original article

Eur Surg (2020) 52:69–73
https://doi.org/10.1007/s10353-019-00622-2

european surgery
ACA Acta Chirurgica Austriaca

Differences in the recommendation of laparoscopic clinical practice guidelines according to the recommendation system—Re-evaluation using GRADE

A pilot study

J. Leung · A. Ng · K. Gurusamy

Received: 3 July 2019 / Accepted: 26 November 2019 / Published online: 10 January 2020

Summary
Background Guidelines are essential for safe and effective treatment. They usually have multiple statements. Since the supporting information for the guidelines varies widely, the degree to which these statements are recommended also differ. We rely on recommendation systems for grading the recommendations for different statements. All recommendation systems have different grading criteria and they could potentially cause confusion and affect the quality of recommendations. Therefore, there is a need to determine the extent of variation and explore the potential reasons behind it.
Methods A purposive sampling on PubMed was conducted to find four different laparoscopic guidelines using different methods to grade the recommendations. Each statement was then re-evaluated using the GRADE recommendation system.
Results The guidelines used GRADE, Oxford Methodology, SIGN, and 'bespoke' systems. The number of statements with similar strength for the different statements as the re-evaluated strengths in the four guidelines were 24.1, 62.2, 35.8 and 50.0% respectively.
Conclusion There were a wide variety of recommendation systems for laparoscopic guidelines and there were differences between the recommendations from the guidelines using GRADE, Oxford Methodology, SIGN and the 'Bespoke' system when re-evaluated by GRADE. A systematic review of recent laparoscopic guidelines might provide the extent and the main reasons of the problem.

Keywords Guideline · Laparoscopy · Surgery · GRADE approach · Recommendation

J. Leung · A. Ng
Royal Free Campus, UCL, University College London Medical Student, Rowland Hill Street, Hampstead, London, NW3 2 PF, UK

J. Leung
zchaokj@ucl.ac.uk

A. Ng
zchangx@ucl.ac.uk

K. Gurusamy (✉)
Department of Surgical Biotechnology, Royal Free Campus, UCL, UCL Division of Surgical and Interventional Sciences, Rowland Hill Street, Hampstead, London, NW3 2 PF, UK
k.gurusamy@ucl.ac.uk

Introduction

In the UK the percentage of people receiving laparoscopic surgery increased from 10 to 28.4% between 2007 and 2009 [1]. In the US, it increased from 13.8 to 42.6% as well [2]. The increasing use of laparoscopy was due to the fact that for certain diseases such as Crohn's disease and procedures like appendectomy, it had become the gold standard with less pain, faster recovery, earlier return of bowel functions, shorter hospital length of stay, fewer scars etc. [3–5]. Moreover, many laparoscopic treatments have been demonstrated as non-inferior compared to open surgery. For example, a 10-year study on colon cancer treatment indicated that there was no difference between laparoscopy and open surgery in terms of survival and recurrences [6]. However, it did have its limitations. It was a steep learning curve for doctors and it was definitely not suitable for all types of surgery.

As laparoscopy became more popular, clinical practice guidelines for laparoscopic surgery were developed as concise instructions to assist practitioners and patients. These guidelines improved both the quality and process of care and outcomes of treatments [7]. They could be referenced in policy making

but they did not have any legal implications. Whilst historically conceived from conferences and expert panels [8], as we approach the era of evidence-based medicine, many countries and medical organisations felt the need for a more reliable system to assess the evidence and recommendations. This was because the main downside of a guideline was that it could be wrong, due to the lack of or misinterpretation of evidence [7]. Different recommendation systems were developed to address these needs: they all had their respective strengths and weaknesses.

The two major systems of recommendations were Grading of Recommendations Assessment, Development and Evaluation (GRADE) and Oxford Methodology. The GRADE system was introduced to the medical field in 2004 [9], and it has been influential since, being endorsed by large medical organisations including World Health Organisation (WHO) and National Institute of Health and Care Excellence (NICE) [9]. The GRADE system was developed with the purpose of grading clinical evidence and creating guidelines suitable for clinical practice based on the evidence [10]. The Oxford Methodology system on the other hand, was updated in March 2009 by experts in the field of evidence-based medicine, to classify and grade recommendations for treatments and diagnostic tests [11]. The general purpose of this system was to offer clinical advice but at the same time make sure the people considering the information were aware of the flaws in the evidence [12]. Other systems include the Scottish Intercollegiate Guidelines Network (SIGN) which aimed to make it easier for doctors to identify the link between evidence and recommendations when reviewing guidelines [13]. Similar to the GRADE system, it was endorsed by NICE and contribute to UK national policies [14]. There were also a lot of guidelines with their own systems to grade their recommendations. The plethora of systems brings the potential of confusion due to variation of recommendation; thus there is a need to assess whether different recommendation systems would produce a different recommendation for a laparoscopic guideline. It is also important to assess the extent of variation and whether this variation could be explained by the guideline authors not complying with the methods used for grading the statements.

Methods

A purposive sampling was performed by searching PubMed for guidelines published in the last 10 years and for which full text was available using the following terms: 'laparoscopy'[MeSH], 'Laparoscopic' [Free text] AND 'guideline' [Free text]; 'guidelines as topic'[MeSH], ('laparoscopic'[Title] OR 'Laparoscopic' [Title]) AND ('guideline' [Title] OR 'guidelines' [Title] OR 'Guideline' [Title] OR 'Guidelines' [Title]). This retrieved 49 papers. Four guidelines that met the following criteria below were selected.

The criteria for the choosing the guidelines were as follows:

- Must relate to laparoscopic surgery
- Different recommendation systems (in order to evaluate whether mismatch was different for the different recommendation systems)
- One of the guidelines must use GRADE (to assess the compliance of the guideline author)
- One of the guidelines must use Oxford Methodology (to assess the compliance of the guideline author)
- The guidelines did not use different systems for different statements

The results were graded as 'strong' or 'weak' by considering the following four factors described in the guideline:

- Balance between desirable and undesirable effects: we looked at variability in importance, baseline risks and relative/absolute effects
- Quality of the evidence [15]
- Costs or resources utilised
- Values and preferences which include variability and absence of information

We then evaluated whether the guideline developers would have arrived at the grade of recommendation as our grade of recommendation, if they had used GRADE rather than the recommendation system they had originally used.

To account for the variability in the terms used in the different guideline grading systems, we converted the terms to equivalent terms (see Table 1), based on the explanation about interpretation as provided in the guidelines.

Two authors (JL and AN) independently reclassified each statement in these guidelines. We assessed the agreement between the two authors by evaluating the proportion of agreement in strong and weak statements and the interrater reliability by using kappa correlation coefficient [16] using GraphPad software. The proportion of agreement between the guideline authors' classification and our consensus classification was also done (see Table 3). We then resolved any differences by discussion.

Results

The following four guidelines were chosen:

- The Clinical Practice Guidelines for Laparoscopic Hysterectomy for Benign Indications developed by the Dutch Society of Obstetrics and Gynecology (NVOG). The recommendations are unclear but they are all implied as strong [17].
- The Guidelines for Laparoscopic Treatment of Ventral and Incisional Abdominal Wall Hernias (International Endohernia Society [IEHS])—Part 1 was developed using the Oxford hierarchy of evidence (Oxford Methodology) [18].

Table 1 Shows the conversion of guidelines recommendations to GRADE recommendations

Guideline recommendation system	Guideline recommendation	GRADE recommendation
GRADE [17]	Strong recommendation	Strong recommendation
	Weak recommendation	Weak recommendation
Oxford Methodology [18]	Grade A	Strong recommendation
	Grade B	Strong recommendation
	Grade C	Weak recommendation
	Grade D	Weak recommendation
SIGN [19]	Strong	Strong recommendation
	Recommended best practice based on the clinical experience of the guideline development group	Strong recommendation
	Conditional	Weak recommendation
SAGES guidelines 'Bespoke' [20]	Grade A	Strong recommendation
	Grade B	Strong recommendation
	Grade C	Weak recommendation

Table 2 Shows the concordance between the two authors who classified the information independently

	Second author: Strong recommendation	Second author: Weak recommendation
First author: Strong recommendation	50	3
First author: Weak recommendation	9	129

- The Southampton Consensus Guidelines for Laparoscopic Liver Surgery used a unique approach to develop and grade the guidelines. Firstly, they used SIGN to assess and develop the statements. Secondly, they used the Delphi method for expert consensus. Finally, the AGREE-II tool was used for validating the statements [19].
- The Guidelines for the Clinical Application of Laparoscopic Biliary Tract Surgery is part of the SAGES guidelines. They have their own system for grading recommendation [20].

In the four guidelines that we selected, there were 191 statements in total.

Table 2 shows the concordance of the strengths of recommendation between JL and AN. By consensus, 57 statements were strong recommendations and 134 statements were weak recommendations. Overall, there was agreement on the strength in 92.6% of strong recommendations and 94.1% of the weak recommendations. Overall, there was agreement in 93.7% of the statements. Interrater reliability was Cohen's kappa = 0.849 (95% confidence intervals 0.766–0.931) and the strength of agreement between the authors was considered to be 'strong' or near perfect reliability [21].

The agreement between the study authors and our consensus grading of recommendation was variable. As shown in Table 3, the guideline that used the GRADE system had the lowest proportion of statements with the same strength of recommendation made between the guideline authors and us: there was only 24.1% concordance in the strength of the recommendations between the guideline authors and us. The guideline that used the Oxford Methodology system had the highest concordance: the concordance was 62.2%. The concordance in the guidelines that used the SIGN and 'bespoke' system were 35.8% and 50.0% respectively.

The classification of each statement in the guideline and the reason for our classification is provided in online supplement 1.

Discussion

This study shows that there were differences in the grade of recommendations made by us and the guideline developers. We have provided a detailed table (online supplement 1) to demonstrate the reasons for the difference. In total, there were four guidelines and 191 statements: 57 were strong recommendations and 134 were weak recommendations. The strength of agreement between the first and second author was 93.7% and interrater reliability as assessed by Cohen's kappa was 0.849, considered to be strong to near perfect reliability [21]. The discordances were mainly due to the fact that there were statements with borderline recommendations of strong or weak: the evidence was judged differently by JL and AN. The overall agreement with each guideline with the consensus grading of recommendation between the two authors were the lowest with the GRADE and SIGN system, which were 24.1% and 35.8% respectively. We did not obtain the results that we initially expected, as one would naturally assume that there would be a high, if not very high agreement in the strength of the recommendations on the GRADE system after re-evaluation. This suggested that the GRADE system of recommendation itself was not used as per the guidance document. On the contrary, the guidelines that used Oxford Methodology and the 'Bespoke' system (SAGES Guidelines for the Clinical Application of Laparoscopic Biliary Tract

Table 3 Agreement between the guideline authors' classification and our consensus classification

Guideline	Recommendation system	Total statements	Number where there was agreement between guideline authors and consensus agreement	Proportion classified correctly (%)
The Clinical Practice Guidelines for Laparoscopic Hysterectomy for Benign Indications	GRADE	29	7	24.1
Guidelines for laparoscopic treatment of ventral and incisional abdominal wall hernias (International Endohernia Society [IEHS])—Part 1	Oxford Methodology	45	28	62.2
Southampton Consensus Guidelines for Laparoscopic Liver Surgery	SIGN	67	24	35.8
The SAGES Guidelines for the Clinical Application of Laparoscopic Biliary Tract Surgery	Bespoke	50	25	50.0

Surgery) had more recommendations with matching strength than different strength when re-evaluated using GRADE, which was 62.2% and 50.0% respectively, mainly due to the fact that that the guideline developers complied better with the criteria of the recommendation system, and that the system included a quality of evidence section similar to GRADE: the lowest level in both the systems would always be inadequate evidence and advises the physician to find alternatives [17, 18, 20].

There are two potential reasons for the differences in agreement between the study authors and us. One possibility is that the guideline developers might have appraised the guidelines correctly but decided to not present the information in a transparent manner. Alternatively, the guideline developers did not fully comply with the judging criteria of the recommendation system they supposedly used. In fact, it is common for organisations to tailor the criteria of the recommendations, which contribute to inconsistencies when grading guidelines [22]. The main reason causing the high number of mismatch of recommendations in the GRADE system and SIGN system was because the guideline developers did not fully comply with the judging criteria of the recommendation system they supposedly used. This can be seen as expert opinion graded as a strong recommendation accounting for 63.6% and 43.9% in the guidelines that used GRADE and SIGN respectively as shown in the online supplement table. This further echoed the criticism of GRADE as to whether GRADE was logical at all [20]. GRADE was meant to 'separate the judgements regarding the quality of evidence from judgements about the strength of recommendation' but some guideline statements were graded as strong recommendation despite low quality of evidence.

The goal of this is an exploratory study was to determine whether the grade of recommendations in the laparoscopy guidelines are reliable. Therefore, we performed only a purposive sampling of four guidelines, which contained a total of 191 guideline statements. This revealed that there may be significant differences in the grading done based on the information provided by the guideline author compared to that performed by the author. In order to find the true extent and reasons of non-transparent or incorrect grading of recommendation, a systematic review has to be performed. In such a systematic review, it is possible to explore whether there is a relationship between non-transparent or incorrect grading of recommendation and AGREE-II tool, specifically, rigor of development and clarity of presentations domains.

Conclusion

There are a wide variety of recommendation systems for laparoscopic guidelines and the recommendations from the guidelines using GRADE, Oxford Methodology, SIGN and the 'Bespoke' system varied when re-evaluated by GRADE. A systematic review of recent laparoscopic guidelines might provide the extent and the main reasons of the problem.

Compliance with ethical guidelines

Conflict of interest J. Leung, A. Ng and K. Gurusamy declare that they have no competing interests.

Ethical standards This is a study on research methodology of guidelines, thus no inclusion of people for the study. All the studies cited by the guidelines also did not include any people for the study.

References

1. Taylor E, Thomas J, Whitehouse L, et al. Population-based study of laparoscopic colorectal cancer surgery 2006–2008. Br J Surg. 2013;100(4):553–60.

2. Kang C, Halabi W, Luo R, et al. Laparoscopic colorectal surgery. Arch Surg. 2012;147(8):724–31.
3. Spanjersberg W, van Sambeeck J, Bremers A, et al. Systematic review and meta-analysis for laparoscopic versus open colon surgery with or without an ERAS programme. Surg Endosc. 2015;29(12):3443–53.
4. Bemelman W, Warusavitarne J, Sampietro G, et al. ECCO-ESCP Consensus on Surgery for Crohn's Disease. J Crohns Colitis. 2018;12(1):1–16.
5. Di Saverio S, Birindelli A, Kelly M, et al. WSES Jerusalem guidelines for diagnosis and treatment of acute appendicitis. World J Emerg Surg. 2016;11:34. https://doi.org/10.1186/s13017-016-0090-5. eCollection 2016.
6. Deijen C, Vasmel J, de Lange-de Klerk E, et al. Ten-year outcomes of a randomised trial of laparoscopic versus open surgery for colon cancer. Surg Endosc. 2016;31(6):2607–15.
7. de Rooij T, van Hilst J, Bosscha K, et al. Minimally invasive versus open pancreatoduodenectomy (LEOPARD-2): study protocol for a randomized controlled trial. Trials. 2018;19(1):1. https://doi.org/10.1186/s13063-017-2423-4.
8. Woolf S, Grol R, Hutchinson A, et al. Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. BMJ. 1999;318(7182):527–30.
9. Grade Working Group. What is GRADE? 2019. http://www.gradeworkinggroup.org. Accessed 19 Dec 2019.
10. Grade Working Group. Grading quality of evidence and strength of recommendations. BMJ. 2004;328(7454):1490.
11. Grondin S, Schieman C. Evidence-based medicine. In: Ferguson M, editor. Levels of evidence and evaluation systems. Difficult decisions in thoracic surgery. London: Springer; 2010. pp. 13–22.
12. Howick J. Oxford centre for evidence-based medicine—Levels of evidence (March 2009). 2009. https://www.cebm.net/2009/06/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/. Accessed 19 Dec 2019.
13. Harbour R, et al. A new system for grading recommendations in evidence based guidelines. BMJ. 2001;323:334.
14. Baird AG, Lawrence JR. Guidelines: is bigger better? A review of SIGN guidelines. Bmj Open. 2014;4:e4278.
15. Guyatt G, Schünemann H, Brożek J, et al. GRADE Handbook. 2013. https://gdt.gradepro.org/app/handbook/handbook.html#h.wsfivfhuxv4r. Accessed 21 Dec 2019.
16. GraphPad. Confidence intervals of proportions were calculated using the GraphPad QuickCalcs Website. 2019. https://www.graphpad.com/quickcalcs/kappa1/. Accessed 22 Dec 2019.
17. Sandberg E, Hehenkamp W, Geomini P, et al. Laparoscopic hysterectomy for benign indications: clinical practice guideline. Arch Gynecol Obstet. 2017;296(3):597.
18. Bittner R, Bingener-Casey J, Dietz U, et al. Guidelines for laparoscopic treatment of ventral and incisional abdominal wall hernias (International Endohernia Society [IEHS])—Part 1. Surg Endosc. 2013;28(1):2–29.
19. Abu Hilal M, Aldrighetti L, Dagher I, et al. The Southampton Consensus Guidelines for Laparoscopic Liver Surgery. Ann Surg. 2018;268(1):11–8.
20. Overby D, Apelgren K, Richardson W, et al. SAGES guidelines for the clinical application of laparoscopic biliary tract surgery. Surg Endosc. 2010;24(10):2368–86.
21. McHugh M. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276–82.
22. Guyatt G, Oxman A, Akl E, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383–94.