Can the Critical Appraisal Skills Programme check-lists (CASP) be used to Grading of Recommendations Assessment, Development, and Evaluation (GRADE) evidence?

Running title: CASP and GRADE

Edward PURSSELL RGN, RSCN, PhD

Senior Lecturer
School of Health Sciences
City, University of London
1 Myddleton Street
London EC1R 4QU

Acknowledgements: none

No conflict of interest has been declared by the author

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

Can the Critical Appraisal Skills Programme check-lists (CASP) be used alongside GRADE (Grading of Recommendations Assessment, Development, and Evaluation) to improve transparency and decision-making?

Abstract

Aims:

The Grading of Recommendations Assessment, Development, and Evaluation Working Group Guidance is widely used to increase the transparency by which evidence is turned into recommendations. However, although the process is clearly defined, it may be difficult to use in nursing education and practice because it uses separate terminology and tools to those sometimes used in education, such as those devised by the Critical Appraisal Skills Programme. This paper aims to show how these tools can be used together.

Design:

Discussion paper.

Data Sources:

Documentation from the Grading of Recommendations Assessment, Development, and Evaluation Working Group and the Critical Appraisal Skills Programme, as of 14th June 2019.

Implications for Nursing:

All of the items from the Critical Appraisal Skills Programme check-list can be incorporated into GRADE which might allow for wider use of its principles in nursing education and practice. Some additions are required however to complete the outcome-level assessment, these being the consistency of the results and possible publication bias. More detail on the extent to which the benefits are worth any harms and costs, and different types of inconsistency (heterogeneity) would

also be useful. This approach is consistent with the Group's Criteria for determining whether the GRADE approach was used.

Conclusion:

The Critical Appraisal Skills Programme tool can be used with minor modification to in a GRADE-like manner. This would allow for GRADE to be taught and used in nursing education and transferred to practice.

Impact:

- This discussion paper addressed the use of the Critical Appraisal Skills Programme Randomised Controlled Trial checklist to undertake GRADE-like assessments of evidence.
- With minor modifications to the way CASP is used, it is possible to use this tool to make GRADE-like assessments of the body of evidence as well as to critique individual studies.
- This finding will allow for the full use of the GRADE approach in healthcare education using the Critical Appraisal Skills Programme tools.

Keywords: CASP; Guideline; GRADE approach [MeSH]; Nursing practice; Nursing theory; Practice Guideline [MeSH], Randomized controlled trial [MeSH]

Introduction

The number of systematic reviews and clinical guidelines that are being written and published has increased greatly in recent years. Transparency in all aspects of research and clinical decision-making is universally recognised as being important, hence the development of a number of standardised study, review and guideline methodologies and reporting standards (Wager & Kleinert, 2013). While the use of guidance in clinical practice is explicitly not mandated, the British National Institute for Health and Care Excellence (NICE) for example stating that "it is not mandatory to apply [the] recommendations"(NICE, 2013); not following what is accepted as best practice through guidance can lead to criticism if negative outcomes result (Mellor, 2014). As important patient decisions and clinical judgements are made using them, the process by which guidelines are developed and recommendations made is therefore of utmost importance.

Background

It is argued by some that in order for nursing to be considered a profession, it must posses a unique perspective and have an identifiable and distinct body of knowledge that is researched, developed and advanced as it is passed on through the members of the nursing profession (Moulton, Wilson, Camargo Plazas, & Halverson, 2019). However, the development of such a body of knowledge in a discipline such as nursing which uses evidence from a wide variety of other professions such as medicine, psychology and sociology is not straightforward; and it may not the the knowledge that is unique, but rather the perspective with which that knowledge is applied (Mckenna, 1993). It has recently been suggested that the fundamental question for nursing research should be around how the well-being of people, families, communities or populations can be improved? (Moulton et al., 2019). The development of nursing research and the use of evidence by nurses to improve patient

outcomes is thought by many to have started with Florence Nightingale in the 1850s, and over time the nature and focus of nursing research has changed. More recently we have seen an increasing emphasis upon evidence-based practice and systematic reviews, perhaps best exemplified by the development of the Cochrane Collaboration in 1993, the Joanna Briggs Institute in 1996 and the Campbell Collaboration in 1999.

The lack of a distinct body of literature and the multiplicity of different methods that are used to examine questions that are of interest to nurses poses a significant problem, which is that it is impossible for researchers and practitioners to be expert in all of the methods used by researchers, even if they wished. Furthermore, the growth in the number of research papers and the variety forms in which they are published may make it difficult for nurses to keep abreast of new developments. Many appraisal tools and check-lists have been developed to help nurses understand the research they are reading (Buccheri & Sharifi, 2017) including those developed by some of the organisations above. Such diversity of tools is not restricted to quantitative research, a recent study of tools designed specifically appraise the quality of quantitative research found 102 different tools covering a total of 22 themes (Munthe-Kaas, Glenton, Booth, Noyes, & Lewin, 2019).

Despite the fact that historically hierarchies of evidence have emphasised randomised controlled trials as being the highest level of evidence (Burns, Rohrich, & Chung, 2011), application of evidence to practice and sound decision-making requires a systematic review of all of the literature in order to make informed judgements; thus in contemporary hierarchies the highest level of evidence for most clinical questions is not an individual study, but rather a systematic review of studies (Oxford Centre for Evidence-Based Medicine, 2016). Although qualitative research is different in many ways to quantitative methods, equivalent hierarchies do exist (Daly et al., 2007a),

and systematic review methodologies for qualitative and mixed-methods are also seen as being increasingly important (Barnett-Page & Thomas, 2009; Gough, 2015).

Although many clinical decisions are made on the basis of evidence from formal systematic reviews or guidance produced by national bodies and specialist groups, others are made on a more local basis or by those with fewer resources than large review or guidance producing organisations. While it is important that systematic reviews are undertaken using a standard methodology to ensure rigour, the number of tools and check-lists involved in this process has made this increasingly complex; these include reporting guidelines for studies, risk of bias tools, evidence summary tools and reporting guidelines for systematic reviews themselves (Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009; Whiting et al., 2016). Additionally the number and complexity of some of these tools appear to be increasing, with the Cochrane Risk of Bias Tool 2 for randomised controlled trials being more complex than the original tool; and a different tool being required for non-random studies for example (Cochrane Methods, 2019).

Although sometimes thought to be a single action, ‘critical appraisal’ is actually a linear process that is comprised of a number of related but distinct parts, these being appraisal of:

1. The quality of reporting for each study
2. The conduct of each study
3. The risk of bias of each study
4. The body of evidence; that is all of the studies together for each outcome. This includes the risk of bias and other factors.
5. The appraised evidence then needs to be transformed in to logical and transparent recommendations.

The first three of these are study-level decisions, while the fourth and fifth are outcome level ones reflecting decisions made across all of the studies that are contributing to that outcome. Transparency throughout this process is of the upmost importance so that the rationale for decisions can be understood. The importance of this can be seen in review and guideline methodologies; explicit inclusion of equalities statements; the increasing publication of original data; and the wider involvement of patients and other stakeholders.

The quality of reporting is quite distinct from the quality of the study, and so will not be considered here apart to note that there are numerous reporting tools which can be found on the Enhancing the QUAlity and Transparency Of health Research or EQUATOR Network webpage (http://www.equator-network.org/). In terms of appraising individual studies as opposed to their reporting there are two distinct concepts within this, that of the conduct of the study and the related but distinct idea of the risk of bias associated with the study. In the Cochrane Handbook it is the latter that is emphasised; that is "the extent to which the results can be believed" (Higgins & Green, 2011). The risk of bias needs to be assessed at both study and review-level, the latter being necessary as part of the process to assess the overall body of evidence used to come to a conclusion or make a recommendation. Other tools such as the Critical Appraisal Skills Programme (CASP) Tools are commonly used for more general assessment of the conduct or quality of a study.

The Critical Appraisal Skills Programme

The Critical Appraisal Skills Programme system is well known to many healthcare practitioners as it is widely used as a pedagogical tool to critique studies and its use is taught in many undergraduate and post-gradate programmes, and the tools within CASP are similar in nature to those produced by a number of other organisations (Buccheri & Sharifi, 2017). CASP comprises a number of check-lists, each of which contains a number of questions presented in a systematic way to guide readers

through the appraisal process, including steps 1 to 3 above, but not the outcome-level steps 4 and 5 (CASP, 2018). In addition to familiarity, it also has the advantage of having a number of different methodological check-lists, including those for different types of observational studies and qualitative research as well as RCTs; the similarity of approach and style enhancing their usability.

The Grading of Recommendations Assessment, Development, and Evaluation

The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) is a system that has been developed to be used to rate the quality of evidence in systematic reviews and guidelines, and then to grade the strength of recommendations made in guidelines in a transparent way (Guyatt et al., 2011). It has a number of features that differentiate it from other similar systems, in particular that it is outcome-centric rather than study-based and that it separates judgements about the confidence in estimates or the quality of the evidence from judgements about the strength of recommendations that come from that evidence. One rates the confidence in effect estimates, that is the quality of the evidence for each outcome, and uses this alongside other factors namely: the balance between desirable and undesirable outcomes; confidence in values and preferences and their variability; and resource use, to develop a nuanced recommendation that can be strong or weak and for or against a treatment option. GRADE therefore defines the quality of evidence as the "extent to which our confidence in an estimate of the effect is adequate to support a particular recommendation" and uses this with these other factors to facilitate transparent decision making (GRADE, 2013). Strong recommendations can therefore come from relatively weak evidence if factors such as the balance between desirable and undesirable consequences of the recommended action suggest this.

Although GRADE has been in existence for some time and has been widely adopted, this has primarily been used with quantitative outcomes. A more recent initiative is GRADE-CERQual (Confidence in the Evidence from Reviews of Qualitative research); which seeks to fulfil a similar role with qualitative research (Lewin et al., 2018) but which is at a much earlier stage of development than GRADE or CASP. Although hierarchies do exist for qualitative studies (Daly et al., 2007b), these are not universally recognised (JBI, 2019) and their use is somewhat less straightforward because qualitative study designs may not be so transparent or easy to identify (Buus & Agdal, 2013; McCrae & Purssell, 2016).

The main criticisms surround the lack of theoretical and/or empirical justification for the GRADE criteria and lack of clarity regarding how some elements should be used in practice (Mercuri & Gafni, 2018a, 2018b, 2018c). Although there has been copious words written on GRADE, in most cases it does still rely on consensus albeit from a large and growing user-base. The authors suggest that until a stronger theoretical base can be provided "enthusiasm for the framework should be tempered" (Mercuri & Gafni, 2018c) (p. 1233). While students should be encouraged to be critical of this and other similar frameworks, it is nonetheless one that they will come across and need to understand, the World Health Organization stating that for guideline development "while the ease of applying the GRADE approach will vary according to the type of evidence being assessed, the circumstances in which GRADE cannot be applied are rare"(World Health Organization, 2014) (p.121). At the very least GRADE forces those making decisions to consider and to be transparent regarding important factors in decision-making, and organisations are adapting it to their specific needs within the broader requirements of the GRADE Working Group (Thornton et al., 2013).

A criticism that can be explored with more advanced users is the theory that a Bayesian approach might be more appropriate, as this would allow one to incorporate prior knowledge and other

information into ones degree of belief in an outcome, the degree of belief being a Bayesian analogue to the confidence in the effect estimate. The authors give the example of a single-large RCT with a clear positive result, but then suggest that the prior knowledge that the results of many clinical studies are not reproducible might affect the degree of belief or confidence that one has in any given study result whatever the apparent methodological strength (Mercuri & Baigrie, 2018). This also reflects a wider debate about the relative merits of Bayesian versus more traditional frequentist approaches to evidence (Gelman, 2008).

Data sources

Papers from the GRADE Working Group and supporting information; most data was taken from the GRADE Handbook and CASP website, as of 14th June 2019.

Discussion

Using CASP

Use of CASP begins with identifying the type of study and choosing the correct checklist. As an example the CASP checklist for RCTs comprises of a number of study-type specific questions designed to help readers consider the three broad issues that need to be considered when appraising a trial which come under three broad headings:

1. "Are the results of the study valid?"
2. "What are the results?"
3. "Will the results help locally?"

Thus the tool leads the user through the process of appraisal, beginning with three screening questions before asking "is it worth continuing?" After this more detailed questions in each of the three categories are posed along with hints and areas for free-text comments. In the case of the RCT tool there are a total of 11 questions.

Although these are designed to be study-level decisions, they could also be used to make a judgement across studies that would apply at the outcome level. A comparison of the criteria used in the two systems are shown in Table 1; this shows that there are two GRADE criteria that are not reflected in CASP as they are specifically cross-study criteria, these being inconsistency of results and publication bias. However CASP also contains a question on "how large was the treatment effect?" This builds upon the GRADE strength of recommendation as adding an explicit statement about the strength of the effect enhances this judgement by being clear about how much benefit is likely to result (CASP, 2018).

Using GRADE

The GRADE process consists of a number of different steps:

1. Defining the question and identifying important and critical outcomes
2. Collecting evidence and generating an estimate of effect for each outcome
3. Rating quality of evidence for each outcome
4. Deciding on the direction and strength of each recommendation (Guyatt et al., 2011).

The use of GRADE begins with formulating the question and identifying the important outcomes, before identifying the studies that can answer the question. In studies of effectiveness randomized trials are seen to provide stronger evidence than non-randomised trials and observational studies, which in turn are evidentially stronger than uncontrolled studies such as case series and expert opinion. However, unlike traditional hierarchies which have a rigid pyramid structure with randomised studies at the top, followed by prospective and retrospective observational studies, and then case series and expert opinion at the bottom (with systematic reviews of each type normally at the top of each of these); GRADE incorporates flexibility about this, thus the evidence provided by

a randomised trial may be downgraded and that from an observational study upgraded (Murad, Asi, Alsawas, & Alahdab, 2016). Case-series which lack a control group are normally considered a low or very low level of evidence; while expert opinion is considered to be an interpretation of evidence rather than a direct form of evidence to be graded (GRADE, 2013).

Randomised study evidence, which starts at a high level can be downgraded on the basis of risk of bias, inconsistency of the results, indirectness of the evidence, imprecision or publication bias.

- Risk of bias is an assessment of possible "systematic error, or deviation from the truth, in results or inferences"(Higgins & Green, 2011), for which a number of different tools exist, including the questions within GRADE, those from the Cochrane Collaboration and elements of CASP. The GRADE tool consists of five items: lack of allocation concealment, lack of blinding, incomplete accounting of patients and outcome events, selective outcome reporting and other factors.
- Inconsistency of results refers to unexplained differences in study results, and is sometimes referred to as heterogeneity.
- Indirectness occurs if the interventions, populations or other important factors are not identical to those in which one is interested.
- Imprecision can be seen usually where sample sizes are small or the event of interest is very rare, leading to wide confidence intervals and uncertainty about the true effect.
- Publication bias occurs when studies with particular results, for example those with a positive finding compared to those with a negative finding have a different probability of being published (GRADE, 2013).

These elements of GRADE are mapped onto CASP in Table 1.

Observational studies on the other hand, which starts at a lower level, can be upgraded on the basis of: a large magnitude of effect; a dose-response gradient; or if all plausible residual confounding effects should reduce the demonstrated effect or increase the effect, if no effect was observed. There are also reasons that one might downgrade observational studies, these being: failure to develop and apply appropriate eligibility criteria; flawed measurement of both exposure and outcome; failure to adequately control confounding; and incomplete or inadequately short follow-up (GRADE, 2013).

The important thing about GRADE is that these judgements are for the body of evidence as a whole (stage 4 above) rather than individual studies, although assessment of factors such as the risk of bias has to be undertaken at both study (stage 3) and outcome (stage 4) levels. For this an appropriate tool must be used; GRADE provides one for RCTs but not other study-types. A recent development has been the Risk Of Bias In Non-randomised Studies - of Interventions (ROBINS-I) which is currently being incorporated into GRADE (Schünemann et al., 2018; Sterne et al., 2016).

Recommendations can then be made on the basis of this assessment of the strength of the evidence for each outcome, alongside a judgement about an assessment of the balance of desirable and undesirable consequences, the resource used, and values and preferences. Recommendations should also have a strength, which reflects the extent to which a one is confident that the desirable effects of an intervention outweigh undesirable effects. This constitutes stage 5. A strong recommendation means that most people would benefit from the recommended course of action; while a weak one that it is thought to be beneficial but not all people would benefit from it; in the case of a negative recommendation these are adjusted accordingly. GRADE also differentiates between guideline development and systematic reviewing with regards to making a recommendation, stating that the former but not the latter make decisions about what outcomes are

critical and the overall quality of evidence. The extent to which this is true is perhaps debatable (GRADE, 2013).

Can CASP be used to GRADE?

The GRADE working-group rightly argues against the confusion that might come from using different systems for grading quality of evidence and strength of recommendations, however CASP could be used to support GRADE, and for clinicians who have learned CASP it may be more straightforward and easier to understand. Furthermore its use would facilitate the use of the GRADE approach to students who might otherwise use informal or less transparent processes, and its transparency of decision-making may help to reduce the theory-practice gap (Purssell, 2019).

Logically to fit in with the systematic approach to critical appraisal and support for evidence outlined above it might be helpful for CASP to be reorientated for this purpose as shown in Table 1, with those items that are study-orientated and those which are outcome orientated differentiated. Although those dealing with the risk of bias are applicable at study and outcome levels, the nature of the assessment is different with the outcome-level assessment being made on the weakest study used in judging each outcome. Thus one might have a choice between more evidence but using weaker studies to support that, or a smaller number of stronger studies.

Another advantage of using CASP is that it allows for systematic reviewers to make recommendations which GRADE does not allow. This is because GRADE states that stakeholders are required to make valid judgements about the trade-offs between the desirable and undesirable consequences of a recommendation (Balshem et al., 2011). Despite this systematic reviewers often do make recommendations with or without stakeholder involvement, the CASP items: can the

results be applied to the local population, or in your context? and are the benefits worth the harms and costs? being appropriately measured responses to this lack of stakeholder involvement.

The main limitation of using CASP in this way is the lack of two important outcome-level criteria, those of the consistency of the results and possible publication bias. These are important and need to be considered, and their absence results from CASP being tools developed to appriase individual studies rather than a body of evidence. Additionally three GRADE criteria are are contained within the CASP question are the benefits worth the harms and costs? It might be useful to be more specific about these (balance of desirable and undesirable consequences, resource use, and values and preferences). It could be further improved by being clear about the different types of inconsistency or heterogeneity, these being: clinical differences and differences due to methodology, as well as simple differences in the results (Higgins & Green, 2011). This would further encourage formal consideration not just of differences in the results, but of the methodology as well.

The GRADE criteria for determining whether the GRADE approach was used is shown in Table 2; this demonstrating that use of CASP alongside a GRADE approach of assessing each study and then the overall body of evidence meets all of these criteria as long as two study-level criteria of inconsistency of results and publication bias are added. In order to make the different levels of analysis clearer a slight reordering of the tool for this purpose might be helpful as outlined in Table 1.

Implications for nursing

The advantage of this approach is that it allows the teaching of GRADE alongside CASP, introducing a widely used method to students and practitioners who might otherwise not be exposed to it. Furthermore if GRADE is seen as a principal as well as a method, this approach is consistent

with the principles underlying GRADE. Although GRADE has been criticised, it is widely used and the principle of transparent decision-making is important. Furthermore extensions of GRADE such as the Evidence to Decision Frameworks are useful both in decision-making and understanding the complexity of practice recommendations by making it clear that evidence and research are not the only factors to be considered (Alonso-Coello et al., 2016).

Conclusion

This paper has shown how CASP can be adapted and used in a GRADE-like way. This is important because GRADE is used by a large number of guideline development and other organisations, and introducing the approach to students and clinicians is important for them to understand how it works. Although GRADE can be seen as a methodology to follow, much as one might use any other; it can also be seen as an approach that provides principles to be followed. If one believes the former one approach would be to cease to use CASP and move over to GRADE entirely, however CASP remains popular and so it would seem premature to take such a step. Another approach however would be to see GRADE as an approach and to use CASP within this. A consensus approach by educators on this subject could refine this further, and perhaps a GRADE for education or GRADE-Ed Group for those interested in this area.

GRADE criteria	CASP RCT question	Comparison
Three categories of outcomes according to importance for decision-making: critical; important but not critical; of limited importance.	Did the trial address a clearly focused issue?	Study and outcome-level factor – CASP equivalent; possible add is it a critical, important or question of limited importance?
Study limitations/risk of bias Selective outcome reporting	Were all clinically important outcomes considered?	Study and outcome-level factor – CASP equivalent
Study limitations/risk of bias Lack of allocation concealment	Was the assignment of patients to treatments randomised?	Study and outcome-level factor – CASP equivalent. CASP asks explicitly about assessing randomisation
Study limitations/risk of bias Incomplete accounting of patients and outcome events	Were all of the patients who entered the trial properly accounted for at its conclusion?	Study and outcome-level factor – CASP equivalent
Study limitations/risk of bias Lack of blinding	Were patients, health workers and study personnel 'blind' to treatment? Aside from the experimental intervention, were the groups treated equally?	Study and outcome-level factor – CASP equivalent
Study limitations/risk of bias Other limitations	Were the groups similar at the start of the trial?	This is a study and outcome-level factor, CASP asks explicitly about the presence of randomisaton
Inconsistency of results		Outcome level factor – CASP has no direct equivalent
Imprecision	How precise was the estimate of the treatment effect?	Study and outcome-level factor – CASP equivalent
Indirectness of evidence	Can the results be applied to the local population, or in your context?	Study and outcome-level factor – CASP equivalent
Publication bias		Outcome level factor – CASP has no direct equivalent
Confidence in the magnitude of estimates of effect (overall quality of evidence for outcomes)		Study and outcome-level factor – CASP equivalent
Balance between desirable and	How large was the treatment	Outcome-level factor – CASP

undesirable outcomes (trade-offs)	effect? Are the benefits worth the harms and costs?	equivalent
Resource use	Are the benefits worth the harms and costs?	Outcome-level factor – CASP equivalent
Confidence in values and preferences and their variability	Are the benefits worth the harms and costs?	Outcome level factor – CASP equivalent

Table 1. Comparison of GRADE and CASP RCT criteria

Criteria for determining whether the GRADE approach was used	Application to this approach
Definition of quality of evidence should be defined consistently with the GRADE definitions. For guideline panels: reflects the extent to which our confidence in an estimate of the effect is adequate to support a particular recommendation. Systematic reviews: reflects the extent to which we are confident that an estimate of the effect is correct.	CASP contains items to assess confidence in study estimates
Consideration given to each of the GRADE criteria for assessing the quality of evidence: risk of bias; directness of evidence; consistency and precision of results; risk of publication bias; magnitude of the effect; dose-response gradient, and influence of residual plausible confounding	CASP does not currently consider consistency of results, publication bias or values and preferences which should be added for outcome-level assessment
Quality of evidence (confidence in the estimated effects) assessed for important outcomes and expressed using three (high, moderate, low) or four categories (high, moderate, low, very low)	This is independent of the tool used
Tables or detailed narrative summaries should be used as the basis for judgements about the quality of evidence and the strength of recommendations	This is independent of the tool used
Consideration given to each of the four criteria for determining the strength of a recommendation (the balance of desirable and undesirable consequences, quality of evidence, values and preferences of those affected, and resource use)	This is independent of the tool used
Strength of recommendations expressed using two categories (weak and strong)	This is independent of the tool used
Decisions about the strength of recommendations transparently reported	This is independent of the tool used

Table 2. Criteria for determining whether the GRADE approach was used

Acknowledgements: none

No conflict of interest has been declared by the author

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

Alonso-Coello, P., Schünemann, H. J., Moberg, J., Brignardello-Petersen, R., Akl, E. A., Davoli, M., … the GRADE Working Group. (2016). GRADE Evidence to Decision (EtD)

frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ*, i2016. https://doi.org/10.1136/bmj.i2016

Balshem, H., Helfand, M., Schünemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., … Guyatt, G. H. (2011). GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology*, *64*(4), 401–406. https://doi.org/10.1016/j.jclinepi.2010.07.015

Barnett-Page, E., & Thomas, J. (2009). Methods for the synthesis of qualitative research: a critical review. *BMC Medical Research Methodology*, *9*(1), 59. https://doi.org/10.1186/1471-2288-9-59

Buccheri, R. K., & Sharifi, C. (2017). Critical Appraisal Tools and Reporting Guidelines for Evidence-Based Practice. *Worldviews on Evidence-Based Nursing*, *14*(6), 463–472. https://doi.org/10.1111/wvn.12258

Burns, P. B., Rohrich, R. J., & Chung, K. C. (2011). The Levels of Evidence and Their Role in Evidence-Based Medicine: *Plastic and Reconstructive Surgery*, *128*(1), 305–310. https://doi.org/10.1097/PRS.0b013e318219c171

Buus, N., & Agdal, R. (2013). Can the use of reporting guidelines in peer-review damage the quality and contribution of qualitative health care research? *International Journal of Nursing Studies*, *50*(10), 1289–1291. https://doi.org/10.1016/j.ijnurstu.2013.02.012

CASP. (2018). CASP - Critical Appraisal Skills Programme. Retrieved 20 May 2019, from CASP - Critical Appraisal Skills Programme website: https://casp-uk.net/

Cochrane Methods. (2019). RoB 2: A revised Cochrane risk-of-bias tool for randomized trials. Retrieved 22 May 2019, from /bias/resources/rob-2-revised-cochrane-risk-bias-tool-randomized-trials

Daly, J., Willis, K., Small, R., Green, J., Welch, N., Kealy, M., & Hughes, E. (2007a). A hierarchy of evidence for assessing qualitative health research. *Journal of Clinical Epidemiology*, *60*(1), 43–49. https://doi.org/10.1016/j.jclinepi.2006.03.014

Daly, J., Willis, K., Small, R., Green, J., Welch, N., Kealy, M., & Hughes, E. (2007b). A hierarchy of evidence for assessing qualitative health research. *Journal of Clinical Epidemiology*, *60*(1), 43–49. https://doi.org/10.1016/j.jclinepi.2006.03.014

Gelman, A. (2008). Rejoinder. *Bayesian Analysis*, *3*(3), 467–477. https://doi.org/10.1214/08-BA318REJ

Gough, D. (2015). Qualitative and mixed methods in systematic reviews. *Systematic Reviews*, *4*(1), 181, s13643-015-0151-y. https://doi.org/10.1186/s13643-015-0151-y

GRADE. (2013). GRADE handbook. Retrieved 20 May 2019, from https://gdt.gradepro.org/app/handbook/handbook.html

Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., … deBeer, H. (2011). GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*, *64*(4), 383–394. https://doi.org/10.1016/j.jclinepi.2010.04.026

Higgins, J., & Green, S. (Eds.). (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. Retrieved from http://handbook-5-1.cochrane.org/

JBI. (2019). 2.3 Introduction to qualitative systematic reviews - JBI Reviewer's Manual - JBI GLOBAL WIKI. Retrieved 23 May 2019, from https://wiki.joannabriggs.org/display/MANUAL/2.3+Introduction+to+qualitative+systematic+reviews

Lewin, S., Booth, A., Glenton, C., Munthe-Kaas, H., Rashidian, A., Wainwright, M., … Noyes, J. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to the series. *Implementation Science*, *13*(S1), 2, s13012-017-0688–3. https://doi.org/10.1186/s13012-017-0688-3

McCrae, N., & Purssell, E. (2016). Is it really theoretical? A review of sampling in grounded theory studies in nursing journals. *Journal of Advanced Nursing*, *72*(10), 2284–2293. https://doi.org/10.1111/jan.12986

Mckenna, G. (1993). Unique theory--is it essential in the development of a science of nursing? *Nurse Education Today*, *13*(2), 121–127.

Mellor, J. (2014). *An avoidable death of a three-year-old child from sepsis*. London: Parliamentary and Health Service Ombudsman.

Mercuri, M., & Baigrie, B. S. (2018). What confidence should we have in GRADE? *Journal of Evaluation in Clinical Practice*, *24*(5), 1240–1246. https://doi.org/10.1111/jep.12993

Mercuri, M., & Gafni, A. (2018a). The evolution of GRADE (part 1): Is there a theoretical and/or empirical basis for the GRADE framework? *Journal of Evaluation in Clinical Practice, 24*(5), 1203–1210. https://doi.org/10.1111/jep.12998

Mercuri, M., & Gafni, A. (2018b). The evolution of GRADE (part 2): Still searching for a theoretical and/or empirical basis for the GRADE framework. *Journal of Evaluation in Clinical Practice, 24*(5), 1211–1222. https://doi.org/10.1111/jep.12997

Mercuri, M., & Gafni, A. (2018c). The evolution of GRADE (part 3): A framework built on science or faith? *Journal of Evaluation in Clinical Practice, 24*(5), 1223–1231. https://doi.org/10.1111/jep.13016

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moulton, E., Wilson, R., Camargo Plazas, P., & Halverson, K. (2019). The central question and the scope of nursing research. *Nursing Philosophy, 20*(1), e12228. https://doi.org/10.1111/nup.12228

Munthe-Kaas, H. M., Glenton, C., Booth, A., Noyes, J., & Lewin, S. (2019). Systematic mapping of existing tools to appraise methodological strengths and limitations of qualitative research: first stage in the development of the CAMELOT tool. *BMC Medical Research Methodology, 19*(1). https://doi.org/10.1186/s12874-019-0728-6

Murad, M. H., Asi, N., Alsawas, M., & Alahdab, F. (2016). New evidence pyramid. *Evidence Based Medicine, 21*(4), 125–127. https://doi.org/10.1136/ebmed-2016-110401

NICE. (2013). Overview | Fever in under 5s: assessment and initial management | Guidance | NICE. Retrieved 20 May 2019, from https://www.nice.org.uk/guidance/cg160

Oxford Centre for Evidence-Based Medicine. (2016, May 1). OCEBM Levels of Evidence. Retrieved 4 September 2019, from CEBM website: https://www.cebm.net/2016/05/ocebm-levels-of-evidence/

Purssell, E. (2019). Using GRADE to reduce the theory-practice gap. *Nurse Education Today, 74*, 82–84. https://doi.org/10.1016/j.nedt.2018.12.008

Schünemann, H. J., Cuello, C., Akl, E. A., Mustafa, R. A., Meerpohl, J. J., Thayer, K., … GRADE Working Group. (2018). GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2018.01.012

Sterne, J. A., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., … Higgins, J. P. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, i4919. https://doi.org/10.1136/bmj.i4919

Thornton, J., Alderson, P., Tan, T., Turner, C., Latchem, S., Shaw, E., … Chamberlain, K. (2013). Introducing GRADE across the NICE clinical guideline program. *Journal of Clinical Epidemiology*, *66*(2), 124–131. https://doi.org/10.1016/j.jclinepi.2011.12.007

Wager, E., & Kleinert, S. (2013). Why do we need international standards on responsible research publication for authors and editors? *Journal of Global Health*, *3*(2), 1–7. https://doi.org/10.7189/jogh.03.020301

Whiting, P., Savović, J., Higgins, J. P. T., Caldwell, D. M., Reeves, B. C., Shea, B., … Churchill, R. (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *Journal of Clinical Epidemiology*, *69*, 225–234. https://doi.org/10.1016/j.jclinepi.2015.06.005

World Health Organization. (2014). *WHO handbook for guideline development*. Retrieved from http://apps.who.int/iris/bitstream/10665/145714/1/9789241548960_eng.pdf