Evaluation of and agreement between Cochrane neonatal reviews and clinical practice guidelines for newborns in Denmark

Ph.D. Thesis

Jesper Brok

Copenhagen Trial Unit, Centre for Clinical Intervention Research, and

Department of Neonatology, Rigshospitalet,

Copenhagen University Hospital, Denmark

CONTENTS

PREFACE

I would like to thank my main supervisor, Christian Gluud, for providing me financial support and facilities, and most importantly, for his inspiring enthusiasm and hard-working effort that made this thesis a milestone in my medical education. I would also thank my supervisor Gorm Greisen who gave essential input about neonatal clinical practice and on my study design and papers. I acknowledge all the Danish neonatologists who kindly provided me with detailed information about their local clinical practice guidelines. I also acknowledge all my co-authors for constructive discussions.

I am grateful for three very pleasant years with the friendly staff at our small and cosy department, Copenhagen Trial Unit. I thank Dimitrinka Nikolova for help to improve my writing skills. I thank Sarah Klingenberg for literature searches, Mette Hansen for secretarial assistance, Nader Salas and Styrbjørn Birch for software support, and Kristian Thorlund for help on statistical issues. I thank Bodil Als-Nielsen for patiently introducing me to the principles of evidence-based medicine and Jørn Wetterslev for our many inspiring 'lunch' discussions about statistics, methodology, and medical ethics. Finally, I recall Peter Gøtzsche's and Henrik Wullf's inspiring lectures about evidence-based medicine which triggered my interest in this field.

To my lovely wife Winnie, who, despite our two wonderful but demanding newborn children and despite moving to a new house, never stressed and has always supported me and made this thesis possible.

ORIGINAL PAPERS

This PhD thesis was based on the following publications and manuscripts:

- Brok J, Greisen G, Jacobsen T, Gluud, LL, Gluud C. Agreement between Cochrane neonatal reviews and guidelines for newborns at a University Hospital. ***Acta Paediatrica*** 2007;96(1):39-43.[23]

- Brok J, Greisen G, Madsen LP, Tilma K, Faerk J, Børch K, Garne E, Christesen HT, Stanchev H, Jacobsen T, Nielsen JP, Henriksen TB, Gluud C. Agreement between Cochrane neonatal reviews and clinical guidelines for newborns in Denmark. ***Arch Dis Child Fetal Neonatal Ed*** 2008;93(3):225-9.[24]

- Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential boundaries improve interpretation of meta-analyses. ***J Clin Epidemiol*** 2008;Apr 12; [Epub ahead of print].[26]

- Brok J, Wetterslev J, Thorlund K, Gluud C. Apparently conclusive meta-analyses may often be inconclusive - a trial sequential analysis adjustment for random error risk in conclusive neonatal meta-analyses. ***Int J Epidemiol*** 2008 (resubmitted).[29]

- Brok J, Greisen G, Madsen LP, Tilma K, Faerk J, Børch K, Garne E, Christesen HT, Stanchev H, Nielsen JP, Henriksen TB, Wetterslev J, Gluud C. Agreement between Cochrane neonatal reviews adjusted for random error risk and guidelines for newborns in Denmark. Submitted 2008.[30]

ABSTRACT

Background Disagreements exist between research evidence and clinical practice. Cochrane systematic reviews with meta-analyses of randomised trials try to facilitate implementation of research evidence into clinical practice. Cochrane reviews may report misleading results due to random errors because of repetitive testing in meta-analyses. Trial sequential analysis (TSA) is a statistical approach that protects against such random errors in meta-analyses analogous to interim monitoring boundaries in a single trial. The AGREE instrument may be used to asses quality of clinical practice guidelines.

Objectives To assess (1) the agreement between Cochrane Neonatal Group reviews and clinical practice guidelines for newborns in Denmark, (2) meta-analyses in reviews using the TSA approach (3) whether TSA affects the conclusion in neonatal reviews recommending an intervention and how the new conclusions influence the level of agreement between reviews and guidelines, and (4) the quality of Danish clinical practice guidelines for newborns using the AGREE instrument.

Results The agreement between Cochrane neonatal reviews and neonatal guidelines is good (*Kappa* 0.56, range 0.52-0.59) despite the fact that guideline authors had rarely used the reviews for guideline development. Many 'conclusive' neonatal meta-analyses may become inconclusive if examined with TSA. TSA-adjustment of meta-analyses changed the conclusion from supporting the use of the intervention to inconclusive evidence in every second review. The TSA-adjustment significantly ($P < 0.05$) reduced the agreement between reviews and Danish clinical practice guidelines (*Kappa* 0.41, range 0.36-0.45). Most guideline authors lacked systematic methods to search for evidence, criteria for selecting the evidence, and references to the supporting evidence.

Conclusions Authors of Cochrane reviews and clinical practice guidelines aim to facilitate implementation of research evidence into clinical practice. However, they should collaborate to improve the quality of reviews and clinical practice guidelines. TSA may reduce the risks of introducing interventions based on statistical significant findings in meta-analysis based on random error because of repetitive testing.

INTRODUCTION

Facts about neonatology

Neonatology is a branch of paediatrics that deals with diseases and care of newborn infants. In the 1960s, the idea of having a special intensive care unit for newborns - a neonatal intensive care unit (NICU) - represented a developmental milestone for neonatology.[1] Subsequently, health-care workers have been able to improve care of many premature or desperately ill newborns. As a result infant mortality in countries with such facilities has shown a steady decrease and survivors have fewer sequelae.[1]

The ideas of evidence-based medicine were introduced early within neonatology. In the 1950s one of the first randomised clinical trials (RCT) was performed: allocation of preterm infants to adrenocorticotropic hormone (ACTH) or no ACTH was randomly decided by drawing marbles from a jar containing an equal number of white and blue marbles.[2] The question was whether ACTH reduced the risk of retinopathy. Although the pre-trial results of ACTH administration were promising, the RCT showed that retinopathy was not significantly different in the two groups. However, the mortality rates were quite disparate: the *untreated* group fared better.

Since then, more than 6,500 neonatal RCTs have been published. Important beneficial (eg, surfactant[3]) as well as harmful (eg, unrestricted oxygen[4]) intervention effects have been observed. In the 1980s, all RCTs were collected and published in one of the first electronical trial databases 'The Oxford Database of Perinatal Trials'.[5] This database including relevant RCTs resulted in a fundamental book in 1992, entitled 'Effective Care of the Newborn Infant'.[6] This book contained reviews of the effects of most therapies used at that time in neonatal care.

In Denmark, there are currently 17 NICU or neonatal sub-units of paediatric departments. Yearly, about 6,000 (10%) newborns are hospitalised with every second hospitalisation being due to premature delivery, ie, delivery before 37 weeks of pregnancy. The preterm delivery rates have been increasing, currently representing 5% to 6% of all births.[7] Preterm infant survival in Denmark is about 30% if gestational age (GA) is 24 weeks, increasing to 80% when GA is 28 weeks.[8]

The Cochrane Collaboration and systematic reviews

The Cochrane Collaboration is an international non-profit and independent organisation, dedicated to making up-to-date, accurate information about the effects of health-care widely available.[9] The Cochrane Collaboration produces systematic reviews of health-care interventions and was founded in 1993. It is named after the physician and epidemiologist Archie Cochrane.

To identify the benefits and harms of health-care interventions, results from similar randomised trials need to be combined mathematically to produce a more powerful and precise estimate of an intervention effect.[10,11] If the design of each study is similar then the results of the individual studies are combined to produce an overall statistic, a meta-analysis.[10,11] Meta-analyses of randomised trials are placed at the top of the evidence hierarchy, due to the ability to increase power and precision.[10,11] Through meta-analyses one may also be able to control bias.[10,11]

The combination of trials needs to be as systematic and trustworthy as possible. A systematic review uses a predefined, explicit methodology specified in a peer-reviewed protocol.[11] The methods used include steps to minimise systematic error (bias) in all parts of the process: identifying relevant studies, selecting them for inclusion, and collecting and combining their data.[11]

Systematic reviews are the cornerstone in the Cochrane Collaboration and aim to present information, rather than offer direct treatment guidelines. In the section 'implication for practice' of each systematic review, readers are helped in understanding the evidence in relation to practical decisions. Clinicians should then decide how applicable the evidence is to their local setting. The Cochrane Library 2004, Issue 2, included 1999 systematic reviews from 51 review groups.[9] The Cochrane Neonatal Review Group had conducted 170 (9%) of these reviews.[12]

Trial sequential analysis

Meta-analyses are not errorless and may produce misleading results due to systematic error (bias)[11,13] or due to random error ('the play of chance').[14] Most discussions have been focusing on bias risks whereas little attention has been paid to the risk of random

error.[13] Random error may arise through repeated significance testing when updating meta-analyses with new trials or through significance testing of multiple outcomes. This thesis focuses only on the risk of random error that is due to repeated significance testing of the same outcome in a meta-analysis.

Meta-analyses with a sparse number of events and participants are at particular risk of producing random error. When conducting a single randomised trial it is essential to have an a priori calculated sample size.[15-18] To avoid random errors, a meta-analysis intuitively needs to include a sample size at least as large as that of an adequately powered single trial. The medical communities have not paid much attention to the issues of sample size consideration and random errors in meta-analyses.

Trial sequential analysis (TSA) is an approach that may provide the required sample size in a meta-analysis.[18] Meta-analyses not reaching this sample size are analysed with trial sequential monitoring boundaries. This represents a more restrictive analysis, analogous to an interim monitoring boundary in a single trial. Trial sequential monitoring boundaries protect against random error.[19] Monitoring boundaries adjust the *P* value that is required for obtaining statistical significance according to the number of events and participants in a meta-analysis. The fewer events and participants, the more restrictive the monitoring boundaries are, and a lower *P* value is required to obtain statistical significance. We will apply TSA retrospectively on conducted meta-analyses, but of course, the idea is, that TSA should try to control random error when prospectively used in cumulative meta-analyses.[18]

Clinical practice guidelines

Clinical practice guidelines aim to help health professionals and patients make the best decisions about treatment or care for a particular condition or situation. International, national, and local groups can develop guidelines. Irrespective of the target group, guidelines have to be valid, usable, and reliable. Thus, it is important to ensure that guidelines are systematically developed like the systematic procedure for conducting Cochrane reviews.[20] Different tools to evaluate the quality of guidelines have been published.[21] Most of them aim to evaluate more comprehensive guidelines rather than the short and concise guidelines, which the clinician carries in his or her

pocket on the ward. Currently, the AGREE instrument, seems to be the only validated tool to assess clinical practice guidelines.[22]

AGREE consists of 23 key items organised in six domains. Each domain tries to capture a separate dimension of guideline quality: scope and purpose; stakeholder involvement; rigour of development; clarity and presentation; applicability; and editorial independence.[22]

OBJECTIVES

1. To assess the agreement between Cochrane Neonatal Group reviews and Danish clinical practice guidelines for newborn infants.

2. To evaluate meta-analyses in Cochrane Neonatal Group reviews using the TSA.

3. To assess whether TSA of meta-analyses affects the overall conclusion in Cochrane Neonatal Group reviews recommending an intervention and how the new conclusions influence the level of agreement between the reviews and Danish guidelines for newborns.

4. To evaluate the Danish neonatal clinical guidelines using the AGREE instrument.

METHODS AND RESULTS

Procedure and strategy

The Cochrane Neonatal Review Group was contacted in order to retrieve all their published reviews up until September 2004 (Review Manager 4.10 file). Two investigators classified the treatment recommendations in the reviews independently with disagreements solved through discussion. Subsequently, all Danish neonatal departments were contacted by e-mail and letter and invited to participate in the project. Concurrently, the objectives (page 9) of the project were presented at a National neonatal symposium in 2004. All departments willing to participate were asked to choose a contact person. This person was to provide the local clinical practice guidelines, provide further information on the guidelines if needed, and comment (in collaboration with other relevant guideline authors) on the assessment of agreement. Correspondence with the contact persons was primarily accomplished through e-mail.

The assessment of the agreement between reviews and guidelines was first carried out as a pilot project at a single department.[23] Experience from this study was used to further improve data extraction sheets, formulations in correspondence, and assessment of guidelines in the national study.[24] After evaluation of each department, a report was sent to the contact person who confirmed the interpretation of the local guidelines and the classification as being in (dis-)agreement with the reviews. The author of the thesis visited the majority of the participating departments and presented and discussed the individual results.

Evaluation of the review meta-analyses with TSA was performed with a computer program (TSA v 0.6) developed by the Copenhagen Trial Unit.[18] Data from the meta-analyses were extracted by the author of this thesis. Data extraction was verified by comparing the *Z*-score for each meta-analysis in TSA v 0.6 with the *Z*-score in Review Manager 4.10. Subsequently, we assessed the agreement between the reviews, which had their conclusions adjusted based on the results of the TSA, and the guidelines.

The final results were presented at the national neonatal symposium in 2007. The overall aim was to highlight the substantial amount of research from the Cochrane Neonatal Group and to initiate a debate about Danish guidelines for newborns.

1. Agreement between Cochrane Neonatal Group reviews and Danish clinical practice guidelines for newborn infants

Sample

We included all published Cochrane Neonatal Group reviews.[23,24] Two investigators independently extracted the treatment recommendations in the reviews applying a six-point scale (Table 1). Disagreements were solved through discussion. We included all Danish guidelines on interventions that had been evaluated in the Cochrane reviews. Guidelines recommending interventions not being evaluated in the reviews were excluded. The assessment of guidelines was done with knowledge of the findings in the Cochrane reviews.

Table 1. Graded classification (text abbreviated) of interventions assessed in the Cochrane Neonatal Group reviews.[23]

(I) Intervention should be abandoned (n = 7).
(II) Intervention should probably be abandoned (n = 23).
(III) Insufficient evidence to support or refute the intervention (n = 52).
(IV) Sparse evidence to support the use of the intervention (n = 38).
(V) Reasonable evidence to use the intervention (n = 31).
(VI) Clear evidence that the intervention should be used (n = 22).

Outcomes

Reviews and guidelines were classified[23,24] as being in

- *agreement:*

- Review and guideline recommend the intervention (grade V-VI, Table 1).
- Review lacks evidence (grade III-IV) or has evidence to refute the intervention (grade I-II) and the guideline does not recommend or address the intervention.

- *partial agreement:*

- Intervention with borderline evidence to use (grade V) and the guideline either recommends or does not recommend the intervention.

- disagreement:

- Review recommends the intervention (grade VI), but the guideline does not.
- Review lacks evidence (grade III-IV) or has evidence to refute the intervention (grade I-II) and the guideline recommends the intervention.

We searched the guidelines covering interventions assessed in the Cochrane neonatal reviews for references to the pertinent review. In cases when the Cochrane review was not mentioned, the guideline authors were asked which evidence they had considered for the guideline development.

Statistical methods

We determined the national median number of *agreements*, *partial agreements*, and *disagreements* between reviews and guidelines. We calculated a weighted kappa (*K)* (3 by 2 tables, partial agreement weighted 0.5) to estimate the agreement beyond chance.[25]

Results

Participation

Of 17 neonatal departments, 14 departments with all regions of Denmark represented agreed to participate in this audit.

Agreement between Cochrane reviews and clinical guidelines

173 interventions were assessed in the Cochrane reviews (Table 1). In our pilot study[23] we included 36 guidelines form one unit. In the national study[24] we included a total of 186 guidelines from 14 departments with a median of 12 (range 5-36) guidelines from each department, which addressed the interventions covered by the Cochrane reviews.

The median number of *agreements* between treatment recommendations in reviews and each department's guideline was 132/173 (76%) interventions (range 129-134).

Of these, 19/132 (14%) interventions were recommended as 'treatment of choice' in both the reviews and the guidelines. 113/132 (86%) interventions were neither recommended in the reviews nor in the guidelines (Table 2). The median number of *partial agreements* was 31/173 (18%) interventions (range 29-33).

The median number of *disagreements* between reviews and each department's guideline was 10/173 (6%) interventions (range 8-13) (Table 2). Of these, most interventions lacked evidence in the reviews (not recommended), but were recommended in the guidelines. A few interventions were recommended in the reviews, but not in the guidelines (Table 2).

Overall, the weighted *K* was 0.56 (range 0.52-0.59), which indicated good agreement (Table 2).

Table 2. Agreement between Cochrane Neonatal Group reviews and clinical practice guidelines in Denmark.

	Guidelines		
Grade*	**Recommended**	**Not recommended or not addressed**	Total
I	***0***	***7***	**120**
II	***0***	***23***	
III	***3***	***49***	
IV	***4***	***34***	
V	***12***	***19***	**31**
VI	***19***	***3***	**22**
Total	**38**	**135**	**173**
Disagreement	**Partial agreement**	**Agreement**	

* Classification of the interventions assessed in Cochrane reviews (Table 1).

Reasons for disagreements between Cochrane reviews and clinical guidelines

The reported reasons from the contact person for recommending interventions that lacked evidence according to the Cochrane reviews (score III-IV, Table 1) were: use of other evidence sources than reviews; single studies (both non-randomised and randomised); textbook recommendations; expert opinion; clinical experience; consensus statements; basic immunology and pathophysiological knowledge;

evidence based on intervention effects on surrogate markers; or intervention without risk of adverse event ('nothing to loose').

The reasons for not recommending interventions that were evidence-based according to the Cochrane reviews (score VI, Table 1) were: unawareness of the review; local consensus ('bad habit'); reservations about the external validity of the review (ie, locally, the basic treatment or the infants' risks differed substantially from the infants in the reviews); use of evidence from single studies; disagreement with the interpretation of the review; easier to administer alternative intervention(s); or economical constraints.

Regarding use of evidence from other studies, none of the studies referred to were published after the last update of the pertinent Cochrane review.

Use of Cochrane reviews for guideline development

We searched the guidelines for references to Cochrane reviews and asked the authors of the guideline whether the review had been considered for the guideline development. The search and the feedback showed that the pertinent Cochrane review was used only in a median of 10% (range 0-36%) of the guidelines.

2. Evaluation of meta-analyses in Cochrane Neonatal Group reviews with trial sequential analysis

Sample

From each Cochrane review we included meta-analyses on mortality and the first two eligible meta-analyses on clinical binary outcome measures according to the authors' priority in the review.[26]

Statistical methods

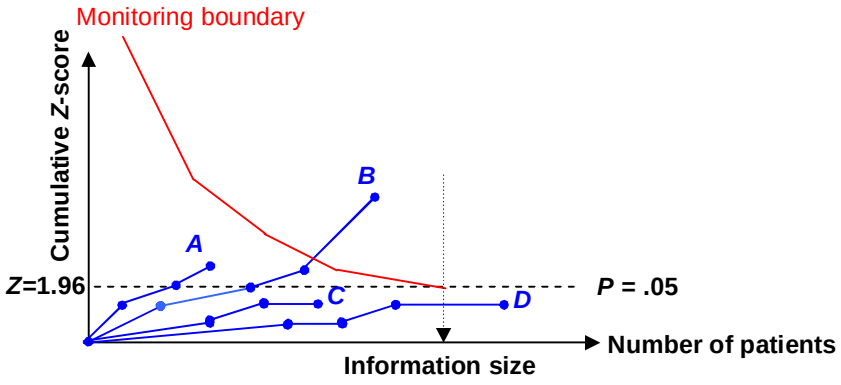
TSA necessitates pre-specification of a relevant (worthwhile) intervention effect and risk of type 1 (α) and type 2 (β) errors.[18] We set two-sided $\alpha = 5\%$, $\beta = 20\%$ ($1 - \beta = 80\%$ power) and used a 25% relative risk reduction (RRR) as an a priori estimate of a realistic treatment effect. With these data, the required information size (ie, the

number of participants in the meta-analysis required to accept or reject the pre-specified intervention effect) could be calculated and the adjacent trial sequential monitoring boundaries could be constructed (Fig 1, page 13). We adjusted the obtained information size and monitoring boundaries for heterogeneity (I^2).[18,27]

In the manuscript[26] we used three different a priori estimates of realistic treatment effects: 15% and 30% RRR and an intervention effect estimated by low bias-risk trials. However, to be consistent throughout this thesis I use only a 25% RRR, although I am aware that this represents a substantial reduction of information (please see Discussion).

For each meta-analysis we calculated the information size and applied the adjacent monitoring boundary. We calculated the cumulative *Z*-curve (ie, the series of consecutive Mantel-Haenszel *Z*-statistics after each trial) of each cumulative meta-analysis and assessed its relation to the information size and the monitoring boundary. The monitoring boundary should be crossed by the cumulative *Z*-curve to obtain evidence for a statistically significant intervention effect (Fig 1). *Z* values of +1.96 or –1.96 correspond to the conventional P = .05 in two-sided hypothesis test.

Figure 1. Examples of One-sided Trial Sequential Analyses.
The cumulative *Z*-curves (A-D) from four different meta-analyses were constructed with each cumulative *Z*-value calculated after including a new trial according to publication date. Crossing of *Z* = 1.96 provides a 'traditionally' significant result (A). Crossing of monitoring boundaries is needed to obtain reliable statistical evidence adjusted for random error risk (B).



Outcomes

The proportion of statistically significant (P < 0.05) meta-analyses that had

- 'potentially spurious evidence of effect', ie, the cumulative *Z*-curve did not cross the monitoring boundary (Fig 1, curve A)
- 'firm evidence of effect', ie, the cumulative *Z*-curve crossed the monitoring boundary (Fig 1, curve B)

The proportion of statistically non-significant (P > 0.05) meta-analyses that had

- 'absence of evidence', ie, the meta-analysis included less patients than the required information size (Fig 1, curve C)
- 'lack of effect', ie, the meta-analysis included more patients than the required information size (Fig 1, curve D).

Results

Characteristics of meta-analyses

We included 363 eligible meta-analyses. These meta-analyses included a median of 3 randomised trials (range 1-21) and a median of 193 participants (range 12-4,986). Of these meta-analyses, 113 (31%) were statistically significant (P < 0.05).

TSA of significant (P < 0 .05) meta-analyses

In 78/113 (69%) meta-analyses the cumulative *Z*-curve did not cross the monitoring boundary indicating potentially spurious evidence of effect. For these meta-analyses, the median additional information size needed to obtain firm evidence of a RRR of 25% was 1664 participants (range, 174-8,376). For 38/113 (31%) meta-analyses the *Z*-curve crossed the monitoring boundary showing 'evidence of effect'. However, none of the meta-analyses became statistically significant after the Z-curve had passed the required information size.

TSA of non-significant (P < 0 .05) meta-analyses

236/250 (94%) meta-analyses had 'absence of evidence' as they included less patients than the required information size in order to accept or reject meta-analytic evidence for a RRR of 25%. 14/250 (6%) meta-analyses showed lack of RRR of 25% as they included more patients than the estimated information size.

3. How trial sequential analysis of meta-analyses influence the conclusion in the reviews recommending an intervention and how the new conclusion influences the agreement between reviews and guidelines

Sample

All Cochrane neonatal reviews that found clear or moderate meta-analytic evidence to recommend an intervention for clinical use (grade V or VI, Table 1) were evaluated.[29] From these reviews we included all statistically significant ($P < 0.05$) meta-analyses assessing a binary outcome (n = 94).

Statistical methods

We applied TSA to all meta-analyses in Cochrane neonatal reviews supporting the use of an intervention.[29,30] According to the results of TSA, we adjusted the results of the meta-analyses and based on this we adjusted the overall conclusion of the review. Then, we assessed the agreement (*K*) between TSA-adjusted conclusions in reviews and the unadjusted recommendations of guidelines.

Results

Trial sequential analysis

Of 54 eligible Cochrane neonatal reviews, we included 45 reviews with 94 significant meta-analyses that favoured the recommended intervention. These meta-analyses included a median of 3 randomised trials (range 1-16) and a median of 394 participants (range 32-4,986).

In 22 of 45 (49%; 95% CI 33-64%) reviews, the cumulative *Z*-curve did not cross the monitoring boundaries in all the included meta-analyses with $P < 0.05$ (Fig 1, curve A) (Table 3). Such TSA-adjustment would change the conclusion of the reviews from clear to inconclusive evidence.

Table 3. Cochrane Neonatal Group reviews that may support the recommendation* of an intervention based on one or more statistically significant meta-analyses. However, TSA found that all the included meta-analyses lacked sufficient evidence.
Caffeine instead of theophylline for apnea in preterm infants[31]
Mechanical ventilation for newborn infants with respiratory failure due to pulmonary disease[32]
Restricted instead of liberal water intake for preterm infants[33]
Endhole instead of side-hole in umbilical catheters[34]

Theophylline instead of continuous positive airway pressure for apnea in preterm infants[35]
Multiple instead of single dose natural surfactant extract for severe neonatal respiratory distress syndrome[36]
Prophylactic vitamin K in neonates[37]
Kangaroo mother care in low birthweight infants[38]
Selenium supplementation in preterm neonates[39]
Synchronised mechanical ventilation for respiratory support in newborn infants[40]
Intravenous dexamethasone for extubation of newborn infants[41]
Elective high frequency jet ventilation instead of conventional ventilation for respiratory distress syndrome[42]
Prolonged instead of short course of indomethacin for the treatment of patent ductus arteriosus in preterm infants[43]
Indomethacin for asymptomatic patent ductus arteriosus in preterm infants[44]
Surfactant for meconium aspiration syndrome in full-term infants[45]
Prophylactic caffeine to prevent postoperative apnea following general anaesthesia in preterm infants[46]
Diuretics acting on the distal renal tubule for preterm infants with (or developing) chronic lung disease[47]
Pentoxifylline for neonatal sepsis[48]
Prophylactic intravenous antifungal agents in very low birth weight infants[49]
Diazepam for treating tetanus[50]
Opiate treatment for opiate withdrawal in newborn infants[51]
Sedatives for opiate withdrawal in newborn infants[52]
** Grade 5 or 6 according to Table 1.*

The agreement between TSA-adjusted review conclusions and guidelines was still good (*K* 0.41, range 0.36-0.45) (Table 4), but significantly lower (P = 0.001, Wilcoxon test) than the unadjusted agreement (*K* 0.56, range 0.52-0.59) (Table 2, page 14).[30]

Table 4 Agreement between TSA-adjusted Cochrane Neonatal Group systematic reviews and clinical practice guidelines in Denmark.

	Guidelines		
Grade*	**Recommended**	**Not recommended**	Total
I	***0***	***7***	**135**
II	***0***	***23***	
III	***3***	***49***	
IV	***9***	***44***	
V	***14***	***9***	**23**
VI	***12***	***3***	**15**
Total	**38**	**135**	**173**

Disagreement	**Partial agreement**	**Agreement**

***** Classification of the interventions assessed in Cochrane reviews (Table 1).

4. Evaluation of clinical practice guidelines with the AGREE instrument

Sample

The 186 clinical practice guidelines from the 14 neonatal departments constituted the sample. Each department had very similar reporting of their own clinical practice guidelines. Therefore, we decided to assess the overall quality of the guidelines for the 14 departments rather than the quality of every single guideline (Table 5).

The author carried out the evaluation of clinical practice guidelines with the AGREE-instrument. The result was not published as it is recommended that four assessors independently assess guidelines with the AGREE-instrument.[22]

Table 5. Assessment of Danish clinical practice guidelines for newborn infants with AGREE[22]

Departments:	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**
AGREE-Items:														
Scope and purpose														
1 Objective	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2 Clinical question	4	4	4	4	4	4	4	4	4	4	4	4	4	4
3 Target population	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Standardised domain score (%)	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*
Stakeholder involvement														
4 Development group	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5 Patients view	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6 Target user	4	4	4	4	4	4	4	4	4	4	4	4	4	4
7 Pre-testing	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Standardised domain score (%)	*50*	*50*	*50*	*50*	*50*	*50*	*50*	*50*	*50*	*50*	*50*	*50*	*50*	*50*
Rigour of development														
8 Evidence search	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9 Evidence selection	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10 Formulation methods	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11 Benefits and harms	2	3	3	1	2	3	3	2	2	3	2	3	2	2
12 References	2	2	2	1	1	3	1	1	1	3	1	1	1	1
13 Peer review	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14 Updating	3	4	1	1	2	4	1	1	1	4	3	2	1	1
Standardised domain score (%)	*14*	*29*	*14*	*0*	*10*	*33*	*10*	*5*	*5*	*33*	*14*	*14*	*5*	*5*
Clarity and presentation														
15 Clarity	4	4	4	4	4	4	4	4	4	4	4	4	4	4
16 Alternatives	3	1	2	2	2	3	2	2	4	2	2	3	2	2
17 Key recommendations	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Standardised domain score (%)	*89*	*67*	*78*	*78*	*78*	*89*	*78*	*78*	*100*	*78*	*78*	*89*	*78*	*78*
Applicability														
18 Tools	4	4	4	4	4	4	4	4	4	4	4	4	4	4
19 Barriers	1	1	1	1	1	1	1	1	1	1	1	1	1	1

20 Costs	2	2	2	2	2	2	2	2	2	2	2	2	2	2
21 Review criteria	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Standardised domain score (%)	*58*	*58*	*58*	*58*	*58*	*58*	*58*	*58*	*58*	*58*	*58*	*58*	*58*	*58*
Editorial independence														
22 Editorial independence	1	1	1	1	1	1	1	1	1	1	1	1	1	1
23 Conflicts of interest	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Standardised domain score (%)	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
Rating: 4; strongly agree, 3; agree, 2; disagree, 1: strongly disagree *Standardised domain score (%): obtained score - minimum score / maximum score - minimum score*														

Five of six standardised domains score were identical or almost identical when comparing the departments, but the 'rigor of development' domain showed variability between the departments. Overall, most guidelines had an adequate description about scope and purpose, stakeholder involvement, and applicability and they were clearly presented. Conversely, most guidelines lacked adequate rigor of development, ie, systematic methods to search for evidence, clear criteria for selecting the evidence, systematic procedure for formulating and updating guidelines, and references list of the supporting evidence. Furthermore, editorial independence and conflicts of interest were never recorded.

DISCUSSION

Summary of the results

There is good agreement between Cochrane Neonatal Group reviews and Danish neonatal guidelines despite the fact that guideline authors have rarely used systematic methods or Cochrane reviews for the guideline development. Many, apparently statistically significant neonatal meta-analyses may become inconclusive if adjusted for random risk error with TSA. Such adjustments revealed that about every second 'positive' Cochrane neonatal review could lack sufficient meta-analytic evidence to support the intervention. TSA-adjustment of meta-analyses significantly decreased the agreement between adjusted review recommendations and unadjusted Danish clinical practice guideline recommendations. Most guidelines lacked systematic methods to search for evidence, clear criteria for selecting the evidence, and references to the supporting evidence.

Strengths

Strengths of this thesis are that we:

- included a complete sample of Cochrane reviews in neonatology, including a large number of interventions and meta-analyses;
- included all neonatal units in Denmark and successfully obtaining data from all but three;
- observed good agreement between the two physicians who independently assessed reviews;
- observed good agreement between Cochrane reviews in neonatology and Danish clinical neonatology guidelines;
- managed to analyse a large number of the Cochrane meta-analyses with the newly developed TSA technique, taking random errors due to repetitive testing into consideration;
- could show that many of the meta-analyses had insufficient information and potentially reached false recommendations based on spurious significant findings;
- consequently found a significant reduction in agreement between adjusted review recommendations and unadjusted guideline recommendations.

Real and potential limitations

The included studies, as well as the present thesis, also contain a number of real or potential limitations, which should be considered when evaluating our results.

Cochrane reviews and meta-analyses

We chose to use only Cochrane systematic reviews to assess the agreement between research and clinical practice guidelines. The reasons were that Cochrane reviews often include meta-analyses of RCTs that are placed at the top of the evidence hierarchy and seem to have better quality than other reviews.[10,53] There may be other reviews that address the same interventions or other interventions, however, the journals in which they are published usually prefer 'hot' topics to satisfy editorial and/or reader interest. Accordingly, Cochrane reviews are only a proportion of the reviews offered to clinicians and other reviews may support or contradict them. We did not assess this aspect. Furthermore, Cochrane reviews take time to prepare and may lack important recent research when published. Thus, evidence from Cochrane reviews should also be accompanied by potential evidence published after the last search date of the review.

Our assessment of agreement should be viewed in the context of Cochrane reviews' limitations related to pooling of results of heterogeneous RCTs, including random error, systematic error (bias), inadequate update, rarely incorporating evidence from non-RCTs,[54] and the reviews of retrospective and observational nature.[55] We observed that only a few Cochrane Neonatal Group reviews considered these issues and that many Cochrane reviews (36%) were not adequately updated. In fact, none of the included reviews considered random error risk as a result of repeated significant testing on the same outcome (updating) or due to significant testing of multiple outcomes. Very few reviews considered heterogeneity and none applied the random-effects model meta-analyses to incorporate heterogeneity in the assessment of estimates of interventions. Heterogeneity was substantial in some of the meta-analyses and we fear that this alone may have caused the review authors to reach the wrong conclusions. However, we did not assess this aspect. Furthermore, only a few of the reviews considered bias. This is in direct contrast to the recommendations of

the Cochrane Handbook.[11] We did not assess this aspect directly, but we indirectly touched upon this bias mechanism through our use of bias and heterogeneity adjusted TSA.[23]

On one hand, authors of reviews are experts on their topic's literature and from a practical point of view it seems reasonable that they allow themselves to give recommendations, as few physicians manage to read a full review. Thus, sometimes statements like 'X is (not) recommended' are seen in the conclusion. On the other hand, it is recommended that Cochrane reviews should provide the evidence rather than direct recommendations, as it is impossible for authors to put the evidence into context of any setting (external validity). Clinicians must then critically assess the internal validity of the review and subsequently decide how applicable the evidence is to their setting. The latter decision depends on the type of patients, costs, resources, and the local assessment of the weights of benefits and harms. Furthermore, other studies on the same intervention for the same condition should also be considered.

By solely classifying the recommendations of interventions on Cochrane reviews we may have by-passed some of the above important considerations. A more systematic and explicit approach to making judgements about the quality of evidence and strengths of recommendations has been suggested by, eg, the GRADE Working Group.[56,57] However, it was impossible to consider all these aspects for all the included interventions.

Assessment of Cochrane reviews and guidelines

The physicians who classified the treatment recommendations in reviews were not neonatologists.[22] Thus, they did not have biased expectations about the benefits and harms of the interventions. The inter-rater agreement was acceptable (***Kappa*** 0.76), but the scale used was not validated and had only been used to classify treatment recommendations in single trials. During the assessment of agreement between reviews and guidelines we received feedback from the contact persons of the departments, who re-confirmed our interpretations of the guidelines. Hence, misinterpretations of local guidelines seem minimal.

Our direct one-to-one comparison of Cochrane reviews and guidelines is debatable because of the limitations mentioned above. Furthermore, both reviews' and guidelines' recommendations are rarely a clear-cut 'yes or no' but associated with modifications. We decided to try to simplify something very heterogeneous and complex in order to provide an overview on many interventions rather than assessing a few interventions more thoroughly. Our categorisation of interventions in Cochrane reviews as 'recommended' or 'not recommended' is solely based on the Cochrane author's interpretation of the data, which is very subjective and not standardised. Some conclusions or recommendations may be based on outcomes being statistically significant ($P < 0.05$) rather than clinically relevant. This could be due to use of unvalidated surrogate outcomes[58] or that the clinical benefit was too small considering the costs. Furthermore, despite agreement (or disagreement) between reviews and guidelines they may both be wrong. We did not assess these aspects.

We decided to categorise absence of a guideline recommendation as 'not recommended' because guidelines rarely address interventions not used. However, absence of a guideline may be due to other reasons, eg, the intervention is less commonly used, time constraints, uncertainty about what to do, that other departments has sufficient guidelines, unawareness, etc. How this influences our overall conclusion is, of course, difficult to estimate.

Trial sequential analysis

Meta-analyses in Cochrane reviews should be updated when valid new evidence emerges or at least every second year.[11] Likewise, trial authors are also encouraged to conduct meta-analyses before and after the conduct of each new trial.[59,60] Hence, repetitive testing is very likely to occur on accumulating data in meta-analysis.[15-18] Therefore, we have suggested TSA as a method to reduce the risk of random errors due to repetitive testing in cumulative meta-analysis. [18]

From a theoretical point of view we found TSA an appealing method to use to apply more stringent standards to avoid false positive results due to random error. When planning a randomised trial the alpha, beta, delta, the variance, and the sample size (n) are closely linked. Likewise, it seems necessary to consider the N (information size) in meta-analyses. If not, the obtained *P* value seems 'homeless' and would reach

statistical significance in the process of adding trials even if there is no treatment effect. A *P* value should be evaluated in the light of which effect size we want to detect or reject. This contradicts the previous standard that meta-analyses supersede the need for predefined selection of intervention effect. Recently, much focus has been paid to trials prematurely terminated for benefit and ignoring predefined sample size.[61] Likewise, we find similar considerations on meta-analysis that were potentially stopped too early highly relevant. Meta-analytic results may themselves become inflated by including trials that were stopped early.[62] We did not investigate this aspect.

TSA does not aim to control for analysis of multiple outcomes. Several methods have been proposed to deal with this problem when comparing several groups in single trial or when having multiple outcomes (Bonferroni, Newman-Keuls, Duncan, etc).[63] These methods could be applied to meta-analyses, but to our knowledge there is currently no golden standard. As cumulative meta-analyses revises information in light of new information on the same outcome (ie, the repeated testing are only partially independent) the TSA needs to penalise Z-scores with a less restrictive factor compared to, eg, Bonferroni-adjustment, which seem too conservative especially if many (more than 5) comparisons are performed. Thus, the risk associated with multiple comparisons is highly important (eg, some neonatal meta-analyses analysed more than 20 outcomes). We can only endorse that this topic is further examined within meta-analyses. However, the need for other adjustments due to problems of multiplicity merely underlines the need to be careful when evaluating meta-analysis when data accumulate.

We used a frequentistic approach. From this frequentistic view repeated testing needs adjustment due to the risk of type 1 error. We are aware that cumulative meta-analysis is naturally amenable to Bayesian methods and within this paradigm it may not be relevant to adjust for multiple looks. However, this reflects a long-standing debate between the two statistical paradigms. We acknowledge that there are other ways to adjust for multiple looks in a single trial, eg, Armitage, Pocock, or Haybittle-Peto methods (all with fixed group interval making them unusable in cumulative meta-analyses), and the O'Brien-Flemming method.[64] The latter allows for a variable

number of looks at the accumulated evidence, exactly a situation we encounter in a meta-analysis. This is why we have implemented this method in TSA.[18]

In meta-analyses the concept to retain type 1 and 2 errors is rarely addressed.[65] Thus, for meta-analysis, we are only aware of a recently suggested method, 'law of iterated logarithm (LIL)' that adjusts for multiple testing.[66,67] LIL is a simulation method for evaluating binary and continuous outcomes in meta-analyses. Calculation of information size is not needed as LIL penalises the *Z*-value to account for multiple tests to maintain the type 1 error. The simulation method, however, represents only meta-analyses corresponding to the criteria chosen for the simulation scenario and this is also reflected in the effect sizes, which may be both larger and smaller depending on the number of trials included in the meta-analysis. The strength of not calculating the information size also represents a weakness. Eg, LIL is not able to advise future trialists of the number of participants they ought to include in their trial in order to obtain firm evidence.

One problem with TSA is whether the interim Z-scores are identically and independently distributed, which is required for a boundary construction assuming a 'true' Brownian motion. Intuitively, this does not seem violated in a single trial as the patients, intervention, setting, etc. are selected and treated according to the same protocol and the outcome of each consecutive patient is independent of previous outcomes. In a meta-analysis the sample sizes, populations, interventions, etc. may differ and the trials may be conducted over a long period. Thus, it may be argued that the settings are not identical among trials in a meta-analysis and the results of a new trial are not independent of the results from previous trials. Whether this issue is important is difficult to assess. In large multi-centre trials analogous problems may arise and they may be analysed undisputed with a group sequential design. One potential way to try to adjust to the assumptions of a true Brownian motion in meta-analyses could be to account for heterogeneity (TSA_{LBHIS}). It has recently been suggested that multicenter trials should also take heterogeneity into consideration when estimating the sample size.[28]

It may be argued that no repetitive testing exists if meta-analyses are carried out only once. Thus, adjustment may not be necessary. This may be correct, but no one can

predict how many times a meta-analysis will be updated. Furthermore, from a practical point of view many meta-analyses, especially on 'hot' topics, are updated several times, eg, 'inhaled nitric oxide for respiratory failure in infants' has been conducted at least five times.[68] Furthermore, meta-analyses ought to be conducted both before and after each new trial.[59,60] Therefore, statistically significant results due to repetitive analyses represent a real problem, already highlighted eleven years ago.[14-16]

We applied TSA retrospectively on meta-analyses to illustrate the concept of TSA and to examine the potential consequences of applying TSA to Cochrane reviews. However, in practice, TSA should be applied prospectively on meta-analyses. We based all TSA on an arbitrary RRR of 25% and our findings are appropriate only under the assumption of this effect size. It could be judged that a smaller or larger pre-specified intervention effect is more relevant for each individual meta-analysis. However, larger effect sizes are rare and a pre-specified assumption of a smaller effect size would only make the TSA even more conservative.

The control event rate may vary substantially among trials in meta-analyses and the calculation of information size depends strongly on the event-rate. Thus, even though the relative risk reduction is identical, the required information size would be large if the event-rate is low and vice versa. This illustrates one weakness associated with meta-analysis in general as well as with TSA. However, the magnitude of this problem depends on the meta-analyst's decision regarding inclusion criteria – is he or she 'lumping or splitting'. Thus, the more strict criteria for including trials, the more similar event-rate among included trials will be observed.

Obviously TSA provides more conservative conclusions compared to 'traditional' meta-analyses.[69] This delays clinicians' use of potentially beneficial interventions but it should be weighted against the risk of introducing harmful intervention based on inconclusive evidence.[70-72] Empirical evidence suggest that this delay is 2-3 years but with large variation (range 0-12).[69] If TSA becomes prospectively applied, strategic planning of several trials may reduce this delay.

Whether TSA is justified is debatable. The method has both pros and cons. Thus, to justify TSA, we need more comparisons with 'traditional' meta-analyses or other ways to perform meta-analyses in, eg, computer simulation or large empirical studies. Currently, one empirical study found that TSA seems to reduce the number of false positive results and overestimates of treatment effects. These beneficial effects occur with some delay in reaching the point where statistically significant evidence is considered conclusive.[69] We need more studies that in depth focus on TSA of single therapeutic areas.[18]

Assessment of guidelines (AGREE)

The AGREE[22] seems the only instrument validated for assessing quality of guidelines, but it has limitations.[21] First, it is based on theoretical assumptions from experts (ie, placed lowest in the evidence hierarchy) rather than on empirical evidence. Second, it can be used to compare guidelines, but does not classify guidelines as adequate or inadequate. Third, it does not assess the quality of evidence supporting the guidelines.[20] Finally, it should be applicable to any guidelines. However, applying the comprehensive AGREE on local guidelines may seem too ambitious as such guidelines should be easy-to-use in clinical practice rather than documenting in detail the development procedure, supporting evidence, editorial independence, etc. If the departments have had an overall systematic process for developing guidelines, the AGREE could have been applied on these, but only one of the included neonatal departments reported such 'guidelines' for guideline development, and it was not possible to get access to this guideline.

Key findings

We found good agreement between evidence in Cochrane Neonatal Group systematic reviews and Danish clinical guidelines for newborns. This contrasts with findings from other fields of medicine that revealed substantial gaps between research evidence and clinical practice.[73,74] The observed agreement may reflect a history of extensive use of evidence-based practice within neonatal care.[6,75] Surprisingly, we observed that just one department reported use of a guideline for guideline development and that authors rarely used Cochrane reviews directly but often used other sources of evidence or evidence coming from lower levels of the 'evidence hierarchy'.[10]

The reasons why clinicians do not consider Cochrane reviews more directly during guideline development are probably diverse. It may be that they are not using systematic and explicit methods to develop guidelines. This is supported by our finding that only one department allegedly had a guideline for guideline development. It may be due to lack of time (and skills) to critically assess the sometimes overwhelming amount of information in Cochrane reviews. Reasons for clinicians not adhering to guidelines include unawareness or lack of familiarity with the intervention, disagreement, lack of self-efficacy (the belief that one masters an intervention), inertia of previous practice, time constraints, as well as other external barriers.[76] Similar barriers may exist for implementation of Cochrane reviews into guidelines.

Most observed disagreements were due to guidelines recommending interventions that lacked evidence in reviews rather than failure to recommend interventions supported by reviews. This may illustrate the potential current societal opinion saying: preferable overuse of potentially beneficial treatment, despite that sometimes it is more right to withhold a treatment.[77]

We directly compared Cochrane reviews with guidelines. Besides research evidence, guidelines should also consider local expertise, cost-benefits, resources, and local values. These factors may be reasons for disagreements. Eg, one reason for not using prophylactic surfactant for preterm infants was due to costs, despite surfactant being recommended in the review according to our classification.

Our TSA showed that the majority of neonatal meta-analyses apparently had insufficient information size to detect or reject a 25% RRR. Auditing such meta-analyses with monitoring boundaries revealed that every second 'positive' Cochrane Neonatal Group review seemed to lack meta-analytic evidence if adjusted for this random error risk.

$TSA_{25\%}$ resulted in statistical interpretation of the meta-analyses very similar to $TSA_{LBHIS.}$[26] $TSA_{15\%}$ provides more conservative results.[26,29] However, different TSAs on meta-analyses do not provide a more different result than if similar group

sequential designs are applied on a single trial, and we know from single trials that estimates of effect size are rarely accurate, but too ambitious.[78] We acknowledge that the decision of the a priori effect size may be a weak aspect in sequential designs in general. We endorse that more research is done on whether to choose a relevant effect size, a realistic effect size, or a minimum relevant effect size. For TSA, the use of effect size from low-bias trials (TSA_{LBHIS}) somehow bypasses this problem and seems most appealing, but it can only be used if enough low-bias trials with enough patients and events have been conducted. If not, TSA_{LBHIS} may be impossible to use or be very uncertain.

TSA is merely a tool for 'scanning' a meta-analysis in order to establish which meta-analysis needs further participants to obtain firm evidence and which has reached sufficient information. Approximately one third of the significant ($P < 0.05$) meta-analysis provided firm evidence in the light of any TSA examined. Thereby one could close the issue of whether further trials should be conducted. This decision must of course be taken on a 'meta-analysis by meta-analysis' basis, incorporating all knowledge on the specific therapeutic area. New trials could examine other important aspects. Furthermore, according to the Declaration of Helsinki all therapeutic areas may need a scientific reassessment now and then. Approximately one third of the significant ($P < 0.05$) meta-analysis did not provide firm evidence irrespective of the TSA applied, thereby crying for more trials to close the 'evidence gap'. These meta-analyses would probably not have met the standards for being 'conclusive' if they had been analysed as a single trial.

Interpreting meta-analyses with TSA decreased the agreement between reviews and Danish clinical practice guidelines. This indicates that potentially false positive results due to random error in some neonatal reviews may, directly or indirectly, have induced use of interventions without beneficial or even with harmful effects.

The interpretation of guideline quality is subjective. In the present evaluation the risk of subjectiveness is even more pronounced as only one appraiser gave AGREE-scores. In this context, the AGREE instrument revealed that the 'rigor of development' domain had noticeable variability among Danish neonatal departments. Especially, reporting of expected (or required) updates and list of references in

guidelines varied from clearly reported to not reported at all. The AGREE instrument showed that the guidelines' scores resembled regarding many quality components. This may indicate that the units are not working independently, but that guideline authors are copying from each other. The latter may be due to several reasons, eg, that the same few specialists in neonatology are moving around between departments or that it is considered reasonable to copy an apparently high quality guideline instead of preparing it once again. It seems reasonable in such instances to give the reference of the other department, but this was rarely reported.

In general, it is very subjective what the appropriate content of individual local guidelines should be. Overall, the assessment of guidelines revealed that there seemed to be room for improvement on many domains. For Danish neonatal practice guidelines, there is a need to: develop and report on systematic methods; to search and select the evidence; develop the systematic procedure for formulating and updating guidelines; and to use references list of the supporting evidence.

Other comments

It was not the aim of this thesis to judge whether guideline recommendations were right or wrong, as no gold standard exists to judge this, but only to reveal the specific disagreements between reviews and guidelines, and hopefully initiate discussions about the basis for current practice. Comprehensive methods for developing guidelines and grading the strengths of recommendations can subsequently be used.[56,57] These methods were published during the preparation of this thesis and thus not considered in the protocol development for the thesis.[56,57] The vast majority of disagreements were due to guideline recommendations of interventions that lacked evidence in reviews. If TSA was the standard for meta-analysis, even more interventions would lack review evidence. Many of the recommended interventions are promising or in a 'grey zone' in which some can 'see the proof', others can only 'sense it', and others may 'not be convinced at all'. During this 'grey zone', willingness for adopting the intervention will differ among clinicians (guideline authors) depending on what is considered sufficient cut-off for 'evidence-level' of research. However, two other aspects must also be considered; ethics and societal opinion.

Regarding ethics, two contrasting points of view exist here, deontology and utilitarianism.[78] Clinicians tend to do their duty, ie, act in their patient's best interest; this is a deontological approach. Thus, if an intervention is marginally promising based on relevant research (ie, an increased probability that the intervention is beneficial, but traditional statistics inform us that too much uncertainty still exists) clinicians often pursue this intervention. However, if all used 'only' promising interventions, the consequences would be that many interventions without benefits could be implemented. The latter ethical approach is utilitarian. Utilitarianism focuses on the consequences of the act rather than the act itself. Utilitarianism looks easier to adopt by researchers and health economists, but utilitarianism should certainly also be consider by clinicians, parents, and patients.

Regarding societal opinion, in a clinical setting with a severely ill newborn, the parents and health-care workers expect the clinician to do something. Thus, as errors of omission seem more reprehensible than errors of commission, most clinicians would find it appropriate to perform a well-intended intervention, irrespective of the research evidence.[77] Thus, they risk doing more harm than good just by doing something.

There are both reasonable and unreasonable causes for disagreement between guidelines and research evidence and it is hard to find definite answers on which interventions ought to or ought not to be implemented into clinical practice. Therefore, it seems important to try to define at what probability level research findings are sufficiently true to be implemented into clinical practice or, alternatively, what level of potentially false research findings we are willing to accept.[79] However, it is important that clinicians only combine reliable research with patients and community values in order to optimise treatment especially in the context of powerful pharmaceutical marketing practices.[80]

Conclusions

Authors of Cochrane reviews and clinical practice guidelines aim to facilitate implementation of research evidence into clinical practice. However, they should also aim to collaborate and further improve the quality of reviews and clinical practice guidelines. Although our studies only focus on Cochrane reviews it illustrates a

method to critically audit guidelines that may be applicable to all specialities. By considering all published reviews it is subsequently easy to cope with updated and new reviews appearing in The Cochrane Library. Hence, clinicians may avoid important disagreements between practice and research evidence in reviews. Likewise, clinicians' comments to authors of Cochrane reviews are needed to improve the clinical relevance and quality of reviews as they are continuously updated. This is possible by using the 'feedback' function linked to each review in The Cochrane Library. Using these initiatives might minimise the gaps between research evidence and clinical practice.

Implication for practice

- Danish neonatal departments should aim to construct adequate guidelines on how to develop clinical practice guidelines. If possible, most clinical practice guidelines should be developed on a national level rather than locally in order to optimise resources and make coherent national treatment of newborns.

- Authors of Cochrane Neonatal Group reviews should consider the risk of random error due to repetitive testing when updating their reviews.

Future research should focus on

- Initiatives that make clinicians develop high quality clinical practice guidelines considering their limited resources.

- Comparing TSA with the law of iterated logarithm or other methods to adjust for the risk of random error in meta-analyses. This could by done by simulation studies or by large prospective empirical studies using the final treatment effect (point estimate and *P* value) of each published meta-analysis as representing the 'truth'.

- Which effect size (the realistic, the minimum clinical relevant, the worthwhile, etc.) to use when estimating the required information size in TSA in meta-analyses.

- How to combine the adjustment for both random errors and systematic errors in meta-analyses.

REFERENCES

1. Neonatal intensive care. A history of excellence. A Symposium Commemorating Child Health Day 1992. http://www.neonatology.org/classics/nic.nih1985.pdf
2. Silverman WA. Personal reflections on lessons learned from randomized trials involving newborn infants, 1951 to 1967. James Lind Library. www.jameslindlibrary.org
3. Soll RF. Prophylactic natural surfactant extract for preventing morbidity and mortality in preterm infants. Cochrane Database Syst Rev. 1997, Issue 4. Art. No.: CD000511. DOI: 10.1002/14651858.CD000511.
4. Askie LM, Henderson-Smart DJ. Restricted versus liberal oxygen exposure for preventing morbidity and mortality in preterm or low birth weight infants. Cochrane Database Syst Rev. 2001, Issue 4. Art. No.: CD001077. DOI: 10.1002/14651858.CD001077.
5. Chalmers I, Hetherington J, Newdick M, et al. The Oxford Database of Perinatal Trials: developing a register of published reports of controlled trials. Control Clin Trials. 1986;7(4):306-24
6. Sinclair JC, Bracken MB. Effective Care of the Newborn Infant, Oxford University Press, 1992.
7. Langhoff-Roos J, Kesmodel U, Jacobsson B, Rasmussen S, Vogel I. Spontaneous preterm delivery in primiparous women at low risk in Denmark: population based study. BMJ. 2006;332(7547):937-9.
8. Molholm Hansen B, Greisen G. Preterm delivery and calculation of survival rate below 28 weeks of gestation. Acta Paediatr. 2003;92(11):1335-8.
9. The Cochrane Library, Issue 4, 2004. Chichester: Wiley.
10. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet. 2002;359(9300):57-61.
11. Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.5. http://www.cochrane.org/resources/handbook/hbook.htm.
12. Soll RF, Sinclair JC, Bracken MB, Horbar JD, Haughton DE, Editorial Team. Cochrane Neonatal Group. About The Cochrane Collaboration (Cochrane Review Groups (CRGs)) 2007, Issue 1. Art. No.: NEONATAL
13. Gluud LL. Bias in clinical intervention research. Am J Epidemiol. 2006;163(6):493-501.
14. Berkey CS, Mosteller F, Lau J, Antman EM. Uncertainty of the time of first significance in random effects cumulative meta-analysis. Control Clin Trials. 1996;17(5):357-71.
15. Pogue J, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. Control Clin Trials. 1997;18:580-93.
16. Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. Lancet. 1998;351(9095):47-52.
17. Devereaux PJ, Beattie WS, Choi PT, Badner NH, Guyatt GH, Villar JC, et al. How strong is the evidence for the use of perioperative beta-blockers in non-cardiac surgery? Systematic review and meta-analysis of randomised controlled trials. BMJ. 2005;331(7512):313-21.
18. Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evi-dence is reached in cumulative meta-analysis. J Clin Epidemiol. 2008; 61(1);64-75.
19. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika. 1983;70: 659-63.
20. Oxman AD, Schunemann HJ, Fretheim A. Improving the use of research evidence in guideline development: 16. Evaluation. Health Res Policy Syst. 2006;4:28.
21. Graham ID, Calder LA, Hebert PC, Carter AO, Tetroe JM. A comparison of clinical practice guideline appraisal instruments. Int J Technol Assess Health Care. 2000;16(4):1024-38.

22. The AGREE Collaboration. Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. Qual Saf Health Care. 2003;12(1):18-23.
23. Brok J, Greisen G, Jacobsen T, Gluud LL, Gluud C. Agreement between Cochrane Neonatal Group reviews and clinical guidelines for newborns at a Copenhagen University Hospital. Acta Paediatr. 2007;96(1):39-43.
24. Brok J, Greisen G, Madsen LP, Tilma K, Faerk J, Børch K, Garne E, Christesen HT, Stanchev H, Jacobsen T, Nielsen JP, Henriksen TB, Gluud C. Agreement between Cochrane neonatal reviews and clinical guidelines for newborns in Denmark. Arch Dis Child Fetal Neonatal Ed. 2008;93(3):225-9.
25. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol. 1987;126(2):161-9.
26. Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential boundaries improve interpretation of meta-analyses. J Clin Epidemiol. 2008;Apr 12; [Epub ahead of print].
27. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002; 21:1539-58.
28. Fedorov V, Jones B. The design of multicentre trials. Stat Methods Med Res. 2005;14(3):205-48.
29. Brok J, Wetterslev J, Thorlund K, Gluud C. Apparently conclusive meta-analyses are often inconclusive - a trial sequential analysis adjustment for random error risk in conclusive neonatal meta-analyses. Int J Epidemiol. 2007 (resubmitted).
30. Brok J, Greisen G, Madsen LP, Tilma K, Faerk J, Børch K, Garne E, Christesen HT, Stanchev H, Jacobsen T, Nilsen JP, Henriksen TB, Wetterslev J, Gluud C. Agreement between Cochrane neonatal reviews adjusted for random error risk and guidelines for newborns in Denmark. (submitted 2008).
31. Steer PA, Henderson-Smart DJ. Caffeine versus theophylline for apnea in preterm infants. Cochrane Database Syst Rev. 1998, Issue 2. Art. No.: CD000273. DOI: 10.1002/14651858.CD000273.
32. Henderson-Smart DJ, Wilkinson A, Raynes-Greenow CH. Mechanical ventilation for newborn infants with respiratory failure due to pulmonary disease. Cochrane Database Syst Rev. 2002, Issue 4. Art. No.: CD002770. DOI: 10.1002/14651858.CD002770.
33. Bell EF, Acarregui MJ. Restricted versus liberal water intake for preventing morbidity and mortality in preterm infants. Cochrane Database Syst Rev. 2001, Issue 3. Art. No.: CD000503. DOI: 10.1002/14651858.CD000503.
34. Barrington KJ. Umbilical artery catheters in the newborn: effects of catheter design (end vs side hole). Cochrane Database Syst Rev. 1999, Issue 1. Art. No.: CD000508. DOI: 10.1002/14651858.CD000508.
35. Henderson-Smart DJ, Subramaniam P, Davis PG. Continuous positive airway pressure versus theophylline for apnea in preterm infants Cochrane Database Syst Rev 2001, Issue 4. Art. No.: CD001072. DOI: 10.1002/14651858.CD001072.
36. Soll RF. Multiple versus single dose natural surfactant extract for severe neonatal respiratory distress syndrome. Cochrane Database Syst Rev. 1999, Issue 2. Art. No.: CD000141. DOI: 10.1002/14651858.CD000141.
37. Puckett RM, Offringa M. Prophylactic vitamin K for vitamin K deficiency bleeding in neonates. Cochrane Database Syst Rev. 2000, Issue 4. Art. No.: CD002776. DOI: 10.1002/14651858.CD002776.
38. Conde-Agudelo A, Diaz-Rossello JL, Belizan JM. Kangaroo mother care to reduce morbidity and mortality in low birthweight infants. Cochrane Database Syst Rev. 2003, Issue 2. Art. No.: CD002771. DOI: 10.1002/14651858.CD002771.
39. Darlow BA, Austin NC. Selenium supplementation to prevent short-term morbidity in preterm neonates. Cochrane Database Syst Rev. 2003, Issue 4. Art. No.: CD003312. DOI: 10.1002/14651858.CD003312.
40. Greenough A, Milner AD, Dimitriou G. Synchronized mechanical ventilation for respiratory support in newborn infants. Cochrane Database Syst Rev. 2004, Issue 3. Art. No.: CD000456. DOI: 10.1002/14651858.CD000456.pub2.

41. Davis PG, Henderson-Smart DJ. Intravenous dexamethasone for extubation of newborn infants. Cochrane Database Syst Rev. 2001, Issue 4. Art. No.: CD000308. DOI: 10.1002/14651858.CD000308.
42. Bhuta T, Henderson-Smart DJ. Elective high frequency jet ventilation versus conventional ventilation for respiratory distress syndrome in preterm infants. Cochrane Database Syst Rev. 1998, Issue 2. Art. No.: CD000328. DOI: 10.1002/14651858.CD000328.
43. Herrera C, Holberton J, Davis P. Prolonged versus short course of indomethacin for the treatment of patent ductus arteriosus in preterm infants. Cochrane Database Syst Rev. 2007, Issue 2 (updated). Art. No.: CD003480. DOI: 10.1002/14651858.CD003480.pub3.
44. Cooke L, Steer P, Woodgate P. Indomethacin for asymptomatic patent ductus arteriosus in preterm infants. Cochrane Database Syst Rev. 2003, Issue 1. Art. No.: CD003745. DOI: 10.1002/14651858.CD003745.
45. Soll RF, Dargaville P. Surfactant for meconium aspiration syndrome in full term infants. Cochrane Database Syst Rev. 2000, Issue 2. Art. No.: CD002054. DOI: 10.1002/14651858.CD002054.
46. Henderson-Smart DJ, Steer P. Prophylactic caffeine to prevent postoperative apnea following general anesthesia in preterm infants. Cochrane Database Syst Rev. 2001, Issue 4. Art. No.: CD000048. DOI: 10.1002/14651858.CD000048.
47. Brion LP, Primhak RA, Ambrosio-Perez I. Diuretics acting on the distal renal tubule for preterm infants with (or developing) chronic lung disease. Cochrane Database Syst Rev. 2002, Issue 1. Art. No.: CD001817. DOI: 10.1002/14651858.CD001817.
48. Haque K, Mohan P. Pentoxifylline for neonatal sepsis. Cochrane Database Syst Rev. 2003, Issue 2. Art. No.: CD004205. DOI: 10.1002/14651858.CD004205.
49. McGuire W, Clerihew L, Austin N. Prophylactic intravenous antifungal agents to prevent mortality and morbidity in very low birth weight infants. Cochrane Database Syst Rev. 2004, Issue 1. Art. No.: CD003850. DOI: 10.1002/14651858.CD003850.pub2.
50. Okoromah CAN, Lesi FEA. Diazepam for treating tetanus. Cochrane Database Syst Rev. 2004, Issue 1. Art. No.: CD003954. DOI: 10.1002/14651858.CD003954.pub2.
51. Osborn DA, Jeffery HE, Cole M. Opiate treatment for opiate withdrawal in newborn infants. Cochrane Database Syst Rev. 2005, Issue 3 (updated). Art. No.: CD002059. DOI: 10.1002/14651858.CD002059.pub2.
52. Osborn DA, Jeffery HE, Cole MJ. Sedatives for opiate withdrawal in newborn infants. Cochrane Database Syst Rev. 2005, Issue 3 (updated). Art. No.: CD002053. DOI: 10.1002/14651858.CD002053.pub2.
53. Moja LP, Telaro E, D'Amico R, Moshetti I, Coe L, Liberati A. Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. BMJ. 2005;330(7499):1053.
54. Brok J, Buckley N, Gluud C. Interventions for paracetamol (acetaminophen) overdose. Cochrane Database of Systematic Reviews 2006, Issue 2. Art. No.: CD003328. DOI: 10.1002/14651858.CD003328.pub2.
55. Egger M, Davey Smith G, Altman DG. Systematic reviews in healthcare. Meta-analysis in context. 2nd ed. London: BMJ Publishing Group, 2001.
56. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations BMJ. 2004;328(7454):1490
57. Atkins D, Briss PA, Eccles M, Flottorp S, Guyatt GH, Harbour RT, et al. Systems for grading the quality of evidence and the strength of recommendations II: pilot study of a new system: BMC Health Serv Res. 2005;5(1):25.
58. Gluud C, Brok J, Gong Y, Koretz RL. Hepatology may have problems with putative surrogate outcome measures. J Hepatol. 2007;46(4):734-42.
59. WHO. Research for health – A Position paper on WHO's Role and Responsibilities in Health Research. 2006. http://www.who.int/rpc/meetings/position_paper.pdf.
60. Smith R. Medical journals are an extension of the marketing arm of pharmaceutical companies. PLoS Med. 2005;2(5):e138.

61. Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. Randomised trials stopped early for benefit: a systematic review. JAMA 2005;294:2203-9.
62. Bassler D, Ferreira-Gonzalez I, Briel M, Cook DJ, Devereaux PJ, Heels-Ansdell D, et al. Systematic reviewers neglect bias that results from trials stopped early for benefit. J Clin Epidemiol. 2007;60(9):869-73.
63. Altmans DG. London, Chapman and Hall. 2. 2001.Practical Statistics for Medical Research
64. Friedman, LM.;Furberg, CD.; DeMets, DL. *St Louis, Mosby.* 3. 1996. Fundamentals of Clinical Trials.
65. Sutton AJ, Higgins JP. Recent developments in meta-analysis. Stat Med. 2007 Jun 25; [Epub ahead of print].
66. Lan KK, Hu M, Cappelleri JCC. Applying the law of iterated logarithm to cumulative meta-analysis of continues endpoint. Statistica Sinica 2003;13:1135-45.
67. Hu M, Cappelleri JCC, 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):329-340.
68. Barrington KJ, Finer NN. Inhaled nitric oxide for respiratory failure in preterm infants. *Cochrane Database Syst Rev.* 2007, Issue 3. Art. No.: CD000509. DOI: 10.1002/14651858.CD000509.pub3.
69. Thorlund K, Devereaux PJ, Wetterslev JW, Guyatt G, Ioannidis JP, Thebana L, et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? (submitted 2008).
70. McGettigan P, Henry D. Cardiovascular risk and inhibition of cyclooxygenase: a systematic review of the observational studies of selective and nonselective inhibitors of cyclooxygenase 2. JAMA. 2006;296(13):1633-44.
71. Jespersen CM, Als-Nielsen B, Damgaard M, Hansen JF, Hansen S, Helo OH, et al. Randomised placebo controlled multicentre trial to assess short term clarithromycin for patients with stable coronary heart disease: CLARICOR trial. BMJ. 2006;332(7532):22-7.
72. Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG, Gluud C. Mortality in randomized trials on antioxidants supplements for primary and secondary prevention. JAMA. 2007;297:842-857
73. Moyer VA, Gist AK, Elliot EJ. Is the practice of paediatric inpatient medicine evidence-based? J Paediatr Child Health. 2002;38(4):347-51.
74. Kürstein P, Gluud LL, Willemann M, Olsen KR, Kjellberg J, Sogaard J, et al. Agreement between reported use of interventions for liver diseases and research evidence in Cochrane systematic reviews. J Hepatol. 2005;43(6):984-9.
75. Sinclair JC, Haughton DF, Bracken MB, Horbar JD, Sol RF. Cochrane Neonatal systematic reviews: a survey of the evidence for neonatal therapies. Clin Perinatol. 2003;30(2):285-304.
76. Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PA, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA.1999;282(15):1458-65.
77. Doust J, Del Mar C. Why do doctors use treatments that do not work? BMJ. 2004;328(7438):474-5.
78. Kumar A, Soares H, Djulbegovic B. High proportion of high quality randomized clinical trials conducted by the NCI are negative or inconclusive.[abstract] 13th International Cochrane Colloquium, Melbourne. 2005.
79. Cadore B, Boitte P, Demijnck G, Greiner D, Jacquemin D. Solidarity in perinatal medicine. Ethical Theory Moral Pract. 2000;3(4):435-54.
80. Djulbegovic B, Hozo I.When should potentially false research findings be considered acceptable? PLoS Med. 2007 ;4(2):e26.
81. Als-Nielsen B, Chen W, Gluud C, Kjaergard LL. Association of funding and conclusions in randomized drug trials: a reflection of treatment effect or adverse events? JAMA. 2003;290(7):921-8.

DANISH SUMMARY

Baggrund Det er velkendt at der findes uoverensstemmelser mellem hvad forskningen viser og hvad kliniske instrukser vejleder sundhedspersonale at gøre. Cochrane litteraturoversigter forsøger at gøre det nemmere at overskue forskningsresultaterne så disse lettere kan implementeres i kliniske instrukser. Oversigterne kan dog rapportere vildledende konklusioner bl.a. pga. tilfældige fund i deres meta-analyser. Forsøgs sekventielle analyser (FSA) er en statistisk metode som begrænser risikoen for tilfældige fund i meta-anlyserne på grund af gentagne analyser når yderligere forsøgsresultater kommer frem. AGREE-instrumentet kan vurdere kvaliteten af kliniske instrukser.

Formål At undersøge (1) overensstemmelsen mellem Cochrane neonatale litteraturoversigter og danske hospitalers kliniske instrukser for nyfødte, (2) meta-analyserne i Cochrane oversigterne med FSA, (3) om FSA ændrede konklusionen i Cochrane oversigter der fandt at en bestemt behandling var gavnlig og om disse nye konklusioner influerede på overensstemmelsen mellem oversigterne og instrukserne, og (4) kvaliteten af danske kliniske instrukser for nyfødte med AGREE-instrumentet.

Resultater Der er god overensstemmelse mellem Cochrane litteraturoversigter og danske kliniske instrukser indenfor neonatologien på trods af at instruksforfatterne sjældent direkte brugte oversigterne. Mange meta-analyser med 'positive' fund bliver inkonklusive når de analyseres med FSA. Dette medførte at ca. halvdelen at alle oversigter som konkluderede at en behandling var gavnlig risikerer at være på baggrund af et tilfældigt fund i meta-analyserne. Med de justerede konklusioner i henhold til FSA-analyserne, så blev overensstemmelsen mellem oversigter signifikant mindre god. De fleste afdelinger bruger ikke systematiske metoder ved udarbejdelsen af kliniske instrukser. De mangler klare kriterier for identifikation og udvælgelse af forskning, der bruges som grundlag for udarbejdelse af kliniske instrukser.

Konklusioner Forfattere af Cochrane oversigter og kliniske instrukser prøver at implementeret forskningsresultater i klinisk praksis. Både kvaliteten af Cochrane neonatal oversigterne samt danske kliniske.