The GRADE for Evidence-Based Decision Making: from Concept to Application in the Field of Pediatric Pharmacotherapy

by

Alla Osadchy

A thesis submitted in conformity with the requirements for the degree of Master of Science in Medical Science
Graduate Department of Institute of Medical Science
University of Toronto

© Copyright by Alla Osadchy 2013
The GRADE for Evidence-Based Decision Making: from Concept to Application in the Field of Pediatric Pharmacotherapy

Alla Osadchy

Master of Science-Medical Science
Institute of Medical Science
University of Toronto
2013

ABSTRACT

Rationale: Methodological quality of systematic reviews (SRs) remains an area of concern with no consensus on the optimal appraisal instrument to assess quality of published research.

Hypothesis: The Grading of Recommendations Assessment, Development and Evaluation (GRADE) is a feasible and useful methodology to assess quality of evidence in pediatric pharmacotherapy.

Aim: To demonstrate the applicability of the GRADE to selected topics.
Methods: To perform two SRs to illustrate the practical use of the GRADE, highlight methodological challenges encountered and compare the GRADE to the alternative appraisal tool.

Results: the GRADE was implemented to conduct two SRs. Judgments on imprecision were challenging. Comparison of the GRADE to alternative tool demonstrated that GRADE may generate different conclusions on overall quality assessment.

Conclusion: The GRADE is a valuable innovative tool for assessing quality of evidence, applicable to the field of pediatric pharmacotherapy, with a potential to impact inferences drawn after applying alternative instruments.
Acknowledgments

This work would not have been completed without the guidance and support of several great people.

I am very grateful to my mentor and supervisor Dr Gideon Koren for his valuable advice, outstanding ongoing support during various stages of my professional life including the execution of this project and confidence in my ability to reach my goals.

I would like to thank my committee members Dr Deborah O'Connor, Dr Anna Taddio, and Dr Bhushan M. Kapur for taking their time and being available to guide me throughout all stages of my project.

I am grateful to Dr Irena Nulman for her encouragement to embark on this project and exceptional support during my fellowship and afterwards.

I thank sincerely Ms Myla E. Moretti, M.Sc. for her technical assistance with statistical analysis and her valuable comments while preparing the manuscript "Effect of Domperidone on Insufficient Lactation in Puerperal Women: A Systematic Review and Meta-Analysis of Randomized Controlled Trials``.

I thank Mr Thirukumaran Ratnapalan, a summer research student, for his contribution to literature search and data extraction for the manuscript ```Ocular toxicity in children exposed in
utero to antimalarial drugs: review of the literature. This project has been supported by an unrestricted research grant from Shoppers’ Drug Mart.
Contributions

The content of the chapter 3 has been published as following:


Dr Alla Osadchy is the principal author of this manuscript. She designed the study, performed literature search, data extraction and quality assessment, analyzed and interpreted the data, and drafted the manuscript. Ms. Myla E. Moretti provided technical training on conducting meta-analysis and read and approved the final manuscript. Dr Gideon Koren checked the extracted data and quality assessment, oversaw the project and coordinated submission of the manuscript. He read and approved the final manuscript. The manuscript was peer-reviewed.

The content of the chapter 4 has been published as following:


Dr Alla Osadchy is the principal author of this manuscript. She designed the study, performed literature search, data extraction and quality assessment, analyzed and interpreted the data, and drafted the manuscript. Mr. Thirukumaran Ratnapalan conducted literature search and data extraction independently and read and approved the final manuscript. Dr Gideon Koren oversaw
the project and coordinated submission of the manuscript. He read and approved the final manuscript. The manuscript was peer-reviewed.
# Table of Contents

The GRADE for Evidence-Based Decision Making: from Concept to Application in the Field of Pediatric Pharmacotherapy ......................................................... ii

Acknowledgments ........................................................................................................ iv

Table of Contents .......................................................................................................... viii

List of Tables .................................................................................................................. X

List of Figures ................................................................................................................ XI

List of Abbreviations ................................................................................................... XII

Chapter 1 Literature Review ......................................................................................... 1

1.1 Background: systematic reviews and evidence-based medicine ....................... 1

1.2 Brief overview of existing quality assessment tools ........................................... 14

1.2.1 Appraisal instruments available – general approach ....................................... 14

1.2.2 Appraisal instruments available – volume ..................................................... 15

1.2.3 Appraisal instruments available – content .................................................. 16

1.2.4 Appraisal instruments available – format ..................................................... 17

1.2.5 Appraisal instruments available – how to choose one? ................................. 19

1.3 The GRADE introduction .................................................................................. 21

Chapter 2 Hypothesis / Aim ....................................................................................... 30

Chapter 3 Effect of domperidone on insufficient lactation in puerperal women: a systematic review and meta-analysis of randomized controlled trials .................................................. 33

1. Introduction ........................................................................................................... 34

2. Methods ............................................................................................................... 36
List of Tables

Table 1- Chapter 3 Characteristics of trials included in analysis ........................................ 44

Table 2 – Chapter 3 Methodological quality of RCTs included in the meta-analysis............. 45

Table 3- Chapter 4 Summary of selected studies ................................................................. 59-61

Table 4- Chapter 5 GRADE evidence profile: domperidone for insufficient lactation in puerperal women .................................................................................................................. 74

Table 5- Chapter 5 GRADE evidence profile: ocular toxicity in children exposed in utero to antimalarial drugs ................................................................. 77

Table 6- Chapter 5 Effect of domperidone on insufficient lactation in puerperal women- assessment of the risk of bias within and across studies utilizing the CCRBT……………… 81

Table 7- Chapter 5 Ocular toxicity in children exposed in utero to antimalarial drugs-risk of bias assessment utilizing CCRBT................................................................. 83

Table 8- Chapter 5 Ocular toxicity in children exposed in utero to antimalarial drugs - assessment of risk of bias within and across studies utilizing CCRBT................................. 84
List of Figures

Figure 1- Chapter 1 Methodology for a systematic review of randomized controlled trials........6
Figure 2- Chapter 1 An updated model for evidence-based clinical decisions ....................... .8
Figure 3 - Chapter 1 Schematic view of GRADE’s process for developing recommendations. .25
Figure 4- Chapter 1 Quality assessment criteria .................................................................27
Figure 5-Chapter 3 Flow chart of selected studies ............................................................... 40
Figure 6- Chapter 3 Percent change in milk volume with domperidone treatment............... 46
Figure 7- Chapter 4 Flow chart of selected articles.............................................................. 58
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADR</td>
<td>Adverse Drug Reaction</td>
</tr>
<tr>
<td>CPG</td>
<td>Clinical Practice Guideline</td>
</tr>
<tr>
<td>CQ</td>
<td>Chloroquine</td>
</tr>
<tr>
<td>EBM</td>
<td>Evidence-Based Medicine</td>
</tr>
<tr>
<td>EP</td>
<td>Evidence Profile</td>
</tr>
<tr>
<td>ERG</td>
<td>Electroretinogram</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
</tr>
<tr>
<td>GRADE</td>
<td>Grading of Recommendations, Assessment, Development, and Evaluation</td>
</tr>
<tr>
<td>HCQ</td>
<td>Hydroxychloroquine</td>
</tr>
<tr>
<td>ISCEV</td>
<td>International Society for Clinical Electrophysiology of Vision</td>
</tr>
<tr>
<td>OIS</td>
<td>Optimal Information Size</td>
</tr>
<tr>
<td>RCT</td>
<td>Randomized Controlled Trial</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>----------</td>
<td>--------------------</td>
</tr>
<tr>
<td>SLE</td>
<td>Systemic Lupus Erythematosus</td>
</tr>
<tr>
<td>SR</td>
<td>Systematic Review</td>
</tr>
<tr>
<td>VEP</td>
<td>Visual Evoked Potential</td>
</tr>
</tbody>
</table>
Chapter 1

Literature Review

1.1 Background: systematic reviews and evidence-based medicine

Systematic reviews (SRs) and meta-analyses have become an increasingly common type of research aimed to summarize scientific evidence on a particular topic. Given the exploding volume of published medical literature and scarcity of clinicians’ time, review articles play a pivotal role in combining results from individual studies to keep clinicians up to date and assist in individual patient’s decision-making process. Furthermore, SRs are essential for policy makers and developers of clinical practice guidelines.¹

The number of published SRs and meta-analyses has been growing steadily since 1990’s.² As reported by Moher et al. in 2007,³ approximately 2,500 new SRs were published annually around that time. More recently, Bastian et al.⁴ suggested a rate of 75 trials and 11 SRs becoming available daily with a continuing trend to grow.

The main advantage of SRs lies in the attempt to minimize bias by applying explicit methodological principles to searching, critical appraisal, summarizing and reporting scientific literature. This rigorous approach, hopefully, allows arriving at accurate and reliable conclusions. For these reasons, SRs of randomized controlled trials (RCTs) are often praised as the highest quality evidence while evaluating effectiveness of treatment interventions.⁵,⁶
Other potential advantages of SRs have been summarized by Greenhalgh as below. 

Advantages of systematic reviews

- Explicit methods limit bias in identifying and rejecting studies
- Conclusions are more reliable and accurate because of methods used
- Large amounts of information can be assimilated quickly by healthcare providers, researchers, and policymakers
- Delay between research discoveries and implementation of effective diagnostic and therapeutic strategies may be reduced
- Results of different studies can be formally compared to establish generalisability of findings and consistency (lack of heterogeneity) of results
- Reasons for heterogeneity (inconsistency in results across studies) can be identified and new hypotheses generated about particular subgroups
- Quantitative systematic reviews (meta-analyses) increase the precision of the overall result

The methodology for performing SRs and meta-analyses has been outlined elsewhere. Among the most quoted guidelines for conducting SRs are those proposed by the Cochrane Collaboration and the Agency for Healthcare Research and Quality (AHRQ). Standards for reporting SRs have been also proposed, such as the PRISMA statement for RCTs, and MOOSE for reporting meta-analyses of observational studies.
A SR usually starts with formulating a specific question which needs to be answered. The PICO format (patient, intervention, comparator, and outcome) is often suggested to frame the research question and guide subsequent literature search. Developing a review protocol is a vital step in conducting SRs and it must include overall objectives of the review, explicit search strategies to implement, pre-defined inclusion/exclusion criteria for selecting primary research studies and studies design(s)/type(s) to consider for inclusion. Furthermore, it is important to define a priori methodological criteria for rigorous quality assessment to be used to appraise each individual study.

The next step involves extraction of the relevant descriptive data from selected publications. Ideally, study selection and data extraction should be done by at least two researchers, independently. It is important to report transparently all included and all excluded studies, with reasoning behind judgments made. Each study, then, should be evaluated in terms of its methodological quality, precision and external validity. Synthesis of the collected data and its interpretation is a next step. At this point, a narrative summary of the data analysis might be presented and this would constitute a SR of the literature on the selected topic. Following SR, statistical methods may apply or not to quantitatively summarize the body of the evidence.

Meta-analysis is a statistical methodology to combine results of individual, similar studies. By pooling results of multiple studies in a statistically valid fashion, meta-analysis has the potential to answer questions when individual RCTs failed to do so, by increasing power and improving precision of the overall estimates of treatment effect.\(^\text{10}\)

In brief, meta-analysis starts with calculation of the treatment effect for the outcome of interest for each individual study being analyzed. Confidence intervals should be reported along with estimates of treatment effects. The most commonly used effect measures (or summary statistics)
are the risk ratio (RR) and the odds ratio (OR) for binary outcomes, and the mean difference and
the standardized mean difference for continuous outcomes. The next step of meta-analysis
involves calculation of overall (pooled) treatment effect across studies as a weighted average of
the intervention effects estimated in the individual studies.

There have been two main methods for performing meta-analysis: a fixed-effect and random-
effects models. The fixed-effect model works under assumption that the effect of treatment is
the same in all studies with the chance being the only explanation for variability among studies.
The summary treatment effect, hence, is the “true” or ‘fixed” one. The random effects model
works under different premises that “true” effects of treatment may differ between studies and
that these various ‘true” effects are normally distributed. From the practical standpoint, the
difference (and choice) of the appropriate model become important when results of individual
studies are not homogeneous and random effects model is a preferable one in such cases. The
results of meta-analysis are usually presented in a graphical form as a forest plot. Each outcome
assessed in a meta-analysis requires an individual forest plot.

Prior to performing a meta-analysis, though, it is important to decide whether studies selected
for SR are sufficiently similar to be combined. While differences among individual studies are
inevitable, it is worthwhile to explore heterogeneity across studies and decide whether meta-
analysis is an appropriate technique to apply. This could be broken further to three questions:
are individual studies clinically similar? Are they methodologically alike? Are the results of
individual reports statistically similar or homogeneous? Respectively, clinical heterogeneity
refers to variability in participants enrolled in different studies, interventions applied,
comparisons used and outcomes studied in various trials. Methodological heterogeneity relates
to variability in study designs and conduct and thus, risk of bias across studies. It is a matter of
clinical judgment and appraisal expertise of the reviewer to decide to which degree studies are sufficiently similar, clinically and methodologically, to pool them in meta-analysis. A statistical test of heterogeneity is often presented as a component of meta-analysis and is used to assess variability between individual studies’ estimates. Of note, significant statistical heterogeneity among studies may not preclude performing meta-analysis. Rather, the effort should be made to explore causes of such heterogeneity. There have been some strategies proposed to deal with heterogeneity. For instance, researches may decide not to perform meta-analysis (when significant heterogeneity is present), ignore heterogeneity and apply fixed-effect model for pooling data, use random-effects model in the attempt to incorporate heterogeneity or explore it by conducting sub-analysis, etc.

Furthermore, sensitivity and subgroup analyses are often performed to test the robustness of the results reported in meta-analysis when appropriate. Finally, in the attempt to address concerns regarding reporting bias, many authors include funnel plots in final reports. Overall methodology of performing SR is schematically shown in Figure 1.
Figure 1- Methodology for a systematic review of randomized controlled trials

Copyright © BMJ Publishing Group Ltd & Royal College of Paediatrics and Child Health.
There is a general consensus that clinical decisions must be evidence-based. Evidence-based medicine (EBM) referred initially to incorporating current best research evidence into management of individual patients.\textsuperscript{14, 15} The cornerstones of EBM are: research evidence which provides information on the particular person’s health concern; clinical circumstances (context) surrounding the patient’s health concern; and patient’s preferences, values, beliefs, etc. All three key elements of EBM are inter-related and likely of equal value. In guiding individual clinical decisions, available research evidence remains important. However, evidence is rarely sufficient. Moreover, individual circumstances (disease related, social, and financial, etc.) and personal values are likely to influence clinical decisions. In the era of patient-centered care, the latter two components may play even a more prominent role in making clinical decisions as physicians strive to put patients’ values first while balancing risks and benefits of proposed interventions. Finally, the clinician’s judgment is essential to integrate all components and provide evidence-based care. Figure 2 illustrates the key elements of the updated evidence-based decision-making model as depicted by Haynes et al.\textsuperscript{16} In this figure clinical expertise represents a novel, the fourth element of EBM, which includes the physician’s knowledge, clinical expertise, skills to appraise research evidence and communicate it effectively and appropriately to the patient while taking into account the patient’s clinical situation and values. That is why clinical expertise component overlaps the remaining three components of EBM.
Figure 2- An updated model for evidence-based clinical decisions\textsuperscript{16}

\textsuperscript{16}©2002 by BMJ Publishing Group Ltd

The concept of EBM appears to be applicable on the both levels, of individual patients, as well as on the population level and thus, is of growing interest to policy makers. One way to promote EBM principles and integrate research evidence, personal values and clinician’s expertise is via development of clinical practice guidelines (CPG). As stated by the College of Physicians and Surgeons of Ontario, CPGs are aimed “to help improve the quality and consistency of care in specified clinical situations”. These documents are usually developed by a panel of the experts and are based on the synthesis of the best scientific literature available. Although the methodology of developing clinical guidelines is beyond the scope of the present work, SR of the existing body of evidence on the topic of interest is an essential step in developing CPGs. As such, high standards of evaluation of the evidence should be followed to ensure the validity of guidelines.

CPGs are often intended to address specific clinical problems which have significant impact on the society (due to high prevalence of certain disease, for instance, or because of the clinical or economic burden, or due to availability of effective treatment or preventive interventions, etc). The examples are numerous: the Canadian Diabetes Association CPG which provides evidence-based recommendations for healthcare professionals for prevention and management of diabetes, breast cancer screening guidelines developed by Canadian Task Force on Preventive Health Care, etc. Clinicians, therefore, are now trained and strongly encouraged in many institutions to follow CPGs; and indeed, multiple existing guidelines are widely adopted and represent a standard of care in multiple clinical areas. However, the process of CPG development is rather lengthy and expensive. The scope of guidelines might be influenced by particular interests or expertise of guidelines developers involved. Furthermore, the majority of existing guidelines are still far away from incorporating patients’ values and preferences in their clinical recommendations. Finally, in daily practice, there have been innumerous clinical
situations where no guidelines exist and physicians must exercise their professional judgment to advise patients.

Systematic review techniques are particularly useful when a large body of evidence exists and/or results of multiple individual studies are unclear, inconsistent or even contradictory. Ideally, SRs of large RCTs demonstrating consistent results would yield the highest quality of evidence and lead to the strongest clinical recommendations as opposed to observational research due to potential risk of bias. However, SRs are even more needed in situations where current evidence has failed to address a clinical question due to scarcity of primary research and/or absence of high-quality evidence, such as RCTs. Furthermore, it is unlikely that CPGs would be available in such situations to assist clinicians in making clinical decisions.

One area where SRs and meta-analysis of epidemiological data are highly desirable is pediatric pharmacotherapy where high-quality research is limited. Children, along with seniors, pregnant and lactating women, are often underrepresented or specifically excluded from clinical trials. Multiple barriers in conducting pediatric research have been cited, including small number of participants when compared to adults, complex consent process, need to investigate various age groups and different formulations, fear of litigations, and ethical and financial concerns.\textsuperscript{17} Therefore, it is not surprising that approximately 70\% of medications prescribed in pediatric population have not been studied and hence are used in unlicensed or off-label fashion.\textsuperscript{17,18} In newborns this number may reach 80-90\%.\textsuperscript{19,20} For comparison, in adult population off label use of medications has been reported around 21\% on average, with higher percentages for certain groups of medications.\textsuperscript{17} Clinicians, thus, make treatment decisions regarding medications use in the pediatric population based on extrapolation from adult data, from clinical experience and
observational research if available. While high quality RCTs conducted in adults are likely to contribute to the knowledge on pharmacotherapy in pediatric population, drugs effectiveness and safety profile may differ between children and adults. This stems from physiological, age-dependent, differences which affect absorption, distribution, metabolism and clearance of medications. In another words, pharmacokinetic and pharmacodynamic profile of the same medication might be quite different in children, especially newborns, compare to adults. An infamous historical example of drug toxicity in children is the “grey baby syndrome”, a rare serious adverse effect exhibited by newborn infants (especially premature) after intravenous administration of chloramphenicol. This toxicity is believed to be due to insufficient glucuronidation of chloramphenicol in liver and reduced renal elimination of active drug and its metabolite in the first days of life with a net result of persistently elevated plasma concentrations of chloramphenicol.

The deficit in information on drug efficacy and safety for children is of concern. The situation where there is little or no reliable evidence available may deprive children from potentially beneficial therapeutic options. On the other hand, they might be at higher risk for adverse drug reactions (ADRs). It has been suggested that unlicensed or off-label use of medications in pediatric population is associated with 2-5 fold increase risk of ADRs. Furthermore, certain routes of drug exposure, namely via maternal drug use during pregnancy or through breast milk, are unique for pediatric population and may, at least theoretically, pose a risk of ADRs.

Hence, there is an urgent need for high quality research on pharmacotherapy in the pediatric population. Although new legislation has been introduced in Europe and USA to stimulate pharmaceutical companies to conduct trials in pediatric population health care professionals still need to make daily decisions related to patients’ care based on the scant evidence available.
This implies that in the absence of high quality RCTs, which is often the case in pediatric pharmacotherapy, clinicians rely on the best available evidence, largely from observational studies, to make informed health care decisions. While high quality RCTs and SRs of RCTs are more likely to lead to strong recommendations, a particular level of quality does not automatically translate into strong recommendations. Hence, SRs of non-randomized studies might be a valuable and pragmatic approach to summarize scant evidence available in pediatric pharmacotherapy. There have been concerns that combining observational studies and applying meta-analytical techniques is problematic due to poor reporting, potential methodological flaws of original studies and significant heterogeneity in study designs and populations. Nevertheless, SRs of observational data are increasingly published and recognized as complementary to RCTs and more attention has been paid to their methodology and reporting. Additionally, certain clinical questions, such as harmful effects of medications or serious ADRs, due to their rare, delayed or unpredictable nature, are unlikely to be addressed in RCTs. Observational data, thus, would be of great value to investigate these clinical outcomes.

The major methodological issue in any type of research synthesis lies in the quality of included studies and overall quality assessment. Thus, evaluation of methodological quality of the primary research and choosing an appropriate appraisal tool for quality assessment becomes a critical element in conducting SRs. Although the importance of quality assessment has been well recognized, it appears that there is no “gold standard” framework to separate high quality from low quality evidence. Moreover, confusion arises from numerous appraisal tools available. For instance, there are multiple quality assessment tools for RCTs. Similarly, various instruments to evaluate quality of observational research exist but seem to be less rigorous. Furthermore, different instruments have been suggested for appraisal of SRs. To bring more confusion, there are separate tools for quality assessment and for reporting research
evidence with many overlapping items included. Finally, numerous appraisal tools for clinical practice guidelines are available.  

There is also evidence of inappropriate use of existing instruments. Recently, da Costa et al. has found frequent misuse (53%) of the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) instrument designed for reporting observational studies and SRs.  

The STROBE is a check list of 22 items which was developed with the purpose to improve reporting of observational research. In a random sample of 100 publications, the authors identified 19 SRs in total which cited the STROBE. Out of 19 SRs, 10 (53%) utilized the STROBE inappropriately, as an instrument to assess methodological quality of included studies. Additionally, out of 32 observational studies quoted the STROBE, 3 (10%) utilized the STROBE to design and conduct research studies which was deemed inappropriate. The authors further suggested that misuse of STROBE might be explained by the lack of appropriate validated tools for quality assessment of research evidence. As also quoted in the same paper, misuse of respective guidelines is not uncommon in the field of RCTs (4-10%). Although the consequences of such misuse require further elucidation, the main concern is the potential to introduce bias into SRs and meta-analysis. As such, there is a need to overcome shortcomings of existing quality assessment tools and confusion with the multiple instruments available and develop an optimal methodological instrument to grade research evidence and convey this to clinicians in the form of recommendations.
1.2 Brief overview of existing quality assessment tools

1.2.1 Appraisal instruments available – general approach

Historically, one of the first proposed methodological tools to grade scientific evidence was that of the Canadian Task Force on Periodic Health Examination published in 1979. In brief, the authors graded evidence based on study design exclusively and ranked evidence as good (if obtained from at least one properly conducted RCT), fair (if obtained from well designed cohort or case-control studies), or poor (coming from expert opinion). Consequently, the strength of recommendations was directly related to the level of evidence. The main advantage of the proposed system was in its simplicity; however, quality of individual studies has not been taken into consideration.

Pertaining quality of individual studies, the term “methodological quality” has been often defined as an extent to which the study design and execution were appropriate to answer the specific research question and focused largely on study’s internal validity or risk of bias. Thus, not surprisingly, various assessment tools for grading quality of research evidence frequently evaluate susceptibility to bias and include such domains as randomization, blinding, completeness of data, selective reporting, etc. However, as some of these domains are not applicable to observational studies, there have been separate instruments proposed for evaluation of RCTs and observational studies. While such design-specific instruments were developed to address methodological differences of respective research, this approach obviously precludes comparison and judgment on quality of evidence across studies with different designs. To overcome this limitation, generic assessment tools aimed to evaluate variety of study designs, have been developed. Furthermore, as SRs became a primary tool to synthesize overwhelming amount of evidence available, separate instruments have been created to assess
Finally, it has been recognized that incomplete reporting of study results in literature might become an issue precluding accurate assessment of study quality and, hence, separate instruments have been produced to report research evidence and to appraise its quality.\textsuperscript{12, 13}

Summary: different appraisal tools are available for different types of research evidence.

1.2.2 Appraisal instruments available – volume

Back in 1995, while quality assessment of existing research was a relatively new field, Moher and colleagues has performed a comprehensive search and identified 34 instruments (25 scales and 9 checklists) available to evaluate quality of RCTs.\textsuperscript{26} The vast majority of instruments identified have not been rigorously developed. Checklists, as judged by authors, are likely to be more suitable for reporting study results rather than for quality assessment. More recent review by Olivo et al. has confirmed similar weaknesses, namely deficiency of rigorous development and lack of reliability testing, for 21 scales used to evaluate quality of RCTs.\textsuperscript{33}

Concerning quality assessment of observational research, Sanderson et al. reported on total of 86 tools available to assess quality and risk of bias in observational studies.\textsuperscript{28} Similarly to RCTs, only half of the tools identified have provided information on their development and validation testing. Furthermore, out of 121 instruments identified by Katrak et al.,\textsuperscript{25} there have been 26 critical appraisal tools aimed to evaluate quality of SRs. In line with previous reports, only 2 instruments designed for SR quality assessment have reported established face validity and inter-rater reliability.

Finally, Vlayen et al. have identified 24 various appraisal tools to evaluate CPGs. Of these instruments, only 4 have been validated.\textsuperscript{30}
Summary: there are a large number of evaluation tools available, for both primary research and SRs; information on their development process, however, is often lacking.

1.2.3 Appraisal instruments available- content

As evident from Katrak et al. review,\textsuperscript{25} there has been considerable variability in items included in 121 appraisal tools identified. The most frequently reported items across tools were eligibility criteria or sample selection, randomization and blinding or study design reported, appropriate data analysis and sample size justification. Similarly, based on review of 86 tools available to assess observational research, as performed by Sanderson et al.,\textsuperscript{28} the key domains included in majority of instruments were methods to select participants (92\% of tools), methods to measure exposure/outcome/confounding (86\%), other design-specific sources of bias (86\%) statistical methods (78\%) and conflict of interest (\%). The number of included items varied from 3 to 36 among tools with a little consistency in a number and content of items across instruments.

Pertaining SR tools, search strategy, inclusion and exclusion criteria, data synthesis and analysis seem to be important domains. While it appears to be some consistency in the content covered by majority of tools, there has been substantial variability in a number and nature of items across domains and among tools, as well as in scoring ranges. The agreement among different tools is rarely tested.

Concerning CPGs assessment tools, the following dimensions have been identified as an important ones: validity, reproducibility, clinical applicability and flexibility, clarity, regular review and multidisciplinary process, dissemination, implementation and evaluation. Out of 24 instruments, three appraisal tools were comprehensive and covered all domains. On a positive note, validity of CPGs has been a universal domain included in all appraisal tools analyzed. On the other hand, none of the tools assessed the evidence base of the guidelines content.\textsuperscript{30}
Summary: tools’ content is based largely on empirical evidence and there is a general agreement on important domains to be included; variability in content domains across different tools, though, exists and poses the risk of inconsistent results of quality assessment performed for the same study while utilizing different appraisal instruments.

1.2.4 Appraisal instruments available – format

As evident from the above-mentioned reviews, the appraisal tools have been developed in two major formats, scales and checklists. In scales, each item measuring quality is given a numeric score with an overall summary score. Checklists typically comprise of a list of items, and no score is produced. Although scoring the quality of the evidence appears to be attractive in assessing individual studies or performing SRs as offering simplicity, comparability and sense of objectivity, the use of scales and scoring is generally discouraged. The main reason is that scale items are varied significantly among scales and likely to be of variable contribution to overall quality. Simply speaking, the quality of the same study might be ranked differently applying different scales. As such, summary scores may not reflect important differences in studies quality and therefore would question reliability of the assessment. As demonstrated by Jüni et al., results and conclusions of the same meta-analysis varied substantially with different scales utilized. Thus, individual judgments on critical methodological quality domains, rather than summary scores, have been advocated.

Such alternative approach to scale format of quality assessment has been proposed by the Cochrane collaboration. Since 2005, the authors have developed and refined the Cochrane Collaboration Risk of Bias Tool (CCRBТ). The tool covers six major domains of bias:

- selection bias
- performance bias
• detection bias
• attrition bias
• reporting bias
• other bias

Selection bias is often stemmed from inappropriate sequence generation and inadequate allocation concealment. Performance bias might be introduced if blinding of participants and personnel is questionable. Detection bias is closely related to performance bias and refers to knowledge of the allocated interventions by outcome assessors which may influence how outcome determined. Attrition bias may result from incomplete outcome data available. Reporting bias might be introduced with selective outcome reporting. Finally, other bias might be present under certain circumstances. For instance, publication bias, which is a type of reporting bias, refers to the tendency for research reporting statistically significant results (or “positive” studies) being published more often or faster as opposed to “negative” studies. One way to explore the existence of publications bias is a visual inspection of the funnel plot, a scatter plot of the treatment effects estimated from individual studies against a measure of study precision. Asymmetry of the funnel plot may suggest publication bias.

Furthermore, each item/domain of the CCRBT tool is judged as having a high, low, or unclear risk of bias and there have been suggestion to provide a short descriptive summary to support a judgement made. It is recommended to summarise risk of bias in a table format. Furthermore, overall conclusion on quality of evidence (risk of bias) across studies also requires degree of judgements and therefore the Cochrane Handbook supplements detailed information to assist with this process.
Summary: use of scoring for quality assessment is of concern; one alternative approach is a qualitative assessment accompanied by a summary of judgements made. Since a degree of subjectivity is likely to be involved in arriving at conclusions, detailed guidance on tool implementation and transparent reporting of judgements is needed.

1.2.5 Appraisal instruments available- how to choose one?

Overall, the choice of quality assessment tool is not straightforward and there are multiple instruments to choose from. While choosing a candidate tool, it appears to be prudent to consider the following:

- clinical question to be addressed
- amount and type of evidence available on the topic of interest
- intent to assess quality of primary (e.g. individual studies) versus secondary research (e.g. SRs)
- key quality domains to be covered by the proposed tool
- evidence of rigorous tool development
- tool validity and reliability testing
- clear instructions on the tool implementation
- relative simplicity

Pertaining the field of pediatric pharmacotherapy, with a relatively scant research evidence available and lack of RCTs on many important topics, SRs appears to be the best way to arrive at conclusions and thus, generic quality assessment tools, as opposed to design-specific ones, or instruments developed specifically for the assessment of SRs might be considered.
One popular instrument being mentioned already is a Cochrane Collaboration Risk of Bias Tool (CCRBRT). It has been widely adopted by the Cochrane Collaboration which is the main producer of SRs of generally higher methodological quality than non-Cochrane reviews. The advantages of the CCRBT include a comprehensive assessment of bias in RCTs and further in SRs or meta-analyses, some evidence of its inter-rater reliability and validity though variability across domains was reported, and guidance on implementation of this tool. The main disadvantage of the CCRBT, though, lies in the fact that this instrument is applicable only to RTCs.

Another popular instrument designed specifically to appraise quality of SRs is the AMSTAR (“Assessment of Multiple Systematic Reviews”) tool. This 11-item tool appears to cover quite comprehensively all methodological aspects of SR such as literature search and study selection, characteristics of included papers and their quality, appropriateness of meta-analysis and publication bias, etc. Importantly, the AMSTAR has been evaluated in terms of its validity and has demonstrated good reliability and face and content validity as reported by authors and external assessors. The AMSTAR appears to be fairly simple to use.

However, it has been tested mainly on SRs of RCTs. Additionally, the AMSTAR follows a checklist format with little directions how to respond to certain items, for instance, concerning quality of included studies. Moreover, the AMSTAR relies on what has been reported by SR authors rather than on actual conduct of SRs. This approach may pose a risk of flawed conclusions on methodological quality of SRs.

Finally, vast majority of existing tools, including the AMSTAR and CCRBT, are focused on the quality of evidence and does not take into account other essential elements of EBM- the patient’s values and circumstances as well as trade-offs between potential benefits and risks of
interventions. Therefore, they are likely to be useful as an initial step and/or a part of the framework to apply in translation research evidence into clinical care.

Overall, it appears that all existing instruments have some deficiencies and there is a room for their improvement. Moreover, there have been ongoing revisions of existing tools and continuing methodological advances toward finding an optimal system for rating quality of evidence. The GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) system is a recently evolved framework which attempts to integrate individual patients’ preferences and balance between desirable and undesirable consequences of the treatment options into clinical decision making process. As a physician, I believe that treatment decisions should be evidence-based and individualized. Therefore, the GRADE seems to be an appealing alternative to aid individual treatment recommendations which are practical and clinically-relevant.

1.3 The GRADE introduction

The Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach appears to be the most recent methodological development in the field of clinical epidemiology and regarded by many as a new paradigm to apply principles of evidence-based medicine to clinical practice.\textsuperscript{41}

The GRADE Working Group started in 2000 as a collaborative effort to overcome drawbacks of existing systems and develop a comprehensive, explicit and transparent framework for rating quality of research evidence. As a first step, the authors evaluated the existing appraisal tools
and deemed that the vast majority of available instruments are modifications of six major systems. Next, the authors critically appraised these six systems and identified important limitations of all of them.\textsuperscript{42} The following systems have been evaluated: American College of Chest Physicians (ACCP), Australian National Health and Medical Research Council (ANHMRC), Oxford Centre for Evidence-based Medicine (OCEBM), Scottish Intercollegiate Guidelines (SIGN), U.S. Preventive Services Task Force (USPSTF), U.S. Task Force on Community Preventive Services (USTFCPS). The important limitations of currently available systems included: missing domains in the most of systems (mainly those assessing preferences and values of patients), discrepancy in grading evidence versus grading recommendations (some systems appears to do better in grading evidence while others perform better on recommendations part). Additionally, all existing systems have required subjective judgments and inevitable tradeoff between systems’ simplicity and clarity appears to be the universal case (the clearer system is, the less simple it is).

In 2005, based on the critical appraisal of the existing systems, the GRADE authors came up with the initial key features of a newly proposed system and conducted a pilot study to test the GRADE on 12 various health care interventions.\textsuperscript{43} Based on the pilot results, there was agreement that the GRADE is a clear and sensible approach and works well for different types of interventions. Low agreements, though, have been reported on judging balance between harm and benefit. The majority of disagreements has been resolved via discussions and has led to consensus on recommendations in 9 out of 12 examples. Accordingly, the tool has been modified slightly to incorporate deficiencies identified/account for limitations and to improve the GRADE approach and its reporting format.
Furthermore, the authors have developed and published comprehensive series describing in detail the methodology of the GRADE.\textsuperscript{44} In 2011, a complementary series have been introduced to highlight step-by-step application of the GRADE.\textsuperscript{45}

Overall, the GRADE appears to be developed more rigorously compared to other appraisal tools, fully addresses key methodological domains considered important for quality and provides clear and explicit guidance as to how implement this tool. Some scientists, however, may argue that the GRADE has not been validated yet which is a common problem with many of available instruments.

Figure 3 summarizes schematically the steps of the GRADE process.\textsuperscript{46} The major key concepts of the GRADE highlighted in the figure 2, deserve mentioning:

- evaluation of quality of evidence is based on SR of the relevant research
- clear and distinct definitions of quality have been proposed: in the context of SR, quality of evidence is defined as “the extent of confidence that an estimate of effect is correct”; in the context of guidelines development, quality refers to ‘the extent of confidence that an estimate of effect is adequate to support recommendations”
- the strength of recommendations is separated from quality of evidence, aimed for guidelines producers and defined as the extent to which one can be confident the desirable effects of intervention outweigh the undesirable effects
- accordingly, the GRADE offers a framework to rate quality of evidence in SRs/aimed for SR authors and to rate strength of recommendations by guideline authors
- conceptually, the GRADE is a process where recommendations are made based on the quality assessment of systematically gathered evidence
• quality of evidence goes beyond risk of bias and includes additional important domains such as inconsistency of results, indirectness, imprecision of evidence and risk of publication bias

• quality of evidence is outcome specific

• four levels of quality assessment are suggested: high, moderate, low, very low

• quality rating can be upgraded or downgraded based on certain criteria

• the results of the quality rating are presented as an evidence profile (EP) and the summary of findings (SoF) table
Figure 3 – Schematic view of GRADE’s process for developing recommendations

© 2011 Elsevier Inc.

With the aim for developing a generic tool, applicable to variety of research study designs, the GRADE is particularly attractive due to its apparent ability to rate quality of evidence of both
RCTs and observational studies. As many previously mentioned tools, the GRADE quality assessment starts from study design with RCTs being judged initially as high-quality evidence. Respectively, observational research starts off as low quality evidence. There are specific criteria, though, to grade evidence up or down. Risk of bias, referred as study limitations in the GRADE, is assessed based on limitations specific to study design. Limitations of RCTs are similar to those stated in the CCRBT tool and include lack of allocation, lack of blinding, incomplete data and selective outcome reporting. While building on previously developed reputable tools, the GRADE seems to put an emphasis on recognizing other potential sources of bias. For instance, it has been suggested that stopping studies early for benefit may lead to overestimation of treatment effects. It is important, therefore, to be aware of this effect and sensitivity analysis should be conducted to estimate the treatment effect more accurately.

Similarly, the GRADE takes into considerations common limitations of observational studies: imbalance between exposed and unexposed groups in cohort studies, under-matching in case-control studies, failure to control confounding, recall bias, incomplete follow up, etc.

The novelty of the GRADE, though, lies in the approach to assess risk of bias within a study, and further across studies, with respect to each particular outcome.

Apart from study limitations, the GRADE pays attention to other domains which may affect a judgment on quality assessment. Inconsistency in effect size, or heterogeneity, must be explored. If inconsistency remains unexplained, this may prompt down rating of overall quality. Furthermore, imprecision should be assessed, primarily by judging the 95% CI, or by calculating the optimal information size (OIS) as described. Next, using indirect evidence in SRs may justify rating quality of evidence down. Indirectness refers to differences in population, intervention, outcome measures and comparisons of interests and those in reported
studies. Finally, publication bias is common and, if suspected, may lead to rating down study quality.

Some factors, however, may increase the quality of evidence, as suggested by the GRADE. Among them are: a large magnitude of effect consistently demonstrated across studies; dose-response gradient observed; and all plausible confounding would diminish the magnitude of apparent treatment effect.

Figure 4 summarizes the quality assessment criteria set by the GRADE.

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Quality of Evidence</th>
<th>Lower if</th>
<th>Higher if</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized trial</td>
<td>High</td>
<td>Risk of bias</td>
<td>Large effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1 Serious</td>
<td>+1 Large</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2 Very serious</td>
<td>+2 Very large</td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>Inconsistency</td>
<td>Dose response</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1 Serious</td>
<td>+1 Evidence of a gradient</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2 Very serious</td>
<td>All plausible confounding</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>Indirectness</td>
<td>+1 Would reduce a demonstrated effect or</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1 Serious</td>
<td>+1 Would suggest a spurious effect when</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2 Very serious</td>
<td>results show no effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Imprecision</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1 Serious</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2 Very serious</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Very low</td>
<td>Publication bias</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1 Likely</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2 Very likely</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4 – Quality assessment criteria**  
© 2011 Elsevier Inc.

Overall, the GRADE appears to offer a number of advantages when grading quality of evidence compared to prior systems: explicit definitions, clear separation of the concepts of quality and recommendations, comprehensive quality assessment beyond risk of bias which allows for the GRADE application to both RCTs and observational data.

In general, there is still a relative paucity of child-relevant SRs compare to adult population. A recent analysis of pediatric SRs in the Cochrane Database of Systematic Reviews has found that only about a half of articles claimed to be relevant to children, actually reported data on children. The age definitions were often omitted or inconsistent across reports. Furthermore, while mixed data on both adults and children is presented, only 25.2% of reviews attempted to perform separate analyses to ascertain differences between the results of the two groups. Additionally, the authors have reported substantial variations in the assessment of methodological quality across pediatric reviews in general. Inadequate reporting was a noticeable concern. Additionally, many authors have utilized the Jadad scale for overall quality assessment, the approach which has been discouraged as described previously. Publication bias was rarely assessed, partly due to small number of included studies. Overall, this study calls for higher standards in performing and reporting SRs in pediatric research.

While the GRADE framework is aimed to conform to high standards of quality assessment and hence, is being adopted quite widely (although there have been no formal validation studies undertaken to compare the GRADE framework to other available tools), the application of this approach appears to have been very scattered in the field of pediatric pharmacotherapy. A brief recent search of MEDLINE and EMBASE has yielded only 3 publications attempting to implement the GRADE for guidelines development in the field of pediatric pharmacotherapy. The work by Agweyu et al. attempted to use the GRADE framework to refine the Kenyan
national pediatric clinical guidelines on treatment of pneumonia. A paucity of high quality evidence on the selected outcomes of interest was evident which is not surprising. Furthermore, the GRADE appeared to be a user-friendly tool to apply. However, judgments as to how down grade evidence might require some training or degree of expertise.

More interestingly, final recommendations made by the panel of experts appear to be influenced largely by local contextual factors (such as limited resources, costs or feasibility, local preferences and values) and to the lesser extent, by best available research evidence summarized in the SR. This corroborates the previous notion that research evidence might be insufficient for making recommendations and other elements of EBM are equally important.

Boluyt and colleagues\textsuperscript{50} have performed assessment of controversial pediatric asthma management recommendations by applying the GRADE approach to re-evaluate existing evidence. The results of this exercise are quite interesting: the overall quality of existing evidence on selected topics was lower than previously assessed (by other grading systems). In fact, evidence for all three outcomes chosen was scored as “very low” quality despite being based on RCTs. Moreover, the recommendations the authors arrived at, were different from those derived from other international guidelines, mainly due to benefits versus harm consequences of available treatment options. Recommendations were deemed as weak and based on low-quality evidence. This study clearly demonstrates a lack of high quality evidence in the field of pediatric pharmacotherapy and the potential role of the GRADE framework to re-shape clinician’s decisions in the clinical cases of uncertainty.

To summarize, there is a substantial inconsistency in current standards in evaluating quality of evidence in pediatric research, with nearly absent experience to apply the latest methodological advancement available, such as the GRADE, and to explore its potential.
Chapter 2

Hypothesis / Aim

The following research question is being proposed: is the GRADE a feasible and useful methodology to evaluate quality of evidence in pediatric pharmacotherapy research?

The hypothesis is that the GRADE is a feasible methodological tool to assess quality of evidence in pediatric pharmacotherapy where high-quality research is scant and has a potential to impact inferences on quality of evidence drawn after applying one of the alternative assessment tools. To test the feasibility of the GRADE, I will implement step-by-step the GRADE to review systematically the research evidence from both RCTs and observational studies relevant to pediatric population. To test the GRADE usefulness, I will compare the GRADE to other tool available to evaluate quality of evidence.

The objective of the proposed work, thus, is to implement the GRADE instrument to systematically review evidence available on a number of topics in pediatric pharmacotherapy. Further, the aim of the proposed work is to demonstrate the applicability of the GRADE tool in rating quality of evidence gathered from randomized and non-randomized trials, to identify potential methodological gaps with the GRADE implementation and to determine whether conclusions on overall quality of evidence are different while applying the GRADE in comparison to other available tools.

Two specific clinical examples, encountered by the author during clinical practice as a clinical fellow at the MOTHERISK, and driven in large part by limited evidence available, were chosen to illustrate the approach, explore methodological challenges encountered and highlight the
practical use of the GRADE methodology. The MOTHERISK is a well recognized program at The Hospital for Sick Children in Toronto, Ontario, Canada dedicated to providing evidence-based information and counseling to pregnant and nursing women and to health care providers concerning potential risks to the fetus or infant from exposure to various factors including drugs. As a part of my clinical duties, I was asked by lactation consultants at the Hospital for Sick Children and by community physicians to summarize the evidence on the effectiveness of domperidone for enhancing lactation. Domperidone is a widely used medication known for its prokinetic and antiemetic properties. Prescribing domperidine as a galactagogue, though, represents off label use but started while back in 1980’s. There has been empiric evidence of its effectiveness and safety. However, the issue has not been addressed systematically in the scientific literature. Thus, I decided to evaluate current evidence on the effectiveness of domperidone for enhancing lactation due to clinical importance and relevance of the topic.

The focus of another topic is on the risk of ocular toxicity in offspring of mothers exposed to antimalarial medications during pregnancy. This clinical question I have encountered multiple times while counseling women with various rheumatic conditions, predominantly systemic lupus erythematosus (SLE), who were planning pregnancy or were pregnant already. Therefore, both examples were chosen initially to address clinically relevant important questions. Additionally, each topic chosen is aimed to answer the real life question with a common challenge of scant research evidence available. Finally, each example represents a unique challenge and illustrates distinct areas of pediatric pharmacology:
1. Domperidone for enhancing lactation- in the light of controversy following the FDA advisory against its use owing to safety concerns;

2. Ocular toxicity in children following in utero exposure to antimalarial medications: a rare but potentially detrimental adverse effect of commonly used drugs.
Chapter 3

Effect of domperidone on insufficient lactation in puerperal women: a systematic review and meta-analysis of randomized controlled trials

Contents of this chapter have been published in the Obstetrics and Gynecology International:


This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Background:** There is a controversy within the medical community regarding the role of domperidone as a galactagogue and the drug has been removed from the US market owing to safety concerns.

**Objective:** To perform a systematic review and meta-analysis of the available data assessing the effect of domperidone on breast milk production in women experiencing insufficient lactation.

**Study selection:** Randomized controlled trials (RCTs) examining the effect of domperidone on breast milk production of puerperal women were eligible for inclusion.

**Data analysis:** Absolute and relative changes from baseline were calculated for individual studies and pooled estimate using a random effects model.
Results: Three RCTs including 78 participants met the inclusion criteria. All showed a statistically significant increase in breast milk production following treatment with domperidone. The analysis of pooled data demonstrated a statistically significant relative increase of 74.72% (95% CI=54.57; 94.86, p< 0.00001) in daily milk production with domperidone treatment compared to placebo. No maternal or neonatal adverse events were observed in any of the trials.

Conclusions: Evidence from a few small RCTs of moderate to high quality suggests that domperidone produces a greater increase in breast milk supply than placebo.

1. Introduction

The benefits of breastfeeding are well recognized for both the mother and baby; thus, efforts should be made to promote initiation, duration and exclusivity of breastfeeding. The recently published survey of Canadian women who gave birth and were residing with their infants at the time of the interview has found that breastfeeding intention and initiating rates were fairly high, 90% and 90.3% respectively, among women of this representative sample. However, reported exclusive breastfeeding rates at three and six months fell substantially - 51.7 % and 14.4%. While factors that affect breastfeeding success are multiple and nonmodifiable at times, the early recognition and timely management of modifiable risk factors is warranted to improve lactation performance. Various nonpharmacological interventions have been shown to be effective and hence are incorporated in the current clinical recommendations for promoting breastfeeding. Among them are individual and group breastfeeding education provided by lactation specialists, peer counseling, in-person or telephone support. Pharmacological
interventions to improve lactation, mainly dopamine antagonists, are usually recommended only after nonpharmacological modalities have failed, and this is largely due to scarcity of available evidence and potential safety issues with pharmaceutical galactagogues.\textsuperscript{54, 55}

Domperidone, a peripheral dopamine receptor antagonist, is believed to enhance breast milk production by increasing prolactin secretion.\textsuperscript{56-58} It has a favorable safety profile when compared to metoclopramide, another dopamine receptor antagonist, with only rare extra-pyramidal side effects owing likely to poor blood-brain barrier penetration of domperidone.\textsuperscript{59-61}

The drug is well tolerated with relatively few side effects reported including headache, dry mouth, and abdominal cramps.\textsuperscript{60, 62} While domperidone is not available for any indication in the United States due to arrhythmia concerns, it is approved in Canada and other countries as a prokinetic agent. Moreover, there is a worldwide experience with domperidone in treating nausea and vomiting. The use of domperidone as a galactagogue, hence, represents an “off-label” indication.

In 2004, the United States Food and Drug Administration (FDA) issued an advisory against the use of domperidone as a milk enhancer due to safety concerns.\textsuperscript{63} There have been a few reports of cardiac arrhythmia and sudden death in cancer patients treated with intravenous domperidone which are often cited in the literature.\textsuperscript{64} Rapid intravenous administration or high doses of domperidone as well as concurrent hypokalemia might be significant contributors to these adverse outcomes leading to discontinuation of the intravenous route of administration. A single case report of reversible QT prolongation associated with oral domperidone administration has been published.\textsuperscript{65} In neonates, oral administration of domperidone was associated with QT prolongation.\textsuperscript{66} Whereas the potential pro-arrhythmic effect of domperidone should not be ignored, the FDA concern over the use of domperidone for promoting lactation has been
regarded by lactation experts as a gross overestimation. Available pharmacokinetic data, although limited, indicates minimal excretion of domperidone into breast milk with extremely low (less than 0.01% of the maternal weight-adjusted dose) infant exposure via breast milk. No side effects have been reported in exposed infants. The American Academy of Pediatrics lists domperidone as compatible with breastfeeding.

Nevertheless, there is a controversy regarding the role of domperidone as a galactagogue: some authors claim no or little effectiveness, largely due to limitations of available data while other researchers suggested that domperidone is a galactagogue of choice based on evidence available. This situation might be a source of confusion in the medical community and, therefore, may compromise clinical management decisions.

The objective of our study was to perform a systematic review and meta-analysis of the available data assessing the effect of domperidone on breast milk supply in women experiencing insufficient breast milk production.

2. Methods

2.1 Eligibility criteria

Randomized controlled trials (RCTs) examining the effect of domperidone on breast milk production were considered for inclusion. We utilized the PICO format (population, intervention, comparison, and outcome) to develop our clinical question, guide the literature search and assess eligibility of potentially relevant studies. The population of interest was puerperal women who had experienced insufficient lactation after delivery. We accepted any definition of insufficient lactation, with the most common definition being milk supply below
the infant’s daily oral feeding requirements. The intervention considered for this paper was domperidone treatment to augment lactation; the comparator considered was placebo or no treatment. The outcome of interest was percent change in daily breast milk volume after domperidone treatment.

2.2 Search strategy

The following electronic databases were searched: Ovid MEDLINE(R) (1948 to May 2011), EMBASE (1947 to May 2011), and Cochrane Library, with no restrictions on language or year of publication. Our search strategy included the following National Library of Medicine Medical Subject Headings (MeSH) terms: “domperidone” combined with “lactation” OR “milk production” OR “galactagogue” OR “breastfeeding”. The search was limited further to human data and clinical trials. Reference lists of relevant review papers and all selected articles were hand searched to identify additional trials.

2.3 Study selection and quality assessment

Literature search and eligibility assessment was performed independently by two reviewers. One reviewer extracted the data and performed quality assessment of included trials. The second reviewer checked the extracted data and quality assessment. Disagreements in judgment between reviewers were resolved by discussion. The following data was extracted: characteristics of trial participants (number, inclusion criteria), type of intervention (dose and duration of domperidone or placebo treatment), outcome measure (type and assessment tool), and maternal and neonatal adverse effects reported.
Study quality was assessed using the GRADE (grading of recommendations assessment, development, and evaluation) system.\textsuperscript{44} The GRADE system was developed by a widely representative group of scientists and adopted by the Cochrane Collaboration to assess the quality of evidence for outcomes reported in systematic reviews. Each individual study was rated as that of high, moderate, low or very low quality. The Cochrane collaboration’s tool has been applied to assess risk of bias across studies. The following domains were evaluated - sequence generation, blinding, allocation concealment, incomplete outcome data, selective outcome reporting, and other sources of bias. Randomized controlled trials are generally rated as a high quality but might be downgraded. Factors that may decrease the quality of evidence include serious limitations in design, imprecision of results, unexplained heterogeneity, and indirectness of evidence and high probability of publication bias.

2.4 Statistical analysis

The primary effect measure for this paper was the change in daily breast milk volume from baseline to the end of the treatment presented as a mean difference and standard deviation. Absolute and relative changes from baseline were recorded for individual studies. Absolute mean differences in daily breast milk volumes before and after treatment were extracted from individual studies. Relative mean differences were calculated as percentage change from baseline.

When the standard deviations of the absolute changes from baseline were not available from individual studies, we imputed them as described in detail in the Cochrane Handbook.\textsuperscript{71-73} In brief, we calculated correlation coefficients from one available study which reported the means and standard deviations for change in breast milk volume from baseline.\textsuperscript{58} Using the imputed
correlation coefficients values, we thereafter calculated a change from baseline standard
deviations for the other studies with missing standard deviations.\textsuperscript{69,70} A sensitivity analysis was
performed utilizing the lowest and highest values of the correlation coefficient to determine the
robustness of the results.

The standard deviations of relative change (\%) were calculated as

\[ SD_{relative \ change} = SD_{absolute \ change} / \text{breast milk volume}_{baseline}. \]

Pooled estimates of the weighted mean differences and 95\% CI were calculated using a random
effects model. The \( I^2 \) statistic was used to assess the extent of heterogeneity among studies. A
priori sub-group analyses were not planned. Due to insufficient number of studies, a formal
assessment of reporting bias by visual inspection of a funnel plot was not possible.

3. Results

3.1 Study selection

The literature search retrieved a total of 24 citations (Figure 5). After duplicate publications
were eliminated, 18 remaining abstracts were screened for eligibility. Of these, six were
excluded (five were deemed not relevant and one was a case report). The full text of the
remaining 12 citations was analyzed further in detail. Nine papers were excluded due to various reasons. Three studies met the inclusion criteria and were included in the systematic review and
meta-analysis.\textsuperscript{58,69,70}
Figure 1. Flow chart of selected studies

Total number of records identified: 24
- Ovid MEDLINE (3)
- EMBASE (11)
- Cochrane Library (10)

Number of records after duplicates removed: 18

Number of records screened: 18
- Number of records excluded: 6
  - not relevant (5)
  - case report (1)

Number of articles assessed for eligibility as a full text: 12
- Number of articles excluded: 9
  - review paper (4)
  - no population of interest (1)
  - no outcome of interest (2)
  - no comparator of interest (1)
  - no full text available (1)

Number of articles included in review and meta-analysis: 3

Figure 5-Flow chart of selected studies
3.2 Study characteristics

All three studies selected for this review were randomized, placebo-controlled trials with a total of 78 patients enrolled (37 in domperidone group and 41 in placebo group). Table 1 summarizes characteristics of included studies. All participants have experienced inadequate breast milk production postpartum and, therefore, were randomized to domperidone or placebo. Of note, all mothers were enrolled after a few weeks postpartum allowing time to establish lactation and/or receive appropriate lactation support. However, only one study mentioned extensive lactation counseling prior to enrolment.

The dose of domperidone used across the studies was 30 mg/d (10 mg orally 3 times daily). The length of the treatment ranged from 7 to 14 days. All studies reported the change in daily milk production from baseline to the end of the treatment. In Petraglia et al., the mothers breastfed their full-term infants, and thus daily milk volumes were assessed by weighing the babies before and after breastfeeding. In two other studies, the mothers pumped breast milk to feed their preterm babies and the amount of milk expressed was recorded.
3.3 Methodological quality of included studies

Table 2 displays the summary of risk of bias for individual studies included in the meta-analysis.

Two of the studies, by Da Silva et al.\textsuperscript{58} and Campbell-Yeo et al.\textsuperscript{70} were ranked as having low risk of bias. The description of randomization, allocation concealment, blinding, and reporting in these two papers was judged as adequate. Da Silva et al.\textsuperscript{58} however, reported incomplete or nonreturned records for three out of 11 participants in the domperidone group which represents missing data for >25% of participants. Overall, both studies were judged as free of serious limitations and were graded as high quality evidence.

The study done by Petraglia et al.\textsuperscript{69} to the contrary, did not provide sufficient information on sequence generation and allocation concealment. Furthermore, the study is described as a double-blind trial; however, there is no information whether placebo and active drug were of similar appearance and taste. It is also unclear from the paper whether all women randomized initially completed the trial. Given the above-mentioned limitations in the study design and implementation, Petraglia et al was downgraded from high- to moderate-quality evidence.

3.4 Results of individual studies

Three included RCTs evaluated the effect of domperidone on a daily breast milk volume in the women with insufficient lactation in comparison to placebo.

Da Silva et al. reported that after 7-day treatment, the mean daily milk volume had increased by 49.5 (SD=29.4) mL/day in the domperidone group compared to 8.0 (SD=39.5) mL/day in the placebo group.\textsuperscript{58} Similarly, Petraglia et al demonstrated that, following 10-day treatment, daily
milk yield was significantly higher in a small group of domperidone-treated mothers than that of the placebo-treated group.69 The mean increase in daily milk yield was 326 (imputed SD= 21.4) mL/day after domperidone versus 63 (imputed SD=23.7) mL/day after placebo treatment. Finally, a significant increase in daily breast milk production was found in Campbell-Yeo et al70: mean increase of 195.8 (imputed SD=98.1) mL/day after a 14-day course of domperidone compare to 33.1 (imputed SD=83.2) mL/day in a placebo-treated group. Overall, in absolute values, all three studies had shown a statistically significant increase from baseline in breast milk production following treatment with domperidone.

Due to substantial differences in baseline milk volumes across the studies, the relative changes from baseline were calculated and used to estimate the pooled effect of domperidone (Figure 6).

3.5 Pooled Analysis

The analysis of pooled data demonstrated a statistically significant increase, of 74.7% (95% CI=54.6, 94.9, p< 0.00001) in daily milk production following treatment with domperidone while comparing to placebo. We observed a moderate statistical heterogeneity among the studies ($I^2 = 50\%$).
<table>
<thead>
<tr>
<th>Reference</th>
<th>N of participant, Intervention/placebo groups</th>
<th>Inclusion criteria</th>
<th>Domperidone dose</th>
<th>Domperidone duration of treatment</th>
<th>Outcome</th>
<th>Outcome assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Petraglia et al</strong> &lt;sup&gt;69&lt;/sup&gt;</td>
<td>9/8</td>
<td>Premiparous mothers of term infants with insufficient lactation &lt;sup&gt;a&lt;/sup&gt;</td>
<td>30 mg/day</td>
<td>10 days</td>
<td>Daily breast milk yield, before and after treatment, mL/day</td>
<td>By weighing the infants before and after breastfeeding using an electronic integrating scale and summarizing the single milk yields for the day</td>
</tr>
<tr>
<td><strong>1985</strong></td>
<td></td>
<td>2 weeks post partum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Da Silva et al</strong> &lt;sup&gt;58&lt;/sup&gt;</td>
<td>7/9</td>
<td>Mothers of preterm infants with low milk production &lt;sup&gt;b&lt;/sup&gt;</td>
<td>30 mg/day</td>
<td>7 days</td>
<td>Daily breast milk volume, before and after treatment, mL/day</td>
<td>Mechanically expressed breast milk by using a double collecting pump</td>
</tr>
<tr>
<td><strong>2001</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Campbell-Yeo et al</strong> &lt;sup&gt;70&lt;/sup&gt;</td>
<td>21/24</td>
<td>Mothers of preterm infants ( &lt; 31 weeks gestation) with lactation failure &lt;sup&gt;c&lt;/sup&gt; ≥ 3 wks after delivery</td>
<td>30 mg/day</td>
<td>14 days</td>
<td>Daily breast milk volume, before and after treatment, mL/day</td>
<td>Mechanically expressed breast milk by using a double collecting system</td>
</tr>
<tr>
<td><strong>2010</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> insufficient lactation defined as milk yields at least 30% lower than those reported as normal

<sup>b</sup> low milk production defined as not meeting the infant’s daily oral feeding requirements

<sup>c</sup> lactation failure defined as one of the following: a decreasing milk supply by > 30% from peak volume based on maternal count; or inability to meet the daily nutritional intake of the infant
<table>
<thead>
<tr>
<th>Studies, year</th>
<th>Sequence generation</th>
<th>Allocation concealment</th>
<th>Blinding of participants and personnel</th>
<th>Blinding of outcome assessment</th>
<th>Incomplete outcome data</th>
<th>Selective outcome reporting</th>
<th>Other bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Petraglia et al⁶⁹ 1985</td>
<td>Insufficient info/</td>
<td>Insufficient info/</td>
<td>Insufficient info/</td>
<td>Insufficient info/</td>
<td>Insufficient info/</td>
<td>No/low risk of bias</td>
<td>No/low risk of bias</td>
</tr>
<tr>
<td></td>
<td>unclear risk of</td>
<td>unclear risk of</td>
<td>unclear risk of</td>
<td>unclear risk of</td>
<td>unclear risk of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>bias</td>
<td>bias</td>
<td>bias</td>
<td>bias</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Da Silva et al⁵⁸ 2001</td>
<td>Yes/low risk of</td>
<td>Yes/low risk of</td>
<td>Yes/low risk of</td>
<td>Insufficient info/</td>
<td>No/low risk of</td>
<td>No/low risk of bias</td>
<td></td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>bias</td>
<td>bias</td>
<td>unclear risk of</td>
<td>bias</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>bias</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Campbell-Yeo et al⁷⁰ 2010</td>
<td>Yes/low risk of</td>
<td>Yes/low risk of</td>
<td>Yes/low risk of</td>
<td>No/low risk of</td>
<td>No/low risk of</td>
<td>No/low risk of bias</td>
<td></td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>bias</td>
<td>bias</td>
<td>bias</td>
<td>bias</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2- Methodological quality of RCTs included in the meta-analysis
Figure 6-Percent change in milk volume with domperidone treatment

<table>
<thead>
<tr>
<th>Study or subgroup</th>
<th>Domperidone Mean (SD)</th>
<th>Placebo Mean (SD)</th>
<th>Mean difference IV, random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campbell-Yeo</td>
<td>106.2 (53.2)</td>
<td>21</td>
<td>29.3% [63.59, 118.41]</td>
</tr>
<tr>
<td>DaSilva</td>
<td>43.9 (26.1)</td>
<td>7</td>
<td>10.5% [-29.45, 84.45]</td>
</tr>
<tr>
<td>Petraglia</td>
<td>94 (6)</td>
<td>9</td>
<td>60.3% [68.76, 81.24]</td>
</tr>
<tr>
<td>Total (95% cl)</td>
<td>37</td>
<td>41</td>
<td>74.72 [54.57, 94.86]</td>
</tr>
</tbody>
</table>

Heterogeneity: $\tau^2 = 165.17; \chi^2 = 3.97, df = 2 (P = .14); I^2 = 50%$

Test for overall effect: $Z = 7.27 (P < .00001)$
4. Discussion

Our findings indicate that domperidone increases inadequate breast milk production in nursing mothers more effectively than placebo. A statistically significant increase in the mean change in daily breast milk volume from baseline was observed in all three studies comparing domperidone and placebo. This consistency of the domperidone effect across the studies enhances the confidence of its beneficial effect as a galactagogue. Importantly, no maternal or neonatal adverse events were observed in any of the three trials. Although not included in this analysis, the study by Wan and colleagues demonstrated a dose-response increase in milk production, further supporting our findings.62

Another strength of the current meta-analysis lies in the fact that, despite the paucity of published reports on effectiveness of domperidone to promote lactation, we attempted to identify and include only randomized placebo-controlled studies which are regarded as higher quality evidence. Two out of the three included trials fulfilled the GRADE criteria for high quality evidence,58,70 while the third study69 was downgraded to moderate-quality evidence due to insufficient details on study design and execution and, therefore, as having a potential risk of bias.

Based on the moderate-high quality of evidence from three RCTs, the pooled effect of a 75 % increase from baseline in daily milk production following treatment with domperidone is deemed to be clinically meaningful.
Our study has several limitations. Only three eligible studies were found with small sample sizes (17, 16, and 45 participants in each trial, resp.). It has been suggested that small trials with an insufficient number of participants and events may produce spurious treatment effects due to random error. Furthermore, there have been reports showing that some meta-analyses become inconclusive when adjusted for random error risk. Hence, the calculation of optimal information size (similar to the concept of sample size calculation for individual studies) and the use of appropriate statistical tools (i.e. trial sequential analysis) have been advocated to judge results of meta-analysis as reliable and conclusive.

On the other hand, it is unclear how many studies are needed to be included in meta-analysis to render results trustworthy. Moreover, some researchers have demonstrated that meta-analysis with a few trials do produce robust results consistent with long-run findings. However, it is difficult to foresee which results might be changed by subsequent large-scale trials. Nonetheless, owing to the above-mentioned limitations, our findings must be interpreted with caution, and generalizable recommendations might be still premature.

Additionally, a moderate statistical heterogeneity was found to exist across the studies’ results. While all three trials have utilized the same doses of domperidone (or placebo) and reasonably similar duration of treatment, the differences in study populations (mothers of pre-term versus full-term infants, breastfeeding or pumping their milk) and outcome measurement instruments (increase in milk supply versus infant weight gain) are likely to explain the observed heterogeneity. We have used a random effects model for the pooled estimate to deal with statistical heterogeneity. However, too few studies available precluded subgroup analyses to further explore the observed heterogeneity. We believe though, that the selected trials were methodologically sound to combine in the present meta-analysis. The clinical relevance of this modest heterogeneity is probably not meaningful as there is a considerable consistency in
domperidone effect across individual studies and no biological reason to suspect that the opposite direction effect might be true. Still, it is sensible to investigate potential sources of heterogeneity as more research on this topic becomes available.

One methodological challenge we encountered in the present meta-analysis is not uncommon and thus deserves special mention. The issue is related to handling missing variance estimates data in primary studies included in meta-analysis. Two out of three RCTs selected for our review failed to provide standard deviations (SDs) for changes from baseline which we selected as a primary effect measure. There have been several methods proposed to impute missing variance estimates for continuous outcomes.\textsuperscript{71-73} Since Da Silva et al\textsuperscript{58} reported SD, we were able to calculate the correlation coefficient, a measurement of similarity between the baseline and final measurements across participants from this study, and then apply the calculated value to impute a change-from-baseline standard deviation for two other studies included in our meta-analysis. In general, correlation coefficient of zero indicates no correlation which is unusual for clinical outcomes as we expect certain degree of association between measurements within an individual. Similarly, a correlation coefficient of one is unlikely due to certain variability present within an individual. The calculated correlations obtained from Da Silva et al\textsuperscript{58} were reasonably similar and close to 1 for the domperidone and placebo groups (0.97 and 0.78, resp.). We used an average correlation coefficient of 0.875 to impute the missing change-from-baseline standard deviations for the remaining studies. A sensitivity analysis was performed utilizing the lowest and highest values of the correlation coefficient and repeating the analysis. This did not change our overall conclusion as the pooled estimates and confidence intervals were not significantly changed in terms of magnitude or directionality (data not shown). Although a certain degree of uncertainty exists regarding the accuracy of the results derived from this
approach, there is a growing body of the literature indicating the validity of results from meta-
alyses utilizing various imputation methods. 78, 79

5. Conclusions

Currently available data from a few small randomized controlled trials suggest that domperidone produces greater increase in breast milk supply than that found with placebo in some puerperal women with insufficient milk production. These results however, should be interpreted in the context of the limitations of available data. Additional randomized clinical trials of adequate sample size are desirable and might have an impact on our confidence in the estimate of domperidone effect as a galactagogue. In the realm of clinical practice, however, while the balance between desirable and undesirable effects often guides treatment decisions, the current analysis supports consideration that domperidone might be an effective treatment option for selected women with inadequate lactation. It appears to be prudent though to try nonpharmacological interventions, for example, maternal lactation education, first.51,55
Chapter 4

Ocular toxicity in children exposed in utero to antimalarial drugs: a systematic review of the literature
ABSTRACT:

Background: The antimalarial drugs chloroquine (CQ) and hydroxychloroquine (HCQ) have been used for decades to treat various rheumatic diseases. CQ is still beneficial for the management of malaria during pregnancy. A growing body of research accumulated to date suggests that antimalarials are safe during pregnancy. However, there have been concerns about adverse long-term effects, mainly retinal toxicity, in offspring of women exposed to antimalarials during pregnancy.

Objective: To systematically review the published evidence on safety of antimalarials during pregnancy focusing on ocular toxicity in the offspring.

Methods: Ovid MEDLINE(R), EMBASE and Cochrane Library databases were searched for the period from their inception to May 2010 inclusive with no restrictions on language or year of publication. Randomized controlled trials (RCTs) and observational studies examining the safety of CQ or HCQ during pregnancy and reporting on visual function or ocular toxicity in the offspring of exposed women at any point of the follow up were eligible for inclusion. The quality of evidence was assessed according to the established criteria the (GRADE approach).

Results: 12 studies with a total of 588 offspring born to mothers treated with CQ or HCQ during pregnancy met the inclusion criteria. Five studies with a total of 251 exposed children reported
no clinical visual abnormalities in any case. In a RCT on malaria prophylaxis, visual acuity in 251 infants exposed to CQ in utero did not differ from the placebo group. Detailed ophthalmological examination was performed in four studies and normal results were reported in all children (n=59). Electrophysiological testing using electroretinogram (ERG) was performed in three small cohorts and were normal in all but six infants 3-7 months of age. All six children had normal fundoscopy before 4 years of age. Heterogeneity in comparison groups and in outcome measures precluded formal meta-analysis.

**Conclusions:** The current evidence suggests no fetal ocular toxicity of antimalarial medications during pregnancy. The clinical significance of early electroretinogram anomalies reported in a small subset of infants remains to be established. Larger follow up studies are warranted to confirm low risk of ocular toxicity in children following antenatal exposure to antimalarial medications.

**Background**

Despite recent advances in the management of rheumatic diseases, antimalarial drugs still have an established beneficial role in the treatment of various rheumatic conditions such as cutaneous and systemic lupus erythematosus (SLE) and rheumatoid arthritis. Hydroxychloroquine (HCQ) and chloroquine (CQ) have been demonstrated to reduce the risk of lupus flares and improve long–term survival of patients with SLE. Additionally, CQ has been used extensively in the past for the prophylaxis and treatment of malaria but currently is replaced by other drugs due to parasite resistance. Still, CQ is recommended for the management of non-falciparum malaria during pregnancy in certain endemic areas. Rheumatic conditions tend to have chronic course and occur more often among women of childbearing age, therefore it is not uncommon to have a successful pregnancy on antimalarial medications. Furthermore, pregnant women are
thought to be particularly susceptible to malaria infection with an estimated 50 million pregnancies exposed annually to malaria worldwide.\textsuperscript{82}

For years, there have been concerns regarding potential harmful effects of antimalarial agents on the developing fetus. A growing body of research accumulated to date suggesting no increased risk of teratogenicity following exposure to antimalarials during pregnancy. Recent systematic reviews including data on more than 300 exposed offspring have shown concordant results: HCQ/CQ use in pregnancy is not associated with increased risk for birth defects, spontaneous abortions, fetal death and prematurity in patients with autoimmune conditions.\textsuperscript{80,83,84} Based on the experience with CQ for malaria prophylaxis, this drug is regarded as safe in pregnancy albeit much lower doses are usually prescribed in comparison to those used for rheumatic conditions.\textsuperscript{81,82}

Although the issue of possible teratogenicity of antimalarials appears to be addressed in the published literature, there is still a concern regarding potential toxic effects, mainly retinal toxicity, in the offspring of women exposed to antimalarials during pregnancy.

Adverse ocular effects of antimalarials in adults are numerous and include keratopathy, ciliary body dysfunction, lens opacities and retinopathy. Keratopathy with corneal deposits is a common, usually benign, ophthalmological finding which is completely reversible upon drug discontinuation with no permanent corneal damage.\textsuperscript{85} To the contrary retinal toxicity, although rare, is a major concern following treatment with antimalarial medications as it may lead to visual field defects, decreased visual acuity and permanent visual loss. It is believed that CQ and HCQ bind to melanin in the retinal pigment epithelium and this may contribute to the drug induced toxicity).\textsuperscript{86,87} CQ binds more tightly and is thought to be more toxic than HCQ.\textsuperscript{88} It has been suggested that daily and cumulative dose of antimalarials are important risk factors for
retinopathy associated with long-term use of these medications. The vast majority of reports described retinal toxicity on daily doses exceeding 6.5 mg/kg/day of HCQ (or 3 mg/kg/day of CQ) or after prolonged treatment ( > 5 years)\textsuperscript{85,89} providing basis for the guidelines for risk assessment of retinopathy in CQ- and HCQ-treated patients. The most recent update on screening recommendations advises not to exceed daily doses of 400 mg HCQ or 250 mg CQ. Even lower doses are suggested for the patients with short stature who are at risk of overdose. For such individuals the “ideal” body weight needs to be taken into consideration while calculating the daily doses. Similarly, obese patients should be medicated on the basis of their “ideal” body weight.\textsuperscript{90}

Given a real risk of retinal toxicity following long-term treatment with antimalarials, it is plausible to consider similar toxicity in the offspring of women taking these medications during pregnancy. There have been studies demonstrating that CQ and HCQ cross the placenta and accumulate in fetal eye tissues.\textsuperscript{87,91-93} Isolated human case reports of retinal degeneration in infants prenatally exposed to CQ have led to further concerns.\textsuperscript{94} There have been a few studies examining visual function of babies exposed in utero to antimalarials with no cases of retinal toxicity reported.\textsuperscript{95,96} Nevertheless, the issue has not been addressed sufficiently in literature and recent systematic reviews have not focused on retinal toxicity in the offspring of exposed women. Hence, the objective of this study was to perform a systematic review of the current literature on safety of antimalarial agents during pregnancy with focus on ocular toxicity in the offspring.

**Methods**

**Search strategy and study selection**
Two authors searched Ovid MEDLINE(R), EMBASE and Cochrane Library electronic databases for the period from their inception to May 2010 inclusive, with no restrictions on language or year of publication. Our search strategy included the following National Library of Medicine Medical Subject Headings (MeSH) terms: “chloroquine” OR “hydroxychloroquine” combined with “pregnancy” OR “congenital, hereditary, and neonatal diseases and abnormalities” OR “prenatal exposure delayed effects” OR “embryo and fetal development” OR “embryonic structures” OR “teratogens”. The search was further limited to human data and we excluded editorials and case reports. Review articles were not included but searched further to identify potentially relevant publications. Reference lists of all retrieved studies and review articles were hand searched to identify additional studies.

Eligible studies included randomized controlled trials and observational studies examining ocular effects of CQ or HCQ exposure during pregnancy and reporting clinical outcomes in the offspring of exposed women. The titles and abstracts were screened for relevance by assessing the population of interest (pregnant women), exposure to antimalarials, and outcome of interest (visual function or ocular toxicity). Potentially relevant articles were thereafter retrieved as a full text and further examined. Two authors reviewed all full text articles independently to determine eligibility for inclusion. Differences in judgment between the reviewers were resolved by consensus discussion. The following information was extracted from studies deemed eligible: study design, drug of exposure, length of follow up, number of subjects exposed, comparison group, outcome of interest, outcome measurement instruments, and reported findings.

**Quality assessment**

The selected studies were critically appraised utilizing the GRADE (grading of recommendations assessment, development and evaluation) approach. The GRADE system
was developed by a group of experts and adopted by the Cochrane Collaboration to assess the quality of evidence for each individual outcome reported in systematic reviews. The GRADE evaluates the risk of bias across six domains—sequence generation, allocation concealment, blinding, incomplete outcome data, selective outcome reporting and other sources of bias; each of the domains is judged as having high, low or uncertain risk of bias. The overall quality of evidence is rated as high, moderate, low and very low. Randomized trials are generally graded as a high quality whereas observational studies as a low quality evidence. However, the reviewers may downgrade or upgrade the quality of evidence based on the specific criteria.

**Results**

The initial search strategy yielded 790 abstracts (Figure 7); 753 were excluded as not relevant. The remaining 37 abstracts were evaluated further as a full text. Of these, 12 studies met the inclusion criteria and were included in this systematic review (Table 3).\(^{95-106}\) They included 588 children born to mothers treated with CQ or HCQ during pregnancy. Eleven out of 12 studies reported on the use of antimalarial medications, predominantly HCQ, for the treatment of rheumatic diseases and included 337 exposed children.\(^{95-105}\)
Figure 7 - Flow chart of selected articles
<table>
<thead>
<tr>
<th>Author, year</th>
<th>Type of study</th>
<th>Follow up (range)</th>
<th>Drug exposure</th>
<th>No of children exposed</th>
<th>No of non-exposed controls</th>
<th>Outcome measure/test used</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Renault et al, 2009</td>
<td>Prospective cohort</td>
<td>11-266 days</td>
<td>HCQ</td>
<td>21</td>
<td></td>
<td>Retinal assessment using ERG and VEP</td>
<td>6-abnormal ERG, of these 4 – delayed VEP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All 6 – normal fundus before age 4</td>
</tr>
<tr>
<td>Villegas et al, 2007</td>
<td>RCT</td>
<td>1 year</td>
<td>CQ</td>
<td>251</td>
<td>246</td>
<td>Visual acuity</td>
<td>No difference in median visual acuity between groups</td>
</tr>
<tr>
<td>Clowse et al, 2006</td>
<td>Prospective cohort</td>
<td>Postpartum NS</td>
<td>HCQ</td>
<td>79</td>
<td>145</td>
<td>Clinical evaluation of vision as a part of routine infant care</td>
<td>No visual impairments in any offspring</td>
</tr>
<tr>
<td>Motta et al, 2005</td>
<td>Prospective cohort</td>
<td>1 year</td>
<td>HCQ</td>
<td>24</td>
<td>No control</td>
<td>Ophthalmological evaluation (anterior segment exam, ocular motility, pupillary size, reaction to light, refraction, indirect ophthalmoscopy)</td>
<td>2 had retinal hemorrhages at birth resolved by the age of 1 month; normal ophthalmological exam in all children</td>
</tr>
<tr>
<td>Study</td>
<td>Design</td>
<td>Follow-up Duration</td>
<td>Treatment</td>
<td>Control</td>
<td>Outcome</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------------</td>
<td>--------------------</td>
<td>-----------</td>
<td>---------</td>
<td>---------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>Cimaz et al, 2004</td>
<td>Prospective cohort</td>
<td>3-30 months</td>
<td>HCQ</td>
<td>No control</td>
<td>ERG</td>
<td>all normal</td>
<td></td>
</tr>
<tr>
<td>Costedoat-Chalumeau et al, 2004</td>
<td>Prospective cohort</td>
<td>4-13 months</td>
<td>HCQ</td>
<td>No control</td>
<td>ERG</td>
<td>all normal</td>
<td></td>
</tr>
<tr>
<td>Costedoat-Chalumeau et al, 2003</td>
<td>Prospective cohort</td>
<td>12-108 months (mean 26 months)</td>
<td>HCQ</td>
<td>117</td>
<td>59</td>
<td>No visual abnormalities as reported by mother, GP or pediatricians</td>
<td></td>
</tr>
<tr>
<td>Levy et al, 2001</td>
<td>RCT</td>
<td>1.5-3 year</td>
<td>HCQ</td>
<td>10</td>
<td>9</td>
<td>All children-normal ophthalmologic exam</td>
<td></td>
</tr>
<tr>
<td>Klinger et al, 2001</td>
<td>Prospective cohort</td>
<td>mean 2.8 years, SD 2.9 years</td>
<td>HCQ / CQ</td>
<td>21</td>
<td>Comparisons made to normal data for healthy children</td>
<td>Normal in all children</td>
<td></td>
</tr>
<tr>
<td>Buchanan et al, 1996</td>
<td>Retrospective cohort</td>
<td>NS</td>
<td>HCQ</td>
<td>32</td>
<td>44</td>
<td>No visual disturbances clinically in exposed infants</td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Design</td>
<td>Follow-up Duration</td>
<td>Treatment</td>
<td>Duration</td>
<td>Control Group</td>
<td>Method of Evaluation</td>
<td>Result</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-----------------</td>
<td>--------------------------</td>
<td>-----------</td>
<td>----------</td>
<td>---------------</td>
<td>-----------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Parke and West, 1996</td>
<td>Prospective cohort</td>
<td>10-72 months (mean 33 months)</td>
<td>HCQ</td>
<td>9</td>
<td>No control group</td>
<td>NS</td>
<td>No visual abnormalities reported</td>
</tr>
<tr>
<td>Levy et al, 1991</td>
<td>Prospective cohort</td>
<td>range 9 - 19 years (mean 5.3 years)</td>
<td>HCQ/CQ</td>
<td>14</td>
<td>No control group</td>
<td>Clinical evaluation of vision</td>
<td>No visual loss as reported by family doctors or pediatricians</td>
</tr>
</tbody>
</table>

HCQ-hydroxychloroquine; CQ- chloroquine; RCT- randomized clinical trial; ERG-electroretinogram; VEP- visual evoked potentials; NS- not specified
There was one trial on CQ use for malaria prophylaxis reporting on 251 exposed infants. Of the 11 studies in rheumatic conditions, one was a RCT and the remaining 10 were cohort studies. Three cohort studies utilized control groups of women with similar diseases, 2 other cohort studies used comparisons to normal values in healthy subjects. The remaining 5 cohorts lack comparison groups. The only one eligible study on malaria prophylaxis was a RCT.

Methods of visual assessment varied greatly among studies. Five studies with a total of 251 exposed children performed clinical evaluation of visual function and found no visual abnormalities in any case as reported by either mother, GP or pediatrician. The time of clinical assessment ranged widely from 10 months to 19 years and was not specified in two studies. In a RCT on malaria prophylaxis, visual acuity was assessed in 251 infants exposed to CQ in utero at 1 year of age and did not differ from placebo group. Detailed ophthalmological examination was performed in four studies and normal results were reported in all children (n=59). Ophthalmological examination was performed during the first year of life or later.

Electrophysiological testing using electroretinogram (ERG) was performed in three small cohorts of infants exposed to HCQ prenatally (n= 31) and were normal in all but six infants of 3-7 months of age. These six children had normal fundoscopy by 4 years of age.

**Study Quality**

The majority (83%) of included reports were observational studies rated as providing low quality evidence. 2 RCTs were ranked as having uncertain risk of bias. The study by Villegas et al demonstrated adequate sequence generation and allocation of concealment,
however it is unclear whether the outcome assessor was blinded to the exposure and no reasons were given for missing outcome data (> 25%). The study of Levy et al \textsuperscript{105} did not provide adequate description of the method of randomization and concealment. Outcome assessors were blinded and outcome data were complete. Overall, both studies were downgraded from high to moderate quality evidence based on the limitations in study design and implementation as described above.

**Discussion**

To the best of our knowledge, this is the first attempt to assess systematically the potential harmful effect of antimalarial medications on visual function in offspring exposed \textit{in utero} to these medications. Collectively, the total number of exposed children (n=588) and the nearly uniform absence of visual abnormalities across the studies suggests low to non existing risk of retinal toxicity in the infants following antenatal exposure to antimalarial medications. The vast majority of studies (92\%) were conducted on offspring of women with rheumatic diseases, mainly systemic lupus erythematosus. Of 337 children born to mothers with rheumatic conditions, 319 (nearly 95 \%) were exposed to HCQ which was continued throughout the entire pregnancy in most cases. The doses of HCQ used were reasonably similar and consistent with traditional doses for the SLE, between 200- 600 mg/day, although a few studies did not report dosage information. \textsuperscript{99,105} Hence, reasonable degree of similarity among the treatment groups and homogeneity in drug exposure across studies can be assumed.

In contrast, variations in the methods used for visual function assessment represent a significant limitation of the current review. Clinical assessment of visual function performed in 5 out of 12 studies was poorly described, making it impossible to assess validity and reliability. \textsuperscript{97-101} Only a small number of exposed infants received a comprehensive ophthalmological evaluation that
included various combinations of the following tests: inspection of anterior/posterior segment, visual acuity testing, color vision, visual fields assessment, and fundoscopic appearance.\textsuperscript{95, 96, 103, 105} Nevertheless, normal findings reported in all children assessed provide a reasonable degree of reassurance.

Given the difficulties in assessing visual function in young children who cannot cooperate or communicate, electrophysiological testing such as electroretinogram (ERG) and visual evoked potential (VEP) recording may be particularly useful for this age group. It has also been suggested that multifocal ERG is more sensitive in detection of early subclinical retinal changes following long-term exposure to antimalarials in adults and is now recommended as one of the objective screening tests.\textsuperscript{90} Importantly, standard protocols for the technical procedures and reporting of ERG have been proposed by the International Society for Clinical Electrophysiology of Vision (ISCEV) to allow comparability of test results.\textsuperscript{107} However, it has been demonstrated that ERG responses mature at different rates in early infancy\textsuperscript{108} and therefore must be ideally compared to the values of the healthy subjects of the same age.\textsuperscript{107} In this review, ERG testing was conducted on a small number of infants, which is not surprising owing to the labor and resource-intensive nature of the testing.

Two small series\textsuperscript{102, 103} have not provided sufficient details on the testing technique and expected normal values for comparison making it difficult to interpret the reported normal ERG results. In contrast, the study by Renault et al.\textsuperscript{104} stated explicitly the method and normative data used, although those seem to deviate from the ISCEV standards. The study demonstrated neurophysiological visual abnormalities in more than 28% of assessed infants. The clinical significance of these findings, however, remains unclear and requires long-term continuing assessments. Future studies with standard protocols are warranted to corroborate these results.
A limitation of the present review is a generally low quality of included studies with the vast majority of them being observational. It is important though to bear in mind the following considerations:

The incidence of the true retinal toxicity in adults following long term treatment with antimalarial medications are low: 2.5% for CQ and 0.1% for HCQ. Hence, it is very unlikely that RCTs of a sufficiently large sample size and adequate length of follow up will be ever conducted in infants to answer the specific research question we attempted to address in this review. The fact that only 2 small RCTs were available for inclusion in this review confirms the claim that observational studies are the main source of evidence to rely on especially for uncommon medical conditions such as SLE or rare adverse effects. Furthermore, observational data gathered in a real clinical setting are more likely to answer relevant clinical problems encountered in daily practice. This may also enhance the external validity and feasibility of future research. Thus, it has been argued that observational studies are a valuable and critical source of data, especially on drug safety, and therefore should complement randomized controlled trials to enable physicians and patients informed decision-making.

The observational studies included in this review were graded at low quality evidence due to high risk of bias. For instance, criteria for selection of participants exposed to antimalarials were not specified and were based on medication use pattern or physician/patient personal preferences. No information on potential confounders and methods to control for them was reported. Thus, selection bias is of particular concern. Furthermore, none of the included observational studies utilized blinded assessment of the visual function. Although detection bias is unlikely to play a role due to virtually no cases of retinopathy detected, it might be a case in Renault et al study. Finally, heterogeneity in study designs and outcome measures,
dissimilarity in comparison groups or the lack of thereof in a great proportion of included reports,\textsuperscript{96,100-103} “zero” events of interest reported in both treatment and control groups made impossible to apply any statistical methods to estimate the effect in separate studies and undertake formal meta-analysis thereafter.

In conclusion, the current evidence from small and relatively low quality studies suggests negligible to no fetal ocular toxicity of antimalarial medications during pregnancy. The clinical significance of early electroretinogram anomalies reported in a small subset of infants remains to be established. Larger follow up studies are warranted to rule out low risk of ocular toxicity in children exposed \textit{in utero} to antimalarial medications.

As mentioned above, RCTs might be impractical to evaluate rare clinical outcomes such as ocular toxicity following in utero exposure to antimalarials. Well designed prospective cohort studies, hence, is a more feasible solution. Case-control studies, another type of observational research, might be even more suitable for clinical situations of rare but serious adverse outcomes. Utilizing large administrative and clinical databases or disease-specific registries to conduct such case-control studies will allow increasing sample size and appears to be a promising innovative approach to ensure drug safety in pediatric population.
Chapter 5

General discussion

Pediatric drug safety is a major public health concern and an unmet need for high quality research. International initiatives have been launched recently to promote knowledge generation and translation and improve pediatric safety of new and old drugs. RCTs remain the main tool used by the research community to evaluate effects of various treatments and hence, provide the strongest evidence of intervention effects. Therefore, RCTs are regarded as the cornerstone of EBM. Still, a significant proportion of medications for children has not been studied in RCTs and thus, is not labeled for use in the pediatric population. However, because these drugs are widely used, it is critical to collect all existing evidence, by conducting SRs, and especially for drugs being in use for years, to provide the best possible understanding of their safety and effectiveness despite lacking the gold standard of RCTs.

While thorough evaluation of clinical evidence is advocated, there is no single formal system to discriminate high quality evidence from low one and convey this information effectively to physicians to assist with decision-making. The Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach appears to be a promising methodology to fill the gap between existing knowledge (interpreting research evidence available) and its application to practice. However, the use of GRADE in the field of pediatric pharmacotherapy is very limited. Hence, the work described above aimed to test the hypothesis that the GRADE is a useful and user-friendly methodological tool, applicable to both RCTs and non-randomized studies, to assess quality of evidence in pediatric pharmacotherapy where the high quality research is scant and likely unattainable.
As the first step, two clinical examples encountered in my practice were chosen to formulate specific research questions and demonstrate the applicability of the GRADE.

Both examples, domperidone for lactation and ocular toxicity after exposure to antimalarials during pregnancy, represent off-label use of medications which is a common situation in pediatric pharmacotherapy. Additionally, unique routes of exposure, such as through maternal prenatal drug use in case of antimalarials, or via breast milk in domperidone case, are specific to the pediatric population and may increase the risk of undesirable side effects in exposed children. In the case of prenatal exposure to antimalarials, the precise mechanism of retinal toxicity caused by these drugs is now well understood but is believed to be due to binding CQ and HCQ to melanin in the retina. It is known that fetal retinal development starts as early as 5-6 weeks of gestational age. However, HCQ binds avidly to fetal tissues and has long half life of 6-8 weeks. This makes difficult to estimate the risk of fetal ocular toxicity in relation to exposure to antimalarials during certain periods of pregnancy. Similarly, it is impossible to comment on the length of follow up needed to rule out possible ocular toxicity in offspring of exposed women, mainly due to rarity of the condition and difficulties in assessing visual function of infants and young children. It is recommended to perform comprehensive eye assessment in exposed children, including visual fields and color vision testing which require a cooperative child (usually after age of 3). Thus, someone can speculate that follow up of exposed children may be as long as few years of age to allow for thorough eye examination. However, direct evidence to support this notion is lacking. Current screening guidelines for adults recommend to perform a baseline eye examination upon initiation treatment with antimalarials and annual screening after 5 years of exposure to antimalarials (or earlier if unusual risk factors are present). Again, it is unclear how to extrapolate these recommendations to children.
The case of domperidone highlights another issue: regulatory actions when the drug has been labeled with a special warning (e.g. the case of domperidone for enhancing breast milk production) adds to the complexity of clinical decision making. In fact, several physicians approached by the author with safety information on domperidone, still felt uncomfortable prescribing domperidone to enhance lactation despite the existing safety data. This may indicate that effective knowledge translation of existing evidence into clinical practice remains a challenge.

Finally, growing public awareness toward drug safety and concerns regarding taking medications during pregnancy and lactation has made these examples challenging clinical dilemmas with an urgent need to resolve them in a timely and systematic fashion. Altogether, two selected examples are among a large group of commonly encountered situations in the field of pediatric pharmacotherapy, are clinically relevant and hence are likely to be of value to both physicians and patients.
5.1 Step-by-step application of the GRADE principles

Formulating health care question and performing SR of relevant literature

In two selected examples, domperidone for lactation and ocular toxicity following exposure to antimalarials during pregnancy, the PICO format was followed to formulate a research question, identify clinical outcome of interest and collect evidence for conducting SRs as described in detail in chapters 3 and 4.

In the example of domperidone, I decided a priori to focus on high quality evidence and therefore set up inclusion criteria limited to RCTs. Based on empirical evidence of using domperidone for years, I expected to have sufficient number of studies to answer my research question.

In the example of antimalarials, I decided to include RCTs and observational studies mainly due to rarity of the outcome of interest (ocular toxicity after prenatal exposure); this also allowed me to test whether the GRADE framework is applicable for assessing quality of observational research which have not been reported before in pediatric research.

Generation of estimate of effect for each outcome of interest

In the case of domperidone, meta-analysis was performed and the best estimate of the effect of domperidone on insufficient lactation has been generated. The methodological challenge encountered in this meta-analysis was related to missing standard deviations (SDs) for the
primary effect measure chosen for the review. The missing variance estimates were imputed as described above.\textsuperscript{111}

In the case of ocular toxicity after in utero exposure to antimalarial medications, it was impossible to perform formal meta-analysis and generate an estimate of the effect owing to the limitations in the available evidence, which precluded combination of data.\textsuperscript{112}

\textbf{Grading the quality of evidence for each outcome of interest across studies and rating overall quality of evidence}

The next step of the GRADE approach is to rate the quality of evidence for each outcome, across the studies, based on explicit criteria which include study design, risk of bias, imprecision, inconsistency, indirectness, and magnitude of effect as published elsewhere.\textsuperscript{46} The end product of the GRADE process is a summary of the evidence (usually quality rating and estimate of effect) presented as the GRADE evidence profile (EP) table. EP tables must include explicit judgments on each factor affecting quality of evidence.

Table 4 illustrates a summary of evidence/EP for the effect of domperidone for insufficient lactation in puerperal women. To judge the risk of bias (or studies limitations), I have followed criteria of the Cochrane Collaboration tool (sequence generation, allocation concealment, blinding, incomplete outcome data, selective reporting, etc).\textsuperscript{10} One study out of 3 was downgraded to moderate quality owing to limitations in study design and implementation.

However, the judgment on imprecision was more difficult to make. I attempted to follow the GRADE guidelines on imprecision as published elsewhere.\textsuperscript{113} In brief, the 95% CI around the estimate of treatment effect is often used as a measure of precision and should be sufficiently narrow. However, 95% CI may become misleading if generated from small sample size trials
with relatively few events. While overall, three included studies are of small sample sizes, they are not similar in this regard. Petraglia et al. provided no details on sample size calculation and recruitment details and thus its quality was downgraded due to potential risk of bias. To the contrary, the study by Da Silva et al. has justified clearly the sample size of 20 participants to be sufficient to demonstrate an increase at least of 25% in milk production from baseline in domperidone group compare to placebo (at power 80% and α of 0.05). While the authors have met their enrollment target, four participants withdrew from the study due to incomplete records (n=3) or death of the infant (n=1) with remaining 16 contributing to the study results. Finally, Campbell-Yeo et al. calculated 22 participants per group as a sample size needed and were able to recruit sufficient number of participants with one drop out in domperidone group prior to treatment initiation. Of note, all studies demonstrated statistically significant difference in the outcome of interest in treatment groups versus placebo. Nonetheless, to deal with the case of small sample size trials, the GRADE authors suggest calculating the optimal information size (OIS), in addition to 95% CI, as a measure of adequate precision. The OIS is usually estimated as a number of patients required for a single adequately powered RCT. According to the GRADE developers, if the both criteria have met, namely the 95% CI excludes no effect and total sample size/number of events exceeds the OIS, precision is judged as adequate.

In the meta-analysis on domperidone for insufficient lactation, the 95% CI around the estimate of treatment effect is reasonably narrow and excludes no effect as documented in table 1. However, the total number of patients in all included studies was only 78. Hence, the OIS was calculated.

To determine OIS, I set α of 0.05, β of 0.2, and assumed treatment difference in means breast milk production between domperidone and placebo of 50 mL/day as a meaningful difference.
This is in line with the reported treatment effect of domperidone in Da Silva et al study.\textsuperscript{58} Additionally, for very premature infants such as those enrolled in Da Silva study (mean gestational age of 29 weeks) and Campbell-Yeo et al study\textsuperscript{70} (mean gestational age 27.8 weeks, mean birth weight 1180.3 g), the 50 ml difference in the daily milk supply might be quite clinically significant assuming daily breast milk requirements of 150 ml/kg and generally low birth weight as reported above.

Choosing an appropriate SD, however, was complicated due to the fact that SDs reported in literature varied significantly. Therefore, I performed repeated calculations of the OIS using SDs of 30mL, 60 mL, and 90 mL, based on the ranges reported in literature. Applying the basic equation for sample size calculation of continuous variables available on-line elsewhere, the number of participants required in each group was 6, 23, and 53, respectively, with estimated OIS varied substantially from as few as 12 to as many as 106 participants. Given a total number of 78 participants included in the present meta-analysis, it appears that imprecision is just borderline in this case. If more conservative approach was used, someone might have concluded that the OIS criterion has not been met, and thus would consider rating the analysis down for imprecision. In contrast, had someone decided to use more liberal SD estimate, the OIS threshold would have easily met. Thus, it is a matter of reviewers’ judgment to decide whether the OIS criterion has been met (or failed) in this particular example, and whether the overall quality of evidence for domperidone for insufficient lactation should be downgraded for imprecision. This situation highlights a particular challenge encountered when judging small studies. Nevertheless, the strength of the GRADE lies in the fact that it allows for the explicit judgments to be made and communicated to readers. Coming to the final decision on the quality of evidence for domperidone for insufficient lactation, I have decided to downgrade the overall quality of evidence from high to moderate owing to the risk of bias and imprecision.
Table 4- GRADE evidence profile: effect of domperidone on insufficient lactation in puerperal women

<table>
<thead>
<tr>
<th>Quality assessment</th>
<th>Summary of findings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome:</strong> relative (%) increase in milk volume with domperidone treatment</td>
<td>Number of patients</td>
</tr>
<tr>
<td><strong>No of studies (Design)</strong></td>
<td>Limitations</td>
</tr>
<tr>
<td>3 (RCT)</td>
<td>Serious limitations (^a)</td>
</tr>
</tbody>
</table>

a. Inadequate description of randomization and concealment of allocation, inadequate information on blinding or incomplete accounting of patients

b. Small sample size

*Abbreviation:* RCT, randomized controlled trial
Table 5 is a summary of evidence/EP for ocular toxicity in children exposed in utero to antimalarial drugs. As evident from the table, the quality of available evidence is low for RCTs and very low for observational studies due to risk of bias and serious imprecision. Interestingly, in our original paper, we have judged two RCTs included as of moderate quality based on uncertain risk of bias. The methodological issue this SR underscores is related to handling sparse event data which is not uncommon for rare medical conditions such as rheumatic diseases in pregnancy or outcomes of concern, namely ocular toxicity followed exposure to antimalarial medications. It is impossible to calculate the risk ratio or the odds ratio if there are no events in the study groups. In such cases, it is often recommended to exclude studies with zero events from meta-analysis as contributing no information on relative probability of the event to pooled estimate.

However, in our case of ocular toxicity of antimalarials, the only data available come from studies where no events occurred. Thus, precise estimation of the effect and drawing inferences seems to be particularly problematic. Though a number of meta-analytical techniques has been proposed for effect measures for rare outcomes, I decided to apply a rather simple and quick formula, known as a rule of three, to estimate the maximum risk of ocular toxicity associated with exposure to antimalarials. In brief, Hanley and Lippman-Hand \(^{118}\) demonstrated years ago that for a sample of given size N, if no actual events occur, a 95% CI for the probability of this event lies between zero and \(3/N\). Furthermore, as stated in the Cochrane Handbook for Systematic Reviews, \(^{10}\) it appears appropriate to apply this rule for SRs, taking N as a sum of sample sizes of all arms receiving intervention from randomized, non-randomized studies and case series. As such, I have applied this rule to my example of SR on ocular toxicity following in utero exposure to antimalarials. Given zero cases of ocular toxicity reported in RCTs and cohort studies with a total number of 489 children exposed to antimalarials, I have easily
constructed an upper limit of the 95% CI as $3/489=0.0061$. In other words, we can be 95% confident that the risk of ocular toxicity in the offspring of women exposed during pregnancy to antimalarials does not exceed 0.61%. Of note, this 95% CI range (0.00-0.61%) is comparable to the estimated incidence of the true retinal toxicity in adults following long-term treatment with antimalarials: 0.1% for HCQ and 2.5% for CQ. It is important to bear in mind, though, that zero event rate reported in literature does not necessarily imply no risk. Thus, translating the results of such studies into the range of risks (via estimated 95% CI) might be particularly helpful while holding discussion with patients and assisting them in making informed decisions regarding their treatments.
Table 5-Quality assessment for ocular toxicity in children exposed in utero to antimalarial drugs

<table>
<thead>
<tr>
<th>Outcome: Ocular toxicity detected at any follow up time point</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of studies</td>
</tr>
<tr>
<td>(design)</td>
</tr>
<tr>
<td>2 (RCT)</td>
</tr>
<tr>
<td>3 (Cohort)</td>
</tr>
<tr>
<td>7 (Case series)</td>
</tr>
</tbody>
</table>

- a. Inadequate description of randomization and concealment of allocation or inadequate information on blinding and incomplete accounting of patients
- b. ZERO events and small sample size
- c. Failure to develop and apply appropriate eligibility criteria, flawed measurement of outcome
- d. No comparison groups
- e. Small sample size

**Abbreviation:** RCT, randomized controlled trial
Decision on the strength of recommendation

The final step of the GRADE algorithm is to decide on the strength of recommendations, based on the balance between desirable and undesirable effects, quality of evidence, cost, and patient’s preferences. The GRADE recommendations are ranked as weak or strong. Strong recommendation refers to the clinical situation while the majority of informed patients would choose the proposed treatment whereas weak recommendation means that patient’s individual decisions are likely to vary according to personal values. Of note, grading the strength of recommendations is a step specific for guidelines development and hence, lies beyond the scope of conducting systematic review of the relevant scientific literature. The panel of guideline developers (but not systematic reviewers) makes a decision regarding overall quality of evidence and direction and strength of recommendations. Nevertheless, due to clinical relevance of the research questions raised in this project and pressing need to address concerns of my patients, I decided to exercise my judgment and implement the GRADE to come up with recommendations for two examples chosen. The first and likely the major determinant for grading recommendations appears to be the safety of proposed treatment. In the SR on domperidone for insufficient lactation, there have been no adverse affects reported in exposed infants and mothers. One woman, exposed to placebo, reported mild abdominal cramping. Nevertheless, the recent Heath Canada advisory on domperidone safety has raised concerns within medical community. Health Canada has quoted two recently published studies which reported association between domperidone use and cardiac arrest and serious ventricular arrhythmia. It appears that people of older age (> 60 years of age), with pre-existing heart conditions (prolonged QT interval or congestive heart failure) or on higher domperidone doses (greater than 30 mg per day) might be at risk. Furthermore, there have been reports of similar
association between oral domperidone administration and QT interval prolongation in infants treated for gastro-esophageal reflux.\textsuperscript{66,117} However, breastfeeding represents a quite different situation with much less exposure of breastfed infants to domperidone during lactation as opposed to oral route of exposure. Limited pharmacokinetic data suggests a minimal systemic exposure of breastfed infants to the drug via breast milk.\textsuperscript{62} Even though Health Canada warning is unlikely to apply to healthy breastfeeding women and clinical and pharmacokinetic data on neonatal exposure to domperidone via breast milk is re-assuring, this advisory may influence individual decision making of female patients and prescribing attitudes of physicians. Therefore, despite relatively favorable safety profile of domeridone for enhancing lactation and reasonable cost of the drug, I believe that current moderate quality evidence warrants a weak recommendation as alternative strategies with virtually no undesirable adverse effects are available (nonpharmacological interventions to enhance lactation, switching to bottle feeding, etc.). Furthermore, some females may be reluctant to use medications while nursing in general, and/or owing to potential negative effect of recent public warning on domperidone, they may choose to avoid this particular medication while breastfeeding. Thus, patients’ variability in values and preferences are expected and supports further a weak recommendation for domperidone as a breast milk enhancer.

Finally, grading the strength of recommendation for the SR of antimalarials is not straightforward: the benefits of treatment with antimalarials during pregnancy versus alternative strategies (no treatment for serious chronic medical conditions such as rheumatic diseases) seem clearly outweigh potential low risk of ocular toxicity in the offspring and therefore favors a strong recommendation. However, overall low quality of evidence available warrants weak recommendation. Thus, individual patients’ values and preferences are likely to be a major
determinant in the rating the strength of recommendations which makes a weak
recommendation more appropriate.

5.2 Comparison of the GRADE and the Cochrane Collaboration Risk
of Bias Tool (CCRBT)

As a final step to test my hypothesis, I decided to apply a different assessment tool, the CCRBT,
to evaluate quality of evidence for two research questions of interest and compare conclusions
reached by the GRADE and the CCRBT. Lack of the “gold standard ‘assessment tool and
advantages of the CCRBT instrument, as described in the Introduction, have guided my choice.
In the case of domperidone, the judgments concerning risk of bias across selected studies are
summarized in Chapter 3, Table 2, page 43. The overall summary of risk of bias within and
across studies included in the meta-analysis is presented in Table 6.
Table 6- Effect of domperidone on insufficient lactation in puerperal women- assessment of the risk of bias within and across studies utilizing the CCRBT

<table>
<thead>
<tr>
<th>Studies, year</th>
<th>Risk of bias within a study</th>
<th>Risk of bias within a study-interpretation</th>
<th>Risk of bias across studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Petraglia et al 69 1985</td>
<td>Unclear risk of bias for key domains</td>
<td>Some doubts about the results</td>
<td>Unclear risk of bias as most information in meta-analysis is from a study at unclear risk of bias</td>
</tr>
<tr>
<td>Da Silva et al 58 2001</td>
<td>Low risk of bias for key domains</td>
<td>Plausible bias unlikely to seriously alter the results</td>
<td></td>
</tr>
<tr>
<td>Campbell-Yeo et al 70 2010</td>
<td>Low risk of bias for key domains</td>
<td>Plausible bias unlikely to seriously alter the results</td>
<td></td>
</tr>
</tbody>
</table>

Overall, I judged the quality of studies included in domperidone meta-analysis as having unclear risk of bias due to relatively heavy weight and, thus, contribution of the Petraglia et al. study 69 to the results of meta-analysis. Furthermore, owing to key quality domains affected in this particular report (sequence generation, allocation concealment and blinding), I decided to downgrade overall quality of studies included from high to moderate level.

Compare to the CCRBT, the GRADE tool categorizes risk of bias (called study limitations in the GRADE) slightly different but conceptually very similar, by assessing the same key domains. The three categories used by the GRADE are -” no serious limitations”, “serious limitations” or “very serious limitations”. As evident from Table 4, I judged domperidone studies as having serious limitations with subsequent judgment to downgrade to moderate quality overall, as per the GRADE framework.
To conclude, the example of domperidone meta-analysis has demonstrated a similar conclusion on risk of bias while applying the GRADE and the CCRBT tools to evaluate quality of evidence. However, the GRADE takes a few steps further and incorporates more factors which may affect quality of evidence (imprecision, inconsistency, indirectness of evidence, publication bias, etc.) which are shown in Table 4. In this particular case of domperidone, my final judgment remains the same- moderate quality of evidence for effectiveness domperidone as galactagogue.

In the case of antimalarials, I applied the CCRBT tool to evaluate risk of bias in two RCTs included in my second SR and details are summarized in Table 7.

Table 8 represents a next step of the CCRBT application- the overall summary of judgments on risk of bias within and across studies reported. I have concluded that evidence from these two RCTs has unclear risk of bias and therefore should be downgraded to moderate quality category. This corresponds to my judgment on quality of RCTs included in the SR of antimalarials while I applied the GRADE- serious limitations and decision to rate down the overall quality of evidence from high to moderate quality. However, while applying rigorously all domains of quality as per the GRADE (beyond the risk of bias), I have downgraded the available RCTs further to low quality due to serious imprecision (zero events rate and small sample sizes) as evident from Table 5. Furthermore, I was unable to apply the CCRBT tool to assess the majority of studies included in the SR of antimalarials due to the fact that this instrument is not applicable to observational research.

To conclude, the example of antimalarials has demonstrated that while the CCRBT tool corroborates conclusions made by the GRADE related to the risk of bias, the GRADE has advantage of more thorough assessment and broader applicability. Furthermore, the GRADE
changed (lowered) overall quality assessment of evidence available and thus has a potential to impact recommendations and decisions on a specific topic of interest.

Table 7–Ocular toxicity in children exposed in utero to antimalarial drugs-risk of bias assessment utilizing CCRBT

<table>
<thead>
<tr>
<th>Studies, year</th>
<th>Sequence generation</th>
<th>Allocation concealment</th>
<th>Blinding of participants and personnel</th>
<th>Blinding of outcome assessment</th>
<th>Incomplete outcome data</th>
<th>Selective outcome reporting</th>
<th>Other bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levy et al\textsuperscript{106} 2001</td>
<td>Unclear risk of bias/no adequate info provided</td>
<td>Unclear risk of bias/no adequate info provided</td>
<td>Low risk of bias/double blind study</td>
<td>Low risk of bias/clearly stated</td>
<td>Low risk of bias/no missing data</td>
<td>Low risk of bias/all outcomes reported</td>
<td>Low risk of bias</td>
</tr>
<tr>
<td>Villegas et al\textsuperscript{106} 2007</td>
<td>Low risk of bias/described in detail</td>
<td>Low risk of bias/described in detail</td>
<td>Low risk of bias/double blind study</td>
<td>Unclear risk of bias/no info if outcome assessors were blind</td>
<td>Unclear risk of bias/no reason provided for &gt; 25% missing outcome data</td>
<td>Low risk of bias/all outcomes reported</td>
<td>Low risk of bias</td>
</tr>
</tbody>
</table>
Table 8-Ocular toxicity in children exposed in utero to antimalarial drugs - assessment of risk of bias within and across studies utilizing CCRBT

<table>
<thead>
<tr>
<th>Studies, year</th>
<th>Risk of bias within a study</th>
<th>Risk of bias within a study-interpretation</th>
<th>Risk of bias across studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levy et al105 2001</td>
<td>Unclear risk of bias for key domains</td>
<td>Some doubts about the results</td>
<td>Unclear risk of bias as most information is from studies at unclear risk of bias</td>
</tr>
<tr>
<td>Villegas et al106 2007</td>
<td>Unclear risk of bias for key domains</td>
<td>Some doubts about the results</td>
<td></td>
</tr>
</tbody>
</table>

A number of limitations highlighted by this work needs to be acknowledged. Presently, use of the GRADE in the field of pediatric pharmacotherapy is rather limited and therefore future work may unveil potential problems, conceptual or practical, with its implementation. The current project, for instance, demonstrated that imprecision was a methodological challenge for both SRs. More challenges may arise with the expansion of the GRADE application.

Furthermore, the GRADE does not offer straightforward answers to some of the questions of interest: a substantial amount of judgment is still involved while interpreting evidence. Disagreements, therefore, are expected. Nevertheless, transparent reporting with the GRADE is felt to be helpful to arrive at individual clinical recommendations. Moreover, this transparency becomes of crucial importance while comparing different SRs, updating existing reviews or using them for subsequent policy making.

Finally, the present work is limited to the GRADE application to the field of individual treatment decisions and has not been tested for diagnostic accuracy and recommendations, prognosis or preventive interventions, or CPG development. Further experience with the
GRADE use in the field of pediatric pharmacotherapy is definitely needed to advance knowledge and bridge the gap between science and clinical practice.
Chapter 6

Conclusion

Our findings indicate that the GRADE is a relatively simple framework which was easy to follow while appraising evidence on selected topics in pediatric pharmacotherapy. Detailed step-by-step implementation of the GRADE algorithm for conducting two SRs as described above confirms its feasibility and applicability to the various topics in the field. A few methodological problems have been encountered during this field testing of the GRADE- small sample size of included reports and sparse event data which made a judgment on imprecision domain challenging. Nevertheless, the GRADE was doing reasonably well in clinical situations of rare medical conditions or outcome of interests, when the number of trials is small and effect estimates are imprecise. Furthermore, both RCTs and observational studies are amenable for evaluation which suggests that the GRADE may serve as a uniform instrument for grading evidence which will enhance its practical use. Comparison of the GRADE to the CCRBT, an alternative tool for assessment of risk of bias recommended by Cochrane Collaboration, has demonstrated that GRADE has broader applicability and may generate different decisions on the overall quality assessment. While this observation requires further confirmation by testing for agreement between the GRADE and alternative instruments, this may substantiate the value of the GRADE as an instrument with a potential to impact conclusions and clinical decision making. The reporting format of the GRADE offers clarity and explicit judgments which enable transparency and reproducibility. Finally, it is probably the first grading system which takes into account personal values and preferences. Overall, the present work fully supports the hypothesis that the GRADE is feasible and valuable methodological instrument to use for the assessment of quality of evidence in the field of pediatric pharmacotherapy when only scant data
available. The GRADE offers an advantage of rigorous methodology combined with relative simplicity.
Future directions

Future research should focus on wider implementation of the GRADE to various types of research questions encountered in the pediatric field, including diagnostic accuracy, preventive interventions, cost-effectiveness, etc, and for different consumers, such as guidelines developers, health policy makers, individual clinical decision makers, etc. Expanding the use of GRADE will contribute to further development of this dynamic framework and confirm its status as a proposed universal appraisal tool. Furthermore, as subsequent studies and reviews utilizing the GRADE become available, formal validation studies are likely to be undertaken to assess the validity of the GRADE, its reliability and perceived practicality. At present, there are very limited published data on validity testing of this methodological tool. In fact, a brief literature search of MEDLINE and EMBASE performed recently, has identified only one report attempting to apply the GRADE to evaluate quality of evidence from three systematic reviews on diverse clinical topics and assess inter-rater reliability of judgments made. Three SRs chosen were diverse in clinical scope, included RCTs and non-RCTs and various interventions (pharmacological and non-pharmacological) and outcomes: 1) steroids and bronchodilators for bronchiolitis, 2) operative and non-operative interventions for rotator cuff tears and 3) pain management for hip fracture. The quality of evidence was rated by two independent assessors across four main domains of the GRADE- risk of bias, consistency, directness and precision. Publication bias domain was not assessed. Inter-rater reliability was evaluated based on kappa statistics for each domain and for overall quality of evidence. The study has demonstrated variable reliability across the risk of bias domain, which ranged from moderate agreement 0.41 to slight, 0.06, with particular challenge of grading non-randomized trials. The precision domain
has created the most disagreements and uncertainty and has had persistently low reliability with kappa 0.18-0.21 which indicates only slight agreement. To the contrary, consistency and directness have had at least fair inter-rater reliability across reviews with some disagreements in cases when meta-analyses were not available.

Interestingly, these findings corroborate the observations made in the current project, with the similar challenges have been experienced while making judgments on precision domain and encountering the clinical situations while sparse research data, mainly from observational research, are available.

Undoubtedly, the judgments about evidence require a great deal of clinical and methodological expertise and are not easy to make; however; as more reports on practical use of the GRADE become available, more guidance and clarity on its application are likely to emerge. This would bring further evidence of the GRADE tool validity which is essential in bridging a gap between evidence synthesis and its application in facilitating personalized decision making.
References


OCEBM Table of Evidence Working Group = Jeremy Howick, Iain Chalmers (James Lind Library), Paul Glasziou, Trish Greenhalgh, Carl Heneghan, Alessandro Liberati, Ivan Moschetti, Bob Phillips, Hazel Thornton, Olive Goddard and Mary Hodgkinson


http://www.ahrq.gov/about/annualmtg08/091008slides/Bass.htm


Copyright Acknowledgements

Permissions are granted for the copyrighted materials used in this thesis and available upon request.