

Reporting in Surgical Research

Inauguraldissertation
zur Erlangung der Würde eines Doktors der Philosophie

vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät der
Universität Basel

von Rachel Rosenthal aus Basel, Schweiz

Basel, 2015

Originaldokument gespeichert auf dem Dokumentenserver der Universität Basel
edoc.unibas.ch



Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung - Nicht kommerziell - Keine Bearbeitungen 4.0 International Lizenz“ (CC BY-NC-ND 4.0). Die vollständige Lizenz kann unter creativecommons.org/licenses/by-nc-nd/4.0/ eingesehen werden.

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von
Prof. Dr. M. Tanner, Prof. Heiner C. Bucher, Prof. Jan van der Meulen

Basel, den 24. Juni 2014

Prof. Dr. Jörg Schibler

Dekan



Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)

This is a human-readable summary of (and not a substitute for) the [license](#).

[Disclaimer](#)

You are free to:

Share — copy and redistribute the material in any medium or format

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:



Attribution — You must give **appropriate credit**, provide a link to the license, and **indicate if changes were made**. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.



NonCommercial — You may not use the material for **commercial purposes**.



NoDerivatives — If you **remix, transform, or build upon** the material, you may not distribute the modified material.

No additional restrictions — You may not apply legal terms or **technological measures** that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable **exception or limitation**.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as **publicity, privacy, or moral rights** may limit how you use the material.

Dedicated to my daughter Valérie

Table of Contents

Summary.....	i
Zusammenfassung.....	v
List of abbreviations.....	x
List of Tables.....	xi
List of Figures.....	xiii
Acknowledgement.....	xv
1. Introduction	1
1.1. Background.....	1
1.2. Aims and objectives	2
2. How to write a surgical clinical research protocol: literature review and practical guide	3
2.1. Introduction	5
2.2. Guide for developing a clinical study protocol for surgical interventions	11
2.3. Conclusion	29
3. The use of systematic reviews when designing and reporting surgical trials.....	30
3.1. Introduction	32
3.2. Methods	33
3.3. Results.....	33
3.4. Discussion	37
4. Completion and publication rates of surgical randomized controlled trials – an empirical study	38
4.1. Introduction	42
4.2. Methods	42
4.3. Results.....	45
4.4. Discussion	53
5. Reporting of adverse events in surgical trials: critical appraisal of current practice	57
5.1. Introduction	59
5.2. Methods	60
5.3. Results.....	62
5.4. Discussion	65
6. Definition and Classification of Intraoperative Complications (CLASSIC): Delphi study and pilot evaluation	70
6.1. Introduction	72
6.2. Methods	72
6.3. Results.....	75
6.4. Discussion	83
7. How to report multiple outcome metrics in virtual reality simulation	87
7.1. Introduction	89

7.2. Methods	90
7.3. Results.....	94
7.4. Discussion	106
8. Disregarding paired data in surgical research – evaluation of current practice and estimation of implications on study results.....	109
8.1. Introduction	110
8.2. Methods.....	110
8.3. Preliminary results	113
8.4. Outlook	113
9. General discussion and conclusion.....	114
9.1. Principal findings.....	114
9.2. Relation to other studies and possible mechanisms	115
9.3. Strengths and limitations.....	117
9.4. Implications for current practice	119
9.5. Implications for future research.....	121
9.6. Conclusions	122
10. References.....	123
11. Appendices	139

Summary

Transparent and standardized reporting in clinical research is a prerequisite for optimal healthcare decision-making. This applies to any clinical discipline, but may be specifically challenging, when complex surgical interventions are involved. There, the impact of surgeons' experience on the size of treatment effects, standardization of interventions and of outcome assessment need to be specifically addressed.

In a previous investigation on selective outcome reporting in surgical trials (Rosenthal and Dwan 2013), we found a remarkably high percentage of unexplained discrepancies between registry entries and final reports of randomized controlled trials (RCTs) that were published in high-impact surgical journals. In order to address problems of reporting in surgical research more in depth, I decided to dedicate my PhD to this topic. Thus, the overall aim of my PhD research is to investigate and promote transparent and standardized reporting in surgical research. As follows, I outline the different aspects I specifically addressed.

Project 1: How to write a surgical clinical research protocol: literature review and practical guide

Any clinical research starts with asking a research question. The question shall be embedded in the existing body of evidence. If the study question and hypothesis is not precise and the study is not designed in a sound manner, this will impact on reporting and overall study conclusions. The core document of the design phase is the study protocol. Our first project was thus to develop a straightforward 10-step practice guideline on how to develop a surgical clinical trial protocol with a focus on methodological aspects. Our project and initiative was highly welcomed by the journal reviewer of the American Journal of Surgery, where the study was published (Rosenthal et al. 2014) and is reflected in his reviewer comment "This is a valuable report that describes the steps necessary to optimize the design of clinical research studies. This information is important and the article should be required reading for all surgeons who are embarking on a research career."

Project 2: The use of systematic reviews when designing and reporting surgical trials

A systematic review (SR) on the existing literature should be part of any clinical research project to justify the planned research from a scientific, ethical and economic point of view as well as to inform the trial design and, finally to sum up the results within the existing body of evidence. In a meta-epidemiological research project we investigated to what extend investigators systematically searched the literature in the context of their project and whether this information was provided in the final publication (Rosenthal et al. 2015a). Of 596 studies, 51 RCTs published in 3 high-impact general surgical journals were identified. SRs were referenced in 65%, either to summarize evidence concerning a related topic (43%) or the study topic (excluding first-in-area RCTs: introduction 24%, discussion 26%). No SRs were used

to inform trial design and no trial updated a SR for the integration of the new results. In conclusion, in the surgical literature SR are rarely used to inform trial design, justify the research and synthesize knowledge for informed decision making.

Project 3: Completion and publication rates of surgical randomized controlled trials – an empirical study

In a next step, we addressed important issues when conducting and reporting results of clinical trials in surgery. We investigated to what extent surgical trials were discontinued early and/or not published. For this purpose, all RCT protocols approved from 2000 to 2003 by six ethics committees in Canada, Germany and Switzerland were screened. We explored risk factors for early trial discontinuation due to slow recruitment and compared surgical to medical trials. Early trial discontinuation due to slow recruitment has an impact on reporting: first, overall trial conclusions due to limited power may be biased if the targeted sample size has not been reached and second early discontinuation may increase the risk of publication bias if results typically derived from underpowered studies are not published. In total, 863 RCT protocols involving adult patients were identified; 127 in surgery (15%) and 736 in medicine (85%). Surgical trials were discontinued for any reason more often than medical trials (43% *versus* 27%, risk difference 16% (95% confidence interval [CI] 5%, 26%); $p=0.001$) and were also more often discontinued for slow recruitment (18% *versus* 11%, risk difference 8% (95% CI 0.1%, 16%); $p=0.020$). The percentage of trials not published as full journal article was similar in surgical and medical trials (44% *versus* 40%, risk difference 4% (95% CI -5%, 14%); $p=0.373$). Discontinuation of surgical trials was a strong risk factor for non-publication (odds ratio 4.18, 95% CI 1.45, 12.06; $p=0.008$) (Rosenthal et al. 2015b).

Project 4: Reporting of adverse events in surgical trials: critical appraisal of current practice

Reporting of surgical outcomes should include detailed reporting of harm. The information needs to be generated according to uniform and reproducible standards that allow for the comparison of results, techniques, centres or surgeons. For postoperative complications, several classifications have been proposed and validated; however, there exists no recognised classification system for intraoperative complications. Therefore, we aimed at assessing the current practice of reporting intra- and postoperative complications in surgical trials that were published in 2010 in three major surgical journals (Rosenthal et al. 2015c). We identified 46 trials that reported intra- and postoperative complications. These complications were reported separately in 42% and pooled in 15%. In 37% intraoperative, in 2% postoperative, and in 4% both intra- and postoperative complications were not reported at all. Exact definitions were provided in 13% for intraoperative and in 50% for postoperative complications. A classification was used in 9% for intra- and in 54% for postoperative complications, most frequently according to severity. As further research, we thus plan to develop and validate a classification of intraoperative complications in order to

facilitate the evaluation of safety and the continuous quality control of surgical interventions with the ultimate goal to contribute to patient safety.

Project 5: Definition and Classification of Intraoperative Complications (CLASSIC): Delphi Study and pilot evaluation

Our abovementioned investigation (Project 4) of reporting of harm clearly showed the urgent need for a classification of intraoperative complications. We therefore conducted a two-stage Delphi study among experts in surgical practice and trial methodology to develop a definition and classification of intraoperative complications (Rosenthal et al. 2015d).

In the Delphi study, a total of 40 out of 52 experts (77% return rate) from 14 countries took part in both rounds. It resulted in a comprehensive definition of intraoperative complications and a straightforward classification. We then conducted a pilot study reviewing 60 records involving surgical interventions of variable complexity by two independent reviewers. It showed good practicability (6 on a 7-point scale) and an 87% agreement with a weighted kappa of 0.83 (95% CI 0.73, 0.94) and an intraclass correlation coefficient of 0.83 (95% CI 0.73, 0.90).

Project 6: How to report multiple outcome metrics in virtual reality simulation

Virtual reality (VR) simulation is increasingly being used for assessment and training purposes in various surgical disciplines. However, there is no consensus on how to report outcomes derived from simulators. VR simulators objectively measure multiple outcomes, which is frequently handled by selective reporting or multiple testing. This compromises comparison between different VR studies and introduces a risk of bias. Thus, suitable methods are needed to either address multiplicity issues or combine evidence from multiple possibly related outcomes into a lower dimensional outcome. We developed an algorithm for summarizing multiple VR outcome metrics into a total score and illustrate the approach with two real data examples (Rosenthal et al. 2015e). Given the increasing number of articles in the field, a standardized and transparent approach for analyzing and reporting VR outcome data is of utmost importance to enhance the validity of VR derived surgical reports.

Project 7 (ongoing): Disregarding paired data in surgical research – evaluation of current practice and estimation of implications on study results

In surgical practice and research, patients may be affected by a pathology and treated on both sides of their body, e.g. in extremity surgery. As a consequence, paired data are generated. In the context of our consultancy service for surgeons we were faced with studies that did not account for the paired design in the statistical analysis. We therefore decided to carry out a systematic literature review based on the example of inguinal hernia repair, one of the most frequently conducted surgical interventions (ongoing). The aim of this project is to first evaluate to what extent paired data are accounted for in the statistical analysis and second to estimate the

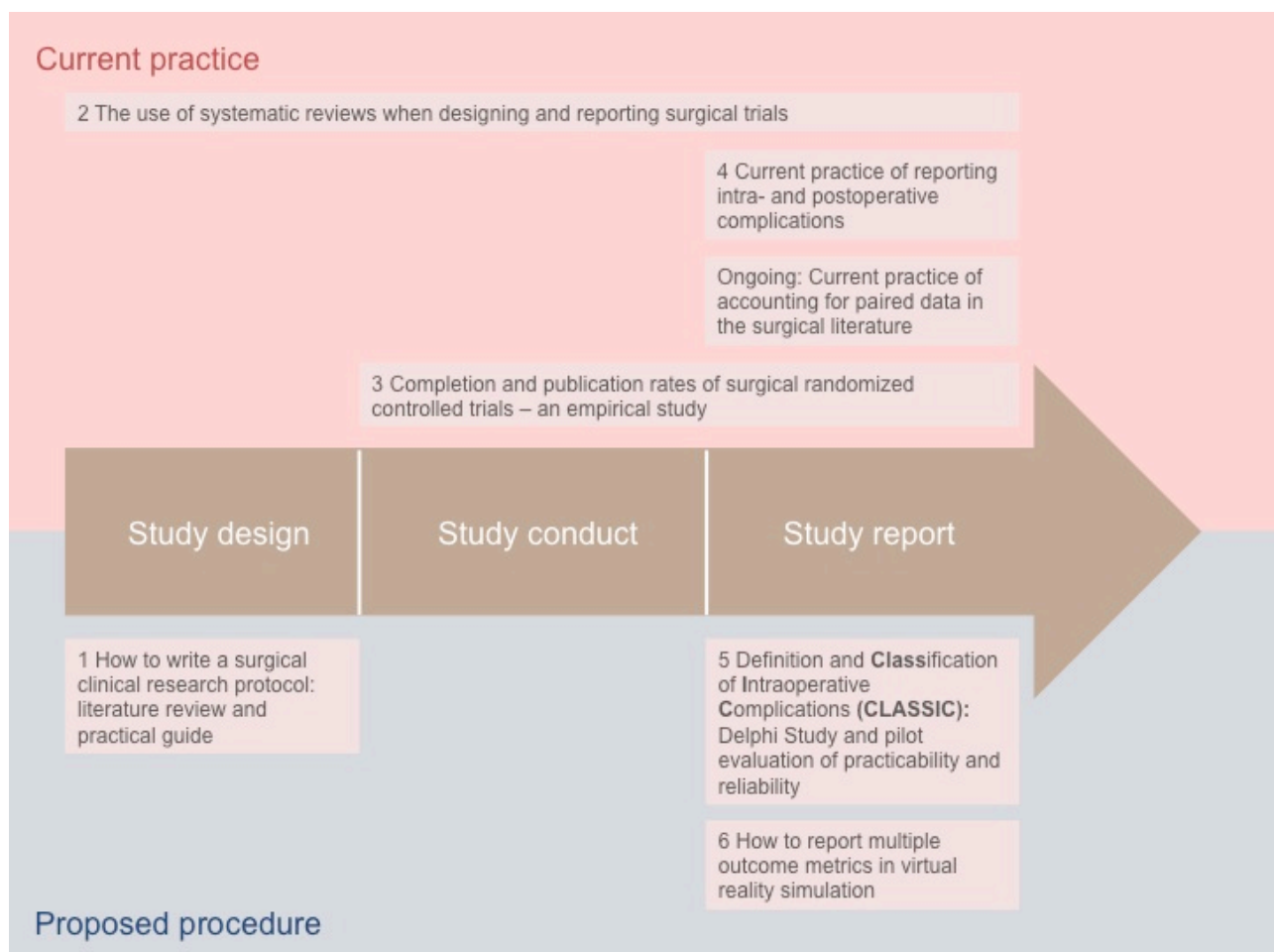
impact on overall study conclusions if this specific design issue is not correctly considered in the statistical analysis.

Conclusion

In conclusion, with the present work the reporting of surgical trials is addressed from two points of view: i) evaluation of current practice and ii) proposal for procedures to address identified challenges in reporting. Since any study report depends on the study conception and conduct, these two phases are included in the evaluation.

Figure 1 gives an overview of the projects within this framework.

Figure 1: Framework of projects evaluating reporting in surgical research



Zusammenfassung

Eine transparente, reproduzierbare und standardisierte Darstellung von klinischen Forschungsergebnissen ist eine Grundvoraussetzung, damit wissenschaftliche Erkenntnisse kritisch evaluiert und – falls valide und relevant - in die Praxis umgesetzt werden können. Sie ist somit eine Grundvoraussetzung für die evidenzbasierte Medizin und für eine optimale Entscheidungsfindung im Gesundheitswesen. Dies gilt für jede klinische Disziplin, ist aber eine besondere Herausforderung, wenn es um komplexe chirurgische Interventionen geht. Dort muss speziellen Gegebenheiten Rechnung getragen werden, wie der Bedeutung der Erfahrung des Chirurgen und deren Einfluss oder Ausmass auf den Behandlungseffekt, der Standardisierung von Interventionen und der Endpunkterfassung.

In einer früheren Untersuchung über die selektive Darstellung von Studienendpunkten (Outcomes) in chirurgischen randomisiert kontrollierten Studien (RCT) (Rosenthal & Dwan 2013) fanden wir einen recht hohen Anteil an unerklärten Diskrepanzen zwischen Ergebnispunkten in Studienregistern und den entsprechenden Publikationen von RCTs, welche in allgemeinchirurgischen Zeitschriften mit hohem Impaktfaktor publiziert wurden. Um mich noch weiteren Aspekten dieses Themas zu widmen, entschloss ich mich, meine Doktorarbeit (PhD) diesem Gebiet zu widmen. Daher ist das übergeordnete Ziel meiner Forschung im Rahmen der Doktorarbeit, die transparente und standardisierte Darstellung und das Berichten von Studienergebnissen in der chirurgischen Forschung zu untersuchen und zu fördern. Nachfolgend fasse ich die spezifischen Fragestellungen, die ich untersucht habe, zusammen.

Projekt 1: Wie man ein chirurgisches klinisches Studienprotokoll abfasst: Literaturreview und Leitfaden

Jede klinische Forschung beginnt mit einer Forschungsfrage. Diese sollte in die bereits bestehende Evidenz eingebettet sein und das adäquate Studiendesign muss dafür gewählt werden. Wenn eine Studie nicht fundiert geplant wird, hat dies schlussendlich Auswirkungen auf die gesamten Schlussfolgerungen der Studie. Das zentrale Dokument in der Studienplanungs- und Durchführungsphase ist das Studienprotokoll. Unser erstes Projekt ist daher die Entwicklung eines 10-Punkte Leitfadens zum Thema „Wie entwickle ich ein Studienprotokoll für eine klinische chirurgische Studie“ mit Fokus auf methodologische Aspekte. Diese Arbeit wurde im American Journal of Surgery publiziert (Rosenthal et al. 2014) und ein Reviewer der Zeitschrift hat die Bedeutung unserer Arbeit wie folgt kommentiert (Zitat übersetzt) „Dies ist ein wertvoller Bericht, der die notwendigen Schritte beschreibt, um die Planung klinischer Studien optimal zu gestalten. Diese Information ist wichtig und der Artikel sollte eine Pflichtlektüre für alle Chirurgen sein, die eine Forscherkarriere planen.“

Projekt 2: Die Verwendung von systematischen Reviewarbeiten für die Planung und das Berichten von Ergebnissen in chirurgischen randomisierten kontrollierten Studien

Als nächstes untersuchten wir, inwiefern Information aus systematischen Reviewarbeiten (SR) verwendet wird, um die geplante Studie aus wissenschaftlicher, ethischer und ökonomischer Sicht zu rechtfertigen sowie um das Studiendesign zu planen. Zudem evaluierten wir, inwiefern am Ende die Resultate in bestehende Information im Sinne einer Synthese integrieren werden (Rosenthal et al. 2015a). Von 596 Studien, welche in allgemeinchirurgischen Zeitschriften mit hohem Impaktfaktor publiziert wurden, konnten 51 RCTs identifiziert werden. SR wurden in 65% verwendet, entweder um Evidenz zu einem verwandten Thema zusammenzufassen (43%) oder zum Studienthema (unter Ausschluss der RCTs, die eine neue Fragestellung untersuchen: Einleitung 24%, Diskussion 26%). In keinem einzigen Fall wurden SR zum Entwickeln des Studiendesigns verwendet und keine RCT hat eine SR aktualisiert unter Integration der neuen Studienresultate. Zusammenfassend lässt sich festhalten, dass SR selten zur Studienplanung, Studienrechtfertigung und Informationssynthese verwendet werden. Dies beeinträchtigt die Beurteilung der Ergebnisse und deren Interpretation im wissenschaftlichen Kontext.

Projekt 3: Studienabschluss –und Publikationsraten bei chirurgischen randomisiert kontrollierten Studien – eine empirische Untersuchung

In diesem Projekt konzentrierten wir uns auf die Aspekte der Studiendurchführung und der Darstellung sowie des Berichtens von Studienresultaten. Der Fokus der Studie lag auf der Untersuchung des Anteils an nicht abgeschlossenen und nicht publizierten Studien. Dafür evaluierten wir alle Studienprotokolle von RCTs, welche zwischen 2000 und 2003 von sechs Ethikkommissionen in Kanada, Deutschland und der Schweiz bewilligt wurden. Wir untersuchten Risikofaktoren für den frühzeitigen Studienabbruch wegen Rekrutierungsproblemen und verglichen chirurgische mit medizinischen Studien. Vorzeitiger Studienabbruch wegen Rekrutierungsproblemen hat einen wichtigen Einfluss auf das Berichten der Resultate: erstens können die Studienresultate wegen der nicht erreichten Fallzahl verzerrt sein, da die statistische Kraft zum Belegen der Studienhypothese fehlen kann und zweitens hat ein frühzeitiger Studienabschluss einen Einfluss darauf, ob die Resultate überhaupt publiziert werden. Studien mit negativen oder unschlüssigen Ergebnissen werden erfahrungsgemäss weniger häufig publiziert.

Insgesamt fanden wir 863 RCT Protokolle von Studien an erwachsenen Patienten, 127 aus der Chirurgie (15%) und 736 aus der Medizin (85%). Chirurgische RCTs wurden insgesamt häufiger vorzeitig abgebrochen als medizinische RCTs (43% versus 27%, Risikodifferenz 16% (95% Vertrauensintervall [CI] 5%, 26%); $p=0.001$) als auch häufiger wegen Rekrutierungsproblemen abgebrochen (18% versus 11%, Risikodifferenz 8% (95% CI 0.1%, 16%); $p=0.020$). Der Prozentsatz an RCTs, welche nicht als volle Zeitschriftenartikel publiziert wurden, war zwischen chirurgischen und

medizinischen RCTs vergleichbar (44% versus 40%, Risikodifferenz 4% (95% CI - 5%, 14%); $p=0.373$). Vorzeitiger Studienabbruch chirurgischer RCTs war ein starker unabhängiger Risikofaktor, dass keine Publikation erfolgte (Odds-Ratio 4.18, 95% CI 1.45, 12.06; $p=0.008$) (Rosenthal et al. 2015b).

Projekt 4: In der Literatur gebräuchliches Berichten von intra- und postoperativen Komplikationen

Die Darstellung und das Berichten von Ergebnissen chirurgischer Studien sollten Komplikationen einschliessen. Es gibt mehrere publizierte und validierte Klassifikationen für postoperative Komplikationen, nicht jedoch für intraoperative Komplikationen. Das Ziel der vorliegenden Untersuchung ist deshalb, das aktuell gebräuchliche Berichten von intra- und postoperativen Komplikationen in chirurgischen RCT zu untersuchen (Rosenthal et al. 2015c). Die Untersuchung von 45 RCTs, welche im Jahr 2002 in den drei höchst zitierten chirurgischen Zeitschriften publiziert wurden, zeigte, dass über intra- und postoperative Komplikationen in 42% separat und in 15% zusammen berichtet wurde. Über intraoperative Komplikationen wurde in 37%, über postoperative in 2% und über intra- und postoperative in 4% überhaupt nicht berichtet. Exakte Definitionen wurden in 13% für intraoperative und in 50% für postoperative Komplikationen angegeben, meist klassifiziert nach Schweregrad der Komplikation. Daher planten wir als nächsten Schritt die Entwicklung und Validierung einer Klassifikation intraoperativer Komplikationen. Ein einheitliches Klassifikationssystem erleichtert die Überprüfung von Sicherheit im Sinne der Qualitätskontrolle chirurgischer Interventionen und leistet schlussendlich einen Beitrag zur Patientensicherheit.

Projekt 5: Definition und Klassifikation intraoperativer Komplikationen (CLASSIC): Delphi Studie und Pilotstudie

Die obengenannte Untersuchung (Projekt 4) zur Darstellung und zum Berichten von Komplikationen zeigt den dringenden Handlungsbedarf und die Notwendigkeit der Entwicklung einer Klassifikation für intraoperative Komplikationen. Wir führten deshalb eine Delphi Studie mit zweimaliger Befragung von Experten in Chirurgie und Studienmethodologie durch zur Entwicklung einer Definition und Klassifikation von intraoperativen Komplikationen (Rosenthal et al. 2015d). Insgesamt 40 von 52 Experten (77% Rücklaufquote) aus 14 Ländern nahmen an beiden Befragungsrunden der Delphi Studie teil. Die Studie führte zu einer umfassenden Definition intraoperativer Komplikationen sowie zu einer überschaubaren Klassifikation. Zwei unabhängige Begutachter wandten dann in einer Pilotstudie mit 60 chirurgischen Eingriffen unterschiedlicher Komplexität die Klassifikation an. Es zeigte sich eine gute Praktikabilität (6 von 7 möglichen Punkten) sowie eine 87-prozentige Übereinstimmung mit einem gewichteten Kappa von 0.83 (95% CI 0.73, 0.94) und einem Intraklasse-Korrelationskoeffizienten von 0.83 (95% CI 0.73, 0.90).

Projekt 6: Wie man multiple unterschiedliche Endpunkte (Outcomes) bei der Simulation mit Virtueller Realität berichtet

Virtuelle Realität (VR) wird zunehmend für die Evaluation und das Training in diversen chirurgischen Disziplinen verwendet. Es gibt allerdings keinen Konsensus, wie man Ergebnisse von Simulatordaten berichten soll. VR Simulatoren messen objektiv eine Vielzahl von Endpunkten (Outcomes), was oft entweder zu selektivem Berichten von Ergebnissen führt oder zum Durchführen multipler statistischer Tests. Dies schränkt die Vergleichbarkeit zwischen Simulatorstudien ein und erhöht das Risiko für eine verzerrte Darstellung der Ergebnisse (Bias). Daher sollten geeignete Methoden zur Verfügung stehen, um die Multiplizität anzugehen oder um Evidenz aus multiplen zum Teil zueinander in Beziehung stehenden Outcomes in einem niedriger dimensional Outcome zusammenzufassen. Wir entwickelten einen Algorithmus um multiple VR Outcomeparameter in einen totalen Score zusammenzufassen und wandten diesen an zwei Datensätzen an (Rosenthal et al. 2015e). Im Hinblick auf die steigende Anzahl an Publikationen auf diesem Gebiet ist ein standardisiertes und transparentes Vorgehen bei der Analyse und dem Berichten von VR Outcomes extrem wichtig, um die Validität von VR Berichten zu stärken.

Projekt 7 (laufend): Vorliegen gepaarter Daten in der chirurgischen Forschung - Evaluation der in der Literatur verwandten Verfahren und Bedeutung für Studienresultate

Im chirurgischen Alltag und der chirurgischen Forschung können Patienten von einer Erkrankung und Behandlung auf beiden Körperseiten betroffen sein, z.B. in der Extremitätenchirurgie. Demzufolge werden bei zweiseitiger Operation gepaarte Daten generiert. Da wir im Rahmen unserer Beratungstätigkeit mehrere Studien fanden, welche das gepaarte Design in der Analyse nicht berücksichtigten, entschlossen wir uns zu einer systematischen Evaluation von Studien im Bereiche der Leistenhernienchirurgie (laufendes Projekt). Das Ziel ist in einem ersten Schritt zu evaluieren, inwiefern dem gepaarten Design in den untersuchten Studien in der statistischen Analyse gerecht wird und in einem zweiten Schritt dann zu untersuchen, inwiefern die fehlende Berücksichtigung der gepaarten Datennatur mit nicht unabhängigen Ergebnissen bei doppelseitigen Eingriffen einen Einfluss auf die Gesamtschlussfolgerungen der Studien hat.

Schlussfolgerung

Zusammenfassend gehen wir mit der vorliegenden Arbeit das Berichtswesen (Reporting) chirurgischer RCTs von zwei Blickrichtungen an: Wir zeigen die aktuelle Praxis des Publikationswesens in der Chirurgie mit ihren Unzulänglichkeiten und Inkonsistenzen und präsentieren Vorschläge für Massnahmen, um die Herausforderung bei der Darstellung und dem Berichten von Studienergebnissen anzugehen. Da das Berichten von Studienresultaten immer auch von der Studienkonzeption und praktischen Durchführung abhängt, haben wir diese beiden

Phasen in unsere Beurteilung integriert. In Abbildung 1 (Figure 1) ist die Projektübersicht in diesem Rahmen grafisch dargestellt.

List of abbreviations

CI	Confidence interval
OR	Odds ratio
RCT	Randomized controlled trial
SR	Systematic Review
VR	Virtual Reality

List of Tables

How to write a surgical clinical research protocol: literature review and practical guide

Table 1: Stages of Surgical Innovation

Table 2: Challenges in Surgical Research

Table 3: Phrasing testable questions

Table 4: Design types

Table 5: Bias types

Table 6: Hypothesis tests and multivariable analysis

Table 7: Sample size calculation

The use of systematic reviews when designing and reporting surgical trials

Table 8: Baseline characteristics of the included studies (n=51)

Table 9: Use of systematic reviews to inform trial design by publication section

Completion and publication rates of surgical randomized controlled trials – an empirical study

Table 10: Baseline characteristics by clinical area

Table 11: Trial completion and publication status by clinical area

Table 12: Reasons for discontinuation by clinical area.

Table 13: Univariable and multivariable logistic regression for trial discontinuation of surgical trials for slow recruitment (using a random effect estimator to account for clustering by the approving research ethics committee)

Table 14: Univariable and multivariable logistic regression for non-publication as full journal article of surgical trials (using a random effect estimator to account for clustering by the approving research ethics committee).

Reporting of adverse events in surgical trials: critical appraisal of current practice

Table 15: Definitions and classification of intra- and postoperative adverse events (n=46)

Definition and Classification of Intraoperative Complications (CLASSIC): Delphi Study and pilot evaluation

Table 16: Baseline characteristics of participants in the Delphi study

Table 17: Proposed *Classification of intraoperative complications* (CLASSIC)

Table 18: Agreement of the pilot study (n=60 cases, n=2 raters)

How to report multiple outcome metrics in virtual reality simulation

Table 19: Algorithm for handling multiple outcomes

Table 20: Difference in mean summary measure or total performance score between the four participant groups: children with low (n=9) and high (n=22) experience in videogames, residents (n=20) and board-certified surgeons (n=14)

Table 21: Difference in mean summary measure or total performance score between the three participant groups: free training group (n=31), structured training group (n=34) and experts (n=3)

List of Figures

Summary

Figure 1: Framework of projects evaluating reporting in surgical research

Completion and publication rates of surgical randomized controlled trials – an empirical study

Figure 2: Flow chart of included studies: Studies eligible for assessment of discontinuation and of non-publication, respectively.

Definition and Classification of Intraoperative Complications (CLASSIC): Delphi Study and pilot evaluation

Figure 3: Delphi study round 1: Answers to the question concerning overall usefulness of a classification of intraoperative complications (1=not useful at all, 9=very useful)

Figure 4: Delphi study round 2: Answers to the questions concerning the distinction of intraoperative complications from sequelae, failure to cure, events related to the underlying disease and error in indication (1= not at all appropriate, 9 = completely appropriate)

Figure 5: Delphi study round 2: Answers to the questions concerning type of grading, classification overall, timepoint of grading, dichotomisation (1= not at all appropriate, 9 = completely appropriate)

How to report multiple outcome metrics in virtual reality simulation

Figure 6: Boxplots comparing children with low (n=9) and high (n=22) experience in videogames (LE and HE), residents (n=20) and board-certified surgeons (n=14) (Board-cert.) (Rosenthal et al. 2011). (A) Accuracy, time and path length mean summary measures. (B) Total virtual reality (VR) performance score with accuracy considered twice as important as time and path length together.

Figure 7: Boxplots comparing all three children with low (n=9) and high (n=22) experience in videogames (LE and HE) and residents (n=20) with board-certified surgeons (n=14) (Board-cert.) with the latter taken as the reference population (Rosenthal, Geuss, Dell-Kuster, Schafer, Hahnloser, & Demartines 2011) (A) Accuracy, time and path length mean summary measures. (B) Total virtual reality (VR) performance score with accuracy considered twice as important as time and path length together.

Figure 8: Dotplots comparing surgical novices in the free training (n=31), structured training (n=34) and expert group (n=3) (von Websky et al. 2013) (A) Accuracy, time and path length mean summary measures. (B) Total virtual reality (VR) performance score with accuracy considered twice as important as time and path length together.

Figure 9: Total virtual reality (VR) performance score – with accuracy considered twice as important as time and path length together – over iterations for participants in the free training (n=31), structured training (n=34) and expert group (n=3) (von Websky, Raptis, Vitz, Rosenthal, Clavien, & Hahnloser 2013) (A) The smooth curves are local averages found using the default loess smoother in R version 3.0.1 (R Foundation for Statistical Computing, Vienna, Austria). (B) The smooth curves are the predicted values (based on the fixed part of the linear mixed-effects model) of the total VR performance score.

Current practice of accounting for paired data in the surgical literature (working project)

Figure 10: Conceptual Framework for handling paired data

Acknowledgement

Completion and publication rates of surgical randomized controlled trials – an empirical study

We would like to thank the presidents and staff of participating research ethics committees from Switzerland (Basel, Lausanne, Zurich, Lucerne), Germany (Freiburg), and Canada (Hamilton, Ontario) for their continuous support and cooperation.

Definition and Classification of Intraoperative Complications (CLASSIC): Delphi Study and pilot evaluation

We are very grateful to all experts who participated in this study. Their input was very important in the development of the proposed definition and classification.

1. Introduction

1.1. Background

Transparent, standardized and accurate reporting in clinical research is of utmost importance for decision making in healthcare. Reporting should follow standardized guidelines as proposed by CONSORT (Consolidated Standards of Reporting Trials Statement) (Schulz et al. 2010). Of specific interest to surgery are the CONSORT extensions for nonpharmacologic treatment (Boutron et al. 2008). There, challenges to surgical trials such as blinding, experience of and clustering by care providers and centers and standardization of interventions are accounted for.

In order to be able to accurately report study results, the study needs to be well-designed and well-conducted. Thus, for reporting in clinical research, all three phases are relevant: the study design, study conduct and study report phase. In the study design phase, numerous methodological issues need to be considered and finally included in a detailed study protocol. Importantly, to be able to justify a planned research from a scientific, ethical and economic point of view, the current research question should be set in context with the existing body of evidence. Information from previous trials and systematic reviews may be used to inform trial design.

In the study conduct phase, the developed protocol should be strictly followed. Additionally, threats to completing an ongoing trial, such as early discontinuation for slow recruitment, should be monitored and anticipated.

Finally, in the study report phase, great care should be given to standardized and transparent reporting, relying on clear definitions and validated classifications. This is a pre-requisite for comparison of study results and healthcare decision making. It applies both to safety and efficacy outcomes. Whereas for postoperative complications several classification systems have been proposed (Clavien et al. 1992;Clavien et al. 2009;Dindo et al. 2004;Pillai et al. 1999;Pomposelli et al. 1997;Strasberg et al. 2009), we are unaware of a validated definition and classification strictly applying to intraoperative complications. An example of efficacy outcomes that should be transparently reported are performance measurements generated by virtual reality (VR) simulators. VR simulators measure multiple outcomes on different scales. Therefore, suitable methods are needed to either address multiplicity issues or combine evidence into a lower dimensional outcome. Additionally, in any type of outcome, the statistical analysis and the report should account for paired data, resulting from several interventions per patient. If not accounted for, effects may be overestimated.

There are two areas to be addressed when further evaluating reporting in surgical research: 1) the evaluation of current practice and 2) the development of guidelines to optimize specific reporting challenges. Based on current literature in

the field, we have thus identified seven research projects addressing study design, conduct and reporting from a perspective of current practice evaluation or guideline development (Figure 1).

1.2. Aims and objectives

The overall aim of the present research is the evaluation and optimization of reporting in surgical research, thus facilitating informed healthcare decision making.

Within seven projects, we specifically address the following objectives:

1. To develop a practice guide outlining key methodological issues important when planning an ethically and scientifically sound research project involving surgical interventions.
2. To evaluate to what extent systematic reviews are used in surgical RCTs to inform trial design and to integrate trial results in the overall body of evidence.
3. To determine the proportion of discontinued surgical trials and the reasons for discontinuation, differences in discontinuation between medical and surgical trials and risk factors for non-publication of surgical trials.
4. To systematically assess the current practice of reporting intra- and postoperative adverse events in RCTs published during one year in three major general surgery journals.
5. To develop a definition and classification for intraoperative complications within a Delphi study and to conduct a pilot study evaluating practicability and interrater agreement.
6. To present an algorithm for summarizing multiple VR outcomes of different dimensions and weighing them within a total score.
7. To analyze in RCTs of patients undergoing hernia repair without exclusion of bilateral surgery i) to what extent the presence of clustered data has been accounted for in the study design and statistical analysis and ii) to estimate the effect of disregarding clustered data on the overall results (ongoing).

2. How to write a surgical clinical research protocol - literature review and practical guide

Rachel Rosenthal^{1*}, MD, MSc, Juliane Schäfer^{1,2}, PhD, Matthias Briel^{2,3}, MD, MSc, Heiner C. Bucher², MD, MPH, Daniel Oertli¹, MD, FACS, Salome Dell-Kuster^{1,2}, MD, MSc

¹Department of Surgery, Basel University Hospital, Switzerland

²Basel Institute for Clinical Epidemiology and Biostatistics, Basel University Hospital, Switzerland

³Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada

*Corresponding author

Disclosure information: There is no funding related to this publication. There are no conflicts of interest to declare.

Publication information: This is a non-final version of an article published in final form in the American Journal of Surgery

(<http://www.americanjournalofsurgery.com/>): Rosenthal R, Schäfer J, Briel M, Bucher HC, Oertli D, Dell-Kuster S. How to write a surgical clinical research protocol: literature review and practical guide. Am J Surg 2014;207(2):299-312 (Rosenthal et al. 2014)

Abstract

Background: The study protocol is the core document of every clinical research project. Clinical research in studies involving surgical interventions presents some specific challenges, which need to be accounted for and described in the study protocol. The aim of this review is to provide a practical guide for developing a clinical study protocol for surgical interventions with a focus on methodological issues.

Data sources: Based on an in-depth literature search of methodological literature and on some cardinal published surgical trials and observational studies, this paper provides a 10-step guide for developing a clinical study protocol in surgery.

Conclusions: This practical guide outlines key methodological issues important when planning an ethically- and scientifically sound research project involving surgical interventions, with the ultimate goal to provide high level evidence relevant for healthcare decision-making in surgery.

2.1. Introduction

The study protocol as a core document in clinical research

The study protocol is the central document of a clinical research project and takes into account scientific, ethical and regulatory considerations. It provides detailed information on all aspects of the planning and conduct of the research project and is the main document for evaluation of the planned research, e.g. by an independent ethics committee and regulatory authorities. It guides study investigators to conduct the study according to standardized criteria and it allows replication in subsequent studies. The protocol includes the justification for the planned research, the objectives, details on the intervention and the study population, information on data management, quality assurance, statistical analyses and ethical considerations. Importantly, the protocol should be developed in an interdisciplinary setting, including clinicians, scientists, statisticians and other involved parties. Study protocols need to be approved by an independent ethics committee (IEC) and by the regulatory authorities according to local guidelines.

Types and phases of surgical research

Surgical clinical research may involve pharmaceuticals, medical devices, surgical procedures and other interventions concerning prevention, diagnostics, treatment and rehabilitation. In drug development, the phases of investigation have been well-defined and most typically range from human pharmacology studies (phase I) through therapeutic exploratory (phase II) to therapeutic confirmatory studies (phase III), followed by post-marketing studies (phase IV) (The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) 2014). Similar phases ranging from pilot, pivotal through to post-marketing surveillance have been described for medical devices, for which requirements for demonstrating safety and efficacy depend on the risk associated with the device (Kaplan et al. 2004). In the example of research involving surgical interventions, the IDEAL framework has been proposed, the acronym standing for the stages (1) **i**dea including proof of concept, (2a) **d**evelopment, (2b) **e**xploration, (3) **a**ssessment and (4) **l**ong-term study (Heikens et al. 2013; McCulloch et al. 2009). An overview of the IDEAL framework with examples is provided in [table 1](#).

In this review article, we focus on surgical interventions; however some of the concepts may be extrapolated to other interventions.

Table 1: Stages of Surgical Innovation (adapted from McCulloch and colleagues (McCulloch, Altman, Campbell, Flum, Glasziou, Marshall, Nicholl, Aronson, Barkun, Blazeby, Boutron, Campbell, Clavien, Cook, Ergina, Feldman, Flum, Maddern, Nicholl, Reeves, Seiler, Strasberg, Meakins, Ashby, Black, Bunker, Burton, Campbell, Chalkidou, Chalmers, de, Deeks, Ergina, Grant, Gray, Greenhalgh, Jenicek, Kehoe, Lilford, Littlejohns, Loke, Madhock, McPherson, Meakins, Rothwell, Summerskill, Taggart, Tekkis, Thompson, Treasure, Trohler, & Vandenbroucke 2009))

IDEAL stage		Purpose	Study design	Example
1	Idea	Proof of Concept	<ul style="list-style-type: none"> · Case report/Case series 	Ileo neorectal anastomosis (INRA) after ulcerative colitis and familial adenomatous polyposis (Heikens, Gooszen, Rovers, & van Laarhoven 2013)
2a	Development	Development	<ul style="list-style-type: none"> · Prospective cohort 	Confirmation of pilot study findings and refining of technique in 26 patients
2b	Exploration	Learning	<ul style="list-style-type: none"> · Research database (Prospective cohort) · Feasibility/explanatory RCT¹ 	Prospective cohort with extended inclusion criteria
3	Assessment	Assessment	<ul style="list-style-type: none"> · RCT · Alternative designs if RCT not applicable <ul style="list-style-type: none"> ○ Matched case-control study ○ Interrupted time series (multiple observations over time, interrupted by intervention) ○ Controlled before-after study (observation before and after intervention in intervention and control group) ○ Step-wedged design (random order of introduction of intervention in a prospective cohort) 	Matched case-control study (INRA versus IPAA (ileal pouch anal anastomosis as gold standard) with long-term results showing comparable morbidity and functional results, no advantage from the patients' point of view, and a disadvantage from the surgeons' point of view (technically demanding and long surgery) ® decision not to offer INRA anymore, use of IPAA as standard
4	Long-term study	Surveillance	<ul style="list-style-type: none"> · Routine database/registry (prospective cohort) · Case report (rare events) 	

¹ RCT=randomized controlled trial

Challenges in clinical research of surgical interventions

When planning surgical research involving surgical interventions, some specific challenges need to be addressed (for an overview, possible solutions and examples refer to table 2) (Bonenkamp et al. 1999; Boutron, Moher, Altman, Schulz, & Ravaud 2008; Clavien, Barkun, de Oliveira, Vauthey, Dindo, Schulick, de, Pekolj, Slankamenac, Bassi, Graf, Vonlanthen, Padbury, Cameron, & Makuuchi 2009; Clinical Outcomes of Surgical Therapy Study Group. 2004; Dindo, Demartines, & Clavien 2004; Finkemeier et al. 2000; Mangram et al. 1999a; Moseley et al. 2002; Tincello et al. 2009).

First, as compared to pharmacological trials, surgical interventions are more complex and may thus be more difficult to standardize. Standardization may be enhanced and controlled by specific surgeon selection (i.e. minimum training requirements) and training, direct and video-recorded supervision as well as by anatomic-pathological quality control, as for instance in the Dutch gastric cancer D1 versus D2 lymphadenectomy trial (Bonenkamp, Hermans, Sasako, van de Velde, Welvaart, Songun, Meyer, Plukker, Van, Obertop, Gouma, van Lanschot, Taat, de Graaf, von Meyenfeldt, & Tilanus 1999) and the COST laparoscopic versus open colectomy colon cancer trial (Clinical Outcomes of Surgical Therapy Study Group. 2004). If applicable, details should be provided in the protocol how interventions are tailored to individual patients (Boutron, Moher, Altman, Schulz, & Ravaud 2008).

Second, the surgeons' expertise or hospital standards may have an impact on the clinical outcome, respectively treatment effect, which needs to be accounted for in the design and analysis phase of the study (Boutron, Moher, Altman, Schulz, & Ravaud 2008; Devereaux et al. 2005). This may, for instance, be addressed by defining eligibility criteria to participate as a care provider and center in a trial, and further be enhanced by foreseeing baseline data on the care providers' and centers case volume, expertise and qualifications as well as by taking into account the clustering effect of care providers and centers in sample size calculation, statistical analysis and reporting (Boutron, Moher, Altman, Schulz, & Ravaud 2008). However, surgeons may tend to be most experienced in one surgical approach, which potentially leads to differential expertise bias, even if they meet minimum criteria for participation in a trial (Devereaux, Bhandari, Clarke, Montori, Cook, Yusuf, Sackett, Cina, Walter, Haynes, Schunemann, Norman, & Guyatt 2005). This problem may be addressed by surgical expertise based randomized controlled trials, in which patients are randomized to different surgeons who are experts in the respective treatment arm (Devereaux, Bhandari, Clarke, Montori, Cook, Yusuf, Sackett, Cina, Walter, Haynes, Schunemann, Norman, & Guyatt 2005). This concept has, for instance, been applied for a trial comparing tibial shaft fracture treatment with intramedullary nails, with versus without reaming (Finkemeier, Schmidt, Kyle, Templeman, & Varecka 2000).

Third, due to the nature of surgical interventions, blinding may be difficult to achieve. If those administering the intervention cannot be blinded, blinding of

outcome assessors and/or patients may still be achieved. In a trial investigating the effect of arthroscopy in patients with knee osteoarthritis, placebo surgery was carried out using skin incisions accompanied with operation room acoustics, comparable to real arthroscopy (Moseley, O'Malley, Petersen, Menke, Brody, Kuykendall, Hollingsworth, Ashton, & Wray 2002). However, such measures to reduce bias need to undergo careful ethical considerations.

Fourth, reporting of adverse events needs to be standardized in order to be comparable between studies (Martin et al. 2002a). Therefore, clear definitions of intra- and postoperative complications in the study protocol are mandatory, including their grading of severity and specification of foreseen follow-up. Surgical site infections for instance may be defined according to the Centers of Disease Control and Prevention and graded in superficial incisional, deep incisional and organ/space (Mangram, Horan, Pearson, Silver, & Jarvis 1999a). A widely used classification of postoperative complications according to severity has been proposed by Clavien and Dindo (Clavien, Barkun, de Oliveira, Vauthey, Dindo, Schulick, de, Pekolj, Slankamenac, Bassi, Graf, Vonlanthen, Padbury, Cameron, & Makuuchi 2009; Dindo, Demartines, & Clavien 2004).

Last, there are some ethical considerations. Whereas equipoise refers to the uncertainty within the scientific community whether one treatment is superior to the other and is an ethical pre-requisite for conducting a randomized controlled trial, patients may not be willing to be randomized to either arm, such as when comparing surgery to medical treatment, potentially leading to selection bias and slow recruitment with early trial termination (McCulloch et al. 2002). A pilot study may be helpful in investigating the informed consent and recruitment process (Lancaster et al. 2004). Additionally, surgeons should be well aware of their potentially conflicting role as clinician versus investigator. Even if clinical equipoise is established within the expert clinical community, an individual surgeon may still have a preference for one treatment. This dilemma may be addressed by recognizing that the overall body of evidence does not suggest any treatment to be superior (McDonald et al. 2010). Moreover, sometimes regular practice, surgical innovation and surgical research may be difficult to discriminate. In such circumstances, ethics committees should be liberally consulted (McDonald, Kulkarni, Farrokhyar, & Bhandari 2010).

The purpose of this paper is to provide a guide for developing a study protocol while focussing on the key methodological issues to consider when investigating a surgical intervention, be it in an observational or interventional setting.

Table 2: Challenges in Surgical Research

Challenge	Meaning	Possible solution	Examples
Standardization	Surgical interventions are complex and difficult to standardize	<ul style="list-style-type: none"> ü Minimum training requirements ü Direct/video supervision ü Anatomico-pathological quality control 	<p>Dutch gastric cancer D1 versus D2 lymphadenectomy trial (Supervision; Monitoring pathological results) (Bonenkamp, Hermans, Sasako, van de Velde, Welvaart, Songun, Meyer, Plukker, Van, Obertop, Gouma, van Lanschot, Taat, de Graaf, von Meyenfeldt, & Tilanus 1999)</p> <p>COST laparoscopic versus open colectomy colon cancer trial (Minimum training requirement) (Clinical Outcomes of Surgical Therapy Study Group. 2004)</p>
Expertise	Surgeons' and hospitals' expertise have an impact on the outcome	<ul style="list-style-type: none"> ü Eligibility criteria to participate as care provider ü Collect baseline characteristics on expertise ü Account for clustering effect in design and analysis ü Expertise-based randomized controlled trial 	Tibial shaft fracture treatment with intramedullary nails, with versus without reaming (Expertise based randomized controlled trial) (Finkemeier, Schmidt, Kyle, Templeman, & Varecka 2000)
Blinding	Blinding not always possible	<ul style="list-style-type: none"> ü Blinded outcome assessors ü Placebo surgery (caveat: ethical considerations) 	Randomized controlled trial with sham surgery to evaluate effect of arthroscopy in patients with knee osteoarthritis (Moseley, O'Malley, Petersen, Menke, Brody, Kuykendall, Hollingsworth, Ashton, & Wray 2002)

Table 2: Challenges in Surgical Research (continued)

Challenge	Meaning	Possible solution	Examples
Adverse events reporting	Standardization of adverse event reporting not always considered	<ul style="list-style-type: none"> ü Clear definitions of intra- and postoperative complications ü Reproducible grading of complications 	<p>Surgical site infections defined according to the Centers of Disease Control and Prevention (Mangram, Horan, Pearson, Silver, & Jarvis 1999a)</p> <p>Classification of postoperative complications according to severity (Clavien, Barkun, de Oliveira, Vauthey, Dindo, Schulick, de, Pekolj, Slankamenac, Bassi, Graf, Vonlanthen, Padbury, Cameron, & Makuuchi 2009; Dindo, Demartines, & Clavien 2004)</p>
Ethical considerations	<p>Patient may not be willing to be randomized to surgical interventions</p> <p>Equipoise versus surgeons' preference</p> <p>Surgical innovation versus surgical research</p>	<ul style="list-style-type: none"> ü Pilot study ü Consider overall body of evidence ü Ethics committee clearance 	<p>Pilot randomized patient-preference study comparing colposuspension with tension-free vaginal tape plus anterior repair in women with incontinence and prolapse (Tincello, Kenyon, Slack, Toozs-Hobson, Mayne, Jones, & Taylor 2009)</p>

2.2. Guide for developing a clinical study protocol for surgical interventions

This paper provides a 10-step practical guide for developing a clinical study protocol investigating a surgical intervention using observational or interventional data. It focuses on methodological issues and may be used as adjunct to existing international guidelines (The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) 2014), local regulations, and the recommendations of the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) initiative (Chan et al. 2013a; Chan et al. 2013b). The informed consent process is beyond the scope of this article.

Step 1: Defining the research question

The heart of every protocol is the research question. It defines the knowledge gap which shall be filled with the planned research. Characteristics of a good research question are easily described by the mnemonic „FINER“, as proposed by Cummings and colleagues, standing for **F**easible in terms of scope, expertise, resources and recruitment, **I**nteresting to the investigator and the scientific community, **N**ovel, targeting new findings or the extension, confirmation or rejection of previous findings, **E**thical with a fair subject selection and a favourable risk-benefit ratio, and **R**elevant to scientific knowledge, daily practice, health policy and future research (Cummings et al. 2007). It is crucial to precisely formulate the research question. This allows to develop a statistical analysis plan and to determine the sample size necessary to attain a targeted power. When formulating a research question, the PICO acronym (Richardson et al. 1995) may be helpful for phrasing testable questions. PICO stands for the **P**atient/problem, the **I**ntervention or exposure, the **C**omparison and the **O**utcome. Some add a “T” (PICOT) as a fifth element, which stands for **T**ime (time frame of outcome assessment) (Haynes 2014), whereas in review questions it may stand for study **T**ype (e.g. randomized controlled trial, cohort study etc.). The relevant points to consider when formulating the research question and examples are presented in [table 3](#).

There may be several research questions, however in general the most important one should be labelled as the primary research question, the other(s) as secondary research question(s). For each research question, a hypothesis should be formulated to pre-specify what results are expected. Since the sample size calculation is based on the primary outcome, secondary research questions may not necessarily be answered with sufficient power. They are thus often more exploratory in nature. In an inguinal hernia trial comparing two surgical techniques, a secondary research question could be, for instance, to compare the postoperative quality of life between the two techniques, whereas the primary research question may be the comparison of recurrence rates.

It is important to define these research questions and outcomes in advance. Post hoc specification with the risk of data-driven selection may firstly introduce considerable outcome reporting bias, i.e. significant results being more likely to be reported than negative results, and secondly lead to the error prone acceptance of an association based on multiple posthoc testing (Chan et al. 2004). Trial registries have been introduced to enhance transparency and to address the problem of publication bias and outcome reporting bias (Zarin et al. 2007). Trial registration includes information on the choice of primary and secondary outcomes (World Health Organization 2013). The International Committee of Medical Journal Editors (ICMJE) (De et al. 2004) and subsequently the Surgical Journal Editors Group (SJEG) (Surgical Journal Editors Group 2007) have published guidelines for mandatory trial registration for all trials as a prerequisite for considering a scientific paper for publication in the respective member journals.

Step 2: Justification of the planned research

It is important for the reader to understand why this research is planned. This involves an overview of the current knowledge in the field («What has been done?») and a presentation of the knowledge gap which will be addressed with the planned research («What needs to be done?»). The ultimate purpose is to justify from a scientific, ethical and economic point of view the conduct of this research. The presentation of the current state of the art and knowledge in the field implies a systematic review of the literature, including published literature, grey literature and consulting trial registries to get information about ongoing trials or past unpublished trials. The presented literature should be critically commented and indicate eventual discrepancies in study results or limitations of study design, methodological quality components such as blinding or extent of follow-up, and the number of included participants. The key information of the cited studies may be presented within a table. This part of the protocol directly leads the reader to the aim of the planned research, which represents the logical consequence of the lack of knowledge previously described.

Table 3: Phrasing testable questions (adapted from Richardson and colleagues (Richardson, Wilson, Nishikawa, & Hayward 1995) and Haynes (Haynes 2014))

PICO(T)	Meaning	Consider	Example	Research question/Hypothesis
Patient/problem	What patient or problem are you planning to address?	<ul style="list-style-type: none"> ü Age ü Gender ü Pathology ü In-/outpatients ü Emergency/elective ü Vulnerable population, e.g. children, cognitively impaired 	All patients aged ³ 18 years with primary unilateral inguinal hernia	<p>Research question:</p> <p>What is the 5-year recurrence rate in adult patients with primary unilateral inguinal hernia undergoing total extraperitoneal versus Lichtenstein hernia repair?</p> <p>Hypothesis:</p> <p>The 5-year recurrence rate in adult patients with primary unilateral inguinal hernia is lower after total extraperitoneal hernia repair than after Lichtenstein hernia repair.</p>
Intervention/exposure	What is the planned intervention?	<ul style="list-style-type: none"> ü Surgical intervention ü Pharmaceutical treatment ü Diagnostic procedure ü Prophylactic procedure ü Management process 	Total extraperitoneal hernia repair	
Comparison	What is your intervention compared to?	<ul style="list-style-type: none"> ü Other intervention ü Standard intervention ü No intervention ü Placebo 	Lichtenstein (open) hernia repair	
Outcome	What will be affected by the intervention?	<ul style="list-style-type: none"> ü Efficacy, e.g. recurrence rate ü Safety, e.g. complication rate ü Mortality rate ü Length of hospital stay ü Patient-reported outcomes, e.g. pain, quality of life 	Hernia recurrence	
(Time)	When will you assess the effect of your intervention?	<ul style="list-style-type: none"> ü At one time point ü At several time points ü Continuously over a certain period ü Is time until reaching the endpoint important? 	5 years	

Step 3: Deciding on outcomes and confounders

Variables may be divided into i) outcome/dependent variables, such as the recurrence rate in a study comparing two different surgical techniques for hernia repair, and ii) independent variables or exposure of interest, in this example surgical technique for hernia repair, and iii) confounders, such as age or ASA (American Society of Anaesthesiologists) classification.

Outcomes

Every outcome (or endpoint) needs to be clearly defined in order to standardize outcome measures. For this purpose, the time point of assessment and unit of the outcome measure should be noted and references to definitions and validations should be included such as the “rate of surgical site infections, defined according to the Centers for Disease Control and Prevention (Mangram, Horan, Pearson, Silver, & Jarvis 1999a)” or “Quality of life, measured using the 36-item short-form health survey (SF-36) (Ware, Jr. and Sherbourne 1992)”. Efficacy and safety outcomes should be labelled as such and standard procedures for reporting and patient follow-up of adverse events need to be described.

When choosing outcomes it is important to be aware of several points that affect the statistical analysis plan and sample size calculation: i) The type of variables that are collected (Whitley and Ball 2002) (e.g. categorical, metric, time-to-event data), ii) in the case of continuous variables, whether they can be expected to be normally distributed or not, and iii) if paired or unpaired data are collected. Categorical (binary in case of two categories) or qualitative variables have no units and may be divided into nominal variables in the case of unordered categories (e.g. blood group) and ordinal variables in the case of ordered categories (e.g. American Society of Anesthesiologists (ASA) classification system). They are most often displayed in frequency tables and bar charts. Metric or quantitative variables are either referred to as discrete variables with integer values and counted units (e.g. number of episodes of angina pectoris per week) or continuous variables with stepless values and measured units (e.g. blood pressure). They are typically displayed reporting their central value and variation, i.e. mean and standard deviation in case of normal distribution, else median and range or interquartile range. In graphs, box plots and histograms are used for displaying metric variables. Typical examples of time-to-event data are overall or progression-free survival. Categorization of continuous variables should be avoided, since this is associated with a loss of information and therefore a loss of power and precision. Right-skewed data (i.e. the mass of the data is concentrated on the left with relatively few high values) are relatively frequent; examples are many laboratory findings, or duration of surgery or of hospitalization. Skewed data may be transformed for statistical analysis to achieve better approximation to normality by, for example, logarithmic transformation. Paired or clustered data are generated if the same measurement is repeated in the same patient over time (e.g. repeated measurements of pain in the same patient after hernia repair) or twice in the same

patient at one point in time, such as evaluation of hernia recurrence on the right and the left side after bilateral inguinal hernia repair. Paired and clustered data will have an impact on the choice of methods for statistical analysis, since the variability of several measurements within one patient is smaller than the variability of measurements between several independent patients.

For some outcome variables and settings it is a prerequisite to get baseline information, such as evaluating the quality of life before and after hernia repair.

Outcomes may be objective, such as mortality or subjective, such as pain. In any case, patient-important outcomes should be considered. For subjective outcomes, blinding is especially relevant.

In the case of rare events with insufficient power to evaluate multiple single outcomes, or when no single outcome optimally represents the outcome of interest, a composite endpoint may be chosen (Mascha and Sessler 2011). In order to enhance feasibility and comparability of RCTs (randomized controlled trials), this has for instance been proposed for liver surgery with a composite endpoint involving «ascites, postresectional liver failure, bile leakage, intra-abdominal haemorrhage, intra-abdominal abscess and operative mortality» (van den Broek et al. 2011). Hereby the individual components of the composite endpoint should be of similar importance to patients, they should occur with similar frequency, and similar treatment effects (e.g. relative risk reductions) should be expected (Montori et al. 2005b). whereas components which are redundant or marginally related to the intervention should be avoided (Mascha & Sessler 2011).

Surrogate endpoints and surrogate biomarkers are frequently used, since they may be easier and faster to assess as compared to the patient-important outcome (Buyse 2009). An example is the surrogate endpoint lipid profile instead of major cardiovascular events. A surrogate endpoint can be defined as “a laboratory measurement or a physical sign used as a substitute for a clinically meaningful endpoint that measures directly how a patient feels, functions or survives” (Temple 2014). The effect of the intervention on the surrogate endpoint should predict the effect on the clinically relevant outcome (Fleming and DeMets 1996). Thus, the use of surrogate endpoints needs to be carefully evaluated (Bucher et al. 1999; Riggs et al. 1990).

Independent and confounding variables

In a randomized controlled trial (RCT), confounders should be equally distributed in the different treatment arms through the process of randomization, if randomization has been correctly conducted and if the number of randomized individuals is sufficiently large. This is not the case in observational data. In a cohort study including patients having undergone laparoscopic or open left colectomy for example, not only the surgical technique, but also age or ASA classification may have an impact on the length of hospital stay and will probably not be equally distributed between groups. To be regarded as a confounder, these

factors need to have an impact not only on the outcome, but also on the choice of intervention (i.e. independent variable), meaning an older patient may be more likely to be assigned to one treatment option than to the other. Which factors qualify to be a confounder, i.e. are associated with the outcome as well as with the exposure of interest, should be pre-specified in the study protocol according to expert opinion and information gathered accordingly (Babyak 2004).

Confounding may be controlled for in the design as well as in the analysis of a study. In the design, randomization should lead to equally distributed known and unknown confounders in the groups. Stratification with/without randomization aims at balancing the groups for specific prognostic patient characteristics (Altman and Bland 1999). In the COST trial comparing laparoscopically-assisted to open colectomy for colon cancer, randomization stratified for the site of primary tumour, ASA classification and surgeon was undertaken (Clinical Outcomes of Surgical Therapy Study Group. 2004). If a randomized study is not possible, matching may be a strategy to reduce confounding. In the statistical analysis, potential confounding may be addressed by adjusting for these variables using multivariable regression analysis. Other ways to control for confounding are using propensity scores (probability of an individual to be treated with an intervention given all available baseline information on the patient) or inverse probability weighting (reciprocal of an individual's probability of receiving the treatment that they actually received) (D'Agostino, Jr. 1998; Hernan and Robins 2006).

Superiority/Equivalence/Non-Inferiority

An *a priori* statement of the overall goal of comparison needs to be provided in the protocol. The reader should know whether the goal is to show that treatment A is superior to treatment B, equivalent, or non-inferior. From a superiority study with non-significant results, one may not conclude that the interventions are equivalent (Alderson 2004). In an equivalence (two-sided hypothesis) or non-inferiority (one-sided hypothesis) setting, the margin of non-inferiority, respectively the two margins of equivalence, need to be pre-specified, i.e. the largest, respectively the largest and smallest value representing a clinically irrelevant difference, need to be defined in advance. A help when defining the margin(s) is the question whether the investigational intervention is equivalent if its efficacy or safety outcome is within the chosen boundaries. This margin has implications on the sample size, i.e. the smaller the margin, the larger the sample size. As a rule of thumb, the required sample size is higher with equivalence/non-inferiority designs than in superiority trials. Examples of non-inferiority trials are trials in surgical oncology with the ultimate goal to assess whether a new intervention with potential benefits such as lower invasiveness, lower toxicity or reduced cost is equivalent or not inferior to an established therapy concerning efficacy, i.e. cancer control (Fueglistaler et al. 2007). Examples are the above mentioned gastric lymphadenectomy trial (Bonenkamp, Hermans, Sasako, van de Velde, Welvaart, Songun, Meyer, Plukker, Van, Obertop, Gouma, van Lanschot, Taat, de Graaf,

von Meyenfeldt, & Tilanus 1999) or the COST trial (Clinical Outcomes of Surgical Therapy Study Group. 2004).

Multiple comparisons, multiple testing and interim analyses

Multiple comparisons between different groups need to be carefully justified; the same is true for multiple testing due to multiple outcomes or multiple time points in the case of interim analyses.

If there is a possibility that the treatment effect might be different in different subgroups of patients, this should be examined through an additional interaction term in the regression model (Assmann et al. 2000), rather than multiple testing of each subgroup (Schulz and Grimes 2005). To give an example, in a long-term comparison of endovascular versus open aortic aneurysm repair, a significant interaction between age and type of treatment was found with a better survival in patients younger than 70 years after endovascular repair versus a borderline better survival in patients 70 years of age or older after open repair (Lederle et al. 2012). Subgroups should be pre-specified in the study protocol; post-hoc subgroup analyses should be declared as such and thus are more explorative in nature. All subgroup analyses should be reported to avoid the risk of selective data-driven reporting (Schulz & Grimes 2005). In case of a continuous covariate which may influence response to treatment, advanced methods to modelling treatment-covariate interactions by fractional polynomials should be considered (Royston and Sauerbrei 2004).

Since multiple comparisons and multiple testing increase the chance of committing an α -error (type I error, i.e. concluding that there is a difference when in fact there is no difference), this needs to be accounted for with a more stringent p-value considered as significant. To give an example, if 20 independent outcomes are compared between 2 groups using hypothesis tests, the global type I error rate will increase to 64%. Various procedures have been described to control for the multiple type I error rate out of which the Bonferroni (α divided by the number of tests) and Bonferroni-Holm procedures are quite common as they strictly control the multiple type I error rate (Neuhauser 2006). Thus, in the case of 20 outcomes, in order to be statistically significant the p-value needs to be below 0.0025 ($0.05/20$) after correction for multiple testing according to the Bonferroni procedure.

In an interim analysis, trial data are analyzed by treatment group for study monitoring purposes before the final analysis. Reasons may be monitoring for superiority, harm or futility. Based on the results of interim analyses, trials may be stopped early, typically as evaluated by an independent data safety and monitoring board (DSMB). The number of interim analyses and definition of stopping rules accounting for multiple testing (e.g. according to O'Brien-Fleming, Peto or Pocock) needs to be pre-specified in the study protocol (Schulz & Grimes 2005). These rules define p-values for considering stopping a trial early depending

on the overall number of planned interim analyses and preserving the overall type I error rate (Schulz & Grimes 2005). In a trial with two interim analyses and one final analysis for instance, the p-value for the interim stopping level would be for the first interim, second interim and the final analysis 0.0005, 0.014 and 0.045 according to the rule of O'Brien-Fleming and 0.001, 0.001 and 0.05 according to the rule of Peto, the latter applying constant stopping levels until the final analysis (Schulz & Grimes 2005). To give an example, a study investigating surgery followed by radiotherapy versus radiotherapy alone for metastatic cancer spinal cord compression was stopped early by the DSMB at a planned interim analysis after recruitment of half of the foreseen patients for superiority of the surgical intervention arm at a p-value of 0.001 according to the rule of O'Brien-Fleming (Patchell et al. 2005). Stopping a trial early for superiority or futility needs however to be carefully evaluated. Empirical evidence indicates that trials having been stopped early for benefit tend to overestimate the underlying true treatment effect. Therefore, interim analysis should be well justified and if possible not be conducted for detecting an early benefit (Bassler et al. 2010).

Step 4: Choosing the appropriate design

Types of study design and potential biases

Depending on the research question, the appropriate study design needs to be chosen and described in the protocol. An overview of study designs (Guralnik 2014) with examples (Champagne et al. 2012; Emami et al. 2011; Law et al. 2010; Lu et al. 2013; Marks et al. 2011; Patel et al. 2012) is provided in table 4. A more detailed description of the pros and cons of different study designs is beyond the scope of the present article.

Bias, i.e. a systematic error, jeopardizes the internal (reliability and accuracy) and external validity (generalizability) of studies. Therefore, methods to minimize the risk of bias need to be outlined in the protocol (Montori, Permyer-Miralda, Ferreira-Gonzalez, Busse, Pacheco-Huergo, Bryant, Alonso, Akl, Domingo-Salvany, Mills, Wu, Schunemann, Jaeschke, & Guyatt 2005b; van den Broek, van Dam, van Breukelen, Bemelmans, Oussoultzoglou, Pessaux, Dejong, Freemantle, & Olde Damink 2011). Many different types of biases have been described and they may be classified in different ways, such as by the direction of resulting change in the estimate or by the stage of research, in which they occur (Delgado-Rodriguez and Llorca 2004). Examples are randomization or matched pairs to address selection bias, blinding to address performance and detection bias, and measures to reduce loss to follow-up to address attrition bias. An outline of biases and possible solutions to address them is provided in table 5 (Akobeng 2008; Bornhoft et al. 2006).

Table 4: Design types (adapted from Guralnik and Manolio (Guralnik 2014))

Study type	Descriptive			Analytical		
	Population	Individual		Observational		Interventional (Experimental)
	Ecological (correlational) study	Case Report/ case series	Cross-sectional (prevalence) study	Cohort study (retro-/prospective)	Case-control study (retro-/prospective)	(Non-) Randomized controlled trial
Exposure/ Outcome			Observation: Exposure and outcome simultaneously	Observation: Exposure® Outcome	Observation: Exposure→ Outcome	Assignment: Exposure
Example	Population level correlation between use of non-steroidal anti-inflammatory drugs and proton pump inhibitors and peptic ulcer bleeding (Lu, Sverden, Ljung, Soderlund, & Lagergren 2013)	Early experience of single-incision laparoscopic colectomy (Law, Fan, & Poon 2010)	Assessment of adoption of laparoscopic colon resection (Patel, Patel, Mahanti, Ortega, Ault, Kaiser, & Senagore 2012)	Single-incision versus standard laparoscopic cholecystectomy (retrospective cohort with historical control) in children (Emami, Garrett, Anselmo, Torres, & Nguyen 2011)	Matched case-control study of single-incision versus standard multiport laparoscopic colectomy (Champagne, Papaconstantinou, Parmar, Nagle, Young-Fadok, Lee, & Delaney 2012)	Randomized controlled trial of single incision versus standard laparoscopic cholecystectomy (Marks, Tacchino, Roberts, Onders, DeNoto, Paraskeva, Rivas, Soper, Rosemurgy, & Shah 2011)
Caveats	<ul style="list-style-type: none"> ü Ecological fallacy (group level association¹ individual level association) 	<ul style="list-style-type: none"> ü Low level of evidence ü Lacking generalizability 	<ul style="list-style-type: none"> ü Not suitable for rare/short duration diseases 	<ul style="list-style-type: none"> ü Confounding ü Choice of control group ü Only one exposure studied 	<ul style="list-style-type: none"> ü Confounding ü Choice of control group ü Recall bias ü Overmatching 	<ul style="list-style-type: none"> ü Equipoise required ü Ethics ü Resources

Table 5: Bias types (adapted from Akobeng (Akobeng 2008) and Bornhöft and colleagues (Bornhoft, Maxion-Bergemann, Wolf, Kienle, Michalsen, Vollmar, Gilbertson, & Matthiessen 2006))

Bias	Meaning	Example	Possible solutions
Selection bias	Intervention group differs from control group regarding baseline characteristics	Difference in age, severity of illness	<ul style="list-style-type: none"> ü Randomization ü Stratified randomization ü Matched pairs
Performance bias	Apart from investigated intervention, the intervention group is treated differently than the control group	When comparing two types of surgery for fracture treatment one group is followed more intensively by physiotherapist	<ul style="list-style-type: none"> ü Blinding ü Documentation of concomitant interventions
Detection bias	The outcomes are assessed differently in the intervention than in the control group	A small postoperative hematoma is regarded as complication in the control group, but not in the intervention group	<ul style="list-style-type: none"> ü Blinding ü Blinded outcome assessors ü Several outcome assessors ü Objective criteria
Attrition bias	The loss of participants from the study (i.e. drop-out, withdrawal for example because of deviation from the protocol) is different in the intervention and control group	A higher drop-out rate in the intervention group than in the control group may underestimate hernia recurrences in a trial comparing two hernia repair techniques	<ul style="list-style-type: none"> ü Measures to reduce drop-outs ü Documentation of patient flow including drop-outs ü Intention-to-treat analysis ü Declaration of strategies to deal with missing data (e.g. last observed value carried forward, best/worst case scenario assumption)

Randomization

The process of randomization comprises the allocation sequence generation, allocation concealment, i.e. neither the participant nor the investigators are able to predict the group assignment, and allocation sequence implementation (Akobeng 2008).

The protocol needs to specify how the randomization sequence is generated, examples are variable block size randomization, stratified randomization, or cluster randomization. Pseudo- or quasi-randomization (e.g. according to the date of birth, date of entry, patient ID, alternating) should be avoided, since allocation will be easily predictable.

Additionally, the method of information transfer needs to be described, e.g. central web-based randomization, central telephone randomization, or serially numbered opaque sealed envelopes. Central randomization is preferred because it is more reliable to ensure allocation concealment.

Since the surgeons' and institutions' expertise may have an impact on the clinical outcome respectively treatment effect (i.e. performance bias), an expertise-based design may be appropriate in certain circumstances (Devereaux, Bhandari, Clarke, Montori, Cook, Yusuf, Sackett, Cina, Walter, Haynes, Schunemann, Norman, & Guyatt 2005).

Blinding

Whenever possible, blinding should be considered and outlined in the study protocol. Studies may be unblinded (open-label), single-blind (i.e. patient-blind) or double-blind (patient and caregiver-blinded) (Day and Altman 2000). Other terms have been used such as triple-blind, referring to the patient, caregiver and assessor. Since there are other persons involved in a trial (data collector, outcome adjudicator, data analyst), it is best to describe in detail who is blinded and for what.

In surgery, blinding may be a challenge. To limit detection bias it is advisable to use blinded outcome assessment, e.g. through a separate team of assessors not involved in surgery. Sham or placebo surgery has been previously conducted (Moseley, O'Malley, Petersen, Menke, Brody, Kuykendall, Hollingsworth, Ashton, & Wray 2002), needs however special ethical justification.

Step 5: Description of the study procedures

A detailed description of all study procedures should be provided. Importantly, it should become clear, what is part of clinical routine and what is study-specific. Since surgical interventions are complex, they need to be standardized as outlined above in order to be able to draw generalizable conclusions from the study. Moreover, the surgeons' expertise or specific hospital standards may have an impact on the outcome, which needs to be accounted for (Boutron, Moher, Altman,

Schulz, & Ravaud 2008). If applicable, measures to assure compliance should be described.

An activity plan in the format of a table describing all activities during the study periods enrolment, allocation, postallocation and closeout such as screening procedures, intervention, and different types of outcome assessments with corresponding timelines and allowed deviance from the foreseen date could be helpful to provide an overview of all involved procedures (Cummings, Browner, & Hulley 2007). Discontinuation criteria for either study participants, parts of the trial, or the trial as a whole, should be described as well.

Step 6: Description of the study population

There are medical, methodological and ethical criteria to define a study population. The choice of participant inclusion and exclusion criteria will have an impact on the internal (reliability and accuracy) and external validity (generalizability) of a study (Akobeng 2008; Bornhoft, Maxion-Bergemann, Wolf, Kienle, Michalsen, Vollmar, Gilbertson, & Matthiessen 2006) and depends on the goal of the study, i.e. an explanatory trial testing efficacy with rigorous control of internal validity versus a pragmatic trial evaluating effectiveness under clinical real-life conditions (Godwin et al. 2003).

The inclusion of a vulnerable population, such as minor children or cognitively impaired adults, e.g. in emergency settings, needs special justification and measures of participant protection (Brody et al. 2005).

Step 7: Development of a statistical analysis plan

The goal of a quantitative assessment entails that the scientific research question is translated into a statistical problem. In the respective protocol section, the statistical methods should be described in sufficient detail including the statistical software to be used, the analysis population (e.g. intention-to-treat (ITT) or per protocol (PP)), descriptive/exploratory statistics, hypothesis-testing indicating the level of significance and taking into account the type of outcome, effect measures (with confidence intervals), type of sample (paired versus unpaired), assumption of data distribution (normally versus not normally distributed data) and modelling if applicable. **Table 6** summarizes the most common hypothesis tests and examples of regression models depending on the type of outcome variable (Kirkwood and Sterne).

Confidence intervals are preferred over p-values, since they provide information not only regarding statistical significance, but also about the smallest and largest plausible value of the effect measure of interest. Importantly, “absence of evidence” commonly does not equal “evidence of absence”, and statistical significance is not to be considered equivalent to clinical relevance (Alderson 2004).

According to the design, the analysis population (ITT versus PP) needs to be pre-specified. ITT refers to the population as randomized, regardless of factors such as compliance, crossover, or loss to follow-up, whereas PP refers to the patients actually treated and followed as foreseen in the protocol. In a superiority trial, the ITT analysis is preferred. It is conservative, since non-compliers generally reduce the treatment effect. In contrast, in an equivalence/non-inferiority design, due to the potentially reduced treatment effect, ITT is no longer conservative. Therefore, the PP analysis is the conservative and preferred primary analysis in the non-inferiority setting, complemented by an ITT analysis (Matilde and Chen 2006).

Procedures of handling missing data (e.g. last observation carried forward, best or worst case scenario imputation, multiple imputation, censoring) should be outlined as well as any planned interim analyses indicating the number, time point, and definition of stopping rule as outlined above. Subgroup investigations/interaction analyses should be determined in advance.

Step 8: Sample size calculation

The sample size is chosen to assure that the study will have sufficient power to allow conclusive inferences regarding the primary outcome, given the assumptions for the sample size calculation happen to be realistic. In a superiority trial, a sample size statement should include the α -level, power (equals $1-\beta$ or type II error), the event rate or value in the control group, the expected (or clinically relevant) effect in the experimental group, one- versus two-sided testing and the expected rate of loss to follow-up. In case of binary outcomes, the effects are estimated in proportions and in case of continuous normally distributed outcome measures, in means and standard deviations (SDs), the latter as a measure of variability in the two groups. **Table 7** summarizes the assumptions required for sample size calculation depending on the type of outcome variable. The source of information for the assumed treatment effects (e.g. literature, pilot study) should be indicated.

In equivalence and non-inferiority trials, the boundaries of equivalence and the non-inferiority margin respectively need to be pre-specified instead of the expected effect in the experimental group.

In case of non-normally distributed data, as a rule of thumb, the sample size may be computed for a two-sample t-test and then inflated by 15% (Lehmann 2014). Alternatively, the sample size may be determined using computer simulations or in case of availability of pilot or historical data, using bootstrap methods (Collings and Hamilton 1988).

Table 6: Hypothesis tests and multivariable analysis (adapted from Kirkwood and Sterne (Kirkwood & Sterne))

Outcome		Parametric	Non-parametric	Multivariable analysis (Examples)
Continuous	Unpaired measurements (comparison of two groups)	Unpaired t-test	Wilcoxon rank sum test (Mann-Whitney-U test)	Continuous: Linear regression Discrete: Poisson regression
	Paired measurements (comparison of two groups)	Paired t-test	Wilcoxon signed rank test, Sign test	
	Comparison of more than two groups	ANOVA (Analysis of Variance)	Kruskall-Wallis test	
Categorical	Unpaired measurements	Chi-squared test	Fisher's exact test ¹	Binary: Logistic regression
	Paired measurements	McNemar's chi-squared test	Methods based on exact probabilities	
Time-to-event		Log-rank test		Regression analysis of survival/time-to-event data (e.g. Cox or Poisson regression or parametric models)

¹ Applicable if the expected value (under the null) of any of the cells in the table <5

Table 7: Sample size calculation

Univariable/ Multi- variable	Assumptions/Sample size		Consider	Example binary outcome variable: Complication yes/no	Example quasi-continuous outcome variable (normally distributed): Quality of life score
Univariable	Assump- tions	Significance level @ type I error (Study erroneously rejects the null, i.e. claims a difference although there is none)	Most typically ü 0.05 ü 0.01	0.05	0.05
		Power @ type II error (Study erroneously accepts the null, i.e. finds no difference although there is one)	Most typically ü 80% ü 90%	80%	90%
		Effect control group	ü Literature ü Pilot study	10%	mean 35.5
		Effect intervention group	ü Clinically relevant difference	5%	mean 45.5
		Standard deviation/variance	ü Assume equal SD ¹ in intervention/control group	Not necessary, will be derived from the chosen effects	SD 9.0 in control and intervention group
		One-/two sided	ü Superiority design most typically two-sided	Two-sided	Two-sided

¹ SD= Standard deviation

Table 7: Sample size calculation (continued)

Univariable/ Multi- variable	Assumptions/Sample size		Consider	Example binary outcome variable: Complication yes/no	Example quasi-continuous outcome variable (normally distributed):Quality of life score
Univariable	Sample size	Calculated sample size ¹	Lack of normality for continuous outcome® ü Inflate sample size by 15% (Lehmann 2014) Estimate sample size with simulation (Collings & Hamilton 1988)	434 per group or 868 in total	18 per group or 36 in total
		Final sample size accounting for drop-outs	ü Literature Pilot study	10% drop out, resulting in 478 per group or 956 in total	20% drop out, resulting in 22 per group or 44 in total
		Sample size statement		Assuming a 10% drop-out rate, a sample size of 478 per group is necessary to have a 80% chance of detecting, as significant at the 5% level, a decrease in complication rate from 10% in the control group to 5% in the intervention group.	Assuming a drop-out rate of 20%, 44 patients are required to have a 90% chance of detecting, as significant at the 5% level, an increase in the primary outcome measure from 35.5 (SD 9.0) in the control group to 45.5 (SD 9.0) in the experimental group.

¹ Using R version 2.14.2 (R Foundation for Statistical Computing, Vienna, Austria)

Table 7: Sample size calculation (continued)

Univariable/ Multi- variable	Assumptions/Sample size		Consider	Example binary outcome variable:Complication yes/no	Example quasi-continuous outcome variable (normally distributed):Quality of life score
Multivariable	Rule of thumb	<ul style="list-style-type: none"> ü binary outcome: 10 events per variable (Peduzzi et al. 1996) ü continuous outcome: 10-15 observations per variable (Babyak 2004) 	<ul style="list-style-type: none"> ü Avoid overfitting (do not include too many variables) 	To adjust for age, surgeon experience (high vs ¹ low) and ASA (≥ 3 vs <3) (i.e. 1 independent variable & 3 confounders), a minimum of 40 patients with/without complication (whichever is the smaller percentage) need to be observed.	To adjust for age, gender and ASA (≥ 3 versus <3) (i.e. 1 independent variable & 3 confounders), a minimum of 40 patients need to be observed.

¹ vs= versus

In multivariable analysis, particularly relevant when analyzing observational data, as a rule of thumb a minimum of 10 to 15 events, respectively non-events, per variable in the model are necessary in order to achieve reliable estimates from logistic regression (binary outcome) (Peduzzi, Concato, Kemper, Holford, & Feinstein 1996) and 10 to 15 observations in multiple linear regression (continuous outcome) (Babyak 2004). For example in a study evaluating risk factors to develop a surgical site infection after hernia repair, a minimum of 10 surgical site infections and of 10 non-surgical site infections are necessary to evaluate one candidate risk factor, whereas to evaluate predictors of length of hospital stay, 10 patients per predictor should be included (table 7).

In retrospective studies and pilot studies, there is generally no formal sample size calculation, but a plausible rationale for the choice of sample size should be provided.

Step 9: Description of data management and quality assurance

The process of data entry, data management, monitoring, quality control and quality assurance should be described. A statement of permitting access to source data for the purpose of audits and inspections by the IEC and regulatory authorities should be included. The process of privacy protection (e.g. reversible anonymization) and the duration of data storage should be described. A detailed description of data management and quality assurance options is beyond the scope of this article.

Step 10: Ethical considerations

Under ethical considerations, a risk-benefit assessment should be presented. Potential benefits, risks, but also inconveniences should be mentioned. These should refer to the individual study participant, may however also include potential benefits for future patients. The inclusion of a vulnerable population should be further elaborated and justified in this section. Other ethical aspects should be mentioned here, such as participation being entirely voluntary and withdrawal being possible at any time without giving any reason and without any impact on patient management. The handling of incidental findings and genetic information as well as the justification of placebo procedures, if applicable, should be included. Additionally, a statement that the study will be conducted according to the study protocol and Good Clinical Practice (GCP) should be included as well as that the study protocol and any potential amendments will be submitted to an IEC and potential regulatory authorities. A funding statement, a description of any potential conflicts of interest, and insurance issues should be provided. Any clinical study with ethical approval should disseminate its results through publication; the agreed publication policy can complete this section.

2.3. Conclusion

The study protocol is the core document when planning and conducting clinical research. It should be created in an interdisciplinary setting, approved and strictly followed. Any changes require an amendment approved by an IEC and the regulatory authorities. The ultimate goal of the protocol is to support the conduct of scientifically and ethically sound research providing high-level evidence relevant for healthcare decision-making.

3. The use of systematic reviews when designing and reporting surgical trials

Rachel Rosenthal¹, MD, MSc, Heiner C. Bucher², MD, MPH, Kerry Dwan³, PhD

¹Department of Surgery, University Hospital Basel, Basel, Switzerland

²Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, Basel, Switzerland

³Department of Biostatistics, The University of Liverpool, Liverpool, United Kingdom

Disclosure information: There is no funding related to this publication. There are no conflicts of interest to declare.

Publication information: This is a non-final version of an article published in final form in the Annals of Surgery (<http://journals.lww.com/annalsofsurgery/>): Rosenthal R, Bucher HC, Dwan K. The use of systematic reviews when designing and reporting surgical trials. Ann Surg 2015 Jan 23 [Epub ahead of print] DOI: 10.1097/SLA.0000000000001092 (Rosenthal et al. 2015a)

Abstract

Background: When planning clinical trials, previous trials in the field should be considered to inform trial design. Likewise, the findings of any trial should be set in context with the overall body of evidence. We aim to investigate to what extent information from systematic reviews (SR) is used to inform trial design and report trial results in general surgical journals.

Methods: All randomized controlled trials (RCTs) published in 2010 in the *Annals of Surgery*, *JAMA Surgery* and the *British Journal of Surgery* were considered.

Results: Of 596 studies, 51 RCTs were identified. SRs were referenced in 65%, either to summarize evidence concerning a related topic (43%) or the study topic (excluding first-in-area RCTs: introduction 24%, discussion 26%). No SRs were used to inform trial design and no trial updated a SR integrating the new results.

Conclusions: SR are rarely used to inform trial design and synthesize knowledge for informed decision-making.

3.1. Introduction

Clinical research projects typically target new research questions or the extension, confirmation, or rejection of previous findings. Therefore, a prerequisite for any clinical research project is an extensive literature review. This is especially of importance when planning clinical trials. Evidence from systematic reviews (SR) and when applicable meta-analyses should be considered to inform the design of clinical trials (Thompson et al. 2013). Moreover, when reporting trial results, these should be used to update previous SR, as recommended for the discussion section by the CONSORT group (“Ideally, we recommend a systematic review and indication of the potential limitation of the discussion if this cannot be completed”) (Altman et al. 2001).

The use of SRs to inform the design of new trials is important from an ethical, scientific, but also economic point of view. Some funding agencies ask for knowledge synthesis in grant applications. The Canadian Institutes of Health Research, requires grant applicants planning a randomized controlled trial (RCT) to include in their application a SR (Graham 2012). Similarly, for the National Institute for Health Research (NIHR) Research for Patient Benefit Programme, “all proposals must show evidence from systematic reviews to ensure patient safety and value for money” (National Institute for Health Research 2013). However, knowledge synthesis prior to conducting an RCT is not uniformly applied. An investigation of RCTs funded by the National Institute for Health Research Health Technology Assessment (NIHR HTA) between 2006 and 2008 including data from 48 trials found 77.1% of the trials referencing a SR (Jones et al. 2013). However, only 41.7% of the trials informed the design of the proposed RCT using information from a SR. SRs were used to define the primary outcome and description of adverse events, as well as for sample size calculation and determination of duration of follow-up. A repeated investigation of RCTs published in the May issue of the *Annals of Internal Medicine*, *BMJ*, *JAMA*, *Lancet* and *New England Journal of Medicine* in 1997 (Clarke and Chalmers 1998), 2001 (Clarke et al. 2002), 2005 (Clarke et al. 2007), 2009 (Clarke et al. 2010), and 2012 (Clarke and Hopewell 2013), evaluated whether the trial reports referred to the existing body of evidence in the discussion section and, since 2005, if SRs were used in the introduction section. The results suggest that the proportion of trials synthesising new findings with previous findings is low (39% in 2012 excluding trials that were the first addressing the question) without an apparent progress over the years (Clarke & Hopewell 2013).

To the best of our knowledge, the use of SR to justify the planned research, inform the design and finally to be updated integrating the new results has not previously been investigated in surgical trials. The aim of the current research is to investigate to what extent information from SR is used 1) to justify and 2) design trials, as well as 3) to synthesize results, evaluating all RCTs published in 2010 in the *Annals of Surgery*, *JAMA Surgery* and the *British Journal of Surgery*.

3.2. Methods

Here we present a secondary analysis of a review investigating discrepancies between registry entries and final reports of RCTs published in 2010 in the *Annals of Surgery*, *JAMA Surgery* and the *British Journal of Surgery* (Rosenthal & Dwan 2013).

The three Journals were screened for RCTs and study characteristics were extracted as previously described (Rosenthal & Dwan 2013). For the present investigation, two reviewers (RR and KD) independently extracted information concerning the use of SRs in the introduction section (justification of the research) and discussion section (synthesis of results), adapted from the extraction forms previously described for five reviews of major medical journals conducted between 1997 and 2012 (Clarke & Chalmers 1998) (Clarke, Alderson, & Chalmers 2002) (Clarke, Hopewell, & Chalmers 2007) (Clarke, Hopewell, & Chalmers 2010) (Clarke & Hopewell 2013). Additionally, the methods sections were screened for statements that the trial design had been informed by a SR, adapting the extracting scheme applied to evaluate the use of systematic reviews to inform the study design in grant applications to the NIHR HTA (Jones, Conroy, Williamson, Clarke, & Gamble 2013). Discrepancies between the two reviewers were resolved by discussion and agreement. For data extraction, an Excel spreadsheet (Microsoft Office XP, Microsoft Corporation, Redmond, WA, USA) was used. Descriptive statistics were conducted using Intercooled Stata Version 12.1; StataCorp LP, College Station, Texas, USA.

3.3. Results

The flow of included studies and individual study details were published previously (Rosenthal & Dwan 2013). In brief, out of 596 studies identified through the search of the three journals, 51 RCTs fulfilled criteria for data extraction. Main baseline characteristics are presented in [Table 8](#).

Overall, out of the 51 RCTs, one or several SRs were referenced anywhere in the final report in 33 studies (65%), in 22 studies (43%) to summarize evidence concerning a different topic (example: RCT investigating early versus late drain removal after pancreatectomy cites a SR about drain versus no drain after colorectal surgery). None of the RCTs mentioned in the introduction or methods section that a SR was used to inform the trial design and no RCT contained an updated SR in the discussion section integrating the new results. Details on the findings of knowledge synthesis in the design and reporting phase, respectively, are presented in [Table 9](#).

Table 8: Baseline characteristics of the included studies (n=51)

Characteristic	Item	Summary measure
Study intervention, n (%)	Surgery	16 (31%)
	Invasive intervention ¹	4 (8%)
	Both surgery and invasive intervention	4 (8%)
	The intervention is neither surgery nor an invasive intervention, but surgery is part of the study ²	22 (43%)
	Conservative treatment	3 (6%)
	Informed consent process	2 (4%)
Primary outcome is safety, n (%)	No	39 (76%)
	Yes	12 (24%)
Sample size ³ , median (IQR)		122 (70, 206)
Funding, n (%)	Trial receiving any support from industry ⁴	19 (37%)
	Trial not receiving any support from industry	22 (43%)
	Missing/insufficient information	10 (20%)

Seventeen RCTs claimed to be the first RCTs in the topic area, two stated this explicitly (i.e. “this is the first trial...”) and 15 implicitly (i.e. “previously no trial investigated...”) and were excluded from further analysis. Out of the remaining 34 trials, eight (24%) RCTs discussed a SR in the topic area in the introduction section without mentioning that it was used to inform the trial design and nine (26%) discussed a previous SR in the discussion section without attempting to integrate the new results and in no trials were findings used to update a SR.

¹ Interventions requiring either percutaneous access (such as radiofrequency ablation of liver metastases) or access through natural orifices (such as endoscopic retrograde cholangio-pancreatography, but not natural orifice transluminal endoscopic surgery)

² Surgery was part of the study, but the intervention in itself was not, such as peri- or intraoperative administration of a drug, nutrition, bowel preparation, or transplant organ preparation

³ total number of randomized patients

⁴ be it by funding the entire trial, part of the trial, or any other support (e.g, support of investigators)

Table 9: Use of systematic reviews to inform trial design by publication section

Characteristic	Item	N(%)
Introduction	Claims to be the first RCT ¹ addressing the question ²	17 (33%)
	Contains an updated SR ³ , which was used to design the new trial	0 (0%)
	Contains an updated SR, not mentioning that it was used to design the new trial	0 (0%)
	Discusses a previous SR in the topic area ⁴ of the trial, which was used to design the new trial	0 (0%)
	Discusses a previous SR in the topic area of the trial, not mentioning that it was used to design the new trial	8 (16%)
	Does not contain a reference to a SR, but contains references to (an) other RCT(s) addressing the topic area	20 (39%)
	Does not contain a reference to a SR, does not contain references to (an) other RCT(s) addressing the topic area, and does not claim to be the first RCT	6 (12%)
Methods	Claims in the methods section to be the first RCT addressing the question	0 (0%)
	Refers to a previous or updated SR, which was used to design the new trial	0 (0%)
	Does not contain reference to a previous or updated SR, which was used to design the new trial and does not claim in the methods section to be the first trial	51 (100%)

35

¹ RCT=randomized controlled trial

² Explicitly stated to be the first or derived from statements that no such study has previously been conducted

³ SR=systematic review

⁴ Topic area= same or similar intervention and condition

Table 9 (continued)

Characteristic	Item	N(%)
Discussion	Contains an updated SR ¹ integrating the new results	0 (0%)
	Discusses a previous SR in the topic area ² of the new trial, but did not attempt to integrate their results	9 (18%)
	No apparent systematic attempt to set the results in the context of other trials	20 (39%)
	Claims in the discussion section to be the first RCT ³ addressing the question ⁴	8 (15%)
	Claims in another section than the discussion section to be the first RCT addressing the question ²	9 (18%)
	Claims in any section to be the first RCT with exactly the same population, intervention, control and outcome (but not necessarily the first RCT in the topic area)	5 (10%)

¹ SR=systematic review

² Topic area=same or similar intervention and condition

³ RCT=randomized controlled trial

⁴ Explicitly stated to be the first or derived from statements that no such study has previously been conducted

3.4. Discussion

Our results show that two-thirds of the RCTs referenced a SR, however none mentioned the use of a SR to inform the trial design and none of the RCTs presented an updated existing SR integrating the new results. Additionally, we found only two RCTs to include explicit statements on whether the trial was the first RCT in the field. Our findings suggest that SRs are considered rather to summarize findings than to inform trial design or for knowledge synthesis after trial conduct. This is in line with previous findings (Clarke & Chalmers 1998) (Clarke, Alderson, & Chalmers 2002) (Clarke, Hopewell, & Chalmers 2007) (Clarke, Hopewell, & Chalmers 2010) (Clarke & Hopewell 2013), although even more marked. However, it is not possible to identify circumstances when a SR may have been used to inform trial design without explicit statements.

In practice, the following four-step framework has been suggested for the use of SR to inform the design of new trials: In step one, the research question of the proposed trial is formulated focussing on definition of the populations, interventions, comparators, outcomes, timing and setting (PICOTS) (Thompson, Tiwari, Fu, Moe, & Buckley 2013). In step two, an up-to-date relevant and valid SR is identified or conducted, and in step three the SR is used to inform the planned trial. If there are several RCTs, but no SR in the field, we propose to first synthesize the body of evidence within a SR and if appropriate conduct a meta-analysis. Finally, in step four the implications for the proposed trial are summarised. This procedure is important to justify any research from a scientific, ethical and economic point of view. Similarly, we propose that trialists attempt to report their results in the context of other trials in the field, whenever possible integrating the new findings by updating a SR. Only when evaluating study results within the overall body of evidence, well-informed decisions in health care are possible.

4. Completion and publication rates of surgical randomized controlled trials – an empirical study

Rachel Rosenthal, MD, MSc¹, Benjamin Kasenda, MD^{2,3}, Salome Dell-Kuster, MD, MSc^{1,2}, Erik von Elm, MD, MSc⁴, John You MD, MSc^{5,6}, Anette Blümle, PhD⁷, Yuki Tomonaga, MSc²⁵, Ramon Saccilotto, MD^{1,2}, Alain Amstutz, BSc², Theresa Bengough, MA⁴, Joerg J. Meerpohl, MD⁷, Mihaela Stegert, MD², Kari A. O. Tikkinen, MD, PhD^{5,8,9}, Ignacio Neumann, MD, MSc^{5,26}, Alonso Carrasco-Labra DDS, MSc^{5,10}, Markus Faulhaber, MD, MSc^{5,27}, Sohail Mulla, MSc⁵, Dominik Mertz MD, MSc^{5,11}, Elie A. Akl, MD, PhD, MPH^{5,12,28}, Dirk Bassler, MD, MSc¹³, Jason W. Busse, DC, PhD^{5,14,15}, Ignacio Ferreira-González, MD, PhD¹⁶, Francois Lamontagne, MD, MSc¹⁷, Alain Nordmann, MD, MSc², Viktoria Gloy, PhD^{2,24}, Kelechi K. Olu, MD, MSc², Heike Raatz, MD, MSc², Lorenzo Moja, MD, MSc¹⁸, Shanil Ebrahim, PhD^{5,14,19,20}, Stefan Schandelmaier, MD^{2,21}, Xin Sun, PhD^{5,22}, Per O. Vandvik MD, PhD²³, Bradley C. Johnston, PhD^{5,19}, Martin A. Walter, MD²⁴, Bernard Burnand, MD, MPH⁴, Matthias Schwenkglenks, PhD, MPH²⁵, Lars G. Hemkens, MD, MPH², Heiner C. Bucher, MD, MPH², Gordon H. Guyatt, MD, MSc⁵, and Matthias Briel, MD, MSc^{2,5}

¹ Department of Surgery, University Hospital Basel, Spitalstrasse 26, 4031 Basel, Switzerland

² Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital of Basel, Hebelstrasse 10, 4031 Basel, Switzerland

³ Department of Oncology, Basel University Hospital, Basel, Switzerland

⁴ Cochrane Switzerland, Institute of Social and Preventive Medicine (IUMSP), Lausanne University Hospital, Lausanne, Switzerland

⁵ Department of Clinical Epidemiology and Biostatistics, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada

⁶ Department of Medicine, McMaster University, 1280 Main Street West, HSC-2C8, Hamilton, Ontario L8S 4K1, Canada

⁷ German Cochrane Centre, Department of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Berliner Allee 29, 79110 Freiburg, Germany

⁸ Department of Public Health, University of Helsinki, Helsinki, Finland

⁹ Department of Urology, Helsinki University Central Hospital and University of Helsinki, Haartmaninkatu 4, 00290 Helsinki, Finland

¹⁰ Evidence-Based Dentistry Unit, Faculty of Dentistry, Universidad de Chile, Santiago, Chile

¹¹ Michael G. DeGroote Institute for Infectious Diseases Research, McMaster University, Hamilton, ON, Canada

¹² Department of Internal Medicine, American University of Beirut, P.O. Box: 11-0236, Riad-EI-Solh, Beirut 1107 2020, Lebanon

¹³ Center for Pediatric Clinical Studies, Department of Neonatology, University Children's Hospital, Calwerstr. 7, 72076 Tuebingen, Germany and Division of

Neonatology, University Hospital Zurich, Frauenklinikstrasse 10, 8091 Zurich, Switzerland

¹⁴ Department of Anesthesia, McMaster University, 1280 Main Street West, HSC-2C8, Hamilton, Ontario L8S 4K1, Canada

¹⁵ Michael G. DeGroot Institute for Pain Research and Care, McMaster University, Hamilton, Ontario, Canada

¹⁶ Epidemiology Unit, Department of Cardiology, Vall d'Hebron Hospital and CIBER de Epidemiología y Salud Pública (CIBERESP), Passeig Vall d'Hebron 119-129, 08005 Barcelona, Spain

¹⁷ Centre de Recherche Clinique Étienne-Le Bel and Department of Medicine, Université de Sherbrooke, CHUS, 3001 12e avenue nord, Sherbrooke, PQ, Canada J1H5N4

¹⁸ IRCCS Orthopedic Institute Galeazzi, Via Riccardo Galeazzi, 4 20161 Milano, Italy

¹⁹ Department of Anesthesia & Pain Medicine, Hospital for Sick Children, 555 University Ave, Toronto, Ontario, Canada; and Institute of Health Policy, Management and Evaluation, University of Toronto, 155 College St, Toronto, Ontario, Canada

²⁰ Stanford Prevention Research Center, Department of Medicine, Stanford University, Stanford, California 94301, USA

²¹ Academy of Swiss Insurance Medicine, University Hospital of Basel, Schanzenstrasse 55, 4031 Basel, Switzerland

²² Chinese Evidence-Based Medicine Center, West China Hospital, Sichuan University, Chengdu, China

²³ Department of Medicine, Innlandet Hospital Trust-Division Gjøvik, Oppland, and Norwegian Knowledge Centre for the Health Services, Oslo, Norway

²⁴ Institute of Nuclear Medicine & Department of Clinical Research, University Hospital Bern, 3010 Bern, Switzerland

²⁵ Institute of Social and Preventive Medicine, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland

²⁶ Department of Internal Medicine, Pontificia Universidad Católica de Chile, Santiago, Chile

²⁷ Department of Family Medicine, McMaster University, Hamilton, Ontario, Canada

²⁸ Department of Medicine, State University of New York at Buffalo, Buffalo, New York, USA

Disclosure information: This study was funded by the Swiss National Science Foundation (grant 320030_133540/1) and the German Research Foundation (grant EL 544/1-2). MB, AN, VG, HR, SDK, LGH, and HCB are supported by Santésuisse and the Gottfried and Julia Bangerter-Rhyner-Foundation. During study preparation, EvE was supported by the Brocher Foundation. KAOT is supported by unrestricted grants from the Finnish Cultural Foundation and the Finnish Medical Foundation. JWB is funded by a New Investigator Award from the

Canadian Institutes of Health Research and Canadian Chiropractic Research Foundation. DM is a recipient of a Research Early Career Award from Hamilton Health Sciences Foundation (Jack Hirsh Fellowship). SE is supported by MITACS Elevate and Sickkids Restrcomp Fellowship Awards. SM is supported by an Ontario Graduate Scholarship.

There are no conflicts of interest to declare.

Publication information: This is a non-final version of an article published in final form in the Annals of Surgery (<http://journals.lww.com/annalsofsurgery/>):

Rosenthal R, Kasenda B, Dell-Kuster S, von Elm E, You J, Blümle A, Tomonaga Y, Saccilotto R, Amstutz A, Bengough T, Meerpohl JJ, Stegert M, Tikkinen KA, Neumann I, Carrasco-Labra A, Faulhaber M, Mulla S, Mertz D, Akl EA, Bassler D, Busse JW, Ferreira-González I, Lamontagne F, Nordmann A, Gloy V, Olu KK, Raatz H, Moja L, Ebrahim S, Schandelmaier S, Sun X, Vandvik PO, Johnston BC, Walter MA, Burnand B, Schwenkglenks M, Hemkens LG, Bucher HC, Guyatt GH, Briel M. Completion and publication rates of randomized controlled trials in surgery: an empirical study. *Ann Surg* 2015;262(1):68-73 (Rosenthal et al. 2015b)

Abstract

Objective: To investigate the prevalence of discontinuation and non-publication of surgical versus medical randomized controlled trials (RCTs) and to explore risk factors for discontinuation and non-publication of surgical RCTs.

Summary background data: Trial discontinuation has significant scientific, ethical, and economic implications. To date, the prevalence of discontinuation of surgical RCTs is unknown.

Methods: All RCT protocols approved 2000-2003 by six ethics committees in Canada, Germany and Switzerland were screened. Baseline characteristics were collected and, if published, full reports retrieved. Risk factors for early discontinuation for slow recruitment and non-publication were explored using multivariable logistic regression analyses.

Results: In total, 863 RCT protocols involving adult patients were identified, 127 in surgery (15%) and 736 in medicine (85%). Surgical trials were discontinued for any reason more often than medical trials (43% *versus* 27%, risk difference 16% (95% confidence interval [CI] 5%, 26%); $p=0.001$) and more often discontinued for slow recruitment (18% *versus* 11%, risk difference 8% (95% CI 0.1%, 16%); $p=0.020$). The percentage of trials not published as full journal article was similar in surgical and medical trials (44% *versus* 40%, risk difference 4% (95% CI -5%, 14%); $p=0.373$). Discontinuation of surgical trials was a strong risk factor for non-publication (odds ratio 4.18, 95% CI 1.45, 12.06; $p=0.008$).

Conclusions: Discontinuation and non-publication rates were substantial in surgical RCTs and trial discontinuation was strongly associated with non-publication. These findings need to be taken into account when interpreting surgical literature. Surgical trialists should consider feasibility studies before embarking on full-scale trials.

4.1. Introduction

Randomized controlled trials (RCTs) can provide high-level evidence about safety and efficacy of interventions. Conducting RCTs involving surgical interventions presents challenges distinct from RCTs investigating pharmacological interventions. Examples are standardization of operative and peri-operative interventions, surgeon and team experience, blinding of study personnel and participants and recruitment (e.g. due to patient preference) (Boutron, Moher, Altman, Schulz, & Ravaud 2008; Devereaux, Bhandari, Clarke, Montori, Cook, Yusuf, Sackett, Cina, Walter, Haynes, Schunemann, Norman, & Guyatt 2005; Solomon and McLeod 1995).

Trials may be discontinued earlier than planned for various reasons, including compelling evidence of treatment benefit or harm, futility, slow recruitment, commercial reasons or the emergence of new evidence from other trials that negates the need for another study (Goodman 2009; Iltis 2005). When there is no early evidence of harm or irrefutable benefit, trial discontinuation has a significant scientific, ethical and economic impact. The involvement of steering committees and Data and Safety Monitoring Boards with members independent of the sponsor is of utmost importance when deciding to stop a trial early. Moreover, reasons for early trial termination should be transparently communicated. To the best of our knowledge, discontinuation of surgical trials as well as the associated risk-factors has not yet been investigated.

We determined the prevalence of, and reasons for, discontinuation of surgical trials. We further explored differences in the prevalence of discontinuation between medical and surgical trials and risk factors for non-publication of surgical trials.

4.2. Methods

Included studies

We identified RCTs conducted in adult patients from a large empirical study investigating trial protocols approved between 2000 and 2003 by six Research Ethics Committees (RECs) in Canada, Germany and Switzerland (Kasenda et al. 2014). Details of the study design have been previously described (Kasenda et al. 2012).

For the current analysis, we used the following pre-specified definitions: All RCTs conducted among adult patients in clinical areas with surgical or peri-operative activities were classified as “surgical”. These included: anaesthesiology, general surgery, vascular surgery, transplantation, orthopaedics, traumatology, cardiothoracic surgery, cardiovascular surgery, neurosurgery, urology, plastic surgery, maxillofacial surgery, ear-nose-throat surgery, obstetrics/gynaecology, ophthalmology and dentistry. All RCTs conducted among adult patients in clinical areas primarily relying on non-surgical activities were classified as “medical”.

Examples are cardiology, gastroenterology, infectious diseases, neurology, and oncology (for full details see AppendixTable S1). In both groups, study interventions were classified as non-invasive (e.g. disinfection with antiseptic agent in a “surgical” trial, antihypertensive medication in a “medical” trial) or invasive (e.g. colon resection in a “surgical” trial, colonoscopy in a “medical” trial).

Data extraction

We used a web-based password-protected database for data extraction (<http://www.squieker.org/>). Collaborators trained in trial methodology signed confidentiality declarations for the participating RECs, completed a calibration process, and then extracted study characteristics such as information concerning the study population, intervention, control, sample size, and funding from the included RCT protocols. The first 310 (30%) trial protocols were extracted independently and in duplicate; disagreements were resolved by discussion. For pragmatic reasons, the remaining protocols were extracted by a single investigator with regular agreement checks using double data extraction. We determined completion status and publication history of RCTs as of April 2013 by using information from REC files if available, by conducting comprehensive searches of electronic databases for corresponding publications, and directly contacting trialists as described previously (Kasenda, von Elm, You, Blumle, Tomonaga, Saccilotto, Amstutz, Bengough, Meerpohl, Stegert, Tikkinen, Neumann, Carrasco-Labra, Faulhaber, Mulla, Mertz, Akl, Bassler, Busse, Ferreira-Gonzalez, Lamontagne, Nordmann, Rosenthal, Schandelmaier, Sun, Vandvik, Johnston, Walter, Burnand, Schwenkglens, Bucher, Guyatt, & Briel 2012). Two investigators independently assessed whether the publications identified by electronic searches matched the corresponding protocol (Kasenda, von Elm, You, Blumle, Tomonaga, Saccilotto, Amstutz, Bengough, Meerpohl, Stegert, Tikkinen, Neumann, Carrasco-Labra, Faulhaber, Mulla, Mertz, Akl, Bassler, Busse, Ferreira-Gonzalez, Lamontagne, Nordmann, Rosenthal, Schandelmaier, Sun, Vandvik, Johnston, Walter, Burnand, Schwenkglens, Bucher, Guyatt, & Briel 2012).

An RCT was considered as discontinued if the investigators indicated discontinuation with a reason in the correspondence with the REC, in a journal publication, or their response to our survey. If we could not elucidate the reason for trial discontinuation or if poor participant recruitment was mentioned, we used a pre-specified cut-off of less than 90% of achieved target sample size to determine discontinuation (Kasenda, von Elm, You, Blumle, Tomonaga, Saccilotto, Amstutz, Bengough, Meerpohl, Stegert, Tikkinen, Neumann, Carrasco-Labra, Faulhaber, Mulla, Mertz, Akl, Bassler, Busse, Ferreira-Gonzalez, Lamontagne, Nordmann, Rosenthal, Schandelmaier, Sun, Vandvik, Johnston, Walter, Burnand, Schwenkglens, Bucher, Guyatt, & Briel 2012). In the rare case that several reasons per study were indicated, the primary reason reported was used. Results were regarded as not published at all if the REC files, the contacted investigators, and the electronic database search yielded no evidence that the data were

published in any format. Additionally, it was evaluated whether results were published as full journal article. We defined sponsorship depending on who took responsibility for the trial: We considered protocols as industry-sponsored, if they were written by industry employees, prominently displayed a company logo or name, reported full funding by the industry without any indication that the trial was investigator-sponsored, or mentioned a specific policy with respect to data ownership and publication rights suggesting sponsorship by the industry. In surgical trials, we explored whether authors reported a primary outcome for harm. One investigator coded this information based on the previously extracted information and any ambiguity was discussed with a second investigator. The rationale to include this variable is the fact that adverse events may be rare or occur late in the sequence of the trial, thus adequately powered studies reaching the targeted sample size and the defined follow-up are of utmost importance.

Statistical analysis

We used descriptive statistics to report the prevalence of, and reasons for discontinuation. We compared features of surgical and medical RCTs using the Chi-squared test or Fisher's exact test, as appropriate, and calculated a risk difference and its 95% confidence interval (CI). We excluded trials that never started (according to REC files or our investigator survey) or that were still ongoing when calculating rates of trial completion and publication status.

We built multivariable logistic regression models to identify predictors for discontinuation due to slow recruitment. Since there are well-justified reasons for trial discontinuation, we did not explore reasons for discontinuation overall, but for discontinuation due to slow recruitment only, which we estimated to be an inappropriate reason for early termination. In order to avoid overfitting, the number of included covariates was limited to one per 10 events/non-events, whichever was the smaller number in all multivariable regression analyses (Peduzzi, Concato, Kemper, Holford, & Feinstein 1996). Covariates were chosen a priori to explore the following two hypotheses: trial discontinuation due to slow recruitment is (1) less likely with industry-sponsored trials than with trials sponsored by an academic investigator, due to differences in the amount of financial and organisational resources between industry and academic sponsors; and (2) less likely with adverse events being the primary outcome than with efficacy outcomes only, due to more patient and health care provider preference for one treatment arm in the latter case hampering recruitment.

Using a multivariable hierarchical logistic regression model, we examined the following pre-specified predictors for non-journal-publication of RCTs: 'industry-sponsorship (*versus* investigator-sponsorship)', 'primary outcome being an adverse event (yes *versus* no)' and 'trial discontinuation for any reason (yes *versus* no)'. We hypothesized (1) that non-publication was more likely in industry-sponsored trials, as suggested by previous literature (Von Elm et al. 2008); (2) that trial non-publication was more likely in discontinued trials (assuming that only few

trials were discontinued due to compelling evidence of treatment benefit or harm); and (3) that trial non-publication was more likely with adverse events being the primary outcome than with efficacy outcomes only, due to the high interest in efficacy outcomes.

We accounted for clustering by the approving REC using a random effect estimator.

We conducted two sensitivity analyses. First, we conducted the multivariable analyses disregarding the clustering by the approving REC. Second, we used multiple imputation to replace missing data for independent and dependent variables (Kenward and Carpenter 2007).

We conducted one posthoc analysis: Since trial comparisons (surgical *versus* medical) refer to the a priori defined clinical areas and not to study interventions, we compared trials with invasive interventions *versus* trials without invasive interventions using descriptive analyses (Appendix Figure S1).

Analyses were conducted using Stata Version 12.1 (StataCorp LP, College Station, Texas, USA).

4.3. Results

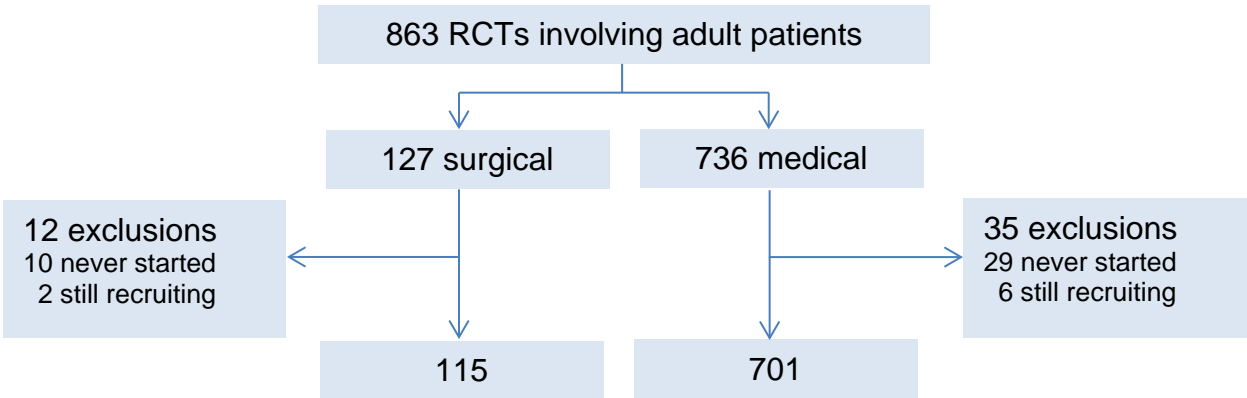
Baseline characteristics of included studies

We identified 863 protocols for RCTs involving 680,019 adult patients; 127 RCTs in the field of surgery (15%) and 736 in medicine (85%) (Figure 2). We excluded 39 (5%) trials that were never started and 8 (1%) that were still recruiting. A total of 816 trials (95%) involving 666,760 adult patients remained for the evaluation of completion and publication status (surgery: 115 (14%), medicine: 701(86%), (Figure 2).

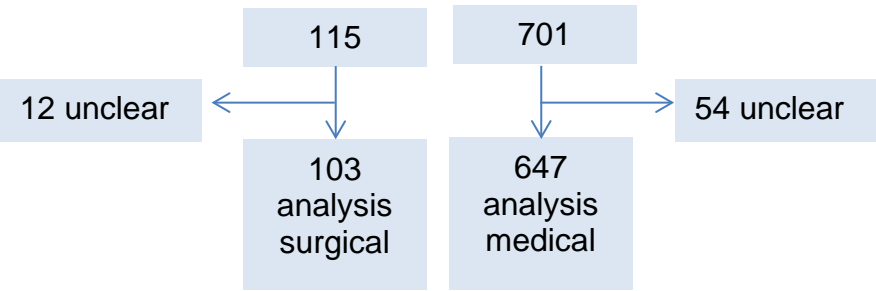
The specialties contributing to at least 10% of the trials were anesthesiology (13%), cardiothoracic surgery (13%), obstetrics/gynecology (16%), ophthalmology (14%) and orthopedics (10%) in surgical trials (Appendix Table S1a) and oncology (21%), cardiovascular (15%) and infectious diseases (11%) in medical trials (Appendix Table S1b).

Table 10 presents an overview of RCT characteristics by clinical area.

Figure 2: Flow chart of included studies: Studies eligible for assessment of discontinuation and of non-publication, respectively



Discontinuation



Non-publication

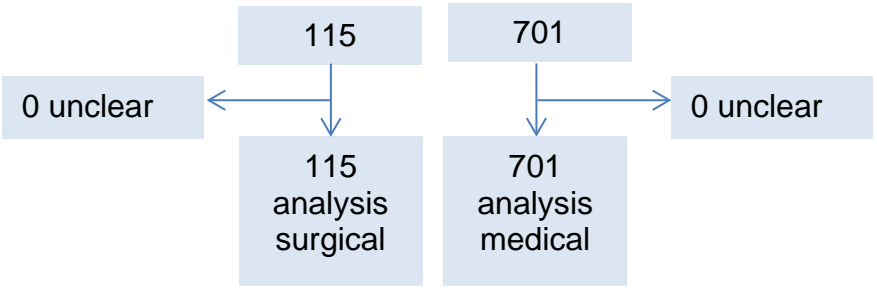


Table 10: Baseline characteristics by clinical area

In categorical variables, numbers (column percentages) are displayed

Characteristic		Clinical area	
		Surgery 127 (100%)	Medicine 736 (100%)
Intervention, n (%)	Medication	85 (67)	618 (84)
	Invasive procedure	31 (24)	25 (3)
	Rehabilitation	2 (2)	14 (2)
	Behavioral	0 (0)	18 (3)
	Diagnostic test	1 (1)	14 (2)
	Other	8 (6)	47 (6)
Sample size ¹ , median (IQR ²)		150 (60, 450)	288 (105, 628)
Industry funding, n (%)	Yes	76 (60)	552 (75)
	No	51 (40)	184 (25)
Trial sponsor, n (%)	Industry	59 (46)	470 (64)
	Academic investigator	71 (56)	276 (38)
Center status ³ , n (%)	Single-center	50 (40)	99 (14)
	Multi-center	75 (60)	633 (86)
Planned interim analysis, n (%)	Yes	30 (24)	249 (34)
	No	97 (76)	487 (66)
Presence of DSMB ⁴ , n (%)	Yes	24 (19)	223 (30)
	No	103 (81)	513 (70)
Planned stopping rule, n (%)	Yes	10 (8)	142 (19)
	No	117 (92)	594 (81)

¹ Surgery: 4 missings, Medicine: 5 missings² IQR=Interquartile range³ Surgery: 2 missings, Medicine: 4 missings⁴ DSMB=Data and Safety Monitoring Board

Discontinuation of trials

We excluded 66 trials (12 surgical, 54 medical) with unclear completion status (Figure 2). In the remaining 750 trials (103 surgical, 647 medical), the overall proportion of discontinued trials was 43% (44/103) in surgical trials and 27% (176/647) in medical trials with a risk difference of 16% (95% CI 5%, 26%; $p=0.001$; Table 11).

Table 11 Trial completion and publication status by clinical area

Numbers (column percentages) are displayed.¹

		Clinical area	
		Surgery	Medicine
		115 (100)	701 (100)
Completion status	Completed	59 (51)	471 (67)
	Discontinued	44 (38)	176 (25)
	Unclear	12 (11)	54 (8)
Publication status	Published	69 (60)	466 (66)
	- Abstract	- 5 (4)	- 31 (4)
	- Journal	- 64 (56)	- 421 (60)
	- Letter	- 0 (0)	- 7 (1)
	- Other	- 0 (0)	- 7 (1)
	Not published	46 (40)	235 (34)

Reasons for RCT discontinuation by clinical area are presented in Table 12. Slow recruitment was the most frequent reason for discontinuation in both surgical (19/44, 43%) and medical (68/176, 39%) trials. Stopping for slow recruitment was more frequent in surgical than in medical trials (19/103 [18%] *versus* 68/647 [11%], risk difference 8%, 95% CI 0.1%, 16%; $p=0.020$).

¹ Exclusion of never started trials (surgery 10/127, 8%; medicine 29/736, 4%) and Exclusion of still recruiting trials (surgery 2/127, 2%; medicine 6/736, 1%)

Table 12 Reasons for discontinuation by clinical area.
Numbers of studies (column percentages) are displayed.

Reason	Clinical area	
	Surgery	Medicine
	44 (100)	176 (100)
Benefit	1 (2)	6 (3)
Futility	5 (12)	28 (16)
Harm	4 (9)	17 (10)
Slow recruitment	19 (43)	68 (39)
External evidence	1 (2)	7 (4)
Lack of funding	0 (0)	4 (2)
Administrative	7 (16)	27 (15)
Other	0 (0)	7 (4)
Unclear	7 (16)	12 (7)

Results of our exploratory analysis comparing invasive to non-invasive trials suggested that RCTs with invasive interventions were more often discontinued for slow recruitment than RCTs with non-invasive interventions (Appendix Figure S1).

Publication of trials

No trials had to be excluded for unclear publication status (Figure 2). The overall percentage of trials not published in any format was 40% (46/115) among surgical trials and 34% (235/701) among medical trials with a risk difference of 6% (95% CI -3%, 16%; $p=0.176$; Table 11). Most trials were published as full journal articles (surgical 64/115 [56%], medical 421/701 [60%]) as opposed to abstracts, letters or other short-forms of publication (surgical 5/115 [4%], medical 45/701 [6%]). The overall percentage of trials not published as full journal article was 44% (51/115) among surgical trials and 40% (280/701) among medical trials with a risk difference of 4% (95% CI -5%, 14%; $p=0.373$).

Of completed surgical trials, 76% (45/59) were published in any format and of completed medical trials 79% (371/471). Of discontinued surgical trials, 55% (24/44) were published and of discontinued medical trials 52% (92/176).

Results of the supplementary exploratory analysis comparing invasive to non-invasive trials suggested no relevant differences in publication rates (Appendix Figure S1).

Risk factors for surgical trial discontinuation due to slow recruitment

Study discontinuation due to slow recruitment was neither influenced by the type of sponsor (industry versus academic investigator, adjusted odds ratio [OR] 0.60, 95% CI 0.20, 1.85, $p=0.377$ (Table 13) nor by the type of primary outcome (i.e. efficacy or harm) (adjusted OR 0.51, 95% CI 0.14, 1.85, $p=0.302$). Sensitivity analyses did not affect these findings (Appendix Table S2).

Risk factors for non-publication of surgical trials as full journal article

In multivariable analysis, trials discontinued for any reason were significantly more likely to remain unpublished than completed trials (adjusted OR 4.18, 95% CI 1.45, 12.06, $p=0.008$) (Table 14). Industry-sponsored trials were significantly more likely to remain unpublished than trials sponsored by an academic investigator (adjusted OR 2.99, 95% CI 1.05, 8.58, $p=0.041$). Trials with a primary harm outcome were less likely to remain unpublished than trials with primary efficacy outcomes only (adjusted OR 0.28, 95% CI 0.09, 0.92, $p=0.035$).

Sensitivity analyses did not affect these findings with the exception of the effect of the sponsor which did not remain statistically significant in the model with multiple imputation (Appendix Table S3).

Table 13 Univariable and multivariable logistic regression for trial discontinuation of surgical trials for slow recruitment (using a random effect estimator to account for clustering by the approving research ethics committee)

n=89¹ trials, 19 were discontinued and 70 completed or discontinued for other reasons.

Potential predictor	Univariable			Multivariable		
	Odds Ratio	95% CI ²	P value	Odds Ratio	95% CI	P value
Industry sponsor, yes <i>versus</i> no	0.66	0.22, 1.97	0.457	0.60	0.20, 1.85	0.377
Primary harm outcome, yes <i>versus</i> no	0.55	0.15, 1.96	0.354	0.51	0.14, 1.85	0.302

51

¹ Out of 115 trials, 12 had an unclear completion status and 14 had at least one missing covariate.

² CI=Confidence interval

Table 14 Univariable and multivariable logistic regression for non-publication as full journal article of surgical trials (using a random effect estimator to account for clustering by the approving research ethics committee).

n=89¹, 56 trials published as full journal article and 33 not published as full journal article.

Potential predictor	Univariable			Multivariable		
	Odds Ratio	95% CI ²	P value	Odds Ratio	95% CI	P value
Industry sponsor, yes <i>versus</i> no	2.02	0.84, 4.88	0.119	2.99	1.05, 8.58	0.041
Primary harm outcome, yes <i>versus</i> no	0.32	0.11, 0.96	0.043	0.28	0.09, 0.92	0.035
Trial discontinuation for any reason, yes <i>versus</i> no	2.26	0.94, 5.44	0.068	4.18	1.45, 12.06	0.008

8

¹ Out of 115 trials, 26 had at least one missing covariate.

² CI=Confidence interval

4.4. Discussion

This study suggests that the discontinuation rate is substantial in surgical RCTs and that discontinuation is more frequent in surgical than in medical RCTs. This applies to discontinuation for any reason as well as to discontinuation for slow recruitment, which is the predominant reason for discontinuation. Trial discontinuation was found to be a strong predictor for non-publication.

Relation to other studies investigating trial discontinuation

Slow recruitment is an important problem in clinical trials. In a cohort of 114 publicly funded multi-center trials, only 31% reached their initial recruitment target, an additional 24% reached 80% of their recruitment target and 53% of the trials were extended (McDonald et al. 2006). The target recruitment size was revised in 34% of trials, of which 86% downward (McDonald, Knight, Campbell, Entwistle, Grant, Cook, Elbourne, Francis, Garcia, Roberts, & Snowdon 2006). Slow recruitment may occur due to a variety of reasons, such as organisational failure, inadequate funding, or unrealistic projections regarding the number of eligible participants (Lievre et al. 2001). It may lead to longer study duration and increased resource consumption and early trial discontinuation with less precise study findings (Trewick et al. 2013). In a Cochrane review investigating methods to improve recruitment, a number of promising strategies have been described, among them telephone reminders and opt-out policies (Trewick, Lockhart, Pitkethly, Cook, Kjeldstrom, Johansen, Taskila, Sullivan, Wilson, Jackson, Jones, & Mitchell 2013). Most importantly, pilot studies may allow to estimate eligibility and consent rates more precisely and thus should be considered when planning a RCT (Lancaster, Dodd, & Williamson 2004).

Our exploratory analysis showed that trials with invasive interventions tended to be more frequently discontinued for any reason and were significantly more often discontinued for slow recruitment. This result suggests that the type of intervention (invasive vs. non-invasive) rather than the setting (surgery vs. medicine) represents a barrier to successful recruitment (Boutron, Moher, Altman, Schulz, & Ravaud 2008).

We found trial discontinuation to be more frequent in surgical than in medical trials, both overall and for slow recruitment. Potential explanations may be surgery-specific aspects and challenges with clinical trials such as requirements concerning surgeon experience, standardization of the intervention, and recruitment (e.g. due to patient or caregiver preference) (Boutron, Moher, Altman, Schulz, & Ravaud 2008; Devereaux, Bhandari, Clarke, Montori, Cook, Yusuf, Sackett, Cina, Walter, Haynes, Schunemann, Norman, & Guyatt 2005; Solomon & McLeod 1995). Additionally, there are some fundamental differences concerning the premarket review and approval process of devices versus drugs. Whereas for U.S. Food and Drug Administration (FDA) approval, safety and efficacy have to be demonstrated in humans for all drugs, this is only required for high risk devices

(Sweet et al. 2011). Similarly in Europe, clinical data are not required for lower risk devices (Sorenson and Drummond 2014). Thus, depending on the risk class of surgical devices, regulatory approval may not be based on clinical evidence, which potentially impacts resource investment and other efforts to achieve trial completion.

Relation to other studies investigating trial non-publication

In the present study, 60% of surgical and 66% of medical trials were published in any format, 56% and 60%, respectively as full journal articles. These proportions are slightly higher compared to a previous investigation of study protocols submitted to a Swiss REC with 52% of trials getting published (233/451) (Von Elm, Rollin, Blumle, Huwiler, Witschi, & Egger 2008). In this previous Swiss study, the odds for publication was higher with non-commercial funding, which is in line with our results, higher with multicentre trials, international collaboration and a high sample size as assessed by median split (Von Elm, Rollin, Blumle, Huwiler, Witschi, & Egger 2008). This investigation comprised, however, mostly medical specialties and thus is not strictly comparable to the surgical trials evaluated in our sample. In a Cochrane review of 79 studies investigating the subsequent publication of abstracts presented at meetings, the weighted full publication rate was 44.5% (95% CI 43.9-45.1) and an estimated cumulative publication rate after 9 years was 52.6% overall and 63.1% for RCTs or controlled clinical trials (Scherer et al. 2007). Abstract acceptance for presentation at a meeting was found to be associated with subsequent publication with an odds ratio of 1.78 (95% CI 1.50-2.12) (Scherer, Langenberg, & Von 2007).

We found trial discontinuation to be an independent risk factor for non-publication. This may be explained by the fact that the most frequent reasons of discontinuation of surgical RCTs were slow recruitment, futility or administrative reasons. These trials were thus potentially more difficult to publish than the few trials stopped early for benefit. Generally, in trials stopped early for benefit, treatment effects tend to be large (Bassler et al. 2008; Montori et al. 2005a), increasing the chance of a trial to be published. This is supported by a review of such trials in which the majority of RCTs (92/143) were published in 5 high-impact medical journals. (Montori, Devereaux, Adhikari, Burns, Eggert, Briel, Lacchetti, Leung, Darling, Bryant, Bucher, Schunemann, Meade, Cook, Erwin, Sood, Sood, Lo, Thompson, Zhou, Mills, & Guyatt 2005a).

Strengths and limitations

The collaborating RECs granted us unrestricted access to trial protocols. As outlined previously (Chan et al. 2006), this reduced the risk of selection bias; asking trialists or sponsors for permission would almost certainly have led to oversampling of successfully completed and published trials. However, some investigators may not have provided useful information about reasons for trial discontinuation. Additionally, our approach allowed us to collect more detailed

information about trial characteristics than investigations relying on trial registry entries only (World Health Organization 2013). Moreover, we obtained study protocols from several RECs in three countries, increasing the generalizability of our findings (Kasenda, von Elm, You, Blumle, Tomonaga, Saccilotto, Amstutz, Bengough, Meerpohl, Stegert, Tikkinen, Neumann, Carrasco-Labra, Faulhaber, Mulla, Mertz, Akl, Bassler, Busse, Ferreira-Gonzalez, Lamontagne, Nordmann, Rosenthal, Schandelmaier, Sun, Vandvik, Johnston, Walter, Burnand, Schwenkglenks, Bucher, Guyatt, & Briel 2012). Additionally, our data extractions were based on a priori definitions (Kasenda, von Elm, You, Blumle, Tomonaga, Saccilotto, Amstutz, Bengough, Meerpohl, Stegert, Tikkinen, Neumann, Carrasco-Labra, Faulhaber, Mulla, Mertz, Akl, Bassler, Busse, Ferreira-Gonzalez, Lamontagne, Nordmann, Rosenthal, Schandelmaier, Sun, Vandvik, Johnston, Walter, Burnand, Schwenkglenks, Bucher, Guyatt, & Briel 2012) and validated through regular quality assessments. Finally, our sensitivity analyses support the robustness of the conducted statistical analyses.

Our investigation has some limitations. Although we included a large number of approved RCTs, we were unable to adjust for additional confounders due to the low number of surgical trials and outcome events (Pezuzzi, Concato, Kemper, Holford, & Feinstein 1996). We thus refrained from adjusting for trial characteristics such as sample size and center status (single-center versus multi-center).

Implication for daily practice and further research

These findings of our study raise ethical concerns, as many patients who consent to enrol in surgical trials do not contribute to scientific knowledge, and suggest that considerable resources are being wasted. Thus, our findings may have the following implications during the different phases of a research pathway:

i) Trial conception and design: Strategies to prevent discontinuation should be adopted already during the trial design phase, including training in research methodology, realistic resource estimation (e.g. time of personnel), creating research infrastructures, and working in networks and interdisciplinary teams involving trial methodologists, statisticians, data managers and trial managers. Pilot studies either as part of the trial (internal pilot) or stand-alone (external pilot) (Lancaster, Dodd, & Williamson 2004), i.e. a small version of the full-scale study or feasibility studies, are effective means to evaluate particular aspects such as recruitment, resource utilization and protocol feasibility (Dolgin 2013). This especially applies to trials involving surgical interventions, where recruitment may be more challenging due to patient preferences. Moreover, pilot studies may identify other challenges such as feasibility of standardization of the surgical interventions. Pilot studies should thus be considered when allocating sparse resources to surgical RCTs. Additionally, trials should be registered prior to enrolment of the first patient. This enhances transparency regarding the targeted

sample size and additionally allows identifying the study while on-going or after close-up before publication.

ii) Trial conduct: Recruitment should be closely monitored and early modification of trial methods foreseen to enhance recruitment if necessary. The decision to discontinue a trial should be made by an independent Data and Safety Monitoring Board based on pre-defined criteria.

iii) Trial reporting: Reasons for discontinuation and results of discontinued RCTs should be transparently communicated and thus made publicly available.

iv) Healthcare decision making: When interpreting surgical literature, be it primary studies or systematic reviews, potential biases introduced by discontinued RCTs or by non-publication of trials need to be considered.

Conclusions

More than one in three surgical RCTs is discontinued. Discontinuation is more frequent among surgical than medical RCTs. Slow recruitment is the predominant reason for early termination. Furthermore, results from discontinued trials are more likely to remain unpublished. Rigorous planning within a multidisciplinary research framework, supported by pilot and feasibility studies, pre-defined criteria for stopping a trial, applied by independent boards, and transparent communication with stakeholders of surgical research and the public could help achieve the ultimate goal of high-quality research that improves the evidence base of surgical interventions.

5. Reporting of adverse events in surgical trials: critical appraisal of current practice

Rachel Rosenthal¹, MD, MSc, Henry Hoffmann¹, MD, Kerry Dwan², PhD, Pierre-Alain Clavien³, MD, PhD, Heiner C. Bucher⁴, MD, MPH

¹Department of Surgery, University Hospital Basel, Spitalstrasse 26, CH-4031 Basel, Switzerland.

²Department of Biostatistics, The University of Liverpool, Liverpool, Liverpool L69 3GS, United Kingdom.

³Department of Surgery, University Hospital Zurich and University of Zurich, Raemistrasse 100, CH-8091 Zurich, Switzerland.

⁴Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, Hebelstrasse 10, CH-4031 Basel, Switzerland.

Disclosure information: There is no funding related to this publication. Heiner C. Bucher is supported by grants from Santésuisse and the Gottfried and Julia-Rhyner-Foundation. There are no conflicts of interest to declare.

Publication information: This is a non-final version of an article published in final form in the World Journal of Surgery

(<http://www.springer.com/medicine/surgery/journal/268>): Rosenthal R, Hoffmann H, Dwan K, Clavien PA, Bucher HC. Reporting of adverse events in surgical trials: critical appraisal of current practice. World J Surg 2015;39(1):80-87 (Rosenthal et al. 2015c)

Abstract

Background: Reporting of surgical outcomes is important for healthcare decision making and includes the reporting of complications. For postoperative complications, however not for intraoperative complications, several classifications have been proposed and validated. The aim of the present study is to assess the current practice of complication reporting in surgical trials.

Methods: We evaluated reporting of intra- and postoperative complications in all registered randomized-controlled trials that investigate surgery or invasive interventions in at least one study arm and were published in 2010 in the Annals of Surgery, JAMA Surgery and the British Journal of Surgery.

Results: Forty six trials were identified and intra- and postoperative complications were reported separately in 42% and pooled in 15%. In 37% intraoperative, in 2% postoperative and in 4% both intra- and postoperative complications were not reported at all. Exact definitions were provided in 13% for intraoperative and in 50% for postoperative complications. A classification was used in 9% for intra- and in 54% for postoperative complications, most frequently according to severity. The type of intervention (surgical versus other) or whether the primary outcome was the assessment of complications had no significant impact on reporting definitions of adverse events.

Conclusions: Intraoperative complications are frequently pooled with postoperative complications, ill-defined or not reported at all, hampering informed decision-making. As further research, we propose to develop and validate a classification of intraoperative complications. This will facilitate the evaluation of safety and the continuous quality control of surgical interventions with the ultimate goal to contribute to patient safety.

5.1. Introduction

High-level evidence is of utmost importance for decision making in healthcare. Therefore, reporting needs to be accurate and transparent. Reporting should follow standardized guidelines as proposed by CONSORT (Consolidated Standards of Reporting Trials Statement) (Schulz, Altman, & Moher 2010). Of specific interest to surgery are the CONSORT extensions for nonpharmacologic treatment (Boutron, Moher, Altman, Schulz, & Ravaud 2008). There, challenges to surgical trials such as blinding, experience of and clustering by care providers and centers and standardization of interventions are accounted for. The CONSORT statement was initially developed to improve reporting of efficacy randomized-controlled trials (RCTs). Since then, the group has published an extension for better reporting of harms in relation to interventions. There, it is specified that the methods section should include a list of addressed adverse events with definitions for each and information on how harms-related data were collected (Ioannidis et al. 2004). If relevant, information on grading and expectancy of adverse events and validation of definitions should be provided (Ioannidis, Evans, Gotzsche, O'Neill, Altman, Schulz, & Moher 2004). In surgery, reporting of adverse events is specifically challenging due to the potential complexity of these events (Boutron, Moher, Altman, Schulz, & Ravaud 2008; Martin, Brennan, & Jaques 2002a). An analysis of the quality of reporting of short-term postoperative surgical complications after pancreatectomy, hepatectomy and esophagectomy involving 119 articles, revealed that definitions of complications were provided in 34% of the articles and the severity was graded in 20% only (Martin, Brennan, & Jaques 2002a).

Whereas several simple and reproducible classification systems for *postoperative* complications have been developed (Clavien, Sanabria, & Strasberg 1992; Clavien, Barkun, de Oliveira, Vauthey, Dindo, Schulick, de, Pekolj, Slankamenac, Bassi, Graf, Vonlanthen, Padbury, Cameron, & Makuuchi 2009; Dindo, Demartines, & Clavien 2004; Pillai, van Rij, Williams, Thomson, Putterill, & Greig 1999; Pomposelli, Gupta, Zacharoulis, Landa, Miller, & Nanda 1997; Slankamenac et al. 2013; Strasberg, Linehan, & Hawkins 2009), to our knowledge there is still no such classification specifically addressing *intraoperative* complications. To date, reporting of intraoperative complications in the literature is frequently either not present at all or intraoperative complications are pooled with postoperative complications (Tou et al. 2011). If intraoperative complications are reported separately, they tend to be ill-defined or only roughly classified according to severity (Morino et al. 2005), for example as „minor“ versus „major“, lacking exact classification criteria and/or classified according to the injured organ (e.g. vascular, visceral, solid organ injury) without accounting for different degrees of severity (Ahmad et al. 2012). A classification system of intraoperative complications is of particular interest and increasing importance in studies evaluating new surgical techniques and devices. It would allow for better transparency and more objective comparison between different surgical

techniques and their outcomes. Such a system would inform a more differentiated judgment of the evidence and provide a better rationale for more individualized decision-making in surgery. Additionally, standardization allows more adequate outcome reporting in educational and training settings (individual surgeon's outcome) and in institution benchmarking (institution outcome).

The aim of the proposed research is to systematically assess the current practice of reporting intra- and postoperative adverse events in randomized controlled trials published during one year in three major general surgery journals. This is of relevance to the subsequent development of a classification system for intraoperative complications.

5.2. Methods

Included studies

This is a secondary analysis of a review investigating discrepancies between registry entries and final reports in three general (non-speciality) surgical journals (Rosenthal & Dwan 2013). The journals with the highest 2010 impact factor under the category "surgery", the *Annals of Surgery*, *Archives of Surgery*, today *JAMA Surgery* and *British Journal of Surgery* were screened for all RCTs published between 01.01.2010 and 31.12.2010. Preclinical studies, duplicate publications, secondary analyses, interim reports and trials with lacking trial registry information (since in the primary review registry entries were compared to final reports) were excluded.

For the current investigation, trials not involving surgery or invasive treatments (be it as study intervention or not), such as trials investigating drug treatment unrelated to surgical or invasive interventions or trials evaluating the informed consent process, were excluded. Invasive interventions were defined as interventions requiring either percutaneous access (such as radiofrequency ablation of liver metastases) or access through natural orifices (such as endoscopic retrograde cholangio-pancreatography, but not natural orifice transluminal endoscopic surgery). The study intervention was defined as perioperative, if surgery was part of the study, but the intervention in itself was not, such as peri- or intraoperative administration of a drug, nutrition, bowel preparation, or transplant organ preparation.

Data extraction

The three Journals were screened for RCTs and data concerning the study population, interventions and outcomes were extracted as previously described (Rosenthal & Dwan 2013). The reviewers were not blinded for the names of the authors and journals of the evaluated RCTs. In the current investigation, two reviewers (RR and HH) independently additionally retrieved the following information concerning adverse events: whether the occurrence of adverse events was reported (or whether there was a statement that none occurred), whether adverse events were the primary outcome, whether both

intraoperative/intra-interventional and postoperative/post-interventional adverse events were reported and if yes, separately or not. Moreover, the reviewers assessed, whether a definition of adverse events was given in the methods section, whether a classification system was used and if yes, according to which specifier the events were classified. Adverse events were considered “reported” if any deviation from the normal intra- or postoperative/-interventional course (excluding sequelae and failures of cure) was reported or if it was reported that no such events occurred. Conversion, such as from laparoscopy to open surgery in itself was not considered as complication (i.e. it could be due to extensive tumour burden or adhesions), but in case of conversion due to complications (e.g. bleeding), the event was considered as complication. Adverse events were classified as “intraoperative”, if they were reported to occur during the surgical or invasive intervention or “postoperative”, if they were reported to occur after the intervention.

Complications were considered as “completely defined” if either an exact definition was provided such as for postoperative complications “any deviation from the normal postoperative course excluding sequelae and failures of cure” (Dindo, Demartines, & Clavien 2004) or if a complete list of events considered as complications was provided. Complications were considered as “not completely defined” if examples, but not a complete list, of events considered as complications were provided. If in the methods section it was mentioned that complications or the postoperative course were recorded without any further elaboration or if complications were listed in the results section only and not mentioned in the methods, complications were considered as “not defined”.

Complications were considered as “classified”, if some type of classification was used such as according to severity or according to affected organ, regardless as to whether the classification had previously been validated and as “not classified” if they were individually listed. Discrepancies between the two reviewers were resolved by discussion and agreement. For data extraction, an Excel spreadsheet (Microsoft Office XP, Microsoft Corporation, Redmond, WA, USA) was used.

Statistical Analysis

Descriptive statistics were used to present the proportions of trials with definition of intraoperative and postoperative complications and with information concerning grading systems, if applicable. In an explorative analysis, i) trials with surgical procedures in at least one study arm as the primary study intervention versus those with invasive interventions or perioperative interventions as study intervention and ii) trials with adverse events being the primary outcome and those with other types of primary outcomes were compared concerning definition of intra- and postoperative adverse events. For the comparisons, Fisher’s exact test was used. The comparisons were posthoc analyses explorative in nature without any correction for multiple testing. Analyses were conducted using Intercooled Stata Version 12.1; StataCorp LP, College Station, Texas, USA.

5.3. Results

Characteristics of included studies

The flow of screened and included studies has been previously described (Rosenthal & Dwan 2013). In brief, out of 596 studies identified through the search of the three journals, 65 were retrieved in full text, while the remaining studies were not RCTs and excluded. Out of these, a further 14 trials were excluded as they were interim reports (n=1), secondary analyses (n=5), not an RCT in contrast to the information provided in the heading or abstract (n=4), not registered RCTs (author information) (n=2) or with no registration information (n=2). Therefore, 51 studies were included (n=21 Annals of Surgery, n= 4 Archives of Surgery and n=26 British Journal of Surgery) (Rosenthal & Dwan 2013). The references of these 51 studies are given in the **Appendix References R1**. For the current investigation, two studies investigating the informed consent process (Online only reference 10*, 24*) and three studies evaluating medical treatment options without any related surgical or invasive intervention, (Online only reference 43*, 45*, 48*) were additionally excluded, leaving a total of 46 studies for this review. Out of these, 20 studies (43%) included surgery, five studies (11%) an invasive intervention and four studies (9%) both surgery and an invasive intervention. In 17 studies (37%), the study intervention consisted of perioperative interventions, but surgery was part of the study.

The baseline characteristics of all 46 studies with details on the population, the intervention, control, primary outcome and result are presented in the **Appendix Tables, Table S4**.

Intraoperative and postoperative adverse event reporting

All 46 studies included some kind of adverse event reporting. The primary outcome was unrelated to adverse events in 67% (n=31), whereas the primary outcome was defined as the occurrence of any kind of adverse event in 33% (n=15). Out of the latter, the adverse event related to postoperative complications in 14 studies, to intraoperative complications in none of the studies and to adverse events unrelated to surgery in one study.

Intraoperative complications were reported separately from postoperative complications in 42% (n=19) and pooled in 15% (n=7), whereas in 37% (n=17) intraoperative, in 2% (n=1) postoperative and in 4% (n=2) both intra- and postoperative complications were not reported at all.

Intraoperative and postoperative complications were completely defined in 13% (n=6) and 50% (n=23), respectively; they were not completely defined in 4% (n=2) and 2% (n=1), respectively; and in 44% (n=20) and 41% (n=19), respectively, they were not defined at all. A classification was used in 9% (n=4) for intraoperative and in 54% (n=25) for postoperative complications, most frequently according to severity (intraoperative complications: 3 out of 4, postoperative complications: 21

out of 25, table 15). Details on definition, classification and occurrence of intraoperative and postoperative adverse events of the individual studies are presented in the Appendix Table S5. Out of the 21 studies grading postoperative complications according to severity, 7 used the generic classification of postoperative complications according to severity by Clavien et al., taking into account the resulting type of complication management, e.g. the possibility of spontaneous resolution versus bedside procedure versus invasive procedure (Appendix Table S6) (Dindo, Demartines, & Clavien 2004). One study referenced the generic classification used at the Memorial Sloan-Kettering Cancer Center (Martin et al. 2002b). Six studies used complication-specific classifications: Three studies graded surgical site infections according to the definitions of the Centers for Disease Control and Prevention (Appendix Table S7) (Horan et al. 1992; Mangram et al. 1999b), one according to the ASEPSIS score (Wilson et al. 1986), one study graded complications after pancreatic surgery according to study group definitions (Bassi et al. 2005; Wente et al. 2007a; Wente et al. 2007b), and one graded bleeding according to criteria defined in a randomized controlled trial (Fox et al. 2004). The remaining 7 studies used own definitions or did not provide any reference, respectively. As for the 3 studies grading intraoperative complications according to severity, one used the classification according to Clavien et al. (Dindo, Demartines, & Clavien 2004), one derived the classification from a previous trial and one used an own definition.

Table 15: Definitions and classification of intra- and postoperative adverse events (n=46)

		Period	
		Intraoperative	Postoperative
		n (%)	n (%)
Definition	Complete definition	6 (13%)	23 (50%)
	No complete definition	2 (4%)	1 (2%)
	No definition	20 (44%)	19 (41%)
	No reporting of intra- respectively postoperative adverse events	18 (39%)	3 (7%)
Classification	Yes	4 (9%)	25 (54%)
	No	24 (52%)	18 (39%)
	No reporting of intra- respectively postoperative adverse events	18 (39%)	3 (7%)
Type of classification	Severity	3 (7%)	21 (45%)
	Organ	1 (2%)	4 (9%)
	No classification	24 (52%)	18 (39%)
	No reporting of intra- respectively postoperative adverse events	18 (39%)	3 (7%)

Comparison of trials with surgical interventions versus those with invasive or perioperative interventions

Three trials (all without surgery as primary study intervention) did not report postoperative adverse events and were therefore excluded. In the remaining trials, 46% (11/24) provided any type of definition of postoperative adverse events in trials with surgery as study intervention versus 68% (n=13/19) in trials without surgery as study intervention (p=0.217).

Eighteen trials (3 with and 15 without surgery as primary study intervention) did not report intraoperative adverse events and were therefore excluded. In the remaining trials, 18% (2/11) provided any type of definition of intraoperative complications in trials with surgical interventions versus 35% (6/17) in trials without surgical interventions (p=0.419).

Comparison of trials with the primary outcome being adverse events versus not

Three trials did not report postoperative adverse events (1 with and 2 without a primary outcome for adverse events) and were therefore excluded. In the remaining trials, 71% (n=10/14) provided any type of definition of postoperative adverse events in trials where an adverse event was the primary outcome versus 48% (n=14/29) in trials with other types of primary outcomes (p=0.199).

Eighteen trials did not report intraoperative adverse events (7 with and 11 without a primary outcome for adverse events) and were therefore excluded. In the remaining trials, 25% (n=2/8) provided any type of definition of intraoperative adverse events in trials where an adverse event was the primary outcome versus 30% (6/20) in trials with other types of primary outcomes (p=1.000).

5.4. Discussion

This review of publications in three major surgical journals shows that intraoperative complications are rarely reported or they are subsumed with postoperative complications. If intraoperative complications are reported, they mostly lack a definition. This is in contrast to postoperative complications, for which about half of the studies provide these details. The type of study intervention or whether the primary outcome was an adverse event did not significantly impact on these results.

Relation to other studies

To our knowledge, this is the first review assessing the quality of reporting adverse events in surgery with a specific focus on intraoperative complications. The assessment of technical errors, mostly in the context of surgeon training and safety evaluation, has however been previously investigated. As pointed out in the presidential address of the European Surgical Association by Clavien, targeting safety and quality in surgery including standardization of outcome measures is of utmost importance (Clavien 2013). In a review of studies describing technical errors during laparoscopy, Bonrath et al. found 8 out of 21 studies investigating surgical performance in terms of errors during routine laparoscopic cholecystectomy, whereas in the remaining 13 studies, errors were evaluated as a surrogate for surgical skills in an educational context (Bonrath et al. 2013a). Error definitions were found to vary considerably and hamper a direct comparison (Bonrath, Dedy, Zevin, & Grantcharov 2013a).

Several studies have investigated the reporting of postoperative surgical complications. A systematic review of the definition of anastomotic leak in gastrointestinal surgery based on 97 studies revealed 56 different definitions of anastomotic leak (Bruce et al. 2001a). None of the studies used the standard definition for anastomotic leak that had previously been proposed at a consensus workshop (Peel and Taylor 1991). Similarly, a systematic review of 90 studies on definitions of surgical wound infection revealed 41 different definitions. Of these, 5

may be considered as “standard” definitions based on work by multidisciplinary groups (Bruce et al. 2001c). Another review on the quality of surgery-related mortality reporting revealed considerable differences in follow-up time and post-discharge mortality was rarely taken into account (Russell et al. 2003).

Thus, surgery-related complications and relevant outcomes in surgery lack standard definitions and therefore do not allow for comparison across studies and between surgical disciplines. However, this problem is not only related to surgery. Adverse event reporting in drug trials is also often lacking appropriate information about drug discontinuation, and adequate reporting on toxicity and severity of adverse events (Ioannidis and Contopoulos-Ioannidis 1998; Ioannidis and Lau 2001).

Definition and classification of surgical complications

Surgical complications are not uniformly defined. One option is to define surgical complications as “any undesirable, unintended, and direct result of an operation affecting the patient, which would not have occurred had the operation gone as well as could reasonably be hoped” (Sokol and Wilson 2008). This definition thus may apply to intra- and postoperative complications, but has been controversially discussed, because it may not always be possible to establish or rule out whether there is a causal relationship between surgery and the undesirable result. Moreover, not all undesirable results are complications, but they may as well be inherent to the intervention or related to the severity of the underlying disease (Dindo and Clavien 2008). Thus, complications may be defined as “any deviation from the normal postoperative course” (Dindo, Demartines, & Clavien 2004) and are distinguished from inherent effects of surgery (sequelae, such as the inability to walk after an amputation of a leg) or failures to cure (such as residual tumor after surgery) (Clavien, Sanabria, & Strasberg 1992). This definition, however, is limited to postoperative complications. In a five-year follow-up of the classification by Clavien et al., it was discussed to capture death during surgery or to “record as complication all events occurring in the operating room from the time of preparation for anesthesia” (Clavien, Barkun, de Oliveira, Vauthey, Dindo, Schulick, de, Pekolj, Slankamenac, Bassi, Graf, Vonlanthen, Padbury, Cameron, & Makuuchi 2009).

The need for a classification system of intraoperative complications

The need for an intraoperative classification of complication is subject to debate. Cunningham and Kavic (Cunningham and Kavic 2009) correctly point out that the definition of complications as deviation from the normal *postoperative* course (Clavien, Sanabria, & Strasberg 1992; Dindo, Demartines, & Clavien 2004) does not capture intraoperative events that do not directly result in clinically relevant postoperative complications. They proposed to monitor any “deviation from the ideal operative course” and to distinguish between “simple errors” (adverse intraoperative events not leading to postoperative manifestations or additional risk

for postoperative complications, e.g. burn to the lateral parietal peritoneum), and “complications” with potential risk for postoperative complications (e.g. inadvertent, but treated enterotomy or splenic injury) (Cunningham & Kavic 2009). The concept was questioned by Wilson and Sokol (Wilson and Sokol 2009). They argue that intraoperative complications which do result in a postoperative complication are anyhow captured as surgical complications, whereas the others theoretically add information, but may complicate monitoring concepts. The authors propose to refer in such cases to “intraoperative errors”, which of course need to be avoided.

Conversely, the documentation of intraoperative events may be especially important when new surgical techniques are introduced. In studies investigating new surgical procedures, the event rate of postoperative complications resulting from intraoperative complications may not be high enough to be captured with sample sizes resulting from calculations for efficacy outcomes. In contrast, documentation of intraoperative adverse events should prompt more extensive safety evaluation before wide application. The Global Harmonisation Task Force on Medical Devices (GHTF, see International Medical Device Regulators Forum IMDRF, www.imdrf.org) has provided guidance to standardize the reporting of adverse events and device failures (Mehran et al. 2004). In their guidelines, the GHTF study group distinguishes death, serious injury and events *potentially* leading to death or serious injuries, for instance if they occurred again (Lalis G for the Global Harmonization Task Force Study Group 2 2006). This concept is in line with a safety culture not limited to investigating actual patient injury in order to prevent reoccurrence of such an injury. Instead, such a concept also addresses critical incidents without actual patient injury and evaluates why such critical incidents and errors occur (Kram 2008;Leape 1994;Reason 2000). This is in accordance with the assumption that adverse events may arise from a build-up of several minor errors (Catchpole 2010). As a matter of fact, an error-outcome hierarchy may be described with the option of being uneventful on every hierarchy level (Bonrath, Dedy, Zevin, & Grantcharov 2013a;Bonrath et al. 2013b). Starting with an error, such as inserting a trocar without visualization, a technical event may or may not occur, such as an enteric injury. This injury may or may not be detected and rectified leading or not to a postoperative complication such as an enteric leak. The latter in turn may or may not be detected and managed and finally may or may not end up in an adverse outcome such as sepsis or death (Bonrath, Zevin, Dedy, & Grantcharov 2013b).

Additionally, some complications are just not captured when focussing on the immediate postoperative phase only, because the consequences are only evident at long-term or immediately and definitively treated during surgery, for instance haemorrhage with intraoperative blood transfusion only.

Limitations

Our study presents some limitations. First, all RCTs published in three major surgical journals except those with strictly medical or informed consent process-interventions were considered for review. Therefore, some of the interventions were not involving surgery only, but as well invasive or perioperative interventions. We addressed this issue by evaluating in an exploratory analysis the impact of studies including surgery as primary study intervention (yes *versus* no) on complication reporting. Results from three journals may not be extrapolated to all surgical journals. We focused on three high-impact journals with endorsed CONSORT reporting guidelines policy and it is therefore unlikely that an investigation in other surgical journals would yield a lower frequencies of lacking definition and classification of intra- and postoperative adverse. Second, reviewers were not blinded regarding authors and journals of the published RCTs. However data abstraction was done by two independent reviewers who used a standardized objective extraction protocol. We consider the lack of blinded data abstraction as a minor risk of bias (Morissette et al. 2011).

Implications for daily practice and further research

Reports on adverse events from surgical interventions often do not use uniform definitions for adverse events or do not distinguish between intra- and postoperative complications and therefore have to be scrutinized. For these reasons, we plan to develop and validate a simple and reproducible classification for intraoperative complications, similar to the existing classifications of postoperative adverse events (Clavien, Sanabria, & Strasberg 1992;Clavien, Barkun, de Oliveira, Vauthey, Dindo, Schulick, de, Pekolj, Slankamenac, Bassi, Graf, Vonlanthen, Padbury, Cameron, & Makuuchi 2009;Dindo, Demartines, & Clavien 2004;Pillai, van Rij, Williams, Thomson, Putterill, & Greig 1999;Pomposelli, Gupta, Zacharoulis, Landa, Miller, & Nanda 1997;Strasberg, Linehan, & Hawkins 2009).

The ultimate goal will be to propose a definition and classification of intraoperative complications that can be used for standardized outcome data collection. Therewith, a core outcome set that may be used as reporting guideline for trials involving surgical procedures may be developed. This is in line with the COMET initiative (Core Outcome Measures in Effectiveness Trials, www.comet-initiative.org) with the goal to develop reporting standards that allow for comparisons across studies, health care providers and patients (Williamson et al. 2012).

Conclusions

There is a lack of universal reporting of adverse events in surgery. This compromises the comparison of reports on intra- and postoperative complications in surgical innovation and clinical research as well as continuous quality assessment and benchmarking. Several straightforward postoperative

complications classifications have been developed and validated, but intraoperative complications are frequently not reported, ill-defined or pooled with postoperative complications. A clear distinction and classification of intraoperative complications is therefore needed to capture intraoperative events that eventually result in postoperative complications or increase the risk for such complications. This would also allow standardizing critical incident reporting systems that may be needed for quality control, trainee assessments or evaluation of surgeons' performance.

6. Definition and *Classification of Intraoperative Complications (CLASSIC)*: Delphi study and pilot evaluation

Rachel Rosenthal¹, MD, MSc, Henry Hoffmann¹, MD, Pierre-Alain Clavien², MD, PhD, FACS, Heiner C. Bucher³, MD, MPH, Salome Dell-Kuster^{1,3,4}, MD, MSc

¹Department of Surgery, University Hospital Basel, Switzerland

²Department of Surgery, University Hospital Zurich and University of Zurich, Zurich, Switzerland

³Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, Basel, Switzerland

⁴Department of Anesthesiology, University Hospital Basel, Switzerland

Disclosure information: There is no funding related to this publication. Heiner C. Bucher is supported by grants from Santésuisse and the Gottfried and Julia-Rhyner-Foundation.

There are no conflicts of interest to declare.

Publication information: This is a non-final version of an article published in final form in the World Journal of Surgery

(<http://www.springer.com/medicine/surgery/journal/268>): Rosenthal R, Hoffmann H, Clavien PA, Bucher HC, Dell-Kuster S. Definition and *Classification of Intraoperative Complications (CLASSIC)*: Delphi study and pilot evaluation. World J Surg 2015;39(7):1663-1671 (Rosenthal et al. 2015d)

Abstract

Background: Standardized reporting of intraoperative adverse events is important to enhance transparency. To the best of our knowledge, there is no validated definition and classification of intraoperative complications.

Study Design: We conducted a two-round Delphi study to develop a definition and classification of intraoperative complications. Experts were contacted by email and sent a link to the online questionnaire. In a pilot study, 2 independent raters applied the definition and classification in a sample of 60 surgical interventions of low, intermediate and high complexity, and evaluated practicability. Interrater agreement of the classification was determined (raw categorical agreement, weighted kappa and intraclass correlation).

Results: In the Delphi study, 40 of 52 experts (77% return rate) from 14 countries took part in each round. The Delphi study resulted in a comprehensive definition of intraoperative complications as any deviation from the ideal intraoperative course occurring between skin incision and skin closure and a straightforward classification with four severity grades. A number of specifiers were elaborated for further distinction of the definition. The pilot study showed good practicability (6 on a 7-point scale) and a high raw agreement of 87%, a weighted kappa of 0.83 (95% confidence interval [CI] 0.73, 0.94) and an intraclass correlation coefficient of 0.83 (95% CI 0.73, 0.90).

Conclusions: While the Delphi process enabled to develop definitions and classification of intraoperative complications by severity, further research including a multicentre international full-scale validation needs to be conducted with the ultimate goal to contribute to standardized reporting in surgical practice and research.

6.1. Introduction

Reporting of surgical outcomes is important for informed decision making. This not only includes efficacy parameters, but also the reporting of adverse events, i.e. intra- and postoperative complications. For postoperative complications, several definitions and classification systems have been proposed (Clavien, Sanabria, & Strasberg 1992; Clavien, Barkun, de Oliveira, Vauthey, Dindo, Schulick, de, Pekolj, Slankamenac, Bassi, Graf, Vonlanthen, Padbury, Cameron, & Makuuchi 2009; Dindo, Demartines, & Clavien 2004; Pillai, van Rij, Williams, Thomson, Putterill, & Greig 1999; Pomposelli, Gupta, Zacharoulis, Landa, Miller, & Nanda 1997; Strasberg, Linehan, & Hawkins 2009), whereas we are not aware of any reporting system addressing intraoperative complications.

In our opinion, standardization of the definition and classification of intraoperative complications is important in a number of settings: first, in research and development, when assessing new surgical techniques and devices; second, in educational and training settings (individual surgeon's performance) and in institution benchmarking (institution performance); third, for risk management, capturing intraoperative events with an increased risk of postoperative complications within a critical incidents reporting system (Kram 2008; Leape 1994; Reason 2000) or in clinical research where postoperative event rates may be too low to capture potential safety concerns; and last for patient management, enhancing quality of patient handover after surgery with anticipated postoperative problems (Manser et al. 2013; Nagpal et al. 2011).

A preliminary investigation of the current practice of reporting intra- and postoperative complications in a cohort of 46 randomized-controlled trials published in 2010 in the *Annals of Surgery*, *JAMA Surgery* and the *BJS* revealed that 41% of the trials failed to report intraoperative complications at all (submitted for publication). Only 13% provided a definition of and 9% used a classification for intraoperative complications.

The aim of the present study is to develop a definition and classification for intraoperative complications within a Delphi study and to conduct a pilot study evaluating practicability and interrater agreement.

6.2. Methods

1. Delphi study

Participants

International experts with a surgical or methodological background were invited to participate in the Delphi study. Prerequisite was board-certification in surgery for at least 2 years or a higher academic degree (MSc/PhD) in trial methodology with experience in trial design for at least 2 years. Different subspecialty surgeons were invited to participate with the intention to develop a classification applicable to any

surgical subspecialty. Experts were recruited through personal contacts and surgical associations with the goal to get a representative sample of different expertise and location of practice across the world.

Questionnaire dissemination

This Delphi study consisted of two rounds (February 21 to March 09, 2014 and March 26 to April 03, 2014). The questionnaires were prepared using SurveyMonkey® (www.surveymonkey.com). The experts were sent a link to the online questionnaire by email. The anonymous responses were downloaded as Excel spreadsheet (Microsoft Office XP, Microsoft Corporation, Redmond, WA, USA). One email reminder per round was sent out before the end of the deadline to all experts regardless whether they had answered or not, the anonymous process not allowing for differentiation between responders and non-responders.

Content first round

Since there is abundant literature on the classification of postoperative complications, we refrained from a first round with open questions, as suggested in such instances (Hsu and Sandford 2013).

Participants were provided with some background information about surgical complications and were asked to answer baseline demographics and background questions. They were then invited to indicate their level of agreement on a 9-point rating scale with the inclusion of a number of items concerning the definition and classification of intraoperative complications. We chose a 9-point scale because of the known limit in accuracy of raters to give an absolute judgment. This span is about seven (MILLER 1956); we rounded up to nine in order to account for a potential end-aversion bias (i.e. have two extra categories to address the tendency to avoid extreme categories). Scales with too low numbers of items, such as two-, three- and four-point scales have been found to show poor reliability, validity and discriminating power (Preston and Colman 2000). Raters could not move on with the questionnaire without having rated every item. Additionally, experts were strongly encouraged to comment on any of the items.

Data analysis first round

Descriptive statistics including graphs were used to analyse quantitative data. All statistical analyses of the present research were conducted using Intercooled Stata Version 12.1; StataCorp LP, College Station, Texas, USA. Comments (qualitative data) were summarized within a narrative review. Since the questionnaire prompted extensive comments and a fruitful extensive discussion, the manuscript only contains the main findings. To enhance transparency and to allow for a detailed insight in the iterative discussion, the entire Delphi report is included in two online-only appendices.

Content second round

The quantitative and qualitative results of the first round were presented to the participants and, where applicable the questions concerning the proposed definition and classification were updated accordingly.

Data analysis second round

Data were analysed as outlined for the first round. Since sufficient consensus was obtained after two rounds, no third round was conducted.

2. Pilot study

Surgical interventions

The pilot study was conducted in a retrospective cohort of patients operated between 01.01.2013 and 31.12.2013 in a tertiary referral and teaching hospital. A random sample of 60 records of patients having undergone surgeries of different complexity (20 type A, 20 type B and 20 type C) was selected. Type A, B and C surgery was defined according to previous investigations (Dindo, Demartines, & Clavien 2004; Klotz et al. 1996). In brief, type A surgery includes surgical procedures without opening the abdominal cavity. Type B surgery includes abdominal procedures with the exception of liver surgery and major retroperitoneal surgery. Type C surgeries include highly specialized procedures such as surgery on the liver, oesophagus, pancreas and rectum. The scheme for the selection of a random sample from individual records was generated with the use of the web page Randomization.com (<http://www.randomization.com>).

Classification

Two investigators (RR and HH), both board-certified surgeons, independently applied the definition and classification resulting from the Delphi process based on the information given in the operation records. Additionally, postoperative complications were classified according to Clavien and Dindo (Dindo, Demartines, & Clavien 2004) based on the entire patient record including the discharge letter. Disagreements were resolved by consensus.

Practicability

We evaluated practicability on a seven-point scale.

Data analysis

Agreement between the judgment of the two raters (before resolving disagreements) was evaluated investigating raw categorical agreement (number of exact categorical matches between two raters divided by the total number of records) (Streiner 1995), kappa with Fleiss-Cohen (quadratic) weights (Fleiss and Cohen 1973; Warrens 2011) and intraclass correlation (Shrout and Fleiss 1979).

6.3. Results

1. Delphi study

Participants

The link to the questionnaire was sent out to 52 experts, of whom 40 answered within the given deadline in both rounds (77% return rate). In Delphi round 1, one participant submitted an incomplete questionnaire containing only baseline characteristics as well as a judgment concerning the overall usefulness of a classification. Two additional contacted experts returned the questionnaire after the given deadline and therefore their answers could not be included in the feedback to round 2 and the pilot evaluation, respectively (81% [42/52] return rate in total). Experts from 14 countries in the US, Europe, Asia and Australia from a large variety of surgical disciplines (n=37) as also experts in epidemiology and statistics (n=3) took part in the study. Their median work experience was 25 years with an interquartile range [IQR] from 12 to 29 years (Table 16).

Results first round

The full report with a workup of all quantitative and qualitative data is given in Appendix Text S1. The overall usefulness of a classification of intraoperative complications was rated as high (Figure 3).

Definition of intraoperative complications

Experts mostly agreed to distinguish intraoperative complications from sequelae, i.e. effects inherent to the surgery (such as the inability to walk after an amputation of a leg), from failures of cure (such as residual tumor after surgery) (Clavien, Sanabria, & Strasberg 1992) and from events related to the underlying disease (such as bleeding from a ruptured abdominal aortic aneurysm). As prerequisite, experts further suggested adding to the definition and classification that the indication for surgery and the interventions has to conform to current guidelines. Additionally, experts pointed out that complications should refer to the index surgery, not to potential follow-up surgeries to treat a complication. Experts discussed controversially whether errors in indication should be included in the definition or not. The discussion of the exact wording for the definition, for example 'any deviation from the ideal intraoperative course' remained inconclusive after the first round, it was however suggested to only consider events that are preventable. The time-period to be taken into consideration was also controversially discussed: while some felt that only surgery (i.e. time between skin incision and skin closure) should be included, others were more in favor of including induction and termination of anesthesia, thus focussing on the entire team and not the individual surgeon.

Table 16: Baseline characteristics¹ of participants in the Delphi study (n=40)

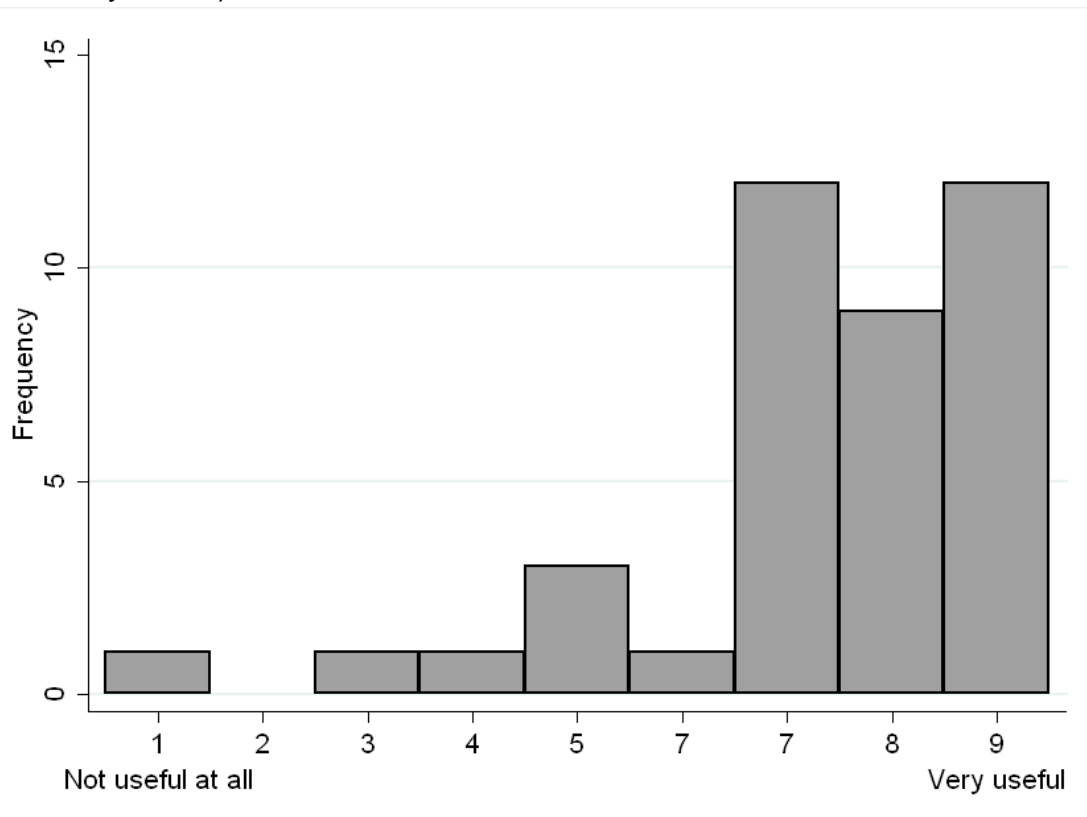
Characteristic	Category	Value
Primary discipline, n (%)	Surgery	37 (92.5%)
	Public health/epidemiology	2 (5.0%)
	Statistics	1 (2.5%)
Primary subspecialty, n (%)	Breast surgery	1 (2.5%)
	Cardiac surgery	1 (2.5%)
	Colorectal surgery	3 (7.5%)
	Neurosurgery	1 (2.5%)
	Orthopedic surgery	4 (10%)
	Pediatric surgery	2 (5.0%)
	Plastic and reconstructive surgery	1 (2.5%)
	Surgical oncology	1 (2.5%)
	Thoracic surgery	2 (5.0%)
	Traumatology	1 (2.5%)
	Urology	1 (2.5%)
	Vascular surgery	2 (5.0%)
	Visceral surgery	14 (35%)
	Other surgical specialty	3 (7.5%)
	Epidemiology	2 (5.0%)
Statistics	1 (2.5%)	
Country, n (%)	Switzerland	23 (57.5%)
	Germany	1 (2.5%)
	Austria	2 (5.0%)
	Netherlands	1 (2.5%)
	Ireland	1 (2.5%)
	United Kingdom	3 (7.5%)
	France	2 (5.0%)
	Italy	1 (2.5%)
	Spain	1 (2.5%)
	Turkey	1 (2.5%)
	USA	1 (2.5%)
	Singapore	1 (2.5%)
	Hongkong	1 (2.5%)
Australia	1 (2.5%)	
Involved in clinical research as (several options applicable), n (%)	Principal investigator	37 (92.5%)
	Co-investigator	35 (87.5%)
	Biostatistician	2 (5.0%)
Years of overall work experience, median (IQR) ²		25 (12, 29)
Years of surgical work experience, median (IQR) ³		24 (12, 29)
Years of clinical research experience, median (IQR)		20 (8, 24)
Gender, n (%)	Male	36 (90%)
	Female	4 (10%)

¹ Assessed in round 1

² IQR=Interquartile range

³ n=3 with primary non-surgical discipline excluded

Figure 3: Delphi study round 1: Answers to the question concerning overall usefulness of a classification of intraoperative complications (1=not useful at all, 9=very useful)



Classification of intraoperative complications

The experts were in favor of considering the severity of complications in the classification system. Experts extensively discussed which time-period should be considered when filling in the classification, i.e. whether it should be filled in directly after surgery or whether it should be updated during follow-up, such as after 30 days, since some complications do not become evident until after surgery. It was argued that in an update during follow-up it would be difficult to distinguish an intraoperative complication becoming evident after surgery from a postoperative complication, since it may be difficult to establish a definitive causal relationship between the intervention and the complication. Next, a draft of a classification with five grades, one of which with two subgrades, was discussed with several examples. The majority of the experts felt that the classification should be simplified further.

Results second round

The full report with a workup of all quantitative and qualitative data is given in Appendix Text S2.

Definition of intraoperative complications

Experts were in favor of the updated definition, which distinguishes intraoperative complications from sequelae, failure to cure, and disease-inherent events and is related to best praxis indication and surgery of the index procedure. The distinction from wrong-side or wrong-patient surgery and errors in indication was not uniformly accepted (Figure 4). For practicability reasons and in order not to underestimate errors in indication and wrong-side/wrong-patient surgery, we decided to exclude these items. Based on the feedback from round 1 we decided to change our wording to 'preventable deviation from the ideal course', but experts gave us mixed feedback and we decided to stay with the former wording and to drop the term 'preventable'. Experts seemed to be rather in favor of only including the time between skin incision and closure, and to record any event during that period, be it surgical or not (e.g. anesthesia-related). As for the number of raters, there was no uniform opinion. Although having two raters seemed favourable to most, provided they were really independent, it was judged not to be very practical having more than one rater, especially in clinical practice. In a research setting, it might however be important to have complications rated by more than one reviewer. We therefore suggest including in the study report a statement on how many raters were involved and whether or not they themselves were involved in the surgical procedure and in the case of multiple raters, how consensus was reached.

Classification of intraoperative complications

Experts mostly were in favor of grading complications according to severity and also of grading immediately after surgery. The reasons were practicability and the distinction from classifications of postoperative complications (Figure 5). Additionally, it was proposed to consider the option of a classification not based on the actual outcome, but on the potential outcome, thus avoiding the underestimation of severe complications. This could be handled by including the term 'potentially' (i.e. 'potentially life-threatening' or 'potentially leading to permanent disability), with the disadvantage however of leaving quite a range for interpretation. Within the pilot study, we decided to test the option of filling in the classification immediately after surgery (i.e. in a retrospective study considering only the operation report for classification). Some experts suggested dichotomizing the severity grading of complications, but we decided to keep the originally proposed classification with four levels. If it is still used for investigating minor versus major complications, both results from the original classification and the contracted classification should be reported providing a clear definition for the cut-off.

Figure 4: Delphi study round 2: Answers to the questions concerning the distinction of intraoperative complications from sequelae, failure to cure, events related to the underlying disease and error in indication:

Based on these answers of round 1, we kindly ask you to rate the following updated definition (1= not at all appropriate, 9 = completely appropriate):

Sequelae: A definition of intraoperative complications needs to distinguish complications from sequelae, i.e. effects inherent to current best practice surgery (such as the inability to walk after amputation of a leg). It should refer to the index surgery and not to potential follow-up surgeries to treat a complication.

Failure to cure: A definition of intraoperative complications needs to distinguish complications from failures of cure (such as residual tumor after surgery), under the condition the indication for surgery and the interventions conform to current guidelines.

Disease: A definition of intraoperative complications needs to distinguish complications from events related to the underlying disease (such as bleeding from a ruptured abdominal aortic aneurysm or intraoperative sepsis due to a purulent peritonitis associated to a perforated appendicitis), under the condition the indication for surgery and the interventions conform to current guidelines.

Indication error: A definition of intraoperative complications should NOT refer to wrong-site or wrong-patient surgery and NOT to errors in indication (such as inappropriate indication for surgery according to current guidelines)

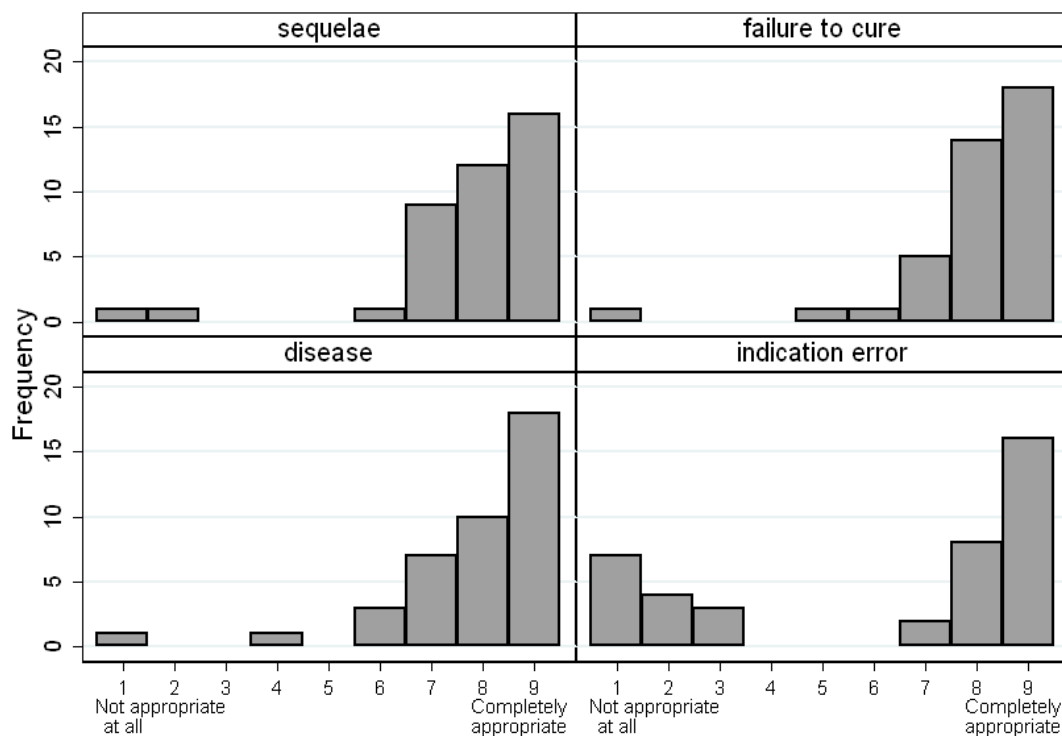


Figure 5: Delphi study round 2: Answers to the questions concerning type of grading, classification overall, timepoint of grading, dichotomisation:

We kindly ask you to rate the following (1= not at all appropriate, 9 = completely appropriate):

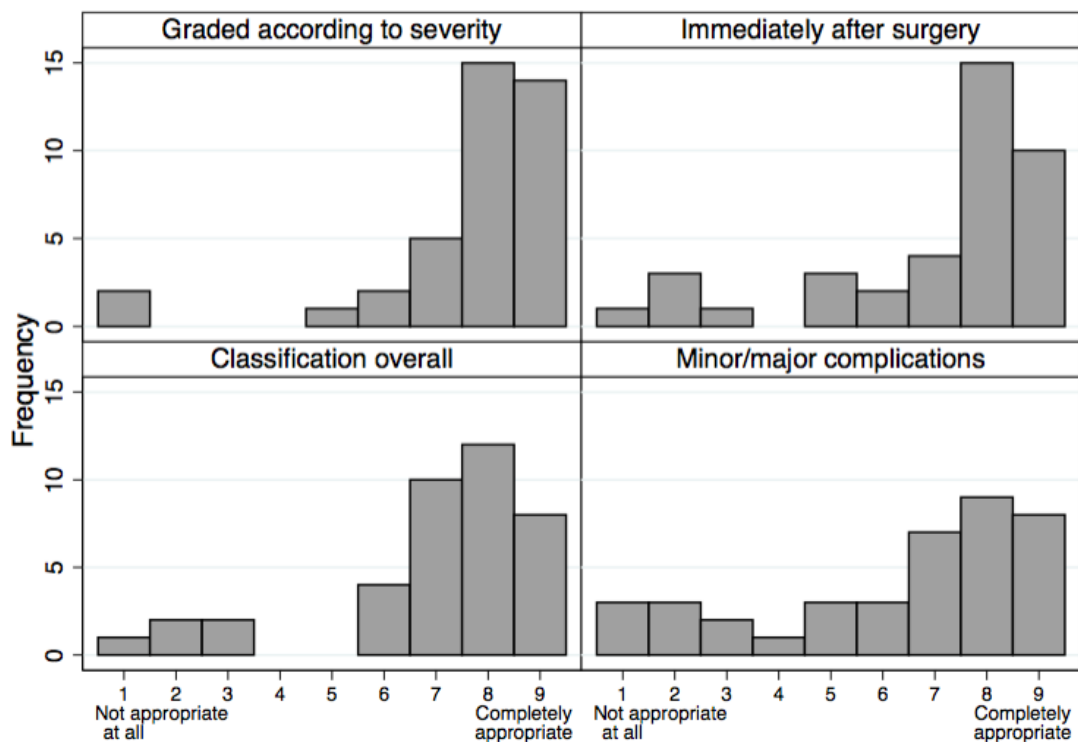
Graded according to severity: A classification of intraoperative complications should consider the severity of the complication and take into account both the necessary treatment and permanent sequelae resulting from the complication.

Immediately after surgery: Should it be possible to definitively fill in a classification of intraoperative complications directly after surgery, whereas complications apparent only after surgery are handled with a classification system for postoperative complications?

Classification overall: We kindly ask you to rate the updated classification:

Prerequisite for all grades: under the condition the indication for surgery and the interventions conform to current guidelines. Grade 0: no preventable deviation (an unpreventable deviation could be for instance a deviation due to an unexpected anatomical situation); grade 1 preventable deviation from the ideal intraoperative course without the need for any additional treatment or intervention; grade 2 and 3 with the need for any additional treatment or intervention, grade 3 if life-threatening and/or leading to permanent disability, else grade 2; grade 4 with patient death.

Minor/major complications: Should complications needed to be dichotomised for further analysis (which we not necessarily advocate), we propose to summarize grades I-II together as “minor complications” versus grades III-IV as “major complications”.



Resulting definition and classification

The Delphi study provided the following definition and classification of intraoperative complications (Table 17): Any deviation from the ideal intraoperative course, given the indication for surgery and the interventions conform to current guidelines. The classification exclusively relates to any event occurring between skin incision and skin closure, and should be rated directly after surgery. Any event during the index-surgery must be considered, regardless whether it is surgery or anesthesia-related. The following events are not defined as intraoperative complications: 1. Sequelae, i.e. effects inherent to current best practice surgery (such as the inability to walk after a leg amputation), 2. Failures of cure (such as residual tumor after surgery), 3. Events related to the underlying disease (such as bleeding from a ruptured abdominal aortic aneurysm or intraoperative sepsis due to a purulent peritonitis associated to a perforated appendicitis), 4. Wrong-side or wrong-patient surgery or errors in indication (such as inappropriate indication for surgery according to current guidelines). Complications evident only after skin closure are captured with the classification for postoperative complications.

The classification is based on severity, considering whether an additional treatment is necessary, the complication is life-threatening or leading to permanent disability or results in death of the patient (Table 17). The classification is not intended to get dichotomized. Any report on intraoperative complications should state the number of raters involved in the assessment and whether they were involved in the procedure. If they were involved in the procedure, it should be stated in which role and if not, whether the classification was based on records or on videotapes of the intervention. In case of multiple raters it should additionally be reported how a consensus was reached.

2. Pilot study

Patient and procedure characteristics

A total of 60 records of patients with a median age of 69 (IQR 54, 77) were reviewed. Out of these, 39 (65%) were male, 23 (38%) had an ASA (American Society of Anesthesiologists) – score of three or higher and 17 (28%) were emergency admissions (Appendix Table S8). Type A surgery was represented by inguinal/femoral hernia repair (n=10) and thyroid/parathyroid surgery (n=10), type B surgery by cholecystectomy (n=10) and colon surgery (n=10), and type C surgery by major liver (n=5), pancreatic (n=5) oesophageal (n=5) and rectal surgery (n=5) (Appendix Table S9).

Table 17: Proposed Classification of *intraoperative complications* (CLASSIC)

Grade	Definition
	The classification exclusively relates to any event occurring between skin incision and skin closure and should be rated directly after surgery. Any event during the index-surgery must be considered, regardless whether it is surgery or anesthesia-related. ¹ Prerequisite: the indication for surgery and the interventions conform to current guidelines
Grade 0	No deviation from the ideal intraoperative course
Grade I	Any deviation from the ideal intraoperative course <ul style="list-style-type: none">Without the need for any additional treatment or intervention
Grade II	Any deviation from the ideal intraoperative course <ul style="list-style-type: none">With the need for any additional treatment or interventionNot life-threatening and not leading to permanent disability
Grade III	Any deviation from the ideal intraoperative course <ul style="list-style-type: none">With the need for any additional treatment or interventionLife-threatening and/or leading to permanent disability
Grade IV	Any deviation from the ideal intraoperative course <ul style="list-style-type: none">With death of the patient

Outcomes

The median length of hospital stay was 8 days (IQR 3, 15). Based on the judgment after resolving disagreement, intraoperative complications were recorded in 19 (32%) patients and postoperative complications in 27 (45%) patients. The spread of complications over the grades was higher for postoperative than for intraoperative complications due to the higher number of low grade intraoperative complications. Details of the grade and outcomes by complexity of the intervention are given in [Appendix Table S10](#).

Agreement

The raw categorical agreement between the two investigators was 87% for both intra- and postoperative complications. Details of the kappa statistics and intraclass correlation coefficient are provided in [Table 18](#)

¹ The following events are not defined as intraoperative complications: sequelae, failures of cure, events related to the underlying disease, wrong-site or wrong-patient surgery or errors in indication

Table 18: Agreement of the pilot study (n=60 cases, n=2 raters)

Characteristic	Category	Intraoperative complications ^{1,2}	Postoperative complications ³ (Dindo, Demartines, & Clavien 2004)
Kappa statistics	Agreement (%)	87%	87%
	Kappa (95% CI ⁴)	0.72 (0.48, 0.89)	0.79 (0.69, 0.94)
	Quadratic weighted kappa (95% CI)	0.83 (0.73, 0.94)	0.97 (0.94, 1.00)
Intraclass correlation	Intraclass correlation (95% CI)	0.83 (0.73, 0.90)	0.97(0.95, 0.98)

Practicability

Both investigators rated the practicability of the definition as well as of the classification as 6 on the scale between one and seven.

6.4. Discussion

The Delphi study provides a comprehensive definition of intraoperative complications as any deviation from the ideal intraoperative course occurring between skin incision and skin closure, given that the indication for surgery and the interventions conform to current guidelines. The classification is intended to be applied directly after surgery. The pilot study demonstrated practicability as well as a good interrater agreement.

Strengths and limitations

To the best of our knowledge, there is no validated definition and classification specifically addressing intraoperative complications. With the involvement of a diverse team of experts in terms of experience in practice and of nationality as well as with a high return rate, we were able to launch an extensive discussion about the definition and classification of intraoperative complications. It became evident that the ideal choices were not uniformly the most practical; hence some decisions on definitions and

¹ Based on agreement of both reviewers

² Classification applied immediately postoperatively without including the option "potentially" life-threatening/leading to permanent disability

³ Based on agreement of both reviewers

⁴ CI=Confidence interval

grading were rather pragmatic, resulting in a comprehensive definition and rather simple classification. We are, however, well aware that we are presenting results of the Delphi process and pilot study, which may be subject to further modification and refinement upon full-scale validation. Moreover, the retrospective setting of the pilot study with its small sample is unlikely to cover all possible complications. This is supported by the findings of a review of over eight-thousand interventions, where operative reports missed 13% and discharge summaries missed 14% of complications (Platz and Hyman 2012). Moreover, we found a lower spread of complications across grades for intra- than for postoperative complications, thus our similar interrater agreement for intra- and postoperative complications needs to be interpreted with care. As further research, we intend to conduct a full-scale validation of this instrument in a prospective international multicenter cohort.

Relation to other studies developing classification systems

Many previous attempts to address intraoperative complications origin in either human factor and system errors perspectives or in individual surgeon performance related to surgical education, surgical experience and technical skills. This is reflected by the findings of a review of studies describing technical errors during laparoscopy, in which error analysis was applied in 38% of the studies, whereas in 62% errors were used to quantify surgical performance in an educational setting (Bonrath, Dedy, Zevin, & Grantcharov 2013a). However, while there are several classification systems for postoperative complications, we are unaware of any validated definition and classification system strictly applying to intraoperative complications only.

Fabri et al. developed a classification system of error in operative therapy in an iterative process, interacting with faculty members of one major surgical academic university department (Fabri and Zayas-Castro 2008). The system was then validated three-fold: by the personal judgement of practicability of four experts (50% return rate), by evaluation of agreement between two senior academic surgeons scoring 72 incidences reported upon departmental morbidity and mortality conferences, and by a national survey rating three scenarios. There, medical errors were evaluated as follows: occurrence during evaluation versus execution, slip ('doing the correct thing, incorrectly') versus mistake ('doing the wrong thing, but correctly'), type of error (11 different types, allowing for up to 5 types per complication, e.g. health system error, error in diagnosis, incomplete understanding of problem, judgment error, equipment failure, etc.), corresponding ACGME (Accreditation Council for Graduate Medical Education) core competency, and percentage of adverse outcome attributable to the error. Additionally, the complication was graded as 1: 'definite complication but no injury to patient and no prolongation of length of stay', 2: 'no injury but prolongation of hospitalization', 3: 'injury occurred with temporary disability', 4: 'injury

occurred with permanent disability' and 5: 'death'. Agreement was found to be over 70% in all error categories in both the paired comparison and the national survey (return rate 29.5%). The system was then applied to over nine thousand surgical procedures taking advantage of an established electronic complication self-reporting system. The results suggested that complications are mainly caused by individual surgeons. The study was however controversially discussed, since it suggested systems errors to play a less important role than human deficiencies, while some of the complications caused by individual surgeons could still have their origin in latent factors hidden in the system (Nagpal et al. 2009). Our study shows similar results concerning interrater agreement. The definition and classification are however not strictly comparable, since our system is focussing on the time between skin incision and closure, relying on a second system to capture postoperative events.

In comparison, the classification for postoperative complications proposed by Clavien and Dindo (Dindo, Demartines, & Clavien 2004) was validated as follows: first, in a sample of over six thousand patients, the classification was prospectively applied and the correlation between the complication grades and the complexity of surgery as well as the length of hospital stay was evaluated and found to be high; second, acceptability and reproducibility based on 14 clinical cases was assessed in an international survey involving 10 centers around the world. Acceptability was high and the answers to the scenarios were correct in 90%, independent of the level of training or origin of the surgeon.

Implications for current practice and further research

In order to enhance patient safety and transparency in reporting, there is an urgent need for standardized reporting of intraoperative events. In an evaluation of surgical adverse event reporting with a specific focus on surgical wound infection, anastomotic leak, deep vein thrombosis and surgical mortality, important differences in definitions and duration of follow-up were found, hampering comparisons (Bruce et al. 2001b). In our own study based on a cohort of 46 randomized-controlled trials published in surgical journals, only half of the studies provided exact definitions of postoperative complications, and only 13% of the studies provided definitions of intraoperative complications (submitted for publication). We believe that this classification can be used prospectively and retrospectively, i.e. with direct rating after surgery, and based on the operation record or for ratings of video-recorded interventions. When used retrospectively, the limitation of not capturing all complications due to the likely absence of low grade complications has to be kept in mind. As further research, we plan to refine and validate the proposed definition and classification of intraoperative complications in a multinational, prospective cohort study. We will evaluate

whether there is an association between the classification and the complexity of surgery, as well as between the classification and the duration of surgery. Furthermore, an explorative analysis shall evaluate whether the classification is predictive for a number of adverse postoperative outcomes, adjusted for the most relevant confounders.

Conclusions

We provide a comprehensive definition of intraoperative complications and a straightforward classification with a high interrater agreement and practicability, as evaluated in a pilot study. This classification shall now undergo full-scale validation.

With the proposed work we are contributing to standardized reporting in clinical practice, research and educational settings, ultimately enhancing patient safety and allowing for informed healthcare decision making.

7. How to report multiple outcome metrics in virtual reality simulation

Rachel Rosenthal¹, MD, MSc, Martin W. von Websky², MD, Henry Hoffmann¹, MD, Martina Vitz³, PhD, Dieter Hahnloser⁴, MD, Heiner C. Bucher⁵, MD, MPH, Juliane Schäfer^{1,5}, PhD

¹Department of Surgery, University Hospital Basel, Basel, Switzerland

²Department of Surgery, University Hospital Bonn, Bonn, Germany

³Laparoscopic Training Center, Zurich, Switzerland

⁴Department of Visceral Surgery, University Hospital Lausanne, Lausanne, Switzerland

⁵Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, Basel, Switzerland

Disclosure information: This research was supported by the Swiss National Science Foundation (grant number 32003B-120722). The Basel Institute for Clinical Epidemiology and Biostatistics is supported by grants from Santésuisse and from the Gottfried and Julia Bangerter-Rhyner-Foundation. There are no conflicts of interest to declare.

Publication information: This is a non-final version of an article published in final form in *European Surgery Acta Chirurgica Austriaca* (<http://www.springer.com/medicine/surgery/journal/10353>): Rosenthal R, von Websky MW, Hoffmann H, Vitz M, Hahnloser D, Bucher HC, Schäfer J. How to report multiple outcome metrics in virtual reality simulation. *Eur Surg* 2015;47:202-205 (Rosenthal et al. 2015e)

Abstract

Background: Virtual reality (VR) simulation is increasingly used in surgical disciplines. Since VR simulators measure multiple outcomes, suitable methods are needed to address multiplicity or combine evidence into a lower dimensional outcome. We aim to present an algorithm for summarizing multiple VR outcome metrics into a total score and to illustrate the approach with two real data examples.

Methods: We present an algorithm for combining multiple VR outcomes into dimension summary measures, which are integrated into a meaningful total score. We re-analyzed the data of two VR studies applying the algorithm. First, 9 children with low, 23 with high experience in videogames, 20 residents and 14 board-certified surgeons were compared concerning VR laparoscopy performance. Second, surgical novices randomized to free versus structured VR training were compared with each other and with experts during simulated cholecystectomy.

Results: The suggested algorithm was successfully applied to both studies. First, video gaming in children improved VR performance, but children were outperformed by residents and board-certified surgeons performed best, as confirmed when taking board-certified surgeons instead of all participants as the reference group. Second, cholecystectomy performance increased from free to structured training to experts, as confirmed by learning curves.

Conclusions: We present and applied an algorithm for combining multiple VR outcome metrics into a meaningful total score. We herewith address selective outcome reporting and multiplicity. Given the increasing number of articles in the field, standardized and transparent analysis and reporting of VR outcomes is important to enhance the validity of VR-related research.

7.1. Introduction

Virtual reality (VR) simulation is increasingly being used for training purposes. A recent Cochrane review investigating the impact of VR training on surgical performance in trainees with limited laparoscopic experience showed an advantage of VR training over no training as well as over box-training: VR training decreased the duration of surgery and improved operative performance (Nagendran et al. 2013).

Tightly linked to the use of VR for training purposes, VR may be used for assessment of surgical skills. A prerequisite is the simulator's construct validity, referring to the ability of the simulator to capture the hypothetical quality it was designed to assess (Oropresa et al. 2010). In other words, the simulator should discriminate between different levels of experience, such as between experts and novices (Feinstein and Cannon 2001;Gallagher et al. 2003;Oropresa, Lamata, Sanchez-Gonzalez, Pagador J.B., Garcia M.E., Sanchez-Margallo, & Gomez 2010). Using VR for assessment may be important to compare the performance (1) of different training groups with each other and (2) of one trainee or different training groups to a reference group (e.g. benchmark). According to a European consensus, a competency-based VR curriculum should consist of tasks of varying difficulty levels, each performed until thresholds - which are pre-defined based on expert performance - are passed twice within a single training session (van Dongen et al. 2011). The consensus was based on a multicenter validated training program developed within a large international team extensively experienced in VR training and assessment (van Dongen, Ahlberg, Bonavina, Carter, Grantcharov, Hyltander, Schijven, Stefani, van der Zee, & Broeders 2011). Alternatively, the upper quartile of peer group performance may serve as target level of performance (von Websky et al. 2012).

VR simulators objectively measure a large number of different outcome metrics. As a consequence, when analyzing and reporting such data, a number of challenges must be faced: Should all outcomes or only parts of them be analyzed and reported and in the latter case based on which selection process? Can outcomes be summarized in a meaningful way? How can outcomes on different scales be handled? Is it possible to weigh outcomes of different clinical importance? Can outcomes from different simulators be compared? Can outcomes from different studies be compared?

The aim of this article is to present an algorithm for summarizing multiple outcomes of different dimensions and weighing them within a total score. We describe a framework for reporting VR outcome metrics, where we (1) summarize results from multiple outcomes of the same dimension in a suitable summary measure, and (2) integrate the dimension summary measures into a meaningful total score. We illustrate the approach with two real data examples. Given the increasing number of articles in the field, a standardized and transparent

approach, minimizing the risk of selective outcome reporting bias and addressing multiplicity issues, is of utmost importance.

7.2. Methods

Framework for analyzing and reporting multiple VR outcome metrics

To summarize multiple VR outcomes into a total score, we propose to proceed as follows. First, dimensions of outcomes need to be defined according to a priori expert judgment. We propose to summarize outcomes under “accuracy” and “efficiency”. With accuracy we refer to “safety” (no harm) and “efficacy” (completion of the task as intended) parameters. With “efficiency” we refer to “time to complete the task” and “economy of movement” (i.e. instrument path length) parameters. Then, the following four-step procedure is carried out (underlying algorithm: [Table 19](#)):

Standardize each original outcome (stated a priori to be incorporated into a summary measure of VR performance) so that outcomes are on comparable scales. To obtain easily comparable scales, we propose to standardize to mean 100 and standard deviation (SD) 15 (Streiner and Norman 2008).

Replace outcomes grouped within the three dimensions (accuracy, time and path length) by their average to obtain summary measures according to the a priori defined standard.

Standardize each new summary measure to mean 100 and SD 15 while reversing the directionality of the measures where necessary, so that higher values are favourable for all three mean summary measures.

Finally, based on a proposal to combine multiple outcomes in clinical trials by Schouten (Schouten 2000) and earlier work by Goldsmith (Goldsmith et al. 1993), compute a total score with the option of weighing the dimensions. In our examples, we considered accuracy twice as important as the efficiency parameters, i.e. time and path length, together.

The rationale for standardizing to mean 100 and SD 15 (rather than mean 0 and SD 1 (z-score)) is to end up with positive numbers to which most people are more accustomed. This procedure is supported by the norming procedure of many intelligence tests to mean 100 and SD 15 (Streiner & Norman 2008). The inclusion of weighing is for illustrative purposes only; the chosen weights need to be interpreted as examples.

Table 19: Algorithm for handling multiple outcomes, step 4 derived from Schouten et al. (Schouten 2000)

<p>1. Standardization</p>	<p>Standardize each original outcome stated a priori to be considered for analysis to mean 100 and standard deviation (SD) 15</p>	<p>Outcomes are on comparable scales prior to calculating mean summary measures of VR performance representing the pre-specified dimensions (e.g. accuracy, time and path length)</p>
<p>2. Calculation of mean summary measures for each dimension</p>	<p>Replace outcomes grouped within the pre-specified dimensions (e.g. accuracy, time and path length) by their average</p>	<p>Reduce multiplicity and obtain mean summary measures defined according to a priori expert judgment</p>
<p>3. Re-Standardization & unification directionality</p>	<p>Standardize each new summary measure (e.g. accuracy, time and path length) to mean 100 and SD 15 while reversing the measures so that higher values correspond to better performance throughout (e.g. for time and path length)</p>	<p>Outcomes are on comparable scales and higher values correspond to better performance for all summary measures</p>
<p>4. Calculation of a weighted average as total performance score</p>	<p>Calculate a total performance score for each study participant as the weighted average of the dimension summary measures</p>	<p>Obtain a clinically meaningful total performance score for each study participant where dimensions are weighted (e.g. considering accuracy twice as important as efficiency, i.e. time and path length together)</p>

Data examples

To illustrate the above framework with real data examples from VR assessment, we re-analyzed the data of two previously published VR studies (Rosenthal, Geuss, Dell-Kuster, Schafer, Hahnloser, & Demartines 2011; von Websky, Raptis, Vitz, Rosenthal, Clavien, & Hahnloser 2013).

In the first study, 9 children with low experience in videogames, 23 with high experience in videogames, 20 residents and 14 board-certified surgeons were compared for VR laparoscopy performance. A total of 23 outcomes of 6 basic VR tasks in the three dimensions accuracy, time and path length were considered (Rosenthal, Geuss, Dell-Kuster, Schafer, Hahnloser, & Demartines 2011). Each task was carried out twice.

In the second study, surgical novices were randomized to either self-controlled basic VR training ("free training", n=32) or VR training based on peer-group-derived benchmarks ("structured training", n=34). Thereafter the two training groups were compared with each other as well as with a group of three experts during a foreseen minimum of 60 iterations of simulated cholecystectomy performance, 10 for each of 6 anatomically different cases (von Websky, Raptis, Vitz, Rosenthal, Clavien, & Hahnloser 2013). Measurement of simulated cholecystectomy performance consisted of 5 outcomes in the three dimensions accuracy, time and path length.

In both studies the Simbionix laparoscopic modules were used (Simbionix USA Corp, Cleveland OH). In the first study, the software was running on a Xitact/Mentice hardware (Mentice SA; formerly Xitact SA, Morges, Switzerland) and in the second study on the LAP Mentor™ (Simbionix USA Corp, Cleveland OH).

Statistical analyses

Using the algorithm described above (Table 19), we calculated mean summary measures within the three dimensions (accuracy, time and path length) and the total performance score with accuracy considered twice as important as time and path length together.

We used analysis of variance to compare the different participant groups in each study. To perform these analyses, we first calculated the mean of the two task repetitions in the first data set and summarized the repeated measures over time into a single mean summary measure per participant in the second data set. For each analysis, we report estimates (with simultaneous 95% confidence intervals) of the difference in outcome for all pairwise comparisons. We used Tukey's honest significant difference to adjust the confidence intervals (Tukey 1953).

To visualize the mean summary measures and total score by participant group, we used box plots. For the second data example, we used dot plots to display the results because only three experts were involved.

We carried out one additional analysis for each data set. In the first data set (Rosenthal, Geuss, Dell-Kuster, Schafer, Hahnloser, & Demartines 2011), we calculated the dimension summary measures and the total score with the results of the board-certified surgeons taken as the reference population, as suggested for proficiency-based learning (van Dongen, Ahlberg, Bonavina, Carter, Grantcharov, Hyltander, Schijven, Stefani, van der Zee, & Broeders 2011). We then contrasted the performance of children with low/high experience in videogames and of residents with that of the board-certified surgeons. In the second dataset (von Websky, Raptis, Vitz, Rosenthal, Clavien, & Hahnloser 2013), we used a linear mixed-effects model, fit to the first 10 iterations of cholecystectomy case number 1 per participant, to assess the learning curves for the three participant groups. For this analysis, the overall mean and standard deviation of all first iterations were taken as the reference. To accommodate curved data, we included a linear and a quadratic effect of iteration in our model. To indicate that the different groups may not only show different linear trends over iterations but also show different quadratic trends, we included interaction effects of both with participant group.

For our analyses, we used R version 3.0.1 (R Foundation for Statistical Computing, Vienna, Austria) and the R add-on packages *lme4* version 1.0-5 (Bates et al. 2013) and *lattice* version 0.20-24 (Sarkar 2008).

7.3. Results

Results example 1: Basic VR laparoscopy performance of children, residents and board-certified surgeons (Rosenthal, Geuss, Dell-Kuster, Schafer, Hahnloser, & Demartines 2011)

Participant characteristics by group are presented in table 1 of the original publication (Rosenthal, Geuss, Dell-Kuster, Schafer, Hahnloser, & Demartines 2011). For the present analysis, one participant who did not complete basic task number 3 was not considered for analysis. The mean summary measures within the three dimensions of accuracy, time and path length and the total score gradually increased from children with low/high experience in videogames to residents and board-certified surgeons (Figure 6a, b). So in summary, our re-analysis of these data confirmed that video gaming in children improves performance on a VR trainer, but expectedly children were outperformed by residents and board-certified surgeons performed best (Table 20). The variance of the mean summary measures and the total score decreased from children with low/high experience in videogames to residents to board-certified surgeons (Figure 6a, b).

Our additional analysis using the board-certified surgeons as the reference population (instead of all study participants) showed that children with low/high experience in videogames and residents both presented less favourable mean summary measures and lower total score than the board-certified surgeons (Figure 7a, b).

Table 20: Difference in mean summary measure or total performance score between the four participant groups: children with low (n=9) and high (n=22) experience in videogames, residents (n=20) and board-certified surgeons (n=14) (Rosenthal, Geuss, Dell-Kuster, Schafer, Hahnloser, & Demartines 2011).

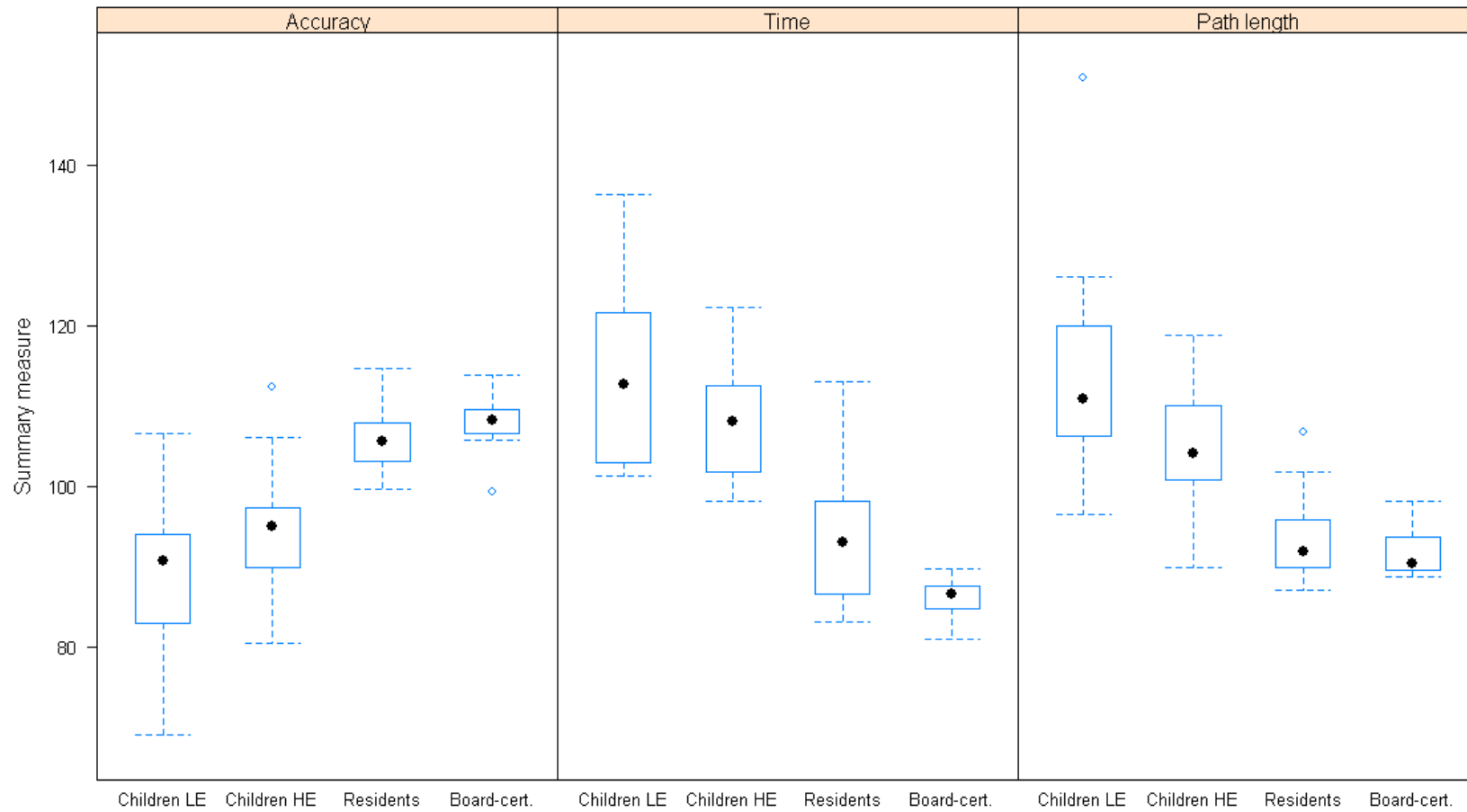
Difference in mean summary measure or total performance score (95% simultaneous [¶] CI)	Accuracy	Time	Path length	Total
Children HE – Children LE	4.9 (-2.1, 11.9)	-6.2 (-14.4, 2.0)	-10.2 (-18.2, -2.2)**	8.4 (-0.7, 17.5)
Residents – Children LE	17.3 (10.2, 24.4)***	-20.4 (-28.8, -12.1)***	-21.9 (-30.0, -13.8)***	26.1 (16.9, 35.4)***
Board-cert. – Children LE	19.4 (11.8, 26.9)***	-28.5 (-37.4, -19.7)***	-23.6 (-32.2, -15.0)***	30.1 (20.3, 40.0)***
Residents – Children HE	12.4 (6.9, 17.9)***	-14.2 (-20.6, -7.8)***	-11.7 (-18.0, -5.5)***	17.7 (10.6, 24.8)***
Board-cert. – Children HE	14.4 (8.4, 20.5)***	-22.3 (-29.4, -15.2)***	-13.4 (-20.3, -6.5)***	21.7 (13.9, 29.6)***
Board-cert. – Residents	2.1 (-4.1, 8.2)	-8.1 (-15.3, -0.9)*	-1.7 (-8.7, 5.4)	4.0 (-4.0, 12.0)

Abbreviations: CI, confidence interval; Children HE/LE, children with high/low experience in videogames; Board-cert., board-certified surgeons.

[¶]Tukey honest significant differences (Tukey 1953)

*** P<0.001; ** P<0.01; * P<0.05.

Figure 6: Boxplots comparing children with low (n=9) and high (n=22) experience in videogames (LE and HE), residents (n=20) and board-certified surgeons (n=14) (Board-cert.) (Rosenthal, Geuss, Dell-Kuster, Schafer, Hahnloser, & Demartines 2011). (A) Accuracy, time and path length mean summary measures.



(B) Total virtual reality (VR) performance score with accuracy considered twice as important as time and path length together.

97

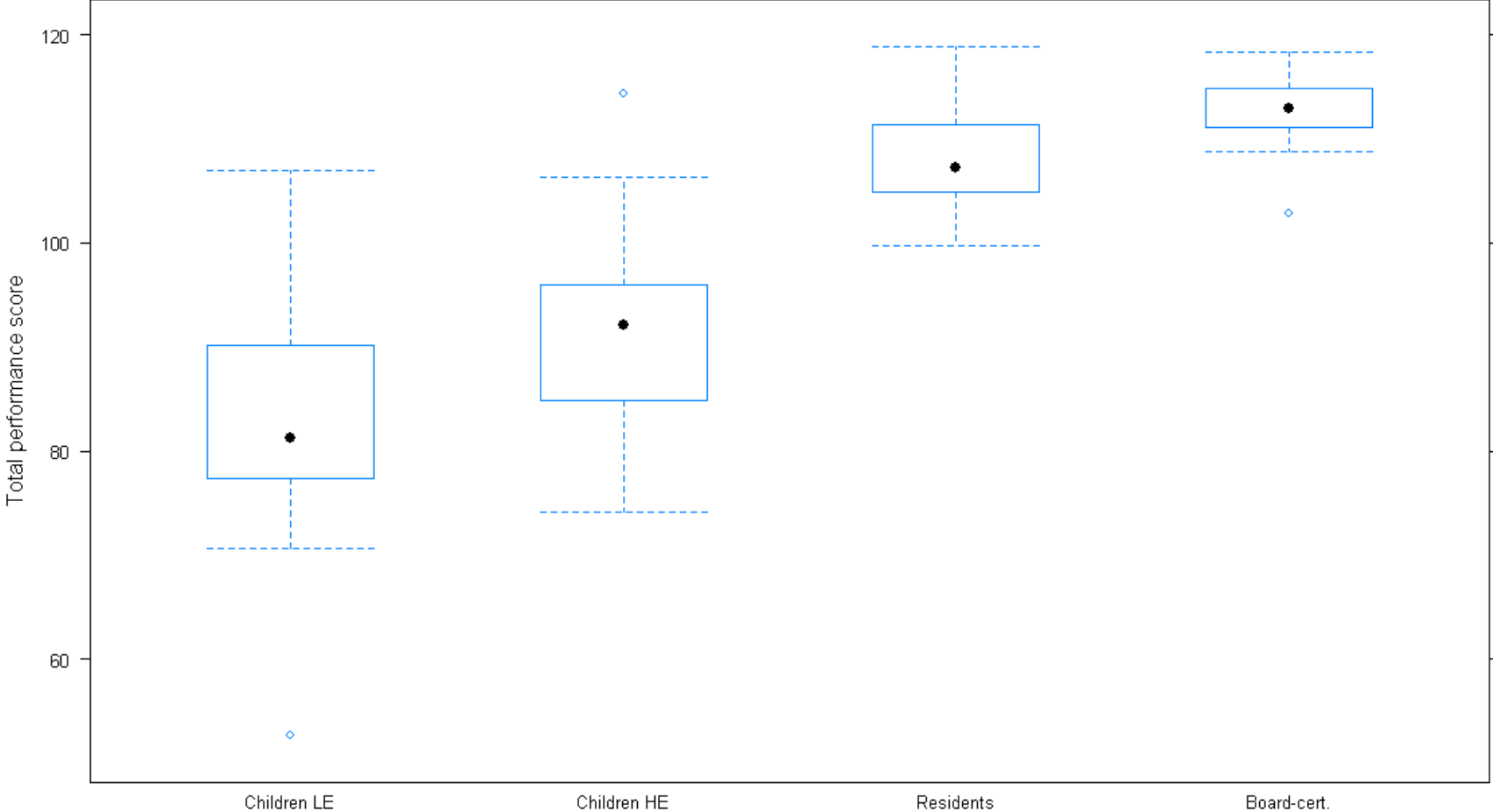
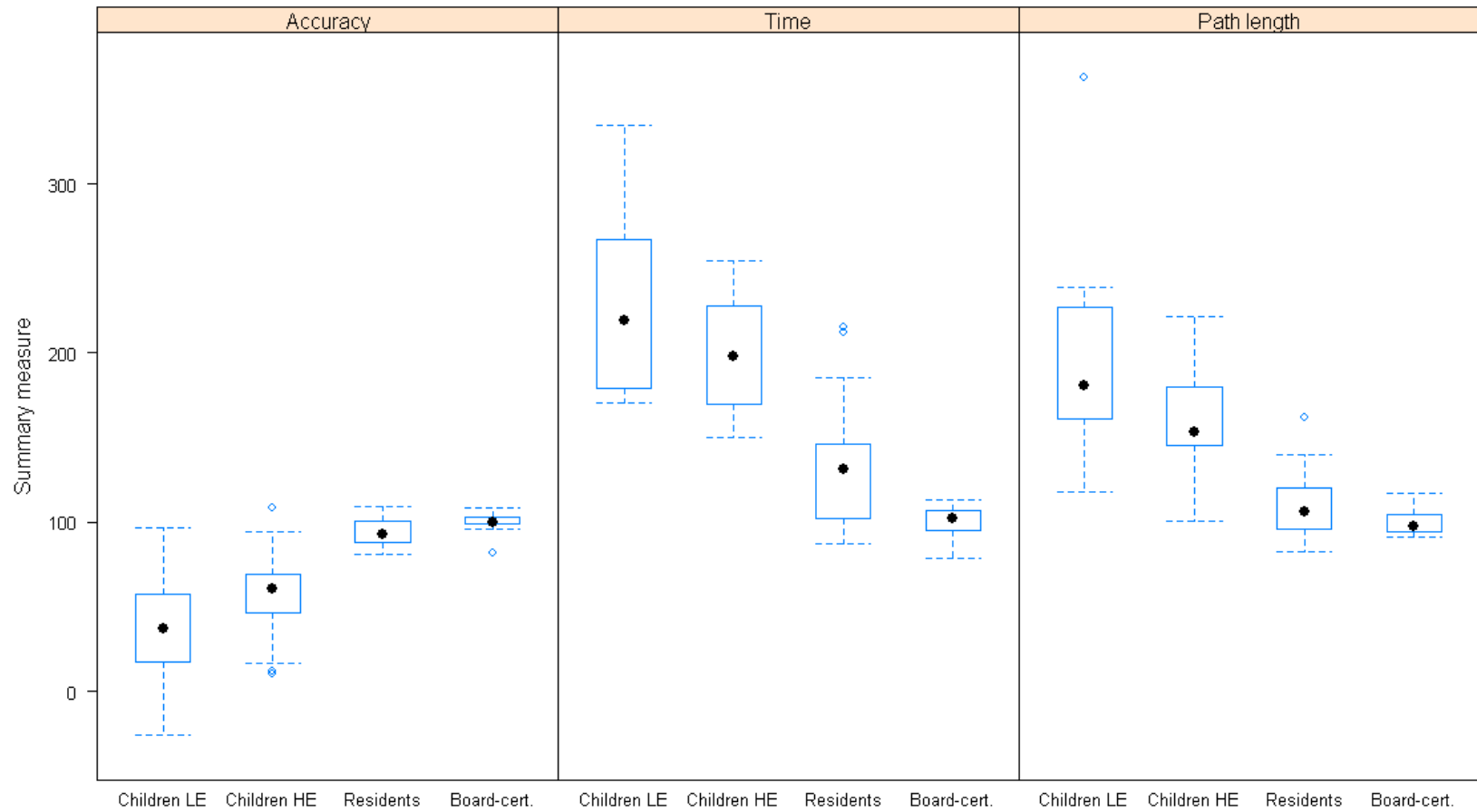
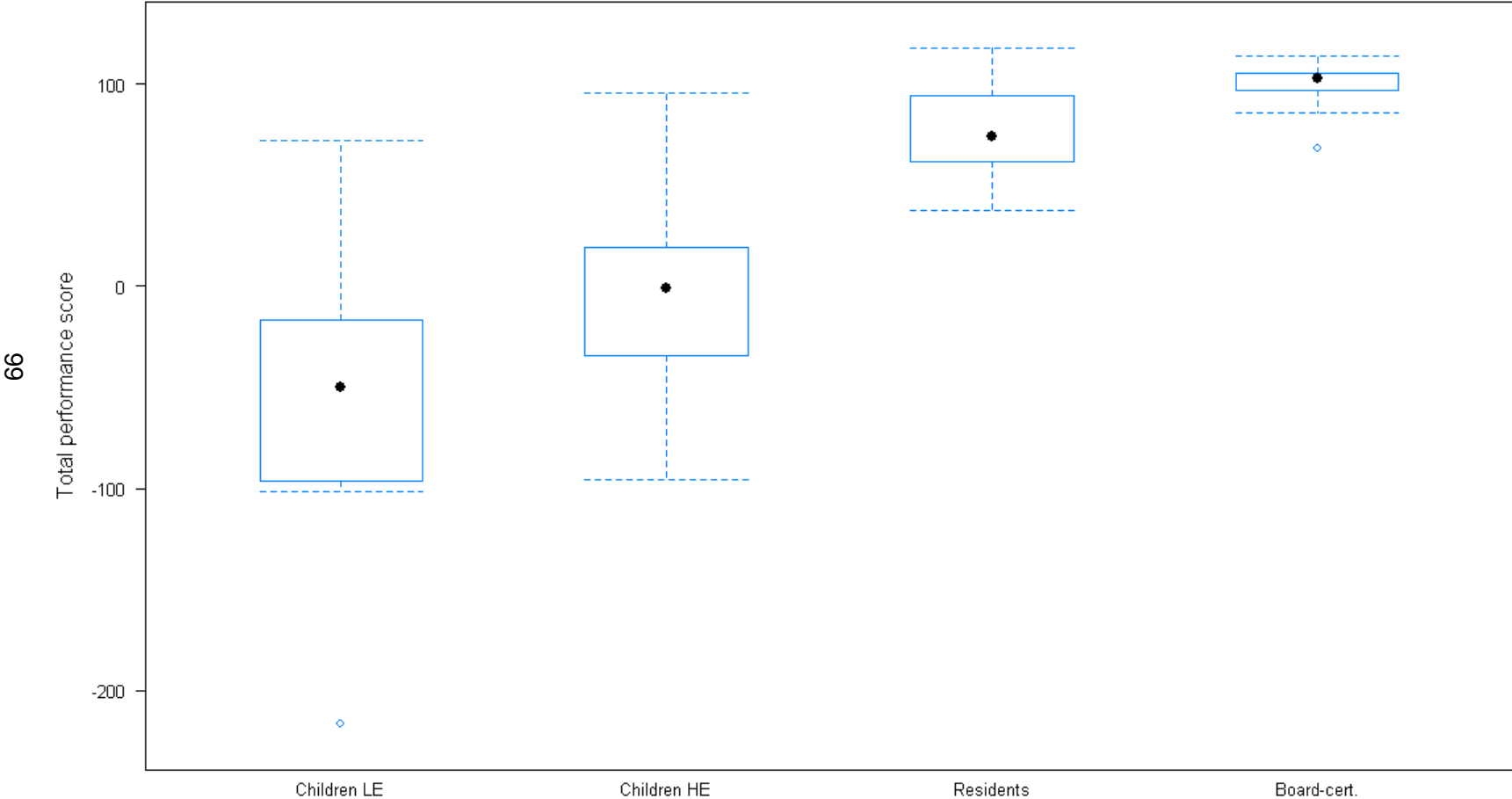


Figure 7: Boxplots comparing all three children with low (n=9) and high (n=22) experience in videogames (LE and HE) and residents (n=20) with board-certified surgeons (n=14) (Board-cert.) with the latter taken as the reference population (Rosenthal, Geuss, Dell-Kuster, Schafer, Hahnloser, & Demartines 2011) (A) Accuracy, time and path length mean summary measures.



(B) Total virtual reality (VR) performance score with accuracy considered twice as important as time and path length together.



Results example 2: VR cholecystectomy performance of the free training group, structured training group and experts (von Websky, Raptis, Vitz, Rosenthal, Clavien, & Hahnloser 2013)

Trainee demographics are described in the original publication (von Websky, Raptis, Vitz, Rosenthal, Clavien, & Hahnloser 2013). For the present analysis, one participant with missing values in the number of serious complications was not considered for analysis. The median (interquartile range) number of iterations was 55 (42.5, 63.5), 57 (35.75, 64.25) and 44 (43.5, 51) for participants in the free training, structured training and expert group, respectively.

With repeated measures over time summarized into a single mean summary measure per participant, VR cholecystectomy performance gradually increased from free training to structured training to experts (Figure 8a, b; Table 21).

When accounting for a maximum of 10 iterations of cholecystectomy case number 1 per participant, the three experts showed the best total performance score and steepest learning curve, followed by the structured and the free training group (Figure 9a). The smooth curves represent local averages. From figure 9b we see the same pattern, with predicted values (smooth curves, based on the fixed part of the model) and actual values of the total score. Experts started off with the best performance – followed by participants in the structured and the free training group – and then showed a modest learning curve over 10 iterations that may reach a plateau, whereas the learning curves of surgical novices in both the free and the structured training group may continue to increase beyond 10 iterations.

Table 21: Difference in mean summary measure or total performance score between the three participant groups: free training group (n=31), structured training group (n=34) and experts (n=3) (von Websky, Raptis, Vitz, Rosenthal, Clavien, & Hahnloser 2013).

Difference in mean summary measure or total performance score (95% simultaneous [¶] CI)	Accuracy	Time	Path length	Total
Structured training – Free training	4.2 (-3.8, 12.3)	-7.1 (-15.5, 1.3)	-9.1 (-17.0, -1.2) [*]	5.9 (-0.7, 12.4)
Experts – Free training	11.6 (-8.1, 31.3)	-25.3 (-45.7, -4.8) [*]	-22.9 (-42.2, -3.7) [*]	16.7 (0.7, 32.7) [*]
Experts – Structured training	7.3 (-12.3, 27.0)	-18.2 (-38.6, 2.1)	-13.8 (-33.9, 5.4)	10.8 (-5.1, 26.8)

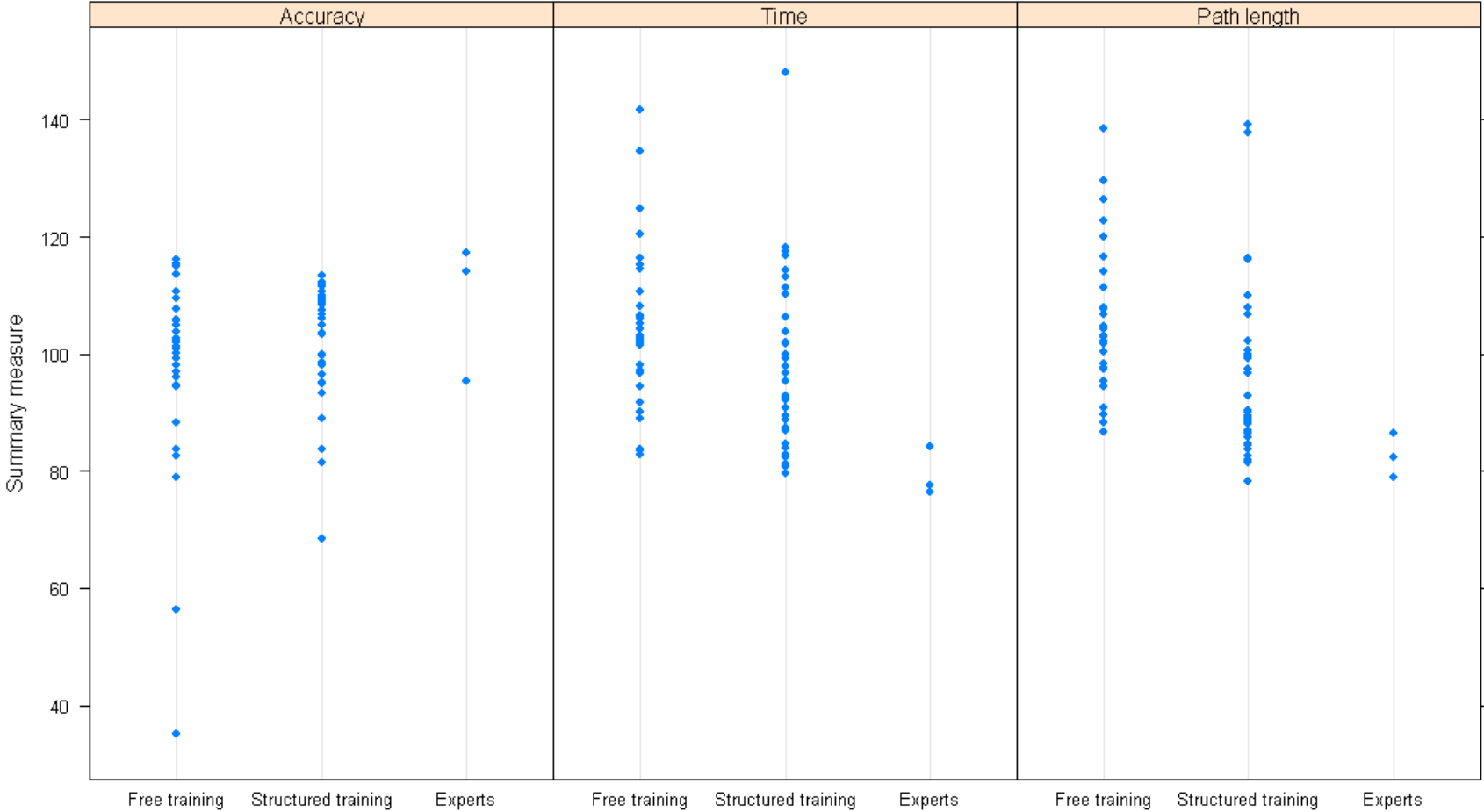
101

Abbreviations: CI, confidence interval.

[¶]Tukey honest significant differences (Tukey 1953)

^{*} P<0.05.

Figure 8: Dotplots comparing surgical novices in the free training (n=31), structured training (n=34) and expert group (n=3) (von Websky, Raptis, Vitz, Rosenthal, Clavien, & Hahnloser 2013) **(A)** Accuracy, time and path length mean summary measures.



(B) Total virtual reality (VR) performance score with accuracy considered twice as important as time and path length together.

103

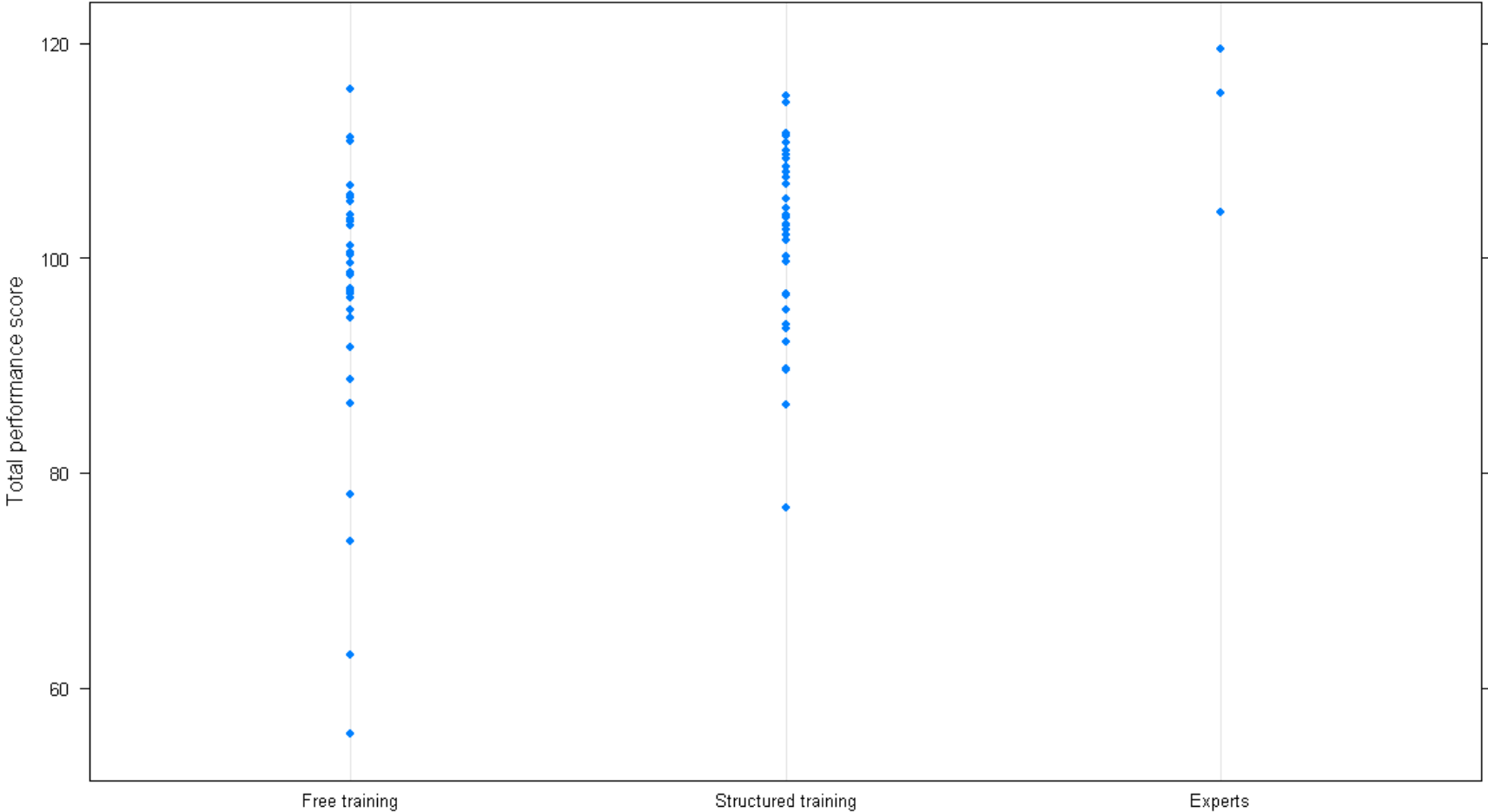
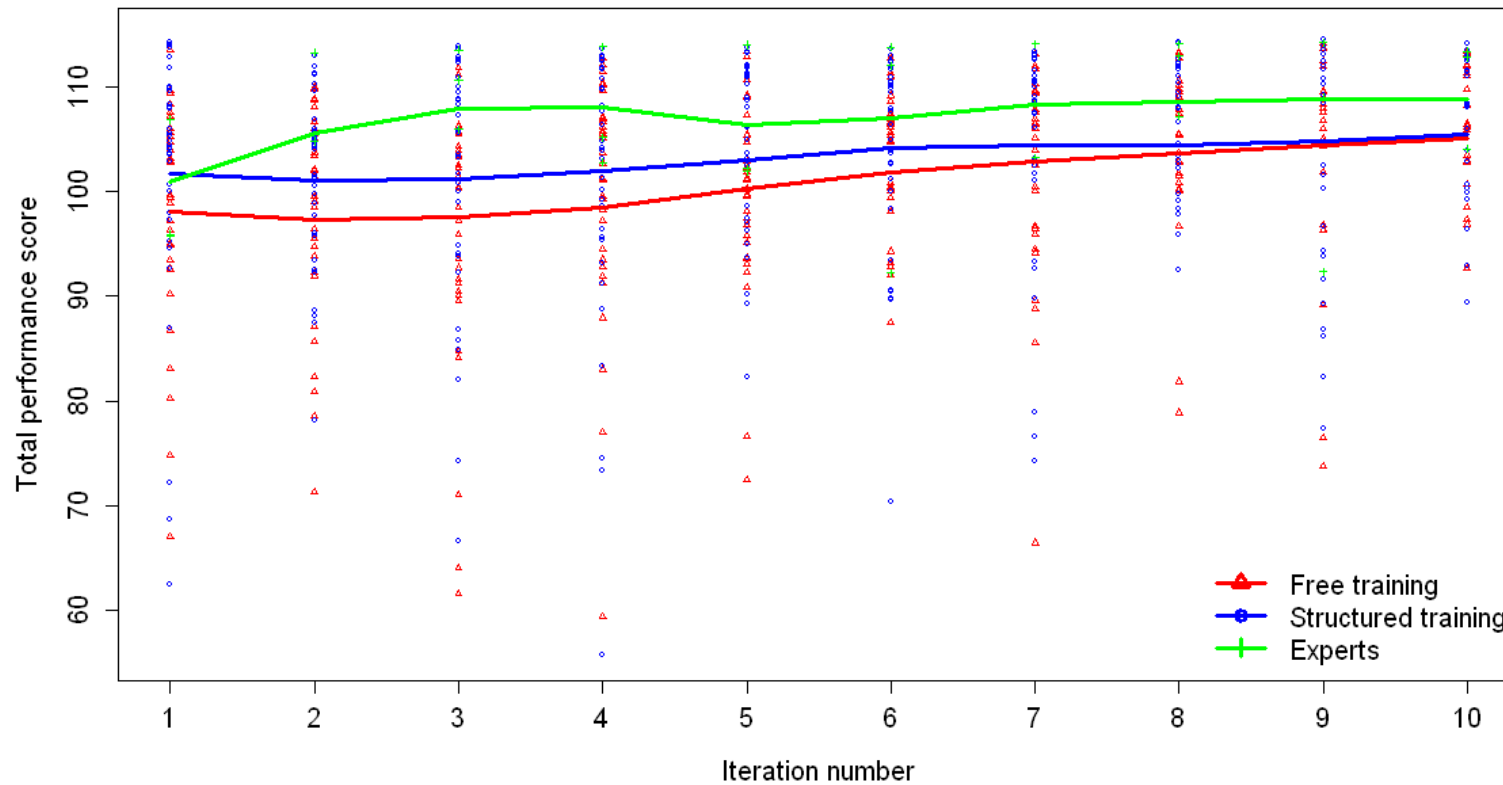
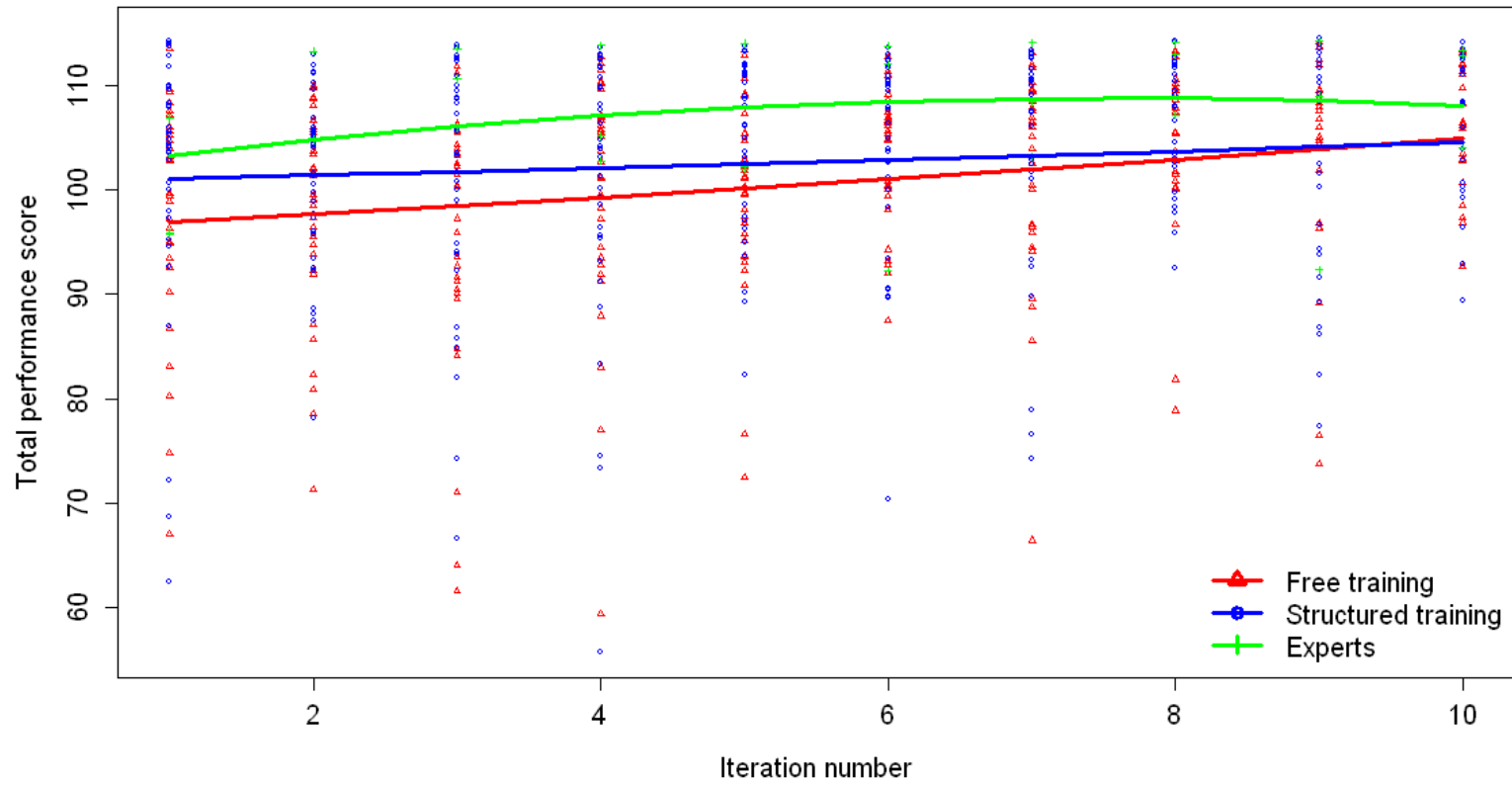


Figure 9: Total virtual reality (VR) performance score – with accuracy considered twice as important as time and path length together – over iterations for participants in the free training (n=31), structured training (n=34) and expert group (n=3) (von Websky, Raptis, Vitz, Rosenthal, Clavien, & Hahnloser 2013) (A) The smooth curves are local averages found using the default loess smoother in R version 3.0.1 (R Foundation for Statistical Computing, Vienna, Austria).



(B) The smooth curves are the predicted values (based on the fixed part of the linear mixed-effects model) of the total VR performance score.



7.4. Discussion

We present an algorithm for (1) combining multiple VR outcome metrics into suitable summary measures of the same dimension, and (2) integrating the dimension summary measures into a meaningful total score. We herewith address the risk of selective outcome reporting bias and multiplicity issues.

When analyzing and reporting multiple outcomes, three different strategies are possible, the choice depending on the study goal and setting: (1) reporting all measured metrics, (2) reporting part of them, or (3) summarizing multiple outcomes to mean summary measures within dimensions or to a total score, as presented within this work. If the results of all measured metrics are presented, a graphical display of all individual outcomes may be an appropriate choice, especially in exploratory studies. If groups are compared for all measured metrics using hypothesis testing with no adjustment for multiplicity, the chance to find at least one significant difference when in fact there is none (family-wise type I error) may increase (Neuhauser 2006; Sterne and Davey 2001). In such instances, procedures to control the family-wise type I error rate are recommended (Neuhauser 2006; Sterne & Davey 2001) or summarizing outcomes as described within our framework. If only parts of the data are reported, there is a risk of data-driven selective reporting. On the other hand, some of the metrics are most likely related to each other, for example economy of movement and path length are similar outcomes. In this case, one of these two variables may be selected as a representative for both variables. Such a decision must be taken a priori based on expert knowledge, and may not be a post-hoc data-driven selection. Importantly, the selection process needs to be transparently communicated in the final report. Our approach of summarizing multiple outcomes is simple and straightforward: We choose outcomes a priori and (1) if they are not on comparable scales, standardize them and (2) then summarize them within a priori defined dimensions. To obtain a total score, (3) we standardize the dimension summary measures, unify their directionality and then (4) calculate the total score. The approach allows for variations: During step 2 and 4, either the average or a weighted average may be calculated. The choice of weights may be based on clinical judgment, by giving more weight to accuracy parameters, i.e. safety and efficacy, than to efficiency parameters, i.e. time to complete a task or economy of movement. In the examples presented in this article, we a priori chose to weigh accuracy, time, and path length in a 4 to 1 to 1 ratio. Alternatively, principal component analysis may be applied. With this approach, the first principal component, which spreads out the scores as much as possible, may be used as a summary measure in step 2 and 4 (Hotelling 1933; Pearson K 1901).

Once the summary measures or total score are calculated, they can be further analyzed as any other continuous outcome. However for data interpretation it has to be kept in mind that the units are standard deviations and the score has been inflated by 15. In our two examples, we used analysis of variance to compare

different participant groups and addressed two additional aspects of data analysis: benchmarking and the analysis of repeated measures over time. In proficiency-based learning, a single new observation from another trainee or several new observations from one or more groups of trainees may be compared to a reference group, for instance a group of experts. Importantly, the reference population needs to be large enough to serve as a benchmark. Benchmark values may be established once and then serve as a reference for subsequent observations. Some intelligence tests for instance are taking advantage of a norming sample from the general population with a mean of 100 and a standard deviation of 15 IQ points (Streiner & Norman 2008). To generate however such a representative reference population, the sample size of the reference population by far needs to exceed the one presented in this paper. Since we standardize all observations with the overall mean and standard deviation of the reference population, no assumptions concerning the distribution of the reference population are being made. The resulting score, as any other continuous outcome, does however not necessarily present a normal distribution. Our results suggest that the total score, as described within the present work, may readily be used for comparing the results of individuals or groups of individuals to a reference group, an important issue in proficiency-based training. The second aspect is as well important in the setting of training: the analysis of repeated measures over time. Repeated observations within the same individual generate clustered data, which need to be addressed with appropriate statistical methods. The reason is that observations within an individual (i.e. the cluster) tend to be more similar than those between individuals, which in turn has an impact on measures of uncertainty associated with effect estimates. Disregarding clustered data may lead to wrong conclusions. In the simplest case, only one task repetition per participant is important and analyzed, such as the first or the last observation, and thus data may be handled as independent observations. However this approach has disadvantages, because most of the available information is not taken into account. If overall performance disregarding training effects from task repetitions is important, repeated measures over time may be summarized into a single summary measure per participant prior to data analysis, as in the first analysis of the second data set. If however the learning curve – with all repeated measures over time – is of interest, a linear mixed-effects model may be used, as illustrated in the second analysis of the second data set. Other options for the analysis of clustered data are beyond the scope of this article.

In conclusion, VR simulation is increasingly being used for assessment and training purposes, especially in minimally invasive surgery. VR allows group comparisons, unlimited task repetition and measurement of a large number of outcome metrics. When analyzing and reporting such outcome data, the use of standardized, statistically sound and transparent methods is of utmost importance. We describe an algorithm to combine the evidence from multiple outcome metrics into a lower dimensional outcome. As further research for the future, test-retest reliability of the mean summary measures and the total score on the same

simulator, followed by test-retest reliability on different simulators should be explored. The ultimate goal of this work and of future research is to enhance the validity of VR-related research reports.

8. Disregarding paired data in surgical research – evaluation of current practice and estimation of implications on study results (working project)

Involved team-members: Raoul A. Droeser^{1¶}, Salome Dell-Kuster^{1,2*¶}, Tarik Delko¹, Viktoria Gloy², Hannah Evald², Lars Hemkens², Juliane Schäfer², Heiner C. Bucher², Rachel Rosenthal^{1§}

¹Department of Surgery, Basel University Hospital, Switzerland

²Basel Institute for Clinical Epidemiology and Biostatistics, Basel University Hospital, Switzerland

[¶]Shared first authors, [§]Senior author

Disclosure information: This work is supported by a grant of the University of Basel, Switzerland

Publication information: This work is ongoing

8.1. Introduction

In surgical studies, paired data resulting from multiple interventions in a single patient (e.g. bilateral hernia, bilateral interventions in extremity surgery) and from multiple outcome assessments over time are relatively common. If clustered data are not accounted for in the statistical analysis, the standard errors of the estimates might be under- or overestimated, which may have an effect on the overall study conclusions (Campbell et al. 2004; Campbell et al. 2012; Lee and Thompson 2005).

Clustered data need be accounted for in both the design phase and the analysis phase of the study. As for the study design, this concerns in particular the definition of the in-/exclusion criteria (allowing or not to generate clustered data), randomization procedure (for instance randomization of the individual versus randomization of the hernia) and sample size calculation (accounting for clustered data) (Lee et al. 2012; Lee & Thompson 2005). Concerning data analysis, the statistical methodology for handling clustered data has been extensively described (Hanley et al. 2003; Zeger and Liang 1986), but not systematically implemented in current practice.

In a systematic review of oral health reports, the clustered nature of data had been accounted for in sample size calculations in 65% of the analyzed studies (Froud et al. 2012). As for data analysis, clustered data do not seem to be systematically accounted for: In a systematic review of implant dentistry, clustered data were not accounted for in 33% of the analyzed RCTs (Cairo et al. 2012). Of note, in a review of 69 ophthalmic RCTs, none of them accounted for clustering in sample size calculation and only 5 studies accounted for non-independence in the statistical analysis (Durkalski et al. 2003; Lee, Cheng, & Fong 2012). To the best of our knowledge, a systematic analysis on the appropriateness of methods used for handling clustered data in surgery has not yet been conducted, neither has the effect of bias from disregarding clustered data been estimated.

The aim of this study is to analyze in randomized-controlled trials in patients undergoing hernia repair without exclusion of bilateral surgery i) to what extent the presence of clustered data has been accounted for in the study design (in-/exclusion criteria, randomization, sample size calculation) and statistical analysis (type and appropriateness of used methods) and ii) to estimate the effect of disregarding clustered data on the overall results.

8.2. Methods

We included in this evaluation RCTs with patients undergoing inguinal hernia repair comparing any type of intervention that is applicable on each side separately, for which the outcome is assessable on each side separately and in which patients with bilateral hernias were not excluded. Studies with ongoing

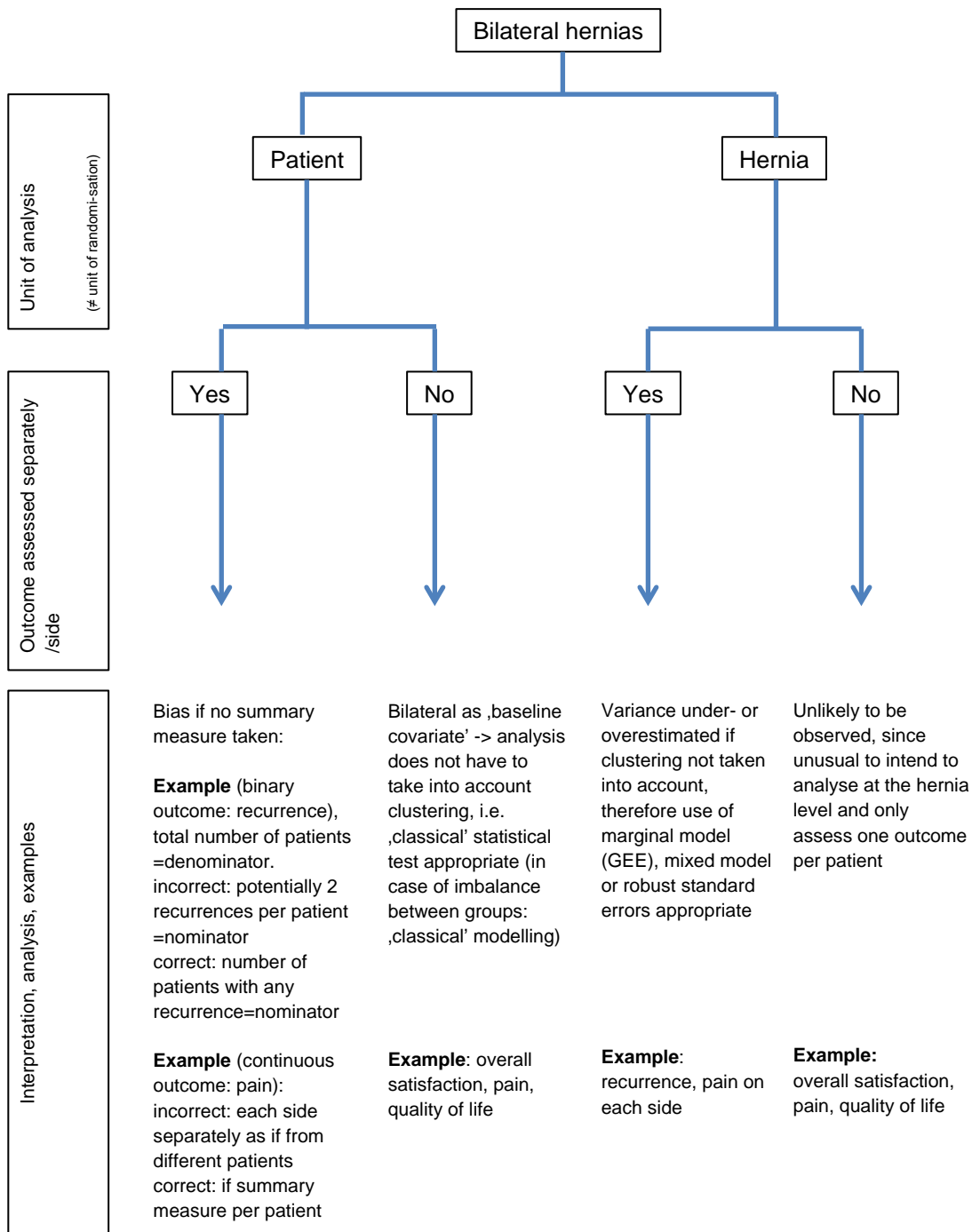
recruitment and follow-up trials were excluded. RCTs were identified through Medline, Embase and Cochrane library search.

Two independent reviewers with medical content expertise judged the eligibility of the RCTs and extracted data on study baseline characteristics, study quality (risk of bias) and handling of clustered data in study design and analysis.

Disagreements were resolved by discussion, Descriptive statistics were used for analysis.

We applied a conceptual framework of different approaches to handle paired data (Figure 10).

Figure 10: Conceptual Framework for handling paired data



8.3. Preliminary results

A total of 4024 records were identified through database search, 2613 after duplicate removal. Out of these, 2326 studies were excluded due to in-/exclusion criteria with 287 studies remaining for full-text eligibility assessment and data extraction, if applicable. The evaluation of these 287 RCTs and the extraction are still ongoing. Preliminary findings suggest that many studies do not use appropriate methods to account for the paired data structure.

8.4. Outlook

To estimate the effect of disregarding clustered data on the overall results, we will attempt to obtain original individual patient data of published RCTs with different proportions of clustered data from the respective authors. The original data of these trials will be analyzed twice, once with appropriate statistical methods accounting for clustered data and once disregarding the presence of clustered data allowing estimating the difference in overall conclusions depending on the type of analysis. Should no original data be available, a simulation study will be conducted to estimate the impact of disregarding paired data on overall trial conclusions.

9. General discussion and conclusion

9.1. Principal findings

Our investigation of reporting in surgical research focused on the assessment of current practice and the development of potential solutions to identified challenges. Our main findings are summarized as follows:

Studies investigating current practice

When designing and reporting surgical trials, trialists referenced *systematic reviews* to summarize findings, but not explicitly to inform trial design and not to integrate the new trial results.

During study conduct, slow recruitment was the predominant cause for *early trial discontinuation*. Trial discontinuation was more frequent in surgical than in medical trials, both for any reason and for slow recruitment. Trial discontinuation in turn was found to be a strong predictor for non-publication and thus had an impact on reporting.

As for adverse intraoperative events, we found *intraoperative complications* rarely to be reported; some were subsumed with postoperative complications. In case they were reported, they mostly lacked a definition. In contrast, for postoperative complications, for which several classification systems have been published, about half of the studies provided this information. Neither the type of study intervention nor whether the primary outcome was an adverse event did significantly impact on these results.

In regard to the handling of *paired data*, preliminary findings suggest that many studies in hernia surgery do not use appropriate methods to account for paired data in the study design, analysis and report.

Studies addressing potential solutions

In order to address some identified challenges when reporting surgical research, we developed a 10-step practical *guide to write a clinical study protocol*. We outlined key methodological issues important when planning an ethically and scientifically sound surgical research project, thus anticipating in the study design phase critical issues of the reporting phase.

As for the reporting of *intraoperative complications*, our Delphi study provided a comprehensive definition of intraoperative complications as any deviation from the ideal intraoperative course occurring between skin incision and skin closure, given the indication for surgery and the interventions conform to current guidelines. The classification comprises four severity grades. The pilot study demonstrated practicability as well as good interrater agreement.

Concerning the handling of *multiple VR outcome metrics* on different scales, we

present an algorithm for combining these outcomes into suitable summary measures of the same dimension and for integrating the dimension summary measures into a meaningful total score.

9.2. Relation to other studies and possible mechanisms

Studies investigating current practice

Our findings concerning the use of *systematic reviews* were in line with previous findings (Clarke & Chalmers 1998) (Clarke, Alderson, & Chalmers 2002) (Clarke, Hopewell, & Chalmers 2007) (Clarke, Hopewell, & Chalmers 2010) (Clarke & Hopewell 2013), although even more marked. However, the findings need to be interpreted in light of the possibility that trialists used systematic reviews to inform trial design, but did not explicitly state so.

We found slow recruitment to be the most frequent cause for *early trial discontinuation*. This is in line with previous findings. In a cohort of 114 publicly funded multi-center trials, only 31% reached their initial recruitment target, an additional 24% reached 80% of their recruitment target and 53% of the trials were extended (McDonald, Knight, Campbell, Entwistle, Grant, Cook, Elbourne, Francis, Garcia, Roberts, & Snowdon 2006). Potential explanations for more surgical than medical trials being discontinued may be surgery-specific aspects and challenges with clinical trials such as requirements concerning surgeon experience, standardization of the intervention, and recruitment (e.g. due to patient or caregiver preference) (Boutron, Moher, Altman, Schulz, & Ravaut 2008; Devereaux, Bhandari, Clarke, Montori, Cook, Yusuf, Sackett, Cina, Walter, Haynes, Schunemann, Norman, & Guyatt 2005; Solomon & McLeod 1995). Additionally, there are some fundamental differences in the premarket review and approval process of devices versus drugs (Sweet, Schwemm, & Parsons 2011), (Sorenson & Drummond 2014). Since depending on the risk class of surgical devices, regulatory approval may not be based on clinical evidence, potentially less resources and other efforts to achieve trial completion may be invested in such instances. The proportions of published trials were slightly higher compared to a previous investigation of study protocols submitted to a Swiss research ethics committee (Von Elm, Rollin, Blumle, Huwiler, Witschi, & Egger 2008). In this previous study, the odds for publication was higher with non-commercial funding, which is in line with our results, higher with multicenter trials, international collaboration and a high sample size as assessed by median split ((Von Elm, Rollin, Blumle, Huwiler, Witschi, & Egger 2008). This investigation comprised, however, mostly medical specialties and thus is not strictly comparable to the surgical trials evaluated in our sample.

We found trial discontinuation to be an independent risk factor for non-publication. This may be explained by the fact that the most frequent reasons for discontinuation of surgical RCTs were slow recruitment, futility or administrative reasons. These trials were thus potentially more difficult to publish than the few trials stopped early for benefit. Generally, in trials stopped early for benefit,

treatment effects tend to be large (Bassler, Montori, Briel, Glasziou, & Guyatt 2008; Montori, Devereaux, Adhikari, Burns, Eggert, Briel, Lacchetti, Leung, Darling, Bryant, Bucher, Schunemann, Meade, Cook, Erwin, Sood, Sood, Lo, Thompson, Zhou, Mills, & Guyatt 2005a), increasing the chance of a trial to be published. This is supported by a review of such trials in which the majority of RCTs (92/143) were published in 5 high-impact medical journals. (Montori, Devereaux, Adhikari, Burns, Eggert, Briel, Lacchetti, Leung, Darling, Bryant, Bucher, Schunemann, Meade, Cook, Erwin, Sood, Sood, Lo, Thompson, Zhou, Mills, & Guyatt 2005a).

To our knowledge, our review of *intraoperative complication* reporting is the first review assessing the quality of reporting adverse events in surgery with a specific focus on intraoperative complications. However, several investigations of reporting of postoperative surgical complications found a lack in standardized definitions (Bruce, Russell, Mollison, & Krukowski 2001b).

We are unaware of any systematic analysis of the appropriateness of methods used for handling *paired data* in surgery and of the effect of bias from disregarding clustered data. Should our preliminary findings of a lack of systematic use of appropriate methods to analyze paired data be confirmed, this would be in line with previous findings in other fields. In a systematic review of oral health reports, the clustered nature of data had been accounted for in sample size calculations in 65% of the analyzed studies (Froud, Eldridge, Diaz, Marinho, & Donner 2012). As for data analysis, in a systematic review of implant dentistry, clustered data were not accounted for in 33% of the analyzed RCTs (Cairo, Sanz, Matesanz, Nieri, & Pagliaro 2012). Importantly, in a review of 69 ophthalmic RCTs, none of them accounted for clustering in sample size calculation and only 5 studies accounted for non-independence in the statistical analysis (Durkalski, Palesch, Lipsitz, & Rust 2003; Lee, Cheng, & Fong 2012).

Studies addressing potential solutions

While there are several classification systems for postoperative complications, we are unaware of any validated definition and *classification system strictly applying to intraoperative complications*. Fabri et al. developed a classification system for errors in operative therapy in an iterative process, interacting with faculty members of one major surgical academic university department (Fabri & Zayas-Castro 2008). Agreement was found to be over 70% in all error categories and thus similar to our findings, which are however not strictly comparable, since our classification is limited to events between skin incision and closure. The classification for postoperative complications proposed by Clavien and Dindo (Dindo, Demartines, & Clavien 2004) showed - in an international survey taking advantage of clinical cases - 90% of the answers to the scenarios to be correct, independent of the level of training or origin of the surgeon.

As for *VR outcome reporting*, we are unaware of any attempts to address the multiplicity issue when reporting simulator outcomes. Some intelligence tests

however are taking advantage of a norming sample from the general population with a mean of 100 and a standard deviation of 15 (Streiner & Norman 2008).

9.3. Strengths and limitations

Studies investigating current practice

To the best of our knowledge, our study represents the first investigation of the use of *systematic reviews* to inform trial design and for evidence synthesis specifically addressing surgical trials. A potential limitation of the investigation is the fact that we did not cross-check any claims of the study reports, such as a claim to be the first RCT addressing a specific research question. Additionally, we refrained from contacting the authors for additional information.

The strength of our investigation concerning *early trial discontinuation* is the fact that the collaborating RECs granted us unrestricted access to trial protocols, thus reducing the risk of a selection bias. Additionally, our approach allowed us to collect more detailed information about trial characteristics than investigating trial registry entries only (World Health Organization 2013). Moreover, we obtained study protocols from several RECs in three countries, increasing the generalizability of our findings (Kasenda, von Elm, You, Blumle, Tomonaga, Saccilotto, Amstutz, Bengough, Meerpohl, Stegert, Tikkinen, Neumann, Carrasco-Labra, Faulhaber, Mulla, Mertz, Akl, Bassler, Busse, Ferreira-Gonzalez, Lamontagne, Nordmann, Rosenthal, Schandelmaier, Sun, Vandvik, Johnston, Walter, Burnand, Schwenkglenks, Bucher, Guyatt, & Briel 2012). Additionally, our data extractions were based on a priori definitions and validated through regular quality assessments. Finally, our sensitivity analyses support the robustness of the conducted statistical analyses. As limitation, the relatively low number of surgical trials and outcome events did not allow us to adjust for additional confounders (Peduzzi, Concato, Kemper, Holford, & Feinstein 1996).

Our investigation concerning the current practice of *reporting intra- and postoperative complications* is limited by the fact that all RCTs published in three major surgical journals were considered for review. Therefore, some of the interventions were not involving surgery only, but as well invasive or perioperative interventions. We addressed this issue by evaluating in an exploratory analysis the impact of studies including surgery as primary study intervention on complication reporting.

Additionally, the results from three journals may not be extrapolated to all surgical journals. We focused on three high-impact journals with endorsed CONSORT reporting guidelines policy and it is therefore unlikely that an investigation in other surgical journals would yield lower frequencies of lacking definition and classification of intra- and postoperative adverse.

Moreover, reviewers were not blinded regarding authors and journals of the published RCTs. However, data abstraction was undertaken by two independent

reviewers who used a standardized objective extraction protocol. We consider the lack of blinded data abstraction as a minor risk of bias (Morissette, Tricco, Horsley, Chen, & Moher 2011).

The evaluation of the use of appropriate methods to account for *paired data* was limited by the fact that many trial reports were lacking clear information in the methods section concerning trial design and analysis, such as concerning the included trial population, the primary outcome, the sample size calculation, and the unit of analysis. We therefore had to infer some of the information from the results section. To enhance reliability of our extractions, all data were extracted in double and disagreements were resolved by discussion. We additionally recorded whether relevant issues were explicitly stated or whether they had to be derived from other information.

Studies addressing potential solutions

Since a prerequisite for adequate reporting in surgical research is a thoroughly planned and sound research protocol, we developed a comprehensive, but hands-on guide to *develop a clinical surgical research protocol*. We illustrated methodological issues with practical examples and provided reference to cardinal trials in surgery. Given the nature of a journal article, not all topics could be addressed in full detail. We thus provided references for further reading.

To the best of our knowledge, there is no validated *definition and classification specifically addressing intraoperative complications*. We were able to launch an extensive discussion about the definition and classification of intraoperative complications, involving an international and interdisciplinary expert team with a high return-rate in both Delphi rounds. It became evident that the ideal choices were not uniformly the most practical; hence some decisions on definitions and grading were rather pragmatic, resulting in a comprehensive definition and simple classification. We are, however, well aware that here we present results of the Delphi process and pilot study, which may be subject to further modification and refinement upon full-scale validation. Moreover, the retrospective setting of the pilot study with its small sample size is unlikely to cover all possible complications. Additionally, we found a lower spread of complications across grades for intra- than for postoperative complications, thus our similar interrater agreement for intra- and postoperative complications needs to be interpreted with care.

We are unaware of any guideline on how to analyse and report multiple *VR outcome metrics*. We here proposed an algorithm to combine the evidence from multiple outcome metrics into a lower dimensional outcome. This concept may be applied in proficiency-based learning, where a single new observation from another trainee or several new observations from one or more groups of trainees may be compared to a reference group, for instance a group of experts. Benchmark values may be established once and then serve as a reference for

subsequent observations. To generate however such a representative reference population, the sample size of the reference population by far needs to exceed the one presented in this paper. Moreover, it needs to be taken into account that the inclusion of weights was for illustrative purposes only; the chosen weights need to be interpreted as examples.

9.4. Implications for current practice

Implications from studies investigating current practice

In practice, the following four-step framework was suggested for the use of *systematic reviews* to inform the design of new trials: In step one, the research question of the proposed trial is formulated focussing on definition of the populations, interventions, comparators, outcomes, timing and setting (PICOTS) (Thompson, Tiwari, Fu, Moe, & Buckley 2013). In step two, an up-to-date relevant and valid systematic review is identified or conducted, and in step three the systematic review is used to inform the planned trial. If there are several RCTs, but no systematic review in the field, we propose to first synthesize the body of evidence within a systematic review and if appropriate conduct a meta-analysis. Finally, in step four the implications for the proposed trial are summarised. This procedure is important to justify any research from a scientific, ethical and economic point of view. Similarly, we propose that trialists attempt to report their results in the context of other trials in the field, whenever possible integrating the new findings by updating a systematic review. Only when evaluating study results within the overall body of evidence, well-informed decisions in health care are possible.

The findings concerning our evaluation of *early trial discontinuation* may have the following implications during the different phases of a research pathway:

i) Trial conception and design: Strategies to prevent discontinuation should be adopted already during the trial design phase, including training in research methodology, realistic resource estimation (e.g. time of personnel), creating research infrastructures, and working in networks and interdisciplinary teams involving trial methodologists, statisticians, data managers and trial managers. Pilot studies either as part of the trial (internal pilot) or stand-alone (external pilot) (Lancaster, Dodd, & Williamson 2004), i.e. a small version of the full-scale study or feasibility studies, are effective means to evaluate particular aspects such as recruitment, resource utilization and protocol feasibility (Dolgin 2013). This especially applies to trials involving surgical interventions, where recruitment may be more challenging due to patient preferences. Moreover, pilot studies may identify other challenges such as feasibility of standardization of the surgical interventions. Pilot studies should thus be considered when allocating sparse resources to surgical RCTs. Additionally, trials should be registered prior to enrolment of the first patient. This enhances transparency regarding the targeted

sample size and additionally allows identifying the study while on-going or after close-up before publication.

- ii) Trial conduct: Recruitment should be closely monitored and early modification foreseen to enhance recruitment if necessary. The decision to discontinue a trial should be made by an independent Data and Safety Monitoring Board based on pre-defined criteria.
- iii) Trial reporting: Reasons for discontinuation and results of discontinued RCTs should be transparently communicated and thus made publicly available.
- iv) Healthcare decision making: When interpreting surgical literature, be it primary studies or systematic reviews, potential biases introduced by discontinued RCTs or by non-publication of trials need to be considered.

Since there is a lack of standardized *reporting of intra- and postoperative complications*, comparisons between adverse events in surgical practice, clinical research, and training settings need to be interpreted with care.

In case our preliminary findings of a lack of systematic use of appropriate methods to analyze *paired data* hold true, published results from trials with a paired data structure need to be critically evaluated for appropriate methodology and results interpreted accordingly.

Implications from studies addressing potential solutions

The *study protocol* is the core document when planning and conducting clinical research and thus has a major impact on the study report. It should be created in an interdisciplinary setting, approved and strictly followed. Any changes require an amendment approved by an IEC and the regulatory authorities. To enhance transparency, trial protocols may be made available, be it as publication in a peer-reviewed journal or be it to journal editors and reviewers, as it is already requested by some journals.

We propose to prospectively collect any *intraoperative complications* with a validated instrument. In our opinion, standardization of definition and classification of intraoperative complications is important in a number of settings: first, in research and development, when reporting assessments of new surgical techniques and devices; second, in educational and training settings (individual surgeon's performance) and in institution benchmarking (institution performance); third for risk management, capturing intraoperative events with an increased risk of postoperative complications within a critical incidents reporting system (Kram 2008;Leape 1994;Reason 2000) or in clinical research where postoperative event rates may be too low to capture potential safety concerns; and last for patient management, enhancing quality of patient handover after surgery concerning anticipated problems (Manser, Foster, Flin, & Patey 2013;Nagpal, Abboudi, Fischler, Schmidt, Vats, Manchanda, Sevdalis, Scheidegger, Vincent, & Moorthy 2011).

When analyzing and reporting *virtual reality outcome* data, the use of standardized, statistically sound and transparent methods is of utmost importance. This will allow for comparison across studies. When interpreting VR outcome metrics from reports, the possibility of selective reporting or multiple testing needs to be critically evaluated.

9.5. Implications for future research

Implications from studies investigating current practice

Concerning the use of *systematic reviews* to inform trial design and to synthesize evidence, future investigations could address the impact of evidence synthesis on successful grant acquisition and regulatory approval in the design phase, on successful and timely study conduct including meeting the assumptions made for sample size calculation, on timely and high-ranked study publication and on treatment coverage (payers' perspective). Additionally, it could be explored, to what extent network meta-analysis is used for comparison of interventions that may or may not have been directly compared against each other (Mills et al. 2013).

To further support our key message from the investigation of *early trial discontinuation* - which may be summarized as 'to conduct pilot studies' - it could be evaluated in the investigated set of trials by contacting the investigators i) to what extent pilot studies were used, ii) for which specific purposes (e.g. evaluation of consent rates, feasibility and standardization of the intervention, generation of preliminary data for sample size calculation), and iii) whether there was an association between conducting a pilot study and successful trial completion and publication.

Our findings of inconsistent reporting of *intraoperative complications* prompted us to conduct a Delphi study with pilot evaluation of a classification of intraoperative complications.

To estimate the effect of disregarding *paired data* on the overall results, we will attempt to obtain original individual patient data of published RCTs with different proportions of paired data from the respective authors. The original data of these trials will be analyzed twice, once with appropriate statistical methods accounting for clustered data and once disregarding the presence of clustered data in order to estimate differences in overall conclusions depending on the type of analysis. Should no original data be available, a simulation study will be conducted to estimate the impact of disregarding paired data on overall trial conclusions.

Implications from studies addressing potential solutions

We believe that a thoroughly planned and well-designed study with a *comprehensive study protocol* is at lower risk for modification throughout the study conduct and reporting phase. Yet, should amendments to the protocol be

necessary, they should transparently be communicated. We have previously investigated discrepancies between trial registry entries and final reports and found a relevant percentages of discrepancies, for instance concerning outcomes (Rosenthal & Dwan 2013). As further research we suggest evaluating discrepancies between clinical trial protocols and final reports. This could be undertaken in the trial set investigated for early discontinuation. Although previous studies have investigated such discrepancies (Dwan et al. 2011), to the best of our knowledge this has never been evaluated in surgical trials.

As further research, we plan to refine and validate the *definition and classification of intraoperative complications* in a multinational, prospective cohort study. We will assess, whether there is an association between the classification and the complexity of surgery, as well as between the classification and the duration of surgery. Furthermore, an explorative analysis shall evaluate, whether the classification is predictive for a number of adverse postoperative outcomes, adjusted for the most relevant confounders.

Concerning our proposed algorithm for combining the evidence *from multiple outcome metrics into a lower dimensional outcome*, we propose to evaluate as further research project test-retest reliability of the mean summary measures and the total score on the same simulator, followed by test-retest reliability on different simulators. The ultimate goal of this work and of future research is to enhance the validity of VR-related research reports.

9.6. Conclusions

In conclusion, we investigated reporting in surgical research by addressing current practice and by developing solutions to identified challenges at different phases of a surgical research project. Our findings are relevant not only in clinical research, but also in clinical practice and educational settings.

Specifically, we addressed standardization and therewith transparency by developing a classification for intraoperative adverse events, by proposing an algorithm for summarizing multiple outcomes, and by investigating the impact of not taking into account paired data. For optimal reporting, we strongly advocate meticulous planning of surgical research and thus present a guide for writing a well-designed study protocol. This allows for justification of the planned research from a scientific, ethical and economic perspective. We recommend setting the planned research in the context of the overall body of evidence using systematic reviews. To lower the risk of early trial discontinuation and therewith as found by our investigation of non-publication of trial results, we suggest conducting pilot studies should there be any doubts concerning trial feasibility. Ultimately, well designed and transparently reported surgical research allows for well-informed healthcare decision making.

10. References

Ahmad, G., O'Flynn, H., Duffy, J.M., Phillips, K., & Watson, A. 2012. Laparoscopic entry techniques. *Cochrane.Database.Syst.Rev.*, 2, CD006583 available from: PM:22336819

Akobeng, A.K. 2008. Assessing the validity of clinical trials. *J.Pediatr.Gastroenterol.Nutr.*, 47, (3) 277-282 available from: PM:18728521

Alderson, P. 2004. Absence of evidence is not evidence of absence. *BMJ*, 328, (7438) 476-477 available from: PM:14988165

Altman, D.G. & Bland, J.M. 1999. Statistics notes. Treatment allocation in controlled trials: why randomise? *BMJ*, 318, (7192) 1209 available from: PM:10221955

Altman, D.G., Schulz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gotzsche, P.C., & Lang, T. 2001. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann.Intern.Med.*, 134, (8) 663-694 available from: PM:11304107

Assmann, S.F., Pocock, S.J., Enos, L.E., & Kasten, L.E. 2000. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*, 355, (9209) 1064-1069 available from: PM:10744093

Babyak, M.A. 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom.Med.*, 66, (3) 411-421 available from: PM:15184705

Bassi, C., Dervenis, C., Butturini, G., Fingerhut, A., Yeo, C., Izbicki, J., Neoptolemos, J., Sarr, M., Traverso, W., & Buchler, M. 2005. Postoperative pancreatic fistula: an international study group (ISGPF) definition. *Surgery*, 138, (1) 8-13 available from: PM:16003309

Bassler, D., Briel, M., Montori, V.M., Lane, M., Glasziou, P., Zhou, Q., Heels-Ansdell, D., Walter, S.D., Guyatt, G.H., Flynn, D.N., Elamin, M.B., Murad, M.H., Abu Elnour, N.O., Lampropoulos, J.F., Sood, A., Mullan, R.J., Erwin, P.J., Bankhead, C.R., Perera, R., Ruiz, C.C., You, J.J., Mulla, S.M., Kaur, J., Nerenberg, K.A., Schunemann, H., Cook, D.J., Lutz, K., Ribic, C.M., Vale, N., Malaga, G., Akl, E.A., Ferreira-Gonzalez, I., Alonso-Coello, P., Urrutia, G., Kunz, R., Bucher, H.C., Nordmann, A.J., Raatz, H., da Silva, S.A., Tuche, F., Strahm, B., Djulbegovic, B., Adhikari, N.K., Mills, E.J., Gwadrý-Sridhar, F., Kirpalani, H., Soares, H.P., Karanicolos, P.J., Burns, K.E., Vandvik, P.O., Coto-Yglesias, F., Chrispim, P.P., & Ramsay, T. 2010. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA*, 303, (12) 1180-1187 available from: PM:20332404

Bassler, D., Montori, V.M., Briel, M., Glasziou, P., & Guyatt, G. 2008. Early stopping of randomized clinical trials for overt efficacy is problematic. *J.Clin.Epidemiol.*, 61, (3) 241-246 available from: PM:18226746

Bates, D., Maechler, M., Bolker, B., & Walker, S. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-5. Available at: <http://CRAN.R-project.org/package=lme4>. Accessed December 09, 2013.

Bonenkamp, J.J., Hermans, J., Sasako, M., van de Velde, C.J., Welvaart, K., Songun, I., Meyer, S., Plukker, J.T., Van, E.P., Obertop, H., Gouma, D.J., van Lanschoot, J.J., Taat, C.W., de Graaf, P.W., von Meyenfeldt, M.F., & Tilanus, H. 1999. Extended lymph-node dissection for gastric cancer. *N.Engl.J.Med.*, 340, (12) 908-914 available from: PM:10089184

Bonrath, E.M., Dedy, N.J., Zevin, B., & Grantcharov, T.P. 2013a. Defining technical errors in laparoscopic surgery: a systematic review. *Surg.Endosc.*, 27, (8) 2678-2691 available from: PM:23436086

Bonrath, E.M., Zevin, B., Dedy, N.J., & Grantcharov, T.P. 2013b. Error rating tool to identify and analyse technical errors and events in laparoscopic surgery. *Br.J.Surg.*, 100, (8) 1080-1088 available from: PM:23754649

Bornhoft, G., Masion-Bergemann, S., Wolf, U., Kienle, G.S., Michalsen, A., Vollmar, H.C., Gilbertson, S., & Matthiessen, P.F. 2006. Checklist for the qualitative evaluation of clinical studies with particular focus on external validity and model validity. *BMC.Med.Res.Methodol.*, 6, 56 available from: PM:17156475

Boutron, I., Moher, D., Altman, D.G., Schulz, K.F., & Ravaud, P. 2008. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. *Ann.Intern.Med.*, 148, (4) 295-309 available from: PM:18283207

Brody, B.A., McCullough, L.B., & Sharp, R.R. 2005. Consensus and controversy in clinical research ethics. *JAMA*, 294, (11) 1411-1414 available from: PM:16174704

Bruce, J., Krukowski, Z.H., Al-Khairy, G., Russell, E.M., & Park, K.G. 2001a. Systematic review of the definition and measurement of anastomotic leak after gastrointestinal surgery. *Br.J.Surg.*, 88, (9) 1157-1168 available from: PM:11531861

Bruce, J., Russell, E.M., Mollison, J., & Krukowski, Z.H. 2001b. The measurement and monitoring of surgical adverse events. *Health Technol.Assess.*, 5, (22) 1-194 available from: PM:11532239

Bruce, J., Russell, E.M., Mollison, J., & Krukowski, Z.H. 2001c. The quality of measurement of surgical wound infection as the basis for monitoring: a systematic review. *J.Hosp.Infect.*, 49, (2) 99-108 available from: PM:11567554

Bucher, H.C., Guyatt, G.H., Cook, D.J., Holbrook, A., & McAlister, F.A. 1999. Users' guides to the medical literature: XIX. Applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. Evidence-Based Medicine Working Group. *JAMA*, 282, (8) 771-778 available from: PM:10463714

Buyse, M. 2009. Use of meta-analysis for the validation of surrogate endpoints and biomarkers in cancer trials. *Cancer J.*, 15, (5) 421-425 available from: PM:19826362

Cairo, F., Sanz, I., Matesanz, P., Nieri, M., & Pagliaro, U. 2012. Quality of reporting of randomized clinical trials in implant dentistry. A systematic review on critical aspects in design, outcome assessment and clinical relevance. *J.Clin.Periodontol.*, 39 Suppl 12, 81-107 available from: PM:22533949

Campbell, M.K., Elbourne, D.R., & Altman, D.G. 2004. CONSORT statement: extension to cluster randomised trials. *BMJ*, 328, (7441) 702-708 available from: PM:15031246

Campbell, M.K., Piaggio, G., Elbourne, D.R., & Altman, D.G. 2012. Consort 2010 statement: extension to cluster randomised trials. *BMJ*, 345, e5661 available from: PM:22951546

Catchpole, K. 2010. Errors in the operating theatre--how to spot and stop them. *J.Health Serv.Res.Policy*, 15 Suppl 1, 48-51 available from: PM:20075129

Champagne, B.J., Papaconstantinou, H.T., Parmar, S.S., Nagle, D.A., Young-Fadok, T.M., Lee, E.C., & Delaney, C.P. 2012. Single-incision versus standard multiport laparoscopic colectomy: a multicenter, case-controlled comparison. *Ann Surg*, 255, (1) 66-69 available from: PM:22104563

Chan, A.W., Hrobjartsson, A., Haahr, M.T., Gotzsche, P.C., & Altman, D.G. 2004. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA*, 291, (20) 2457-2465 available from: PM:15161896

Chan, A.W., Tetzlaff, J.M., Altman, D.G., Laupacis, A., Gotzsche, P.C., Krleza-Jeric, K., Hrobjartsson, A., Mann, H., Dickersin, K., Berlin, J.A., Dore, C.J., Parulekar, W.R., Summerskill, W.S., Groves, T., Schulz, K.F., Sox, H.C., Rockhold, F.W., Rennie, D., & Moher, D. 2013a. SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Ann Intern.Med.*, 158, (3) 200-207 available from: PM:23295957

Chan, A.W., Tetzlaff, J.M., Gotzsche, P.C., Altman, D.G., Mann, H., Berlin, J.A., Dickersin, K., Hrobjartsson, A., Schulz, K.F., Parulekar, W.R., Krleza-Jeric, K., Laupacis, A., & Moher, D. 2013b. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ*, 346, e7586 available from: PM:23303884

Chan, A.W., Upshur, R., Singh, J.A., Ghersi, D., Chapuis, F., & Altman, D.G. 2006. Research protocols: waiving confidentiality for the greater good. *BMJ*, 332, (7549) 1086-1089 available from: PM:16675819

Clarke, M., Alderson, P., & Chalmers, I. 2002. Discussion sections in reports of controlled trials published in general medical journals. *JAMA*, 287, (21) 2799-2801 available from: PM:12038916

Clarke, M. & Chalmers, I. 1998. Discussion sections in reports of controlled trials published in general medical journals: islands in search of continents? *JAMA*, 280, (3) 280-282 available from: PM:9676682

Clarke, M. & Hopewell, S. 2013. Many reports of randomised trials still don't begin or end with a systematic review of the relevant evidence. *J Bahrain Med Soc* 24[3], 145-148.

Clarke, M., Hopewell, S., & Chalmers, I. 2007. Reports of clinical trials should begin and end with up-to-date systematic reviews of other relevant evidence: a status report. *J.R.Soc.Med.*, 100, (4) 187-190 available from: PM:17404342

Clarke, M., Hopewell, S., & Chalmers, I. 2010. Clinical trials should begin and end with systematic reviews of relevant evidence: 12 years and waiting. *Lancet*, 376, (9734) 20-21 available from: PM:20609983

Clavien, P.A. 2013. Targeting quality in surgery. *Ann.Surg.*, 258, (5) 659-668 available from: PM:24121256

Clavien, P.A., Barkun, J., de Oliveira, M.L., Vauthey, J.N., Dindo, D., Schulick, R.D., de, S.E., Pekolj, J., Slankamenac, K., Bassi, C., Graf, R., Vonlanthen, R., Padbury, R., Cameron, J.L., & Makuuchi, M. 2009. The Clavien-Dindo classification of surgical complications: five-year experience. *Ann.Surg.*, 250, (2) 187-196 available from: PM:19638912

Clavien, P.A., Sanabria, J.R., & Strasberg, S.M. 1992. Proposed classification of complications of surgery with examples of utility in cholecystectomy. *Surgery*, 111, (5) 518-526 available from: PM:1598671

Clinical Outcomes of Surgical Therapy Study Group. 2004. A comparison of laparoscopically assisted and open colectomy for colon cancer. *N.Engl.J.Med.*, 350, (20) 2050-2059 available from: PM:15141043

Collings, B.J. & Hamilton, M.A. 1988. Estimating the power of the two-sample Wilcoxon test for location shift. *Biometrics*, 44, (3) 847-860 available from: PM:3203133

Cummings, S. R., Browner, W. S., & Hulley, S. B. Conceiving the research question. In Hulley SB, Cummings SR, Browner WS et al, eds. *Designing Clinical Research*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2007. p.17-26.

Cunningham, S.C. & Kavic, S.M. 2009. What is a surgical complication? *World J.Surg.*, 33, (5) 1099-1100 available from: PM:19130118

D'Agostino, R.B., Jr. 1998. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat.Med.*, 17, (19) 2265-2281 available from: PM:9802183

Day, S.J. & Altman, D.G. 2000. Statistics notes: blinding in clinical trials and other studies. *BMJ*, 321, (7259) 504 available from: PM:10948038

De, A.C., Drazen, J.M., Frizelle, F.A., Haug, C., Hoey, J., Horton, R., Kotzin, S., Laine, C., Marusic, A., Overbeke, A.J., Schroeder, T.V., Sox, H.C., & Van Der Weyden, M.B. 2004. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N.Engl.J.Med.*, 351, (12) 1250-1251 available from: PM:15356289

- Delgado-Rodriguez, M. & Llorca, J. 2004. Bias. *J.Epidemiol.Community Health*, 58, (8) 635-641 available from: PM:15252064
- Devereaux, P.J., Bhandari, M., Clarke, M., Montori, V.M., Cook, D.J., Yusuf, S., Sackett, D.L., Cina, C.S., Walter, S.D., Haynes, B., Schunemann, H.J., Norman, G.R., & Guyatt, G.H. 2005. Need for expertise based randomised controlled trials. *BMJ*, 330, (7482) 88 available from: PM:15637373
- Dindo, D. & Clavien, P.A. 2008. What is a surgical complication? *World J.Surg.*, 32, (6) 939-941 available from: PM:18414942
- Dindo, D., Demartines, N., & Clavien, P.A. 2004. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Ann.Surg.*, 240, (2) 205-213 available from: PM:15273542
- Dolgin, E. 2013. Publication checklist proposed to boost rigor of pilot trials. *Nat.Med.*, 19, (7) 795-796 available from: PM:23836203
- Durkalski, V.L., Palesch, Y.Y., Lipsitz, S.R., & Rust, P.F. 2003. Analysis of clustered matched-pair data. *Stat.Med.*, 22, (15) 2417-2428 available from: PM:12872299
- Dwan, K., Altman, D.G., Cresswell, L., Blundell, M., Gamble, C.L., & Williamson, P.R. 2011. Comparison of protocols and registry entries to published reports for randomised controlled trials. *Cochrane.Database.Syst.Rev.* (1) MR000031 available from: PM:21249714
- Emami, C.N., Garrett, D., Anselmo, D., Torres, M., & Nguyen, N.X. 2011. Single-incision laparoscopic cholecystectomy in children: a feasible alternative to the standard laparoscopic approach. *J.Pediatr.Surg*, 46, (10) 1909-1912 available from: PM:22008326
- Fabri, P.J. & Zayas-Castro, J.L. 2008. Human error, not communication and systems, underlies surgical complications. *Surgery*, 144, (4) 557-563 available from: PM:18847639
- Feinstein, A. H. & Cannon, H. M. Fidelity, verifiability, and validity of simulation: constructs for evaluation. Wayne State University Marketing Department Working Paper 2001-006. Available at: sbaweb.wayne.edu/~marketing/wp/008HC.pdf (Accessed November 26, 2013).
- Finkemeier, C.G., Schmidt, A.H., Kyle, R.F., Templeman, D.C., & Varecka, T.F. 2000. A prospective, randomized study of intramedullary nails inserted with and without reaming for the treatment of open and closed fractures of the tibial shaft. *J.Orthop.Trauma*, 14, (3) 187-193 available from: PM:10791670
- Fleiss, J.L. & Cohen, J. 1973. Equivalence of Weighted Kappa and Intraclass Correlation Coefficient As Measures of Reliability. *Educational and Psychological Measurement*, 33, (3) 613-619 available from: ISI:A1973Q863000009
- Fleming, T.R. & DeMets, D.L. 1996. Surrogate end points in clinical trials: are we being misled? *Ann Intern.Med.*, 125, (7) 605-613 available from: PM:8815760

- Fox, K.A., Mehta, S.R., Peters, R., Zhao, F., Lakkis, N., Gersh, B.J., & Yusuf, S. 2004. Benefits and risks of the combination of clopidogrel and aspirin in patients undergoing surgical revascularization for non-ST-elevation acute coronary syndrome: the Clopidogrel in Unstable angina to prevent Recurrent ischemic Events (CURE) Trial. *Circulation*, 110, (10) 1202-1208 available from: PM:15313956
- Froud, R., Eldridge, S., Diaz, O.K., Marinho, V.C., & Donner, A. 2012. Quality of cluster randomized controlled trials in oral health: a systematic review of reports published between 2005 and 2009. *Community Dent.Oral Epidemiol.*, 40 Suppl 1, 3-14 available from: PM:22369703
- Fueglistaler, P., Adamina, M., & Guller, U. 2007. Non-inferiority trials in surgical oncology. *Ann Surg Oncol.*, 14, (5) 1532-1539 available from: PM:17294075
- Gallagher, A.G., Ritter, E.M., & Satava, R.M. 2003. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. *Surg.Endosc.*, 17, (10) 1525-1529 available from: PM:14502403
- Godwin, M., Ruhland, L., Casson, I., MacDonald, S., Delva, D., Birtwhistle, R., Lam, M., & Seguin, R. 2003. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. *BMC.Med.Res.Methodol.*, 3, 28 available from: PM:14690550
- Goldsmith, C.H., Smythe, H.A., & Helewa, A. 1993. Interpretation and power of a pooled index. *J.Rheumatol.*, 20, (3) 575-578 available from: PM:8478877
- Goodman, S.N. 2009. Stopping trials for efficacy: an almost unbiased view. *Clin. Trials*, 6, (2) 133-135 available from: PM:19342465
- Graham, I.D. 2012. Knowledge synthesis and the Canadian Institutes of Health Research. *Syst.Rev.*, 1, 6 available from: PM:22587985
- Guralnik, J. M. M. T. A. Design and Conduct of Observational Studies and Clinical Trials. In: Gallin JI, Ognibene FP, eds. *Principles and practice of clinical research*. Oxford: Elsevier, 2007. p 197-217.
- Hanley, J.A., Negassa, A., Edwardes, M.D., & Forrester, J.E. 2003. Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am.J.Epidemiol.*, 157, (4) 364-375 available from: PM:12578807
- Haynes, R. B. Forming research questions. In: Haynes RB, Sackett DL, Guyatt GH, Tugwell P. *Clinical epidemiology: How to do clinical practice research*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2006. p 3-14.
- Heikens, J.T., Gooszen, H.G., Rovers, M.M., & van Laarhoven, C.J. 2013. Stages and evaluation of surgical innovation: a clinical example of the ileo neorectal anastomosis after ulcerative colitis and familial adenomatous polyposis. *Surg Innov.*, 20, (5) 459-465 available from: PM:23242517
- Hernan, M.A. & Robins, J.M. 2006. Estimating causal effects from epidemiological data. *J.Epidemiol.Community Health*, 60, (7) 578-586 available from: PM:16790829

Horan, T.C., Gaynes, R.P., Martone, W.J., Jarvis, W.R., & Emori, T.G. 1992. CDC definitions of nosocomial surgical site infections, 1992: a modification of CDC definitions of surgical wound infections. *Infect.Control Hosp.Epidemiol.*, 13, (10) 606-608 available from: PM:1334988

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *J Educational Psychol*, 24, 417-441

Hsu, C.-C. & Sandford, B. A. The Delphi Technique: Making Sense of Consensus. *Practical Assessment, Research & Evaluation* 12[10], 1-8. 2013.

Ref Type: Generic

Iltis, A.S. 2005. Stopping trials early for commercial reasons: the risk-benefit relationship as a moral compass. *J.Med.Ethics*, 31, (7) 410-414 available from: PM:15994362

Ioannidis, J.P. & Contopoulos-Ioannidis, D.G. 1998. Reporting of safety data from randomised trials. *Lancet*, 352, (9142) 1752-1753 available from: PM:9848355

Ioannidis, J.P., Evans, S.J., Gotzsche, P.C., O'Neill, R.T., Altman, D.G., Schulz, K., & Moher, D. 2004. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann.Intern.Med.*, 141, (10) 781-788 available from: PM:15545678

Ioannidis, J.P. & Lau, J. 2001. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *JAMA*, 285, (4) 437-443 available from: PM:11242428

Jones, A.P., Conroy, E., Williamson, P.R., Clarke, M., & Gamble, C. 2013. The use of systematic reviews in the planning, design and conduct of randomised trials: a retrospective cohort of NIHR HTA funded trials. *BMC.Med Res.Methodol.*, 13, 50 available from: PM:23530582

Kaplan, A.V., Baim, D.S., Smith, J.J., Feigal, D.A., Simons, M., Jefferys, D., Fogarty, T.J., Kuntz, R.E., & Leon, M.B. 2004. Medical device development: from prototype to regulatory approval. *Circulation*, 109, (25) 3068-3072 available from: PM:15226221

Kasenda, B., von Elm, E.B., You, J., Blumle, A., Tomonaga, Y., Saccilotto, R., Amstutz, A., Bengough, T., Meerpohl, J., Stegert, M., Tikkinen, K.A., Neumann, I., Carrasco-Labra, A., Faulhaber, M., Mulla, S., Mertz, D., Akl, E.A., Bassler, D., Busse, J.W., Ferreira-Gonzalez, I., Lamontagne, F., Nordmann, A., Rosenthal, R., Schandelmaier, S., Sun, X., Vandvik, P.O., Johnston, B.C., Walter, M.A., Burnand, B., Schwenkglenks, M., Bucher, H.C., Guyatt, G.H., & Briel, M. 2012. Learning from failure--rationale and design for a study about discontinuation of randomized trials (DISCO study). *BMC.Med.Res.Methodol.*, 12, 131 available from: PM:22928744

Kasenda, B., Von, E.E., You, J., Blumle, A., Tomonaga, Y., Saccilotto, R., Amstutz, A., Bengough, T., Meerpohl, J.J., Stegert, M., Tikkinen, K.A., Neumann, I., Carrasco-Labra, A., Faulhaber, M., Mulla, S.M., Mertz, D., Akl, E.A., Bassler, D., Busse, J.W., Ferreira-Gonzalez, I., Lamontagne, F., Nordmann, A., Gloy, V., Raatz, H., Moja, L., Rosenthal, R., Ebrahim, S., Schandelmaier, S., Xin, S., Vandvik, P.O., Johnston, B.C., Walter, M.A., Burnand, B., Schwenkglenks, M., Hemkens, L.G., Bucher, H.C.,

- Guyatt, G.H., & Briel, M. 2014. Prevalence, characteristics, and publication of discontinued randomized trials. *JAMA*, 311, (10) 1045-1051 available from: PM:24618966
- Kenward, M.G. & Carpenter, J. 2007. Multiple imputation: current perspectives. *Stat.Methods Med.Res.*, 16, (3) 199-218 available from: PM:17621468
- Kirkwood, B. R. & Sterne, J. A. C. *Essential Medical Statistics*. Second Edition, Oxford: Blackwell Science Ltd, 2003.
- Klotz, H.P., Candinas, D., Platz, A., Horvath, A., Dindo, D., Schlumpf, R., & Largiader, F. 1996. Preoperative risk assessment in elective general surgery. *Br.J.Surg.*, 83, (12) 1788-1791 available from: PM:9038571
- Kram, R. 2008. Critical incident reporting system in emergency medicine. *Curr.Opin.Anaesthesiol.*, 21, (2) 240-244 available from: PM:18443496
- Lalis G for the Global Harmonization Task Force Study Group 2. Medical Devices Post Market Surveillance: Global Guidance for Adverse Event Reporting for Medical Devices. <http://www.imdrf.org/documents/doc-ghrf-sg2.asp> (Accessed October 23, 2013).
- Lancaster, G.A., Dodd, S., & Williamson, P.R. 2004. Design and analysis of pilot studies: recommendations for good practice. *J.Eval.Clin.Pract.*, 10, (2) 307-312 available from: PM:15189396
- Law, W.L., Fan, J.K., & Poon, J.T. 2010. Single-incision laparoscopic colectomy: early experience. *Dis.Colon Rectum*, 53, (3) 284-288 available from: PM:20173474
- Leape, L.L. 1994. Error in medicine. *JAMA*, 272, (23) 1851-1857 available from: PM:7503827
- Lederle, F.A., Freischlag, J.A., Kyriakides, T.C., Matsumura, J.S., Padberg, F.T., Jr., Kohler, T.R., Kougias, P., Jean-Claude, J.M., Cikrit, D.F., & Swanson, K.M. 2012. Long-term comparison of endovascular and open repair of abdominal aortic aneurysm. *N.Engl.J.Med.*, 367, (21) 1988-1997 available from: PM:23171095
- Lee, C.F., Cheng, A.C., & Fong, D.Y. 2012. Eyes or subjects: are ophthalmic randomized controlled trials properly designed and analyzed? *Ophthalmology*, 119, (4) 869-872 available from: PM:22226885
- Lee, K.J. & Thompson, S.G. 2005. Clustering by health professional in individually randomised trials. *BMJ*, 330, (7483) 142-144 available from: PM:15649931
- Lehmann, E. L. Comparing two treatments or attributes in a population model. In: *Nonparametrics: Statistical Methods Based on Ranks. Revised first edition*. New York: Springer, 2006. p 76-81.
- Lievre, M., Menard, J., Bruckert, E., Cogneau, J., Delahaye, F., Giral, P., Leitersdorf, E., Luc, G., Masana, L., Moulin, P., Passa, P., Pouchain, D., & Siest, G. 2001. Premature discontinuation of clinical trial for reasons not related to efficacy, safety, or feasibility. *BMJ*, 322, (7286) 603-605 available from: PM:11238162

- Lu, Y., Sverden, E., Ljung, R., Soderlund, C., & Lagergren, J. 2013. Use of non-steroidal anti-inflammatory drugs and proton pump inhibitors in correlation with incidence, recurrence and death of peptic ulcer bleeding: an ecological study. *BMJ Open.*, 3, (1) available from: PM:23293249
- Mangram, A.J., Horan, T.C., Pearson, M.L., Silver, L.C., & Jarvis, W.R. 1999a. Guideline for Prevention of Surgical Site Infection, 1999. Centers for Disease Control and Prevention (CDC) Hospital Infection Control Practices Advisory Committee. *Am.J.Infect.Control*, 27, (2) 97-132 available from: PM:10196487
- Mangram, A.J., Horan, T.C., Pearson, M.L., Silver, L.C., & Jarvis, W.R. 1999b. Guideline for prevention of surgical site infection, 1999. Hospital Infection Control Practices Advisory Committee. *Infect.Control Hosp.Epidemiol.*, 20, (4) 250-278 available from: PM:10219875
- Manser, T., Foster, S., Flin, R., & Patey, R. 2013. Team communication during patient handover from the operating room: more than facts and figures. *Hum.Factors*, 55, (1) 138-156 available from: PM:23516799
- Marks, J., Tacchino, R., Roberts, K., Onders, R., DeNoto, G., Paraskeva, P., Rivas, H., Soper, N., Rosemurgy, A., & Shah, S. 2011. Prospective randomized controlled trial of traditional laparoscopic cholecystectomy versus single-incision laparoscopic cholecystectomy: report of preliminary data. *Am.J.Surg*, 201, (3) 369-372 available from: PM:21367381
- Martin, R.C., Brennan, M.F., & Jaques, D.P. 2002a. Quality of complication reporting in the surgical literature. *Ann.Surg.*, 235, (6) 803-813 available from: PM:12035036
- Martin, R.C., Jaques, D.P., Brennan, M.F., & Karpeh, M. 2002b. Achieving RO resection for locally advanced gastric cancer: is it worth the risk of multiorgan resection? *J.Am.Coll.Surg.*, 194, (5) 568-577 available from: PM:12025834
- Mascha, E.J. & Sessler, D.I. 2011. Statistical grand rounds: design and analysis of studies with binary- event composite endpoints: guidelines for anesthesia research. *Anesth.Analg.*, 112, (6) 1461-1471 available from: PM:21543784
- Matilde, S.M. & Chen, X. 2006. Choosing the analysis population in non-inferiority studies: per protocol or intent-to-treat. *Stat.Med.*, 25, (7) 1169-1181 available from: PM:16397861
- McCulloch, P., Altman, D.G., Campbell, W.B., Flum, D.R., Glasziou, P., Marshall, J.C., Nicholl, J., Aronson, J.K., Barkun, J.S., Blazeby, J.M., Boutron, I.C., Campbell, W.B., Clavien, P.A., Cook, J.A., Ergina, P.L., Feldman, L.S., Flum, D.R., Maddern, G.J., Nicholl, J., Reeves, B.C., Seiler, C.M., Strasberg, S.M., Meakins, J.L., Ashby, D., Black, N., Bunker, J., Burton, M., Campbell, M., Chalkidou, K., Chalmers, I., de, L.M., Deeks, J., Ergina, P.L., Grant, A., Gray, M., Greenhalgh, R., Jenicek, M., Kehoe, S., Lilford, R., Littlejohns, P., Loke, Y., Madhock, R., McPherson, K., Meakins, J., Rothwell, P., Summerskill, B., Taggart, D., Tekkis, P., Thompson, M., Treasure, T., Trohler, U., & Vandenbroucke, J. 2009. No surgical innovation without evaluation: the IDEAL recommendations. *Lancet*, 374, (9695) 1105-1112 available from: PM:19782876

McCulloch, P., Taylor, I., Sasako, M., Lovett, B., & Griffin, D. 2002. Randomised trials in surgery: problems and possible solutions. *BMJ*, 324, (7351) 1448-1451 available from: PM:12065273

McDonald, A.M., Knight, R.C., Campbell, M.K., Entwistle, V.A., Grant, A.M., Cook, J.A., Elbourne, D.R., Francis, D., Garcia, J., Roberts, I., & Snowdon, C. 2006. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. *Trials*, 7, 9 available from: PM:16603070

McDonald, P.J., Kulkarni, A.V., Farrokhyar, F., & Bhandari, M. 2010. Ethical issues in surgical research. *Can.J.Surg.*, 53, (2) 133-136 available from: PM:20334746

Mehran, R., Leon, M.B., Feigal, D.A., Jefferys, D., Simons, M., Chronos, N., Fogarty, T.J., Kuntz, R.E., Baim, D.S., & Kaplan, A.V. 2004. Post-market approval surveillance: a call for a more integrated and comprehensive approach. *Circulation*, 109, (25) 3073-3077 available from: PM:15226222

MILLER, G.A. 1956. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol.Rev.*, 63, (2) 81-97 available from: PM:13310704

Mills, E.J., Thorlund, K., & Ioannidis, J.P. 2013. Demystifying trial networks and network meta-analysis. *BMJ*, 346, f2914 available from: PM:23674332

Montori, V.M., Devereaux, P.J., Adhikari, N.K., Burns, K.E., Eggert, C.H., Briel, M., Lacchetti, C., Leung, T.W., Darling, E., Bryant, D.M., Bucher, H.C., Schunemann, H.J., Meade, M.O., Cook, D.J., Erwin, P.J., Sood, A., Sood, R., Lo, B., Thompson, C.A., Zhou, Q., Mills, E., & Guyatt, G.H. 2005a. Randomized trials stopped early for benefit: a systematic review. *JAMA*, 294, (17) 2203-2209 available from: PM:16264162

Montori, V.M., Permanyer-Miralda, G., Ferreira-Gonzalez, I., Busse, J.W., Pacheco-Huergo, V., Bryant, D., Alonso, J., Akl, E.A., Domingo-Salvany, A., Mills, E., Wu, P., Schunemann, H.J., Jaeschke, R., & Guyatt, G.H. 2005b. Validity of composite end points in clinical trials. *BMJ*, 330, (7491) 594-596 available from: PM:15761002

Morino, M., Rimonda, R., Allaix, M.E., Giraudo, G., & Garrone, C. 2005. Ultrasonic versus standard electric dissection in laparoscopic colorectal surgery: a prospective randomized clinical trial. *Ann.Surg.*, 242, (6) 897-901, discussion available from: PM:16327500

Morissette, K., Tricco, A.C., Horsley, T., Chen, M.H., & Moher, D. 2011. Blinded versus unblinded assessments of risk of bias in studies included in a systematic review. *Cochrane.Database.Syst.Rev.* (9) MR000025 available from: PM:21901737

Moseley, J.B., O'Malley, K., Petersen, N.J., Menke, T.J., Brody, B.A., Kuykendall, D.H., Hollingsworth, J.C., Ashton, C.M., & Wray, N.P. 2002. A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *N.Engl.J.Med.*, 347, (2) 81-88 available from: PM:12110735

Nagendran, M., Gurusamy, K.S., Aggarwal, R., Loizidou, M., & Davidson, B.R. 2013. Virtual reality training for surgical trainees in laparoscopic surgery. *Cochrane.Database.Syst.Rev.*, 8, CD006575 available from: PM:23980026

Nagpal, K., Abboudi, M., Fischler, L., Schmidt, T., Vats, A., Manchanda, C., Sevdalis, N., Scheidegger, D., Vincent, C., & Moorthy, K. 2011. Evaluation of postoperative handover using a tool to assess information transfer and teamwork. *Ann.Surg.*, 253, (4) 831-837 available from: PM:21475027

Nagpal, K., Vats, A., Vincent, C., & Moorthy, K. 2009. A systems approach to errors. *Surgery*, 145, (6) 689-690 available from: PM:19486778

National Institute for Health Research. Research programs. Available at: http://www.nihr.ac.uk/research/Pages/programmes_research_programmes.aspx (Accessed December 17, 2013) .

Neuhauser, M. 2006. How to deal with multiple endpoints in clinical trials. *Fundam.Clin.Pharmacol.*, 20, (6) 515-523 available from: PM:17109645

Oropresa, I., Lamata, P., Sanchez-Gonzalez, P., Pagador J.B., Garcia M.E., Sanchez-Margallo, F. M., & Gomez, E. J. Virtual Reality Simulators for Objective Evaluation on Laparoscopic Surgery: Current Trends and Benefits. -Virtual Reality, Prof. Jae-Jin Kim (Ed.), ISBN: 978-953-307-518-1, InTech, Available from: <http://www.intechopen.com/books/virtual-reality/virtual-reality-simulatorsfor-objective-evaluation-on-laparoscopic-surgery-current-trends-and-benef> (Accessed November 26, 2013).

Patchell, R.A., Tibbs, P.A., Regine, W.F., Payne, R., Saris, S., Kryscio, R.J., Mohiuddin, M., & Young, B. 2005. Direct decompressive surgical resection in the treatment of spinal cord compression caused by metastatic cancer: a randomised trial. *Lancet*, 366, (9486) 643-648 available from: PM:16112300

Patel, S.S., Patel, M.S., Mahanti, S., Ortega, A., Ault, G.T., Kaiser, A.M., & Senagore, A.J. 2012. Laparoscopic versus open colon resections in California: a cross-sectional analysis. *Am.Surg*, 78, (10) 1063-1065 available from: PM:23025941

Pearson K 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2, 559-572

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., & Feinstein, A.R. 1996. A simulation study of the number of events per variable in logistic regression analysis. *J.Clin.Epidemiol.*, 49, (12) 1373-1379 available from: PM:8970487

Peel, A.L. & Taylor, E.W. 1991. Proposed definitions for the audit of postoperative infection: a discussion paper. Surgical Infection Study Group. *Ann.R.Coll.Surg.Engl.*, 73, (6) 385-388 available from: PM:1759770

Pillai, S.B., van Rij, A.M., Williams, S., Thomson, I.A., Putterill, M.J., & Greig, S. 1999. Complexity- and risk-adjusted model for measuring surgical outcome. *Br.J.Surg.*, 86, (12) 1567-1572 available from: PM:10594507

- Platz, J. & Hyman, N. 2012. Tracking intraoperative complications. *J.Am.Coll.Surg.*, 215, (4) 519-523 available from: PM:22727607
- Pomposelli, J.J., Gupta, S.K., Zacharoulis, D.C., Landa, R., Miller, A., & Nanda, R. 1997. Surgical complication outcome (SCOUT) score: a new method to evaluate quality of care in vascular surgery. *J.Vasc.Surg.*, 25, (6) 1007-1014 available from: PM:9201161
- Preston, C.C. & Colman, A.M. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychol.(Amst)*, 104, (1) 1-15 available from: PM:10769936
- Reason, J. 2000. Human error: models and management. *BMJ*, 320, (7237) 768-770 available from: PM:10720363
- Richardson, W.S., Wilson, M.C., Nishikawa, J., & Hayward, R.S. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP J.Club.*, 123, (3) A12-A13 available from: PM:7582737
- Riggs, B.L., Hodgson, S.F., O'Fallon, W.M., Chao, E.Y., Wahner, H.W., Muhs, J.M., Cedel, S.L., & Melton, L.J., III 1990. Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis. *N.Engl.J.Med.*, 322, (12) 802-809 available from: PM:2407957
- Rosenthal, R., Geuss, S., Dell-Kuster, S., Schafer, J., Hahnloser, D., & Demartines, N. 2011. Video gaming in children improves performance on a virtual reality trainer but does not yet make a laparoscopic surgeon. *Surg.Innov.*, 18, (2) 160-170 available from: PM:21245068
- Rosenthal, R. & Dwan, K. 2013. Comparison of randomized controlled trial registry entries and content of reports in surgery journals. *Ann.Surg.*, 257, (6) 1007-1015 available from: PM:23478520
- Rosenthal, R., Schafer, J., Briel, M., Bucher, H.C., Oertli, D., & Dell-Kuster, S. 2014. How to write a surgical clinical research protocol: literature review and practical guide. *Am.J.Surg*, 207, (2) 299-312 available from: PM:24262932
- Rosenthal, R., Bucher, H. C., Dwan, K. 2015a. The use of systematic reviews when designing and reporting surgical trials. *Ann.Surg.* Jan 23 [Epub ahead of print], DOI: 10.1097/SLA.0000000000001092 available from: PM:25970024
- Rosenthal, R., Kasenda, B., Dell-Kuster, S., Von Elm, E., You, J., Blumle, A., Tomonaga, Y., Saccilotto, R., Amstutz, A., Bengough, T., Meerpohl, J., Stegert, M., Tikkinen, K. A., Neumann, I., Carrasco-Labra, A., Faulhaber, M., Mulla, S., Mertz, D., Akl, E. A., Bassler, D., Busse, J. W., Ferreira-Gonzalez, I., Lamontagne, F., Nordmann, A., Gloy, V., Olu, K. K., Raatz, H., Moja, L., Ebrahim, S., Schandelmaier, S., Sun, X., Vandvik, P. O., Johnston, B. C., Walter, M. A., Burnand, B., Schwenkglenks, M., Hemkens, L. G., Bucher, H. C., Guyatt, G., & Briel, M. 2015b. Completion and publication rates of randomized controlled trials in surgery: an empirical study. *Ann.Surg.*, 262, (1) 68-73 available from: PM:24979608

- Rosenthal, R., Hoffmann, H., Dwan, K., Clavien, P. A., Bucher, H. C. 2015c. Reporting of adverse events in surgical trials: critical appraisal of current practice. *World J.Surg.*, 39, (1) 80-87 available from: PM:25201471
- Rosenthal, R., Hoffmann, H., Clavien, P. A., Bucher, H. C., Dell-Kuster. S. 2015d. Definition and Classification of Intraoperative Complications (CLASSIC): Delphi study and pilot evaluation. *World J.Surg.*, 39, (7) 1663-1671 available from: PM:25665678
- Rosenthal, R., von Websky, M.W., Hoffmann, H., Vitz, M., Hahnloser, D., Bucher, H. C., Schäfer, J. 2015e. How to report multiple outcome metrics in virtual reality simulation. *Eur.Surg.*, 47, 202-205
- Royston, P. & Sauerbrei, W. 2004. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Stat.Med.*, 23, (16) 2509-2525 available from: PM:15287081
- Russell, E.M., Bruce, J., & Krukowski, Z.H. 2003. Systematic review of the quality of surgical mortality monitoring. *Br.J.Surg.*, 90, (5) 527-532 available from: PM:12734856
- Sarkar, D. 2008, Lattice: Multivariate Data Visualization with R. New York, USA: Springer.
- Scherer, R.W., Langenberg, P., & Von, E.E. 2007. Full publication of results initially presented in abstracts. *Cochrane.Database.Syst.Rev.* (2) MR000005 available from: PM:17443628
- Schouten, H.J. 2000. Combined evidence from multiple outcomes in a clinical trial. *J.Clin.Epidemiol.*, 53, (11) 1137-1144 available from: PM:11106887
- Schulz, K.F., Altman, D.G., & Moher, D. 2010. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, c332 available from: PM:20332509
- Schulz, K.F. & Grimes, D.A. 2005. Multiplicity in randomised trials II: subgroup and interim analyses. *Lancet*, 365, (9471) 1657-1661 available from: PM:15885299
- Senders W, M. N. e. Human Error (Cause, Prediction, and Reduction): Analysis and Synthesis. Hillsdale, NJ: Lawrence A Erlbaum Associates; 1991.
- Shrout, P.E. & Fleiss, J.L. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychol.Bull.*, 86, (2) 420-428 available from: PM:18839484
- Slankamenac, K., Graf, R., Barkun, J., Puhan, M.A., & Clavien, P.A. 2013. The comprehensive complication index: a novel continuous scale to measure surgical morbidity. *Ann.Surg.*, 258, (1) 1-7 available from: PM:23728278
- Sokol, D.K. & Wilson, J. 2008. What is a surgical complication? *World J.Surg.*, 32, (6) 942-944 available from: PM:18266027

Solomon, M.J. & McLeod, R.S. 1995. Should we be performing more randomized controlled trials evaluating surgical operations? *Surgery*, 118, (3) 459-467 available from: PM:7652679

Sorenson, C. & Drummond, M. 2014. Improving medical device regulation: the United States and Europe in perspective. *Milbank Q.*, 92, (1) 114-150 available from: PM:24597558

Sterne, J.A. & Davey, S.G. 2001. Sifting the evidence-what's wrong with significance tests? *BMJ*, 322, (7280) 226-231 available from: PM:11159626

Strasberg, S.M., Linehan, D.C., & Hawkins, W.G. 2009. The accordion severity grading system of surgical complications. *Ann.Surg.*, 250, (2) 177-186 available from: PM:19638919

Streiner, D.L. 1995. Learning how to differ: agreement and reliability statistics in psychiatry. *Can.J.Psychiatry*, 40, (2) 60-66 available from: PM:7788619

Streiner, D. L. & Norman, G. R. From items to scales. In: Streiner DL, Norman GR, eds. Health measurement scales - a practical guide to their development and use. 4th ed.Oxford: Oxford University press; 148. 2008.

Surgical Journal Editors Group 2007. Consensus statement on mandatory registration of clinical trials. *Br.J.Surg*, 94, (4) 511-512 available from: PM:17385182

Sweet, B.V., Schwemm, A.K., & Parsons, D.M. 2011. Review of the processes for FDA oversight of drugs, medical devices, and combination products. *J.Manag.Care Pharm.*, 17, (1) 40-50 available from: PM:21204589

Tang, B., Hanna, G.B., Joice, P., & Cuschieri, A. 2004. Identification and categorization of technical errors by Observational Clinical Human Reliability Assessment (OCHRA) during laparoscopic cholecystectomy. *Arch.Surg.*, 139, (11) 1215-1220 available from: PM:15545569

Temple, R. J. A regulatory authority's opinion about surrogate endpoints. In: Nimmo WS, Tucker GT, eds. *Clinical Measurement in Drug Evaluation*. New York: John Wiley and Sons Inc., 1995. p 57.

The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). ICH Harmonised Tripartite Guideline. Guideline for Good Clinical Practice E6 (R1). Geneva, Switzerland. Available at: <http://www.ich.org/>. Accessed April 04, 2012.

The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for HumanUse (ICH). ICH Harmonised Tripartite Guideline. General considerations for clinical trials. E8. Geneva, Switzerland. Available at: <http://www.ich.org/>. Accessed January 30, 2013.

Thompson, M., Tiwari, A., Fu, R., Moe, E., & Buckley, D. I. A Framework to Facilitate the Use of Systematic Reviews and Meta-analyses in the Design of Primary Research Studies. (Prepared by the Oregon Evidence-based Practice Center under Contract HHS 290-2007-10057-I.) AHRQ Publication No. 12-EHC009-EF.Rockville,

MD: Agency for Healthcare Research and Quality. January 2012. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm. (Accessed December 04, 2013).

Tincello, D.G., Kenyon, S., Slack, M., Toozs-Hobson, P., Mayne, C., Jones, D., & Taylor, D. 2009. Colposuspension or TVT with anterior repair for urinary incontinence and prolapse: results of and lessons from a pilot randomised patient-preference study (CARPET 1). *BJOG.*, 116, (13) 1809-1814 available from: PM:19781044

Tou, S., Malik, A.I., Wexner, S.D., & Nelson, R.L. 2011. Energy source instruments for laparoscopic colectomy. *Cochrane.Database.Syst.Rev.* (5) CD007886 available from: PM:21563161

Treweek, S., Lockhart, P., Pitkethly, M., Cook, J.A., Kjeldstrom, M., Johansen, M., Taskila, T.K., Sullivan, F.M., Wilson, S., Jackson, C., Jones, R., & Mitchell, E.D. 2013. Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis. *BMJ Open.*, 3, (2) available from: PM:23396504

Tukey, J. W. The problem of multiple comparisons (unpublished manuscript). In: Braun H, ed. *The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948-1983*, New York, USA: Chapman & Hall. 1953.
Ref Type: Generic

U.S. Department of Health and Human Services, N. I. o. H. N. C. I. Common Terminology Criteria for Adverse Events (CTCAE) Version 4.03. Available at: <http://evs.nci.nih.gov/ftp1/CTCAE/About.html>. Accessed May 10, 2014. 2014.
Ref Type: Generic

van den Broek, M.A., van Dam, R.M., van Breukelen, G.J., Bemelmans, M.H., Oussoultzoglou, E., Pessaux, P., Dejong, C.H., Freemantle, N., & Olde Damink, S.W. 2011. Development of a composite endpoint for randomized controlled trials in liver surgery. *Br.J.Surg.*, 98, (8) 1138-1145 available from: PM:21557208

van Dongen, K.W., Ahlberg, G., Bonavina, L., Carter, F.J., Grantcharov, T.P., Hyltander, A., Schijven, M.P., Stefani, A., van der Zee, D.C., & Broeders, I.A. 2011. European consensus on a competency-based virtual reality training program for basic endoscopic surgical psychomotor skills. *Surg.Endosc.*, 25, (1) 166-171 available from: PM:20574856

Von Elm, E., Rollin, A., Blumle, A., Huwiler, K., Witschi, M., & Egger, M. 2008. Publication and non-publication of clinical trials: longitudinal study of applications submitted to a research ethics committee. *Swiss.Med.Wkly.*, 138, (13-14) 197-203 available from: PM:18389392

von Websky, M.W., Raptis, D.A., Vitz, M., Rosenthal, R., Clavien, P.A., & Hahnloser, D. 2013. Access to a simulator is not enough: the benefits of virtual reality training based on peer-group-derived benchmarks--a randomized controlled trial. *World J.Surg.*, 37, (11) 2534-2541 available from: PM:23942532

von Websky, M.W., Vitz, M., Raptis, D.A., Rosenthal, R., Clavien, P.A., & Hahnloser, D. 2012. Basic laparoscopic training using the Symbionix LAP Mentor: setting the

- standards in the novice group. *J.Surg.Educ.*, 69, (4) 459-467 available from: PM:22677582
- Ware, J.E., Jr. & Sherbourne, C.D. 1992. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med.Care*, 30, (6) 473-483 available from: PM:1593914
- Warrens, M.J. 2011. Weighted kappa is higher than Cohen's kappa for tridiagonal agreement tables. *Statistical Methodology*, 8, (2) 268-272 available from: ISI:000304292900012
- Wente, M.N., Bassi, C., Dervenis, C., Fingerhut, A., Gouma, D.J., Izbicki, J.R., Neoptolemos, J.P., Padbury, R.T., Sarr, M.G., Traverso, L.W., Yeo, C.J., & Buchler, M.W. 2007a. Delayed gastric emptying (DGE) after pancreatic surgery: a suggested definition by the International Study Group of Pancreatic Surgery (ISGPS). *Surgery*, 142, (5) 761-768 available from: PM:17981197
- Wente, M.N., Veit, J.A., Bassi, C., Dervenis, C., Fingerhut, A., Gouma, D.J., Izbicki, J.R., Neoptolemos, J.P., Padbury, R.T., Sarr, M.G., Yeo, C.J., & Buchler, M.W. 2007b. Postpancreatectomy hemorrhage (PPH): an International Study Group of Pancreatic Surgery (ISGPS) definition. *Surgery*, 142, (1) 20-25 available from: PM:17629996
- Whitley, E. & Ball, J. 2002. Statistics review 1: presenting and summarising data. *Crit Care*, 6, (1) 66-71 available from: PM:11940268
- Williamson, P.R., Altman, D.G., Blazeby, J.M., Clarke, M., Devane, D., Gargon, E., & Tugwell, P. 2012. Developing core outcome sets for clinical trials: issues to consider. *Trials*, 13, 132 available from: PM:22867278
- Wilson, A.P., Treasure, T., Sturridge, M.F., & Gruneberg, R.N. 1986. A scoring method (ASEPSIS) for postoperative wound infections for use in clinical trials of antibiotic prophylaxis. *Lancet*, 1, (8476) 311-313 available from: PM:2868173
- Wilson, J. & Sokol, D.K. 2009. Do we need a concept of intraoperative complication? *World J.Surg.*, 33, (5) 1102 available from: PM:19198931
- World Health Organization. WHO Trial Registration Data Set (Version 1.2.1) [WHO Web site]. Available at: <http://www.who.int/ictrp/network/trds/en/index.html>. Accessed August 15, 2013 .
- Zarin, D.A., Ide, N.C., Tse, T., Harlan, W.R., West, J.C., & Lindberg, D.A. 2007. Issues in the registration of clinical trials. *JAMA*, 297, (19) 2112-2120 available from: PM:17507347
- Zeger, S.L. & Liang, K.Y. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42, (1) 121-130 available from: PM:3719049

11. Appendices

Completion and publication rates of surgical randomized controlled trials – an empirical study

Appendix Table S1: Subspecialties of RCTs

S1a: Subspecialties of surgical RCTs (n=127)

S1b: Subspecialties of medical RCTs (n=736)

Appendix Table S2: Multivariable regression for trial discontinuation for slow recruitment of surgical trials not accounting for clustering by the approving research ethics committee (n=89) and after multiple imputation of missing predictor and outcome variables (n=115)

Appendix Table S3: Multivariable regression for non-publication as full journal article of surgical trials not accounting for clustering by the approving research ethics committee (n=89) and after multiple imputation of missing predictor and outcome variables (n=115)

Appendix Figure S1: Flow chart of included studies and supplementary information concerning invasive versus non-invasive trials

Reporting of adverse events in surgical trials: critical appraisal of current practice

Appendix References R1: References of the 51 studies included in the analysis

Appendix Table S4: Baseline characteristics of all studies included in the summary statistics (n=46)

Appendix Table S5: Adverse event reporting in all studies included in the summary statistics (n=46)

Appendix Table S6: Classification of postoperative complications

Appendix Table S7: Definition of Surgical Site Infection (SSI)

Definition and Classification of Intraoperative Complications (CLASSIC): Delphi Study and pilot evaluation

Appendix Text (including illustrations) S1: Questions and summary results round 1, questions round 2

Appendix Text (including illustrations) S2: Results round 2

Appendix Table S8: Patient characteristics of the pilot study (n=60)

Appendix Table S9: Procedure characteristics of the pilot study (n=60)

Appendix Table S10: Outcomes of the pilot study (n=60)

Appendix Table S1: Subspecialties of RCTs

S1a: Subspecialties of surgical RCTs (n=127)

Numbers (column percentages) are displayed.

Subspecialty	Frequency (%)
total	127 (100)
anesthesiology	17 (13)
cardiothoracic	16 (13)
dentistry	2 (2)
ear-nose-throat (ENT)	6 (5)
general surgery	8 (6)
maxillofacial surgery	2 (2)
neurosurgery	6 (5)
obstetrics / gynecology	20 (16)
ophthalmology	18 (14)
orthopedics	13 (10)
plastic surgery	1 (1)
transplantation	3 (2)
traumatology	3 (2)
urology	8 (6)
vascular surgery	4 (3)

S1b: Subspecialties of medical RCTs (n=736)

Numbers (column percentages) are displayed.

Subspecialty	Frequency (%)
total	736 (100)
alternative medicine	1 (<1)
cardiovascular	111 (15)
dermatology	22 (3)
emergency medicine	2 (<1)
endocrinology	57 (8)
gastroenterology	48 (6)
haematology	35 (5)
immunology	6 (1)
infectious diseases	80 (11)
intensive care	13 (2)
nephrology	15 (2)
neurology	58 (8)
oncology	155 (21)
physiotherapy	1 (<1)
psychiatry	41 (5)
psychotherapy	3 (<1)
radiology	7 (1)
rehabilitation	6 (1)
pneumology	44 (6)
rheumatology	30 (4)
sports medicine	1 (<1)

Appendix Table S2: Multivariable regression for trial discontinuation for slow recruitment of surgical trials disregarding clustering by the approving research ethics committee (n=89¹) and after multiple imputation of missing predictor and outcome variables (n=115²)

Potential predictor	Multivariable, disregarding clustering by research ethics committee			Multivariable, multiple imputations		
	Odds ratio	95% CI ³	P value	Odds ratio	95% CI	P value
Industry sponsor, yes <i>versus</i> no	0.53	0.19, 1.51	0.235	0.64	0.23, 1.81	0.402
Primary harm outcome, yes <i>versus</i> no	0.55	0.16, 1.91	0.348	0.50	0.14, 1.80	0.291

¹ Out of 115 trials, 12 had an unclear completion status and 14 had at least one missing covariate.

² Exclusion of 10 trials that never started and of 2 that were still recruiting

³ CI=Confidence interval

Appendix Table S3: Multivariable regression for non-publication as full journal article of surgical trials disregarding clustering by the approving research ethics committee (n=89¹) and after multiple imputation of missing predictor and outcome variables (n=115²)

Potential predictor	Multivariable, disregarding clustering by research ethics committee			Multivariable, multiple imputations		
	Odds ratio	95% CI ³	P value	Odds ratio	95% CI	P value
Industry sponsor, yes <i>versus</i> no	2.99	1.05, 8.58	0.041	1.92	0.81, 4.53	0.137
Primary harm outcome, yes <i>versus</i> no	0.28	0.09, 0.92	0.035	0.30	0.10, 0.90	0.031
Trial discontinuation for any reason, yes <i>versus</i> no	4.18	1.45, 12.05	0.008	3.32	1.33, 8.33	0.010

143

¹ Out of 115 trials, 26 had at least one missing covariate.

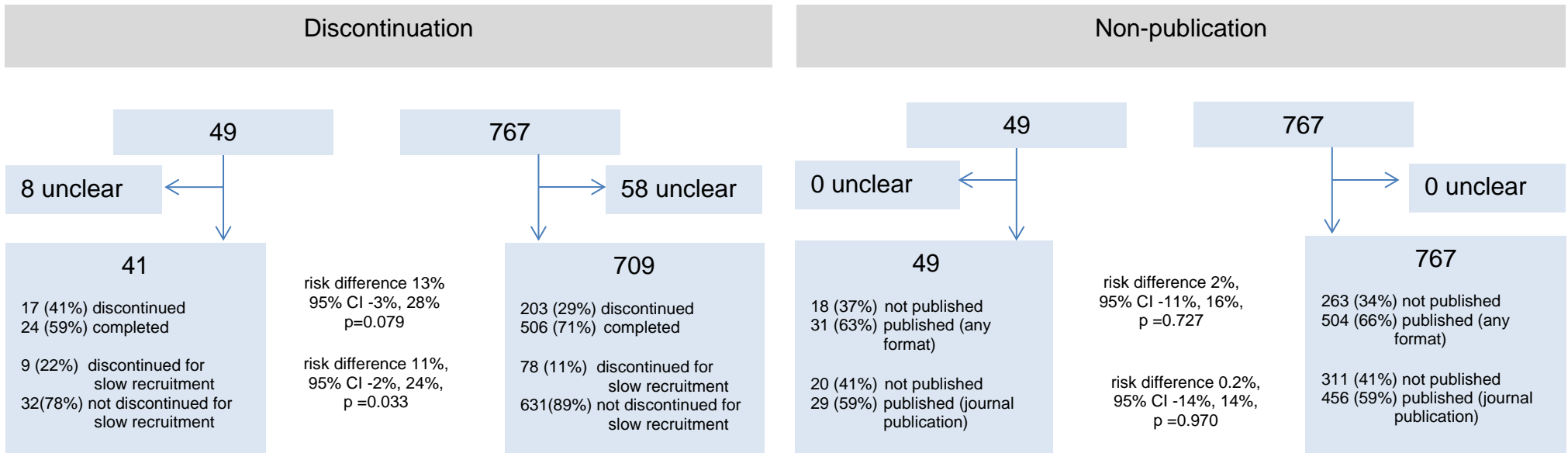
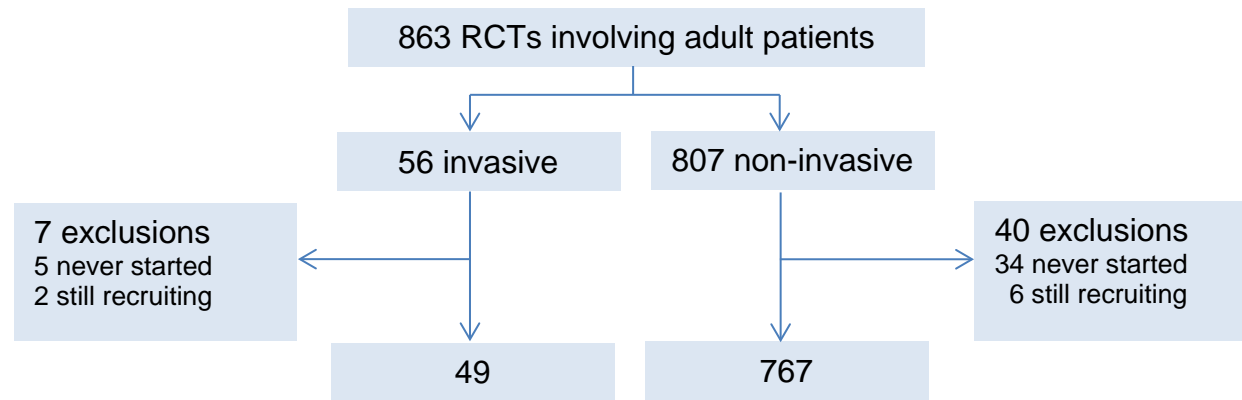
² Exclusion of 10 trials that never started and of 2 that were still recruiting

³ CI=Confidence interval

Appendix Figure S1

Flow chart of included studies, eligibility, descriptive results and risk differences (95% confidence intervals) comparing invasive to non-invasive trials

144



Appendix References R1: References of the 51 studies included in the analysis

* references of the studies included in the analysis (other references in the manuscript are presented without *)

^{add} additional publications concerning a specific RCT

1* Schouten R, Rijs CS, Bouvy N, et al. A multicenter, randomized efficacy study of the EndoBarrier Gastrointestinal Liner for presurgical weight loss prior to bariatric surgery. *Ann Surg* 2010 ; 251(2):236-243.

2* Aboulian A, Chan T, Yaghoobian A, et al. Early cholecystectomy safely decreases hospital stay in patients with mild gallstone pancreatitis: a randomized prospective study. *Ann Surg* 2010; 251(4):615-619.

3* Ishikawa Y, Kiyama T, Haga Y, et al. Maximal sterile barrier precautions do not reduce catheter-related bloodstream infections in general surgery units: a multi-institutional randomized controlled trial. *Ann Surg* 2010; 251(4):620-623.

4* Parés JO, Juan J, Tellez R, et al. Varicose vein surgery: stripping versus the CHIVA method: a randomized controlled trial. *Ann Surg* 2010; 251(4):624-631.

5* Langeveld HR, van't Riet M, Weidema WF, et al. Total extraperitoneal inguinal hernia repair compared with Lichtenstein (the LEVEL-Trial): a randomized controlled trial. *Ann Surg* 2010;251(5):819-824.

6* Danielsen PL, Agren MS, Jorgensen LN. Platelet-rich fibrin versus albumin in surgical wound repair: a randomized trial with paired design. *Ann Surg* 2010;251(5):825-831.

7* Morris GS, Grubbs EG, Hearon CM, et al. Parathyroidectomy improves functional capacity in "asymptomatic" older patients with primary hyperparathyroidism: a randomized control trial. *Ann Surg* 2010; 251(5):832-837.

8* Daraï E, Dubernard G, Coutant C, et al. Randomized trial of laparoscopically assisted versus open colorectal resection for endometriosis: morbidity, symptoms, quality of life, and fertility. *Ann Surg* 2010;251(6):1018-1023.

9* Sammour T, Kahokehr A, Hayes J, et al. Warming and humidification of insufflation carbon dioxide in laparoscopic colonic surgery: a double-blinded randomized controlled trial. *Ann Surg* 2010; 251(6):1024-1033.

10* Fink AS, Prochazka AV, Henderson WG, et al. Enhancement of surgical informed consent by addition of repeat back: a multicenter, randomized controlled clinical trial. *Ann Surg* 2010; 252(1):27-36.

11* Burdess A, Nimmo AF, Garden OJ, et al. Randomized controlled trial of dual antiplatelet therapy in patients undergoing surgery for critical limb ischemia. *Ann Surg* 2010; 252(1):37-42.

12* Peeters E, Spiessens C, Oyen R, et al. Laparoscopic inguinal hernia repair in men with lightweight meshes may significantly impair sperm motility: a randomized controlled trial. *Ann Surg* 2010; 252(2):240-246.

13* Awad S, Constantin-Teodosiu D, Constantin D, et al. Cellular mechanisms underlying the protective effects of preoperative feeding: a randomized study investigating muscle and liver glycogen content, mitochondrial function, gene and protein expression. *Ann Surg* 2010; 252(2):247-253.

14* Jochmans I, Moers C, Smits JM, et al. Machine perfusion versus cold storage for the preservation of kidneys donated after cardiac death: a multicenter, randomized, controlled trial. *Ann Surg* 2010; 252(5):756-764.

15* Petrowsky H, Breitenstein S, Slankamenac K, et al. Effects of pentoxifylline on liver regeneration: a double-blinded, randomized, controlled trial in 101 patients undergoing major liver resection. *Ann Surg* 2010; 252(5):813-822.

16* Bretagnol F, Panis Y, Rullier E, et al.; French Research Group of Rectal Cancer Surgery (GRECCAR). Rectal cancer surgery with or without bowel preparation: The French GRECCAR III multicenter single-blinded randomized trial. *Ann Surg* 2010; 252(5) :863-868.

17* Fischer M, Matsuo K, Gonen M, et al. Relationship between intraoperative fluid administration and perioperative outcome after pancreaticoduodenectomy: results of a prospective randomized trial of acute normovolemic hemodilution compared with standard intraoperative management. *Ann Surg* 2010; 252(6):952-958.

18* Bernard SA, Nguyen V, Cameron P, et al. Prehospital rapid sequence intubation improves functional outcome for patients with severe traumatic brain injury: a randomized controlled trial. *Ann Surg* 2010; 252(6):959-965.

19* Gervaz P, Inan I, Perneger T, et al. A prospective, randomized, single-blind comparison of laparoscopic versus open sigmoid colectomy for diverticulitis. *Ann Surg* 2010; 252(1):3-8.

20* Bassi C, Molinari E, Malleo G, et al. Early versus late drain removal after standard pancreatic resections: results of a prospective randomized trial. *Ann Surg* 2010; 252(2):207-214.

21* Huang J, Yan L, Cheng Z, et al. A randomized trial comparing radiofrequency ablation and surgical resection for HCC conforming to the Milan criteria. *Ann Surg* 2010; 252(6):903-912.

22* Rogers SJ, Cello JP, Horn JK, et al. Prospective randomized trial of LC+LCBDE vs ERCP/S+LC for common bile duct stone disease. *Arch Surg* 2010;145(1):28-33.

23* Itani KM, Hur K, Kim LT, et al. Comparison of laparoscopic and open repair with mesh for the treatment of ventral incisional hernia: a randomized trial. *Arch Surg* 2010; 145(4):322-328.

24* Mishra PK, Mathias H, Millar K, et al. A randomized controlled trial to assess the effect of audiotaped consultations on the quality of informed consent in cardiac surgery. *Arch Surg* 2010; 145(4):383-388.

25* Futier E, Constantin JM, Petit A, et al. Conservative vs restrictive individualized goal-directed fluid replacement strategy in major abdominal surgery: A prospective randomized trial. *Arch Surg* 2010; 145(12):1193-1200.

26* Schmittner MD, Schreiber H, Janke A, et al. Randomized clinical trial of perianal surgery performed under spinal saddle block versus total intravenous anaesthesia. *Br J Surg* 2010; 97(1):12-20.

27* Mekako AI, Chetter IC, Coughlin PA, et al.; Hull, Antibiotic prophylaxis in varicose Vein Surgery Trialists (HARVEST). Randomized clinical trial of co-amoxiclav versus no antibiotic prophylaxis in varicose vein surgery. *Br J Surg* 2010; 97(1):29-36.

28* Søvik TT, Taha O, Aasheim ET, et al. Randomized clinical trial of laparoscopic gastric bypass versus laparoscopic duodenal switch for superobesity. *Br J Surg* 2010; 97(2):160-166.

29* Nyström PO, Qvist N, Raahave D, et al. Randomized clinical trial of symptom control after stapled anopexy or diathermy excision for haemorrhoid prolapse. *Br J Surg* 2010; 97(2):167-176.

30* Hessman O, Westerdahl J, Al-Suliman N, et al. Randomized clinical trial comparing open with video-assisted minimally invasive parathyroid surgery for primary hyperparathyroidism. *Br J Surg* 2010; 97(2):177-184.

31* Wang ZG, Wang Q, Wang WJ, et al. Randomized clinical trial to compare the effects of preoperative oral carbohydrate versus placebo on insulin resistance after colorectal surgery. *Br J Surg* 2010; 97(3):317-327.

32* Subramonia S, Lees T. Randomized clinical trial of radiofrequency ablation or conventional high ligation and stripping for great saphenous varicose veins. *Br J Surg* 2010; 97(3):328-336.

33* Maggiori L, Rullier E, Meyer C, et al. Randomized controlled trial of pelvic calcium alginate following rectal cancer surgery. *Br J Surg* 2010; 97(4):479-484.

34* Mathur S, Plank LD, McCall JL, et al. Randomized controlled trial of preoperative oral carbohydrate treatment in major abdominal surgery. *Br J Surg* 2010; 97(4):485-494.

35* Degiuli M, Sasako M, Ponti A; Italian Gastric Cancer Study Group. Morbidity and mortality in the Italian Gastric Cancer Study Group randomized clinical trial of D1 versus D2 resection for gastric cancer. *Br J Surg* 2010; 97(5):643-649.

36* O'Hare JL, Stephens J, Parkin D, et al. Randomized clinical trial of different bandage regimens after foam sclerotherapy for varicose veins. *Br J Surg* 2010; 97(5):650-656.

37* Jiang ZM, Wilmore DW, Wang XR, et al. Randomized clinical trial of intravenous soybean oil alone versus soybean oil plus fish oil emulsion after gastrointestinal cancer surgery. *Br J Surg* 2010; 97(6):804-809.

38* Shepherd AC, Gohel MS, Brown LC, et al. Randomized clinical trial of VNUS ClosureFAST radiofrequency ablation versus laser for varicose veins. *Br J Surg* 2010; 97(6):810-818.

39* Steffen T, Warschkow R, Brändle M, et al. Randomized controlled trial of bilateral superficial cervical plexus block versus placebo in thyroid surgery. *Br J Surg* 2010; 97(7):1000-1006.

40* Braga M, Frasson M, Zuliani W, et al. Randomized clinical trial of laparoscopic versus open left colonic resection. *Br J Surg* 2010; 97(8):1180-1186.

41* Carli F, Charlebois P, Stein B, et al. Randomized clinical trial of prehabilitation in colorectal surgery. *Br J Surg* 2010; 97(8):1187-1197.

42* Hendry PO, van Dam RM, Bukkems SF, et al.; Enhanced Recovery After Surgery (ERAS) Group. Randomized clinical trial of laxatives and oral nutritional supplements within an enhanced recovery after surgery protocol following liver resection. *Br J Surg* 2010; 97(8):1198-1206.

43* Maeda Y, Ng SC, Durdey P, et al.; Topical Metronidazole in Perianal Crohn's Study Group. Randomized clinical trial of metronidazole ointment versus placebo in perianal Crohn's disease. *Br J Surg* 2010; 97(9):1340-1347.

44* Bevis PM, Windhaber RA, Lear PA, et al. Randomized clinical trial of mesh versus sutured wound closure after open abdominal aortic aneurysm surgery. *Br J Surg* 2010;97(10):1497-1502.

45* Momsen AH, Jensen MB, Norager CB, et al. Randomized double-blind placebo-controlled crossover study of caffeine in patients with intermittent claudication' *Br J Surg* 2010; 97(10):1503-1510.

46* Reid K, Pockney P, Pollitt T, et al. Randomized clinical trial of short-term outcomes following purse-string versus conventional closure of ileostomy wounds. *Br J Surg* 2010; 97(10):1511-1517.

47* Nthumba PM, Stepita-Poenaru E, Poenaru D, et al. Cluster-randomized, crossover trial of the efficacy of plain soap and water versus alcohol-based rub for surgical hand preparation in a rural hospital in Kenya. *Br J Surg* 2010; 97(11):1621-1628.

48* Gatt M, MacFie J. Randomized clinical trial of gut-specific nutrients in critically ill surgical patients. *Br J Surg* 2010; 97(11) :1629-1636.

49* Jayne DG, Thorpe HC, Copeland J, et al. Five-year follow-up of the Medical Research Council CLASICC trial of laparoscopically assisted versus open surgery for colorectal cancer. *Br J Surg* 2010; 97(11):1638-1645.

49*^{add1} Guillou P, Quirke P, Thorpe H, et al.; MRC CLASICC trial group. Short-term endpoints of conventional versus laparoscopic-assisted surgery in patients with colorectal cancer (MRC CLASICC trial): multicentre, randomised controlled trial. *Lancet* 2005; 365(9472):1718-1726.

49*^{add2} Jayne DG, Guillou PJ, Thorpe H, et al.; UK MRC CLASICC Trial Group. Randomized trial of laparoscopic-assisted resection of colorectal carcinoma: 3-year results of the UK MRC CLASICC Trial Group. *J Clin Oncol* 2007; 25(21):3061-3068.

50* Chalmers RT, Darling III RC, Wingard JT, et al. Randomized clinical trial of tranexamic acid-free fibrin sealant during vascular surgical procedures. *Br J Surg* 2010; 97(12):1784-1789.

51* Watanabe M, Murakami M, Nakao K, et al. Randomized clinical trial of the influence of mechanical bowel preparation on faecal microflora in patients undergoing colonic cancer resection. *Br J Surg* 2010; 97(12):1791-1797.

Appendix Table S4: Baseline characteristics of all studies included in the summary statistics (n=46)

* References of the studies included in the analysis mentioned in the Online Only References. NA=not applicable

First Author	Condition/ Population	Intervention	Control	Primary outcome (Result)	Adverse event reporting	Comments
Schouten ^{1*}	Planned bariatric surgery	Endoscopic implant mimicking duodenal-jejunal bypass	Diet	Weight loss (intervention higher than control)	Yes	
Aboulian ^{2*}	Mild pancreatitis	Early laporoscopic cholecystectomy	Control laparoscopic cholecystectomy	Length of hospital stay (intervention shorter than control)	Yes	Interim report, which is the final report, since the study was terminated thereafter.
Ishikawa ^{3*}	Central venous catheter insertion	Maximal sterile barrier	Standard sterile barrier	Incidence of catheter-related bloodstream infections (intervention = control)	Yes	
Parés ^{4*}	Varicose vein surgery	Ambulatory conservative hemodynamic management of varicose veins	Stripping (2 groups: with clinical marking and with duplex marking)	Clinical recurrence within 5 years (intervention less than control)	Yes	
Langeveld ^{5*}	Inguinal hernia	Laparoscopic total extraperitoneal repair	Open mesh repair (Lichtenstein)	Postoperative pain, length of hospital stay, time until complete recovery, quality of life (pain and time to recovery: intervention less/shorter than control, length of hospital stay and quality of life: intervention= control)	Yes	

Table S4: Baseline characteristics (continued)

First Author	Condition/ Population	Intervention	Control	Primary outcome (Result)	Adverse event reporting	Comments
Daniel- sen ^{6*}	Trocar site wounds in laparoscopic chole- cystectomy	Autologous platelet-rich fibrin	Human albumin	Incisional wound breaking strength (intervention = control) and subcutaneous collagen deposition (intervention lower concentration than control)	Yes	
Morris ^{7*}	Asymptomatic primary hyperpara- thyroidism	Parathyreoidectomy	Observation	Functional/ physical capacity (Six-minute walking distance improved in intervention group, unchanged in control group, Fifty foot walk and repeated sit- to-stand times unchanged in both groups)	Yes	
Darai ^{8*}	Colorectal endometriosis	Laparoscopically assisted colorectal resection	Open colorectal resection	Improvement in dyschesia (intervention = control)	Yes	
Sam- mour ^{9*}	Laparoscopic colon surgery	Warming and humidification of insufflation carbon dioxide	Standard dry carbon dioxide	Total opiate analgesia use (intervention = control)	Yes	

Table S4: Baseline characteristics (continued)

First Author	Condition/ Population	Intervention	Control	Primary outcome	Adverse event reporting	Comments
Burdess ^{11*}	Surgery for critical limb ischemia	Preoperative dual anti-platelet therapy	Preoperative single anti-platelet therapy	Platelet-monocyte aggregation (reduction by intervention, not control)	Yes	
Peeters ^{12*}	Laparoscopic inguinal hernia repair	Lightweight Vypro II or TiMesh prostheses	Heavyweight Marlex prosthesis	Fertility aspects (intervention decreased sperm motility (vs. preoperatively) compared with control)	Yes	
Awad ^{13*}	Laparoscopic cholecystectomy	Preoperative oral nutritional supplement	Placebo	Drink-related side-effects (non occurred)	Yes (nutritional supplement intervention). No (operation).	Labelled as pilot study. For summary statistics, primary outcome was regarded as assessment of an adverse event.
Jochmans ^{14*}	Kidneys donated after cardiac death	Machine perfusion	Cold storage	Delayed graft function (Intervention less frequently than control)	Yes	
Petrovsky ^{15*}	Major liver surgery	Pentoxifylline	Placebo	Liver regeneration (overall intervention = control, intervention better regeneration in small remnant livers)	Yes	

Table S4: Baseline characteristics (continued)

First Author	Condition/ Population	Intervention	Control	Primary outcome	Adverse event reporting	Comments
Bre-tagnol ^{16*}	Elective rectal cancer surgery	No bowel preparation	Bowel preparation	Overall 30-day morbidity rate (higher in intervention than control)	Yes	
Fischer ^{17*}	Pancreatico-duodenectomy	Normovolemic hemodilution	Standard management	Proportion of patients requiring allogeneic red blood cell products \leq 30 days (intervention = control)	Yes	In summary statistics, primary outcome was regarded as assessment of an adverse event.
Bernard ^{18*}	Severe traumatic brain injury	Prehospital rapid sequence intubation	Hospital intubation	Median extended Glasgow Outcome Scale (GOSe) (intervention = control. Proportion of patients with favourable GOSe: intervention higher than control (2° outcome)	Yes	
Gervaz ^{19*}	Diverticulitis scheduled for elective sigmoidectomy	Laparoscopic sigmoid colectomy	Open sigmoid colectomy	Postoperative pain, duration of postoperative ileus, total morphine intake (intervention similar pain compared to control, shorter ileus and lower morphine intake)	Yes	Third primary outcome: discrepancy between abstract and rest of the manuscript.
Bassi ^{20*}	Standard pancreatic resections	Early (postoperative day 3) drain removal	Standard (postoperative day \geq 5) drain removal	Pancreatic fistula (intervention lower rate than control)	Yes	

Table S4: Baseline characteristics (continued)

First Author	Condition/ Population	Intervention	Control	Primary outcome	Adverse event reporting	Comments
Huang ^{21*}	Small hepatocellular carcinoma	Radiofrequency ablation	Surgical resection	Overall survival (intervention lower than control)	Yes	
Rogers ^{22*}	Common bile duct stone disease	Laparoscopic cholecystectomy plus laparoscopic common bile duct exploration	Endoscopic retrograde cholangiopancreatography, sphincterotomy plus laparoscopic cholecystectomy	Efficacy of stone clearance (Intervention = control)	Yes	
Itani ^{23*}	Ventral incisional hernia	Laparoscopic mesh repair	Open mesh repair	% Patients with at least 1 complication within 8 weeks (intervention overall less complications than control; in patients with complications, more severe complications more frequently in intervention than control)	Yes	Primary outcome: discrepancy between abstract and rest of the manuscript.
Futier ^{25*}	Major abdominal surgery	Restrictive fluid administration	Conservative fluid administration	Incidence of overall postoperative complications (intervention higher than control)	Yes	

Table S4: Baseline characteristics (continued)

First Author	Condition/ Population	Intervention	Control	Primary outcome	Adverse event reporting	Comments
Schmittner ^{26*}	Perianal surgery	Spinal saddle block	Total intravenous anaesthesia	Analgesic consumption within 24h after surgery (intervention lower than control)	Yes	
Mekako ^{27*}	Varicose vein surgery	Co-amoxiclav	No antibiotic prophylaxis	Satisfactory wound outcome, defined as an ASEPSIS score of 10 or less (intervention more frequent than control)	Yes	
Søvik ^{28*}	Superobesity	Laparoscopic biliopancreatic diversion with duodenal switch	Laparoscopic Roux-en-Y gastric bypass	1-year weight loss (intervention greater than control)	Yes	Planned duration of follow-up is 5 years (registry information). This report includes 1-year data.
Nyström ^{29*}	Haemorrhoid prolapse	Stapled anopecty	Milligan-Morgan haemorrhoidectomy	Resolution of symptoms 1 year after operation (control higher percentage than intervention)	Yes	
Hessman ^{30*}	Primary hyperparathyroidism	Video-assisted minimally invasive parathyroidectomy	Open parathyroidectomy	Duration of surgery and postoperative pain (duration: control quicker than intervention; pain: intervention=control)	Yes	

Table S4: Baseline characteristics (continued)

First Author	Condition/ Population	Intervention	Control	Primary outcome	Adverse event reporting	Comments
Wang ^{31*}	Colorectal cancer surgery	Preoperative oral carbohydrate supplement	Placebo or fasting	Whole-body insulin sensitivity (postoperative insulin resistance placebo/fasting higher than intervention)	Yes	Outcome used for sample size calculation assumed to be the primary outcome
Subramonia ^{32*}	Great saphenous varicose veins	Radiofrequency ablation	Conventional high ligation and stripping	Time to return to normal household activities (intervention earlier than control)	Yes	Planned long-term follow-up. This report includes short-term results.
Maggiore ^{33*}	Elective sphincter-saving rectal cancer surgery	Pelvic calcium alginate	No haemostatic agent	Volume of fluid collected by the pelvic suction drain (intervention lower than control)	Yes	
Mathur ^{34*}	Major abdominal surgery	Oral carbohydrate drink	Placebo	Postoperative fatigue and length of stay (intervention=control)	Yes	

Table S4: Baseline characteristics (continued)

First Author	Condition/ Population	Intervention	Control	Primary outcome	Adverse event reporting	Comments
Degiuli ^{35*}	Gastric cancer	D2 gastrectomy	D1 gastrectomy	Overall survival (long-term follow-up). Current report: morbidity (prespecified 12% difference non-inferiority margin) and 30d-mortality. Intervention slightly exceeding non inferiority margin (95% confidence interval for the difference 0-13%).	Yes	Planned long-term follow-up. This report includes short-term results. For summary statistics primary outcome regarded as evaluation of adverse even.t
O'Hare ^{36*}	Foam sclerotherapy for varicose veins	Compression bandaging for 24 h	Compression bandaging for 5 days	Change in Aberdeen Varicose Vein Severity Score 6 weeks after foam sclerotherapy compared with pretreatment score (intervention=control)	Yes	
Jiang ^{37*}	Gastro-intestinal cancer surgery	Intravenous soybean oil plus fish oil emulsion	Intravenous soybean oil alone	Number of in-hospital postoperative infectious complications (intervention <control, p = 0.066) and Occurrence of SIRS (Intervention significantly less than control)	Yes	
Shepherd ^{38*}	Varicose veins	Radiofrequency ablation	Endovenous laser	Mean postprocedural pain over first 3 days (intervention lower pain score than control)	Yes	Primary outcome definition: discrepancy in abstract versus main text

Table S4: Baseline characteristics (continued)

First Author	Condition/ Population	Intervention	Control	Primary outcome	Adverse event reporting	Comments
Steffen ^{39*}	Thyroid surgery	Pre- or postoperative superficial cervical block with bupivacaine	Pre- or postoperative cervical placebo	Postoperative pain (intervention less than control)	Yes	
Braga ^{40*}	Elective left colonic resection	Laparoscopic surgery	Open surgery	30-day morbidity rate (outcome used for sample size calculation) (intervention = control)	Yes	A subgroup of patients had been included in a previous study.
Carli ^{41*}	Colorectal surgery	Prehabilitation regimen of stationary cycling and strengthening	Simple regimen of walking and breathing	6-minute walk test (6MWT) (intervention= control)	Yes	
Hendry ^{42*}	Liver resection	Laxatives or oral nutritional supplements (ONS) or both	Control	Time to first passage of stool (Laxatives reduced time. ONS not significant trend for reduced time)	Yes	
Bevis ^{44*}	Open abdominal aortic aneurysm surgery	Mesh wound closure	Sutured wound closure	Incisional hernia within 3 years of surgery (intervention <control)	Yes	In summary statistics, incisional hernia considered as postoperative complication
Reid ^{46*}	Closure of ileostomy wounds	Purse string closure	Linear closure	Surgical site infection (intervention fewer than control)	Yes	Interim report, which is the final report, since the study was terminated thereafter.

Table S4: Baseline characteristics (continued)

First Author	Condition/ Population	Intervention	Control	Primary outcome	Adverse event reporting	Comments
Nthumba ^{47*}	Rub for surgical hand preparation in rural hospital in Kenya	Alcohol based	Plain soap and water	Surgical site infection (intervention=control)	Yes	
Jayne ^{49*}	Colorectal cancer	Laparoscopically assisted surgery	Conventional surgery	Report includes long-term secondary outcomes 5-year overall survival, disease-free survival (DFS), and recurrences (intervention = control). [Primary short term outcomes ^{49*add1} were positivity rates of resection margins, proportion of Dukes' C2 tumours, and in-hospital mortality]	Yes (^{49*add1})	This report includes 5-year data. Short-term ^{49*add1} and 3-year ^{49*add2} results were published previously. For summary statistics, short-term results are considered ^{49*add1}
Chalmers ^{50*}	Polytetra-fluoro-ethylene arterial anastomoses	Tranexamic acid-free fibrin sealant	Manual compression	Absence of bleeding at the anastomosis at 4 min after randomization (intervention higher percentage than control)	Yes	

Table S4: Baseline characteristics (continued)

First Author	Condition/ Population	Intervention	Control	Primary outcome	Adverse event reporting	Comments
Watanabe ^{51*}	Colonic cancer resection	Mechanical bowel preparation	No mechanical bowel preparation	Bacterial microflora and faecal organic acid content of faecal material obtained at operation (bacterial microflora lower in inter-vention than control, levels of faecal organic acids lower in intervention than control)	Yes	

Appendix Table S5: Adverse event reporting in all studies included in the summary statistics (n=46)

* References of the studies included in the analysis mentioned in the Online Only References.

**postoperative and intraoperative refers not only to operations in the operating room, but as well to other invasive interventions. Conversions were not regarded as complication. However, some reasons for conversion were regarded as complications, e.g. bleeding.

M=retrieved from the methods section, R=retrieved from the results section. NA=not applicable. AE=adverse events.

First Author	Condition/ Population	Definition and occurrence of intraoperative** adverse event	Definition and occurrence of postoperative** adverse event	Classification of **intra-operative adverse event	Classification of **postoperative adverse event	Comments
Schouten ^{1*}	Planned bariatric surgery	No definition (M). None occurred. In 4 patients no implantation due to anatomical difficulties (R).	Yes, definition given with classification of AE (M). List of AE provided (R).	No classification (M, R).	Yes, according to severity in mild, moderate and severe (M). Percentages of patients with AE per severity grade reported (R).	
Aboulian ^{2*}	Mild pancreatitis	Yes, pooled with postoperative AE and defined by complete list (bleeding requiring transfusion) (M). None occurred (R).	Yes, pooled with intraoperative AE and defined by complete list (bile duct injury, bleeding requiring transfusion or reoperation, wound infection, pneumonia, and need for readmission within 30d) (M). None occurred (R).	No classification (M, R).	No classification (M, R).	
Ishikawa ^{3*}	Central venous catheter insertion	No exact definition, examples given (pneumo-, hemo- and hydro-thorax) (M). Pneumo-thorax and malposition rates reported (R).	Catheter-related (blood-stream) infection defined, other complications not defined (M). Infection and occlusion rates reported (R).	No classification (M, R).	No classification (M, R).	In summary statistics, postoperative complications counted as not defined.

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Parés ^{4*}	Varicose vein surgery	No definition (M). Not reported (R).	Yes, given with classification of AE (M). There were no major AE, rates for minor AE reported (R).	No classification (M, R).	Yes, according to severity in minor and major (M). Percentages with minor AE reported, no major complications (R).	
Langeveld ^{5*}	Inguinal hernia	No definition (M). Rates for a list of intraoperative complications reported (R).	Definition for wound infection given with classification (M). No definition for other complications. Rates for a list of postoperative AE reported (R).	No classification (M, R).	Yes, for wound infection according to severity in light, medium and severe) (M). Rates of patients with light wound infection reported (R).	In summary statistics, complications counted as not defined.
Daniel-sen ^{6*}	Trocar site wounds in laparoscopic chole-cystectomy	No definition, no invasive intervention (M). Not reported (R).	No definition, indication that wounds were monitored for clinical signs of infections (M). None occurred (R).	No classification (M, R).	No classification (M, R).	
Morris ^{7*}	Asymptomatic primary hyperpara-thyroidism	No definition (M). None occurred, pooled with postoperative adverse events (R).	No definition, indication that patients were checked for bleeding, ability to swallow, nausea and pain prior discharge (M). None occurred, pooled with intraoperative AE (R).	No classification (M, R).	No classification (M, R).	In summary statistics, complications counted as not defined.

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Darai ^{8*}	Colorectal endometriosis	Yes, definition given with classification of adverse events, although this classification originally was published for postoperative complications (M). Number and grading of AE provided (R).	Yes, definition given with classification of AE (M). Number and grading of AE provided (R).	Yes, according to severity in Grade I to V, although this classification originally was published for postoperative complications (M, R).	Yes, according to severity in Grade I to V (M, R).	
Sammour ^{9*}	Laparoscopic colon surgery	No definition (M). Two intraoperative complications reported in the context of reasons for conversion to open surgery (R).	Yes, defined based on a reference (minor/ intermediate/ major septic/ nonseptic), ileus was additionally defined (M). Number and grading of complications provided (R).	No classification (M, R).	Yes, according to severity in Grade I to V (M, R).	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Burdess ^{11*}	Surgery for critical limb ischemia	Yes, definition given for bleeding events with classification (M). Numbers, grading of bleeding complications, pooled with postoperative bleeding events reported (R).	Yes, definition given for bleeding events with classification. Indication that patients were checked for gastrointestinal bleeding, persistent (>3 days) wound leak, hematoma, or infection (M). Numbers, grading and numbers of reoperations for bleeding events reported, pooled with intra-operative bleeding events. Numbers for gastrointestinal bleeding and wound leaks reported. Indication that there were no in-patient deaths, intracranial haemorrhages, incidences of inotrope use or early graft failures (R).	Yes, classification given for bleeding according to severity in major (life-threatening or nonlife-threatening) and minor (M, R).	Yes, classification given for bleeding according to severity in major (life-threatening or nonlife-threatening) and minor (M, R).	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Peeters ^{12*}	Laparoscopic inguinal hernia repair	No definition (M). Numbers of post-operative suction drain in case of intraoperative oozing reported (R).	No definition, indication that patients were followed concerning complications (M). Numbers of haematoma and seroma reported. No wound complications occurred (R).	No classification (M, R).	No classification (M, R).	In summary statistics, intraoperative oozing regarded as reporting of intraoperative complication, since labelled as intraoperative morbidity (R)
Awad ^{13*}	Laparoscopic cholecystectomy	Examples for drink-related side-effects given. No definition for intraoperative AE (M). There were no drink-related complications. Intraoperative complications not reported (R).	Examples for drink-related side-effects given. No definition of postoperative AE (M). Postoperative complications not reported (R).	No classification (M, R).	No classification (M, R).	In summary statistics, complications counted as not reported.

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Jochmans ^{14*}	Kidneys donated after cardiac death	No definition (M). Potential intraoperative complications pooled with postoperative: indication that no vascular complications of the graft were seen and that cardiovascular, gastrointestinal, infectious, metabolic, urinary, and technical complications were comparable between the groups and to literature (data not shown) (R).	No definition (M). Postoperative complications pooled with intraoperative: indication that no vascular complications of the graft were seen and that cardiovascular, gastrointestinal, infectious, metabolic, urinary, and technical complications were comparable between the groups and to literature (data not shown) (R).	No classification (M, R).	No classification (M, R).	
Petrovsky ^{15*}	Major liver surgery	No definition (M). Numbers of intra-operatively transfused patients and central venous pressure reported (R).	Yes, definition given with classification of AE (M). Number and grading of AE provided (R).	No classification (M, R).	Yes, according to severity in Grade I to V (M, R). Grade I, II and IIIa were considered as minor, Grade IIIb, IV and V as major. Additionally, list of complications sorted according to organ system provided.	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Bre-tagnol ^{16*}	Elective rectal cancer surgery	No definition (M). Numbers of intra-operative complications and rates of intra-operative faecal spillage reported (R).	Yes, definition given with classification of AE. Definition for anastomotic leakage and wound abscess given (M). Rates of overall morbidity and mortality and of major morbidity, infectious and non-infectious morbidity as well as of types of infectious and non-infectious morbidity reported (R).	No classification (M, R)	Yes, according to severity in Grade I to V. Grade I and II were considered as minor, Grade III, IV and V as major (M, R).	
Fischer ^{17*}	Pancreatico-duodenectomy	Definition only of intra-operative transfusion, not of other intraoperative complications (M). Number of intraoperative transfusion rates reported (R).	Yes, definition given with classification of AE. In addition, definitions and grading schema for pancreatic leak, fistula, and abscess (M). Rates of overall morbidity and mortality and of Grade 3 morbidity reported (R).	No classification (M, R).	Yes, according to severity in Grade 0 to 5 overall and for pancreatic leak, fistula and abscess in Grade 1 to 5 (M, R).	In summary statistics, intraoperative complications regarded as defined.

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Bernard ^{18*}	Severe traumatic brain injury	No exact definition. Indication that esophageal intubation was recorded as example of complication (M). Numbers of failed and esophageal intubation reported (R).	No definition of postinterventional AE. Extended Glasgow outcome scale (GOSe) defined. Indication that prehospital cardiac arrest was recorded (M). Number of prehospital cardiac arrests and GOSe reported (R).	No classification (M, R).	No classification for postinterventional adverse events. GOSe classification from 1 to 8 (M, R).	GOSe and cardiac arrest not considered as measure of intubation-related complication, since strongly related to brain injury.
Gervaz ^{19*}	Diverticulitis scheduled for elective sigmoidectomy	No definition (M). Only number of patients with small bowel perforation reported not indicating whether this corresponds to an intra-operative complication (R).	No definition. Indication that morbidity, mortality and reoperations were recorded (M). Numbers for a list of major (small bowel perforation, intra-abdominal abscess and small bowel obstruction) and minor complications provided (R).	No classification (M, R).	No classification (M). Classification in major (requiring reoperation) and minor (R).	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Bassi ²⁰	Standard pancreatic resections	No definition (M). Numbers of intraoperatively transfused patients reported (R).	Definitions for haemorrhage, pancreatic and enteric fistula, abdominal collection and abscess, acute pancreatitis and delayed gastric emptying provided. Indication that rates of pulmonary complications, reinterventions and mortality were assessed (M). Rates for these complications as well as for the occurrence of any complication reported, no deaths occurred (R).	No classification (M, R)	Grading of postpancreatotomy haemorrhage, pancreatic fistula and delayed gastric emptying according to severity provided in the papers referred to for definitions of complications (M, R).	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Huang ^{21*}	Small hepatocellular carcinoma	No definition (M). Rates of procedure-related haemorrhage reported (R).	No definition. Indication that patients were followed until disappearance of AE and complications (M). No mortality, rates for various AE (e.g. hepatic failure, bile leakage, postoperative bleeding/procedure-related haemorrhage) reported (R).	No classification (M, R).	No classification (M, R).	
Rogers ^{22*}	Common bile duct stone disease	No definition (M). Pooled with postoperative complications: numbers for minor and major (none major occurred) complications reported as well as that no patient developed bleeding, minor/major not defined (R).	No definition, morbidity (no definition) and mortality listed as secondary outcomes (M). Pooled with intraoperative complications. Numbers for minor and major (none major occurred) complications reported as well as that no patient developed bleeding, pancreatitis, cholangitis, aspiration pneumonia. (R)	No classification (M, R).	No definition (M). Minor/major classification without definition (R).	In summary statistics, postoperative complications regarded as not defined, but classified

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Itani ^{23*}	Ventral incisional hernia	No definition, indication that intraoperative complications were recorded (M). Rates of intraoperative complications (Injury to bowel, problems related to anaesthesia, other and overall) reported (R).	Yes, definition given with classification of AE (M). Rates of short-term and long-term (8weeks) postoperative complications and of serious complications (Sepsis, urinary tract infection, other, overall) within 30 days reported (R).	No pre-specified classification, referenced severity grading refers to postoperative complications (M). Might contribute (pooled with postoperative complications) to the figure reporting numbers of severe/nonsevere complications (R).	Yes, according to severity in Grade I to V (M). According to severity in 2 grades: nonsevere and severe, cutoff unclear, and additionally in serious (R).	In summary statistics, intraoperative complications not considered as classified.

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Futier ²⁵	Major abdominal surgery	No definition. Indication that intraoperative hypovolemia and need for vasoactive support were recorded (M). Mean hypovolemic episodes reported, numbers of patients with vasoactive support and blood transfusion reported (R).	Yes, definition given with classification of AE. Definitions provided for wound infection, for anastomotic leak and intraabdominal perianastomotic abscess, definitions for urinary tract infection, pneumonia, sepsis, acute lung injury, acute respiratory distress syndrome, renal dysfunction, recording information for reinterventions, cardiac failure, pulmonary embolism, pneumothorax, neurological complications, postoperative haemorrhage and death (M).Numbers of patients with any/>1 complication, and complications per individual reported (R).	No classification (M, R).	Yes, according to severity in Grade I to V (M, R).	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Schmittner ^{26*}	Perianal surgery	No definition (M). Number of attempts to place laryngeal mask/for spinal puncture and numbers of patients with allergic reaction and need for intubation reported. Indication that bradycardia and hypotension after disinfection and no respiratory problems occurred. No surgery-related intraoperative complications reported (R).	No definition. Indication that side effects were monitored with a patient questionnaire (M). Numbers of patients with post-operative nausea and vomiting reported and indication that no life-threatening complication occurred. Numbers of patients with a list of side effects reported (R).	No classification (M, R).	No classification (M, R).	In summary statistics, intraoperative complications not considered as reported.
Mekako ^{27*}	Varicose vein surgery	No definition (M). Not reported (R).	Definition and grading of wound complications (M). Number of patients with different grades of wound complications and with new-onset lower leg paraesthesia reported (R).	No classification (M, R).	Classification of wound healing according to severity in 5 groups (M, R).	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Søvik ^{28*}	Superobesity	No definition. Indication that perioperative complications (occurring within 30 days of surgery) were recorded (M). Number for complications reported, without separating intra- and postoperative, e.g. haemorrhage (R).	No definition. Indication that perioperative and late complications were recorded (M). Numbers for complications reported, e.g. haemorrhage, anastomotic leak, leak from duodenal stump, abscess (intraabdominal, cutaneous), total number of complications. No deaths occurred (R).	No classification (M, R).	No classification (M, R).	
Nyström ^{29*}	Haemorrhoid prolapse	No definition (M). Percentages of patients with blood loss >10ml reported. Percentage of patients with complete doughnut after stapled anopexy reported (R).	No definition (M). Types of common complications indicated, rates for major complications (excessive or protracted pain, excessive bowel function disturbance, postoperative bleed, anal stenosis, thrombosed residual haemorrhoid, anal spasm, fissure) recorded (R).	No classification (M, R).	No classification (M). Common and major complications reported (R).	In summary statistics, intraoperative complications not considered as reported.

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Hessman ^{30*}	Primary hyperparathyroidism	No definition (M). Indication that there were no differences in perioperative data. Numbers of patients with bleeding as reason for conversion reported (R).	No definition. Indication that complications were recorded (M). Rates for postoperative haematoma without intervention, wound infection, urinary tract infection, deep vein thrombosis, hypocalcaemia requiring supplementation and unilateral temporary or permanent vocal cord paresis reported (R).	No classification (M, R).	No classification (M, R).	
Wang ^{31*}	Colorectal cancer surgery	No definition (M). Intraoperative blood loss and fluid transfusion reported, intraoperative complications not reported (R).	No definition. Indication that patient-reported discomfort was re-corded (M). Tolerance of intervention and subjective discomfort reported, indication that there were no pulmonary aspirations. No information on postoperative surgical complications (R).	No classification (M, R).	No classification (M, R).	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Subramonia ^{32*}	Great saphenous varicose veins	No definition. Intra-operative complications listed as secondary outcome (M). In the intervention group, rates for unsuccessful tumescent infiltration, great saphenous vein perforation, catheter tip thrombus reported, skin burns (none) reported, in the control group rates for incomplete stripping reported (R).	Postoperative morbidity (pain (VAS-scale and as none, very mild, mild, moderate, severe, very severe), analgesic requirements, sensory abnormalities, wound problems, phlebitis, skin burns, pigmentation) listed as secondary outcome (M). Median (IQR) for pain, analgesic requirements, rates for pain category ratings and under minor complications sensory abnormalities, groin wound problems (mild inflammation, serous wound discharge, haematoma, wound breakdown, all with spontaneous resolution), thigh haematomas, pigmentation reported (R).	No classification (M, R).	No classification (M). Complications listed under minor complications, without definition and without referral to major complications (R).	In summary statistics, complications considered as non-classified.

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Maggiori ^{33*}	Elective sphincter-saving rectal cancer surgery	No definition (M). Number of patients requiring intraoperative blood transfusion reported, intraoperative surgical complications not reported (R).	No definition. Total and daily pelvic suction drain volume, postoperative mortality and morbidity recorded. Indication that anastomotic leakage included clinical leaks (symptomatic, requiring specific medical, surgical or radiological treatment) and abnormalities diagnosed solely by contrast enema (M). Numbers of patients for mortality and morbidity (anastomotic leakage, isolated pelvic abscess, rectal haemorrhage, prolonged ileus, intestinal obstruction, anastomotic stricture, stoma complication, medical morbidity, ≥ 1 complication) reported (R).	No classification (M, R).	No classification (M, R).	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Mathur ^{34*}	Major abdominal surgery	No definition (M). No reporting (R).	Definition of infectious complications occurring within 28 days (clinical sepsis (³ 2 of: temperature >38.5°C, heart rate > 100 beats/min, arterial partial pressure of carbon dioxide <32 mm Hg or respiratory rate >20/min, white cell count >12×10 ⁹ /l or >10% immature forms) or temperature >38°C for 24 h and no alternative source for the fever, and antibiotics were commenced). Drink-associated side effects recorded (M). Percentage of patients with postoperative infectious complications reported. Drink-associated side-effects reported (R).	No classification (M, R).	No classification (M, R).	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Degiuli ³⁵	Gastric cancer	No definition (M). No reporting (R).	No definition, indication that procedure-related morbidity and mortality were recorded (M). Total morbidity and 30d mortality rates reported as well as numbers of post-operative non-surgical and surgical (pancreatic leakage, peritoneal haemorrhage, bleeding from anastomosis, duodenal leakage, colonic perforation, abdominal abscess, intestinal ischaemia, acute pancreatitis, gastric atonia) complications reported (R).	No classification (M, R).	No classification in methods section (M). Classification in surgical and non-surgical, sub-classified according to organ (R).	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra- operative adverse event	Classification of postoperative adverse event	Comments
O'Hare ^{36*}	Foam sclerotherapy for varicose veins	No definition (M). Numbers of patients with foam extravasation, transient visual distur- bances within 30 min of treatment, allergic re-action to the bandages manifested as red and itchy skin, and sensation of a swollen throat for 6h after treatment reported (R).	No definition, examples given: Indication that patients were followed for possible complications such as deep vein thrombosis or new reflux and that pain, residual phlebitis (novel grading system) and skin staining as well as any other possible complications of treatment were recorded (M). Numbers of patients with staining or bruising, bulges due to blood in the treated vein addressed by aspiration, phlebitis, skin discoloration, ulcer and new deep venous reflux reported. No deep vein thrombosis (R).	No classification (M, R).	No classification (M, R).	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Jiang ^{37*}	Gastro-intestinal cancer surgery	No definition (M, R)	Definition of infectious complications with reference (US Centers for Disease Control and Prevention) and of SIRS (criteria of the American College of Chest Physicians/Society of Critical Care Medicine consensus conference definition of 1992). Indication that safety of nutritional support was recorded (M). Numbers for infectious complications (wound, respiratory tract, urological, skin infections and bacteraemia) and SIRS as well as of polyhidrosis and facial blush reported. Indication that there were no serious adverse events (R).	No definition (M, R).	No classification (M, R).	Intervention started postoperatively

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Shepherd ^{38*}	Varicose veins	No definition (M). Report of hypotension secondary to general anaesthesia (R).	No definition. Indication that any complications at 1 and 6 weeks were recorded (M). Numbers of patients with complications reported as major (pulmonary embolus, lymphatic leak) versus minor (wound infection, haematoma, thrombophlebitis, saphenous nerve paraesthesia, skin staining, nausea, hypotension secondary to general anaesthesia, pain) (R).	No classification (M, R).	No pre-specified classification in methods section (M). In results, division in minor and major (R).	In summary statistics, postoperative complications considered as classified according to severity.
Steffen ^{39*}	Thyroid surgery	No definition (M). No report (R)	Definition according to referenced paper (M). Haemorrhage on postoperative day 1 with reoperation reported. Indication that there were no other adverse events or side-effects (R).	No classification (M, R).	No classification (M, R).	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra- operative adverse event	Classification of postoperative adverse event	Comments
Braga ^{40*}	Elective left colonic resection	No definition. Indication how the volume of operative blood loss was calculated and in which cases perioperative transfusion of blood products was undertaken (M). No explicit reporting, only mean operative blood loss and number of transfused patients reported and number of patients with conversion due to bleeding reported (R).	Yes, a distinct list of postoperative compli- cations was defined according to previous literature. Indication that microbiological analysis/culture was used to confirm infectious complications and that follow-up was carried out for 30 days after discharge (M). Number of patients with a list of complications within 30d given as well as long-term morbidity including numbers with incisional hernia, intestinal obstruction, abdominal abscess, urinary dysfunction, peristomal abscess, anastomotic stricture including numbers of late readmi- ssion/reoperation (R).	No classification (M, R).	No classification (M), classification according to organ and timepoint (R).	For summary statistic regarded as classified according to organ

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra- operative adverse event	Classification of postoperative adverse event	Comments
Carli ^{41*}	Colorectal surgery	No definition (M). Not reported (R).	Definitions given with classification (M). Numbers of patients with no, grade 1, 2 and ³ 3 complications reported (R).	No classification (M, R).	Classification according to severity.	
Hendry ^{42*}	Liver resection	Definition according to literature reference (M). Reporting pooled with postoperative complications (30d morbidity and mortality) (R).	Definition according to literature reference (M). Reporting pooled with intraoperative compli- cations (30d morbidity and mortality). Numbers of patients with reoperation for haemo- rrhage, readmission for intraabdominal collec- tion, severe constipation, bile leak and pneumonia and of deaths from myocar- dial infarction reported (R).	Classification according to type (cardiorespiratory, surgical, infectious).	Classification according to type (cardiorespiratory, surgical, infectious).	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Bevis ^{44*}	Open abdominal aortic aneurysm surgery	No definition (M). No reporting (R).	No definition. Indication that postoperative complication rates were recorded (M). Number of perioperative deaths (unrelated to intervention) reported, reasons indicated (cardiac complication, intra-abdominal bleeding within 24h after surgery) and numbers of wound infections, small bowel obstruction requiring relaparotomy, conservatively treated mesh seroma and incisional hernia (R).	No classification (M, R).	No classification (M, R).	

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Reid ^{46*}	Closure of ileostomy wounds	No definition (M). Indication that there were no intraoperative complications (R).	Surgical site infection (ssi) was defined according to the Centers for Disease Control (CDC) providing a reference paper (M). Indication that there were no anastomotic complications. Indication that postoperative complications were similar in both groups. Numbers with ssi, superficial wound dehiscence, drainage, intravenous/oral antibiotics and community nurse visits reported. Indication that there were no ssi-related reoperations and no readmissions (R).	No classification (M, R).	Surgical site infection classified as superficial or deep or organ/space infection. Other complications not classified (M, R).	For summary statistics complications regarded as classified.

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra- operative adverse event	Classification of postoperative adverse event	Comments
Nthumba 47*	Rub for surgical hand preparation in rural hospital in Kenya	No definition (M). No reporting (R).	Surgical site infection was defined referring to the Centers for Disease Control and Prevention (CDC) providing a refe- rence paper. Tolerability and side-effects of skin preparation (M). Numbers with super-ficial, deep, organ/space infection reported, no skin reactions (R).	No classification (M, R)	Surgical site infection classified as superficial or deep or organ/space infection (M, R).	For summary statistics complications regarded as classified.

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra- operative adverse event	Classification of postoperative adverse event	Comments
Jayne ^{49*} and ^{49*add1}	Colorectal cancer	No definition. Indication that complications during surgery were recorded (M). Numbers provided for clinically significant haemorrhage, cardiac/pulmonary insufficiency, bowel/ureteric/vessel/bladder injury (R).	No definition (M). Indication that complications were recorded at 30d and 3 months (M). At 30d, numbers provided for wound/chest infection, anastomotic dehiscence, deep vein thrombosis. At 3 months, numbers provided for major and minor with indication that most frequently intestinal obstruction and wound infection occurred. In-hospital mortality mentioned (R).	No classification (M, R).	Classification in major and minor (R).	^{49*add1} considered for this information

Table S5: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra- operative adverse event	Classification of postoperative adverse event	Comments
Chalmers ^{50*}	Polytetra- fluoroethylene arterial anastomoses	No exact definition of intraoperative complications. Persisting bleeding at the anastomosis requiring adjunctive haemostatic measures as determined by the surgeon defined as eligibility criterion. Bleed- ing at the anastomosis at 4, 7 and 10min defined as efficacy criterion (M). Achievement of haemo- stasis reported, no intra- operative complications except from number of patients fulfilling inclusion criteria reported (R).	AE defined and classi-fied according to the Medical Dictionary for Regulatory Activities (MedDRAÒ) codes. Indication that AE potentially related to bleeding were recorded at 5 weeks (M). Number of patients with AE (nausea, anaemia, cardiac failure, hypotension, constipation, graft infection, graft occlusion/ thrombosis, peripheral oedema, urinary tract infection) and complications potentially related to bleeding (anaemia/low haemo- globin/low haematocrit, haematoma, bleeding, increased sanguinous drainage, seroma and bruising) reported (R).	No classification (M, R).	Medical Dictionary for Regulatory Activities (MedDRAÒ) codes, System Organ Class being the highest level of the hierarchical terminology, distinguishing by anatomical or physiological system, etiology, or purpose (M, R).	In summary statistics, intraoperative complications considered as defined, taking the eligibility criterion as definition.

Table S52: Adverse event reporting (continued)

First Author	Condition/ Population	Definition and occurrence of intraoperative adverse event	Definition and occurrence of postoperative adverse event	Classification of intra-operative adverse event	Classification of postoperative adverse event	Comments
Watanabe ^{51*}	Colonic cancer resection	No definition (M). Median operative blood loss indicated and notion that no problems with the stapling device occurred and that the frequency of spillage of bowel content during surgery did not differ between groups (R).	Definition of surgical site infections (ssi) with reference. Indication that daily records of postoperative course were obtained, and infectious complications were recorded for up to 30 days after surgery. Numbers of patients with complications (ssi or paralytic ileus) reported (R).	No classification (M, R)	Classification of surgical site infections according to severity in superficial incisional, deep incisional and organ/space (M, R).	For summary statistics complications regarded as classified.

Appendix Table S6: Classification of postoperative complications(Dindo, Demartines, & Clavien 2004)

Grade	Definition
Grade I	Any deviation from the normal postoperative course without the need for pharmacological treatment or surgical, endoscopic, and radiological interventions. Allowed therapeutic regimens are: drugs as antiemetics, antipyretics, analgetics, diuretics, electrolytes, and physiotherapy. This grade also includes wound infections opened at the bedside
Grade II	Requiring pharmacological treatment with drugs other than such allowed for grade I complications. Blood transfusions and total parenteral nutrition are also included
Grade III	Requiring surgical, endoscopic or radiological intervention
- Grade IIIa	- Intervention not under general anesthesia
- Grade IIIb	- Intervention under general anesthesia
Grade IV	Life-threatening complication (including CNS ¹ complications) ² requiring IC ³ /ICU ⁴ management
- Grade IVa	- Single organ dysfunction (including dialysis)
- Grade IVb	- Multiorgan dysfunction
Grade V	Death of a patient
Suffix "d"	If the patient suffers from a complication at the time of discharge, the suffix "d" (for "disability") is added to the respective grade of complication. This label indicates the need for a follow-up to fully evaluate the complication.

¹ CNS= Central nervous system

² Brain hemorrhage, ischemic stroke, subarachnoidal bleeding, but excluding transient ischemic attacks.

³ IC= Intermediate care

⁴ ICU= Intensive care unit

Appendix Table S7: Definition of Surgical Site Infection (SSI)(Mangram, Horan, Pearson, Silver, & Jarvis 1999b)

Type	Definition
<p>Superficial Incisional SSI</p>	<p>Infection occurs within 30 days after the operation and infection involves only skin or subcutaneous tissue of the incision and at least one of the following:</p> <ol style="list-style-type: none"> 1. Purulent drainage, with or without laboratory confirmation, from the superficial incision. 2. Organisms isolated from an aseptically obtained culture of fluid or tissue from the superficial incision. 3. At least one of the following signs or symptoms of infection: pain or tenderness, localized swelling, redness, or heat and superficial incision is deliberately opened by surgeon, unless incision is culture-negative. 4. Diagnosis of superficial incisional SSI by the surgeon or attending physician. <p>Do not report the following conditions as SSI:</p> <ol style="list-style-type: none"> 1. Stitch abscess (minimal inflammation and discharge confined to the points of suture penetration). 2. Infection of an episiotomy or newborn circumcision site. 3. Infected burn wound. 4. Incisional SSI that extends into the fascial and muscle layers (see deep incisional SSI). <p>Note: Specific criteria are used for identifying infected episiotomy and circumcision sites and burn wounds.</p>
<p>Deep incisional SSI</p>	<p>Infection occurs within 30 days after the operation if no implant¹ is left in place or within 1 year if implant is in place and the infection appears to be related to the operation and infection involves deep soft tissues (e.g., fascial and muscle layers) of the incision and at least one of the following:</p> <ol style="list-style-type: none"> 1. Purulent drainage from the deep incision but not from the organ/space component of the surgical site. 2. A deep incision spontaneously dehisces or is deliberately opened by a surgeon when the patient has at least one of the following signs or symptoms: fever (>38°C), localized pain, or tenderness, unless site is culture-negative. 3. An abscess or other evidence of infection involving the deep incision is found on direct examination, during reoperation, or by histopathologic or radiologic examination. 4. Diagnosis of a deep incisional SSI by a surgeon or attending physician.

¹ National Nosocomial Infection Surveillance definition: a nonhuman-derived implantable foreign body (e.g., prosthetic heart valve, nonhuman vascular graft, mechanical heart, or hip prosthesis) that is permanently placed in a patient during surgery.

Appendix Table S7: Definition of Surgical Site Infection (continued)

Type	Definition
Deep incisional SSI (continued)	<p>Notes:</p> <ol style="list-style-type: none"> 1. Report infection that involves both superficial and deep incision sites as deep incisional SSI. 2. Report an organ/space SSI that drains through the incision as a deep incisional SSI.
Organ/space SSI	<p>Infection occurs within 30 days after the operation if no implant¹ is left in place or within 1 year if implant is in place and the infection appears to be related to the operation and infection involves any part of the anatomy (e.g., organs or spaces), other than the incision, which was opened or manipulated during an operation and at least one of the following:</p> <ol style="list-style-type: none"> 1. Purulent drainage from a drain that is placed through a stab wound¹ into the organ/space. 2. Organisms isolated from an aseptically obtained culture of fluid or tissue in the organ/space. 3. An abscess or other evidence of infection involving the organ/space that is found on direct examination, during reoperation, or by histopathologic or radiologic examination. 4. Diagnosis of an organ/space SSI by a surgeon or attending physician.

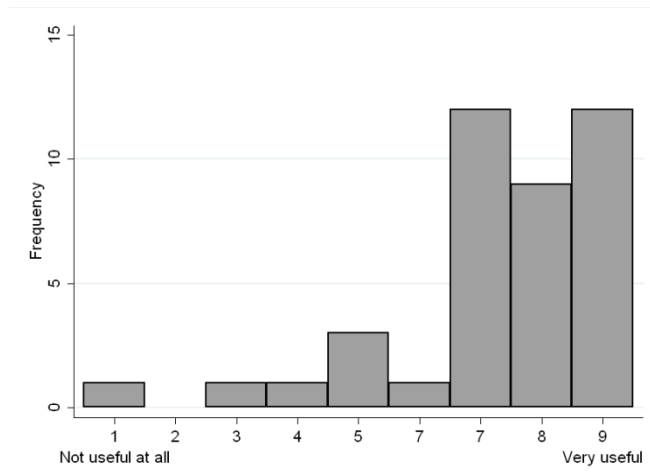
¹ If the area around a stab wound becomes infected, it is not an SSI. It is considered a skin or soft tissue infection, depending on its depth.

Appendix Text (including illustrations) S1: Questions and summary results round 1, questions round 2

1. Overall usefulness of a classification of intraoperative complications

Question round 1

How do you rate the overall usefulness of a classification of intraoperative complications (1=not useful at all, 9=very useful)?



Summary results

The experts seemed to be in favour of a classification (which is as well supported by the tick box answers), although they thought that it is a challenge to find a uniform system. One expert suggested the use of CTCAE (Common Terminology Criteria for Adverse Events)(U.S.Department of Health and Human Services 2014), which could be well-applicable to postoperative complications, however might lead to an upgrading of events when used for intraoperative complications, since many of these will be resolved surgically and thus graded III or higher. We are well aware that most importantly complications should be avoided and that confounding factors associated to the patient and the surgeon should be taken into account.

We think that a classification system for intraoperative complications will allow for standardized reporting of intraoperative events and deviations from the ideal intraoperative course. This is important in the following context:

- Standardization of reporting of evaluations of new surgical techniques and devices
- Standardization of reporting in educational and training settings
- Capture of intraoperative events with increased risk of postoperative complications for a critical incidents reporting system
- Enhanced patient handover quality after surgery concerning anticipated problems.

2. Definition of intraoperative complications of intraoperative complications

2.1. Distinction

Questions round 1

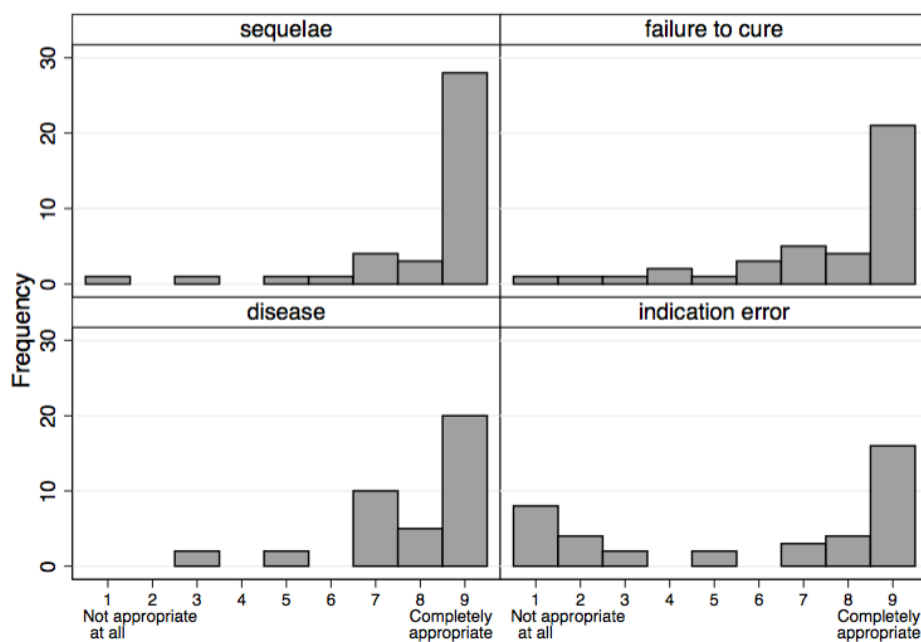
Please rate on a scale between 1 and 9 the appropriateness of including the following items in a definition of intraoperative complication (1=not appropriate at all, 9= completely appropriate).

Sequelae: A definition of intraoperative complications needs to distinguish complications from sequelae, i.e. effects inherent to the surgery (such as the inability to walk after an amputation of a leg)

Nocure: A definition of intraoperative complications needs to distinguish complications from failures of cure (such as residual tumor after surgery)

Disease: A definition of intraoperative complications needs to distinguish complications from events related to the underlying disease (such as bleeding from a ruptured abdominal aortic aneurysm)

Indication: A definition of intraoperative complications should NOT refer to errors in indication (such as wrong-side surgery)



Summary results sequelae

Most experts highly agreed with this distinction (tick box answers), however suggested a more precise description to distinguish effects inherent to the “index surgery” as initially planned for the respective pathology (e.g. aimed life-saving leg amputation) versus the same intervention to treat a complication (“consecutive surgery” e.g. leg amputation after an infected femoro-popliteal graft).

Additionally, it was pointed out that not every intraoperative complication leads to a long-term disability, to which we entirely agree. The developed system is however intended to capture both immediate and long-term consequences.

The differentiation between an effect inherent to the intervention and a suboptimal result, e.g. in trauma surgery, may not be clear cut, especially with differences in surgeons' expertise. We suggest considering those effects as sequelae which are evident (i.e. lack of joint function after arthrodesis) or which are inherent given optimal treatment according to current guidelines.

We are aware that after leg amputation a patient is still able to walk with crutches or a prosthesis, yet think that this example is intuitively illustrative.

Question round 2

Based on these answers, we kindly ask you to rate the following updated definition (1= not at all appropriate, 9 = completely appropriate):

A definition of intraoperative complications needs to distinguish complications from sequelae, i.e. effects inherent to current best practice surgery (such as the inability to walk after amputation of a leg). It should refer to the index surgery and not to potential follow-up surgeries to treat a complication.

Summary results nocure

Again, most experts seemed to agree with this distinction (tick box answers, see above), but pointed out the importance of expertise in the field.

We agree with the experts that a failure of planning a complex intervention should be accounted for, however we think this should be rather accounted for as misjudgement in indication than as intraoperative event (although this may actually become evident only intraoperatively).

If a patient is judged as inoperable due to an advanced tumor stage, we would not apply the classification, since he/she has not been operated on. Therefore, we cannot compare failure of cure rates between more or less "aggressive" surgeons.

Question round 2

Based on these answers, we kindly ask you to rate the following updated definition (1= not at all appropriate, 9 = completely appropriate):

A definition of intraoperative complications needs to distinguish complications from failures of cure (such as residual tumor after surgery), under the condition the indication for surgery and the interventions conform to current guidelines.

Summary results disease

Experts seemed to agree that this distinction should be made (tick box answers, see above). Again, failures in indication were pointed out to be relevant as well, to which we entirely agree. As mentioned above, we suggest rather accounting for these instances as misjudgement in indication than as intraoperative event.

Additionally, it was suggested to refer to optimal treatment, since some of these events potentially could have been prevented given optimal treatment.

Question round 2

Based on these answers, we kindly ask you to rate the following updated definition (1= not at all appropriate, 9 = completely appropriate):

A definition of intraoperative complications needs to distinguish complications from events related to the underlying disease (such as bleeding from a ruptured abdominal aortic aneurysm or intraoperative sepsis due to a purulent peritonitis associated to a perforated appendicitis), under the condition the indication for surgery and the interventions conform to current guidelines.

Summary results indication

Here, the expert opinions were contradictory. While some experts felt that indication belongs to the surgical procedure itself and thus wrong-site surgery is an intraoperative complication, others felt that errors in indication do not belong to the intraoperative phase. Moreover, it was suggested to distinguish between errors such as wrong site surgery and other errors in indication. We are well aware that indication is a crucial step of surgery; we propose however to limit this classification to the “gesture” of surgery itself and to describe errors in indication separately and not as intraoperative complications.

Question round 2

Based on these answers, we kindly ask you to rate the following updated definition (1= not at all appropriate, 9 = completely appropriate):

A definition of intraoperative complications should NOT refer to wrong-side or wrong-patient surgery and NOT to errors in indication (such as inappropriate indication for surgery according to current guidelines).

2.2. Wording

Questions round 1

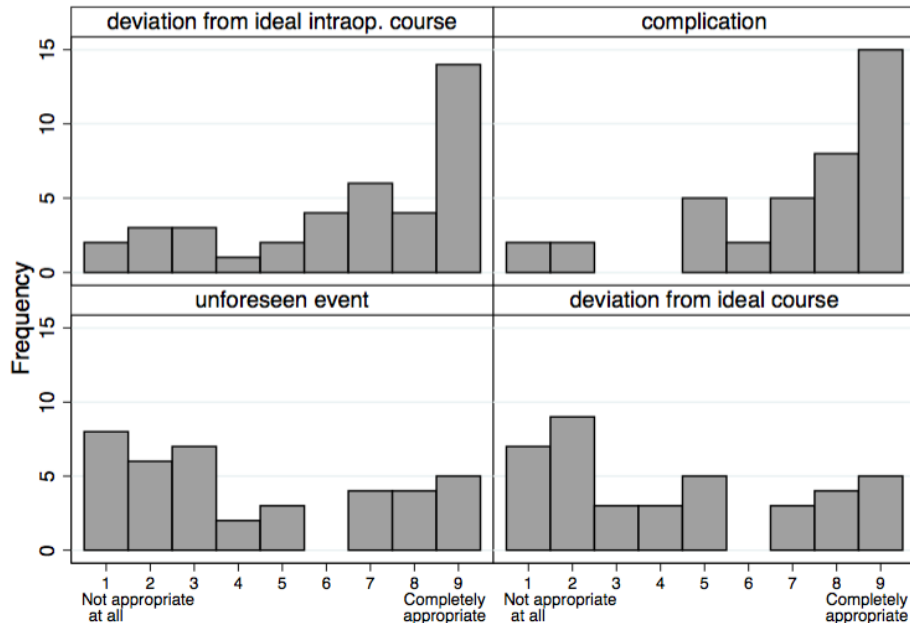
Please rate on a scale between 1 and 9 the appropriateness of including the following items in a definition of intraoperative complication (1=not appropriate at all, 9= completely appropriate).

Deviation from ideal intraop. course: A definition of intraoperative complications may be formulated as ‘any deviation from the ideal intraoperative course’

Complication: The definition (and classification) should contain the word ‘complication’ (see as well next two questions)

Unforeseen event: The definition (and classification) should contain the word ‘unforeseen event’ instead of ‘complication’

Deviation from ideal course: The definition (and classification) should contain the word ‘deviation from the ideal course’ instead of ‘complication’



Summary deviation from ideal intraoperative course

Although quite some of the experts felt that this definition was completely appropriate, many disagreed. The rationale was that it may be difficult to define what is ideal and that not every deviation from the ideal course is a complication, for instance in case of an unexpected anatomical difficulty found upon surgery. Moreover, some minor deviations such as a thread rupture upon knot tying that is corrected may be a deviation, but without any consequence. Else an intervention taking longer than planned may not actually be a complication, if not patient-relevant (disregarding cost). We agree with the issue that many complications may not be patient-relevant, but propose to keep the definition as such, however to grade these events differently according to their consequences and severity. Additionally, it was suggested to limit the definition to factors that are avoidable such as avoidable bleeding.

Summary complication:

The word complication was more uniformly accepted, since it is less subject to interpretation. It was however suggested to add defined qualifiers.

Summary unforeseen event

Regarding the tick box answers, this terminology was rather not accepted, although some felt unforeseen event and complication rather to be synonyms, whereas others disagreed on that issue. It was pointed out that such a definition would as well include instances not necessarily related to the operative procedure, such as on table myocardial infarction. As a matter of fact, many definitions for postoperative complications include both surgical and medical complications;

therefore this principle could apply as well to intraoperative complications. Additionally, it was pointed out that such a definition would be the most neutral encouraging surgeons to record it, while the term complication would be avoided, since it might be associated with legal problems. It was also pointed out that many complications are predictable and included in the informed consent process, although they are definitively unwanted.

Summary deviation from the ideal course

Regarding the tick box answers, reviewing this question after considering different other options seems not to be in favour of this wording. Once more, it was pointed out that such a deviation may be well due to the underlying disease and not to the operation itself, while others felt that this could be a definition or description of or for “complication”. Additionally, it was pointed out that not every undesired event is actually preventable.

Question round 2

Based on these answers, we kindly ask you to rate the following updated definition (1= not at all appropriate, 9 = completely appropriate):

A definition of intraoperative complications may be formulated as ‘any preventable deviation from the ideal intraoperative course’

2.3. Time period

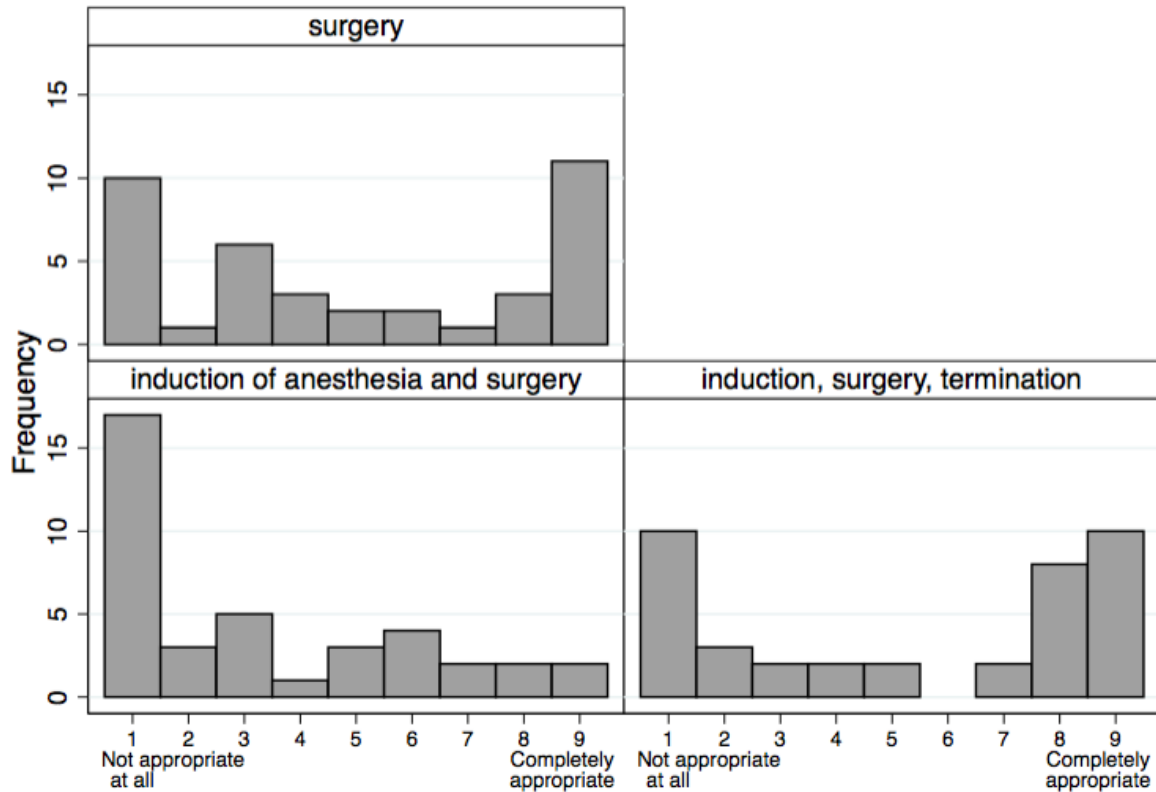
Questions round 1

Please rate on a scale between 1 and 9 the appropriateness of including the following items in a definition of intraoperative complication (1=not appropriate at all, 9= completely appropriate).

Surgery: A definition of intraoperative complications should include the following time period: Surgery only (i.e. between skin incision and skin closure) (see as well next two questions)

Induction of anesthesia and surgery: A definition of intraoperative complications should include the following time period: Induction of anesthesia and surgery

Induction, surgery, termination: A definition of intraoperative complications should include the following time period: Induction of anesthesia, surgery and termination of anesthesia (defined as end of continuous presence of a member of the anesthesiology team (for example transfer to the recovery room or transfer to the ICU)



Summary surgery

Here, the results were conflicting, some strongly suggested to include the period between skin incision and closure only (however including anesthesia-related complications during this period), whereas others focussed on the team performance including anesthesia. Others preferred to decide individually, i.e. if the focus was to compare two surgical techniques, not to include anesthesia, as opposed to when special anesthesia techniques were investigated. Some commented that positioning on the table belongs to surgery, while others disagreed. It was as well pointed out that some intraoperative complications may become evident only postoperatively and in that case the follow-up should even be longer (this was discussed in a separate question).

Summary induction of anesthesia and surgery

When specifically asked, if induction should be included, most disagreed (tick box answers), which was well reflected by the free text comments. Some suggested calling these perioperative complications, whereas others would call them anesthesia-related complications. Here again, it was stressed that complications that become evident only later, but have their origin in the surgical procedure, should be included, such as an intraabdominal abscess secondary to inappropriate intraoperative handling of a bowel injury.

Summary induction, surgery, termination

Again, both the tick box answers as well as the comments suggested quite contradictory views. While some strongly advocated limiting the period to the time

between skin incision and closure (e.g. not to include a tubus-related, central line or arrhythmia problem upon induction), others thought it to be mandatory to overview the whole process of surgical treatment including both induction and anesthesia (which the former then would rather call perioperative or periprocedural complication than intraoperative complication). It was pointed out that if induction was included, termination consequently should also be included.

Summary time period overall

It was recognized that the issue is controversial and each definition has advantages and disadvantages, most frequently it was suggested to distinguish between intraoperative complications (regardless when they become manifest) and perioperative complications (the latter including induction and termination). As major drawback it was pointed out that including induction and termination would render such a classification more complex and potentially less useful to surgeons.

Questions round 2

Based on these answers, we kindly ask you to rate the following updated definition (1= not at all appropriate, 9 = completely appropriate):

A definition of intraoperative complications should include the following time period of event occurrence (disregarding time point of event manifestation): Surgery only (i.e. between skin incision and skin closure) and should be distinguished from perioperative complications including anesthesia induction and termination.

It additionally seems important to clarify, what sort of complications should be captured, i.e. both “surgical” and “medical” (as it is the case in the Clavien-Dindo complication classification (Dindo, Demartines, & Clavien 2004)) or only “surgical”. In our opinion, making a distinction between surgical and medical is not always obvious; therefore it seems more practical to include both. If for instance a patient shows signs of cardiac ischemia during abdominal surgery, this could be due to an underlying coronary disease or this could be in the context of severe bleeding during surgery or both. Similarly, intraoperative death would be recorded regardless the origin.

Therefore, please rate the following question (1= not at all appropriate, 9 = completely appropriate):

A definition of intraoperative complications should include any event occurring during surgery, regardless whether it is related to the surgical gesture itself or not.

3. Classification of intraoperative complications

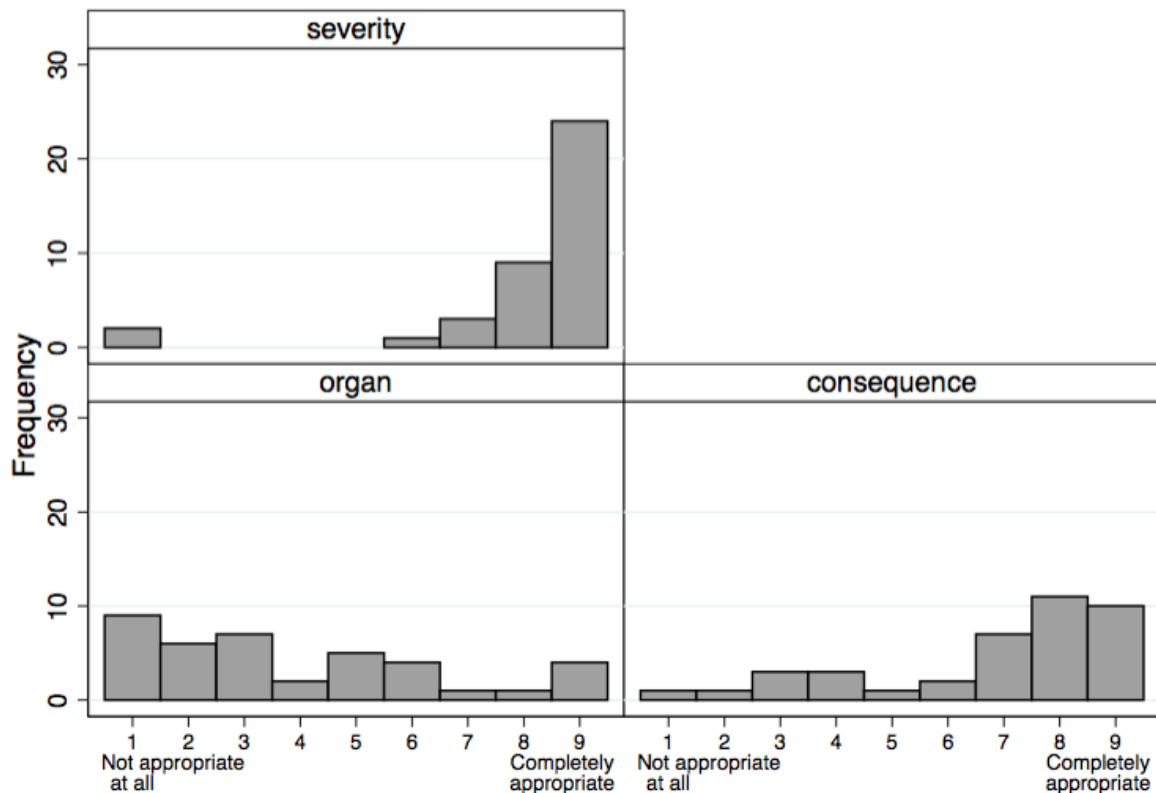
3.1. Type

Questions round 1

Severity: A classification of intraoperative complications should be undertaken according to severity (see as well the next question)

Organ: A classification of intraoperative complications should be undertaken according to the injured organ regardless of severity

Consequence: If a classification of intraoperative complications is undertaken according to severity, shall the classification be undertaken according to the resulting treatment (similar to the classification of postoperative complications according to Clavien-Dindo (Dindo, Demartines, & Clavien 2004))?



Summary severity

Here, the tick box answers were in favour of a classification according to severity. It was pointed out that the severity should take into account the potential resulting morbidity, since for instance a millimetric injury to the duodenum could be much more dangerous than to the ileum or colon.

One expert suggested having two surgeons classify events. We believe that this is a very important point, although sometimes there are not two surgeons available, but for instance one surgeon and one student or junior resident or sometimes the assistant may not be the same throughout the intervention. Moreover, in strongly hierarchical settings it might be difficult to find consensus. In an ideal setting, for instance in a clinical research setting, one could have two blinded raters who independently carry out the classification, whereas in every day practice, this might be more difficult.

Question round 2

Based on these answers, we kindly ask you to rate the following:
(1= not at all appropriate, 9 = completely appropriate):

A classification of intraoperative complications should be undertaken by two independent surgeons.

Summary organ

The experts were not in favour of a classification according to the injured organ, as supported as well by the free text options favouring a patient-outcome centred approach with optimal generalizability.

Summary consequence

Whereas the tick box answers seemed rather in favour of a classification according to the resulting treatment, because the degree of invasiveness seems to go along with the degree of severity, it was advocated to consider as well the patient outcomes such as sequelae and quality of life. Additionally, it was suggested to distinguish between minor versus major complications as well as between complications with transient versus permanent consequences.

Question round 2

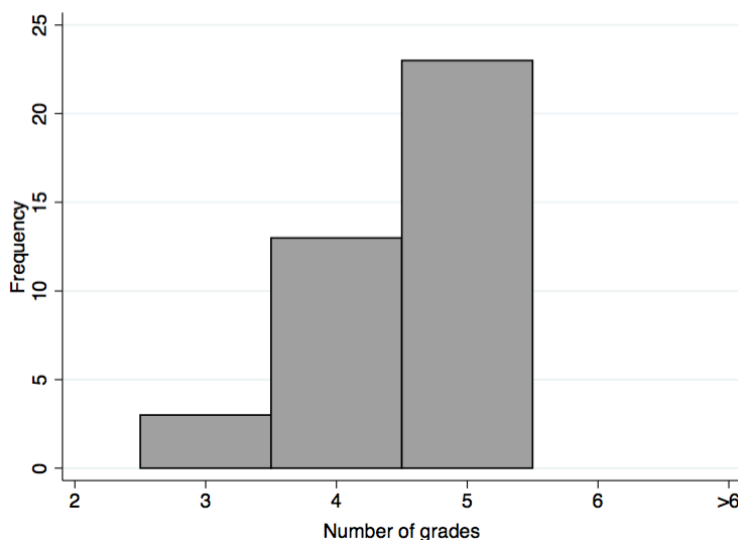
Based on these answers, we kindly ask you to rate the following updated definition
(1= not at all appropriate, 9 = completely appropriate):

A classification of intraoperative complications should consider the severity of the complication and take into account both the necessary treatment and permanent “sequelae” (i.e. disability) resulting from the complication.

3.2. Grades

Question round 1

How many grades should a classification of intraoperative complications according to severity include (see also the next question)?

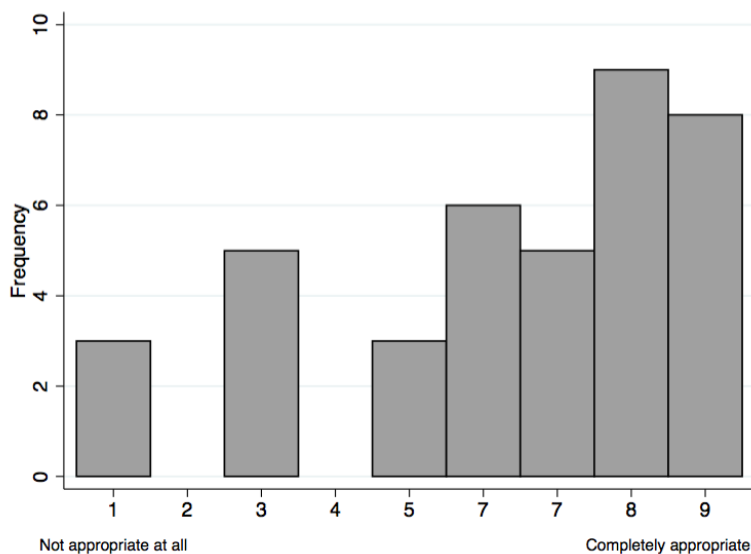


Summary grades

The tick-box answers were in favour of 4 or 5 degrees, although we were well aware that it was difficult to decide on this issue at that point in time. Some suggested to keep it as simple as possible, while others felt that less than four would not be enough differentiated. It was suggested to align the number with the one in the postoperative complication scoring. Additionally, it was referred to the CTCAE (U.S.Department of Health and Human Services 2014), which foresee five degrees of severity, however allow for some not to be applicable depending on the type of adverse event.

Question round 1

Should sub-classifications be allowed in a classification of intraoperative complications according to severity?



Summary subgrades:

The inclusion of sub-classifications was rather preferred, however only if adding important information and if not rendering the classification too complex.

3.3. Time

Questions round 1

Directly after surgery: Should it be possible to definitively fill in a classification of intraoperative complications directly after surgery, i.e. only intraoperative, but not postoperative outcomes are considered (see also next 3 questions)?

Short-term: Should information collected during the short-term postoperative course (up to 30 days postoperatively) be accounted for in the classification of intraoperative complications?

Example: Should intraoperative complications (e.g. thermic small bowel injury with perforation or foreign body left in place) that are not recognized intraoperatively, but only postoperatively, be taken into account in the classification of

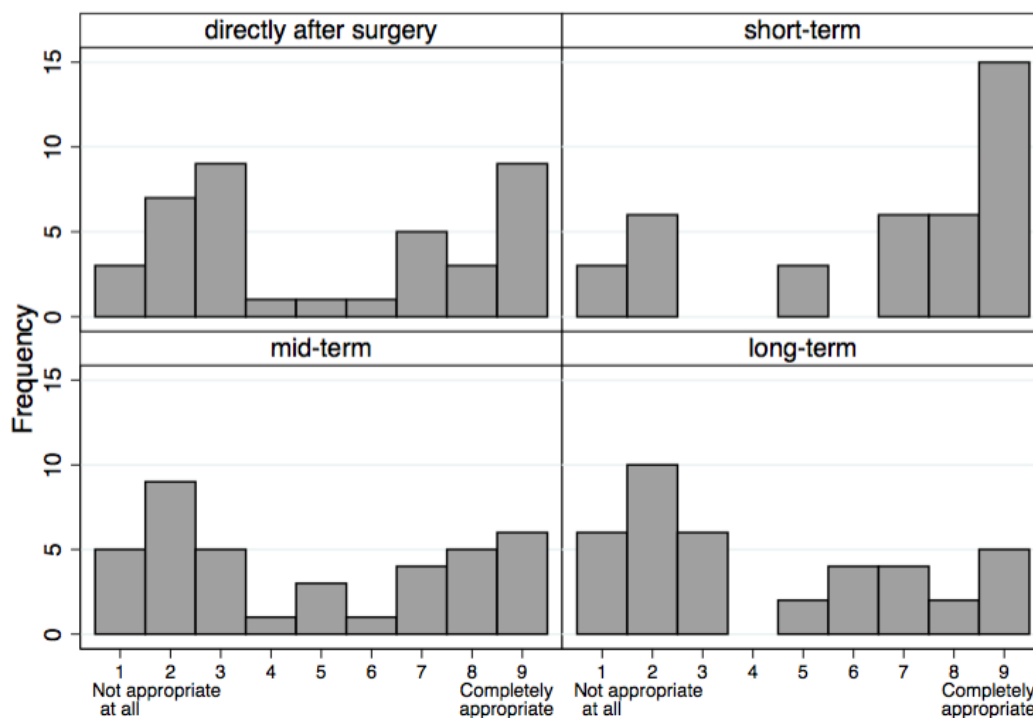
intraoperative complications (consequently the classification may not be definitively filled in immediately postoperatively)?

Mid-term: Should information collected during the mid-term postoperative course (up to 6 months postoperatively) be accounted for in the classification of intraoperative complications?

Example: Should permanent organ dysfunction (e.g. brain, kidney) resulting from severe hypotension due to intraoperative bleeding be accounted for in the classification of intraoperative complications? Should hoarseness after thyroidectomy be followed up to distinguish transient from permanent recurrent nerve injury? Consequently, the classification may not be definitively filled in immediately postoperatively.

Long-term: Should information collected during the long-term postoperative course (>6 months postoperatively) be accounted for in the classification of intraoperative complications?

Example: Should permanent organ dysfunction (e.g. liver) resulting from common bile duct dissection be accounted for in the classification of intraoperative complications? Consequently the classification may not be definitively filled in immediately postoperatively.



Summary directly after surgery

The tick box answers reflect the free-text answers suggesting that some intraoperative complications are recognized only later. It was suggested to fill in the classification immediately after surgery and thus including it in the report of the intervention, however leaving an option to “correct” the classification, in case a complication during surgery becomes evident later. It was however also warned

that an extension of the time-period considered may lead to a confusion with classification systems for postoperative complications, such as the Clavien-Dindo system (Dindo, Demartines, & Clavien 2004).

Summary short term

The tick box answers seemed rather in favour of an inclusion of the short-term period, whereas the free-text answers again pointed out the advantages and disadvantages of the two approaches. Whereas it would be more accurate to include an option of “updating” the classification later (in analogy to the TNM system), feasibility might be hampered and thus complications becoming evident later could be handled with a classification of postoperative complications. It was pointed out that if a classification at a later point in time shall be allowed, initial classification should in any case take place as well to avoid a potential recall bias. Moreover, it was suggested to take into account that a differentiation between intraoperative complication becoming evident only later and a postoperative complication independent from the initial surgery might sometimes be very challenging, not always allowing causal inference.

Summary mid-term

Here, experts seemed to be slightly more in favour of not including such a long follow-up period. One expert pointed out to end at 3 months instead of 6 months, whereas others felt that more than 30 days is not realistic. It was again argued that causal inference is not clear, for instance hoarseness after thyroid surgery may be due to oedema or an injury. Moreover, permanent disability could be captured with a classification for postoperative complications. On the other hand, it was stressed that some complications becoming evident only postoperatively may be very clearly attributed to the initial surgery (e.g. foreign body left in place) and thus should be captured with the classification of intraoperative complications.

Summary long-term

Here, again experts rather felt that this might be too long, whereas others felt that functional long-term results are important. It was referred to the APGAR score as example of a score that may be immediately filled in regardless long-term outcome, although it is not strictly comparable, because all the information relevant to fill in the APGAR score is immediately available and does not become apparent only days thereafter, but we agree that the score does not take into account long-term outcomes, even though it may be associated with them.

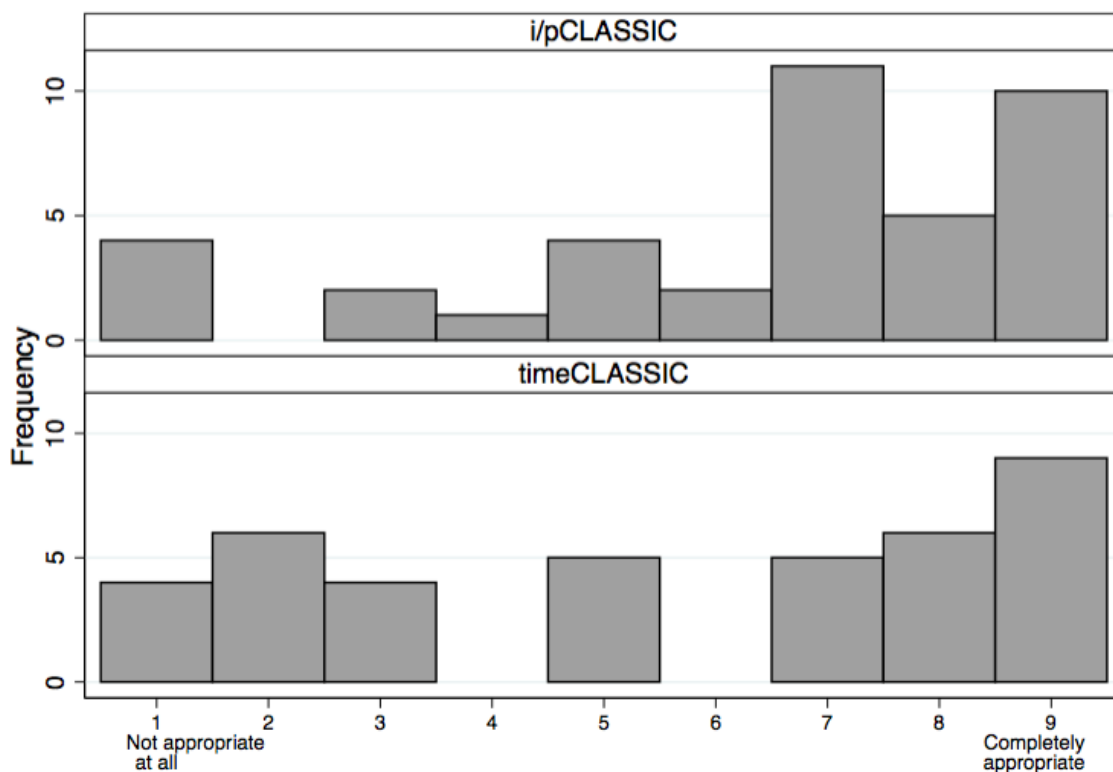
Comment and questions round 1

In order to address the issue of different follow-up times, we proposed to include a prefix indicating the follow-up time: iCLASSIC for filling in the classification immediately after surgery and accounting for the intraoperative course only versus pCLASSIC for filling in the classification accounting for the postoperative course.

To include the exact time-point of assessment, one could imagine replacing the i and p by the time-point, e.g. 0CLASSIC for immediately after surgery, 30dCLASSIC for 30 days, and 6mCLASSIC for 6 months after surgery. Repeated measurements could be taken into account per patient.

*i/p*CLASSIC: Should the classification allow for differentiation of the timepoint of assessment of intraoperative complications by a prefix distinguishing between *i*CLASSIC for immediately after surgery and *p*CLASSIC accounting for the postoperative course (see as well the next question)?

timeCLASSIC: Should such a prefix include more precise information about the duration of follow-up, e.g. 30dCLASSIC for a follow-up up to 30 days postoperatively?



Summary *i/p*CLASSIC

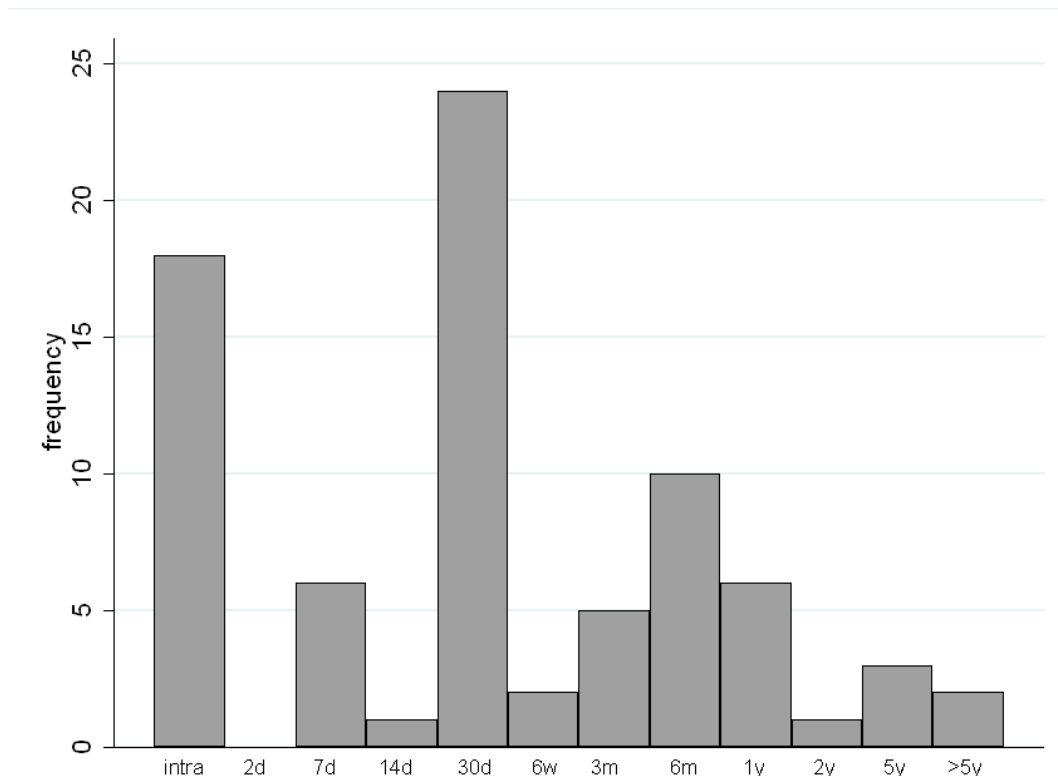
Experts were rather in favour for such a distinction, although comments pointed out again the difficult distinction between *p*CLASSIC and any system of classification of postoperative complications (with the only difference being the cause, which anyhow sometimes would be difficult to evaluate). Moreover, it was doubted that classifications evaluated at different time-points allow for comparisons. Clearly, in our opinion, if we allow for immediate and later classification, everyone should fill in the *i*CLASSIC and *p*CLASSIC would be an add-on.

Summary timeCLASSIC

Here, quite some felt that this is inappropriate, as reflected by additional comments, since it is too complex, not everyone might choose the same time-point hampering comparison and it would end-up in different grades over time. One expert suggested taking into account only the highest grade to correct over time.

Question round 1

For which follow-up periods should intraoperative complications be assessed (several options possible)?



Summary

Whereas the tick box answers suggested the inclusion of the time of surgery and the 30d follow-up, the free-text answers were more in favour of only including the intraoperative time-period, for the reasons mentioned above. It was additionally suggested that this may not be generalized, since every surgical branch might be different.

Question round 2

Based on these answers, we kindly ask you to rate the following updated definition (1= not at all appropriate, 9 = completely appropriate):

Should it be possible to definitively fill in a classification of intraoperative complications directly after surgery, whereas complications apparent only after surgery are handled with a classification system for postoperative complications?

3.4. Classification

Proposed classification and question round 1

The following questions refer to the initially proposed classification of intraoperative complications below:

Table: Proposed Classification of intraoperative complications (CLASSIC)

Please note that the classification could be denoted with a prefix indicating the time-point of assessment. Examples indicated mostly refer to abdominal surgery; however, the classification is intended to be used in any surgical discipline.

IC=intermediate care; ICU=intensive care unit.

Grade	Definition	Examples
Grade I	Any deviation from the ideal intraoperative course <ul style="list-style-type: none"> Without the need for any additional treatment or intervention 	<ul style="list-style-type: none"> Bleeding: Self-limiting bleeding at trocar site Injury: Electrocautery injury to parietal peritoneum Non-surgical: self-limiting arrhythmia
Grade II	Any deviation from the ideal intraoperative course <ul style="list-style-type: none"> Requiring pharmacological treatment, plasma volume expansion, coagulation products or blood transfusion Not life-threatening and not leading to permanent disability <p>Note: Routine medications and infusions to maintain anaesthesia are <u>not</u> considered as treatment. Blood transfusions for preoperative anaemia are not considered as treatment</p>	<ul style="list-style-type: none"> Bleeding: Plasma volume expansion for transient abundant venous bleeding from presacral venous plexus at rectal surgery Non-surgical: Antiarrhythmic treatment
Grade III	Any deviation from the ideal intraoperative course <ul style="list-style-type: none"> With the need for an additional surgical/endoscopic intervention Not life-threatening and not leading to permanent disability 	
<ul style="list-style-type: none"> Grade IIIa 	With organ preservation and low risk of long-term disability	<ul style="list-style-type: none"> Bleeding: Redo-vascular anastomosis for bleeding Injury: Suture of iatrogenic bladder injury at rectum resection Injury: Direct closure with suture of iatrogenic small bowel injury Injury: Chest tube insertion for accidental iatrogenic pneumothorax
<ul style="list-style-type: none"> Grade IIIb 	With loss of organ or part of organ and low risk of long-term disability	<ul style="list-style-type: none"> Bleeding: Splenectomy for bleeding complication at left hemicolectomy Injury: Short segmental resection of small bowel for iatrogenic injury Injury: Cholecystectomy for incidental gallbladder injury
Grade IV	Any deviation from the ideal intraoperative course <ul style="list-style-type: none"> With the need for an additional intervention Life-threatening (requiring IC/ICU management) and/or leading to permanent disability or with a high risk of permanent disability 	<ul style="list-style-type: none"> Bleeding: Hypovolemic shock due to laceration of the caval vein at right adrenalectomy Injury: Roux-en-y reconstruction for major bile duct injury Injury: Accidental transection of femoral nerve Injury: Bilateral transection of recurrent laryngeal nerve Non-surgical: anaphylactic shock from routine antimicrobial prophylaxis
Grade V	Death of a patient	<ul style="list-style-type: none"> Bleeding: Death from hypovolemic shock Non-surgical: Death from myocardial infarction

Classification overall: OVERALL, how appropriate is the proposed classification of intraoperative complications (see as well next five questions)?

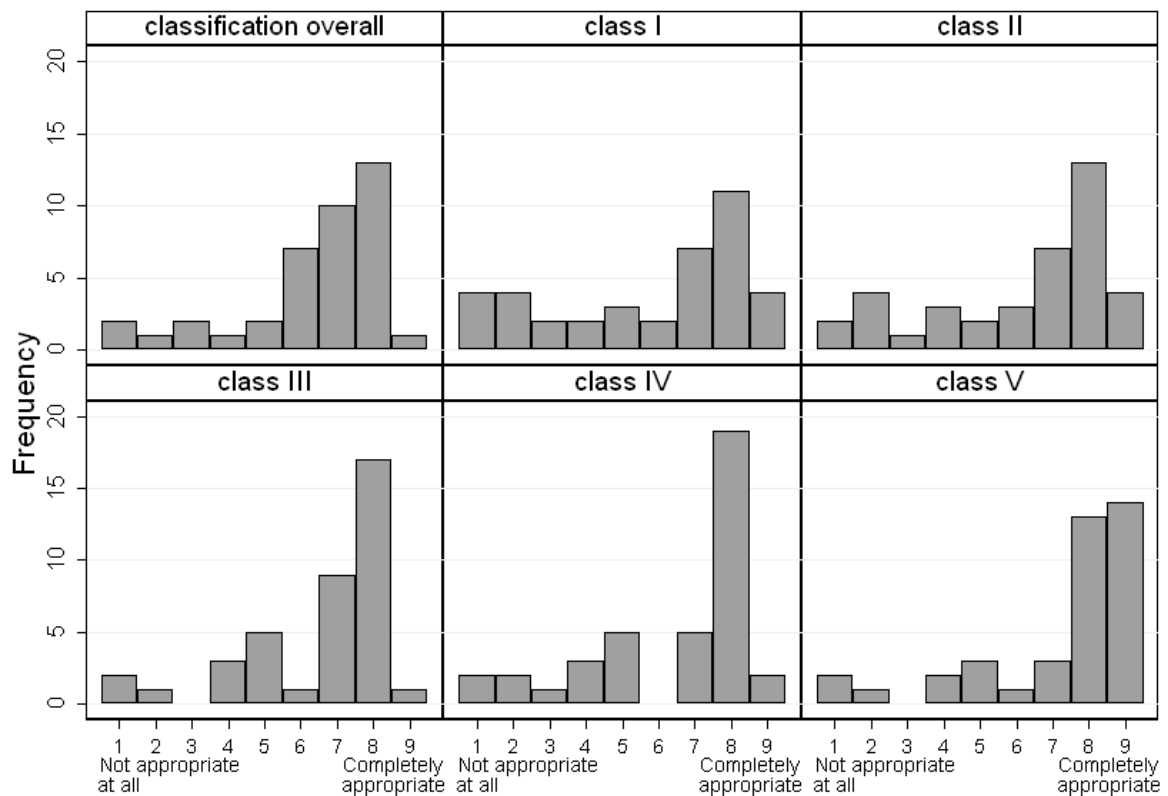
Class I: How appropriate is the proposed CLASS I type of intraoperative complications?

Class II: How appropriate is the proposed CLASS II type of intraoperative complications?

Class III: How appropriate is the proposed CLASS III type of intraoperative complications?

Class IV: How appropriate is the proposed CLASS IV type of intraoperative complications?

Class V: How appropriate is the proposed CLASS V type of intraoperative complications?



Summary classification overall

Concerning the tick box answers, the experts were rather in favour of the proposed classification and the grades, although to a lower extent of grade I and II.

In general, the classification was criticised for being too complicated, whereas the contrary was as well the case, one expert pointing out that it was too broad for instance compared to the CTCAE(U.S.Department of Health and Human Services 2014).

Again, the wording ‘any deviation from the ideal intraoperative course’ was felt to be inappropriate, since ideal or normal is difficult to define and since some

deviations are not preventable and not considered as complication. In contrast, they are important, such as a deviation due to an unexpected anatomical situation. As mentioned above, this could be handled by adding any 'preventable' deviation from the ideal course, 'given optimal treatment according to current guidelines'. It was pointed out that in some specialties such as cardiac surgery, grade I-II complications are very frequent, for instance bleeding due to preoperative administration of anti-platelet drugs; that in major liver surgery for instance bleeding is difficult to define, since it may be part of a routine course and a definition over transfusion requirement would not seem to be appropriate in such a case. In neurosurgery, a single non-life-threatening, but persistent deficit should be taken into account with the classification.

Some felt the classification to be too similar to the Clavien-Dindo classification for postoperative complications (Dindo, Demartines, & Clavien 2004) and it was therefore suggested to use the Clavien-Dindo Classification for intraoperative complications as well. It was as well suggested to test practicability in a retrospective cohort, as foreseen. The use of the Clavien-Dindo classification for intraoperative complications would again have the disadvantage of "upgrading" complications requiring a surgical intervention, since intraoperatively an additional surgical gesture would more deliberately be carried out as compared to after completion of the surgery.

Summary class I

This and the following questions refer to "grade", although it was misleadingly called "class". It was pointed out by the experts that class I complications are irrelevant to the patients' health and that they happen so often that most likely they will not be listed anyhow. Conversely, it was proposed to include subcategories for grade I. It was as well suggested to introduce a grade 0 for no complication. In our opinion, a grade I might still be interesting in research settings, although we agree that in daily practice such complications most likely would not be mentioned. For analysis, grades 0 and 1 could be taken together anyhow.

Summary class II

It was questioned whether a bleeding managed with plasma volume expansion is less severe than a suture of a iatrogenic small bowel injury and that a surgically managed bleeding was not listed to belong to grade II. Additionally, it was mentioned that plasma volume expansion is difficult to define. Moreover, it was suggested to include a measure of preoperative patient morbidity, since indication for interventions such as blood transfusions depend on comorbidity.

Summary class III:

It was pointed out to be unclear whether a conversion to open procedures is regarded as additional surgical intervention. Additionally, low versus high risk of long-term disability was criticised for a lack in definition.

Summary class IV:

Experts seemed to feel the grading somewhat arbitrary or in other words that it is difficult to decide whether a bile duct injury is “worse” than a massive bleeding with massive blood transfusions. The definition of an additional intervention seemed not clear, i.e. is revision surgery included in this definition? The requirement for an additional intervention seemed to be estimated as rather a lower grade than grade IV.

Summary class V

While some felt that at least death was a clear definition, it was pointed out that the definition was lacking which time period should be included. This was handled previously in the questions referring to which time interval should be included.

Questions round 2

Based on these answers, we updated the classification as follows:

Table: Proposed updated Classification of intraoperative complications (CLASSIC)

Grade	Definition
Grade 0	No preventable ¹ deviation from the ideal intraoperative course, under the condition the indication for surgery and the interventions conform to current guidelines.
Grade I	Any preventable deviation from the ideal intraoperative course, under the condition the indication for surgery and the interventions conform to current guidelines. Without the need for any additional treatment or intervention
Grade II	Any preventable deviation from the ideal intraoperative course, under the condition the indication for surgery and the interventions conform to current guidelines. With the need for any additional treatment or intervention Not life-threatening and not leading to permanent disability
Grade III	Any preventable deviation from the ideal intraoperative course, under the condition the indication for surgery and the interventions conform to current guidelines. With the need for any additional treatment or intervention Life-threatening and/or leading to permanent disability
Grade IV	Any preventable deviation from the ideal intraoperative course, under the condition the indication for surgery and the interventions conform to current guidelines. With death of the patient

We kindly ask you to rate the updated classification as presented in this table (1= not at all appropriate, 9 = completely appropriate).

Should complications needed to be dichotomised for further analysis (which we not necessarily advocate), we propose to summarize grades I-II together as “minor complications” versus grades III-IV as “major complications” (1= not at all appropriate, 9 = completely appropriate):

¹ an unpreventable deviation could be for instance a deviation due to an unexpected anatomical situation

Appendix Text (including illustrations) S2: Results round 2

1. Definition of intraoperative complications

1.1. Distinction

Questions round 2

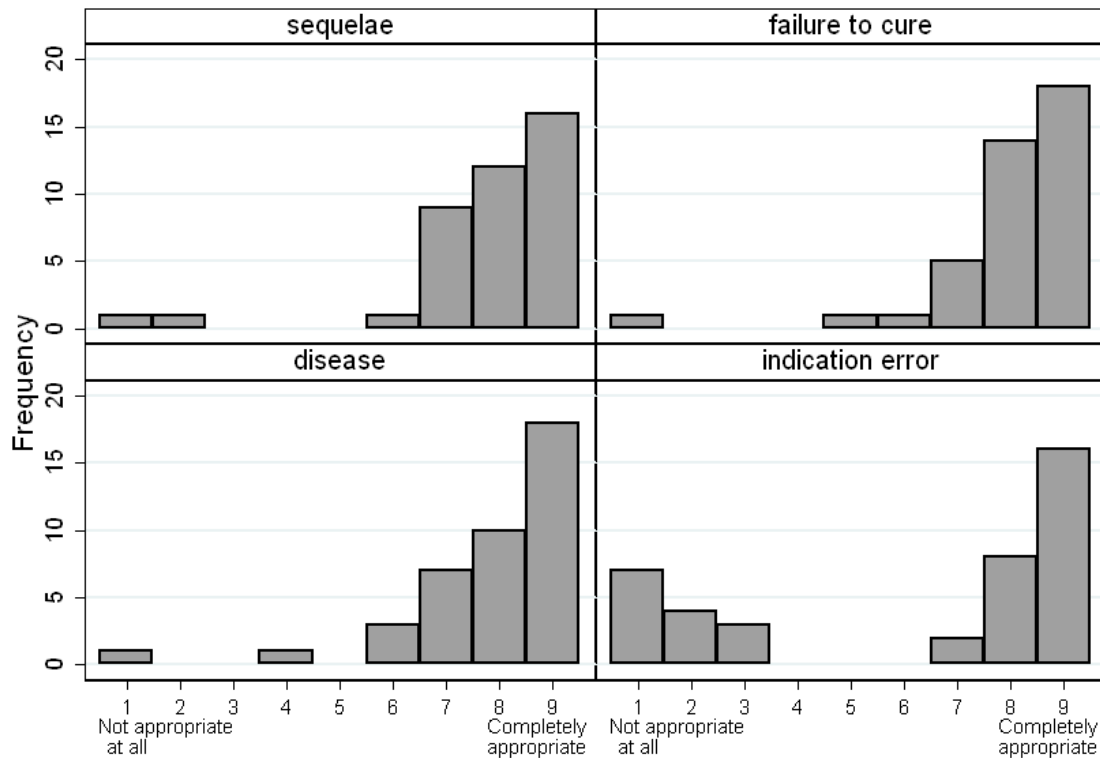
Based on these answers of round 1, we kindly ask you to rate the following updated definition (1= not at all appropriate, 9 = completely appropriate):

Sequelae: A definition of intraoperative complications needs to distinguish complications from sequelae, i.e. effects inherent to current best practice surgery (such as the inability to walk after amputation of a leg). It should refer to the index surgery and not to potential follow-up surgeries to treat a complication.

Failure to cure: A definition of intraoperative complications needs to distinguish complications from failures of cure (such as residual tumor after surgery), under the condition the indication for surgery and the interventions conform to current guidelines.

Disease: A definition of intraoperative complications needs to distinguish complications from events related to the underlying disease (such as bleeding from a ruptured abdominal aortic aneurysm or intraoperative sepsis due to a purulent peritonitis associated to a perforated appendicitis), under the condition the indication for surgery and the interventions conform to current guidelines.

Indication error: A definition of intraoperative complications should NOT refer to wrong-side or wrong-patient surgery and NOT to errors in indication (such as inappropriate indication for surgery according to current guidelines)



Summary sequelae

Experts seemed largely to agree with this definition (tick box answers). One expert again pointed out that sequelae could be the result of poor surgical practice. We entirely agree, but think that the addendum ‘best practice’ should rule out such cases. Another expert pointed out that the sentence about index surgery was unclear.

Update

We thus update the definition as follows:

A definition of intraoperative complications needs to distinguish complications from sequelae, i.e. effects inherent to current best practice surgery (such as the inability to walk after amputation of a leg). It should refer to complications occurring during the index surgery and not to those occurring during potential follow-up surgeries to treat a complication.

Summary nocure

Again, the tick box answers suggested that the reviewers agreed with this distinction. The importance of adequate staging was pointed out and is already accounted for in the provided definition. We agree that failure to cure is an important predictive parameter for long-term outcome. It was additionally suggested to slightly shorten the definition (replace ‘under the condition’ by ‘provided’). It was highlighted that ‘cure’ does not broadly apply to all interventions and that it could be replaced by ‘intended or proposed aim of surgery’, as for instance in aesthetic surgery. We agree with the expert that cure is less broadly

applicable, feel however that if we distinguish failures of reaching the intended or proposed aim of surgery from intraoperative complications, many complications would not be captured, since recorded as failure of reaching the aim instead of as complication. It additionally was stressed that guidelines are not always appropriate or available and thus it was suggested to write 'under the condition the indication for surgery was properly planned and indicated'. Again, we agree that there are not always established high-level evidence guidelines. However when using 'properly', a definition should be provided, which may not be clear-cut. Additionally, the question was highlighted on how to handle failures of adherence to the indicated plan or guidelines. In our opinion, this may not be uniformly answered. Failure of adherence to guidelines may be at the level of indication and is outlined below or can occur during surgery and manifest as intraoperative complication, which then is captured as such.

Update

We thus update the definition as follows:

A definition of intraoperative complications needs to distinguish complications from failures of cure (such as residual tumor after surgery), provided the indication for surgery and the interventions conform to current guidelines.

Summary disease

Experts seemed to agree on this point (tick box). Besides the same shortening as mentioned above, there were no more comments.

Update

We thus update the definition as follows:

A definition of intraoperative complications needs to distinguish complications from events related to the underlying disease (such as bleeding from a ruptured abdominal aortic aneurysm or intraoperative sepsis due to a purulent peritonitis associated to a perforated appendicitis), provided the indication for surgery and the interventions conform to current guidelines.

Summary indication

Here, similar to the first round, the tick box answers suggested that some experts rather regarded errors in indication as intraoperative complications. These contradictory answers were as well reflected by the comments. On one hand, it was pointed out that indication is taking place before surgery and for practical reasons, such a classification should account only for the time from skin incision to closure. Conversely, the operation was as well considered as part of a whole system highlighting that the systems' behaviour dictates the outcome. Additionally, it was pointed out that errors in indication may lead to intraoperative complications, such as in the case of an inadvertent vessel injury while attempting to remove a non-resectable tumour. However, an intervention could be carried out without any intraoperative 'technical' complications, but for instance on the wrong side. So we rather see error in indication as a risk factor for intraoperative complications, which

then may or may not occur. We agree with one expert that wrong side surgery is not the same as error in indication, since the indication could be entirely appropriate, had the intervention been carried out on the proper side. It is evident that the controversy on how to handle these errors in indication, side or patient may not be resolved. We propose however to handle such errors separately in order to enhance practicability. Moreover, there are interventions that – if carried out on the wrong patient - are neither life-threatening nor leading to permanent disability, so they will be graded low, whereas intuitively it seems to be a very serious complication, if you carry out a surgery on the wrong patient.

Update

We thus update the definition as follows:

A definition of intraoperative complications should NOT refer to wrong-site or wrong-patient surgery and NOT to errors in indication (such as inappropriate indication for surgery according to current guidelines). Wrong-site, wrong-patient surgery and errors in indication should be reported separately.

1.2. Wording

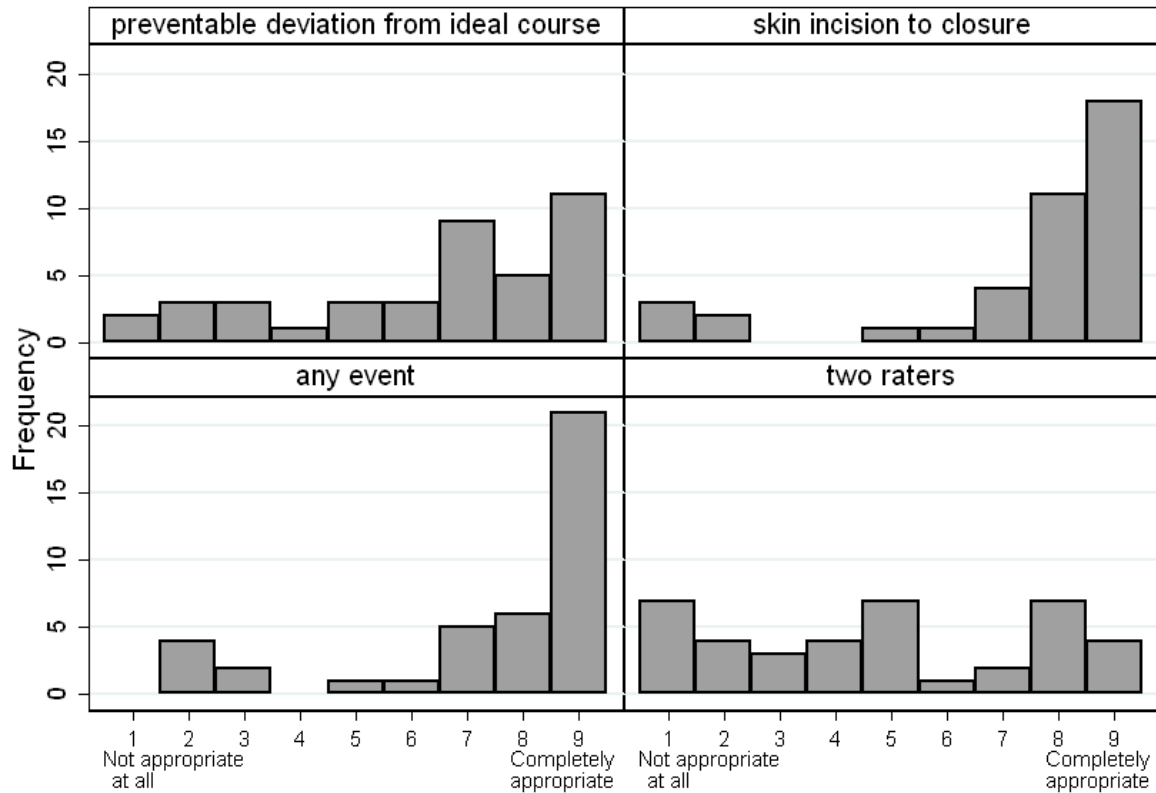
Questions round 2

Preventable deviation from ideal course: A definition of intraoperative complications may be formulated as “any preventable deviation from the ideal intraoperative course”

Skin incision to closure: A definition of intraoperative complications should include the following time period of event occurrence (disregarding time-point of event manifestation): Surgery only (i.e. between skin incision and skin closure) and should be distinguished from perioperative complications including anesthesia induction and termination.

Any event: A definition of intraoperative complications should include any event occurring during surgery, regardless whether it is related to the surgical gesture itself.

Two raters: A classification of intraoperative complications should be undertaken by two independent surgeons.



Summary preventable deviation from ideal course

Given the tick-box answers, experts seemed more to agree, although in this case there were quite some which were rather neutral or disagreeing. The word preventable was added after the first round, because experts suggested that not every deviation from the ideal course is a complication, for instance in case of an unexpected anatomical difficulty found upon surgery. It was however argued in this round that even given unexpected anatomical difficulties, there could be an ideal course (which may be different from the originally planned ideal course) and that deviations from the course would be an intraoperative complication. Additionally, it was argued that preventable is highly dependent on individual (subjective) judgment and that it is not always clear-cut to assess whether a deviation was preventable or not, as for instance an enterotomy during extensive adhesiolysis. It was as well argued that some intraoperative events are not actually preventable, such as fat embolism when reaming a bone or cement embolism in cemented hip arthroplasty, but still should be recorded as serious complications. It was additionally pointed out that 'ideal' implies a non-real world situation and suggested rather to use 'deviation from the planned/expected/correct course'. With 'planned' we are facing again the problem that sometimes plans need to be changed due to unexpected events, such as anatomical difficulties. If 'deviation from the expected course' is used, it may be attempted to list some relevant deviations as expected and not capture them anymore as such. One expert pointed out to use the OCHRA system for definitions of error. OCHRA stands for Observational Clinical Human Reliability Assessment.(Tang et al. 2004)

n the cited work, human error was defined according to the Bellagio Conference on Human error as “. . . something that has been done which was: (i) not intended by the actor, (ii) not desired by a set of rules or an external observer, or (iii) that led the task or system outside acceptable limits”.(Senders W 2014) The authors considered “any action or omission that resulted in a negative consequence or increased the time of the surgical procedure by necessitating a corrective action, that fell outside of the “acceptable limits” and was, therefore, registered as a consequential error.” The authors defined inconsequential error as “action or omission that increased the likelihood of negative consequence and under slightly different circumstances could have had a consequential effect”.(Tang, Hanna, Joice, & Cuschieri 2004)

We suggest not including the concept of ‘intention’ in the definition, i.e. omitting ‘not intended’. Complications may result from intended actions resulting from misinterpretation of a situation. Similarly, desired seems to be linked to some sort of intention. We agree entirely on the point that such a deviation may be consequential or inconsequential, which is however covered by the grading (i.e. inconsequential is similar to ‘no need for additional action’).

Update

We thus update the definition as follows:

A definition of intraoperative complications may be formulated as “any deviation from the ideal intraoperative course”

Summary skin incision to closure

Experts seemed largely to agree on this point (tick box answers). It was however pointed out that this definition would then not cover any bowel injury manifest only some days later. Based on the discussion of round 1, we felt that in many cases it is difficult to establish such a causal relationship, although in some it is obvious (i.e. foreign body left back). However in the majority of cases, it will be difficult to definitively advocate a causal relationship. In the rare event that causal relationship between a surgical intervention and a complication becoming evident only some days after surgery (e.g. surgical instrument left in place), we suggest capturing this event with the postoperative complication tool and additionally transparently reporting the specific event as such.

It was additionally pointed out that anaesthetic complications should not be considered separately and the example of an anaphylactic response occurring during surgery based on an antibiotic given before surgery should be captured. We entirely agree on this point and such an event would be captured as long as occurring during surgery. One expert asked how to handle mispositioning complications, for instance compartment syndrome or nerve injury. We propose to capture these events as perioperative events, similar to events occurring during induction of anaesthesia prior skin incision.

Update

We thus update the definition as follows:

A definition of intraoperative complications should include the following time period of event occurrence (disregarding time-point of event manifestation): Surgery only (i.e. between skin incision and skin closure) and should be distinguished from perioperative complications occurring between the beginning of anaesthesia induction and patient positioning (whichever is first) and skin incision and between skin closure and anaesthesia termination.

Summary any event

Here, most of the experts seemed to be in favour of including any event (tick box answers). One expert argued that this approach is in accordance with the patient's view not to distinguish between the surgeon and any other reason during surgery.

Update

We thus maintain the definition unchanged:

A definition of intraoperative complications should include any event occurring during surgery, regardless whether it is related to the surgical gesture itself.

Summary number of raters

In this regard, there was no uniform opinion at all. As a matter of fact, most of the tick box answers were chosen at about the same frequency.

The comments revealed a number of arguments for and against a rating by two independent surgeons. Quite some experts were in favour of the idea, stressing that this would be the optimal way to proceed, felt however that it was impractical. It was pointed out that an easy and pragmatic classification should not need to implicate two surgeons; especially not once it has been validated. Additionally, it was highlighted that with such a prerequisite, some events might not be captured because of a lack of a second independent assessor. In any case, a second surgeon having taken part in the surgery would not really be independent. So it was rather proposed to seek an agreement of all surgeons present during the surgery with the option to include a suffix if not in agreement. One expert suggested opting for an agreement between 2 members of the operating team, not necessarily 2 surgeons. Similarly, one expert pointed out the option of including the scrub-nurse. It was additionally mentioned that there is a risk of underreporting in surgeon-led non-training services. Finally, one expert pointed out that in future "litigation-happy societies" anyhow all procedures would be videotaped solving the problem of independent review.

Update

We thus update the definition as follows:

Any report on intraoperative complications should state how many raters were classifying the events and whether they were involved in the procedure. If they were involved in the procedure, it should be stated in which role and if not, whether the classification was based on records or on videorecordings of the

intervention. In case of multiple raters, it should additionally be reported how a consensus was reached.

2. Classification of intraoperative complications

Questions round 2

We kindly ask you to rate the following (1= not at all appropriate, 9 = completely appropriate)

Graded according to severity: A classification of intraoperative complications should consider the severity of the complication and take into account both the necessary treatment and permanent sequelae resulting from the complication.

Immediately after surgery: Should it be possible to definitively fill in a classification of intraoperative complications directly after surgery, whereas complications apparent only after surgery are handled with a classification system for postoperative complications?

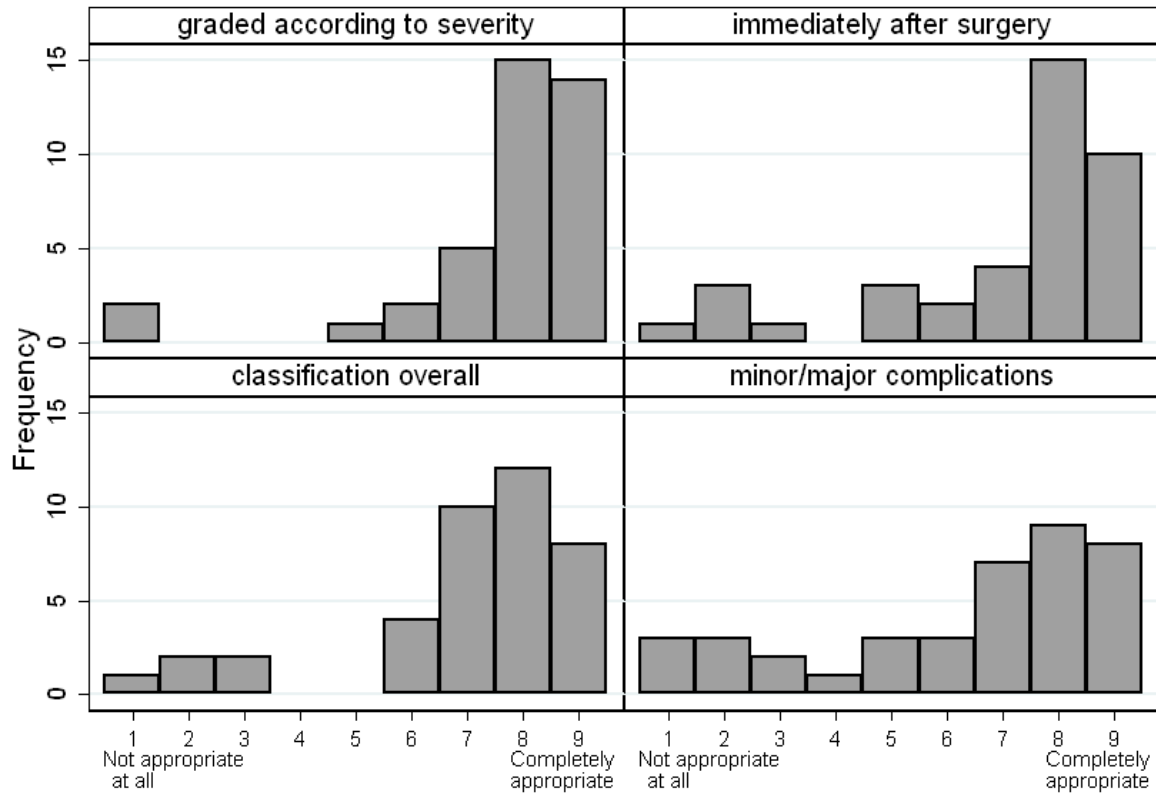
Classification overall: We kindly ask you to rate the updated classification as presented in the table below.

Minor/major complications: Should complications needed to be dichotomised for further analysis (which we not necessarily advocate), we propose to summarize grades I-II together as “minor complications” versus grades III-IV as “major complications”.

Table : Proposed updated Classification of intraoperative complications (CLASSIC)

Grade	Definition
Grade 0	No preventable ¹ deviation from the ideal intraoperative course, under the condition the indication for surgery and the interventions conform to current guidelines.
Grade I	<p>Any preventable deviation from the ideal intraoperative course, under the condition the indication for surgery and the interventions conform to current guidelines.</p> <ul style="list-style-type: none"> · Without the need for any additional treatment or intervention
Grade II	<p>Any preventable deviation from the ideal intraoperative course, under the condition the indication for surgery and the interventions conform to current guidelines.</p> <ul style="list-style-type: none"> · With the need for any additional treatment or intervention · Not life-threatening and not leading to permanent disability
Grade III	<p>Any preventable deviation from the ideal intraoperative course, under the condition the indication for surgery and the interventions conform to current guidelines.</p> <ul style="list-style-type: none"> · With the need for any additional treatment or intervention · Life-threatening and/or leading to permanent disability
Grade IV	<p>Any preventable deviation from the ideal intraoperative course, under the condition the indication for surgery and the interventions conform to current guidelines.</p> <ul style="list-style-type: none"> · With death of the patient

¹ an unpreventable deviation could be for instance a deviation due to an unexpected anatomical situation



Summary graded according to severity

Here, experts mostly agreed with the definition (tick box answers), although some felt that wording was quite complex and that the interplay with any classification of postoperative complications should be clarified.

Update

We thus update the definition as follows:

A classification of intraoperative complications should consider the severity of the complication. The severity is graded taking into account whether an additional treatment is resulting from the complication and whether it is life-threatening and/or leading to permanent disability.

Summary immediately after surgery

Experts largely support the proposition of grading directly after surgery (tick box answers), because more practical and for a clear distinction from any postoperative classification system. However, one expert preferred keeping a 30day-evaluation.

Update

We thus maintain the definition unchanged:

It should be possible to definitively fill in a classification of intraoperative complications directly after surgery, whereas complications apparent only after surgery are handled with a classification system for postoperative complications.

Summary Classification overall

The tick box answers showed that the experts mostly were in favour of the presented classification. Similar to the comments on the definition of intraoperative complications, many experts felt that the word 'preventable', which was added due to suggestions of the first round, was inappropriate (see above) and should be deleted. It was pointed out that 'preventable' is difficult to define and that many surgeons would claim many errors as not preventable. Some events are not preventable, even in very experienced hands, so the only way to prevent them is not to carry out the intervention at all. The example given by one expert was dural tear in decompression spine surgery requiring an additional intervention (suture). Another example given by another expert was that cemented hip hemiarthroplasty in frail patients is associated with a certain mortality rate, which is not really preventable. If we kept 'preventable' in the definition, such a death would not be captured.

Some experts still felt the grading system to be too complicated, whereas others would like to add more subgroups such as for grade III, because they felt this group might then unify complications with different consequences. In the example given, small bowel injury, splenectomy and bile duct injury were graded as III, whereas in our opinion, for instance a small bowel injury not qualifies in all instances to be graded as III. Additionally, it was pointed out that a bleeding without requiring blood substitution might not be captured as grade I. However, a bleeding not requiring blood substitution, but additional volume (e.g. crystalloids), would still be captured.

One expert criticised the classification for being an adaption of the Clavien-Dindo classification (Dindo, Demartines, & Clavien 2004) and thus not original and the fact that it seems to require the observer to await the outcome (such as death) not allowing immediate classification. We intended however to capture with the classification only deaths during the time of surgery, whereas deaths thereafter would be captured with the classification of postoperative complications, regardless the aetiology of death (since it might not be obvious to establish a causal relationship between a postoperative event and the index surgery). The experts suggested rather classifying complications not based on the actual outcome, but based on the probability of the outcome. This would allow for immediate classification and accounting for the fact that some severe complications remaining 'silent' might then erroneously be graded as I, although a relatively small complication might lead to severe consequences in exceptional circumstances.

We actually discussed the issue in our group before launching the Delphi process and initially foresaw including a term 'potentially' to account for the fact that we are unable to await the outcome for practical reasons. The consequence would then be to adapt the wording into 'life-threatening/potentially life-threatening and/or leading/potentially leading to permanent disability'. The drawback of such a wording then however is that 'potentially' allows for quite some range of

interpretation and thus hampering again uniform reporting. This is the reason why we dropped again this option.

Update

We thus update the grading omitting 'preventable' as follows:

Proposed updated Classification of intraoperative complications (CLASSIC), not allowing for potential consequences

Grade	Definition
	Prerequisite: the indication for surgery and the interventions conform to current guidelines
Grade 0	No deviation from the ideal intraoperative course
Grade I	Any deviation from the ideal intraoperative course <ul style="list-style-type: none"> · Without the need for any additional treatment or intervention
Grade II	Any deviation from the ideal intraoperative course <ul style="list-style-type: none"> · With the need for any additional treatment or intervention · Not life-threatening and not leading to permanent disability
Grade III	Any deviation from the ideal intraoperative course <ul style="list-style-type: none"> · With the need for any additional treatment or intervention · Life-threatening and/or leading to permanent disability
Grade IV	Any deviation from the ideal intraoperative course <ul style="list-style-type: none"> · With death of the patient

Summary minor/major complications

The tick box answers were not conclusive, although more experts found the suggestion appropriate. While one expert preferred dichotomia, others felt 0-IV categories to be more appropriate and not too complex to memorize. While it was suggested first addressing the pilot study before deciding on this issue, it was also criticised that grade II complications should not be regarded as minor.

Update

We thus update the definition as follows:

We propose not to dichotomise the grading for further analysis. If it shall still be used for investigating minor versus major complications, both results from the original classification and the contracted classification should be reported providing a clear definition for the cut-off.

Summary comments overall

The initiative to undertake this study was well received and it was proposed to validate the classification within a multicentre and international setting (if possible including the involved experts again).

3. Summary overall

3.1. Definition of intraoperative complications

3.1.1. Distinction

A definition of intraoperative complications

- Needs to distinguish complications from sequelae, i.e. effects inherent to current best practice surgery (such as the inability to walk after amputation of a leg).
- Refers to complications occurring during the index surgery and not to those occurring during potential follow-up surgeries to treat a complication.
- Needs to distinguish complications from failures of cure (such as residual tumor after surgery), provided the indication for surgery and the interventions conform to current guidelines
- Needs to distinguish complications from events related to the underlying disease (such as bleeding from a ruptured abdominal aortic aneurysm or intraoperative sepsis due to a purulent peritonitis associated to a perforated appendicitis), provided the indication for surgery and the interventions conform to current guidelines.
- Does NOT refer to wrong-site or wrong-patient surgery and NOT to errors in indication (such as inappropriate indication for surgery according to current guidelines). Wrong-site, wrong-patient surgery and errors in indication should be reported separately.

3.1.2. Wording

An intraoperative complication

- Is defined as “any deviation from the ideal intraoperative course”
- Includes the following time period of event occurrence (disregarding time-point of event manifestation): Surgery only (i.e. between skin incision and skin closure) and should be distinguished from perioperative complications occurring between the beginning of anesthesia induction and patient positioning (whichever is first) and skin incision and between skin closure and anesthesia termination.
- Includes any event occurring during surgery, regardless whether it is related to the surgical gesture itself or not.

Any report on intraoperative complications should state how many raters were classifying the events and whether they were involved in the procedure. If they were involved in the procedure, it should be stated in which role and if not, whether the classification was based on records or on videotapings of the intervention. In case of multiple raters it should additionally be reported how a consensus was reached.

3.2. Classification of intraoperative complications

A classification of intraoperative complications

- Considers the severity of the complication.
- Concerns deviations from the ideal intraoperative course, provided the indication for surgery and the interventions conform to current guidelines.
- Takes into account,
 - § whether a complication occurs (no [grade 0] versus yes)
 - § whether an additional treatment is resulting from the complication (no [grade I] versus yes)
 - § whether it is life-threatening and/or leading to permanent disability (no [grade II] versus yes [grade III]).
 - § whether it is resulting in death of the patient [grade IV])
- Is definitively filled in directly after surgery, whereas complications apparent only later after surgery are handled with a classification system for postoperative complications.

We propose not to dichotomise the grading for further analysis. If it shall still be used for investigating minor versus major complications, both results from the original classification and the contracted classification should be reported providing a clear definition for the cut-off.

Appendix Table S8: Patient characteristics of the pilot study (n=60)

Characteristic	Category	Value	
Median (IQR ¹) age in years		69 (54, 77)	
Gender, n (%)	Male	39 (65%)	
	Female	21 (35%)	
ASA ² classification, n (%)	1	5 (8%)	
	2	32 (53%)	
	3	20 (33%)	
	4	3 (5%)	
	5	0 (0%)	
Admission type, n (%)	Elective	43 (72%)	
	Emergency	17 (28%)	
Diagnosis, n (%)	Benign thyroid disease	7(12%)	
	Thyroid cancer	2 (3%)	
	Primary hyperparathyroidism	1 (2%)	
	Inguinal hernia	- unilateral	5 (8%)
		- bilateral	3 (5%)
	Unilateral femoral hernia	- without incarceration	1 (2%)
		- with incarceration	1 (2%)
	Symptomatic cholecystolithiasis	3 (5%)	
	Acute cholecystitis	5 (8%)	
	Cholangitis/cholangiosepsis	2 (3%)	
	Colorectal cancer	9 (15%)	
	Colorectal metastasis	1 (2%)	
	Volvulus of sigmoid colon	1 (2%)	
Diverticulitis	- free perforation	1 (2%)	
	- perforation in mesentery	2 (3%)	
	- vesicoenteric fistula	1 (2%)	
Hepatocellular carcinoma	2 (3%)		
Liver metastasis	3 (5%)		
Pancreatic cancer	4 (7%)		
Intraductal papillary mucinous neoplasm	1 (2%)		
Esophageal cancer	5 (8%)		

¹ IQR=Interquartile range

² ASA= American Society of Anesthesiologists

Appendix Table S9: Procedure characteristics of the pilot study (n=60)

Characteristic	Complexity ¹ of surgery(Dindo, Demartines, & Clavien 2004;Klotz, Candinas, Platz, Horvath, Dindo, Schlumpf, & Largiader 1996)		
	Type A n=20	Type B n=20	Type C n=20
Type of surgical intervention, n (%)	4 (7%) Hemithyroidectomy 1 (2%) Hemithyroidectomy with parathyroidectomy	1 (2%) Open cholecystectomy 1 (2%) Open cholecystectomy & cholangiography	1 (2%) Right hemihepatectomy 1 (2%) Left hemihepatectomy
	3 (5%) Total thyroidectomy 1 (2%) Total thyroidectomy with neck dissection	8 (13%) Laparoscopic cholecystectomy	2 (3%) Open bisegmentectomy/atypical liver resection 1 (2%) Laparoscopic bisegmentectomy/atypical liver resection
	1 (2%) Parathyroidectomy	1 (2%) Open right hemicolectomy 2 (3%) Laparoscopic right hemicolectomy	1 (2%) Laparoscopic left pancreatectomy
	4 (7%) Lichtenstein unilateral 1 (2%) Lichtenstein bilateral	1 (2%) Open left hemicolectomy 0 (0%) Laparoscopic left hemicolectomy	4 (7%) Pyloruspreserving pancreatectomy (Traverso-Longmire)
	1 (2%) McVay repair 1 (2%) McVay repair with small bowel resection	4 (7%) Open sigmoidectomy 1 (2%) Laparoscopic sigmoidectomy	2 (3%) Transhiatal esophagus resection
	1 (2%) TEP ² unilateral 2 (3%) TEP bilateral	1 (2%) Hartmann's procedure	1 (2%) Open abdominothoracal esophagus resection 2 (3%) Laparoscopic abdominothoracal esophagus resection
			1 (2%) Total mesorectal excision
			1 (2%) Open rectosigmoid resection 3 (5%) Laparoscopic rectosigmoid resection
Wound class(Mangram, Horan, Pearson, Silver, & Jarvis 1999a)			
Clean	19 (95%) 1 (5%)	0 (0%) 10 (50%)	0 (0%) 18 (90%)
Clean-contaminated	0 (0%)	8 (40%)	1 (5%)
Contaminated	0 (0%)	2 (10%)	1 (5%)
Dirty-infected			

¹ Complexity A=low, B=intermediate, C=high

² TEP=Total extraperitoneal repair

Appendix Table S9: Procedure characteristics (continued)

Characteristic	Complexity ¹ of surgery(Dindo, Demartines, & Clavien 2004;Klotz, Candinas, Platz, Horvath, Dindo, Schlumpf, & Largiader 1996)		
	Type A n=20	Type B n=20	Type C n=20
Surgeon experience, n (%)			
Resident	14 (24%)	7 (35%)	7 (35%)
Junior consultant	23 (38%)	9 (45%)	7 (35%)
Senior consultant	23 (38%)	4 (20%)	6 (30%)
Assistant experience, n (%)			
Resident	6 (30%)	5 (25%)	1 (5%)
Junior consultant	11 (55%)	12 (60%)	8 (40%)
Senior consultant	3 (15%)	3 (15%)	11 (55%)
Median (IQR ²) duration of surgery in minutes	93 (75, 125)	133 (70, 195)	253 (183, 320)

232

¹ Complexity A=low, B=intermediate, C=high

² IQR=Interquartile range

Appendix Table S10: Outcomes of the pilot study (n=60)

Characteristic	Category	All n=60	Complexity ¹ of surgery(Dindo, Demartines, & Clavien 2004;Klotz, Candinas, Platz, Horvath, Dindo, Schlumpf, & Largiader 1996)		
			Type A n=20	Type B n=20	Type C n=20
Median (IQR ²) length of hospital stay in days		8 (3, 15)	3 (2,3)	8 (7,15)	14 (9,31)
Intraoperative complications, n (%)	0	41 (68%)	17 (85%)	13 (65%)	11 (55%)
	1	2 (3%)	0 (0%)	1 (5%)	1 (5%)
	2	16 (27%)	2 (10%)	6 (30%)	8 (40%)
	3	1 (2%)	1 (5%)	0 (0%)	0 (0%)
	4	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Postoperative complications, n (%)(Dindo, Demartines, & Clavien 2004)	0	33 (55%)	15 (75%)	12 (60%)	6 (30%)
	1	7 (12%)	0 (0%)	3 (15%)	4 (20%)
	2	8 (13%)	2 (10%)	1 (5%)	5 (25%)
	3a	2 (3%)	0 (0%)	0 (0%)	2 (10%)
	3b	5 (8%)	2 (10%)	1 (5%)	2 (10%)
	4a	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	4b	1 (2%)	0 (0%)	1 (5%)	0 (0%)
	5 ³	4 (7%)	1 (5%)	2 (10%)	1 (5%)

¹ Complexity A=low, B=intermediate, C=high

² IQR=Interquartile range

³ Related to patient factors (comorbidities) and/or severe illness