

一个新的文献评价系统——GRADE 评价系统

王一飞, 何少茹

(广东省人民医院儿科、广东省医学科学院, 广州 510080)

[摘要] 循证医学要求对现有临床证据进行综合评价以指导临床实践。临床文献的正确评价不但有利于临床指南的制定,也有利于临床决策。目前存在的诸多文献评价系统虽然方便了临床文献的评价,但对于文献评价过程中有争议证据的评价却没有合适的解决方法。本文介绍一种新的文献评价系统,将该系统的文献评价方式及有争议证据的评价方式进行介绍,为我国临床工作人员和政策制定者开展文献评价工作提供参考。

[关键词] 循证医学; 文献; 评价

[中图分类号] R-3 **[文献标识码]** A **DOI:** 10.12019/j.issn.1671-5144.2018.05.013

A New Literature Evaluation System—GRADE System

WANG Yi-fei, HE Shao-ru

(Pediatrics Department, Guangdong General Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510080, China)

Abstract: Evidence-based medicine is an approach to medical practice intended to optimize decision-making by evaluating currently available evidence. Proper evaluation of clinical literature could not only facilitate development of clinical guidelines, but also help make good decision of medical practice. Although current literature-evaluating systems improve the progress of literature evaluation, but it is still unsolvable to correctly resolve the uncertainty between different opinions. This article is to introduce a new literature-evaluating system by elaborating how it grades literature and how it resolves the dispute between different opinions, which might help clinical practitioners and policy makers make better decisions.

Key words: evidence-based medicine; literature; evaluation

随着循证医学在临床工作中的渗透,很多专业组织或杂志都出现了相应的文献评价系统。这些评价系统旨在对现有文献进行评价,并根据评价结果制定疾病诊治指南以指导临床工作。但疾病认识过程的曲折性决定了临床指南的易变性。很多根据当时临床证据做出的所谓正确判断在经过后续的证据积累后被证实是错误的。如何让临床工作者更好地理解临床证据并指导临床实践一

直是制定文献评价系统的挑战。虽然目前已经有10多种文献评价系统,但该问题一直未得到有效解决^[1]。

Guyatt 等从2000年开始,对这个问题进行了一系列的探讨,并建立了GRADE (Grading of Recommendations Assessment, Development and Evaluation)评价系统^[2]。该系统是一个高度结构化、全覆盖、全透明的文献评价系统。它使用结构化方式对证据等级进行分类并设定了相应的推荐强度,使用全覆盖方式保证获得最广泛的证据,使用全透明方式保证所有读者对有争议评价的争议原因有所了解,进而自行判断评价结果的有效性和可行性。截至2016年底,已经有19个国家超过100个国际组织同意或正在使用GRADE评价系

[作者简介] 王一飞(1979-),男,福建古田人,主治医师,医学博士,从事早产儿管理和循证医学研究。

[通讯作者] 何少茹, Tel:020-83827812-61812; E-mail: hsr1605@126.com

统,其中包括世界卫生组织和DynaMed、Cochrane组织、国际复苏联合委员会ILCOR、美国医师协会ACP、英国国家健康中心NHS等^[3]。循证医学数据库UpToDate更是将其汇总的近万条推荐意见使用GRADE进行评价^[4]。我国于2009年引入GRADE概念,并于2011年在兰州大学成立GRADE中心^[5]。同年,《中国循证医学杂志》也对GRADE做了系列报道^[6]。但伴随着国外各大学术组织对GRADE的接受,国内使用GRADE进行文献评价和指南制定的文章屈指可数。究其原因,主要与指南制定者对GRADE不熟悉有关。此外,GRADE工作组近年来也对GRADE进行了持续更新^[7-15],而国内的期刊尚未涉及这部分内容。本文拟对GRADE评价系统(简称GRADE)进行再次综述,补充更新内容,以提高我国临床工作者的文献评价效率及对临床指南的准确掌握水平。

1 制定评价证据等级及推荐强度的原因

临床工作中,对临床证据和指南的评判非常复杂。临床指南不但要明确哪些治疗目的是必须的,明确每一个具体目标有哪些证据支持,还要明确如何评价研究证据以及如何判断该治疗利大于弊。同时,由于资源有限,用于治疗一种疾病的资源将不能用于其他疾病的治疗。如何决定资源分配也很重要。制定评价证据等级能够更有效地判断临床证据的好坏。同样,临床指南也会给出推荐等级。全面、系统、透明地评价证据和决策不仅能够避免错误,还可以促进信息传递。

在GRADE中,证据等级用来评价现有研究证据的可信度,而推荐强度用于评价该研究证据对临床实践的影响。等级高的临床证据可信度高,推荐强度强的干预措施治疗效果好。反之亦然。但凡事并非绝对。有时候可信度高的证据等级不一定会被指南强烈推荐;而可信度低的证据等级也不一定不会被强烈推荐。例如,早产儿肺发育不成熟。静脉给予糖皮质激素能够改善肺功能,减少机械通气时间。虽然糖皮质激素改善肺功能的证据等级为高,但由于生后1周内使用糖皮质激素会对早产儿的脑发育产生不良影响。目前并不推荐早产儿出生后早期使用糖皮质激素帮助撤机^[16]。同样,虽然世界卫生组织不确定寨卡病毒携带者是否会通过性生活传播,证据等级很低,但考虑到寨卡病毒会导致胎儿小头畸形,仍建议孕妇采用有效措施来避免性传播的可能^[17]。

2 GRADE评价系统介绍

GRADE是一种对系统综述和临床指南的证据进行分级以及对指南推荐强度进行评估的评价系统。该系统主要用于干预性研究相关综述或指南的评估。其覆盖内容包括疾病的诊断、筛查、预防和治疗。另外,GRADE也可用于公共卫生和健康相关问题的评价^[18]。

GRADE不仅仅是一个评价系统。它提供了一个透明的、结构化的临床证据汇总方式和指南的推荐方式。特别是在指南制定过程中,它对如何提出临床问题、如何选择理想的研究终点并评估其重要性、如何评估证据等级、如何考虑患者的自身利益及社会价值等方面做了详细阐述。此外,它也给临床医生、患者甚至是政策制定者提供了一种使用临床指南的方法。无论相关研究的质量如何,GRADE都可以使用。

2.1 GRADE评价系统的操作流程

GRADE的流程包括确定临床问题、收集相关证据、评价证据质量、评价推荐等级。

2.1.1 确定临床问题^[19]

在确定临床问题方面,GRADE要求清晰地描述相关的背景、研究人群(population)、干预措施(intervention)、对照(comparator)以及研究结果(outcome)。相关的背景是指指南所应用的环境,如发达国家或发展中国家;研究人群是指指南所针对的人群,如不同年龄、不同基础疾病状态等;干预措施是指指南中所包含的所有干预措施,包括不同的治疗方案、不同的药物剂量、用药途径和治疗疗程等;研究结果应该是对患者很重要的结果,如死亡或生存质量。研究结果可分为三类:对临床决策很关键的结果、重要非关键的结果和不重要的结果。GRADE将其分为9级。7~9级为关键结果;4~6级为重要非关键结果;1~3级为不重要结果。例如,在使用吸入一氧化氮治疗新生儿持续肺动脉高压时,死亡率就属于关键结果,氧合指数改善属于重要非关键结果,而使用一氧化氮的花费则属于不重要结果。如果使用不重要的结果进行研究,则需要降低证据等级。另外,一个临床问题的解决可能需要多个系统综述的证据支持。

2.1.2 相关证据的收集

循证医学的三个原则:①最佳的临床决策需要获得最佳的证据;②循证医学指导人们判断证据的可信度;③单纯证据不足以决定临床决

策^[20]。收集最佳证据是完成临床决策的第一步。最佳的证据常常来源于设计良好的大样本、多中心、随机对照的临床试验或对这些研究的系统综述。在没有相关系统综述或有相关综述质量较低或较陈旧的情况下,研究人员就需要自己获得所有相关的研究结果并分析评价证据的质量(可靠性)。

2.1.3 评价证据质量

2.1.3.1 证据等级的划分

在评价证据质量方面,GRADE通过汇总现有证据并进行等级划分后,将证据等级分为高、中、低和极低。证据等级的高低代表了文献评估者对该证据可靠性的认可程度。高等级指将来的研究不太可能改变文献评估者对估计结果好坏的信心;中等级指将来的研究可能会影响文献评估者对评估结果的信心,甚至改变评估结果;低等级指将来的研究很可能对评估结果产生重要影响,很可能改变评估结果;极低等级指任何效果的评估结果都很不确定。随机对照试验以高等级起始,观察性研究以低等级起始^[21-22]。在GRADE中,证据的质量受研究偏倚的影响。研究自身缺陷、研究结果的不一致性、研究结果的不一致性以及文章发表的偏倚都可能影响证据的可靠性^[23-27]。另外,也有些情况会提高证据的可靠性,如样本量非常大、剂量-反应关系很明显以及可能的偏倚都倾向于削弱治疗效果。正因为存在以上诸多情况,证据的等级会因此升高或降低。

2.1.3.2 降级原因^[21]

证据等级降级主要是因为研究的设计到发表环节中存在严重问题。GRADE根据发生问题的类型将降级原因分为五类。

第一类:研究自身缺陷。分组方式没有保密;没有使用盲法,特别是需要主观判断重要结果,而主观因素很可能影响判断结果时;失访人数多;没有按照意向性分析原则分析研究结果;因为治疗有效而提前停止试验以及没有在文章中报道某些结果,特别是没有治疗结果的结果。

第二类:不一致性。这种不一致性可能因为研究人群、干预措施以及研究结果的不同引起。例如,2006年唯一的一项研究咖啡因治疗早产儿呼吸暂停的研究显示,在欧美及澳洲34家单位的1 006例极早产儿随机接受咖啡因治疗后支气管肺发育不良的发生率明显低于1 000例的对照组^[28]。但2018年另一项纳入83例极早产儿的单

中心随机对照研究显示,使用咖啡因会增加支气管肺发育不良的发生率^[29]。此时,如果不能找到合适的原因来解释这种不一致性,证据等级就会下降。

第三类:间接性。两种情况:第一种,没有两种干预措施直接对比的随机对照研究,而采用都与空白对照对比的方法得出结论。例如,对比布洛芬和对乙酰氨基酚关闭早产儿未闭动脉导管的效果时,只能找到使用布洛芬对比空白对照以及使用对乙酰氨基酚对比空白对照的研究,而没有直接对比布洛芬和对乙酰氨基酚治疗效果的研究。第二种,研究的人群、干预措施、对照以及研究终点指标不同。例如,研究药物关闭早产儿未闭动脉导管时,不同的干预年龄、不同的药物使用途径、对照的选择以及研究的终点指标是短期内动脉导管是否关闭还是长期的出院时死亡与否。这些间接的对比都可能是导致等级降级的原因。

第四类:不够精确。由于纳入研究的人数较少或研究的事件发生率太低等。以研究某新疗法的疾病治愈率为例。虽然新疗法比经典疗法的治愈率有所提高且差异有统计学意义,但由于纳入样本量较小(精确度不够),结果的95%可信区间较大,跨过中线。或者由于该疾病的治愈率极低(如1‰),1 000例患者采用新疗法治愈2例,而对照组治愈1例。治愈率升高100%。这两种情况都可能导致对治疗效果的过度判断。

第五类:报道偏倚。阴性结果较难发表,大部分研究是由相关公司资助完成等。

2.1.3.3 升级原因^[30]

虽然很多随机对照临床试验会因为存在上述问题而让研究证据等级降低,但也存在以下三种情况,可以升级观察性研究的证据等级。第一,样本量非常大(非常大的样本量得出了与其他研究一致的结论时)。第二,有明显的剂量-反应关系(治疗效果越好,证据等级越高)。第三,所有可能的偏倚都倾向于削弱干预组的治疗效果[例如,干预组患者存在多种预后不良的高危因素(病情更重,治疗开始时间更晚),但治疗效果仍好于对照组]。

2.1.4 推荐等级的评判^[12-13]

2.1.4.1 临床指南推荐等级的划分

GRADE除了对研究结果进行评估外,也提出了相应的临床指南推荐等级。根据推荐意见,指

南推荐等级分为强和弱两种,推荐方向分为推荐使用和不推荐使用两种。强推荐是指指南制定的专家委员会相信遵从指南执行后的获益明显优于不良反应。弱推荐指遵从指南执行后的获益可能优于不良反应,但指南制定的专家委员会不太确定。对于不推荐的意见也分为强和弱两种。其意义同推荐意见的强弱类似。推荐意见的强弱要参考证据等级进行。关键研究结果的证据等级决定了推荐等级。而且,每个关键结果都要进行证据等级划分。根据证据等级最低的标准进行评判。

2.1.4.2 影响推荐等级的因素

GRADE 提出了采用6项标准来决定推荐的强度和方向。这6项标准包括:对重要结果(如评估机械通气时间,而不是吸氧浓度的改变)进行的干预治疗的效果、对干预效果评估的可信度、对患者价值观的评估、对患者偏好的评估、这些评估的可信度以及资源的使用情况。这6项标准可以整合为4条:预期效果和不良反应的利弊权衡,重要结果干预效果的可信度,患者价值观、偏好的确定性以及资源的使用。指南制定委员会需要整合以上信息来做出推荐意见。

在权衡预期效果和不良反应前,需要考虑两个问题。对预期效果和不良反应做出最好的评估并确定预期效果和不良反应的平衡点(例如,使用糖皮质激素改善早产儿肺功能从而存活与神经系统不良预后的平衡点。目前较为接受的平衡点是,对于危重症早产儿如果不接受糖皮质激素治疗,由于病情得不到改善,出现神经系统不良预后的风险高于使用糖皮质激素改善病情带来的风险)。要有专门的专家小组评估平衡预期效果和不良反应的问题。而且,指南中必须详细、透明地说明决策过程,并尽可能进行量化。

另一个影响推荐强度的因素是评估效果的可信度。指南不太可能给可信度低的治疗效果一个强推荐。对于没有直接检测关键结果的研究,仅使用替代结果很可能会降低推荐等级。但以下五种情况例外:①如果研究结果显示,某项治疗能够降低病死率。即使研究质量较低,也可能被强推荐。②当某项治疗的预期获益不确定,但副作用和资源消耗非常明显时,该治疗可能会被列为强不推荐。③虽然不确定某项治疗方法的效果如何,但副作用却很小,该治疗方法可能会被强推荐。④虽然两种治疗方法的效果相似,但副作用小的治疗方法可能会被强推荐。⑤当某项治疗方

法确定可以带来少许治疗效果,但其严重不良反应的发生情况尚不明确时,指南会强烈不推荐使用该项治疗方法。

目前关于患者价值观和偏好的研究较少。但对其确定性越差,越可能给予弱推荐。当患者的变异性较大时(如年龄范围0~80岁),很难让一个指南适用于所有人群。积极的人和消极的人对治疗的选择是完全不一样的。

资源的使用也是影响推荐等级的因素^[31]。有些指南之所以不考虑资源消耗问题是因为这方面没有可靠的数据支持、治疗效果明显而资源的使用不会影响临床决策或临床决策由其他人完成。不管如何,指南应该详细说明不考虑资源消耗的原因。

2.1.4.3 推荐强度的特殊考虑

基础发生率会影响治疗利弊的评估。推荐的强度会因是否为高危人群而有不同。死亡是所有人都关心的重要结果。普通流感对于成人不构成威胁,病死率极低;但严重急性呼吸窘迫综合征 SARS 的高病死率使得副作用明显的大剂量激素都不反对使用。

治疗的效果也可能会因地区的差异而有所不同。不同地区会有经济水平的差异、提供相关治疗能力的差异、不同人种的体质差异。对于严重的脓毒血症,在发达国家已经使用监测血流动力学的方法指导治疗,而不发达地区仍以经验性治疗为主。患者的价值观和偏好也会因文化背景不同而有所差异。主流文化和非主流文化也会对价值观产生影响。资源的消耗及在所在地区的价值也会影响推荐强度。

另外,即使证据等级相同,由于不同人群的期望不同,推荐强度也会有所差异。患者的期望与政策制定者和社会学家的期望会有差异。

2.1.4.4 仅限于研究的推荐意见

有时候,指南会对某个方案使用“仅限于研究”的推荐意见。虽然指南没有对其进行推荐,但该方法却可以引导研究人员完成进一步研究以便获得足够的证据来支持或反对该方案。使用“仅限于研究”的推荐意见存在三个前提:①目前的证据尚不足以支持推荐该方案。②进一步研究很有可能降低这种不确定性。③进一步的研究很可能会取得进展,甚至突破。但以下三种情况不推荐使用“仅限于研究”:①该方案的治疗效果完全来自于推测。②虽然该方案的疗效比较明确,

但利弊得失相差不多。很难决定推荐或不推荐该方案。③两种方案各有利弊,而患者由于各自情况不同,会做出不同选择。进一步研究两个方案的治疗效果意义不大。

指南的推荐意见通常需要明确推荐的人群。有时候,还要包括相关的资源配置情况,如发达国家和发展中国家。通常,指南会倾向于推荐某种治疗方案,而不是反对。除非有充足证据证明这个治疗方案无效,甚至有害。为了避免对指南的误解,GRADE建议同时使用符号和文字来表达推荐意见。

需要注意的是,强推荐并不一定是最佳选择。这种情况主要根据指南推荐的对象来确定。某项治疗方案可能治疗效果很好,但要花费大量社会资源。对于患者来说,应该首先考虑使用,但对政策制定者或保险公司来说,则会有所选择。或者说某种稀缺药物可以治疗很多疾病,但有些疾病治疗效果好,有些疾病治疗效果没那么好。这时候,对该药物的分配也会有所不同。影响强推荐意见执行的因素包括疾病的流行情况(优先考虑常见疾病)、使用的难易度(优先选择可以立即执行的方案)、对社会的负担(优先选择负担小的方案)以及执行力度的好坏(执行力差的需要更多关注)。

2.1.4.5 推荐意见的整合

GRADE要求指南中列出决定推荐强度的相关问题以及指南如何整合这些问题的方法。指南制定者之间会对不同的问题有不同的判断。指南详细说明如何整合不同意见可以增加指南的透明度,也可以引起相关人员的关注。

目前不同的指南使用不同的方法表达证据等级和推荐强度,这种方法方便了学术交流。GRADE也推荐使用数字或符号来表示推荐等级。

2.2 GRADE评价系统的优点

GRADE由具有广泛代表性的国际指南制定者制定。它清晰地地区分出证据的质量和推荐等级,详细地评估其他替代治疗结果的重要性,详细地、深度地上调和下调证据质量评级的标准,透明地体现从证据到指南的过程,认同患者的价值观和意愿。它对指南推荐等级强弱的解释清晰、可行,适用于临床工作者、患者以及政策制定者,也适用于系统评价、健康技术评估和指南的评价^[32-33]。

2008年“Surviving Sepsis Campaign”使用了GRADE评价系统。来自10个国家超过50名专家

参与了指南的制定。所有这些专家在参加指南制定前都接受了GRADE应用培训。结果显示,在对相关任务进行介绍和讨论环节,花费时间不到10分钟;而在对指南关注的环节进行讨论仅用了不到2分钟^[34-35]。

2.3 GRADE评价系统的局限性

①GRADE主要用于干预性研究和政策制定,对于风险评估和预后类问题没有涉及。

②对于临床操作流程类的指南,如采集病史、体格检查、帮助患者决策、签署知情同意书等,无法使用GRADE评价;对于效果非常明显的治疗也无需GRADE评价。

③在进行GRADE评价时,需要严格按照推荐的步骤完成指南的制定。

④目前使用GRADE效果最好的是评价预防性和干预性治疗以及解决临床问题。使用GRADE解决诊断性研究和公共卫生问题还存在不足。

⑤使用GRADE无法解决对证据解释有争议的问题。GRADE只是全面地、透明地进行证据评估和推荐。对于证据的采纳与否仍需要自行判断。

2.4 使用GRADE评价系统的标准

GRADE制定的一个目标是减少文献评价和推荐过程中出现的不必要的混乱。GRADE制定小组建议在使用GRADE时要符合以下标准:

①对证据等级的确定要符合GRADE制定的定义。

②要对GRADE中每个项目仔细评估证据的可靠性。

③证据总的判断应该使用(高、中、低和极低)四个等级对每个重要结果进行评估,且评估的每个项目都应该按照GRADE制定的标准进行。

④应以用于总结和决策的证据作为基础来判断证据的可信度和推荐强度。理论上讲,应该使用证据的基本属性来评估证据可靠性,而证据的基本属性应该来自系统综述。至少,用于评价的证据和用来评价的方法应该清晰表达。

⑤在决定推荐或决策的方向和强度上,应该对GRADE的每个标准进行详细考虑。理论上,用于决策的GRADE证据应该记录所有相关的研究证据、相关的考虑和判断。而且,全过程要完全透明。

⑥应该使用与GRADE一致的两个选项(推荐或不推荐)和每个等级(强烈推荐、可以考虑/有条件考虑)来评价推荐强度。

3 其他

GRADE不仅可以用在临床试验中,也可用于诊断试验^[36]。在诊断试验中,研究的设计和存在的偏倚对研究结果的准确性影响较大。有效的诊断试验需要具有代表性且连续获得的患者。这些患者都存在诊断不明确的情况。如果仅选择重症患者和健康人,研究的准确性会被人为升高。另外,对于诊断试验带来的不良反应也应该关注。有些诊断试验虽然精确度高,但由于不良反应可能严重影响病情而应该慎重选择。如有创介入血管造影虽然可以清楚判断冠状动脉狭窄情况,但心肌梗塞甚至死亡属于严重不良反应。在诊断试验的选择时,需要小心。指南选择诊断试验时,也存在研究方法不一致的问题(没有在同一项研究中进行比较,没有使用相同的对照组)。研究结果的不一致性也会影响诊断试验的质量。对患者重要的结果以及诊断成本都是影响诊断试验应用的因素。虽然有些诊断方法的不良反应发生率很低,但其较高的假阴性率会延误疾病的救治时机。而且,诊断试验只是患者利益相关重要预后的一个替代指标。不管诊断效果如何,只有在改善患者病情的前提下,诊断试验才有存在价值。

当然,GRADE只是一种文献评价系统,是一种方法学,无法对专业知识的正确性及适用性进行判断。所以,在进行GRADE评价前,必须完成相关概念的清晰定义以及临床问题的界定(如针对的人群、使用的干预措施以及所采用的对照措施等)。指南制定委员会也要对如何界定推荐等级强、弱或不推荐作出说明,保证讨论的顺利进行。此外,本文仅对GRADE的评价原则和要点进行综述,而《中国循证医学杂志》只对GRADE的部分内容进行了介绍。对使用GRADE感兴趣的研究人员仍需要查阅相关文献或登陆GRADE工作组网址(<http://gdt.guidelinedevelopment.org/app/handbook/handbook.html#h.ged5uqebmir9>)全面了解使用GRADE的细则。

[参 考 文 献]

- [1] ATKINS D, ECCLES M, FLOTTORP S, et al. Systems for grading the quality of evidence and the strength of recommendations I : Critical appraisal of existing approaches The GRADE Working Group [J]. BMC Health Serv Res, 2004, 4(1): 38.
- [2] ATKINS D, BRISS P A, ECCLES M, et al. Systems for grading the quality of evidence and the strength of recommendations II : Pilot study of a new system [J]. BMC Health Serv Res, 2005, 5(1): 25.
- [3] KAHN J M, GOULD M K, KRISHNAN J A, et al. An official American thoracic society workshop report: Developing performance measures from clinical practice guidelines [J]. Ann Am Thorac Soc, 2014, 11(4): S186-S195.
- [4] AGORITSAS T, MERGLEN A, HEEN A F, et al. UpToDate adherence to GRADE criteria for strong recommendations: An analytical survey [J]. BMJ Open, 2017, 7(11): e018593.
- [5] 陈耀龙, 杨克虎, 姚亮, 等. GRADE系统方法学进展 [J]. 中国循证儿科杂志, 2013, 8(1): 64-65.
- [6] GUYATT G H, OXMAN A D, SCHUNEMANN H J, et al. GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology [J]. 中国循证医学杂志, 2011, 11(12): 1427-1429.
- [7] BRIGNARDELLO-PETERSEN R, BONNER A, ALEXANDER P E, et al. Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis [J]. J Clin Epidemiol, 2018, 93:36-44.
- [8] BRUNETTI M, SHELMT I, PREGNO S, et al. GRADE guidelines: 10. Considering resource use and rating the quality of economic evidence [J]. J Clin Epidemiol, 2013, 66(2): 140-150.
- [9] GUYATT G, OXMAN A D, SULTAN S, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes [J]. J Clin Epidemiol, 2013, 66(2): 151-157.
- [10] GUYATT G H, OXMAN A D, SANTESSO N, et al. GRADE guidelines: 12. Preparing summary of findings tables - binary outcomes [J]. J Clin Epidemiol, 2013, 66(2): 158-172.
- [11] GUYATT G H, THORLUND K, OXMAN A D, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles - continuous outcomes [J]. J Clin Epidemiol, 2013, 66(2): 173-183.
- [12] ANDREWS J, GUYATT G, OXMAN A D, et al. GRADE guidelines: 14. Going from evidence to recommendations: The significance and presentation of recommendations [J]. J Clin Epidemiol, 2013, 66(7): 719-725.
- [13] ANDREWS J C, SCHUNEMANN H J, OXMAN A D, et al. GRADE guidelines: 15. Going from evidence to recommendation - determinants of a recommendation's direction and strength [J]. J Clin Epidemiol, 2013, 66(7): 726-735.
- [14] SCHUNEMANN H J, MUSTAFA R, BROZEK J, et al. GRADE guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health [J]. J Clin Epidemiol, 2016, 76: 89-98.
- [15] GUYATT G H, EBRAHIM S, ALONSO-COELLO P, et al. GRADE guidelines 17: Assessing the risk of bias associated with missing participant outcome data in a body of evidence

- [J]. *J Clin Epidemiol*, 2017, 87: 14-22.
- [16] DOYLE L W, EHRENKRANZ R A, HALLIDAY H L. Early (<8 days) postnatal corticosteroids for preventing chronic lung disease in preterm infants [J]. *Cochrane Database Syst Rev*, 2014, (5): CD001146.
- [17] World Health Organization. Pregnancy management in the context of Zika virus infection (Interim guidance update) [EB/OL]. 2016. <http://www.who.int/csr/resources/publications/zika/pregnancy-management/en/> (accessed Oct 11, 2017).
- [18] GUYATT G, OXMAN A D, AKL E A, et al. GRADE guidelines: 1. Introduction - GRADE evidence profiles and summary of findings tables [J]. *J Clin Epidemiol*, 2011, 64 (4): 383-394.
- [19] GUYATT G H, OXMAN A D, KUNZ R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes [J]. *J Clin Epidemiol*, 2011, 64(4): 395-400.
- [20] GUYATT G. What is evidence-based medicine // GUYATT G. Users' Guides to the Medical Literature [M]. 3rd Edition. San Francisco: McGraw Hill Education, 2014: 8.
- [21] BALSHEM H, HELFAND M, SCHUNEMANN H J. GRADE guidelines: 3. Rating the quality of evidence [J]. *J Clin Epidemiol*, 2011, 64(4): 401-406.
- [22] GUYATT G H, OXMAN A D, KUNZ R, et al. What is "quality of evidence" and why is it important to clinicians? [J]. *BMJ*, 2008, 336(7651): 995-998.
- [23] GUYATT G H, OXMAN A D, VIST G, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias) [J]. *J Clin Epidemiol*, 2011, 64 (4): 407-415.
- [24] GUYATT G H, OXMAN A D, MONTORI V, et al. GRADE guidelines: 5. Rating the quality of evidence—publication bias [J]. *J Clin Epidemiol*, 2011, 64(12): 1277-1282.
- [25] GUYATT G H, OXMAN A D, KUNZ R, et al. GRADE guidelines 6. Rating the quality of evidence—imprecision [J]. *J Clin Epidemiol*, 2011, 64(12): 1283-1293.
- [26] GUYATT G H, OXMAN A D, KUNZ R, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency [J]. *J Clin Epidemiol*, 2011, 64(12): 1294-3302.
- [27] GUYATT G H, OXMAN A D, KUNZ R, et al. GRADE guidelines: 8. Rating the quality of evidence—indirectness [J]. *J Clin Epidemiol*, 2011, 64(12): 1303-1310.
- [28] SCHMIDT B, ROBERTS R S, DAVIS P, et al. Caffeine therapy for apnea of prematurity [J]. *N Engl J Med*, 2006, 354 (20): 2112-2121.
- [29] AMARO C M, BELLO J A, JAIN D, et al. Early caffeine and weaning from mechanical ventilation in preterm infants: A randomized, placebo-controlled trial [J]. *J Pediatr*, 2018, 196: 52-57.
- [30] GUYATT G H, OXMAN A D, SULTAN S, et al. GRADE guidelines: 9. Rating up the quality of evidence [J]. *J Clin Epidemiol*, 2011, 64(12): 1311-1316.
- [31] GUYATT G H, OXMAN A D, KUNZ R, et al. Incorporating considerations of resources use into grading recommendations [J]. *BMJ*, 2008, 336(7654): 1170-1173.
- [32] GUYATT G H, OXMAN A D, VIST G E, et al. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations [J]. *BMJ*, 2008, 336(7650): 924-926.
- [33] VANDVIK P O, SANTESSO N, AKL E A, et al. Formatting modifications in GRADE evidence profiles improved guideline panelists comprehension and accessibility to information. A randomized trial [J]. *J Clin Epidemiol*, 2012, 65 (7): 748-755.
- [34] DELLINGER R P, LEVY M M, CARLET J M, et al. Surviving sepsis campaign: International guidelines for management of severe sepsis and septic shock: 2008 [J]. *Crit Care Med*, 2008, 36(1): 296-327.
- [35] JAESCHKE R, GUYATT G H, DELLINGER P, et al. Use of GRADE grid to reach decisions on clinical practice guidelines when consensus is elusive [J]. *BMJ*, 2008, 337: a744. DOI: 10.1136/bmj.a744.
- [36] SCHÜNEMANN H J, OXMAN A D, BROZEK J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies [J]. *BMJ*, 2008, 336 (7653): 1106-1110.

[收稿日期] 2017-12-05