Resuscitation 95 (2015) e33–e41

Contents lists available at ScienceDirect

Resuscitation

journal homepage: www.elsevier.com/locate/resuscitation

Part 2: Evidence evaluation and management of conflicts of interest 2015 International Consensus on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science with Treatment Recommendations☆,☆☆

Peter T. Morley*, Eddy Lang, Richard Aickin, John E. Billi, Brian Eigel, Jose Maria Ferrer, Judith C. Finn, Lana M. Gent, Russell E. Griffin, Mary Fran Hazinski, Ian K. Maconochie, William H. Montgomery, Laurie J. Morrison, Vinay M. Nadkarni, Nikolaos I. Nikolaou, Jerry P. Nolan, Gavin D. Perkins, Michael R. Sayre, Andrew H. Travers, Jonathan Wyllie, David A. Zideman

ARTICLE INFO

Keywords:
Cardiac arrest
Conflict of interest
Evidence evaluation
Resuscitation

> Facts are stubborn things; and whatever may be our wishes, our inclinations, or the dictates of our passions, they cannot alter the state of facts and evidence.—John Adams, second President of the United States

Introduction

The international resuscitation community, under the guidance of the International Liaison Committee on Resuscitation (ILCOR), has continued its process to identify and summarize the published resuscitation science in the documents known as the ILCOR Consensus on Science with Treatment Recommendations (CoSTR). The accompanying articles represent the culmination of many years work, where a total of 250 evidence reviewers from 39 countries completed 165 systematic reviews on resuscitation related questions.

Process before 2015

The processes previously used by ILCOR in the development of their CoSTR were specifically tailored to the complex needs of resuscitation science. At the time that the evidence evaluation was undertaken for the 2010 publication, there were still no other processes which could deal with the complexity of literature that we need to evaluate: from randomized controlled trials to case series, and from mathematical models to animal studies. The 2010 evidence evaluation process required completion of an electronic worksheet,[1] that included a table, summarizing the evidence addressing individual questions. It included 3 options for the direction of support (supportive, neutral and opposing), 5 Levels of Evidence, and a quality assessment of the individual studies (good, fair or poor).[2,3]

Improvements for the 2015 process

When developing the process to be adopted for the 2015 CoSTR, ILCOR made a commitment to use the best available methodological tools to conduct its evaluation of the published resuscitation literature. To this end, ILCOR agreed to perform systematic reviews based on the recommendations of the Institute of Medicine of the

☆ The European Resuscitation Council requests that this document be cited as follows: Peter T. Morley, Eddy Lang, Richard Aickin, John E. Billi, Brian Eigel, Jose Maria E. Ferrer, Judith C. Finn, Lana M. Gent, Russell E. Griffin, Mary Fran Hazinski, Ian K. Maconochie, William H. Montgomery, Laurie J. Morrison, Vinay M. Nadkarni, Nikolaos I. Nikolaou, Jerry P. Nolan, Gavin D. Perkins, Michael R. Sayre, Andrew H. Travers, Jonathan Wyllie, David A. Zideman. Part 2: Evidence evaluation and management of conflicts of interest. 2015 International Consensus on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science with Treatment Recommendations. Resuscitation 2015;95:e33–e41.

☆☆ This article has been copublished in "Circulation".

* Corresponding author.
E-mail address: peter.morley@mh.org.au (P.T. Morley).

http://dx.doi.org/10.1016/j.resuscitation.2015.07.040

National Academies,[4] and to use the methodological approach proposed by the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) Working Group.[5]

In addition, ILCOR leveraged technologic innovations, with the support of science and technology specialists at the American Heart Association, to build a Web-based information system that would support the creation of scientific statements and recommendations that adhere to the GRADE methodology. An online platform known as the Scientific Evaluation and Evidence Review System (SEERS: www.ilcor.org/seers) was developed to guide the taskforces and their individual evidence reviewers, and enabled those responsible for tasks to better monitor progress in real time and receive assignments as indicated by the progression in work flow. One key feature of the SEERS system is the ability to open all components of the process to the public for comments and suggestions. SEERS functions as the repository of all the information and reviews processed since 2012 by the task forces, and Evidence Reviewers and discussions at the C2015 Conference. It remains the home for the 15 GRADE tutorials and 13 GRADE "ask the expert" seminars, as well as housing the training videos produced by AHA staff.

The GRADE process

Why introduce the GRADE process?

The methodological approach proposed by the GRADE Working Group has been developed over the past decade by key health professionals, researchers, and guideline developers in an attempt to provide a consistent and transparent process for use in guideline development.[6] It provides guidance for the rating of quality of evidence and the grading of strength of recommendations in health care. It is now widely used in the guideline development processes throughout the world including by organizations such as the Cochrane Collaboration, the World Health Organization, the National Institute for Health and Care Excellence (NICE), the Scottish Intercollegiate Guidelines Network (SIGN), and the American Thoracic Society.[7] The GRADE approach has been refined to the point that it is now able to incorporate the variety of studies that make up the body of resuscitation science.

What is different about the GRADE process?

The GRADE process outlines a systematic and explicit consideration of study design, study quality, consistency, and directness of evidence to be used in judgments about the quality of evidence for each outcome of each specific question. The GRADE process is, therefore, much more outcome-centric than our previous processes. GRADE considers evidence as a function of the totality of data that informs a prioritized outcome across studies, as opposed to information evaluated at the level of the individual study. The GRADE approach facilitates appropriate consideration of each outcome when grading overall quality of evidence and strength of recommendations, and it reduces the likelihood of mislabeling the overall quality of evidence when evidence for a critical outcome is lacking.[6]

The 2015 ILCOR evidence evaluation process

The 2015 ILCOR evidence evaluation followed a complex but systematic process. In general, the steps followed are consistent with those outlined by the Institute of Medicine.[4] During the development of this process, a transition was made to a more complete online process, using a combination of existing and newly developed tools. The steps in the evidence review process are outlined in Table 1.

Table 1
Summary outline of the evidence evaluation process for the ILCOR 2015 CoSTR.

- Task forces select, prioritize, and refine questions (using PICO format)
- Task forces allocate level of importance to individual outcomes.
- Task forces allocate PICO question to task force question owner and 2 evidence reviewers
- Task force works with information specialists to develop and fine-tune search strategies (for PubMed, Embase, and Cochrane)
- Public invited to comment on PICO question wording, as well as the proposed search strategies
- Revised search strategies used to search databases (PubMed, Embase, and Cochrane)
- The articles identified by the search are screened by the evidence reviewers using inclusion and exclusion criteria
- Evidence reviewers agree on final list of studies to include
- Evidence reviewers agree on assessment of bias for individual studies
- GRADE evidence profile table created
- Draft consensus on science statements and treatment recommendations created
- Public invited to comment on draft consensus on science and treatment recommendations
- Detailed iterative review of consensus on science and treatment recommendations to create final version
- Peer review of final CoSTR document

CoSTR indicates Consensus on Science with Treatment Recommendations; GRADE, Grading of Recommendations, Assessment, Development, and Evaluation; ILCOR, International Liaison Committee on Resuscitation; and PICO, Population, Intervention, Comparator, Outcome.

Task forces, task force question owners, evidence reviewers, evidence evaluation specialist/GRADE/methodology experts

Seven task forces evaluated the resuscitation literature: Acute Coronary Syndromes; Advanced Life Support; Basic Life Support; Education, Implementation, and Teams; First Aid; Neonatal Resuscitation; and Pediatric Life Support. Each task force appoints Task Force Question Owners and Evidence Reviewers to oversee the evidence evaluation process for each question. The task forces were supported by online resources[5,8] as well as telephone, face-to-face, and Web-based educational sessions provided by a GRADE methodologist and an evidence evaluation expert, with advice from a specifically formed ILCOR Methods Group.

Components of the 2015 ILCOR systematic reviews

The evidence evaluation follows a standard format. The key components of this format are described in detail below.

Agree on PICO-formatted question and prioritizing outcomes

Each task force identified the potential questions to be addressed on the basis of known knowledge gaps, priorities as part of previous recommendations, current issues raised by individual resuscitation councils, the known published literature, and areas of controversy. The task forces were then required to prioritize these questions for formal review, and to develop agreed-upon wording by using the PICO (population, intervention, comparator, outcome) format.[9]

As part of the PICO question development, the GRADE process required designation of up to 7 key outcomes for each PICO question. The task force then allocated a score for each outcome on a scale from 1 to 9.[10] Critical outcomes were scored 7 to 9, important outcomes were scored 4 to 6, and those of limited importance were scored 1 to 3. The types of outcomes used (and their possible relevant importance score) included neurologically intact survival (e.g., critical 9), discharge from hospital alive (eg, critical 8), and return of spontaneous circulation (e.g., important 6).

The explicit preference of this process was that if evidence was lacking for a key outcome, this was acknowledged rather than excluding that outcome.

Develop search strategy

Detailed strategies to search the published literature were developed in conjunction with information specialists. Initial draft search strategies were developed for each of 3 databases: PubMed (National Library of Medicine, Washington, DC), Embase (Elsevier B.V., Amsterdam, The Netherlands), and the Cochrane Library (The Cochrane Collaboration, Oxford, England). These strategies were developed to optimize the sensitivity and specificity of the search and then refined on the basis of feedback from the resuscitation community and public comment. The articles identified by the final search strategies were combined into a single database for more detailed analysis by the evidence reviewers.

Identify articles for inclusion and exclusion

Each evidence reviewer used the SEERS online process to screen the identified articles for further review. The initial screening, based on formal inclusion and exclusion criteria, was performed by using each article's title and abstract, and then a review of the full text of the article was performed if needed. Specific inclusion and exclusion criteria varied according to the individual PICO questions, but generic criteria included such items as a requirement for the study to be published in the peer-reviewed literature (not just in abstract form) and to specifically address the individual components of the PICO question. The evidence reviewers were also asked to check for studies that may have been missed in the initial search, by reviewing the references of the identified studies, and performing a forward search on key studies (e.g., by the use of "cited by" in PubMed).

Bias assessment of individual studies

The Cochrane Collaboration's tool was used for assessing the risk of bias for randomised controlled trials.[11] The GRADE tool was used to assess the risk of bias of observational studies (for both therapy and prognosis questions) (Table 2).[12,13]

The Quality Assessment of Diagnostic Accuracy Studies (QUADAS)-2 tool was used for assessing risk of bias in studies of diagnostic accuracy.[14] If there were significant differences in the risks of bias for different outcomes, evidence reviewers were instructed to create a separate row in the table for each outcome. Individual studies can be allocated an overall "low" risk of bias if most or all key criteria listed above are met, and any violations are not crucial. Individual studies that have a crucial limitation in 1 criterion or some limitations in multiple criteria, sufficient to lower the confidence in the estimate of effect, are considered at "moderate" risk of bias. Individual studies that have a crucial limitation in 1 or more criteria, sufficient to substantially lower the confidence in the estimate of effect, are considered at "high" risk of bias.

The two (or more) individual evidence reviewers for each question created a reconciled (agreed) risk of bias assessment for each of the included studies, which was recorded by using an electronic template (Fig. 1).

GRADE evidence profile tables

The GRADE working group has developed validated evidence tables known as *evidence profile tables*. These tables incorporate information on the quality of evidence for each outcome-dedicated row and provide information on effect size and precision, and they can provide information about varying effects across a variety of baseline risks.[15] The evaluation of the evidence supporting each outcome incorporates the information from study design and the five core GRADE domains: risk of bias, imprecision, indirectness, inconsistency, and other considerations (e.g., publication bias).[5] An overall assessment is then made of the quality of evidence to support each outcome (high, moderate, low, or very low).

Table 2
Bias assessment tools.

Randomized controlled trials	
Selection bias	• Was the method used to generate the allocation sequence described in sufficient detail to allow an assessment of whether it should produce comparable groups? • Was the method used to conceal the allocation sequence described in sufficient detail to determine whether intervention allocations could have been foreseen in advance of, or during, enrollment?
Performance bias	• Were measures used to blind study participants and personnel from knowledge of which intervention a participant received? • Was the intended blinding effective?
Detection bias	• Were measures used to blind outcome assessors from knowledge of which intervention a participant received? • Was the intended blinding effective?
Attrition bias	• Were the outcome data complete for each main outcome, including attrition and exclusions from the analysis?
Reporting bias	• Did the study report appropriate outcomes (ie, to avoid selective outcome reporting)?
Other bias	• Was the study otherwise free of important sources of bias not already reported previously?
Observational studies	
Selection bias	• Were appropriate eligibility criteria developed and applied to both the cohort of interest and the comparison cohort? • Was confounding adequately controlled for?
Detection bias	• Was measurement of exposure and outcome appropriate and consistently applied to both the cohort of interest and the comparison cohort?
Attrition bias	• Was follow-up complete?

The completion of these evidence profile tables was facilitated by online access to the Guideline Development Tool (GDT).[16] See Fig. 2.

GRADE evidence profile tables: Study design. The methodological type of study is used by the GRADE process as the starting point for the estimate of overall risk of bias. The rating for each type of study varies according to type of question being asked.

For PICO questions related to therapeutic interventions, evidence supported by RCTs starts as high-quality evidence (⊕⊕⊕⊕). Evidence supported by observational studies starts as low-quality evidence (⊕⊕).[17] For PICO questions related to diagnostic accuracy, evidence supported by valid diagnostic accuracy studies (cross-sectional or cohort studies, in patients with diagnostic uncertainty and direct comparison with an appropriate reference standard) starts as high-quality evidence (⊕⊕⊕⊕).[18] The overwhelming majority of outcomes for the PICO questions were associated with very low quality of evidence (⊕).

GRADE evidence profile tables: Core domains.

Risk of bias. The overall risk of bias for each study relevant to each key outcome was allocated in the bias assessment in individual studies process. In the evidence profile table, a summary assessment is required across the included studies for each outcome. The 3 possible categories are as follows:

- No serious limitations: most information is from studies at low risk of bias.
- Serious limitations: most information is from studies at moderate risk of bias.
- Very serious limitations: most information is from studies at high risk of bias.

RCT bias assessment

Study	Year	Design	Total Patients	Population	Industry Funding	Allocation: Generation	Allocation: Concealment	Binding: Participants	Binding: Assessors	Outcome: Complete	Outcome: Selective	Other Bias
Jones	2002	RCT	152	OHCA	Partly	Low	Low	High	Low	Low	Low	Unclear
Stevens	2002	RCT	36	OHCA	No	High	High	High	Low	Low	Low	Unclear
Laurence	2005	RCT	74	OHCA	No	Low	Low	High	High	Low	Low	Unclear
Zhang	2005	RCT	188	OHCA	Yes	High	High	High	High	Low	High	High
Lopez	2012	RCT	34	OHCA	No	Low	Low	High	Low	Low	Low	Unclear
Simons	2013	RCT	202	OHCA	No	Low	Low	High	Low	Low	Low	Low

Non-RCT bias assessment

Study	Year	Design	Total Patients	Population	Industry Funding	Eligibility Criteria	Exposure/Outcome	Confounding	Follow up
Jinas	2013	Non-RCT	65	OHCA	No	High	High	High	Low
Ruessel	2014	Non-RCT	69	OHCA	No	Unclear	Low	Low	Low

Fig. 1. Example of bias assessment tables (RCTs and non-RCTs).

Evidence across studies may be ranked down for risk of bias by either one level, for serious limitations, or two levels, for very serious limitations.

Inconsistency. Inconsistency is a concept that considers the extent to which the findings of studies that look at the same outcomes agree with each other in a consistent way. Variability in the magnitude of effect may be because of differences in PICO or other differences in study design. Reviewers were asked to document limitations when (1) point estimates varied widely across studies, (2) confidence intervals (CIs) showed minimal or no overlap (ie, studies appear to have different effects), or (3) statistical tests of heterogeneity were suggestive of inconsistency.[19] Again reviewers were asked to assess the studies that report that outcome as having:

- No serious inconsistency.
- Serious inconsistency.
- Very serious inconsistency.

Evidence across studies may be ranked down for inconsistency (by either 1 [for serious limitations] or 2 levels [for very serious limitations]).

Indirectness of evidence. The GRADE process describes direct evidence as "research that directly compares the interventions in which we are interested, delivered to the populations in which we are interested, and measures the outcomes important to patients.".[20] Concerns about directness therefore arise when there are differences in the Population (e.g., patients in cardiac arrest versus not in cardiac arrest), Intervention (e.g., different techniques to induce therapeutic hypothermia), Comparison (e.g., conventional CPR using 2010 guidelines versus conventional CPR using 2000 guidelines), or outcomes (e.g., return of spontaneous circulation versus termination of ventricular fibrillation for 5 s), or where

Author(s): Peter Morley, Eddy Lang
Date:
Question: Drug X compared to Standard Care for Out-of-Hospital Cardiac Arrests
Setting: Prehospital Arrests in Victoria, Australia
Bibliography (systematic reviews): Ruessel, 2014 75; Jinas, 2013 342

Quality assessment							No. of patients		Effect		Quality	Importance
No. of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Drug X	Standard Care	Relative (95% CI)	Absolute (95% CI)		
Survival to hospital discharge (Ruessel, 2014 75; Jinas 2013 342)												
2	cohort studies	serious[1]	not serious	not serious	serious[2,3]	none	17/64 (26.6%)	23/70 (32.9%)	**RR 0.81** (0.48 to 1.37)	62 fewer per 1000 (from 122 more to 171 fewer)	⊕◯◯◯ VERY LOW	CRITICAL
Survival to hospital (Ruessel, 2014 75; Jinas 2013 342)												
2	cohort studies	serious[1]	not serious	not serious	serious[2,3]	none	30/64 (46.9%)	30/70 (42.9%)	**RR 1.09** (0.75 to 1.59)	39 fewer per 1000 (from 107 fewer to 253 more)	⊕◯◯◯ VERY LOW	IMPORTANT

MD – mean difference, RR – relative risk

1. Inadequate control for confounders
2. 95% CI unable to exclude significant harm
3. Total numbers <100 patients

Fig. 2. Example of GRADE evidence profile table completed by using the guideline development tool.

there are no head-to-head comparisons between interventions. Important differences in outcome measures include time frame (e.g., hospital discharge vs 6-month survival) or other surrogate outcomes (e.g., hospital admission vs neurologically intact survival). Usually data that rely on surrogate outcomes would result in an allocation of serious or very serious limitations.

Limitations in more than one type of directness may suggest a need to rate the studies as having very serious limitations.

In general, allocating limitations as serious or very serious should be considered only where there is a compelling reason to think that the biology in the population of interest is so different that the magnitude of effect will differ substantially (eg. cardiac arrest victim vs stroke victim). Evidence from animal studies, manikins or other models would generally be rated as having very serious limitations (but this would be dependent on the key outcomes listed).

Again reviewers are asked to assess the studies that report that outcome as

- No serious indirectness.
- Serious indirectness.
- Very serious indirectness.

Any of these concerns may result in a rating down of the quality of evidence for directness (by either one [serious limitations] or two levels [very serious limitations]).

Imprecision. The assessment of precision and imprecision is complex. The CI around a result enable us to assess the range in which the true effect lies. If the CIs were not sufficiently narrow (such as overlap with a clinical decision threshold, eg. a 1% absolute difference in survival to hospital discharge), the quality would be rated as having serious limitations (or as very serious limitations if the CI is very wide). Another way of describing this is where the recommendation would be altered if the upper boundary of the CI or the lower boundary of the CI represented the true effect. Factors that may further influence this decision include the importance of the outcome, the adverse effects, the burden to the patient, the resources required, and the difficulty of introducing a technique into practice.[21] If the total number of patients included in the evidence for each outcome being evaluated does not exceed the number of patients generated by a conventional sample size calculation for a single adequately powered trial, evidence reviewers were advised to consider rating down for imprecision. This "optimal information size" can be estimated using calculators and tables.[21] Even if the optimal information size is met, and the CI overlaps no effect (ie, CI includes relative risk [RR] of 1.0) evidence reviewers were instructed to rate down the quality of the evidence for imprecision if the CI fails to exclude important benefit or important harm (e.g., a 25% increase in mortality).[21]

Reviewers were asked to assess the studies that reported that outcome as

- No serious imprecision.
- Serious imprecision.
- Very serious imprecision.

If problems with precision were detected, the quality of evidence for precision was rated down (by either one [for serious limitations] or two levels [for very serious limitations]).

Publication bias. Unidentified studies may yield systematically different estimates of beneficial effects of an intervention. Studies with positive results are much more likely to be published (odds ratio, 3.9; 95% CI, 2.68–5.68).[22] Biased conclusions can result from early review (missing studies with delayed publication [even more likely with negative studies]), restricting the search to English language journals, or not including grey literature (e.g., clinical trial registers, abstracts, theses). Discrepancies between meta-analyses of small studies and subsequent large RCTs occur in approximately 20% of cases, in part due to publication bias.

Reviewers should allocate strongly suspected (bias) when the evidence consists of a number of small studies, especially if these are industry sponsored or if the investigators share another conflict of interest.[23] The risk of publication bias in observational studies is probably larger than in RCTs (particularly small studies, data collected automatically, or data collected for a previous study). The use of graphical or statistical testing for publication bias may be useful but has limitations, and is not routinely recommended. Additional information about unpublished trials can be found in databases such as www.clinicaltrials.gov. GRADE suggests that the rating for publication bias across studies should be allocated:

- undetected, or
- strongly suspected.

If publication bias is strongly suspected the quality of evidence is rated down by one level.

Rating up the quality of observational studies. The GRADE group recommends that methodologically rigorous observational studies may have their quality rated up where there is a large magnitude of effect, where there is a dose–response gradient, or when all plausible confounders or biases would reduce the demonstrated effect. Obviously consideration for rating down the quality of evidence (risk of bias, imprecision, inconsistency, indirectness, and publication bias) must precede considerations for rating up the quality.[24] Only a very small number of the systematic reviews identified evidence that met these criteria.

Magnitude of effect. A large magnitude effect would be considered justification to increase the rating by 1 level (eg, from low to moderate) if the RR was 2 to 5, or 0.2 to 0.5 with no plausible confounders. The reviewer would be more likely to rate up if the above size of effects occurred rapidly and out of keeping with prior gradient of change; in these situations, they would usually be supported by indirect or lower levels of evidence. If above criteria are all met, and the RR is very large (e.g., greater than 5–10) or very low (RR less than 0.2), rating up by 2 levels (from low to high) could be considered.

Dose–response effect. A dose–response gradient, such as increased effect with an increased dose, or decreased time to intervention, or increased intensity or duration of an educational intervention, increases the confidence in the findings of observational studies. In this setting, rating up the quality of evidence by 1 level could be considered.

Issues around confounding. If all plausible prognostic factors are accurately measured in observational studies, and if all the observed residual confounders and biases would diminish the observed effect, then the effect estimate would be strengthened. In this setting, rating up the quality of evidence by 1 level could be considered.

GRADE evidence profile tables: Estimate of effect. We asked evidence reviewers to complete the effect size column for each row in the evidence profile tables with an estimate for both relative and absolute effects. For example, binary outcomes required RR (or odds ratio), of the intervention compared to control, with 95% CIs and absolute effect of intervention - control as absolute percentage, with 95% CIs. It is the absolute differences that allow accurate assessment of precision.

There was significant discussion about the exact principles to be employed to determine whether a meta-analysis of data should be performed. There are statistical concerns about the simple combining of results from trials,[25] but there are also significant

concerns about performing a meta-analysis when it would not be appropriate.[26]

If several RCTs or observational studies were identified that published results for outcomes considered critical or important, and these studies were closely matched to the PICO question, the evidence reviewers were encouraged to complete an Assessing the Methodological Quality of Systematic Reviews (AMSTAR) checklist to ensure that the appropriate principles for performance of the meta-analysis were considered.[27] In scenarios where it was thought that the data should not be combined into a meta-analysis, the authors were instructed to list the outcomes for each study, or, if a simple mathematical combination of data was performed, this would be accompanied by a statement suggesting that the data were simply pooled (combined without being weighted).

Guideline development tool

The GRADE process takes a very comprehensive approach to the determination of the direction and strength of any recommendations. During the conduct of the systematic reviews, an updated online tool developed by the GRADE Working Group became available for use. An online Guideline Development Tool[16] developed by the GRADE Working Group was used to help assess the overall balance between benefits and risks or harms for each option, including consideration of dimensions such as patient values and preferences and resource considerations.[28] The ILCOR task forces were encouraged to use this tool to assist in their deliberations.

Creation of consensus on science statements

The completed evidence profile tables were then used to create a written summary of evidence for each outcome: the consensus on science statements. The structure of the new 2015 consensus on science statement was developed as a means of providing an explicit narrative to communicate the evidence synthesis and quality judgments found in the evidence profiles. These statements are supported by a categorization of the overall quality of the evidence (high, moderate, low, or very low) and include reasons for their downgrading or upgrading. The recommended standard consensus on science format was as follows:

> For the important outcome of *Z* (e.g., return of spontaneous circulation), we have identified very-low-quality evidence (downgraded for risk of bias and imprecision) from 2 observational studies (#1, #2) enrolling 421 patients showing no benefit (RR, 0.81; 95% CI, 0.33–2.01).

Creation of agreed treatment recommendations

Consensus-based treatment recommendations were then created whenever possible. These recommendations were accompanied by an overall assessment of the evidence and a statement from the task force about the values and preferences that underlie the recommendations. These are supported by a categorization of the overall quality of the evidence (high, moderate, low, or very low) and strength of recommendation (strong or weak).

The recommended standard treatment recommendation format was as follows:

> We suggest/recommend for/against *X* in comparison with *Y* for out-of-hospital cardiac arrest (weak/strong recommendation, very low/low/moderate/high quality of evidence).

The GRADE process encourages organizations to commit to making a recommendation by using "we recommend" for strong recommendations and "we suggest" for weak recommendations in either a positive or negative direction (ie, "suggest/recommend," "for/against"). In the unusual circumstances in which task forces chose not to make recommendations, they were encouraged to specify whether this was because they had very low confidence in effect estimates (very limited data), because they felt that the balance between desirable and undesirable consequences was so close they could not make a recommendation (data exists, but no clear benefits), or because the two management options had very different undesirable consequences (and local values and preferences would decide which direction to take).

Values and preferences and task force insights

The task forces were encouraged to provide a values and preferences statement whenever a treatment recommendation was made. This is an overarching term that includes perspectives, beliefs, expectations, and goals for health and life as well as the processes used in considering the potential benefits, harms, costs, limitations, and inconvenience of the management options in relation to one another.[28] Task forces were encouraged to provide additional explanatory comments whenever possible to help readers gain more insight into the perspectives of the discussion.

Developing consensus

Each task force used regular audio conferencing and webinars, where the systematic reviews were electronically presented for discussion and feedback. Additional face-to-face meetings were held at least once each year to provide opportunities to learn about the process and to facilitate collaboration between the seven task forces. Consensus was obtained through detailed discussion and feedback provided by the ILCOR task force members, the GRADE and evidence evaluation experts, the ILCOR methods group, the public, and the individual international resuscitation councils.

Public consultation

To ensure as much broad input as possible during the evidence evaluation process, public comment was sought at two main points. Initial feedback was sought about the specific wording of the PICO questions and the initial search strategies. Subsequent feedback was sought after creation of the initial draft consensus on science statements and treatment recommendations.[29] A total of 492 comments were received. At each of these points in the process, the public comments were made available to the evidence reviewers and task forces for their consideration.

Challenging areas

Lower levels of evidence

In many resuscitation scenarios, there are no RCTs or even good observational studies, so there is a need to explore other population groups. The GRADE process is very explicit about the allocation of quality of evidence to support the individual outcomes. Extrapolation of data from other patient groups (e.g., adult versus pediatric, cardiac arrest versus shock), mathematical models, and animal studies means that this evidence, irrespective of methodological quality, would be downgraded for at least serious indirectness. This usually resulted in a very low quality of evidence, and many task forces found this initially challenging.

Diagnostic and prognostic questions

The GRADE process has been developed specifically to deal with questions that address alternative management strategies. It has been modified to enable consideration of questions that relate to diagnosis,[18] but it was not developed to address questions about risk or prognosis.

A few diagnostic questions were addressed in the 2015 process, and ideally the best diagnostic questions relate their outcomes

to when a particular diagnostic strategy is used or not used (i.e., actually an intervention question).

The first of a series of GRADE articles about studies addressing prognosis has been published only recently,[30] so, unfortunately, these details were not available to the evidence reviewers for this process. A couple of approaches to prognosis were used, including the use of existing observational study bias assessment tools or a modification of these.

Discordant recommendations

There were several situations when task forces were keen to use a strong recommendation when the quality of evidence did not support this. This is not unexpected given the few published RCTs and good observational studies available in the resuscitation literature. Task forces were made aware of the importance of clarifying their rationale when they wished to make such discordant recommendations. They were encouraged to use standardized wording (e.g., "Intervention may reduce mortality in a life-threatening situation, and adverse events not prohibitive" or "A very high value is placed on an uncertain but potentially life-preserving benefit").[31] In keeping with this approach, the number of discordant recommendations in ILCOR was limited in the 2015 process, as were the number of strong recommendations.

Management of conflicts of interest throughout the CoSTR process

To ensure the integrity of the evidence evaluation and consensus on science development process, ILCOR followed its rigorous conflict of interest (COI) management policies at all times. A full description of these policies and their implementation can be found in Part 4 of the 2010 CoSTR.[32,33] All persons involved in any part of the process disclosed all commercial relationships and other potential conflicts, and in total, the AHA processed more than 1000 COI declarations. These disclosures were taken into account in assignment of task force co-chairs and members, writing group co-chairs, and other leadership roles. Relationships were also screened for conflicts in assigning task force question owners and evidence reviewer roles for individual PICO questions. Individuals were reassigned when potential conflicts surfaced. Participants, co-chairs, and staff raised COI questions and issues throughout the process and referred them to the COI co-chairs if they could not be resolved within their group. The COI co-chairs kept a complete log of all COI-related issues and their resolutions. None of the issues required serious intervention, such as replacement of any leader roles. As a result of commercial relationships, however, several PICO questions were reassigned to evidence reviewers or question owners without potential conflicts. As in 2010, the phone number for the COI hotline was broadly disseminated throughout the 2015 Consensus Conference for anonymous reporting; no calls were received.

As in 2010, the dual-screen projection method was used for all sessions at the 2015 Consensus Conference. One screen displayed the presenter's COI disclosures continuously throughout his or her presentation. The same was true for all questions or comments from participants or task force members: whenever they spoke, their relationships were displayed on one screen, so that all participants could see potential conflicts in real time, while slides were projected on the second screen. Individuals also abstained from voting on any issue for which they had a conflict. Such abstentions, along with any other issues that arose, were recorded on a COI attestation completed by the COI monitor for each session. As in 2010, the COI system ran smoothly and did not impede the progress of the evidence discussions.

Summary

The process for evaluating the resuscitation science has evolved considerably over the past 2 decades. The current process, which incorporates the use of the GRADE methodology, culminated in the 2015 CoSTR publication, which in turn will inform the international resuscitation councils' guideline development processes. Over the next few years, the process will continue to evolve as ILCOR moves toward a more continuous evaluation of the resuscitation science.

Disclosures

2015 CoSTR Part 2: Evidence evaluation: writing group disclosures.

	Employment	Research grant	Other research support	Speakers' bureau/honoraria	Expert witness	Ownership interest	Consultant/advisory board	Other
Writing group member								
Peter T. Morley	University of Melbourne	None	None	None	None	None	American Heart Association†	None
Eddy Lang	University of Calgary	None	None	None	None	None	American Heart Association†	None
Richard Aickin	Starship Children's Hospital	None	None	None	None	None	None	None
John E. Billi	The University of Michigan Medical School	None	None	None	None	None	None	None
Judith C. Finn	Curtin University	NHMRC (Australia)†	None	None	None	None	None	St John Ambulance Western Australia†
Ian K. Maconochie	St. Mary's Hospital	None	None	None	None	None	None	None
Laurie J. Morrison	University of Toronto	None	None	None	None	None	None	None
Vinay M. Nadkarni	Children's Hospital Philadelphia	NIHAHRQ†; Zoll Corporation†; Nihon-Kohden Corporation*	None	None	None	None	None	None
Nikolaos I. Nikolaou	Konstantopouleio General Hospital	None	SANOFI*; AMGEN*	None	None	None	None	None

	Employment	Research grant	Other research support	Speakers' bureau/honoraria	Expert witness	Ownership interest	Consultant/advisory board	Other
Jerry P. Nolan	Royal United Hospital, Bath	NIHR Programme Development Grant*; NIHR Health Technology Assessment Programme Grant*	None	None	None	None	None	None
Gavin D. Perkins	Warwick Medical School and Heart of England NHS Foundation Trust	None	None	None	None	None	None	None
Michael R. Sayre	University of Washington	None	None	None	None	None	None	None
Jonathan Wyllie	James Cook University Hospital	MRC*	None	None	None	None	None	None
David A. Zideman	Imperial College Healthcare NHS Trust	None	None	None	None	None	None	None
Staff								
Brian Eigel	American Heart Association	None	None	None	None	None	None	None
Jose Maria Ferrer	American Heart Association	None	None	None	None	None	None	None
Lana M. Gent	American Heart Association	None	None	None	None	None	None	None
Russell E. Griffin	American Heart Association	None	None	None	None	None	None	None
Consultants								
Mary Fran Hazinski	Vanderbilt	None	None	None	None	None	American Heart Association†	None
William H. Montgomery	American Heart Association	None	None	None	None	None	American Heart Association†	None
Andrew H. Travers	Emergency Health Services, Nova Scotia	None	None	None	None	None	American Heart Association†	None

This table represents the relationships of writing group members that may be perceived as actual or reasonably perceived conflicts of interest as reported on the Disclosure Questionnaire, which all members of the writing group are required to complete and submit. A relationship is considered to be "significant" if (a) the person receives $10,000 or more during any 12-month period, or 5% or more of the person's gross income; or (b) the person owns 5% or more of the voting stock or share of the entity, or owns $10,000 or more of the fair market value of the entity. A relationship is considered to be "modest" if it is less than "significant" under the preceding definition.

* Modest.

† Significant.

Acknowledgements

The writing group gratefully acknowledges the leadership and contributions of the late Professor Ian Jacobs, PhD, as both ILCOR Co-Chair and inaugural Chair of the ILCOR Methods Group. Ian is greatly missed by the international resuscitation community.

References

1. Morley PT. Evidence evaluation worksheets: the systematic reviews for the evidence evaluation process for the 2010 International Consensus on Resuscitation Science. Resuscitation 2009;80:719–21.
2. Morley PT, Atkins DL, Billi JE, et al. Part 3: Evidence evaluation process: 2010 International Consensus on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science With Treatment Recommendations. Circulation 2010;122:S283–90.
3. Morley PT, Atkins DL, Billi JE, et al. Part 3: Evidence evaluation process: 2010 International Consensus on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science with Treatment Recommendations. Resuscitation 2010;81:e32–40.
4. Institute of Medicine. Standards for systematic reviews; 2011. ⟨http://www.iom.edu/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews/Standards.aspx⟩ (accessed May 6, 2015).
5. Schünemann H, Brożek J, Guyatt G, Oxman A. GRADE handbook; 2013. ⟨http://www.guidelinedevelopment.org/handbook/⟩ (accessed May 6, 2015).
6. Guyatt G, Oxman AD, Akl EA<ET-AL. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64:383–94.
7. GRADE Working Group. Organizations that have endorsed or that are using GRADE. GRADE Working Group; 2015. ⟨http://www.gradeworkinggroup.org/society/index.htm.⟩ (accessed May 10, 2015).
8. American Heart Association, American Stroke Association, International Liaison Committee on Resuscitation (ILCOR). GRADE presentations: SEERS Presentation Library; 2015. ⟨https://volunteer.heart.org/apps/pico/Pages/ILCOR—GRADE-Presentations.aspx.⟩ (accessed May 10, 2015).
9. The Cochrane CollaborationO'Connor D, Green S, Higgins J. Defining the review questions and developing criteria for including studies. In: Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions. 2015. Version 5.1.0. 2011. ⟨http://handbook.cochrane.org/⟩ (accessed May 6), (Chapter 5).
10. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;64:395–400.
11. Higgins JPT, Altman DG, Sterne J. The cochrane collaboration's tool for assessing risk of bias. The cochrane collaboration. In: Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions. 2015. Version 5.1.0. 2011. ⟨http://handbook.cochrane.org/⟩ (accessed May 6), (Chapter 8.5).
12. Schünemann H, Brożek J, Guyatt G, Oxman A. 5 2 1 Study limitations (risk of bias). In: GRADE handbook; 2013. ⟨http://www.guidelinedevelopment.org/handbook/#h.m9385o5z3li7⟩ (accessed May 6, 2015).
13. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). J Clin Epidemiol 2011;64:407–15.
14. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529–36.
15. Vandvik PO, Santesso N, Akl EA, et al. Formatting modifications in GRADE evidence profiles improved guideline panelists comprehension and accessibility to information. A randomized trial. J Clin Epidemiol 2012;65:748–55.
16. Evidence Prime Inc. GRADEpro guideline development tool; 2015. ⟨http://www.guidelinedevelopment.org/⟩ (accessed May 6, 2015).
17. Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64:401–6.
18. Schünemann HJ, Schünemann AH, Oxman AD, et al. GRADE Working Group. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336:1106–10.

19. Guyatt GH, Oxman AD, Kunz R, et al. GRADE Working Group. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. J Clin Epidemiol 2011;64:1294–302.
20. Guyatt GH, Oxman AD, Kunz R, et al. GRADE Working Group. GRADE guidelines: 8. Rating the quality of evidence—indirectness. J Clin Epidemiol 2011;64:1303–10.
21. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 6. Rating the quality of evidence—imprecision. J Clin Epidemiol 2011;64:1283–93.
22. Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database Syst Rev 2007:MR000010.
23. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence—publication bias. J Clin Epidemiol 2011;64:1277–82.
24. Guyatt GH, Oxman AD, Sultan S, et al. GRADE Working Group. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011;64:1311–6.
25. Rücker G, Schumacher M. Simpson's paradox visualized: the example of the rosiglitazone meta-analysis. BMC Med Res Methodol 2008;8:34.
26. Sharpe D. Of apples and oranges, file drawers and garbage: why validity issues in meta-analysis will not go away. Clin Psychol Rev 1997;17:881–901.
27. Shea BJ, Hamel C, Wells GA, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol 2009;62:1013–20.
28. Andrews JC, Schünemann HJ, Oxman AD, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. J Clin Epidemiol 2013;66:726–35.
29. American Heart Association. American Stroke Association, International Liaison Committee on Resuscitation (ILCOR). In: ILCOR Scientific Evidence Evaluation and Review System (SEERS). American Heart Association; 2015. ⟨https://volunteer.heart.org/apps/pico/Pages/default.aspx⟩ (accessed May 10, 2015).
30. Iorio A, Spencer FA, Falavigna M, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. BMJ 2015;350:h870.
31. Alexander PE, Bero L, Montori VM, et al. World Health Organization recommendations are often strong based on low confidence in effect estimates. J Clin Epidemiol 2014;67:629–34.
32. Billi JE, Shuster M, Bossaert L, et al. International Liaison Committee on Resuscitation; American Heart Association. Part 4: Conflict of interest management before, during, and after the 2010 International Consensus Conference on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science with Treatment Recommendations. Circulation 2010;122:S291–7.
33. Shuster M, Billi JE, Bossaert L, et al. International Liaison Committee on Resuscitation; American Heart Association. Part 4: Conflict of interest management before, during, and after the 2010 International Consensus Conference on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science with Treatment Recommendations. Resuscitation 2010;81:e41–7.