FACULTY OF HEALTH AND MEDICAL SCIENCES
UNIVERSITY OF COPENHAGEN

Ph.D. Thesis

Multiplicity and sparse data in systematic reviews of anaesthesiological interventions cause increased risk of random errors and lack of reliability of conclusions

Georgina Imberger

The Copenhagen Trial Unit, Centre for Clinical Intervention Research, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark

This thesis has been submitted to the Graduate School of The Faculty of Health and Medical Sciences, University of Copenhagen, on the 1st of March 2014

Academic advisors:

Jørn Wetterslev, Chief Physician, Associate Professor, MD, Ph.D. (Faculty advisor)
Copenhagen Trial Unit, Centre for Clinical Intervention Research
Rigshospitalet
Copenhagen University

Ann M Møller, Professor, MD, DMSc (Project advisor)
Department of Anaesthesia & Intensive Care
Herlev Hospital
Copenhagen University

Evaluating committee:

Jørgen B.Dahl, Professor, Chief Physician, MD, DMsc, MBAex (Chairperson)
Department of Anaesthesiology, Copenhagen University Hospital, Rigshospitalet
Faculty of Health and Medical Sciences
University of Copenhagen

Ian Roberts, Professor, MB, BCh, FRCP, FPH
Department of Population Health
London School of Hygiene and Tropical Medicine
United Kingdom

Hans Kirkegaard, Professor, MD, DMsc, Ph.D, DEAA
Department of Clinical Medicine
Aarhus University

This Ph.D. thesis is based on the following four studies:

STUDY 1

Imberger G, Vejlby AD, Hansen SB, Møller AM, Wetterslev J. Statistical multiplicity in systematic reviews of anaesthesia interventions: a quantification and comparison between Cochrane and non-Cochrane reviews. *PLoS ONE* 2011; **6**: e28422

STUDY 2

Imberger G, Thorlund K, Gluud C, Wetterslev J. False positive findings in cumulative meta-analysis with and without application of trial sequential analysis. To be submitted in 2014.

STUDY 3

Imberger G, Gluud C, Boylan J, Wettterslev J. Results in systematic reviews of anaesthesiological interventions that claim to be statistically significant often contain risk of type 1 error greater than 5%. To be submitted in 2014.

STUDY 4

Imberger G, Orr A, Thorlund K, Wetterslev J, Myles P, Møller AM. Does anaesthesia with nitrous oxide affect mortality or cardiovascular morbidity? A systematic review with meta-analysis and trial sequential analysis. *British Journal of Anaesthesia*. 2014;112:410-26.

Preface

During the time I worked on this Ph.D., I have lived in three countries and worked in four. The personalities I would like to thank are as broad and rich as my experiences during these years.

I start with the families – urban and real – who have supported me with such style and effect. In Copenhagen, I thank my fellow peddlers, with whom I spent long happy nights in blues bars, for their loyal and fabulous friendship. In Dublin, I thank the boys, with whom I learnt that 3am is an early finish, for keeping everything in shining perspective. In Melbourne, I thank them all, walkers of the tan, chasers of cows, lovers of The Beatles, for everything. In particular, I thank both my parents, who have been quite extraordinary in their unwavering support of my left-hand turns.

Professionally, I thank Ann Møller and everyone at CARG in Copenhagen, for welcoming me, enabling my start on the project, and continuing their valued involvement as the project grew. I thank John Boylan and everyone in the St. Vincent's Anaesthesia Department, Dublin, for taking me on, and being so charming about being interested in my research. And I thank Paul Myles and everyone at the Department of Anaesthesia at Monash University, Melbourne, for giving me a lovely lodge for the final chapters.

Finally, I thank everyone at The Copenhagen Trial Unit, where I was lucky enough to find a research home and have been very happy. I thank Mette Hansen, for all her practical magic. I thank Christian Gluud, for that sharp, sharp mind and his willingness to use it to help others. Most importantly, I thank Jørn Wetterslev, who has been unfaltering as my supervisor, as a practical guide, a teacher, a role model and a friend. I feel very fortunate to have had his wisdom in my reach and I am enormously grateful for everything he has done.

Georgina Imberger
February 2014

Contents

Abbreviations

CARG	Cochrane Anaesthesia Review Group
D^2	Diversity
EBM	Evidence based medicine
EER	Experimental error rate
ENIGMA	Evaluation of nitrous oxide in the gas mixture for anaesthesia
GRADE	Grading of Recommendations Assessment, Development and Evaluation
I^2	Inconsistency
IQR	Inter-quartile range
MI	Myocardial Infarct
NNT	Number needed to treat
PICO	Population, intervention, comparison, outcome
PE	Pulmonary embolus
RIS	Required information size
RRR	Relative risk reduction
TSA	Trial sequential analysis
95% CI	95% confidence interval

Definitions

95% Confidence Interval	An estimate of precision around a sample parameter, such that if independent samples were taken repeatedly from the same population, the confidence interval calculated for each sample would contain the unknown population parameter 95% of the time
Design error	Error resulting from decreased applicability of a body of evidence to a given clinical question (or vice versa)
D^2 (Diversity)	A measurement of between-trial variation (heterogeneity) in meta-analysis
Experimental error rate	The probability of rejecting at least one of k independent null hypotheses when all are true
Internal multiplicity	Multiplicity resulting from multiple statistical tests within a single study or investigation
I^2 (Inconsistency)	A statistic used to quantify inconsistency across studies in a meta-analysis
Meta-analysis	The process of statistically combining results from different studies
Multiplicity	The presence of multiple statistical tests
P-value	The probability that an observed test statistic, or one more extreme, comes from a population where the null hypothesis is true
Random error	Error resulting from the ‘play of chance’
Sequential multiplicity	Multiplicity resulting from repeated statistical tests of the same hypothesis over time
Systematic error	Error resulting from methodological conduct causing an increase in the risk of bias in a final conclusion
Type 1 error	Incorrect rejection of a null hypothesis
Type 2 error	Incorrect rejection of an alternative hypothesis
Z-value	Standardized test statistic

Summary

Background

Systematic reviews with meta-analyses have often been considered as the highest level of evidence and their conclusions hold much relative influence. The reliability of these conclusions is therefore an extremely important agenda. In this Ph.D. project, we aimed to explore the effect of random error on the reliability of conclusions. Specifically, we focused on the issue of multiple statistical comparisons (multiplicity) and sparse data in systematic reviews of anaesthesiological interventions. Multiplicity increases the risk of type 1 random error. Trial sequential analysis (TSA) is a methodology that aims to adjust for the multiplicity caused by repeated updates in meta-analyses with sparse data.

Objectives and methods

The first objective was to quantify the internal multiplicity present in systematic reviews in anaesthesia research and compare this quantity between Cochrane reviews and comparable non-Cochrane reviews. We matched systematic reviews published by the Cochrane Anaesthesia Review Group with comparable non-Cochrane reviews, counted the total number of meta-analysed statistical comparisons, and compared this number between Cochrane and non-Cochrane reviews.

The second objective was to investigate the proportion of real-life cumulative meta-analyses that produce false positive findings and explore the ability of trial sequential analysis to prevent these false positives. We selected 100 Cochrane meta-analyses that were large enough to have demonstrated, to a reasonable level, that the given intervention does not cause a clinically relevant effect on the outcome in question. We conducted retrospective cumulative meta-analysis using conventional techniques and measured the proportion of false positives that would have occurred had these meta-analyses been updated after each new trial had been completed. For these false positive findings, we performed TSA, using three different approaches, mimicking how a prospective analysis could have been performed had it been done at the time the cumulative meta-analysis showed a false positive result.

The third objective was to estimate the proportion of reported statistically significant findings in systematic reviews of anaesthesiological interventions that preserve their risk of type 1 error below 5% when TSA is performed. We conducted a search to identify all systematic reviews with meta-analysis investigating anaesthesiological interventions and randomly selected 50 that reported a statistically significant categorical outcome in their abstract. We applied TSA to these meta-analyses, using two main TSA approaches (relative risk reduction (RRR) 20% and the border of the conventional 95% confidence interval (CI) closest to null) and a further two approaches as sensitivity analyses. We calculated the proportion of meta-analyses that maintained statistical significance with TSA.

The fourth objective was to perform a systematic review of general anaesthesia with nitrous oxide versus without and use trial sequential analysis as a demonstration of how to estimate and communicate uncertainty in meta-analytic conclusions in the context of repeated updates and sparse data. ENIGMA II is a large randomized clinical trial currently underway which is investigating nitrous oxide versus no nitrous oxide and cardiovascular complications. Before the completion of this trial, we performed a systematic review and meta-analysis, using Cochrane methodology, with conventional meta-analysis and TSA, on the five outcomes that make up the composite primary outcome in ENIGMA II.

Results

Regarding the quantity of internal multiplicity in Cochrane and non-Cochrane systematic reviews, the median number of statistical tests overall was 10 (IQR 6 to18). The median was 12 in Cochrane and 8 in non-Cochrane reviews with a difference in medians of 4 (95% CI 2.0–19.0). The issue of multiplicity was addressed in only 6% of all the reviews.

Regarding false positives in cumulative meta-analysis with and without TSA, using conventional retrospective cumulative meta-analysis, one or more false positives were present in at least 7% (95% CI 3%-14%) of the meta-analyses. This estimate of the false positive rate may be an underestimate as 57% more meta-analyses with P-values greater than or equal to 0.05 are updated compared with meta-analyses with P-values less than 0.05. Using the three TSA approaches, TSA prevented the false positive error 13 of the 14 times it occurred (93%, 95% CI 64%-100%).

Regarding the reliability of conclusions of anaesthesiological systematic reviews, using the two TSA approaches, only 30% (95% CI 18-45%) of the meta-analyses preserved the risk of type 1 error below 5%. In 98% (88-100%) of the systematic reviews, either a formal assessment of risk of bias was not conducted or the selected meta-analysis included trials assessed as having increased risk of bias.

Regarding the nitrous oxide review, using conventional meta-analysis, the relative risk of short-term mortality in the nitrous oxide group was 1.38 (95% CI 0.22–8.71) and the relative risk of long-term mortality in the nitrous oxide group was 0.94 (95% CI 0.80–1.10). In both cases, TSA demonstrated that the data were far too sparse to make any conclusions. There were insufficient data to perform meta-analysis for stroke, myocardial infarct, pulmonary embolus, or cardiac arrest.

Conclusions

The quantity of internal multiplicity in systematic reviews is high overall and higher in Cochrane reviews than in non-Cochrane reviews. Few systematic reviews, whether Cochrane or non-Cochrane, address the issue of multiplicity.

False positives most likely occur more frequently in real-life cumulative meta-analysis than the desired 5%. TSA is a helpful statistical methodology to prevent these false positives and to assess the reliability of early nominally statistically significant findings in cumulative meta-analyses.

TSA demonstrates that, due to the increased risk of type 1 random error present in early meta-analyses, a large proportion of published conclusions from meta-analyses of anaesthesiological interventions may be unreliable.

Systematic review and meta-analysis with TSA demonstrated that, prior to ENIGMA II, evidence is far too sparse to make any conclusions about how nitrous oxide used as part of general anaesthesia affects mortality and cardiovascular complications.

Dansk resumé

Baggrund

Systematiske litteraturoversigter med meta-analyser er ofte blevet anset for at repræsentere den højeste grad af evidens og konklusioner på baggrund af disse har opnået tiltagende indflydelse. At fastslå pålideligheden af disse konklusioner er derfor overordentlig vigtigt. Formålet med dette ph.d. projekt har været at undersøge effekten af risikoen for tilfældige fejl på pålideligheden af konklusionerne. I særdeleshed har vi fokuseret på problemerne ved multiple sammenligninger, multiplicitet, og utilstrækkeligt antal randomiserede deltagere, kaldet informationsstørrelsen, i systematiske litteraturoversigter af anæstesiologiske interventioner. Multiplicitet og en informationsstørrelse mindre end den nødvendige øger risikoen for type 1 fejl, et problem der synes at have opnået ringe opmærksomhed i den systematiske litteraturoversigts kontekst. Forsøgssekventielle meta-analyser (engelsk: trial sequential analysis (TSA)) er en metodologi, der har som mål at justere risikoen for tilfældige fejl for den multiplicitet, der skyldes utilstrækkelig information og repetitiv testning på akkumulerende data, i opdaterede meta-analyser.

Formål og metoder

Det første formål var at kvantitere graden af tilstedeværelsen af intern multiplicitet i systematiske litteraturoversigter i anæstesiologisk forskning og sammenligne graden af multiplicitet mellem Cochrane litteraturoversigter og tilsvarende non-Cochrane litteraturoversigter. Vi matchede systematiske litteraturoversigter publiceret af Cochrane Anaesthesia Review Group med tilsvarende non-Cochrane litteraturoversigter. Vi opgjorde det totale antal meta-analyserede statistiske sammenligninger i henholdsvis Cochrane og non-Cochrane litteraturoversigter og testede forskellen.

Det andet formål var at undersøge proportionen af kumulative meta-analyser med falsk positive fund, trods $P<0.05$, og TSA's evne til at forhindre disse falsk positive. Vi udvalgte 100 Cochrane meta-analyser som med rimelig sikkerhed, 80% power, var store nok til udelukke en klinisk relevant effekt på det effektmål der var undersøgt. Vi gennemførte herefter retrospektive, konventionelle kumulative meta-analyser og udregnede proportionen af falsk positive ($P<0.05$) som ville have forekommet når disse meta-analyser var blevet

opdateret efter inklusion af hver nyt forsøg. For hver af disse falsk positive kumulative meta-analyser udførte vi TSA, under 3 forskellige forudsætninger, for at vurdere hvordan en prospektiv TSA ville være faldet ud på det tidspunkt hvor den kumulative meta-analyse opnåede en $P<0.05$.

Det tredje formål var at estimere proportionen af rapporterede statistisk signifikante fund i systematiske litteraturoversigter af anæstesiologiske interventioner efter at TSA var blevet udført med en bevaret risiko for type 1 fejl mindre 5%. Vi identificerede alle systematiske litteraturoversigter med meta-analyser af anæstesiologiske interventioner og udvalgte tilfældigt 50 meta-analyser af disse som rapporterede et statistisk signifikant kategorisk effektmål i det tilhørende resumé. Vi udførte TSA på disse 50 meta-analyser idet vi hovedsagelig anvendte TSA analyser med henholdsvis en relative risiko reduktion (RRR) på 20% og en RRR hidrørende fra den konventionelle 95% konfidensgrænse tættest på 0. Vi udførte yderligere sensitivitets analyser med henholdsvis RRR=10% og RRR=30%. Herefter udregnede vi proportionen af meta-analyser der forblev statistisk signifikante efter en TSA analyse.

Det fjerde formål var at udføre en systematisk litteraturoversigt vedrørende en anæstesiologisk intervention og anvende TSA til at demonstrere estimering og kommunikation af usikkerheden i konklusionen på en meta-analyse i konteksten af gentagne opdateringer og utilstrækkelig informationsstørrelse. ENIGMA II er et stort randomiseret klinisk forsøg som i øjeblikket undersøger effekten af kvælstofforilte på forekomsten af kardiovaskulære komplikationer. Inden færdiggørelsen af dette forsøg, har vi med Cochrane metodologi udført en systematisk litteraturoversigt med konventionelle meta-analyser og TSA, på de 5 effektmål som udgør det sammensatte primære effektmål i ENIGMA II.

Resultater

Den interne multiplicitet i Cochrane og non-Cochrane systematiske litteraturoversigter opgjort som det mediane antal statistiske test udført på sammenligninger mellem interventioner var 10 (IQR: 6 til18). Medianen var 12 i Cochrane og 8 i non-Cochrane litteraturoversigter, differencen mellem medianerne var 4 (95% CI 2.0–19.0).

Problemet med multiplicitet blev kun adresseret i 6% af disse litteraturoversigter.

I 7% (95% CI 3%-14%) af meta-analyserne forekom falsk positive kumulative meta-analyser ved anvendelse af konventionelle retrospektive kumulative meta-analyser, uden TSA. Ved anvendelse af de 3 TSA modeller, forhindrede TSA en falske positiv konklusion i 13 af 14 tilfælde hvor den konventionelle type 1 fejls grænse ($P<0.05$) var overskredet, d.v.s. i 93% (95% CI 64%-100%) af tilfældene blev signifikansgrænsen i TSA ikke overskredet.

Pålideligheden af konklusionerne i anæstesiologiske litteraturoversigter kunne kun bekræftes i 30% (95% CI 18-45%) af meta-analyserne ved anvendelse af 2 af TSA modeller for at holde den samlede type 1 fejls risiko under 5%. I 98% (88-100%) af de systematiske litteraturoversigter var der enten ikke foretaget en formel bias vurdering eller også inkluderede meta-analyserne forsøg med høj risiko for bias.

I den systematiske litteraturoversigt af effekten af kvælstofforilte var den the relative risiko for død indenfor 30 dage i kvælstofforilte gruppen sammenlignet med kontrolgruppen 1.38 (95% CI 0.22–8.71) og den relative risiko for død i løbet af maximale opfølgning i kvælstofforilte gruppen sammenlignet med kontrolgruppen var 0.94 (95% CI 0.80–1.10) ved en konventionel meta-analyse. I begge tilfælde demonstrerede TSA at informationsstørrelsen var alt for lille for at kunne konkludere pålideligt. Der var ikke tilstrækkelige data til at kunne udføre meta-analyser på effekten af kvælstofforilte versus kontrolinterventionen på forekomsten af slagtilfælde, myokardieinfarkt, lungeemboli eller hjertestop.

<u>Konklusioner</u>

Gaden af intern multiplicitet i systematiske litteraturoversigter er høj og højere i Cochrane end i non-Cochrane systematiske litteraturoversigter. Det er meget få litteraturoversigter, hvad enten det er Cochrane eller non-Cochrane systematiske litteraturoversigter, der adresserer dette multiplicitetsproblem.

Falsk positive meta-analyser forekommer hyppigere i kumulative meta-analyser end de ønskede og konventionelt fastlagte 5%. TSA er en anvendelig statistisk metode til at forebygge falsk positive konklusioner og til at at vurdere pålideligheden af tidlige nominelt statistisk signifikante fund i kumulative meta-analyser.

Anvendelsen af TSA demonstrerer at på grund af den forøgede risiko for type 1 fejl i meta-analyser, der ikke har nået den nødvendige informationsstørrelse, er en stor andel af publicerede konklusioner fra meta-analyser af anæstesiologiske interventioner ikke pålidelige.

En systematisk litteraturoversigt med meta-analyse og TSA forud for ENIGMA II viser at vi ikke har robust evidens for at kvælstofforilte anvendt som en del af generel anæstesi påvirker mortaliteten og forekomsten af kardiovaskulære komplikationer.

Background

Systematic reviews and the reliability of conclusions

A systematic review aims to collate all the available evidence in order to answer a specific research question. Meta-analysis refers to the statistical combining of results. Systematic reviews with meta-analyses have often been considered as the highest level of evidence(1-3) and their conclusions hold much relative influence(4). The reliability of these conclusions is therefore an extremely important agenda. Reliability implies consistency and reproducibility and reliable conclusions require error and consequential uncertainty to be minimised when possible and accurately estimated and communicated when not. Error can come from design error, systematic error, or random error(5). 'Design error' refers to the assessment of the applicability of a body of evidence to a given clinical situation, describing the situation where the trials that make up the body of evidence were not designed to investigate the exact clinical question of interest(6). Systematic error occurs when methodological conduct causes an increase in the risk of bias in the final conclusion. Random error is the 'play of chance' and is an inevitable part of inference, representing a permanent barrier to certainty.

Multiplicity in systematic reviews

Multiplicity is a noun that means a 'multitude' or a 'great number'. Multiplicity can also mean the state of being manifold, of many kinds, numerous and varied, or having many different parts, elements, features, or forms. In this thesis, I define 'multiplicity' as the presence of multiple statistical tests. I define 'internal multiplicity' as multiplicity resulting from multiple statistical tests within a single study or investigation, and 'sequential multiplicity' as multiplicity resulting from repeated statistical tests on accumulating data of the same hypothesis over time.

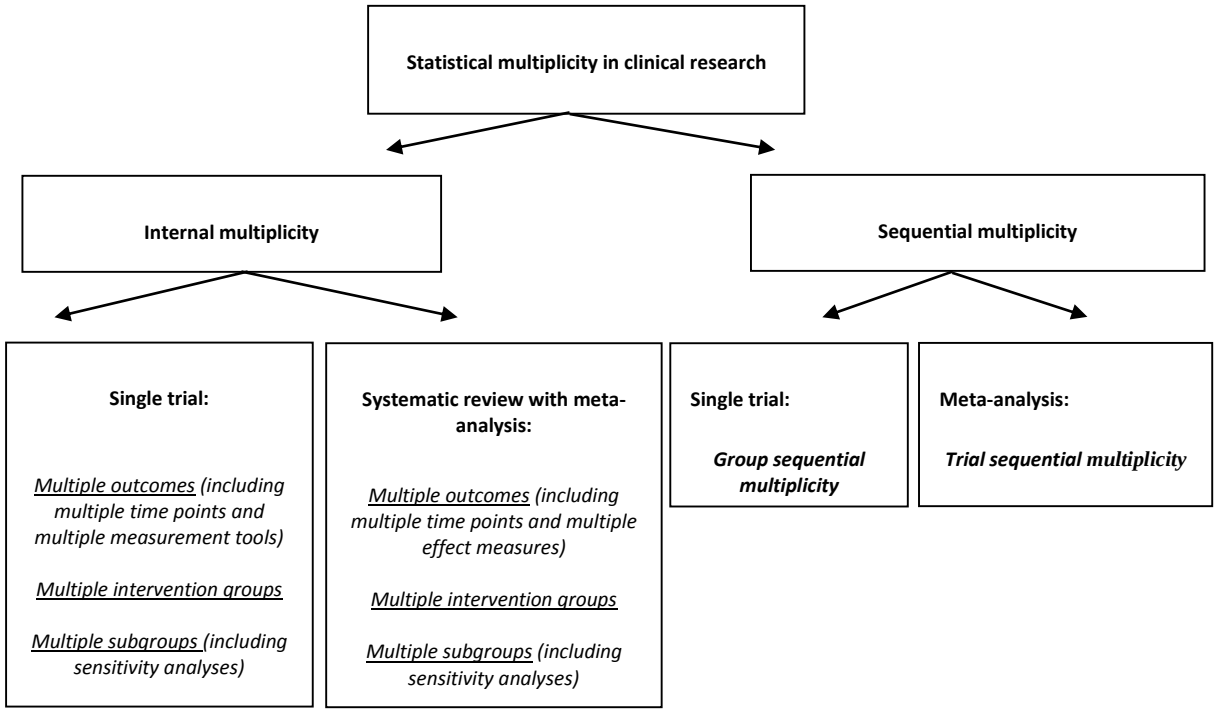
In medical research, it is common for many statistical comparisons to be made. This multiplicity can arise for various reasons: multiple outcomes may be compared, the same outcome may be measured at different time points, there may be multiple intervention groups, more than one measurement may be used to compare a single treatment effect, there may be analyses made of subgroups, or accumulating data may be compared before the final,

so-called fixed, sample size has been reached. Figure 1 summarises the classification of multiplicity that I use for this thesis.

While it is often a necessary consequence of analysing information, the problem with multiplicity is that it increases the risk of type 1 errors. Type 1 error is an incorrect rejection of a null hypothesis. When performing a statistical test, we make an inference that has an inherent risk of random type 1 error. If that risk is estimated at less than a certain value (usually 5%), it is considered reasonable to reject a null hypothesis. When two or more statistical tests are performed, using the same data, and each test estimates the risk of type 1 error as 5%, then the probability of making a type 1 error overall may end up being higher than 5%. In practical terms, statistical multiplicity increases the risk of false positive findings(7, 8).

The issue of multiplicity has received much attention in the context of single clinical trials (9-19). In systematic reviews, however, few attempts have been made to address the problem(20). In the systematic reviews themselves, the presence of this multiplicity is rarely mentioned (21). A review on the topic of multiplicity in systematic reviews in 2008 concluded that the issue requires recognition and further research is required(8). The issue of sequential multiplicity and sparse data in meta-analysis and its effect on the risk of random error has indeed begun to receive some attention(22-31). Overall, however, the published discussion of the importance of the issue of multiplicity in systematic reviews remains limited and the difficult issue of how it should be interpreted and managed remains largely unaddressed (5, 8, 20).

Figure 1. Classification of multiplicity



Interpretation and management of multiplicity in systematic reviews

Internal multiplicity

Internal multiplicity in systematic reviews presents a difficult challenge. The question of how to handle internal multiplicity in clinical trials has been a contentious one for many years(11, 12, 15, 16, 19). Many argue that multiplicity decreases the reliability of conclusions and quantitative adjustments should be made(8, 10, 18). There are others who think that the issue should be handled with good trial design, transparent reporting, and qualitative discussion(15-17). In systematic reviews, the question of how to manage internal multiplicity is potentially even more challenging than in a single trial. Authors of systematic reviews often aim to cover a topic thoroughly, sometimes with many outcomes, many subgroups, and many sensitivity analyses. When the authors of a systematic review meta-analyse data from comparisons that have already been made and published, one could argue that the results do not represent any *increased* multiplicity. Instead, one might argue that a meta-analysis provides a summary of the multiplicity that already existed. Alternatively, one might argue

that the number of meta-analytic comparisons in a systematic review do represent new multiplicity, with the new meta-analysed result being considered as a type of sequential statistical comparison. Moreover, somewhat more simply, truly new comparisons are often done as part of a meta-analysis – testing new sub-groups, performing sensitivity analyses or comparing outcomes that have not previously been tested. Given the complexity of the issue, and the existing divergence of opinions that exists about internal multiplicity in the context of a single trial, it seems likely that differences of opinion will exist about how to handle internal multiplicity in systematic reviews.

Sequential multiplicity

While the issue of internal multiplicity in a single trial remains contentious, there has long been consensus that statistical adjustment needs to be performed to correct for the sequential multiplicity produced by interim analyses done in a single trial before the sample size has been reached(9, 13). The higher the number of statistical tests that are performed as additional data accumulates, the higher the probability of observing a false positive result because of random error(32, 33). Early stopping can be problematic and monitoring boundaries, incorporating the sample size calculation, are commonly used to control the risk of random error at desired levels(9, 32, 34). When meta-analyses are updated over time, the sequential multiplicity is analogous to that created by interim analyses in a single trial and the risk of random error is similarly increased(35).

The implications of sequential multiplicity are particularly striking in systematic reviews when we note that the majority of published meta-analyses are underpowered. For example, testing for a relative risk reduction of 30%, 78% of Cochrane meta-analyses with a binary outcome have power less than 80%(36). 50% have power less than 27%(36). Given that The Cochrane Collaboration recommends updating systematic reviews at least every two years(5), the published conclusions from these under-powered meta-analyses are very likely to be updated in the future. That is, the majority of conclusions published from meta-analysis can be thought of as interim analyses, complete with the associated increased risk of type 1 error. In line with this argument, increased risk of type 1 error in early meta-analyses has been demonstrated by theoretical argument (8, 26), evidence from simulations studies (22, 25, 30, 35), and evidence from empirical work(3).

However much we would like our conclusions to be definitive, good clinical decisions require accurate estimation of uncertainty. It is better for meta-analysts to communicate greater error more accurately than it is to infer less error inaccurately. With the goal of more accurate estimation of the risk of random error in the context of sparse data and repeated updates in cumulative meta-analysis, several techniques have been proposed. Examples include trial sequential analysis (TSA) (23, 28, 29, 37), a semi-Bayes procedure (24), sequential meta-analysis using Whitehead's triangular test (38), and an application of the law of iterated logarithm(25). There is, however, a lack of consensus about the necessity to use these techniques(5, 8, 20).

Trial Sequential Analysis

TSA provides an approach to the problem of sequential multiplicity by adjusting the threshold for declaring statistical significance(39). TSA uses methodology developed for repeated significance testing in single trials(23, 28, 29, 37). The required information size (RIS) estimates the size of a meta-analysis that would be adequate for a given clinical question and scenario, using a specified control event proportion, a specified intervention effect, chosen risks of type 1 and type 2 errors, and an estimation of heterogeneity(39, 40). Thresholds are then constructed for statistical significance using an alpha-spending function. Alpha-spending functions have been used previously in interim-analyses of single trials, where they are used to create thresholds for statistical significance that are more conservative when the data are sparse and become progressively more lenient as the accrued information gets closer to the RIS(32).

Sequential multiplicity also increases the risk of type 2 error and TSA is able, using the same methodology, to construct adjusted thresholds for non-superiority and non-inferiority(39). The importance of declaring futility at the earliest time reasonable is clear, potentially preventing unnecessary ongoing trials. However, issues associated with type 2 error were not included in the scope of this Ph.D. project and are therefore not discussed in this thesis.

TSA boundaries are based on methodology developed by Lan and DeMets, initially intended for flexible repeated significance testing in a single trial(32, 41, 42), and then applied in the context of cumulative meta-analysis by Pogue and Yusuf(26). The thresholds are constructed

as a function of the strength of the evidence and the underlying statistical methodology depends on an assumption that data will continue to accumulate until the RIS is passed(32, 33, 41-44).

A sample size calculation in a single trial is designed to allow a trial to be large enough to reliably answer a defined hypothesis in the construct of a frequentist probability model. Conventional frequentist meta-analysis uses the same underlying probability model to produce confidence intervals and P-values. Moreover, meta-analyses are likely to include more variation in population and intervention, known as heterogeneity, decreasing the precision of the results. RIS for a meta-analysis, therefore, needs to be at least as large as the sample size for a single trial asking the same question(23, 26, 28, 29, 37, 45-47). TSA estimates the RIS for a meta-analysis by calculating a conventional sample size for a given set of parameters and then increasing it depending on the amount of heterogeneity present, using an estimate of between trial variance called 'diversity'(40).

The goal of the adjusted thresholds is to maintain overall risk of type 1 error at 5%, independent of how many times a hypothesis is repeatedly tested. TSA uses the alpha spending function to achieve this goal(32). The alpha spending function uses accumulated information as an independent variable (accumulated number of participants in a meta-analysis) and the maximum allowed cumulative type 1 error as the dependent variable(32, 41). The maximal allowed cumulative type 1 error describes the amount of error that should be maximum for a given accumulated number of participants (in order to ensure that the overall risk of type 1 error stays below 5%)(39).

The z-value is the standardized test statistic, it summarises the information contained in the meta-analysis. A higher z-score is consistent with a lower probability that the data came from a population where the null hypothesis is true (a lower p-value). The TSA thresholds are constructed using z-values for the dependent variable on the TSA curve, representing the z-value that has to be crossed in order to reach statistical significance. The threshold z-value for a given number of participants corresponds to the maximal allowed cumulative type 1 error for that number of participants. As the number of participants increases, the alpha spending function increases – that is, the maximal allowed cumulative type 1 error increases. By design, the alpha spending function is 0 at a minimal (when 0 participants have accumulated) and 1 at a maximum (when the number of participants equals RIS). The value of the y-axis

(between 0 and 1) represents the proportion of overall type 1 error allowed. For example, if an alpha spending function equals 0.25 after i number of participants (n_i), that means that the maximum allowed cumulative type 1 error at n_i is 0.25 of the overall type 1 error allowed. That is, if the maximal overall type 1 error allowed is 5%, then the maximal error at n_i is 1.25% (0.25 x 5%).

The actual TSA boundaries are constructed by translating this proportion of the overall type 1 error allowed into a z-score that can be used as a threshold for significance. With the above example, if the alpha spending function results in 1.25% being the maximum type 1 error allowed at n_i, the TSA boundary will give the z-value at that point with corresponds to a type 1 error of 1.25%. For a two-sided test, the z-value that corresponds to a type 1 error 1.25% is 2.50. So, if the z-value of the cumulative meta-analysis at n_i is greater than 2.50, then the overall risk of type 1 error is estimated as less than 5%, and statistical significance is declared using TSA(39).

Figure 2. Schematic demonstration of TSA

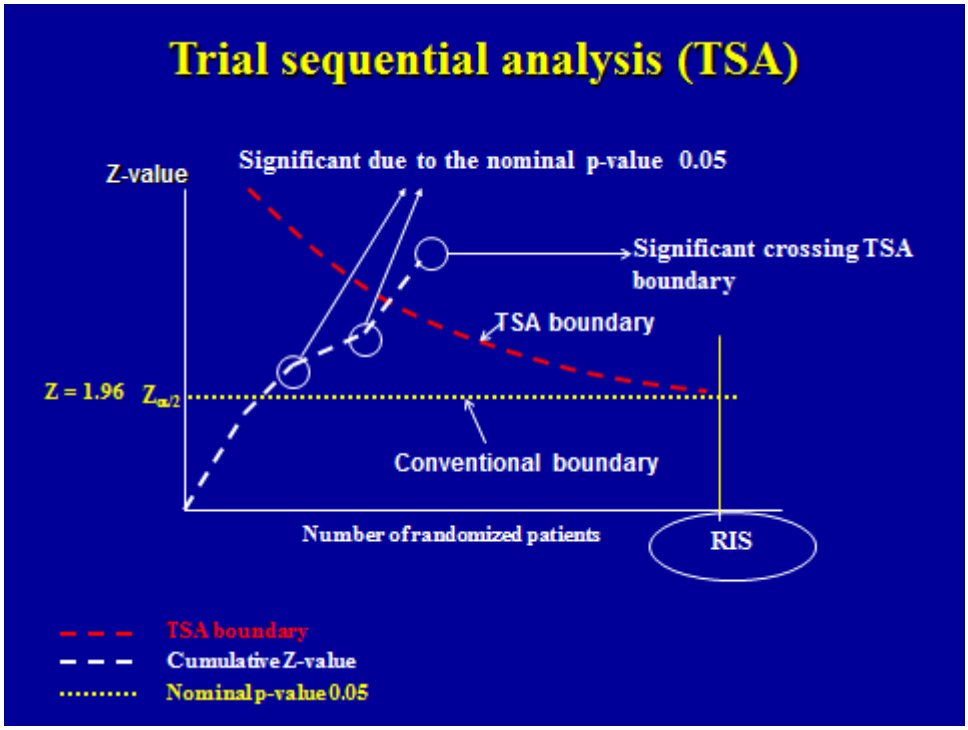


Figure 2 shows a schematic demonstration of TSA. As the number of participants increases, the z-value calculated from accumulating trials is plotted. The straight line where z=1.96

represents the conventional boundary for declaring significance for a type 1 error of 5%. The curved TSA boundary is constructed such that for any number of randomized participants (relative to RIS), anything below that z value represents an over-all risk of type 1 error greater than 5% and anything above that z-value represents an over-all risk of type 1 error less than 5%. The cumulative z-value is shown here to be increasing as the number of randomised participants increases, first crossing the conventional boundary, but requiring more precision before the TSA boundary is crossed and the overall risk of type 1 error can be considered to be less than 5%.

TSA has the potential to be a powerful and useful tool in the goal of summarizing evidence. In a hypothetical model, where assumptions are known, using TSA to construct thresholds for significance can control type 1 errors at the desired levels. In the real world, assumptions are not known, evidence is rarely sufficient, and it is therefore often not clear which parameters should be used to estimate RIS. Altering RIS will alter the entire TSA approach, and may change the overall conclusion. By exploring the different inferential results provided by using different TSA approaches (that is, using different estimates for the RIS calculation and creating probability models based on different assumptions), one of TSA's strong advantages is that it can be a dynamic tool, allowing for a realistic and changeable model with which to consider statistical significance in cumulative meta-analysis(48).

Anaesthesia with nitrous oxide and cardiovascular complications

As part of the exploration of random error in systematic review and how to best estimate and communicate it, especially in the context of sparse data, we conducted a systematic review. We chose a clinical question in anaesthesia that is important, was thought to currently have sparse data, and is due to be updated with new trial results in the immediate future. The goal with this choice was not just to summarise the body of evidence for an important clinical question, but to demonstrate cumulative meta-analysis as a prospective activity and how the use of trial sequential analysis can help better estimate and communicate the uncertainty in an early meta-analysis with sparse data.

Nitrous oxide has been used as a general anaesthetic for over 160 years. Collective anecdotal experience with this drug must be larger than with any other drug used in anaesthesia. Despite this experience, opinions about the role of nitrous oxide in modern-day practice

continue to diverge(49, 50). One concern is that exposure to nitrous oxide may increase the risk of cardiovascular complications. Nitrous oxide oxidises the cobalt atom in vitamin B12, inactivating methionine synthase, causing a decrease in folate metabolism and an increase in homocysteine(51). Homocysteinaemia after exposure to nitrous oxide has been well demonstrated in vivo(52-54), and long-term homocysteinaemia is known to be associated with an increased risk of ischaemic heart disease(55). It remains unclear, however, whether this information about this surrogate outcome translates into real clinical risk. To investigate the possible causal association between mortality, cardiac morbidity and nitrous oxide, the Enigma trial group has designed a large, multi-centre randomized clinical trial: ENIGMA II is enrolling at-risk patients and is powered to investigate a composite primary outcome of mortality, non-fatal acute myocardial infarction, cardiac arrest, pulmonary embolism and stroke(56).

Summary of objectives

The overall objective for this Ph.D. project was to explore the issue of multiplicity in systematic reviews with meta-analysis and the consequential increased risk of random error and lack of reliability of conclusions. A focus of this objective was a consideration of trial sequential analysis and its potential to better estimate risk of error in cumulative meta-analysis. To achieve this overall goal, we conducted four projects with the objectives outlined below.

1. Quantification of internal multiplicity in systematic reviews

   Statistical multiplicity in systematic reviews of anaesthesia interventions: a quantification and comparison between Cochrane and non-Cochrane reviews(57)

   The objective of this project was to help clarify the size of the issue of internal multiplicity in systematic reviews in anaesthesia research. A secondary objective was to compare this quantity between Cochrane reviews and comparable non-Cochrane reviews.

2. False positives in cumulative meta-analysis with and without TSA

   False positive findings in cumulative meta-analysis with and without application of trial sequential analysis(58)

   The objective of this study was to identify a population of Cochrane meta-analyses for which a reasonable RIS had been reached and the final conclusion was that there was no effect of the assessed intervention, identify how many cumulative meta-analyses, on their way to the final meta-analysis, would have produced early false positives had they been updated each time a new trial was published using conventional meta-analysis, and then explore how TSA could have contributed in making a more accurate assessment of error in these early meta-analyses.

3. Reliability of conclusions of anaesthesiological systematic reviews

   Results in systematic reviews of anaesthesiological interventions that claim to be statistically significant often contain risk of type 1 error greater than 5%(59)

The objective of this study was to examine apparently statistically significant conclusions in meta-analyses of anaesthesiological interventions and to assess what proportion of these conclusions would maintain statistical significance if TSA boundaries were used.

4. Nitrous oxide systematic review

Does anaesthesia with nitrous oxide affect mortality or cardiovascular morbidity? A systematic review with meta-analysis and trial sequential analysis(60)

The objective of this study was to perform a systematic review and meta-analysis, including TSA, on the outcomes used as the composite outcome in ENIGMA II(56), prior to the completion of this trial, with the goal of representing the first of a series of cumulative meta-analyses and demonstrating how random error caused by sequential multiplicity can be estimated and communicated.

Summary of methods and findings

1. Quantification of internal multiplicity in systematic reviews

Statistical multiplicity in systematic reviews of anaesthesia interventions: a quantification and comparison between Cochrane and non-Cochrane reviews(57)

We took the 43 systematic reviews published in November 2009 in The Cochrane Library by the Cochrane Anaesthesia Review Group (CARG) that contained a meta-analysis and matched them with 43 comparable non-Cochrane reviews. We counted the total number of meta-analysed statistical comparisons, whether a primary outcome was defined, the number of comparisons done under the primary outcome, and noted whether the authors addressed the issue of multiplicity in the text.

The median number of statistical comparisons in all the reviews was 10 (IQR 6 to 18) and was higher in the Cochrane than in the non-Cochrane reviews (12 compared to 8; adjusted p = 0.04). In the Cochrane reviews, the number of meta-analysed comparisons ranged from 1 to 1872. In the non-Cochrane reviews, the number of meta-analytic comparisons ranged from 1 to 98. In 87% of all reviews, there were more than 4 meta-analytic comparisons. In 24% of all reviews, there were greater than 20 meta-analytic comparisons.

In 49 (57%, 95% CI 46-67%) of the reviews, the primary outcome was clearly defined. In those where the primary outcome was not defined, we considered that all the meta-analysed comparisons in the review were part of the primary analysis. With this definition, the median number of meta-analysed comparisons done under the primary outcome overall was 6 (IQR 4 to 13), and there was still a trend for more multiplicity in the Cochrane reviews than in the non-Cochrane reviews (median 8 compared to 6; adjusted p value = 0.24).

Only five (6%, 95% CI 2-14%) of the reviews addressed the issue of multiplicity in their own review in some way, either mentioning it as a source of error or implementing some type of statistical adjustment for some effect of multiplicity.

The conclusion was that the quantity of multiplicity is high in systematic reviews. Multiplicity may be greater in Cochrane reviews than in non-Cochrane reviews. Many of the reasons for the increase in multiplicity may well represent improved methodological approaches and greater transparency, but multiplicity may also cause an increased risk of

spurious conclusions. Few systematic reviews, whether Cochrane or non-Cochrane, address the issue of multiplicity.

2. False positives in cumulative meta-analysis with and without TSA

False positive findings in cumulative meta-analysis with and without application of trial sequential analysis(58)

We screened the Cochrane Database of Systematic Reviews and selected 100 meta-analyses that were large enough to have demonstrated, to a reasonable level, that the given intervention does not cause a clinically relevant effect on the outcome in question. We conducted retrospective cumulative meta-analysis using conventional techniques and measured the proportion of false positives that would have occurred. For these false positives, we performed TSA, mimicking how a prospective analysis could have been performed had it been done at the time of publication of each new trial. We used three different TSA approaches to mimic this prospective analysis. The first approach – a 'credible parameters TSA approach' – used the parameter estimates that actually did exist in the final updated analysis and aimed to mimic what might have been chosen as credible and reasonable choices for the clinical questions being asked. For the second approach, we used parameter estimates from the trials included when the false positive occurred, using the border of the 95% confidence interval closest to the null at the time of the false positive as the estimate of effect. For the third approach, we again used parameter estimates from the trials included at the time when the false positive occurred, but this time we used the point estimate at the time of the false positive as the estimate of effect. As a post-hoc analysis, we surveyed three years of Cochrane systematic reviews and calculated the relative risk of a meta-analysis being updated if it was not significant relative to if it was significant.

Using conventional retrospective cumulative meta-analysis, one or more false positives were present in seven of the meta-analyses (7%; 95% CI 3%-14%). This estimate of the false positive rate may be an underestimate as our post-hoc analysis showed that 57% more meta-analyses with P-values greater than or equal to 0.05 are updated compared with meta-analyses with P-values less than 0.05. Using the three TSA approaches, TSA prevented the false positive type 1 error 13 of the 14 times the conventional threshold was crossed (93%, 95% CI 64%-100%).

The conclusions were that the proportion of false positives in cumulative meta-analysis is likely to be higher than 5% and that TSA is a helpful statistical methodology when assessing the reliability of early nominally statistically significant findings in cumulative meta-analysis.

3. Reliability of conclusions in anaesthesiological systematic reviews

Results in systematic reviews of anaesthesiological interventions that claim to be statistically significant often contain risk of type 1 error greater than 5%(59)

We conducted a search to identify all systematic reviews with meta-analysis investigating anaesthesiological interventions. We randomly selected 50 meta-analyses that reported a statistically significant categorical outcome in their abstract. We applied TSA to these meta-analyses, using two main TSA approaches (RRR 20% and the border of the conventional 95% CI closest to null) and a further two approaches as sensitivity analyses. We calculated the proportion of meta-analyses that maintained statistical significance with TSA. We also reviewed the assessments of risk of bias for the included trials in the selected meta-analyses.

From 11,870 titles, there were 682 systematic reviews that investigated anaesthesiological interventions. In the 50 randomly selected from these 682, the median number of trials included in the meta-analyses was 8 (IQR 5-14), the median number of participants was 964 (IQR 523-1736) and the median number of events was 202 (IQR 96-443). Using the two TSA approaches, only 30% (95% CI 18-45%) meta-analyses preserved the risk of type 1 below 5%. In 98% (88-100%) of the systematic reviews, either a formal assessment of risk of bias wasn't done or the included meta-analysis included trials assessed as having increased risk of bias.

The conclusion was that a large proportion of published conclusions from meta-analyses of anaesthesiological interventions may be unreliable.

4. Nitrous oxide systematic review

Does anaesthesia with nitrous oxide affect mortality or cardiovascular morbidity? A systematic review with meta-analysis and trial sequential analysis (60).

We performed a systematic review and meta-analysis, using Cochrane methodology, on the five outcomes that make up the composite primary outcome in ENIGMA II: mortality, stroke, myocardial infarction, pulmonary embolus, and cardiac arrest(56). We used conventional meta-analysis and TSA. We reviewed 8282 abstracts and selected 138 that fulfilled the criteria for study type, population and intervention. We attempted to contact the authors of all selected publications to check for unpublished outcome data. 13 trials had outcome data eligible for the five outcomes. We assessed three of these trials as having a low risk of bias. Using conventional meta-analysis, the relative risk for short-term mortality in the nitrous oxide group was 1.38 (95% CI 0.22 to 8.71) and the relative risk for long-term mortality in the nitrous oxide group was 0.94 (95% CI 0.80 to 1.10). In both cases, TSA demonstrated that the data were far too sparse to make any conclusions. There were insufficient data to perform meta-analysis for stroke, myocardial infarct, pulmonary embolus, or cardiac arrest.

The conclusion was that this systematic review demonstrated that we currently have far too sparse data to make any conclusions about how nitrous oxide, versus no nitrous oxide, used as part of general anaesthesia, affects mortality and cardiovascular complications.

Discussion

Interpretation of methodological findings

The issue of multiplicity and sparse data in systematic reviews, and its influence on random error and reliability of conclusions is both complex and broad. With the studies in this Ph.D. project, we hoped to contribute and further the discussion about the presence of the issue, its importance, and how we might better estimate and communicate uncertainty in conclusions of meta-analysis. The principal methodological findings were that the quantity of internal multiplicity in systematic reviews is high overall and higher in Cochrane reviews than in their non-Cochrane counterparts, that false positives are probably more frequent in cumulative meta-analysis than the generally accepted rate of 5%, that TSA is able to exclude most of these false positives and may provide better estimates of confidence in the context of sequential multiplicity, and that a large proportion of the apparently statistically significant meta-analytic conclusions published in anaesthesia research may not be reliable.

The first methodological finding was that the quantity of internal multiplicity in systematic reviews is large and higher in Cochrane reviews than in their non-Cochrane counterpart. We found an overall median of 10 (IQR 6 to 18) meta-analysed comparisons in each review and that the internal multiplicity was higher in Cochrane reviews than in their non-Cochrane counterparts. With regard to interpreting this finding, I discuss here whether this quantity of internal multiplicity is high enough to be a concern, the implications of the higher quantity in the Cochrane reviews, and give a brief outline of some possible approaches for addressing this issue of internal multiplicity in systematic reviews.

The experimental error rate (EER) is the probability of rejecting at least one of k independent null hypotheses when in fact all are true. EER is given by:

$$EER = 1 - (1-\alpha)^k$$

Where k equals the number of independent comparisons and α = the assigned type I error.

If each calculated risk of α is 0.05, then the probability of rejecting at least one of 10 null hypotheses incorrectly (assuming that they are all true) is 0.40 ($1 - (1-0.05)^{10}$). There are

substantial limitations to this calculation for an estimation of EER. Importantly, it is highly unlikely that the comparisons in a meta-analysis are independent – the accurate EER will lie somewhere between 0.05 and 0.40 and will depend on the quantity of the correlation between the comparisons, being difficult to estimate. Moreover, this measure is also only a description of how the type 1 error in a frequentist model operates under the assumption of a universal null, it does not describe the probabilities involved when one or more of the null hypotheses is rejected. These limitations in the estimation of EER give a brief indication of the difficulties in estimating and adjusting for internal multiplicity generally. Without addressing further the complexities involved for estimations for internal multiplicity, the simplified equation for EER serves here to demonstrate that the presence of 10 statistical comparisons has the potential to cause a large increase in the overall risk of type 1 error. For this reason, we concluded that the quantity of internal multiplicity in systematic reviews is high.

There are many explanations for why Cochrane reviews may have more internal multiplicity than there non-Cochrane counterparts and this increase is likely to represent methodological trends that are positive. It may be that as we improve the quality and breadth of systematic reviews, we cannot help but increase the multiplicity. For example, Cochrane encourages the investigation of adverse events(5). Such investigation is clearly important, but also leads to an increase in the number of outcomes. Similarly, Cochrane encourages subgroup analyses including studies with varying risks of bias(5). This finding is not a criticism of Cochrane reviews. Rather, it may be that methodological improvements in the way systematic reviews are conducted may lead, as a type of adverse side-effect, to an increase in internal multiplicity. As such, this finding re-iterates the importance of addressing the issue of how to handle internal multiplicity in the context of systematic reviews.

The problem of internal multiplicity in systematic reviews is a challenge. Opinions vary about its importance, whether to adjust for it, when to adjust for it, and, if so, how to adjust for it(8). In single clinical trials, many statistical procedures have been used to adjust for the increased random type 1 error caused by internal multiplicity(7, 12, 14, 20, 61). Examples include Bonferroni(62), Bonferroni derivatives such as Holm and Hochberg(12), non-Bonferroni procedures such as non-sampling procedures(63), and gatekeeping procedures(64).

Bender et al. published a review on the topic in the context of systematic review and offered a set of guidelines, providing suggestions about design, such as defining a primary outcome, keeping the total number of outcomes as small as possible, and focusing on patient important outcomes only(8). With regard to adjustment for statistical multiplicity in meta-analysis using aggregrated data, Bender et al. suggested that more research is needed and they propose multivariate meta-analysis as one promising approach(8).

The Cochrane handbook includes a discussion about multiplicity in systematic reviews(5), stating that the 'issues of multiplicity apply just as much to systematic reviews as to other types of research'. The Cochrane Collaboration, overall, do not recommend using adjustments. Mostly, they suggest preventative approaches when designing the systematic review and they also refer readers to the Bender review for further guidance(5). The conclusion in The Cochrane Handbook – that 'there is no simple or completely satisfactory solution to the problem of multiple testing and multiple interval estimation in systematic reviews'(5) is no doubt appropriate.

The next methodological findings in this Ph.D. project related to sequential multiplicity in systematic reviews. First, we found that false positives are probably more frequent in cumulative meta-analysis than the generally accepted rate of 5% and that using TSA can exclude most of these false positives. Second, when using TSA to adjust for sparse data and sequential multiplicity, we found that only 30% of apparently statistically significant conclusions (p-value<0.05) in anaesthesiological meta-analyses preserve a risk of type 1 error below 5%. With regard to interpreting these findings, I discuss here whether why we concluded that the risk of false positives is higher than 5%, and the challenges involved in conducting TSA with regard to the choice of parameters to inform the RIS.

Using conventional retrospective cumulative meta-analysis, we found that one or more false positives were present in 7% (95% CI 3%-14%). We concluded that the rate is probably higher than 5%, despite the 95% confidence interval reaching to 3%. The reasoning for this conclusion was based on a consideration of the population of meta-analyses that we were able to examine. We wanted to find a population of Cochrane systematic review meta-analyses where a clinical question had been reasonably clearly answered with adequate power and where the final conclusion was one of no or only a small clinical effect. To do this, we selected meta-analyses that fulfilled a RIS for a defined set of parameter estimates. We found

that such meta-analyses are very rare in Cochrane systematic reviews. Using the RRR of 10% as the criteria, only 1.8% (95% CI 1.3-2.3%) of Cochrane systematic reviews were eligible. Using NNT of ≥100 participants as the criteria for inclusion, the proportion was only 2.6% (95% CI 2.0-3.5%).

Given the rarity of these meta-analyses, we questioned whether the population that we selected were unusual, whether they were typical of how an average meta-analysis would end up when it became large enough to reach a reasonable RIS. In particular, we questioned whether a Cochrane meta-analysis is less likely to be updated if its conclusion was statistically significant. We hypothesised that a significant conclusion reduced the probability of consequential updating, and that our selected population therefore represented meta-analyses that were less likely to have had statistically significant results early on. If this hypothesis were true, then any false positives present in early Cochrane meta-analyses would be less likely to reach a reasonable diversity-adjusted RIS and less likely to be included in our study. To test this hypothesis, we reviewed all the Cochrane systematic reviews from 2005, 2006, and 2007 and found that a non-significant Cochrane meta-analysis is 1.57 times more likely to be updated (95% CI 0.92-2.68). While this observation did not reach statistical significance, it does suggest support to our hypothesis and leads to a suspicion that the proportion of early Cochrane meta-analyses with false positives may be higher than the 7% that we found. Adjusted for this risk of lack of updating in Cochrane reviews, the 7% converts to 11% (95% CI 5-22%) and suggests that the proportion of false positives in meta-analyses when the required information size has not been reached is higher than 5%.

We were also concerned about the heterogeneity of the population of meta-analyses that we included. Our selection criteria required a meta-analysis to have reached a defined RIS, it was more likely for a meta-analysis to be included if it had a lower heterogeneity (and consequently a lower RIS). In our selected population, the meta-analyses that did produce false positives had a higher diversity than those that did not (difference in means 24, 95% CI -4 – 53). This characteristic of our selected population represents a second reason why the actual false positive rate in early Cochrane meta-analyses may be higher than the one we found.

The selection of the included meta-analyses was further limited because we did not incorporate an assessment of risk of systematic error in this selection process. Consideration

of the risk of systematic errors according to the seven domains outlined in The Cochrane Handbook (5) is an extensive process and was not included in the scope if this study. Therefore, the result of the TSA analyses could be regarded as assessing the risk of random error under the assumption that all included trials had low risk of bias. Under this assumption, negative TSA analyses would conclude that even assuming low risk of bias for all the included trials, the risk of random error for concluding at least a beneficial effect would exceed the 5% type 1 error risk. If the boundary for benefit were crossed, then it becomes crucial that this apparently low risk of random error is not induced by an undue bias risk due to one or more trials having high risk of systematic errors. The omission of a full bias assessment represents a major limitation in our investigation. In order to define an early crossing of a threshold for statistical significance as a false positive, logic holds that the final conclusion must be that the intervention has no effect. If there were trials included in the final meta-analysis which had inflated effect estimates due to bias (which there undoubtedly were), this classification – of these meta-analyses as being ones where the question was reasonably answered – is not valid. Unfortunately, if we had undertaken a formal assessment of risk of bias, given the rarity of the meta-analyses we sought and the high prevalence of increased risk of bias in included trials, it is highly unlikely we would have found sufficient meta-analyses to conduct any investigation. Consequently, we concede this currently unavoidable limitation of the potential effect of systematic error.

It can be difficult to define credible parameter estimates to inform RIS for a TSA. The challenge of this definition is similar to that when calculating power for a clinical trial, but where the power calculation for a clinical trial occurs as a controlled prospective exercise, the probability modelling associate with TSA takes place as an uncontrolled exercise, often at multiple time points during the accumulation of data. In the setting of meta-analysis, a clear clinical question can inform values for the proportion of events in the control group. The decision with regard the a relevant effect size will always be challenging, in the same way as it is for a clinical trial, resting a reasonable quantity in the context of the given clinical scenario. With regard to heterogeneity, we know that estimates in early meta-analysis are unreliable(65) and using a range may turn out to be the most appropriate approach.

For both of the studies investigating TSA in this Ph.D., we used different TSA approaches, varying the parameters used for calculating RIS. When investigating the ability of TSA to prevent false positives, we used three approaches. For the first approach, we used an

approach that aimed to represent a 'credible parameters TSA approach', aiming to mimic what might have been chosen as credible and reasonable choices for the clinical question being asked. For the second approach, we used parameter estimates from the trials included when the false positive occurred and the D^2 estimated from the trials included up until that point. We used an un-weighted mean of the proportion of events in the control groups at that as the estimate of the proportion of events in the control group, and we used the border of the 95% confidence interval closest to the null at the time of the false positive as the estimate of effect. For the third approach, we again used parameter estimates from the trials included at the time when the false positive occurred, but this time we used the point estimate at the time of the false positive as the estimate of effect. The second and third approaches represented 'existing data TSA approaches', where parameter estimates for the TSA approach are chosen from the trials that have been included up until that point in time.

All three approaches performed well in preventing the false positives, but the third approach was least effective. Approach one prevented the false positives 93% of the time (95% CI 64-100%). Approach two prevented the false positives 86% of the time (95% CI 56-97%). Approach three prevented the false positives 79% of the time (95% CI 49-94%). It makes sense that the TSA approaches using the point estimate were not as helpful. We know that an early nominally statistically significant result increases the probability and an effect estimates is inflated (3, 66). Using an inflated effect estimate to do a TSA creates a probability model that is not consistent with the situation that it is intended to model. However, using the full range of uncertainty in the 95% CI and the least probable RRR seems to eradicate more early false positive meta-analyses.

We used the findings in Project 2 to choose which TSA approaches we used in Project 3, where we assessed the reliability of conclusions in meta-analyses of anaesthesiological interventions. We used two main TSA approaches for this study – a 'credible parameters TSA approach' using a RRR of 20% and a 'existing data TSA approaches' using the border of the conventional 95% confidence interval closest to the null. We didn't vary the estimates for control event proportion and heterogeneity, using the values present in the included trials and this omission does represent a limitation. We found, using two reasonable TSA approaches, only 30% (95% CI 18-45%) preserve a risk of type 1 error less than 5%, considering sparse data and repetitive testing on accumulating data. Moreover, our brief examination of the risk of systematic error in this population of systematic reviews revealed

that 98% (95% CI 88-100%) of anaesthesiological meta-analyses either include trials that have been assessed as having increased risk of bias or have had no formal assessment of bias risk. Despite the challenges and dynamic nature of the choice of parameters for the RIS calculation, our exploration was sufficient to allow us to conclude that a large proportion of published conclusions from meta-analyses of anaesthesiological interventions may be unreliable.

Findings in the context of clinical decision making and research

The focus of the three methodological papers in this Ph.D. project was how multiplicity in systematic reviews can affect the reliability of conclusions. Our findings suggest that the sequential multiplicity in systematic reviews may cause traditional estimates of uncertainty to be overly optimistic. That is, the conclusions may not be as precise as they claim to be. Furthermore, we demonstrated that TSA – a methodology designed to correct for the increased random error caused by sequential multiplicity and sparse data – can assist us in summarising with more accurate estimates of uncertainty.

This focus on sequential multiplicity, and its effect on precision in results, represents only one aspect of the overall picture of evidence-based medicine and how evidence is interpreted and implemented in clinical practice. It is the implementation or rejection of the intervention into clinical practice that is the important endpoint in the whole process. Multiplicity in systematic reviews may compromise the reliability of the conclusions of systematic reviews. Inaccurate assessments of uncertainty have implication for guideline development and for every-day clinical decision making. In anaesthesia, guideline developers are increasingly aiming to incorporate systematic evidence-based techniques(67-70). Several systematic approaches are available to assist in this goal and I describe here the structure provided by the GRADE working group(71). I discuss GRADE in the context of how evidence can be used to answer a clinical question generally, and how the issue of multiplicity fits in to this picture. Following, I discuss how the issue of multiplicity in meta-analysis can affect research decisions, and I provide some examples from published literature of cases where the more accurate estimate of uncertainty provided by TSA has altered a future research agenda.

The GRADE Working Group has developed an approach to grading the quality of evidence and the strength of recommendations when producing clinical guidelines(71). GRADE is a structured and thorough approach, and while it was intended for the development of the guidelines and recommendations, it is a helpful template on which to consider how to incorporate evidence into clinical decision making generally(67). Indeed, the Cochrane Collaboration recommends using the principles of the GRADE system for just this purpose, using it in the overall summary of the findings of a systematic review, allowing for clinical interpretation of what the systematic review has shown(5).

To use GRADE to apply evidence to a clinical question, the first step is to clearly define that clinical question. GRADE defines a clinical question using PICO. The PIC stands for population, intervention, and comparison(72). This process does not incorporate open-ended questions such as: how should I anaesthetise a patient having a liver transplant? Rather, this process aims to answer a defined question such as: for patients with hepatitis C presenting for a liver transplant, should I give drug x routinely or only when clinically indicated?

The O in PICO stands for outcomes. In order to answer the clinical question, a decision has to be made about which clinical outcomes are important(72). This judgement may be subjective, may alter for different clinical contexts, and may include issues regarding resource management. Independent of the subjectivity, in order to answer the question, the important outcomes – both advantageous and disadvantageous – need to be defined.

For the defined PICO, the next step in the GRADE process is to collect all the available evidence and then assess the quality of that evidence. The collection of a body of evidence for a defined clinical question is – by definition - a systematic review. As used by GRADE, the quality of a body of evidence refers to the extent to which we can be confident of the estimated effect of an intervention on a specific outcome(73). For a body of evidence to be considered as high quality, you need: a reasonable number of trials with a low risk of bias(74), evidence that publication bias was unlikely(75), precision in the overall results when a meta-analysis is performed(76), consistency between the results of the trials(77), and directness(78).

When the quality of the evidence has been assessed for each important outcome, then the advantages of the intervention can be weighed up against the disadvantages and a decision can be made about whether to use drug x routinely or not.

The issue of multiplicity is relevant in this GRADE process when considering the precision for the assessment of the quality of the evidence for each outcome. It is only one part of this overall process, but it is a very important part. Specifically, sequential multiplicity in systematic reviews with meta-analysis increases the risk of random type 1 error(58). Internal multiplicity in systematic reviews may also contribute to increased risk of type 1 error(57). This increased risk translates into less precision and, using GRADE nomenclature, therefore a lower quality of evidence. With regard to systematic reviews of anaesthesiological interventions quoting statistically significant conclusions, the majority of the meta-analyses are under-powered and many do not maintain their statistical significance when the increased risk of random type 1 error is included in the assessment using TSA. In the context of the GRADE structure, these meta-analyses represent bodies of evidence for clinical questions that have less precision than their authors claim.

Issues relating to multiplicity and increased random type 1 error may result in meta-analyses with conclusions of statistical significance followed by subsequent convincing research that contradicts these results. Methodology that corrects for the increased risk of random type 1 error – such as TSA – has the ability to prevent these early spurious conclusions by better estimating the uncertainty in these findings. The improved estimate of uncertainty has implications for guideline development and individual physicians making clinical decisions, as outlined above. The improved estimate also has implications for decision making about future trials. One of the findings of the second project in this thesis suggested that a Cochrane meta-analysis that is statistically significant is less likely to be updated subsequently than one that is not statistically significant(58). The demonstration, using TSA, that the risk of type 1 error is greater than 5% (despite nominal statistical significance using conventional techniques) may contribute to the decision about conducting further randomized clinical trials.

Apart from our own systematic review done as part of this Ph.D. project, TSA has already been helpful in anaesthesia research assessing the strength of evidence and planning future clinical trials. Examples of anaesthesia–related interventions that have been explored using TSA include perioperative beta-blockade(79) and hydroxyethyl starch(80, 81) A recent example from outside the anaesthesia literature is the effect of vitamin D supplementation on skeletal, vascular, or cancer outcomes (82).

I present here two specific examples, relevant to anaesthesia, where early meta-analyses report a statistically significant conclusion, TSA at that time disputed the statistical significance and suggested that the risk of random type 1 error was higher than 5%, and then further well-powered trials have gone on to conclude that the intervention has no such effect.

The first example involves the question of perioperative inspired oxygen and wound infection. In 2009, a systematic review published by Qadan et al. reported in its abstract a statistically significant reduced infection rate for patients receiving a hyperoxic gas mixture compared with control(83). At a similar time, Meyhoff et al. published a meta-analysis using TSA for the same clinical question(84). In conjunction with an alteration in meta-analytic technique, the TSA demonstrated that evidence for this question was not definitive(84).

Figure 3. Hyperoxia vs control and wound infection, prior to the PROXI trial(85). TSA reproducing that published by Meyhoff et al. in 2008(84). Using random-effects meta-analysis and based on a diversity-adjusted RIS of 4500, calculated using a RRR of 33%, control event proportion of 14%, a type 1 error risk of 5% and a power of 80%.

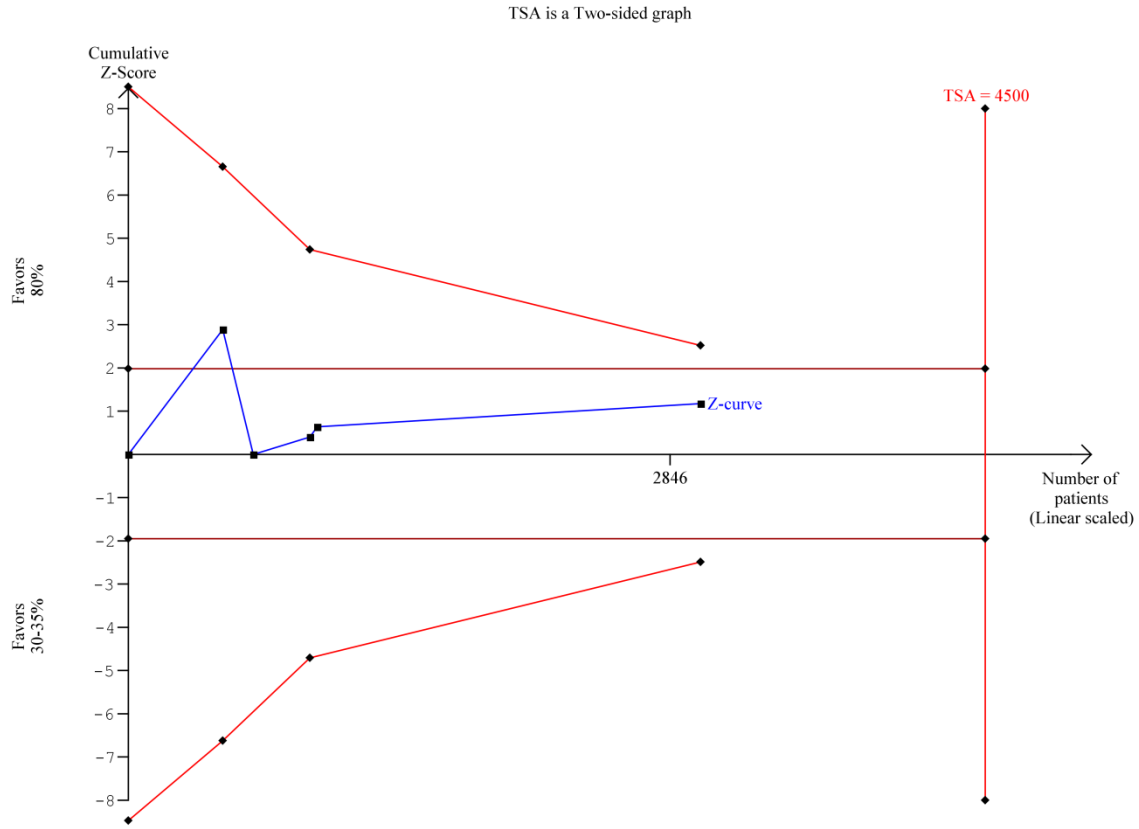
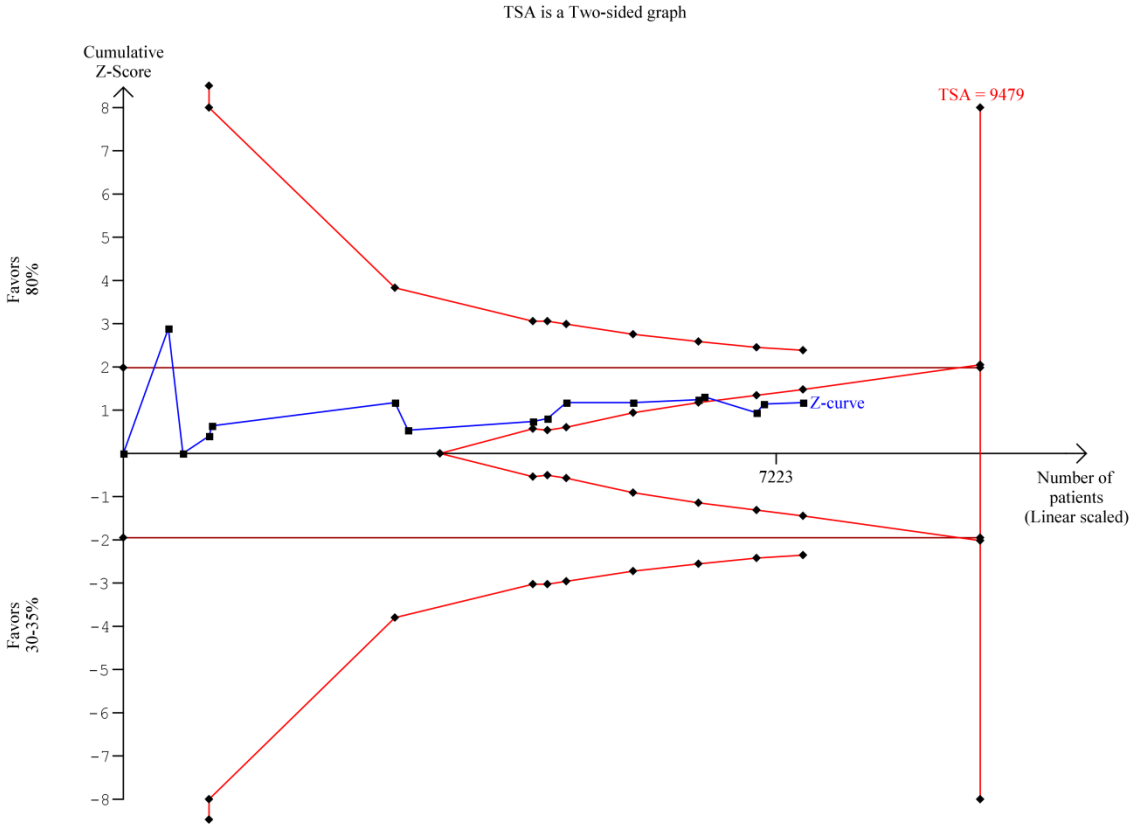


Figure 3 shows a reproduction of the TSA that was presented in this publication. The analysis was performed using a random effects meta-analysis model, as there was substantial heterogeneity present. While the change to the random effects model contributed to altering the conclusion here, this TSA also demonstrates how early crossings of the conventionally boundary for statistical significance can be misleading.

Figure 4. Hyperoxia vs control and wound infection, including the PROXI trial(85) and those since. Using random-effects meta-analysis and based on a diversity-adjusted RIS of 9479, calculated using a RRR of 20%, control event proportion of 16%, a type 1 error risk of 5% and a power of 80%.

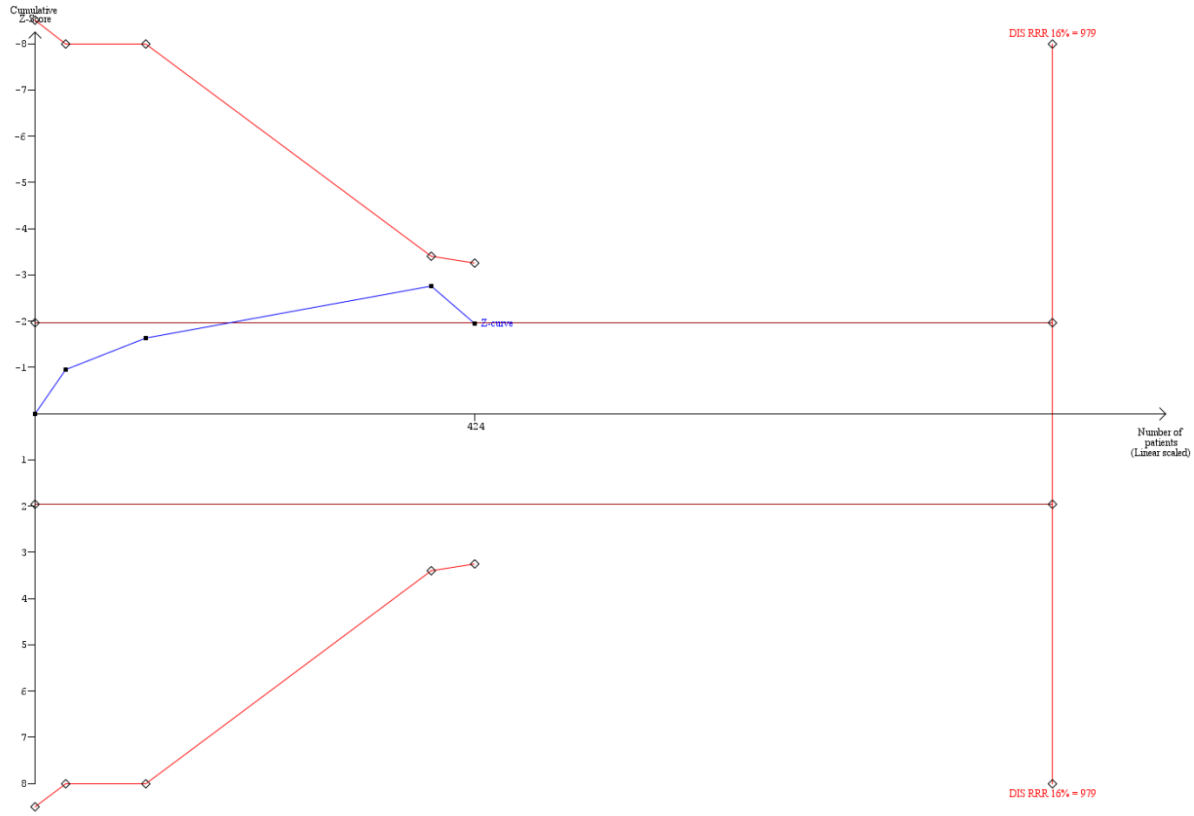


Subsequently, based on this assessment of uncertainty, Meyhoff et al. conducted a well-powered randomized clinical trial with overall low risk of bias and found no statistically significant difference when comparing 80% oxygen with 30% oxygen in the risk of surgical site infection after abdominal surgery(85). Other randomized trials followed(86-91). The TSA in Figure 4 shows that with these extra trials added, the evidence now passes a boundary for futility, a clear suggestion that the early conclusions of statistical significance were

spurious and an illustration of the utility of TSA. The evolution of this cumulative meta-analysis also demonstrates how the parameters used to estimate RIS will change as more evidence accumulates, and that the TSA approach is therefore a dynamic one.

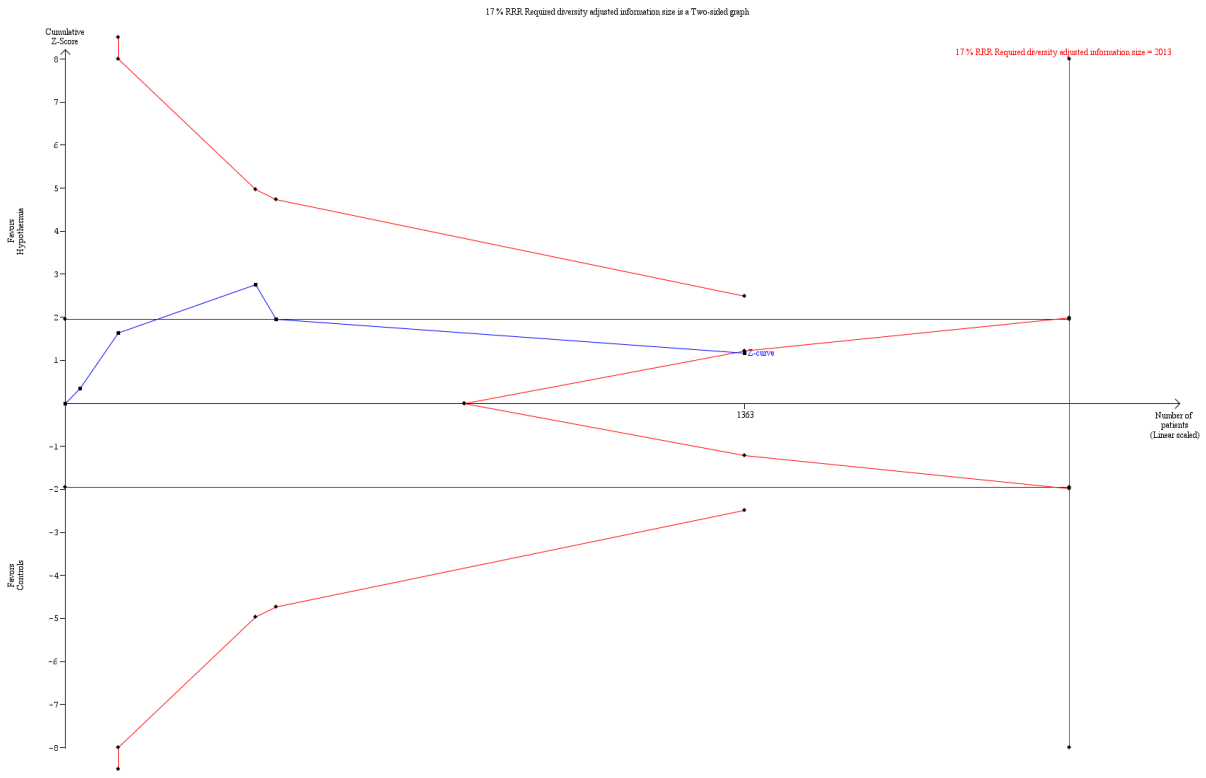
The second example involves the question of targeted temperature management at 33°C degrees versus 36°C after cardiac arrest. A meta-analysis with TSA was published in 2011 by Nielsen et al. comparing mild hypothermia with control after cardiac arrest, with all-cause mortality as the outcome(92). Despite crossing the conventional threshold for statistical significance, this TSA demonstrated that there was insufficient evidence to reject or detect the intervention effect of RRR of 16% which was present in the existing trials with low risk of bias(92). Figure 5 shows a reproduction of the TSA in this publication.

Figure 5. Hypothermia vs control after cardiac arrest and all-cause mortality.
TSA reproducing that published by Nielsen et al. in 2011(92) Using random-effects meta-analysis and based on an diversity-adjusted RIS of 979, calculated using a RRR of 16%, control event proportion of 59%, a type 1 error risk of 5% and a power of 80%.



After this TSA, with a consideration of the uncertainty reflected here, Nielsen et al. conducted a well-powered randomized clinical trial with low risk of bias and found that in unconscious survivors of out-of-hospital cardiac arrest, hypothermia at a targeted temperature of 33°C did not confer a benefit as compared with a targeted temperature of 36°C. The TSA in Figure 6 shows the TSA after the results of the extra trial were added, again demonstrating an example of an early statistically significant conclusion being spurious and TSA's ability to identify this uncertainty. The TSA now actually shows that we can refute a 17% RRR as the cumulative z-curve has reached the futility area for such an effect. This TSA also demonstrates the dynamic nature of the RIS estimate.

Figure 6. Hypothermia vs control after cardiac arrest and all-cause mortality.
TSA including the data from the randomized control trial by Nielsen et al (93) Using random-effects meta-analysis and based on an diversity-adjusted RIS of 2013, calculated using a RRR of 17%, a control event proportion of 62%, a type 1 error risk of 5% and a power of 80%.



Implications for future research

The findings in the methodological work in this thesis are explorative and they stimulate further questions about the complex issue of multiplicity in the context of systematic reviews. The mathematics of how multiple statistical tests will alter the probability of error is clear: the more times you roll a dice, the more likely you are to roll a six. The difficulty, in the context of making conclusions in systematic review, is how to place meaning in the context of such probability.

In this project, we did not aim to answer the question of how we should handle internal multiplicity in systematic reviews. Rather, we provided quantitative evidence that the issue is present and an argument that it warrants more attention. Further research needs to focus on this discussion of when and how internal multiplicity should be approached in systematic reviews.

The investigation of sequential multiplicity suggests several important directions for further research. The TSA approach requires estimates for heterogeneity, control event proportion, and effect size. These estimates work like assumptions in a probability model, but may be loaded by data accumulated until the time point when the TSA is performed, mimicking some kind of adaptive TSA corresponding to an adaptive trial design. Future research needs to explore how to best choose these assumptions, how to incorporate and communicate the uncertainty in these choices, and how they alter the inferential conclusion.

In this project, the focus was on the risk of type 1 error, and much of the observation was in the context of statistical significance being declared prematurely. Issues of sequential multiplicity also affect type 2 error; the risks are of course intertwined. TSA methodology is able to estimate the risk of type 2 error in the context of repeated testing and sparse data, using the same theoretical premise as it uses for estimating type 1 error. The accurate assessment of type 2 error and futility in meta-analysis is a further, important future direction of methodological research in this area.

As mentioned in the background section, there have been other methodologies proposed to address the issue of sequential multiplicity in cumulative meta-analysis(24, 25, 30, 38). We need comparisons and exploration of these various approaches in the context of them being used and interpreted by the authors and consumers of systematic reviews, focusing on the performance of each approach and its usability.

Conclusions

In this Ph.D. project, we explored the issue of random type I error in systematic reviews with meta-analysis and the effect on the reliability of conclusions. Specifically, we focused on the issue of multiple statistical comparisons (multiplicity) and sparse data in systematic reviews of anaesthesiological interventions.

With regard to the methodological investigations, we concluded that the quantity of internal multiplicity in systematic reviews is high over-all and higher in Cochrane reviews than in non-Cochrane reviews. Few systematic reviews, whether Cochrane or non-Cochrane, address the issue of multiplicity. We also concluded that false positive findings are likely to occur more frequently in real-life cumulative meta-analysis than the desired 5% and that TSA is a helpful statistical methodology to prevent increased risk caused by sequential multiplicity and to assess the reliability of early nominally statistically significant findings in cumulative meta-analyses. Finally, we concluded that TSA demonstrates that due to the increased risk of type 1 random error present in early meta-analyses, a large proportion of published conclusions from meta-analyses of anaesthesiological interventions may be unreliable.

With regard to the application of the investigated methodology, we conducted our own systematic review with meta-analysis and applied TSA. We concluded that, prior to the completion of ENIGMA II, we do not have robust evidence for how nitrous oxide used as part of general anaesthesia affects mortality and cardiovascular complications.

An overriding premise of the investigation in this Ph.D. project was that uncertainty in conclusions should be estimated and communicated as accurately as possible. Reducing the risk of type 1 error in our estimates of precision will always result in a conclusion having more uncertainty. The thing about uncertainty is that physicians do not like it. Patients do not like it either. When it comes to health interventions, we prefer to have distinct answers, clarity in our course of action. Unfortunately, however, underestimating uncertainty doesn't diminish its presence. A p-value of 0.05 means that there is a 5% probability of having obtained the data we have, or data more extreme, if the truth is that the null hypothesis in the population is true. Similarly, if we hypothetically repeated the whole meta-analysis process an infinite number of times, sampling independently from the same population, and

calculated a 95% confidence interval for a summary parameter for each meta-analysis, then 95% of these intervals should include the true (unknown) population parameter. The definitions of this frequentist approach to estimating uncertainty are certainly convoluted, and their limitations and difficulties are well and widely discussed(94, 95). Leaving the contentious discussion of these limitations aside, p-values and confidence intervals remain the main techniques used in conclusions of systematic reviews with meta-analyses. If we want to continue to use these techniques to estimate precision in meta-analyses, and if we wish this estimate to be an accurate assessment of the uncertainty caused by random error, we need to address the issue of multiplicity and sparse data in systematic reviews.

References

1. Lyman GH, Kuderer NM**.** The strengths and limitations of meta-analyses based on aggregate data. BMC Medical Research Methodology. 2005;5:14.
2. Olkin I**.** Meta-analysis: current issues in research synthesis. Statistics in Medicine. 1996;15(12):1253-7; discussion 9-62.
3. Pereira TV, Ioannidis JP**.** Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. Journal of Clinical Epidemiology. 2011;64(10):1060-9.
4. Patsopoulos NA, Analatos AA, Ioannidis JP**.** Relative citation impact of various study designs in the health sciences. JAMA. 2005;293(19):2362-6.
5. Higgins JPT GS, eds,**.** Cochrane Handbook for Systematic Reviews of Interventions. Chichester, UK: JohnWiley & Sons; 2011.
6. Keus F, Wetterslev J, Gluud C, van Laarhoven CJ**.** Evidence at a glance: error matrix approach for overviewing available evidence. BMC Medical Research Methodology. 2010;10:90.
7. Bender R, Lange S**.** Adjusting for multiple testing--when and how? Journal of Clinical Epidemiology. 2001;54(4):343-9.
8. Bender R, Bunce C, Clarke M, Gates S, Lange S, Pace NL, et al. Attention should be given to multiplicity issues in systematic reviews. Journal of Clinical Epidemiology. 2008;61(9):857-65.
9. Bassler D, Montori VM, Briel M, Glasziou P, Guyatt G**.** Early stopping of randomized clinical trials for overt efficacy is problematic. Journal of Clinical Epidemiology. 2008;61(3):241-6.
10. Gordi T, Khamis H**.** Simple solution to a common statistical problem: interpreting multiple tests. Clinical Therapeutics. 2004;26(5):780-6.
11. Gracely E**.** So, why do I have to correct for multiple comparisons? Concepts and commentary on Turk et al. Pain. 2008;139(3):481-2.
12. Levin B**.** On the Holm, Simes, and Hochberg multiple test procedures. American Journal of Public Health. 1996;86(5):628-9.
13. McPherson K**.** Statistics: the problem of examining accumulating data more than once. The New England journal of medicine. 1974;290(9):501-2.

14. O'Brien PC, Fleming TR. A Multiple Testing Procedure for Clinical Trials. Biometrics. 1979;35(3):549-56.
15. Perneger TV. What's wrong with Bonferroni adjustments. BMJ. 1998;316(7139):1236-8.
16. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology. 1990;1(1):43-6.
17. Schulz KF, Grimes DA. Multiplicity in randomised trials I: endpoints and treatments. Lancet. 2005;365(9470):1591-5.
18. Streiner DL. From the corrections officer: why we correct for multiple tests. Canadian Journal of Psychiatry - Revue Canadienne de Psychiatrie. 2009;54(6):351-2.
19. Young SS. Acknowledge and fix the multiple testing problem. International Journal of Epidemiology. 2010;39(3):934; author reply -5.
20. Benda N, Bender R. Multiplicity issues in clinical trials. Biometrical Journal. 2011;53(6):873-4.
21. Biester K, Lange S. The multiplicity problem in systematic reviews [abstract]. XIII Cochrane Colloquium; 2005 Oct 22-26; Melbourne, Australia; 2005:153.
22. Berkey CS, Mosteller F, Lau J, Antman EM. Uncertainty of the time of first significance in random effects cumulative meta-analysis. Controlled Clinical Trials. 1996;17(5):357-71.
23. Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may be inconclusive - Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. International Journal of Epidemiology. 2009;38(1):287-98.
24. Higgins JPT, Whitehead A, Simmonds M. Sequential methods for random-effects meta-analysis. Statistics in Medicine. 2011;30(9):903-21.
25. Hu M, Cappelleri JC, Lan KK. Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. Clinical Trials. 2007;4(4):329-40.
26. Pogue JM, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. Controlled Clinical Trials. 1997;18(6):580-93; discussion 661-6.
27. Thorlund K, Anema A, Mills E. Interpreting meta-analysis according to the adequacy of sample size. An example using isoniazid chemoprophylaxis for tuberculosis in

purified protein derivative negative HIV-infected individuals. Clinical Epidemiology. 2010;2(1):57-66.

28. Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JPA, Thabane L, et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? International Journal of Epidemiology. 2009;38(1):276-86.
29. Wetterslev J, Thorlund K, Brok J, Gluud C**.** Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. Journal of Clinical Epidemiology. 2008;61(1):64-75.
30. Whitehead A**.** A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. Statistics in Medicine. 1997;16(24):2901-13.
31. Wetterslev J, Engstrøm, J, Gluud, C & Thorlund, K Trial sequential analysis: methods and software for cumulative meta-analyses. Cochrane Methods. Cochrane DB Syst Rev. 2012;vol Suppl 1(no. 1-56):29-31.
32. DeMets DL, Lan KKG**.** Interim analysis: The alpha spending function approach. Statistics in Medicine. 1994;13(13-14):1341-52.
33. Pocock SJ**.** Interim analyses for randomized clinical trials: the group sequential approach. Biometrics. 1982;38(1):153-62.
34. SJ. P**.** Clinical trials, a practical approach. First edition. Chichester: John Wiley & Sons; 1983.
35. Borm GF, Donders ART**.** Updating meta-analyses leads to larger type I errors than publication bias. Journal of Clinical Epidemiology. 2009;62(8):825-30.e10.
36. Turner RM, Bird SM, Higgins JP**.** The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. PLoS ONE [Electronic Resource]. 2013;8(3):e59202.
37. Brok J, Thorlund K, Gluud C, Wetterslev J**.** Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. Journal of Clinical Epidemiology. 2008;61(8):763-9.
38. van der Tweel I, Bollen C**.** Sequential meta-analysis: an efficient decision-making tool. Clinical trials (London, England). 2010;7(2):136-46.
39. Thorlund K EJ, Wetterslev J, Brok J, Imberger G, Gluud C. User manual for trial sequential analysis (TSA). Copenhagen, Denmark: Copenhagen Trial Unit; 2011.
40. Wetterslev J, Thorlund K, Brok J, Gluud C**.** Estimating required information size by quantifying diversity in random-effects model meta-analyses. BMC medical research methodology. 2009;9:86.

41. Lan KK, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika. 1983;70(3):659-63.
42. Jennison C, Turnbull BW. Group Sequential Methods with Applications to Clinical Trials: Taylor & Francis; 2010.
43. Armitage P. Sequential Analysis in Therapeutic Trials. Annual Review of Medicine. 1969;20(1):425-30.
44. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977;64(2):191-9.
45. Borm GF, Donders AR. Updating meta-analyses leads to larger type I errors than publication bias. J Clin Epidemiol. 2009;62(8):825-30 e10.
46. Ioannidis J, Lau J. Evolution of treatment effects over time: empirical insight from recursive cumulative metaanalyses. Proc Natl Acad Sci U S A. 2001;98(3):831-6.
47. Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. Lancet. 1998;351(9095):47-52.
48. Imberger G, Wetterslev J, Gluud C. Trial sequential analysis has the potential to improve the reliability of conclusions in meta-analysis. Contemporary Clinical Trials. 2013;36(1):254-5.
49. Baum VC, Willschke H, Marciniak B. Is nitrous oxide necessary in the future? Paediatric Anaesthesia. 2012;22(10):981-7.
50. Myles PS, Leslie K, Silbert B, Paech MJ, Peyton P. A review of the risks and benefits of nitrous oxide in current anaesthetic practice. Anaesthesia & Intensive Care. 2004;32(2):165-72.
51. Weimann J. Toxicity of nitrous oxide. Best Practice & Research. Clinical Anaesthesiology. 2003;17(1):47-61.
52. Badner NH, Drader K, Freeman D, Spence JD. The use of intraoperative nitrous oxide leads to postoperative increases in plasma homocysteine. Anesthesia & Analgesia. 1998;87(3):711-3.
53. Myles PS, Chan MT, Kaye DM, McIlroy DR, Lau CW, Symons JA, et al. Effect of nitrous oxide anesthesia on plasma homocysteine and endothelial function. Anesthesiology. 2008;109(4):657-63.
54. Rao LK, Francis AM, Wilcox U, Miller JP, Nagele P. Pre-operative vitamin B infusion and prevention of nitrous oxide-induced homocysteine increase. Anaesthesia. 2010;65(7):710-5.

55. Homocysteine Studies C**.** Homocysteine and risk of ischemic heart disease and stroke: a meta-analysis. JAMA. 2002;288(16):2015-22.
56. Myles PS, Leslie K, Peyton P, Paech M, Forbes A, Chan MT, et al. Nitrous oxide and perioperative cardiac morbidity (ENIGMA-II) Trial: rationale and design. American Heart Journal. 2009;157(3):488-94.e1.
57. Imberger G, Vejlby AD, Hansen SB, Moller AM, Wetterslev J**.** Statistical multiplicity in systematic reviews of anaesthesia interventions: a quantification and comparison between Cochrane and non-Cochrane reviews. PLoS ONE [Electronic Resource]. 2011;6(12):e28422.
58. Imberger G, Thorlund K, Gluud C, Wetterslev J**.** False positive findings in cumulative meta-analysis with and without application of trial sequential analysis. Version for submission. 2014.
59. Imberger G, Gluud C, Boylan J, Wetterslev J**.** Results in systematic reviews of anaesthesiological interventions that claim to be statistically significant often contain risk of type 1 error greater than 5%. Version for submission. 2014.
60. Imberger G, Orr A, Thorlund K, Wetterslev J, Myles P, Møller AM**.** Does anaesthesia with nitrous oxide affect mortality or cardiovascular morbidity? A systematic review with meta-analysis and trial sequential analysis. British Journal of Anaesthesia. 2014;112(3):410-26.
61. Dmitrienko A, Hsu JC**.** Multiple Testing in Clinical Trials. Encyclopedia of Statistical Sciences: John Wiley & Sons, Inc.; 2004.
62. Bland JM, Altman DG**.** Multiple significance tests: the Bonferroni method. BMJ. 1995;310(6973):170.
63. Westfall PH, Young SS**.** Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment: Wiley; 1993.
64. Dmitrienko A, Soulakova JN, Millen BA**.** Three methods for constructing parallel gatekeeping procedures in clinical trials. J Biopharm Stat. 2011;21(4):768-86.
65. Thorlund K, Imberger G, Johnston BC, Walsh M, Awad T, Thabane L, et al. Evolution of heterogeneity (I2) estimates and their 95% confidence intervals in large meta-analyses. PLoS ONE [Electronic Resource]. 2012;7(7):e39471.
66. Thorlund K, Imberger G, Walsh M, Chu R, Gluud C, Wetterslev J, et al. The number of patients and events required to limit the risk of overestimation of intervention effects in meta-analysis--a simulation study. PloS one. 2011;6(10):e25491.

67. Imberger G**.** Clinical guidelines and the question of uncertainty. British Journal of Anaesthesia. 2013;111(5):700-2.
68. Kozek-Langenecker SA, Afshari A, Albaladejo P, Santullano CA, De Robertis E, Filipescu DC, et al. Management of severe perioperative bleeding: guidelines from the European Society of Anaesthesiology. European Journal of Anaesthesiology. 2013;30(6):270-382.
69. Kozek-Langenecker SA, Imberger G, Rahe-Meyer N, Afshari A**.** Reply to: ESA guidelines on the management of severe perioperative bleeding: Comments on behalf of the Subcommittee on Transfusion and Haemostasis of the European Association of Cardiothoracic Anaesthesiologists. European Journal of Anaesthesiology (EJA). 9000;Publish Ahead of Print:10.1097/EJA.0000000000000029.
70. De Hert S, Imberger G, Carlisle J, Diemunsch P, Fritsch G, Moppett I, et al. Preoperative evaluation of the adult patient undergoing non-cardiac surgery: guidelines from the European Society of Anaesthesiology. European Journal of Anaesthesiology. 2011;28(10):684-722.
71. Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A**.** GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. Journal of Clinical Epidemiology. 2011;64(4):380-2.
72. Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. Journal of Clinical Epidemiology. 2011;64(4):395-400.
73. Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. Journal of Clinical Epidemiology. 2011;64(4):401-6.
74. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). Journal of Clinical Epidemiology. 2011;64(4):407-15.
75. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. Journal of Clinical Epidemiology. 2011;64(12):1277-82.
76. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. Journal of Clinical Epidemiology. 2011;64(12):1283-93.

77. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. Journal of Clinical Epidemiology. 2011;64(12):1294-302.
78. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. Journal of Clinical Epidemiology. 2011;64(12):1303-10.
79. Bangalore S, Wetterslev J, Pranesh S, Sawhney S, Gluud C, Messerli FH**.** Perioperative beta blockers in patients having non-cardiac surgery: a meta-analysis. Lancet;372(9654):1962-76.
80. Haase N, Perner A, Hennings LI, Siegemund M, Lauridsen B, Wetterslev M, et al. Hydroxyethyl starch 130/0.38-0.45 versus crystalloid or albumin in patients with sepsis: systematic review with meta-analysis and trial sequential analysis. BMJ. 2013;346:f839.
81. Perner A, Haase N, Guttormsen AB, Tenhunen J, Klemenzson G, Aneman A, et al. Hydroxyethyl starch 130/0.42 versus Ringer's acetate in severe sepsis. N Engl J Med. 2012;367(2):124-34.
82. Bolland MJ, Grey A, Gamble GD, Reid IR**.** The effect of vitamin D supplementation on skeletal, vascular, or cancer outcomes: a trial sequential meta-analysis. The Lancet Diabetes & Endocrinology. 2014.
83. Qadan M, Akca O, Mahid SS, Hornung CA, Polk HC, Jr. Perioperative supplemental oxygen therapy and surgical site infection: a meta-analysis of randomized controlled trials. Archives of Surgery. 2009;144(4):359-66; discussion 66-7.
84. Meyhoff CS, Wetterslev J, Jorgensen LN, Henneberg SW, Simonsen I, Pulawska T, et al. Perioperative oxygen fraction - effect on surgical site infection and pulmonary complications after abdominal surgery: a randomized clinical trial. Rationale and design of the PROXI-Trial. Trials. 2008;9:58.
85. Meyhoff CS, Wetterslev J, Jorgensen LN, Henneberg SW, Hogdall C, Lundvall L, et al. Effect of high perioperative oxygen fraction on surgical site infection and pulmonary complications after abdominal surgery: the PROXI randomized clinical trial. JAMA. 2009;302(14):1543-50.
86. Williams N, Crisp C, Glover M, Downing C, McKenna D**.** 676: Randomized controlled trial evaluating the effect of variable FiO2 on cesarean delivery surgical site infection. American journal of obstetrics and gynecology. 2009;201(6):S244.

87. Scifres CM, Leighton BL, Fogertey PJ, Macones GA, Stamilio DM**.** Supplemental oxygen for the prevention of postcesarean infectious morbidity: a randomized controlled trial. Am J Obstet Gynecol. 2011;205(3):267 e1-9.
88. Schietroma M, Cecilia EM, Carlei F, Sista F, De Santis G, Piccione F, et al. Prevention of anastomotic leakage after total gastrectomy with perioperative supplemental oxygen administration: a prospective randomized, double-blind, controlled, single-center trial. Ann Surg Oncol. 2013;20(5):1584-90.
89. Bickel A, Gurevits M, Vamos R, Ivry S, Eitan A**.** Perioperative hyperoxygenation and wound site infection following surgery for acute appendicitis: a randomized, prospective, controlled trial. Arch Surg. 2011;146(4):464-70.
90. Duggal N, Poddatoori V, Noroozkhani S, Siddik-Ahmad RI, Caughey AB**.** Perioperative oxygen supplementation and surgical site infection after cesarean delivery: a randomized trial. Obstet Gynecol. 2013;122(1):79-84.
91. Golfam F, Golfam P, Golfam B, Mortaz SSS, Pahlevani P**.** PP-036 Effects of supplemental oxygen for prevention of wound infection after breast surgery. International Journal of Infectious Diseases. 2011;15, Supplement 1(0):S55.
92. Nielsen N, Friberg H, Gluud C, Herlitz J, Wetterslev J**.** Hypothermia after cardiac arrest should be further evaluated--a systematic review of randomised trials with meta-analysis and trial sequential analysis. Int J Cardiol. 2011;151(3):333-41.
93. Nielsen N, Wetterslev J, Cronberg T, Erlinge D, Gasche Y, Hassager C, et al. Targeted temperature management at 33degreeC versus 36degreeC after cardiac arrest. New England Journal of Medicine. 2013;369(23):2197-206.
94. Nuzzo R**.** STATISTICAL ERRORS. Nature. 2014;506(7487):150-2.
95. Goodman SN**.** Toward evidence-based medical statistics. 1: The P value fallacy. Annals of Internal Medicine. 1999;130(12):995-1004.