TITLE PAGE

Original Study: Development of a Reliable Surgical Quality Assurance System for Two-stage Esophagectomy in Randomized Controlled Trials

AUTHORS

Alexander Harris PhD[1]

James Butterworth MSc[1]

Piers R Boshier PhD[1]

Hugh MacKenzie PhD[1]

Masanori Tokunaga PhD[2]

Hideki Sunagawa PhD[3]

Stella Mavroveli PhD[1]

Melody Ni PhD[1]

Sameh Mikhail MD[4]

Chi-Chuan Yeh PhD[5]

Natalie S Blencowe PhD[6,7]

Kerry N L Avery PhD[7]

Richard Hardwick MD[8]

Arnulf Hoelscher PhD[9]

Manuel Pera PhD[10]

Giovanni ZaninottoMD[1]

Simon Law PhD[11]

Donald E Low MD[12]

JJB van Lanschot PhD[13]

Richard Berrisford ChM[14]

C Paul Barham MD[6]

Jane M Blazeby MD[6,7]

George B Hanna PhD[1]

[1] *Department of Surgery & Cancer, Imperial College London, United Kingdom*

[2] *Department of Gastrointestinal Surgery, Tokyo Medical and Dental University, Japan*

[3] *Department of Gastroenterological Surgery, New Tokyo Hospital, Japan*

[4] *Department of General Surgery, Faculty of Medicine University of Cairo, Egypt*

[5] *Department of Surgery, National Taiwan University Hospital, Taiwan*

[6] *Division of Surgery, University Hospitals Bristol NHS Foundation Trust, United Kingdom*

[7] *National Institute for Health Research Bristol Biomedical Research Centre, University of Bristol, United Kingdom*

[8] *Upper gastrointestinal Unit, Cambridge University Hospitals NHS Foundation Trust, United Kingdom*

[9] *Center for Esophageal and Gastric Surgery, Agaplesion Markus Hospital, Germany*

[10] *Department of Surgery, Hospital del Mar, Spain*

[11] *Department of Esophageal and Upper Gastrointestinal Surgery, The University of Hong Kong Queen Mary Hospital, Hong Kong*

[12] *Department of Thoracic Surgery, Virginia Mason Medical Center, USA*

[13] *Department of Surgery, Erasmus MC University Medical Center, The Netherlands*

[14] *Department of Surgery, University Hospitals Plymouth NHS Trust, United Kingdom*

ACKNOWLEDGEMENTS

The authors would like to acknowledge the input of Dr Chris Metcalfe, Professor of Medical Statistics & Co-director Bristol Randomized Trials Collaboration, with this research. We are grateful to Professor Mitsuru Sasako (Japan) and Professor David Farley (USA) for facilitating the interviews and task analysis conducted at the Hyogo College of Medicine in Japan and the Mayo Clinic in Rochester, USA. We also thank Jane M Blazeby, C Paul Barham, Dan Titcomb, Christopher Streets, Andrew Hollowood, Richard Krysztopik, Richard Berrisford, Tim Wheatley, and Grant Sanders who submitted operative data during the pilot phase of the study.

Corresponding Author

Professor George Hanna,

Department of Surgery and Cancer,

Imperial College London,

10th Floor QEQM Building,

St Mary's Hospital,

London, W2 1NY UK

E-mail: g.hanna@imperial.ac.uk

Telephone No: +44 (0)207 886 2125

Fax No: +44 (0)207 886 2125

Re-prints will not be available from the corresponding author.

SOURCES OF SUPPORT

NIHR-HTA Grant 10/50/65, Randomised Oesophagectomy – Minimally Invasive or Open (ROMIO) Trial.

London Deanery Simulation and Technology enhanced Learning Initiative (STeLI).

The Great Britain Sasakawa Foundation.

This work was undertaken with the support of the Medical Research Council ConDuCT-II (Collaboration and innovation for Difficult and Complex randomized controlled Trials In Invasive procedures) Hub for Trials Methodology Research (MR/K025643/1) (http://www.bristol.ac.uk/social-community-medicine/centres /conduct2/) and Royal College of Surgeons of England Bristol Surgical Trials Centre. KNLA and JMB were supported by the National Institute for Health Research (NIHR) Biomedical Research Centre (BRC) at the University Hospitals Bristol NHS Foundation Trust and the University of Bristol. JMB holds an NIHR Senior Investigator award. The NIHR Imperial Biomedical Research Centre also provided support and infrastructure. PB is a NIHR Lecturer. The views expressed are those of the authors and not necessarily those of the MRC, UK National Health Service, National Institute for Health Research or Department of Health and Social Care.

RUNNING HEAD

Surgical quality assurance in trials

INTRODUCTION

Surgery remains the cornerstone in the multimodality management of esophageal cancer. The outcome of esophagectomy is surgeon-dependent with wide variability in surgical technique and perioperative management. Systematic reviews of open and minimally invasive surgery have shown that formal lymphadenectomy was not performed in most studies and lymph node harvest fell below the minimum number recommended to achieve survival benefits [1, 2]. The problem of surgical variability is heightened in randomized controlled trials (RCTs) where the quality of surgery could influence the final outcome and might compromise the generalizability of results [3, 4]. A systematic review assessing surgery within RCTs for the treatment of esophagogastric cancer demonstrated significant heterogeneity in study design and quality assurance [5]. Surgical quality indicators identified in this systematic review were: pre-trial standardization of surgical technique, credentialing of surgeons to enter into the trial, and monitoring of surgical performance. Those measures influence in-hospital mortality, the quality of lymphadenectomy, and loco-regional recurrence.

Surgical quality assurance (SQA) was developed within the context of the pilot phase of the multicenter Randomised Oesophagectomy - Minimally Invasive or Open (ROMIO) trial, which is comparing minimally invasive esophagectomy with open surgery [6], where surgery is the trial intervention within a multimodality approach to the treatment of esophageal cancer. The specific objectives were to standardize the performance of two-stage esophagectomy and develop a competency assessment tool for trial. The intended deliverables were an operation manual and reliable competency assessment tools.

METHODS

(i) Standardization of two-stage esophagectomy

Semi-structured interviews and structured observations

Semi-structured interviews and structured observations [7] were performed with surgeons at specialist centers for esophagogastric cancer surgery in the United Kingdom, United States of America and Japan (Glossary of terms used provided in Supplemental Digital Content 1). These investigated similarities and variations in clinical practice of two-stage esophagectomy performed worldwide. Interviews were digitally audio-recorded (with consent) before being transcribed, checked for accuracy, and qualitatively analyzed using Thematic Analysis [8, 9]. For contingency, shorthand written records were also kept in case of digital data loss. Structured observations were written in a research diary kept by the primary researcher (AHa). A second researcher (PB) assisted with intraoperative data collection in Japan and the United Kingdom. A debrief held at the end of each observed operation permitted comparison of notes between researchers, with video recordings of selected operations used to support further in-depth analysis remotely.

Hierarchical task analysis

Findings from the published literature and on-line digital media were combined with Thematic Analysis of the semi-structured interviews and structured observations in order to create a hierarchical task analysis (HTA) for two-stage esophagectomy [10]. Several iterations were written and revised in consultation with a panel of senior

esophagogastric cancer surgeons. The accuracy of each HTA was verified against live and video recorded two-stage esophagectomies performed by senior esophagogastric cancer surgeons until no additional changes were identified. The final HTA was then tested against a series of subsequent operations, for which the primary researcher was present as an observer.

Delphi consensus process for the esophagectomy HTA

Ten peer-nominated expert esophagogastric cancer surgeons were electronically invited to participate in a Delphi consensus process [11]. This method was selected to ensure the underlying face and content validity of the final assessment tool, as well as surgeon acceptance of the procedural HTA. Esophagogastric cancer surgeons involved during the development of the HTA were excluded. In total, nine surgeons consented to participate in the Delphi process.

In the first Delphi round, each surgeon was provided with the final HTA and a questionnaire where they were requested to rate the requirement for each of the defined steps as either *mandatory*, *optional*, or *prohibited* [12]. This method of rating was in accordance with the requirements of the ROMIO trial protocol [6]. Additional free space was available for comments to be made regarding each individual step within the HTA. Completed questionnaires were returned electronically and analyzed by the primary researcher. An arbitrary consensus agreement of 75% was sought for each step [13].

This process was repeated in the second Delphi round, during which the nine respondents from the first Delphi round were electronically sent the percentage agreement and anonymized comments for each step of the HTA. The original HTA was

re-sent, along with a new Delphi questionnaire. If respondents' answers remained outside of the majority agreement, they were asked to provide reasons for this in the comments section. For steps where consensus agreement could not be reached by the end of the second Delphi round, it was deemed acceptable for a majority opinion to be upheld if it reflected the findings of the HTA and its original evidence base (i.e. triangulation of the published literature, semi-structured interviews, and structured observations).

(ii) Development of an operation manual and note

A complete operation manual was constructed for surgeons based on the Delphi consensus approved HTA. The key operative steps for open and minimally invasive two-stage esophagectomy were identical. Each step of the operation, both mandatory and optional, was described in detail with photographs illustrating the required en-bloc lymphadenectomy.

Given the length of the full manual, a separate summary document describing ten essential steps for the abdominal and thoracic phases of the operation was also produced. The operation manual and summary of essential steps were approved by the ROMIO pilot phase steering committee.

A standardized operation note was constructed for two-stage esophagectomy in an iterative process. It was designed to reflect the clinical requirement of providing a formal record of the operation performed and as a requirement of SQA. The body of the operation note included a tick box version of the operation manual, permitting surgeons to rapidly provide a detailed outline of the procedure performed, with white space boxes available for additional information.

(iii) Development of video and photographic assessment tools

The details for image capture and data transfer are supplied in Supplemental Digital Content 2.

Video assessment tool

In accordance with the Systems Engineering Initiative for Patient Safety model that describes a *structure-process-outcome* approach [14], the results obtained from the semi-structured interviews and observations confirmed the importance of technical performance (*process*) and oncological quality (*outcome*) of the operation. Following consideration of techniques that could permit independent remote blind evaluation of the technical performance and oncological quality of surgery, it was determined that a video assessment tool would best address all of these aspects.

An existing validated, consultant-level, surgical assessment tool [15, 16] was deconstructed and its underlying principles adapted during the structural development of this video assessment tool. Elements relevant to the safety and efficiency of the operative process, as well as the oncological quality of the end product, were identified from Thematic Analysis of the semi-structured interviews and structured observations such that clear definitions for each of the terms used were composed.

Several different video assessment tools were written and piloted at St Mary's Hospital, London, UK, over the course of three months. Each version placed a different emphasis on rating the element being assessed, with the intention to balance the operating surgeon's technical safety, efficiency and oncological quality of their dissection in the final tool. (Please note that tasks have been labelled differently in the HTA and

assessment tool, as the research evolved). The video assessment tool was approved for use by the ROMIO pilot phase steering committee and tested at the two centers involved in the pilot RCT.

Photographic assessment tool

The outcome section of the video assessment tool was purposefully written as a stand-alone photographic assessment tool, focusing on the completeness of the lymphadenectomy and exposure of the relevant anatomical landmarks, should video submission have not proved feasible. The photographic tool was also approved for use by the ROMIO pilot phase steering committee and piloted as above.

(iv) Examining reliability of the video and photographic assessment tools

Independent assessment of video and photographic records from the ROMIO Trial

Three esophagogastric cancer surgeons (one based in the UK and two in Japan) were invited to assess and rate the intraoperative videos and photographs submitted by surgeons within the pilot ROMIO trial. Prior to commencing data analysis, the three assessors were trained by the senior author (GH) in two videoconference meetings on the pre-defined terms for using the assessment tools and to clarify any conceived variability. Prior to the second videoconference, each assessor was asked to independently rate two videos that had been chosen at random. These assessments formed a focal point for the discussion held during the second videoconference in order to minimize the discrepancy in assessments.

These three independent blinded assessors then applied the video and photographic esophagectomy assessment tools to rate each of the video and photographic records

submitted to the pilot ROMIO trial. Assessments were completed on paper forms, which were subsequently submitted as a scanned PDF file and later transcribed into Excel (Microsoft office, Redmond, WA, USA).

Statistical analysis

Generalizability (G) theory was used to assess the reliability of the video and photographic assessment tools because, in contrast to the classical test theory, G-theory includes several aspects of reliability (e.g. inter-rater, inter-test, and intra-test) in the same model. A decision (D) study was performed to determine the combination of components that yielded the maximum generalizability [17]. G-string software was used to conduct the generalizability theory, inter-rater reliability and internal consistency analysis [18]. Cronbach's alpha for internal consistency was performed using IBM SPSS statistics (Ver. 24, SPSS Inc., Chicago, IL, USA) as part of cross-validation.

RESULTS

(i) Standardization of two-stage esophagectomy

Semi-structured interviews and structured observations

In total, eight separate semi-structured interviews were performed with surgeons from the UK (n=6) and the USA (n=2). Themes arising from the qualitative analysis of these interviews are summarized in Supplemental Digital Content 3. Greater than fifty esophagectomies, performed by sixteen surgeons from the UK (n=9), USA (n=6), and Japan (n=1), were observed in seven different hospitals. Structured observation notes

were combined with findings from the “operative procedure” theme identified from the semi-structured interviews and incorporated into the HTA.

Hierarchical task analysis

Full details of the HTA are provided in Supplemental Digital Content 4. The abdominal component of two-stage esophagectomy comprised seven tasks, and the thoracic component comprised six tasks. Each task was then sub-divided into multiple steps. Overall, fifty-four steps were identified.

Delphi consensus process

Full details of the Delphi consensus process are provided in Supplementary Digital Content 5. In round one of the Delphi consensus process, nine of the ten invited surgeons responded representing the UK, Germany, Spain, Italy, the Netherlands, the USA and Hong Kong. They reached consensus (75%) agreement on forty of the fifty-four steps comprising the HTA for esophagectomy according to whether each task was mandatory, optional or prohibited.

In round two of the Delphi consensus process, the same nine surgeons who had previously responded were provided with the results from the first round and asked to re-rate each task. This time, only six surgeons responded in the allotted timeframe. Thirty-nine tasks reached consensus agreement. Seven of the fifteen tasks without consensus agreement were the same in both rounds.

Given the diminishing number of responses, it was felt that a third round would not be beneficial. Furthermore, it was not logistically possible to gather all invited surgeons in person, or via conference call, to discuss their decisions. Accordingly, for tasks that did

not reach consensus agreement in round two, a majority decision was upheld if it reflected the findings of the evidence-based HTA. Thirteen of the fifteen tasks without agreement had a majority decision in the second round. Two of the fifteen tasks were equally split by respondents in the second round, but had had a majority decision in round one. The earlier majority decision (based on a greater number of respondents) was therefore followed.

(ii) Development of an operation manual and note

A summary describing ten essential steps for both the abdominal and the thoracic phases of the operation can be found in Table 1. The final versions of the operation manual and note are provided in Supplemental Digital Contents 6 and 7.

(iii) Development of video and photographic assessment tools

The video and photographic assessment tools are shown in Figures 1 and 2 respectively.

(iv) Examining reliability of the video and photographic assessment tools

31 videos and 53 photographic series from patients undergoing two-stage esophagectomy were submitted for assessment. The length of submitted video recordings varied widely, ranging from 1.0 minute to 447.0 minutes. In total, 4464.3 minutes of video footage were received, with a median duration of 119 minutes submitted per esophagectomy. Photographic submissions ranged from 2 images to 35 images per esophagectomy. In total, 451 images were received, with a median of 9 photographs per case.

Despite a large volume of data being submitted, the three assessors identified that there was also a significant amount of data missing. Following an interim review, a videoconference between the three reviewers explored the possible reasons for missing data and potential strategies to mitigate its impact on data analysis. The original three-point lymphadenectomy rating system of *complete*, *incomplete* and *not performed* was deemed to be inadequate. Alternative solutions that were considered included making missing assessment values a mean value or coding them as *not performed* [19]. However, assessors were concerned that this would introduce bias and skew results. There was consensus that two additional categories could be used to re-code those parts of the assessment in which assessors were unable to provide a rating. The new categories acknowledged *insufficient evidence* for assessors to provide a rating (e.g. videos with an obstructed field of view or blurred photographs) and *absent data* (i.e. no video or photograph submitted). Overall, 32.3% of video and photographic data were absent. 6.8% of video data were insufficient for assessors to provide a rating, compared with 23.4% of photographic series.

G-theory results for the video assessment tool

Generalizability analyses were performed to evaluate reliability of the 35-item video assessment tool with a fully crossed design using videos (V), items (I) and assessors (A), such that (V x I x A). In total, 93 assessment forms (31 videos rated by 3 assessors) of the 35-item video assessment tool were used in the analysis. Raw scores of the 35-item video assessment tool were generalized over the assessor and item. Overall reliability of video assessment was represented by a generalizability coefficient of G(AI) = 0.744. D-studies were performed to examine the effect of increasing numbers of assessors and video

esophagectomies that they assessed (Figure 3). The critical G coefficient of 0.8 was reached with 4 assessors each rating 26 video esophagectomies or 6 assessors each rating 16 video esophagectomies.

G-theory results for the photographic assessment tool

To evaluate the reliability of the 27-item photographic assessment tool, a fully crossed design using photographs (P), items (I) and assessors (A) such that (P x I x A) was used.

In total, 159 ratings (53 sets of operative photographs rated by 3 assessors) of a 27-item photographic assessment tool were used in the analysis. Raw scores of the 27-item photographic assessment tool were generalized over the assessor and item. Overall reliability of photographic assessment was represented by a generalizability coefficient of G(AI) = 0.700. D-studies were once again performed to examine the effect of increasing numbers of assessors and sets of photographs of esophagectomies that they assessed (Figure 4). The critical G coefficient of 0.8 was reached with 6 assessors each rating 38 sets of esophagectomy photographs or 8 assessors each rating 33 sets of esophagectomy photographs.

Generalizability coefficients were also calculated separately and demonstrated consistently high reliability coefficients within the video and photographic assessment tools respectively. It was noted that video assessment had consistently higher reliability coefficients compared to photographic assessment across all tasks (Table 2).

Inter-rater reliability and internal consistency

By treating one facet at a time as random, whilst fixing the other facets, it was possible to determine the equivalent of inter-rater reliability for the video assessment tool as: Ep^2

=0.492. By setting the item as random and the assessor as fixed for the video assessment tool, it was possible to determine the equivalent of internal consistency as Ep^2 =0.991, which was similar to the value calculated using SPSS with Cronbach's alpha 0.986.

Through the same process, the inter-rater reliability and internal consistency for the photographic assessment tool were calculated as Ep^2 = 0.438 and Ep^2 =0.948 respectively. Again, the internal consistency was similar to the value calculated using SPSS with Cronbach's alpha 0.942.

Further analysis was performed to determine redundancy of items within each of the video and photographic assessment tools. Cronbach's alpha remained constant at 0.986 on removal of any anatomical item from the video assessment tool, thereby demonstrating high inter-item reliability. In the photographic assessment tool, removal of the same anatomical items resulted in small improvements in Cronbach's alpha. However, this variation was attributed to the high occurrence of *absent data* or *insufficient evidence* within this cohort.

DISCUSSION

This SQA system has been developed to assess the anatomy and tissues that remain after oncological resection in order to complement histological examination of the removed specimen, including resection margins and lymph node yield, which are the traditional markers of surgical quality. The study describes the development of key components of a SQA system that defines the operative standard for two-stage esophagectomy and assesses operative competency. The deliverables were an operation manual and reliable video and photographic assessment tools for use within surgical oncology RCTs. Those

deliverables address the required SQA measures identified in two systematic reviews to standardize surgical techniques, credential surgeons before entry into RCTs and monitor performance during the trial [5, 20].

The output of the standardization process provided the structure for the operation manual and basis for the competency assessment tools. A set of mandatory and optional steps was specified as a guide for surgical performance and a framework to examine variability of task execution between surgeons during the trial. Categorizing steps into mandatory and optional tasks allows for flexibility in surgical performance whilst maintaining a minimum quality standard [12]. Seventeen international esophago-gastric surgeons with trial experience participated in the interviews, observations and Delphi consensus process in order to set a proposed standard for the trial. This international composition affords a degree of generalizability for the developed tools to be applied in esophageal RCTs worldwide. Nevertheless, this operative template may need to be modified by leading surgeons in specific trials to provide a balance between standardization and practicalities in exploratory and pragmatic trials. Although not ideal, the differences in surgical rigour between countries, cancer centers and trial designs are a reality in surgery. The real performance of surgeons in the trial will be the practical level against which trial outcomes will be judged and explained.

The developed tools have a high level of content and face validity as they are based on a hierarchical task analysis and a Delphi consensus process of surgical performance. Both video and photographic assessment tools have a high reliability score using generalizability theory. However, capturing video or photographic data during open esophagectomy presents a challenge. This research shows that a high proportion of video and photographic data was absent or insufficient to rate performance. The

operative time for esophagectomy and the potential intrusiveness of audio-visual recordings, given the restricted surgical access and limited operative field in open surgery, explains the challenge in capturing the image data. Clear instructions for data capturing as well as adequate resources and strong engagement from participating surgeons are required. An anecdotal observation from fieldwork performed in Japan, was how motivated surgeons were to capture high-quality video and photographic evidence of the procedure.

An alternative approach that could be explored is a short video recording, performed at key stages of the operation, in order to demonstrate the extent of the dissection and characteristics of the reconstruction. The benefits of this approach would be the avoidance of long video recordings and the frequent inadequacies of photographic images. The short recording would allow a dynamic snapshot of the operative field, permitting visualization of anatomical structures from multiple angles, as well as better assessment of conduit health and tension at the anastomosis. A limitation of such an approach would be the inability to assess the process, including safety and efficiency of operative tasks, as it would only show the quality of the end product. In addition, a feedback system will be developed for surgeons participating in the trial [21]. Nevertheless, the adoption of surgical assessment tools would be enhanced by overcoming potential practical challenges in routine practice that create the perception of being impractical and time consuming. Provision of measures that facilitate the ease and convenience of video/photo capture, sharing and assessment is critical for the uptake of SQA in clinical trials. Recording systems and instructions for imaging the operative field should be provided and tested at the outset of the trial. Future iterations of the tool could be hosted on a web-based platform to support the exchange and

assessment through digital media. The time required for assessment could be shortened by limiting the assessment to short videos and/or photographs of the operative field at the end of the procedure. The ROMIO trial does however provide an opportunity to examine the tools' practicality and to consider any changes required to make such an approach more feasible.

The study has several limitations. The development of SQA was not based on clinical outcomes, but on a hierarchical task analysis of surgical procedures and a consensus view that is constrained by the definition and selection of surgical expertise. However, in the absence of established SQA methods, it is reasonable to start the process with observational data and expert consensus. The Delphi methodology has advantages not observed in other traditional qualitative methods [22]. Whilst anonymity was preserved across panel members and only the primary researcher could identify the responses, the expert nomination process could have biased the results.

In conclusion, a reliable surgical quality assurance system for two-stage esophagectomy has been developed for surgical oncology randomized controlled trials. Key components of SQA include standardization of two-stage esophagectomy and assessment of competent performance. The predictive clinical validity of these assessment tools is still to be examined.

REFERENCES

1. Hanna, G., S. Arya, and S. Markar, *Variation in the standard of minimally invasive esophagectomy for cancer - systematic review.* Seminars in Thoracic and Cardiovascular Surgery, 2012. **24**: p. 176-187.
2. Boshier, P., O. Anderson, and G. Hanna, *Transthoracic versus transhiatal esophagectomy for the treatment of esophagogastric cancer: a meta-analysis.* Annals of Surgery, 2011. **254**: p. 894-906.
3. Macdonald, J., et al., *Chemoradiotherapy after surgery compared with surgery alone for adenocarcinoma of the stomach or gastroesophageal junction.* New England Journal of Medicine, 2001. **345**: p. 725-730.
4. Macdonald, J., et al., *Postoperative combined radiation and chemotherapy improves disease-free survival and overall survival in resected adenocarcinoma of the stomach and GE junction. Results of intergroup study INT-0116 (SWOG 9008).* European Journal of Cancer, 2001. **37**(S6): p. S10.
5. Markar, S., et al., *Assessment of the quality of surgery within randomised controlled trials for the treatment of gastro-oesophageal cancer: a systematic review.* Lancet Oncology, 2015. **16**: p. e23-31.
6. Avery, K., et al., *The feasibility of a randomized controlled trial of esophagectomy for esophageal cancer - the ROMIO (Randomized Oesophagectomy: Minimally Invasive or Open) study: protocol for a randomized controlled trial.* Trials, 2014. **15**: p. 200.
7. Marshall, C. and G. Rossman, *Designing qualitative research*. 2nd ed1995, London: SAGE Publications.

8. Flick, U., *An introduction to qualitative research*. 4th ed2009, London: SAGE publications.

9. Braun, V. and V. Clarke, *Using thematic analysis in psychology.* Qualitative research in psychology, 2006. **3**: p. 77-101.

10. Sarker, S., et al., *Constructing hierarchical task analysis in surgery.* Surgical Endoscopy, 2008. **22**: p. 107-111.

11. Palter, V., H. MacRae, and T. Grantcharov, *Development of an objective evaluation tool to assess technical skills in laparoscopic colorectal surgery: a Delphi methodology.* The American Journal of Surgery, 2011. **201**: p. 251-259.

12. Blencowe, N., et al., *Standardizing and monitoring the delivery of surgical interventions in randomized clinical trials.* British Journal of Surgery, 2016. **103**: p. 1377-1384.

13. Vernon, W., *The Delphi technique: A review. .* International Journal of Therapy and Rehabilitation, 2009. **16**: p. 69-76.

14. Carayon, P., et al., *Work system design for patient safety: the SEIPS model.* Qual Saf Health Care, 2006. **15 (suppl I)**: p. i50-58.

15. Miskovic, D., et al., *Observational clinical human reliability analysis (OCHRA) for competency assessment in laparoscopic colorectal surgery at the specialist level.* Surgical Endoscopy, 2012. **26**: p. 796-803.

16. Miskovic, D., et al., *Is competency assessment at the specialist level achievable? A study for the national training programme in laparoscopic colorectal surgery in England.* Annals of Surgery, 2013. **257**: p. 476-482.

17. Bloch, R. and G. Norman, *Generalisability theory for the perplexed: A practical introduction and guide: AMEE Guide No 68.* Medical Teacher, 2012. **34**: p. 960-992.

18. *Generalisability Theory*. [cited 1st December 2017; Available from: http://fhsperd.mcmaster.ca/g_string/index.html.

19. Ware, J., et al., *Missing Data.* New England Journal of Medicine, 2012. **367**: p. 1353-1354.

20. Foster, J., et al., *Methods of quality assurance in multicenter trials in laparoscopic colorectal surgery: a systematic review.* Annals of Surgery, 2014. **260**: p. 220-229.

21. Blencowe, N., et al., *Protocol for developing quality assurance measures to use in surgical trials: an example from the ROMIO study.* BMJ Open, 2019. **9**: p. e026209.

22. Avella, J., *Delphi panels: Research design, procedures, advantages, and challenges.* International Journal of Doctoral Studies, 2016. **11**: p. 305-321.

Legends for Tables, Figures and Supplemental Files

Table 1.docx: Essential tasks for two-stage esophagectomy

Table 2.docx: Generalizability coefficients by task

Figure 1.docx: Video assessment tool

Figure 2.docx: Photographic assessment tool

Figure 3.docx: D-study for the video assessment tool

Figure 4.docx: D-study for the photographic assessment tool

Supplemental Digital Content 1.docx: Glossary of terms used

Supplemental Digital Content 2.docx: Image capture and data transfer techniques

Supplemental Digital Content 3.docx: Synopsis of Thematic Analysis of the semi-structured interviews

Supplemental Digital Content 4.docx: HTA for two-stage esophagectomy

Supplemental Digital Content 5.docx: Delphi consensus process results

Supplemental Digital Content 6.docx: Operation manual for two-stage esophagectomy

Supplemental Digital Content 7.docx: Operation note