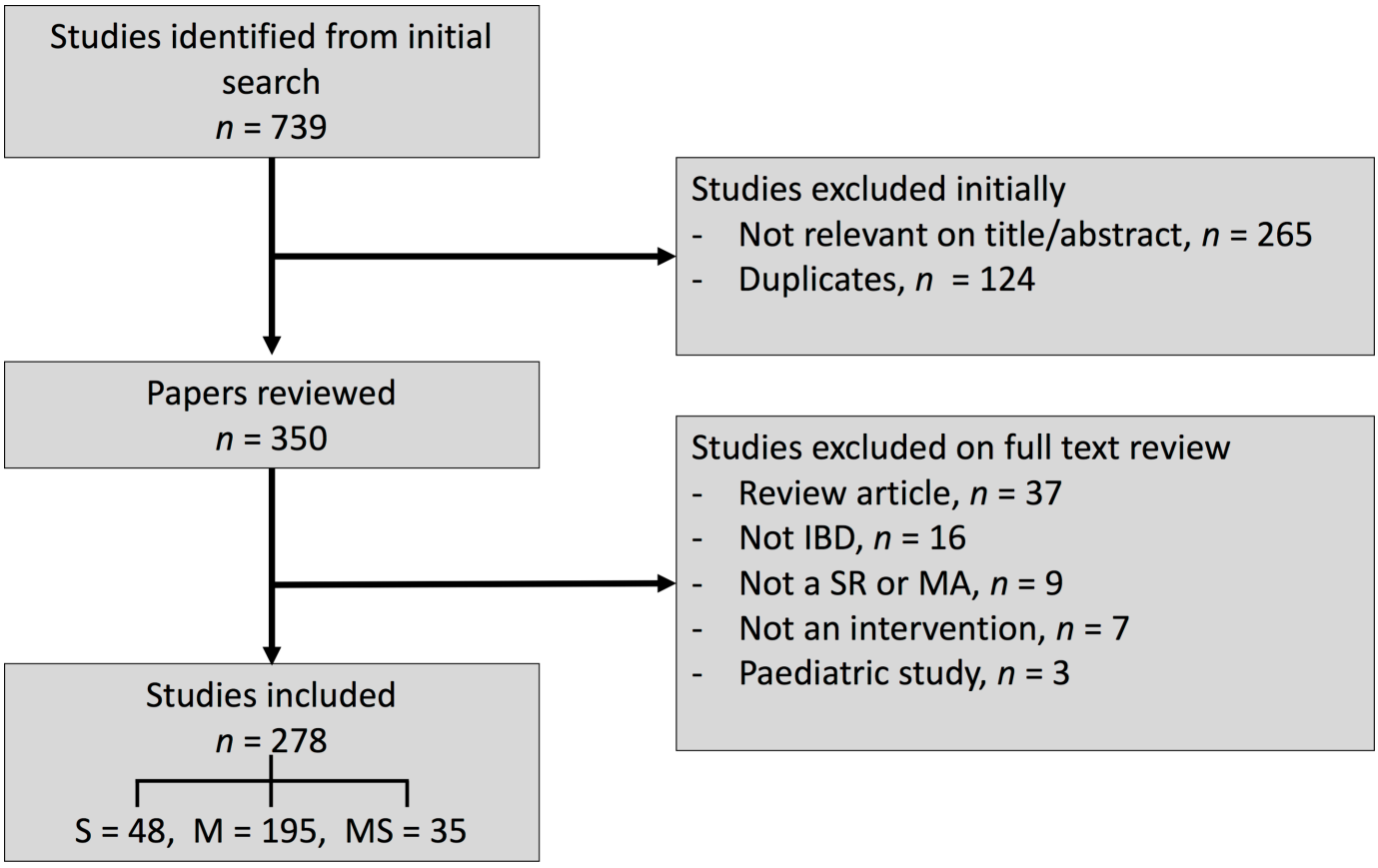

Figure 1 PRISMA paper inclusion and exclusion flow diagram. IBD, inflammatory bowel disease; M, medical intervention group; MA, meta-analysis; MS, medical and surgical intervention group; n, number of papers; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; S, surgical intervention group; SR, systematic review.

for this have been well espoused elsewhere.[30] This audit, by focusing on the ability of meta-evidence to provide a recommendation, raises two questions: first, given the prior probability of a clinical recommendation within surgical meta-evidence is 2.5 times less than in medical literature, is aggregate analysis of surgical evidence a worthwhile investment of limited resources?, and second, in light of this, should meta-evidence in surgery still be regarded as the 'best' available evidence?

The purpose of aggregation in level 1 evidence is to maximise our approximation of reality, but considering the findings shown here, is it possible that in surgical meta-literature, where input quality is poorer, aggregation leads to attenuation? High-quality trials will always be well regarded, but one wonders as to the influence of suboptimal trials and equivocal meta-evidence on the acceptance and application of evidence-based surgery. In this setting, a challenge is created for any surgeon attempting to practice 'best evidence'. This is perhaps best reflected when one looks at the degree of confidence that the authors of each paper have shown in their conclusions; higher levels of uncertainty are expressed in clinical recommendations for surgical procedure when compared with medical therapies by a factor of two.

Great effort, intellect and perseverance have given us the present surgical evidence and reviews on IBD, but the results of the present study suggest a higher level of scepticism towards surgical evidence and meta-evidence may be warranted. The lack of difference across the metrics studied in this paper, save for type of intervention (surgical vs medical), suggests an unresolved challenge to successfully combining original surgical research. An increase in error appears to be associated with the surgical research process when compared with equivalent medical research, which is exacerbated when combined analysis is performed. Continuation of surgical research that is of inferior quality to medical research, with less predictive power in the meta-evidence setting, weakens the standing of evidence-based surgical practice. However, equally, so too does surgical meta-evidence that must equivocate when presented with the available literature and whose calls for improved methodology in original studies have not been sufficiently heeded, excellent examples of which may be seen in sequential Cochrane reviews.[31 32]

Our financial analysis reveals a striking discrepancy in funding between surgical and medical meta-evidence, most notably in the government sector. This is despite a quarter-of-a-billion surgical cases worldwide annually.[33] How may these funding shortfalls, compounding the unique challenges of surgical research, be addressed? And in doing so, how may we create a surgical output more cohesive and clinically useful? The role of the

Table 3 Financial support of papers

	Intervention			
	Medical	Surgical	Medical and surgical	Total
Corporate				
Count	6	2	4	12
% Corporate	50.0	16.7	33.3	100.0
Government				
Count	42	1	4	47
% Government	89.4	2.1	8.5	100.0
Academia				
Count	55	21	16	92
% Academia	59.8	22.8	17.4	100.0
Other				
Count	14	3	0	17
% Other	82.4	17.6	0.0	100.0
Government and university				
Count	31	3	3	37
% Government and university	83.8	8.1	8.1	100.0
Corporate and u niversity				
Count	4	1	2	7
% Corporate and university	57.1	14.3	28.6	100.0
Unclear				
Count	43	17	6	66
% Unclear	65.2	25.8	9.1	100.0
Total				
Count	195	48	35	278
% All	70.1	17.3	12.6	100.0

international community of surgical academia to address this issue is paramount. In addition to petitioning government, increased levels of collaboration and consolidation may prove valuable.[34] Resources may be used in a more focused manner; for instance, the publication requirements of those who aspire to become academic surgeons provides a ready example of a resource that could be used more effectively towards targeted scientific questions.[34] Lastly, surgical journals must continue to insist on higher levels of methodology in surgical trials and a greater degree of focus on uniformity of trial design,[35] enhancing the reputation of surgical science and hence the argument for funding.

Strengths and limitations

The strengths of this 'overview-of-reviews' are the large sample of papers, the use of multiple independent reviewers and the validity of AMSTAR as a quality assessment tool. The methods used in search and data-retrieval has been clearly outlined, with explicit inclusion and exclusion criteria.

The limitations of the study include the inability of AMSTAR to discriminate between poor methodological quality of a study and poor reporting quality within the paper (internal validity). The use of IBD as a framework may introduce selection bias, particularly given that surgical intervention typically represents a failure of medical therapy in IBD. The findings of this 'review-of-reviews' are limited in their application outside of IBD research. Similar studies in differing fields will provide a useful basis for comparison. The assessment of a paper's level of equivocation is subjective and open to bias. An author's bias towards a subject may also contribute to a paper's self-reported level of equivocation and the reasons for equivocation. Subtle changes in the language may influence the perceived level of confidence and the rationale for uncertainty.

CONCLUSION

This paper has demonstrated that surgical meta-evidence within the IBD domain is 2.5 times more likely than medical meta-evidence to be unable to provide a recommendation for clinical practice. Whether the intervention being assessed was surgical or medical was the only significant predictor of equivocation when considered against meta-evidence methodology, number of

papers, number of patients or level of data heterogeneity. Surgical research also experiences resource limitations where compared with medical research, notably in government funding. We suggest that a discussion should be undertaken within the surgical community, including in this and other journals, about the evolution of the surgical research paradigm; how best to design a system of hypothesis testing that will generate robust results from the unique clinical, moral and human environment of the surgical intervention.

Contributors JDD and AFE were the designers of the work. The acquisition of the data was performed by JDD, JTH, RKD and PJL. JDD and AFE contributed to the analysis and interpretation of the data. The work was drafted by JDD, with critical revision by JTH, RKD, PJL and AFE. All authors gave final approval for the published version and agreed to be held to its accuracy and integrity.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement There are no unpublished data for this study. Any enquiries relating to the paper are welcome via email: jdel2642@uni.sydney.edu.au.

REFERENCES

1. Egger M, Ebrahim S, Smith GD. Where now for meta-analysis? *Int J Epidemiol* 2002;31:1–5.
2. Sacks HS, Berrier J, Reitman D, *et al*. Meta-analyses of randomized controlled trials. *N Engl J Med* 1987;316:450–5.
3. OCEBM Levels of Evidence Working Group. *The Oxford 2011 levels of evidence*. Oxford, UK: Oxford Centre for Evidence-Based Medicine, 2011.
4. Petticrew M. *Systematic reviews in the social sciences: a practical guide*. Boston: Blackwell Publishing, 2006.
5. Shemilt I, Mugford M, Vale L, *et al*. Evidence synthesis, economics and public policy. *Res Synth Methods* 2010;1:126–35.
6. Brooke BS, Nathan H, Pawlik TM. Trends in the quality of highly cited surgical research over the past 20 years. *Ann Surg* 2009;249:162–7.
7. Buchwald H. Surgical procedures and devices should be evaluated in the same way as medical therapy. *Control Clin Trials* 1997;18:478–87.
8. Ergina PL, Cook JA, Blazeby JM, *et al*. Challenges in evaluating surgical innovation. *Lancet* 2009;374:1097–104.
9. McLeod RS, Wright JG, Solomon MJ, *et al*. Randomized controlled trials in surgery: Issues and problems. *Surgery* 1996;119:483–6.
10. Rangel SJ, Kelsey J, Henry MC, *et al*. Critical analysis of clinical research reporting in pediatric surgery: justifying the need for a new standard. *J Pediatr Surg* 2003;38:1739–43.
11. Hall JC, Platell C, Hall JL. Surgery on trial: an account of clinical trials evaluating operations. *Surgery* 1998;124:22–7.
12. Delaney J, Laws P, Wille-Jørgensen P, *et al*. Inflammatory bowel disease meta-evidence and its challenges: is it time to restructure surgical research? *Colorectal Dis* 2015;17:600–11.
13. Dellinger EP. Increasing inspired oxygen to decrease surgical site infection: time to shift the quality improvement research paradigm. *JAMA* 2005;294:2091–2.
14. Sellke FW, DiMaio JM, Caplan LR, *et al*. Comparing on-pump and off-pump coronary artery bypass grafting: numerous studies but few conclusions: a scientific statement from the American Heart Association council on cardiovascular surgery and anesthesia in collaboration with the interdisciplinary working group on quality of care and outcomes research. *Circulation* 2005;111:2858–64.
15. Lee E, Dobbins M, Decorby K, *et al*. An optimal search filter for retrieving systematic reviews and meta-analyses. *BMC Med Res Methodol* 2012;12:51.
16. The Cochrane Collaboration. *Cochrane handbook for systematic reviews of interventions*: Wiley-Blackwell, 2011.
17. Talley NJ, Abreu MT, Achkar JP, *et al*. An evidence-based systematic review on medical therapies for inflammatory bowel disease. *Am J Gastroenterol* 2011;106(Suppl 1):S2–25.
18. Cima RR, Pemberton JH. Medical and surgical management of chronic ulcerative colitis. *Arch Surg* 2005;140:300–10.
19. Fichera A, Michelassi F. Surgical treatment of Crohn's disease. *J Gastrointest Surg* 2007;11:791–803.
20. Jones DW, Finlayson SR. Trends in surgery for Crohn's disease in the era of infliximab. *Ann Surg* 2010;252:307–12.
21. Quan H, Parsons GA, Ghali WA. Validity of procedure codes in International Classification of Diseases, 9th revision, clinical modification administrative data. *Med Care* 2004;42:801–9.
22. Shea BJ, Grimshaw JM, Wells GA, *et al*. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol* 2007;7:10.
23. Shea BJ, Hamel C, Wells GA, *et al*. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *J Clin Epidemiol* 2009;62:1013–20.
24. Thompson Reuters. *Journal citation reports. JCR science*, 2012.
25. Microsoft Corporation. *Microsoft excel for Mac*. 14.3.2 edn: Microsoft Corporation, 2011.
26. IBM Corp. *IBM SPSS statistics for macintosh*. 24.0 edn: IBM Corp, 2016.
27. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med* 2010;7:e1000326.
28. Tebala GD. What is the future of biomedical research? *Med Hypotheses* 2015;85:488–90.
29. Sinha S, Sinha S, Ashby E, *et al*. Quality of reporting in randomized trials published in high-quality surgical journals. *J Am Coll Surg* 2009;209:565–71.
30. Rosenthal R, Kasenda B, Dell-Kuster S, *et al*. Completion and publication rates of randomized controlled trials in surgery: an empirical study. *Ann Surg* 2015;262:68–73.
31. Lustosa SA, Matos D, Atallah AN, *et al*. Stapled versus handsewn methods for colorectal anastomosis surgery. *Cochrane Database Syst Rev* 2001:CD003144.
32. Neutzling CB, Lustosa SA, Proenca IM, *et al*. Stapled versus handsewn methods for colorectal anastomosis surgery. *Cochrane Database Syst Rev* 2012:CD003144.
33. Weiser TG, Regenbogen SE, Thompson KD, *et al*. An estimation of the global volume of surgery: a modelling strategy based on available data. *Lancet* 2008;372:139–44.
34. Søreide K, Alderson D, Bergenfelz A, *et al*. Strategies to improve clinical research in surgery through international collaboration. *Lancet* 2013;382:1140–51.
35. Wynne KE, Simpson BJ, Berman L, *et al*. Results of a longitudinal study of rigorous manuscript submission guidelines designed to improve the quality of clinical research reporting in a peer-reviewed surgical journal. *J Pediatr Surg* 2011;46:131–7.