

随机对照试验与 Meta 分析中效应修饰可信度评估工具 (ICEMAN) 的解读



任相颖^{1#}, 王 宇^{1, 2#}, 王天堃¹, 王诗淳¹, 黄 桥¹, 王永博¹, 靳英辉¹, 曹 越², 阎思宇¹

1. 武汉大学中南医院循证与转化医学中心 (武汉 430071)

2. 武汉大学中南医院中南医学期刊社 (武汉 430071)

【摘要】在随机对照试验与 Meta 分析中, 效应修饰 (EM) 常被用于探索干预效果在不同人群中的差异。但大量的 EM 声明实际源于随机误差、多重检验或选择性报告, 缺乏可靠验证, 易误导临床决策, 亟需可信度评估工具加以甄别。EM 可信度评估工具 ICEMAN 针对单个随机对照试验与 Meta 分析分别设计了差异化的评估框架, 涵盖多个核心维度, 最终综合评定可信度为极低、低、中等或高 4 个等级。本文旨在解读 ICEMAN 的概念框架、适用范围及操作要点, 为国内研究人员使用该工具提供参考。

【关键词】随机对照试验; Meta 分析; 效应修饰; 亚组分析; ICEMAN; 解读

【中图分类号】 R195.1 **【文献标识码】** A

Interpretation of the Instrument for assessing the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and Meta-analysis

REN Xiangying^{1#}, WANG Yu^{1,2#}, WANG Tiankun¹, WANG Shichun¹, HUANG Qiao¹, WANG Yongbo¹, JIN Yinghui¹, CAO Yue², YAN Siyu¹

1. Evidence-Based and Translational Medicine Center, Zhongnan Hospital of Wuhan University, Wuhan 430071, China

2. Zhongnan Medical Journal Press, Zhongnan Hospital of Wuhan University, Wuhan 430071, China

[#]Co-first authors: REN Xiangying and WANG Yu

Corresponding authors: CAO Yue, Email: caoyue124@foxmail.com; YAN Siyu, Email: 15927071586@163.com

【Abstract】In randomized controlled trials (RCTs) and Meta-analysis, effect modification (EM) is often used to explore the differences in intervention effects across populations. However, a large number of EM claims actually stem from random error, multiple testing or selective reporting, and lack reliable verification, which may mislead clinical decision-making and thus urgently necessitates dedicated credibility assessment tools for identification. The Instrument for assessing the Credibility of Effect Modification Analyses (ICEMAN) has developed differentiated assessment frameworks for individual RCTs and Meta-analysis respectively, covering multiple core dimensions, and ultimately comprehensively rates the credibility into four grades: very low, low, moderate and high. This paper aims to interpret the conceptual framework, scope of application and key operational points of ICEMAN, and to provide a reference for domestic researchers in

DOI: 10.12173/j.issn.1004-5511.202607051

基金项目: 国家中医药管理局岐黄学者支持项目 (国中医药人教函 [2026] 9 号); 国家自然科学基金青年科学基金项目 (82505373)

[#]共同第一作者

通信作者: 曹越, 博士, 副编审, Email: caoyue124@foxmail.com

阎思宇, 助理研究员, Email: 15927071586@163.com

yxxz.whuzhnmhj.com

applying this tool.

【Keywords】 Randomized controlled trial; Meta-analysis; Effect modification; Subgroup analysis; ICEMAN; Interpretation

在随机对照试验 (randomized controlled trial, RCT) 和 Meta 分析等研究中, 研究者常通过效应修饰 (effect modification, EM) 评估干预措施的效果是否会因另一变量而存在差异^[1-3]。EM 又称为亚组效应 (subgroup effect)、统计交互作用 (statistical interaction) 或干预效应异质性 (heterogeneity of treatment effects) 等^[4], 是指在统计分析中, 某一协变量 (如年龄、疾病严重程度、研究年份) 会影响另一干预变量 (通常是干预措施, 如药物、手术等) 对研究结局 (如疗效、不良反应) 的作用效果, 即干预措施的效果会因变量的不同而呈现差异。

然而, 很多 EM 是不可尽信的^[5-9], 基于此制定的医疗决策可能会使患者无法获得有益治疗, 并可能导致医疗费用的增加^[10]。例如, 在针对短暂性脑缺血发作患者开展的首次大规模阿司匹林试验中, 研究者通过事后数据探索得出“阿司匹林能预防男性卒中, 但对女性无效”的结论^[11]。多年来, 这一发现使许多医生拒绝对患有脑血管疾病的女性患者使用阿司匹林。但后续研究未能稳定重复这一 EM 结果, 提示最初发现可能主要受到随机误差或事后分析的影响^[12-14], 而这一基于不可信 EM 的医疗决策, 直接让女性患者错过了阿司匹林预防卒中的有益治疗, 造成了健康损害与医疗资源的无效消耗。许多理论分析和模拟研究表明误导性 EM 的出现是由于随机误差^[15]: 即使治疗效果对所有患者相同, 但对足够多的样本进行研究时也必然会发现一种看似存在但实则具有误导性的 EM。此外, 选择性报告^[16]、缺乏先验证据^[17], 以及未采用恰当的统计分析方法^[5]也可能导致错误的 EM。由于大多数 EM 未经重复验证以确认其真实性, 因此迫切需要一套标准来区分真实与错误的 EM, 确定在哪些情况下存在真正的效果差异, 以为患者提供正确的治疗方案。为解决这些局限性, Gordon H Guyatt 团队 Stefan Schandelmaier 等开发了效应修饰分析可信度评估工具 (Instrument for assessing the Credibility of Effect Modification Analyses, ICEMAN)^[4]。本文旨在对 ICEMAN 工具进行系统

解读, 介绍工具的使用方法, 为国内相关领域的学者及目标用户群体 (如临床研究者、方法学家) 提供参考。

1 ICEMAN 概念界定、适用范围及注意事项

1.1 概念界定及适用范围

(1) EM 指干预措施对结局的影响会随另一变量的水平不同而变化。

(2) 该工具旨在帮助使用者评估关于“存在 EM”声明的可信度, 若要评估“不存在 EM”的声明, 则需采用不同判定标准。

(3) 若某一 EM 不太可能由随机误差或偏倚导致, 则可认为其具有可信度。

(4) 患者重要性不属于可信度的评估范畴; EM 主要界定 EM 因素与“干预措施对结局的因果效应”之间的关联, 即使修饰变量与结局之间不存在因果关系, 该声明仍可能成立; 且 EM 可基于任意效应尺度 [如相对危险度 (relative risk, RR)、危险差 (risk differences, RD)] 进行评估。

(5) 该工具适用人群包括临床医护人员、临床试验研究者、系统评价作者、卫生技术评估者、期刊编辑、指南制定者及卫生政策制定者。

(6) 该工具适用于单个 RCT 及 RCT 的 Meta 分析, 包含汇总数据和个体参与者数据 (individual participant data, IPD) 的 Meta 分析。

(7) 工具包含 8~12 个问题, 并为每个问题提供明确的评判选项。

(8) 针对不同 EM 时, 对各条目的评判结果应有所区分; 过于严苛或宽松且缺乏区分度的条目无法有效辨别声明可信度的高低。

(9) 工具设置整体评估模块, 以呈现 EM 的整体可信度评级。

1.2 ICEMAN 使用注意事项

研究人员在应用 ICEMAN 工具前需注意以下重要说明。

(1) 应根据每个 EM 因素 (如年龄、合并症、药物剂量等)、每个结局指标 (如死亡率、卒中、

住院时长)、每个时间点(如3个月、6个月)以及每个效应指标(如RR、RD)分别进行可信度评估。

(2) 若交互作用 P 值 ≥ 0.1 ,即几乎无统计学证据支持存在EM,则不适用ICEMAN工具。

(3) 位于左侧的评判选项表示可信度“肯定或可能降低”;位于右侧的评判选项表示可信度“可能或肯定升高”。

(4) 若评判结果为“完全不清楚”,应将其解读为“可能降低可信度”。

(5) 为确保评价过程的透明度,需在每个条目下提供文字说明,阐明评分依据,并将填写完整的ICEMAN评价表作为附件附在文章中。

1.3 界定待评估目标EM需考虑的核心前置填报项

在使用ICEMAN工具评估RCT或Meta分析时,研究人员在定义目标EM前,需先考量以下4个条目。

(1) 明确界定单一候选EM因素(如年龄或合并症)。每份ICEMAN评估表仅评估1个EM因素。

(2) 明确该EM因素是否在随机化时或随机化前测量。若“是”,继续评估;若“否”,停止评估并参阅工具手册以获取进一步说明。ICEMAN仅适用于随机分组前或分组同期测量的EM因素,如基线变量、随机化分层变量。若EM在随机化之后才形成或受干预影响(如中间结局指标),分析会变得复杂,且结果可能存在误导性。当存在以下情况时,ICEMAN可适用于随机化后测量的EM因素:①EM因素为在随机化前已客观存在的特征,如性别、年龄等,其取值在随机化前已确定,不因干预实施而改变;②对于Meta分析,EM因素为研究层面特征,如偏倚风险、随访时长、平均给药剂量。

综上所述,适用性的判断核心在于区分“变量在随机化前已客观存在”,还是“变量本身在随机化后形成或受干预影响”;前者原则上可适用ICEMAN,后者通常不适用。

(3) 明确界定单一的结局指标和观测时间点。多数研究仅涉及一种研究人群、干预措施与对照措施,但会报告多个结局指标。因ICEMAN工具每次仅针对单一结局指标开展评估,使用者必须明确目标结局指标,以及结局指标的观测时

间点(如随访1年的全因死亡率)。

(4) 明确界定单一的效应指标。效应指标的类型是核心评判前提,因EM的幅度通常随效应指标的不同而变化,相对效应指标与绝对效应指标间的差异尤为显著,不同效应尺度下的结论不能直接相互替代。因此,选取的效应指标不同,可信度评级结果也可能不同。开展ICEMAN评价时需明确界定本次分析使用的效应指标,如风险比、比值比、RD等。

1.4 ICEMAN 中文版翻译与解读过程

由2名研究者独立对ICEMAN原始论文及官方使用手册(<https://iceman.help/>)进行翻译,形成初稿后由第3名研究者进行核对与整合,对分歧之处通过讨论达成一致。邀请方法学专家对翻译稿进行审阅,重点关注术语一致性和条目的可理解性。根据中文读者的阅读习惯对部分条目的表述进行微调,但未改变任何条目的评判标准和判定规则。本文内容以ICEMAN原始论文和官方使用手册为依据,未对工具本身进行任何删改或增补。

2 用于评估RCT的ICEMAN工具

用于评估RCT的ICEMAN工具共包含5个条目及1个可选条目;除可选条目外,其他条目下均设有4个选项及相应解释,“说明”栏用于记录勾选对应评价结果的原因。具体内容见附件表1,以下对各条目进行逐项解读。

2.1 条目1 EM的方向是否在事前被正确假设?

若研究者正确预判了EM的方向(如某种干预措施对年轻患者的疗效优于老年患者),则可信度较高;若未能预判方向,则可信度降低;若预判方向与实际结果相反,则可信度最低。此条目综合考量了多项可信度相关要素:①正确预判EM意味着研究者基于生物学机制,或其他因果逻辑,或既往研究证据,形成了明确的研究假设;若数据与研究假设一致,则可信度增加,反之降低。②若研究方案中仅提出“疗效可能随年龄变化”,但未明确EM具体方向(未说明疗效在老年人群还是年轻人群中更优),则该假设的参考价值较弱,其可信度与完全无先验假设的情况相差无几。在贝叶斯统计体系中,明确的先验假设对应采用有信息先验分布,而非无信息先验分

布。此外,事先假设(如基于生物学原理)要比事后解释更可信。事后解读往往是研究者结合试验结果尝试多种推论,易引发多重检验偏倚。若事先假设在研究方案或统计分析计划中得到验证,则可信度达到最高^[4]。

需要注意的是,仅标注“已预设分析内容”并不能完全保证其真实性,也不代表形成了明确的研究假设,更无法规避多重分析带来的偏倚问题。若某一EM因素为随机分层变量,仅能说明研究预先关注该指标,并不等同于已制订了针对性研究假设,但会提升存在预设假设的可能性。此外,当EM因素同时也是预后因素(绝大多数EM因素均具备预后属性)时,EM的作用方向可能会随效应量指标类型的不同而发生改变。

2.2 条目2 EM作用是否有先验证据的支持?

若EM有既往直接或间接证据的支持,则可信度更高;若该EM为首次发现(通常为探索性分析),则可信度降低;若与既往证据相悖,则可信度最低。理想情况下,RCT的证据能够被复现会降低EM是由偶然因素导致的可能性。但开展重复验证研究并得到阳性重复结果的案例目前较为少见,多数情况下也难以获取明确的既往佐证证据。

与上条条目类似,EM方向是重要评判依据。若2项试验显示的EM方向不同,会显著降低可信度。不过受随机误差影响,研究者不应期望所有研究的EM幅度完全一致。许多试验的检验效能不足,部分试验也可能会由于偶然因素而呈现出相反的EM方向^[4]。

2.3 条目3 交互作用检验是否提示偶然性不太可能是造成明显EM作用的原因?

若交互作用检验结果提示,在原假设(即不存在EM)成立的前提下,观察到的EM不太可能由偶然因素导致,则可信度较高;若交互作用检验显示EM与偶然误差相符,或未开展交互检验、或无法推算检验结果,那么可信度则偏低。

本条目仅采用原文报告的交互作用检验结果(通常为 P 值),不考量该 P 值是否经过多重检验校正,也不区分EM因素是联合分析还是逐一分析。多重检验相关问题将在下一条目中单独评判。值得注意的是,仅依据某一亚组结果有统计学意义、另一亚组无统计学意义进行判断,参考价值十分有限,该方式无法判断亚组间的疗效差

异是否源于偶然因素。

目前已有多种交互作用检验方法,RCT最常用的方式是在回归模型中纳入交互项。多数RCT并未采用单一量化指标[如RR值及其置信区间(confidence interval, CI)]来描述EM,而是通过图表展示各亚组的效应值,规范的报告会同步附上交互作用检验 P 值。若原文未报告交互作用 P 值,有时可根据各亚组的效应点估计值与CI进行推算。经验法则为若各亚组效应值的95%CI互不重叠,则交互作用 P 值通常 <0.05 。

本条目以0.05、0.01、0.005这3个常用 P 值界值划分评价等级,其中 P 值 ≤ 0.005 对应可信度最高的等级。除 P 值外,交互作用CI、贝叶斯因子等统计指标往往比 P 值更具信息量和直观性,但这类统计量在报告中并不常见^[4]。

2.4 条目4 研究者是否仅检验了少量EM因素,或在统计分析中考虑了检验的数量?

多重检验是EM分析中需重点关注的问题。临床试验研究者通常会收集大量基线指标,其中多项指标都可作为候选因素开展EM检验。由于多重检验会增大偶然得出阳性结果的风险,即可信度会随着检验的候选EM因素数量增加而降低,因此,若研究者仅检验了少量EM因素,其可信度更高。此条目建议统计研究中候选EM因素的数量,若有研究方案,需以方案记录为准进行核对。

多重检验偏倚的产生存在多种情形。最常见的情况是研究者检验多个候选EM因素,但仅挑选有统计学意义的结果进行报告;其他潜在问题包括连续型EM因素的截断值选取等。多个观测时间点、多种效应量尺度、多个结局指标、多种交互检验方法,均会引发多重检验问题。因此,即使检验的EM因素数量很少,也应考虑是否存在其他可能造成多重检验偏倚的问题。

除了限制分析数量外,也可通过统计学方法校正多重检验问题。若经多重检验校正后,EM结果依然成立,则可信度更高。常用校正方法包括:基于整体I类错误率对 P 值进行校正、将所有候选EM因素纳入同一模型分析、构建风险评分等复合指标,以及采用压缩估计法。上述所有方法均会不可避免地降低检验效能。

多重检验的评估高度依赖完整的结果报告。若无法获知实际开展EM分析的数量,则难以评

估多重检验带来的影响。理想情况下,研究者应在研究方案中明确列出所有候选EM因素,同时注明指标定义与分析细则。若无法获取研究方案,则需查阅原文中关于EM因素数量的明确说明。需注意的是,即使某一项EM因素是预先设定的,也不代表不存在多重检验问题,因研究者可能预先设定了大量其他EM因素。

综上,本条目评判流程为:统计EM因素数量(同时兼顾其他多重检验相关问题);条件允许时结合研究方案进行核对;判断研究者是否在统计分析中针对多重检验采取了相应处理措施^[4]。

2.5 条目5 若EM因素为连续变量,是否避免了随意设定截断值?

对连续型EM因素进行分组划分是临床研究中的常见做法,但人为设定截断值可能会引入多重检验偏倚、降低检验效能、掩盖变量间线性或非线性关联的问题,并增加不同研究间结果比对的难度。因此,避免使用截断值并使用完整数值范围的分析是最可信的。

但研究者经常不使用完整数据,而是使用截断值将连续变量划分为2个或多个类别,其中有充分实证依据的分类方式可信度最高。若研究者为了凸显EM结果,根据数据特征择优选取拟合度最佳的截断值(即数据驱动型截断值),则分析可信度极低,也会大幅提升假阳性结论的发生风险。

有序分类(如低、正常、高血压)同样依赖截断值划分,因此也存在主观随意性问题。然而,若多个有序亚组显示出明确的趋势,则可增强可信度。需注意的是,尽管名义变量的分组不涉及截断值,但分组也可能存在人为偏倚。

在对连续型EM因素进行模型拟合时存在一些挑战,有可能降低可信度,如若变量间的关联规律仅由少数极端观测值主导,可能导致模型设定偏误;事后根据数据反复调整模型则会导致过拟合。因此,可信度最高的分析方式为研究者提前明确治疗效应与连续变量的函数关系,如线性关系、对数关系,或预先限定少数几种备选函数形式。重叠亚组分析法(如滑动窗口法)可作为截断值法与复杂模型拟合的替代方案,该方法的可信度远高于人为设定截断值,但结果解释难度较大。制作回归函数曲线并附带CI,同时对模型

开展充分验证,连续变量分析的可信度会进一步提升。针对连续型EM因素,还应规范开展交互作用检验。

此外,判断连续型EM分析的可信度,还可参考其他补充依据,例如是否存在明确的剂量-反应关系、分析结果是否在敏感性分析中保持稳健(详见后续条目)^[4]。

2.6 条目6 是否存在可能提高或降低可信度的其他考量因素?

此条目为可选填条目,即不填写此条目也不影响可信度评估。需注意的是,若核心项目已判定结论可信度为低或极低,则无需再考虑这些额外因素。

这些额外考量因素由于相关性较低、适用场景有限,或难以评估而未被纳入核心条目中,因此答案选项仅为“可能降低”和“可能提高”,以下是可能相关的额外考量因素。

(1) 敏感性分析显示结果对相关假设具有稳健性:敏感性分析有助于提高EM结论的可信度,若EM分析基于分组后的连续变量,当更换不同截断值时EM仍然存在,则可信度提高。

(2) EM因素不同水平间存在“剂量-反应关系”:若治疗效应随EM因素水平变化呈单调递增或递减趋势,则可信度提升;反之,若将研究对象划分为3个及以上有序亚组后,疗效变化无合理解读规律,即便结果存在统计学意义,也需高度警惕。典型情况为连续变量分组后,仅某一个亚组出现疗效异常,相邻亚组却无类似表现。

(3) 校正其他潜在EM因素后,结论仍然成立:若多因素分析显示,某一EM因素独立于其他候选EM因素,则可信度更高,如可采用逐步回归法判断研究中是否存在多个EM因素。需注意的是,多个EM因素在统计学上相互独立,并不等同于存在因果关联,但会小幅提升因果推断的可能性。然而,大多数EM分析检验效能不足,难以开展有效的多因素分析,且研究往往无法提前锁定全部关键EM因素。

(4) RCT主效应的偏倚风险:若原始研究在随机分配、盲法、数据缺失和报告方面存在高偏倚风险,则其所有二次分析结论的可信度均会下降。目前学界普遍使用Cochrane偏倚风险评估工具评价RCT整体偏倚风险,但关于RCT整体偏倚风险与EM偏倚关联性的研究较少。部分方法学

家认为,交互作用检验通常不易受主效应混杂因素和EM因素测量误差的干扰;但也有研究表明,相较于非企业资助的试验,企业资助的临床试验更易得出虚假的EM结论,尤其是当主效应不显著时。

(5) 试验具有极高的EM检验效能:有方法学家认为,研究事前设定的检验效能越高,所得EM结论的可信度越高。但大多数EM分析的检验效能较低,且研究方案中很少专门针对EM开展样本量与效能计算。

(6) 相关结局指标的EM结果保持一致:若多项存在生物学关联(或其他内在关联)的结局指标均得到相似的EM结果,则可信度更高。需注意的是,判断一致性需结合EM的大小与方向,而非仅依赖统计学显著性,例如不同结局样本量存在差异,会直接影响P值结果。同时需警惕某些偏倚可能在多个相关结局中均有体现,从而错误提示可信度提高^[4]。

2.7 整体评级 您如何评价所提出的EM的整体可信度?

本工具通过汇总全部评价条目所得结论,综合获得整体可信度评级。整体评分采用连续型视觉模拟量表,划分为极低、低、中等、高可信度4个区间,反映了研究者对EM真实存在(而非偶然误差或偏倚所致)的信心程度,依次对应<25%、25%~50%、50%~75%及>75%。具体判定规则见附件表1,建议在评级下方预留区域,对综合评级结果、各条目权重分配作出说明。

3 用于评估Meta分析的ICEMAN工具

用于评估Meta分析的ICEMAN工具共包含8个条目及1个可选条目,其中4个条目与用于评估RCT的ICEMAN工具重叠;除可选条目外,其他条目下均设有4个选项及相应解释,“说明”栏用于记录勾选对应评价结果的原因。具体内容见附件表2,以下对各条目进行逐项解读。

3.1 条目1 EM分析是基于试验内比较还是试验间比较?

基于试验间比较(即以各项研究作为亚组)得出的EM结论,其可信度通常远低于基于试验内比较得出的结果。

试验间比较的一个重要问题是研究水平混杂偏倚。若某项研究变量(如干预方案类型)与结

局指标存在关联,而该关联可能被其他研究变量(如偏倚风险)干扰,则最终得到的EM可能是虚假关联。当假定的EM因素是由受试者个体指标汇总而成的研究变量(如平均年龄)时,研究水平混杂偏倚可能更具误导性。这类汇总指标在不同试验间的差异通常较小,无法反映试验内部受试者的真实个体差异,导致检验真实的试验内EM的效能大幅降低,观察到的EM也大概率是由研究水平混杂因素所致。

汇总数据Meta分析是最常见的Meta分析类型,其EM分析通常完全基于试验间比较(如Meta回归),这些分析存在较高研究水平混杂偏倚风险,因此可信度较低。

部分研究者会结合试验内与试验间数据开展分析,常用方式包括以下3类:①分别估算试验内和试验间的EM结果,再对两类结果进行合并;②在单阶段IPD Meta分析模型中直接纳入交互项;③先按亚组对各项临床试验数据分别合并,再对比不同亚组的汇总效应量。第3种方式灵活性最强,可纳入仅报告单一亚组数据的临床试验,但相应也会提升研究水平混杂的发生风险。

上述3种分析方式均存在一定的研究水平混杂风险。为此,ICEMAN设置了2项中间评价选项:若权重最高的核心试验仅提供单一亚组数据,提示该EM结果大概率由试验间差异主导,可信度可能下降(判定为以试验间比较为主);若核心试验均可提供试验内对比数据,则EM结果主要依托试验内信息得出,可信度可能上升(判定为以试验内比较为主)。若无特殊说明,IPD Meta分析中的EM分析一般都融合了试验内和试验间信息,无法完全规避研究水平混杂偏倚。

若EM分析完全基于试验内数据,则不存在研究水平混杂偏倚,此时需满足以下情况:Meta分析中纳入的所有试验均提供或可推算出试验内EM,再单独对各项研究结果进行汇总合并;也可采用更复杂的建模方法,在单阶段模型中单独提取出研究内的EM效应,该方法同样适用于IPD Meta分析^[4]。

3.2 条目2 对于试验内比较,各试验间的EM是否具有-致性?

若EM在独立试验中得到重复验证,则其可信度会提高。重复验证可降低随机误差,也能降

低混杂因素造成偏倚的可能性。由于不同研究的重复结果难以做到完全一致，本条目结合 EM 的作用方向与效应大小划分了多级评价标准。若适用本条目，建议逐一量化每项临床试验的 EM 幅度。需注意，本项评价内容与利用 P 统计量评估不同研究间治疗效应的一致性（或异质性）并非同一概念^[4]。

3.3 条目3 对于试验间比较，纳入的试验数量是否充足？

对于基于试验间比较的 EM，可信度会随纳入试验数量的增加而提高。若纳入试验数量较少，EM 可能由过拟合或混杂偏倚所致。充足的试验数量还能提高分析检验效能，也能更好地拟合随机效应模型中的研究间离散程度。

本条目评价标准以样本量最少亚组的最低研究数确定；对于连续变量的 Meta 回归，则以纳入研究的总数量作为评判依据。该参考标准在特定场景下可调整：当 EM 因素包含 3 个及以上有序分组时，允许其中某个亚组仅纳入少量研究。对于连续变量的 Meta 回归，除统计研究总数量外，评价者还需考虑各项研究在 EM 因素不同水平上的分布情况。如共纳入 20 项研究，但其中 18 项集中分布在变量区间的一端，那么其证据力度会远低于研究均匀分布于整个区间的情况^[4]。

3.4 条目4 EM 方向是否在事前被正确假设？

在 2.1 基础上需补充说明：由于 Meta 分析是回顾性的，研究者在进行分析时，可能已了解试验结果及可能存在的 EM 因素；此情形下，本条目即使提示可信度提高，其参考价值也会降低。若所有关键原始试验均未检验目标 EM 因素，且 EM 分析完全基于试验间比较，则本条目的评判意义更为突出^[4]。

3.5 条目5 交互作用检验是否提示偶然性不太可能是造成明显 EM 作用的原因？

与 RCT 工具条目 3 解释相同。

3.6 条目6 研究者是否仅检验了少量 EM 因素，或在统计分析中考虑了检验的数量？

在 2.4 基础上需补充说明：本条目的潜在局限性是 Meta 分析研究者在规划 Meta 分析前，可能已筛选出相关 EM 因素。此情形下，即使 Meta 分析最终仅检验少量 EM 因素，也会掩盖各原始试验前期筛选环节产生的多重检验偏倚^[4]。

3.7 条目7 研究者是否使用了随机效应模型？

若研究者在亚组内采用了随机效应模型开展分析，EM 的可信度会更高。若研究者采用了以下模型，可信度会较低：①共同效应模型：该模型假定同一亚组内所有研究基于同一人群，关键特征完全一致，但对基于公开发表的文献汇总的 Meta 分析而言，这种情况几乎不成立；②固定效应模型：其计算方式与上述共同效应模型相同，但解释不同，该模型假设亚组结果仅适用于该亚组内已纳入的研究，无法推广到其他研究，这违背大多数 Meta 分析的研究目的。因此，上述 2 个模型无法合理量化研究间异质性导致的结果不确定性。

模拟研究证实，无论是汇总数据 Meta 分析还是 IPD 水平 Meta 分析，未采用随机效应模型会大幅提升假阳性结论风险。若研究者声明采用了混合效应模型但未进一步说明，通常代表他们对亚组内的研究差异采用了随机效应模型，对亚组间差异采用了固定效应模型，这种情况下可显著提高可信度。

本条目也适用于 IPD Meta 分析，实证研究表明，大多数此类分析并未采用随机效应模型^[4]。

3.8 条目8 若 EM 因素为连续变量，是否避免了随意设定截断值？

与 RCT 工具条目 5 解释相同。

3.9 条目9 是否存在可能提高或降低可信度的其他考量因素？

可选条目，与用于评估 RCT 时的可选条目考量相同，即这些额外因素由于相关性较低、适用场景有限，或难以评估未被纳入核心条目中，因此答案选项仅为“可能降低”和“可能提高”。值得注意的是，若核心条目已判定结论可信度为低或极低，则无需再考虑这些额外因素。以下是可能相关的额外考量因素 [在“用于评估 RCT 的 ICEMAN 工具”条目 6 的 (1) (2) (3) (6) 基础上补充说明]。

(1) EM 得到外部证据支持：若所提出的 EM 与 Meta 分析未纳入的研究（如高质量队列研究）结果一致，则可信度可能更高。

(2) 单个 RCT 或 Meta 分析主效应的偏倚风险：若单个原始研究或 Meta 分析本身存在高偏倚风险，则 EM 可信度下降。常用的正式评估工具

包括针对单个临床试验的 Cochrane 偏倚风险评估工具, 以及针对系统评价的 ROBIS 偏倚风险评估工具。但现有文献极少探讨整体偏倚风险与 EM 分析偏倚之间的关联。若仅部分研究报告某一 EM 因素、其余研究未报告, 则会产生发表偏倚, 即研究者往往仅在得到阳性结果后才报道对应亚组数据。

(3) Meta 分析具有极高的效能以检测 EM: 方法学研究指出, 研究预设的检验效能越高, 所得 EM 结论可信度越高。但绝大多数 EM 分析检验效能偏低, 研究方案中也极少专门针对亚组交互

开展效能计算。

3.10 整体评级 您如何评价所提出的 EM 的整体可信度?

与 RCT 工具整体评级条目解释相同。

4 用于评估 RCT 和 Meta 分析的 ICEMAN 工具比较

ICEMAN 工具中 RCT 和 Meta 分析 2 个版本的核心问题对比见表 1, 两版本核心条目存在部分重叠, 均可用于评估 EM 可信度, 但评估侧重点存在不同考量。

表 1 ICEMAN 工具两版本核心问题对比

Table 1. Comparison of the core questions of the 2 versions of the ICEMAN

条目	版本*	
	RCT	Meta 分析
EM 分析是基于试验内比较还是试验间比较?	NA	1
对于试验内比较, 各试验间 EM 是否具有有一致性?	NA	2
对于试验间比较, 纳入的试验数量是否充足?	NA	3
EM 的方向是否在事前被正确假设?	1	4
EM 作用是否有先验证据的支持?	2	NA
交互作用检验是否提示偶然性不太可能是造成明显 EM 作用的原因?	3	5
研究者是否仅检验了少量 EM 因素, 或在统计分析中考虑了检验的数量?	4	6
研究者是否使用了随机效应模型?	NA	7
若 EM 因素为连续变量, 是否避免了随意设定截断值?	5	8

注: *数字表示各问题在完整工具中的出现顺序; NA. 不适用 (not applicable)。

5 讨论

ICEMAN 工具的提出旨在针对目前 EM 研究中普遍存在的可信度不足问题, 为研究者提供一套结构化的可信度评估方法, 本文对该工具的概念框架、适用范围及具体评估条目进行了系统解读。

EM 在 RCT 和 Meta 分析中非常普遍, 一项针对 35 篇发表于顶级综合医学期刊并开展亚组分析的 RCT 的研究显示, 21 篇 (60%) 宣称存在 EM^[18]; 而多项实证研究揭示了 EM 声明的可信度普遍偏低^[19-20]。在一项基于 ICEMAN 工具针对 2004 至 2020 年间发表的 III 期肿瘤学试验的系统调查中, 101 项声称存在治疗效果差异的亚组分析中仅 7 项 (7%) 被评为中等可信度, 93% 的声明被评为低或极低可信度^[19]。另一项针对 Meta 分析中不同性别治疗效果差异的亚组分析显示, 研究纳入 216 篇基于 RCT 的 Meta 分析, 其中 97 篇提出 115 项“性别影响治疗效果”的论断, 结果

显示仅 27 项 (23.5%) 具备统计学意义, 39 项 (33.9%) 存在明显统计谬误, 剩余 49 项 (42.6%) 因数据不足无法验证, 且无一被评为高可信度^[20]。在卒中领域, 一项纳入 74 篇文献、647 项亚组分析的系统综述显示, 65% 的 EM 被 ICEMAN 评为低可信度^[21]。研究者指出, 统计方法的不规范是导致可信度缺失的首要原因。

Schandelmaier 等^[15]通过系统综述系统梳理识别出 36 个 EM 标准、150 篇文献及 30 个清单, 研究发现, 虽然学界已经认识到需为 EM 声明建立评估框架, 但这些标准之间缺乏共识, 表述方式不一致, 并且在操作层面缺乏明确的指导。这易使 RCT 及 Meta 分析研究者、临床医师、卫生决策人员产生困惑, 也极大阻碍了研究者对 EM 声明进行统一、透明的可信度评估, 使虚假的 EM 结论得以持续出现在已发表的文献中, 进而对循证决策产生潜在误导。

GRADE 工作组在 2023 年发表的《GRADE 指南 36》中纳入 ICEMAN, 明确其可用于评估 EM

可信度，与 GRADE 评价体系完全兼容^[22]。ICEMAN 与 GRADE 均采用高、中、低、极低 4 个评级，并且均要求评价者基于多项标准进行综合判断而非依赖单一指标。当 ICEMAN 判定 EM 为“高”或“中等”可信度时，建议研究者分亚组呈现效应估计值及证据确定性评级；当可信度为“低”或“极低”时，则应使用整体效应估计值。

ICEMAN 也存在一定局限性。首先，部分条目的评分阈值存在争议。例如，在用于评估 RCT 的 ICEMAN 工具中，条目 3 将交互作用 P 值 ≤ 0.005 作为“偶然性是不太可能的解释”的阈值，这一阈值是否适用于不同样本量、不同结局类型的研究场景，尚需进一步验证；条目 5 对连续变量截断值的评价标准也存在类似问题。其次，整体评级仍包含一定主观判断成分。工具目前尚无法全面覆盖所有复杂场景，如数据驱动的亚组发现算法、多重 EM 因子间相关结构的考量等。对此，ICEMAN 提供了“可选条目”，允许评价者在核心条目之外纳入额外考量因素。此外，ICEMAN 的可操作性高度依赖于原始研究的报告质量。国内许多 RCT 和 Meta 分析未公开发布研究方案或统计分析计划，也未明确报告实际检验的候选 EM 因素数量、截断值的选取依据等关键信息。在此类情形下，评价者往往只能选择“不清楚”选项，可能导致最终评级更多地反映原始研究的报告不充分，而非分析本身的方法学质量较差。ICEMAN 开发团队在工具说明中也承认了这一局限，建议评价者在说明栏中详细记录“不清楚”判断的具体原因，并与“可能否”等负面判断加以区分。未来研究在使用 ICEMAN 时，应特别注意区分“分析未执行”与“分析已执行但未报告”2 种情形，并在评价报告中予以明确说明。

ICEMAN 开发团队在工具发表时已明确指出了未来的研究方向。首先是工具信效度的系统验证，需要更多独立研究团队在不同临床领域和不同研究类型中应用 ICEMAN，以评估其评价者间信度和预测效度。其次，随着 EM 分析方法学的持续发展，工具可能需要适时修订和扩展。例如，针对多个 EM 因子间的交互作用、不同效应尺度（如 RR 与 RD）下 EM 的可信度比较，以及真实世界数据中 EM 的评估等更复杂场景，可考虑开发 ICEMAN 的扩展版。Schandelmaier 与

Guyatt 等^[15, 22]进一步指出，除应用 ICEMAN 外，还可在 SPIRIT 和 CONSORT 指南中增加亚组分析方法学细节，从源头上提高报告质量。

综上所述，ICEMAN 工具不仅为评估 EM 声明的可信度提供了系统框架，也为未来研究提供了明确指引。建议研究者在开展 EM 分析时，规范 EM 分析的研究设计与报告标准，详细报告先验假设、EM 因素数量、交互作用检验结果、连续变量分析方法等关键信息，同时建议将 ICEMAN 评估结果作为研究补充材料，作为判断 EM 可信度的重要依据，以提升临床决策依据的质量。

附件见《医学新知》官网附录 (<https://yxxz.whuznhmedj.com/futureApi/storage/appendix/202607051.pdf>)

伦理声明：不适用

作者贡献：论文撰写：任相颖、王宇；论文修改：王天堃、王诗淳、黄桥、王永博；文章指导与审阅：靳英辉、阎思宇、曹越

数据获取：不适用

利益冲突声明：无

致谢：不适用

其他声明：ICEMAN 工具的中文版解读已获得原文第一作者与通信作者的正式授权，中文版解读并非由原文发表杂志 *Canadian Medical Association Journal* (CMAJ) 翻译及审核，CMAJ 不对中文版的翻译准确性及解读内容承担任何责任，特此声明

参考文献

- Schuit E, Li AH, Ioannidis JPA. How often can Meta-analyses of individual-level data individualize treatment? A Meta-epidemiologic study[J]. *Int J Epidemiol*, 2019, 48(2): 596–608.
- Gabler NB, Duan N, Ranases E, et al. No improvement in the reporting of clinical trial subgroup effects in high-impact general medical journals[J]. *Trials*, 2016, 17(1): 320.
- Zhang S, Liang F, Li W, et al. Subgroup analyses in reporting of phase III clinical trials in solid tumors[J]. *J Clin Oncol*, 2015, 33(15): 1697–702.
- Schandelmaier S, Briel M, Varadhan R, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and Meta-analyses[J]. *CMAJ*, 2020, 192(32): E901–E906.
- Sun X, Briel M, Busse JW, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review[J]. *BMJ*, 2012, 344: e1553.
- Lagakos SW. The challenge of subgroup analyses—reporting without distorting[J]. *N Engl J Med*, 2006, 354(16): 1667–9.
- Wang R, Lagakos SW, Ware JH, et al. Statistics in medicine—reporting of subgroup analyses in clinical trials[J]. *N Engl J Med*, 2007, 357(21): 2189–94.
- Wang SS, Ou YC, Cheng CL, et al. Evidence-based urology in practice: when to believe a subgroup analysis?[J]. *BJU Int*, 2010, 105(2): 162–4.

- 9 Wallach JD, Sullivan PG, Trepanowski JF, et al. Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials[J]. *JAMA Intern Med*, 2017, 177(4): 554–60.
- 10 Sun X, Ioannidis JP, Agoritsas T, et al. How to use a subgroup analysis: users' guide to the medical literature[J]. *JAMA*, 2014, 311(4): 405–11.
- 11 Canadian Cooperative Study Group. A randomized trial of aspirin and sulfinpyrazone in threatened stroke[J]. *N Engl J Med*, 1978, 299(2): 53–59.
- 12 Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy—I: prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients[J]. *BMJ*, 1994, 308(6921): 81–106.
- 13 Gdovinova Z, Kremer C, Lorenzano S, et al. Aspirin for primary stroke prevention; Evidence for a differential effect in men and women[J]. *Front Neurol*, 2022, 13: 856239.
- 14 Ridker PM, Cook NR, Lee IM, et al. A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women[J]. *N Engl J Med*, 2005, 352(13): 1293–304.
- 15 Schandelmaier S, Chang Y, Devasenapathy N, et al. A systematic survey identified 36 criteria for assessing effect modification claims in randomized trials or Meta-analyses[J]. *J Clin Epidemiol*, 2019, 113: 159–67.
- 16 Donegan S, Williams L, Dias S, et al. Exploring treatment by covariate interactions using subgroup analysis and Meta-regression in Cochrane reviews: a review of recent practice[J]. *PLoS One*, 2015, 10(6): e0128804.
- 17 KASENDA B, SCHANDELMAIER S, SUN X, et al. Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications[J]. *BMJ*, 2014, 349: g4539.
- 18 Assmann SF, Pocock SJ, Enos LE, et al. Subgroup analysis and other (mis)uses of baseline data in clinical trials[J]. *Lancet*, 2000, 355(9209): 1064–9.
- 19 Sherry AD, Hahn AW, Mccaw ZR, et al. Differential treatment effects of subgroup analyses in phase 3 oncology trials from 2004 to 2020[J]. *JAMA Netw Open*, 2024, 7(3): e243379.
- 20 Kastrati L, Farina S, Valzgris A, et al. Evaluation of reported claims of sex-based differences in treatment effects across Meta-analyses: a Meta-research study[J]. *BMJ Evid Based Med*, 2026, 31(2): 114–124.
- 21 Ademola A, Thabane L, Adekanye J, et al. The credibility of subgroup analyses reported in stroke trials is low: a systematic review[J]. *Int J Stroke*, 2023, 18(10): 1161–1168.
- 22 Guyatt G, Zhao Y, Mayer M, et al. GRADE guidance 36: updates to GRADE's approach to addressing inconsistency[J]. *J Clin Epidemiol*, 2023, 158: 70–83.

收稿日期: 2026 年 07 月 09 日 修回日期: 2026 年 08 月 14 日

本文编辑: 杨室淞 杨 燕

引用本文: 任相颖, 王宇, 王天堃, 等. 随机对照试验与 Meta 分析中效应修饰可信度评估工具(ICEMAN)的解读[J]. 医学新知, 2026, 36(10): 1081–1090. DOI: 10.12173/j.issn.1004–5511.202607051.

Ren XY, Wang Y, Wang TK, et al. Interpretation of the Instrument for assessing the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and Meta-analysis[J]. *Yixue Xinzhi Zazhi*, 2026, 36(10): 1081–1090. DOI: 10.12173/j.issn.1004–5511.202607051.