COSMIN 研究设计清单：用于患者报告结局测量工具

版本：2019 年 7 月

作者：

Lidwine B Mokkink
Cecilia AC Prinsen
Donald L Patrick
Jordi Alonso
Lex M Bouter
Henrica CW de Vet
Caroline B Terwee

译者：

胡雁
贾凌莹
郑苏娜
徐蕾
成磊
王童瑶
王静

联系方式：

Lidwine B Mokkink 博士
流行病学与生物统计学系
阿姆斯特丹公共卫生研究所
阿姆斯特丹大学医学中心，VUmc
邮政信箱 7057
1007 MB 阿姆斯特丹
荷兰
网站： www.cosmin.nl
电子邮箱：w.mokkink@amsterdamumc.nl

胡雁 教授
护理学院
复旦大学
上海市徐汇区枫林路 305 号
邮政信箱 200032
中国
电话：+86 21 6443 1273
传真：+86 21 6416 1784
电子邮箱： huyan@fudan.edu.cn

目录

缩略语表

CTT:	classical test theory, 经典测量理论
IRT/Rasch:	Item Response Theory and Rasch analyses, 项目反应理论和 Rasch 分析
NA:	not applicable, 不适用
Original CC:	original COSMIN checklist, 初版 COSMIN 清单
PROM:	patient-reported outcome measure, 患者报告结局测量工具
RoB:	Risk of Bias, 偏倚风险；它指的是 COSMIN 偏倚风险清单[2]

前言

建议使用 COSMIN 研究设计清单来设计评估现有患者报告结局测量工具 (PROM) 测量属性的研究。本清单可供研究人员和临床医生或其他设计研究的健康相关专业人员使用，用于评估现有 PROM 的测量属性；或由科学委员会及医学伦理委员会使用，用以对测量属性相关研究方案进行评估；抑或供发表 PROM 测量属性相关研究方案/协议的科学期刊审稿人使用。

COSMIN 研究设计清单基于 COSMIN 清单的初始版本[1,3]，以及最近开发的对于 PROM 的 COSMIN 偏倚风险清单[2]。在反复讨论的基础上，COSMIN 指导委员会决定对研究设计清单进行改编。讨论形式包括面对面会议（LM、CP、HdV 和 CT）和电子邮件讨论（COSMIN 指导委员会全体成员，即所有作者）。

COSMIN 研究设计清单由十个模块组成。第一个模块，即*测量属性研究设计的总体建议*。总体建议模块包含了在设计测量属性相关研究时应考虑的一般标准，且需要注意的是，所有运用本清单的研究都应参考第一个模块。其余模块则针对特定研究给出了可能需要使用到的九个测量属性，即内容效度、结构效度、内部一致性、跨文化效度\测量不变性、测量误差和稳定性、效标效度、构念效度的假设检验和反应度[2,4]。此外，我们在清单中还提供了对现有患者自我报告结局测量工具进行引进及跨文化调适的标准流程。

在这份清单中，每一条目均附有一个新增的 4 级评分量表。该评分量表基于 COSMIN 偏倚风险清单[2]，用以进一步的解释说明条目内容，帮助使用者更好地理解每一步实验设计的选择对研究的方法学质量所造成的影响。这份评分表并非在实践中为您提供一个总体的研究设计评分（即基于最差评分原则），而只是帮助您检查是否已全面考虑到所有重要问题。有关如何设计并分析这些研究，则在 *Measurement in Medicine* 一书中有详细描述[5]；而大多数个别标准的澄清和解释可以在 2 篇使用手册（www.cosmin.nl）以及 COSMIN 偏倚风险清单[6,7]中找到。 另，参考文献及一些对于样本量要求的举例，也包含在 COSMIN 使用手册中[6,7]。

本清单中的条目均与研究在偏倚问题、报告问题或样本量问题上的潜在风险有关。在本文中，每一条目（standard）后都增添了备注（justification），指的是板块（box）的编号（括号内的数字是指具体板块中具体条目的编号）是来自 COSMIN 偏倚风险清单 (RoB)[2]、初版 COSMIN 清单(Original CC)[1]、与样本量有关，或为新增条目。

测量属性研究设计的总体建议

*测量属性研究设计的总体建议*模块与所有测量属性研究相关。PROM 测量属性研究的目的在于评价 PROM（一个或多个方面）的质量。测量属性研究的目的应当明确（即关注什么测量属性）；对 PROM 本身以及研究对象的描述应当清晰。
由于测量属性研究的结果取决于研究的样本，因此 PROM 的质量应在使用该 PROM 的目标人群的样本中确定。

测量属性研究设计的总体建议	很好	良好	模糊	不良	备注
研究目的					
1 提供明确的研究目的，包括： （1）PROM 的名称和版本 （2）目标人群 （3）所关注的测量属性	明确描述了研究目的			未明确描述研究目的	新增条目
患者报告结局测量工具（PROM）					
2 提供所测构念的清晰描述	明确描述了构念			未明确描述构念	偏倚风险清单 框目 1
3 提供 PROM 开发过程的清晰描述，包括描述 PROM 的目标人群	明确描述了开发过程		未明确描述开发过程		偏倚风险清单 框目 1
4 应明确构念的来源：提供一个理论、概念框架（即反映模型或形成模型）、疾病模型或明确合理地定义所测构念	构念的来源明确		构念的来源不明确		偏倚风险清单 框目 1

5	提供对 PROM 结构（包括条目和分量表的数量、指导语、选项）及其计分方法的清晰描述	明确描述了 PROM 的结构和计分方法			未明确描述 PROM 的结构和计分方法	偏倚风险清单 框目 1
6	提供并清晰描述了 PROM 质量的现有证据	明确描述了 PROM 质量的现有证据		未明确描述 PROM 质量的现有证据		新增条目
7	提供对于使用情境的清晰描述*	明确描述了使用情境		未明确描述使用情境		偏倚风险清单 框目 1
目标人群						
8	提供对受试者的纳入和排除标准的明确描述，例如疾病状况、年龄、性别、语言或国家、健康背景（如一般人群、初级护理、医院/康复护理）	明确描述了受试者的纳入排除标准			未明确描述受试者的纳入排除标准	该研究的人口学特征[6]
9	提供对研究对象的抽样方法的明确描述，例如方便、连续或随机抽样	明确描述了研究对象的选样方法		未明确描述选样方法		新增条目
10	描述所选样本是否具有代表性，例如在年龄、性别、疾病重要特征（例如严重程度、疾病状态、持续时间）方面是否能代表 PROM 的目标使用人群	明确描述了研究样本能够代表目标人群	可以认为研究样本能够代表目标人群，但是没有明确地描述	不清楚研究样本是否能够代表目标人群	研究样本无法代表目标人群	偏倚风险清单 框目 1

*使用情境（context）：使用情境可以指 PROM 测量的目的（如用于诊断、评价或预测），也可以指开发 PROM 的特定环境（如在医院或家中使用）或特定的使用方法（如纸张或计算机）。如果 PROM 是为多个使用情境而开发，那么对此也应清楚描述。

内容效度

PROMs 的内容效度评估可以通过询问患者和健康领域专家关于条目、选项和指导语的相关性、全面性以及可理解性的评价来实现。内容效度模块为患者及健康领域专家参与内容效度研究提供了参考标准。

内容效度	很好	良好	模糊	不良	不适用	备注
设计要求						
1 从患者的角度：使用合适的方法评估 （1） 各条目与患者病情体验的相关性 （2） PROM 的全面性 （3） PROM 指导语、条目、选项和回忆期的可理解性	将使用广泛认可或有充分依据的质性研究方法评估这三个方面	将评估这三个方面，但仅使用量性（调查）研究方法；或可以认为使用的方法恰当，但未清晰描述	不清楚患者是否会被询问：每一项条目的相关性、可理解性、全面性；或会否（由于信息不充分等原因）怀疑评估内容效度的方法不恰当	使用的方法不恰当；或患者将不会被问及所有条目的相关性、全面性或可理解性		偏倚风险清单框目 2（1，8，15）
2 从健康领域专家的角度：使用合适的方法评估 （1） 每个条目与所测量的构念相关 （2） PROM 的全面性	将使用广泛认可或合理的质性研究方法评估这两个方面	将评估这两个方面，但仅使用了量性（调查）研究方法；或可以认为使用的方法恰当，但未清晰描述	不清楚专家是否会被询问：每一项条目是否相关；**以及**所有条目组合起来的全面性(即存在有条目未被询问的情况)；或会否（由于信息不充分等原因）怀疑	使用的方法不恰当；或专家将不会被问及所有条目的相关性或全面性	不适用	偏倚风险清单框目 2（22，27）

	评估内容效度的方法不恰当	

3	纳入所有相关学科的专家	将纳入所有相关学科的专家	可以认为将纳入所有相关学科的专家，但未清晰描述	怀疑是否将纳入所有相关学科的专家；或没有纳入相关学科的专家		不适用	偏倚风险清单框目 2（28，23）
4	评估每个条目的受试者、专家数量是否合适 对于质性研究 对于量性（调查）研究	 ≥7 名 ≥50 名	 4-6 名 30-49 名	 ＜4 名或不清楚 ＜30 名或不清楚			偏倚风险清单框目 2（2，9，16，24，29）
5	选用经验丰富的小组会议主持人/访谈者	将选用经验丰富的小组会议主持人/访谈者	小组会议主持人/访谈者经验有限，或将针对这项研究接受过专门的培训	不清楚小组会议主持人/访谈者是否将被培训；或小组会议主持人/访谈者既没有接受过培训也没有经验		不适用	偏倚风险清单框目 2（3，10，17）
6	小组会议/访谈的主题/访谈提纲合适	将选用合适的主题/访谈提纲	可以认为话题/提纲合适，但未清晰描述	不清楚是否将使用访谈提纲；或怀疑访谈提纲是否合适；或没有使用访谈提纲		不适用	偏倚风险清单框目 2（4，11，18）
7	录音并逐字转录小组会议/访谈内容	所有小组会议或访谈内容将被录音并逐字转录	可以认为所有小组会议或访谈都将被录音和逐字转录，但未清晰描述	不清楚是否所有小组会议或访谈都将被录音和逐字转录； 或只有录音而没有逐字转录； 或小组会议或访谈期间只做笔记	没有录音，且没有笔记	不适用	偏倚风险清单框目 2（5，12，19）

分析方法						
8　使用合适的方法分析数据	使用广泛认可或合理的方法	可以认为方法合适，但未清晰描述	不清楚使用什么方法；或怀疑方法是否合适	方法不合适		偏倚风险清单框目 2（6，13，20，25，30）
9　至少有两名研究者参与分析	至少有两名研究者将参与分析	可以认为至少有两名研究者将参与分析，但未清晰描述	不清楚是否有两名研究者参与分析；或只有一名研究者将参与分析		不适用	偏倚风险清单框目 2（7，14，21，26，31）

结构效度

PROM 可以基于反映模型或形成模型 [8-10]。在反映模型中，PROMs 的所有条目都是同一潜在构念的表现形式，这些条目被称为效应指标（effect indicators），各条目之间具有高度相关性及可互换性。而形成模型则正相反，是由模型中的条目共同形成了构念，这些条目之间不需要互相关联。在测量属性研究的设计方案（protocol）中应明确描述 PROM 是基于反映模型或是形成模型。结构效度仅适用于基于反映模型构建的 PROMs。当研究目的是评估多维 PROM 的结构效度时，则需要对整个量表进行整体的因子分析。当研究目的在于评估分量表的单维性时，可以分别对每个分量表进行因子分析。

结构效度	**很好**	**良好**	**模糊**	**不良**	**不适用**	**备注**
统计方法						
1 对于经典测量理论（CTT）:进行验证性因子分析	将进行验证性因子分析	将进行探索性(公)因子分析		将不会使用验证性因子分析和探索性因子分析	不适用	偏倚风险清单框目 3（1）
2 对于 CTT：提供有关如何进行因子分析的明确信息。 例如，选用的软件程序、估计方法、模型拟合指标、是否以及如何进行检验假设等。	提供了有关分析方法的明确信息		提供了有关分析方法的部分信息	不清楚将采用何种分析方法	不适用	初版 COSMIN 清单
3 对于项目反应理论(IRT) /Rasch 分析：选择适合该研究问题的模型	所选模型非常适合研究问题	可以认为所选模型适合研究问题	不清楚所选模型是否适合研究问题	所选模型不适合研究问题	不适用	偏倚风险清单框目 3（2）
4 对于 IRT/Rasch 分析：提供有关如何进行 IRT 或 Rasch 分析的明确信息。例如，选用的软件程序、使用的 IRT 或 Rasch 模型、估计方法、模	提供了有关分析方法的明确信息		提供了有关分析方法的部分信息	不清楚将采用何种分析方法	不适用	初版 COSMIN 清单

型拟合指标、是否以及如何检验假设等				

5	分析中包含足够的样本量（考虑到预估的样本流失和缺失值）	因子分析：条目数的 7 倍且≥100 名	因子分析：至少是条目数的 5 倍且≥100 名； 或至少是条目数的 6 倍但<100 名	因子分析：条目数的 5 倍但<100 名	因子分析：少于条目数的 5 倍		偏倚风险清单框目 3（3）
		Rasch/单参数 IRT 模型：≥200 名受试者	Rasch/单参数 IRT 模型：100-199 名受试者	Rasch/单参数 IRT 模型：50-99 名受试者	Rasch/单参数 IRT 模型：＜50 名受试者		
		双参数 IRT 模型或 Mokken 量表分析：≥1000 名受试者	双参数 IRT 模型或 Mokken 量表分析：500-999 名受试者	双参数 IRT 模型或 Mokken 量表分析：250-499 名受试者	双参数 IRT 模型或 Mokken 量表分析：＜250 名受试者		
6	明确说明了将如何处理缺失的数据、条目等信息	清晰描述了缺失信息的处理方式		未清晰描述缺失信息的处理方式			初版 COSMIN 清单

内部一致性

与结构效度相同，内部一致性仅适用于基于反映模型构建的 PROMs。此外，仅单维（分）量表有必要评估内部一致性，因此，（评估内部一致性之前）应该对研究中的每个量表或分量表进行单维性检验或结构效度评估，抑或提供在可比目标人群的研究中获得的结构效度的证据。

内部一致性	很好	良好	模糊	不良	不适用	备注
设计要求						
1 检查是否每一个（分）量表都具有单维性	证据表明每一个量表或分量表都具有单维性		不清楚是否每一个量表或分量表都具有单维性	量表或分量表不具有单维性		偏倚风险清单框目 4（1）
2 分析中包含足够的的样本量（考虑到预估的样本流失和缺失值）	≥100 名受试者	50-99 名受试者	30-49 名受试者	＜30 名受试者		样本量
3 明确说明了将如何处理缺失的数据、条目等信息	清晰描述了缺失信息的处理方式		未清晰描述缺失信息的处理方式			初版 COSMIN 清单
统计方法						
4 对于定量数据：计算每一个单维（分）量表的 Cronbach's alpha 系数或 Omega 系数	将计算 Cronbach's alpha 系数或 Omega 系数		只计算条目-总分相关系数	将不会计算 Cronbach's alpha 系数和条目-总分相关系数	不适用	偏倚风险清单框目 4（2）

5	对于二分类数据：计算每一个单维（分）量表的 Cronbach's alpha 系数或 KR-20 值	将计算 Cronbach's alpha 系数或 KR-20 值		只计算条目-总分相关系数	将不会计算 Cronbach's alpha 系数或 KR-20 值，且未计算条目-总分相关系数	不适用	偏倚风险清单框目 4（3）
6	对于基于项目反应理论的数据： 计算 θ 的标准误，即 SE(θ)； 或，每个单维（分）量表潜在特质估计值的其他信度系数，如受试者（或项目）分离指数	将计算 SE(θ)或其他信度系数			将不会计算 SE(θ)或其他信度系数	不适用	偏倚风险清单框目 4（4）

跨文化效度\测量不变性

此测量属性旨在评价 PROM 各条目在不同人群中的效能是否相似，例如在不同种族或语言组、不同性别或年龄组、不同疾病人群之间等。因此，跨文化效度\测量不变性的评价需要来自多个组（比如多个语言组）的数据。

跨文化效度\测量不变性	很好	良好	模糊	不良	不适用	备注
设计要求						
1 提供对分组变量的分类方法的明确描述，包括二分类或多分类	明确描述了分组变量的分类方法	可以假定分组变量的分类方法但描述不清晰		未明确描述分组变量的分类方法		偏倚风险清单框目 5（1）
2 提供对于亚组间应保持相同/相似的相关变量的明确描述，比如人口统计学变量或疾病特征	明确描述了相关变量			未明确描述相关变量		偏倚风险清单框目 5（2）
3 分析中包含足够的样本量（考虑到预估的样本流失和缺失值）						偏倚风险清单框目 5（3）
- 回归分析或基于 IRT/Rasch 的分析：	每组 200 名受试者	每组 150-199 名受试者	每组 100-149 名受试者	每组<100 名受试者		
- 多组验证性因子分析（MGCFA）：	条目数的 7 倍且≥100 名受试者	条目数的 5 倍且＞100 名受试者；或条目数的 5-7 倍且＜100 名受试者	条目数的 5 倍但＜100 名受试者	＜条目数的 5 倍		

统计方法						
4	对于经典测量理论(CTT)：进行多组验证性因子分析（MGCFA）	将进行多组验证性因子分析			将不会进行验证性因子分析	不适用	偏倚风险清单框目 5（2）
5	对于经典测量理论(CTT)：提供有关如何进行因子分析的明确信息。例如，选用的软件程序、估计方法、模型拟合指标、是否以及如何进行检验假设等。	提供了有关分析操作方法的明确信息		提供了有关分析操作方法的部分信息	不清楚分析的操作方法	不适用	偏倚风险清单框目 5（2）
6	对于 IRT/Rasch 分析：进行项目功能差异分析（DIF）	将进行项目功能差异分析			将不会进行项目功能差异分析	不适用	偏倚风险清单框目 5（2）
7	对于 IRT/Rasch 分析：提供有关如何进行 IRT 或 Rasch 分析的明确信息。例如，选用的软件程序、使用的 IRT 或 Rasch 模型、估计方法、模型拟合指标、是否以及如何检验假设等	提供了有关分析方法的明确信息		提供了有关分析方法的部分信息	不清楚将采用何种分析方法	不适用	偏倚风险清单框目 5（2）
8	明确说明了将如何处理缺失的数据、条目等信息	清晰描述了缺失信息的处理方式		未清晰描述缺失信息的处理方式			初版 COSMIN 清单

测量误差和稳定性

测量误差和稳定性的研究设计和数据收集方式相同。一般需要在所测构念稳定的一组人中进行两次测量。因为这两种测量属性的研究设计和数据收集方式相同，只有统计参数不同，所以我们在同一个模块里介绍。我们非常鼓励研究人员在报告稳定性参数外，另外报告测量误差。

测量误差与稳定性	很好	良好	模糊	不良	不适用	备注
设计要求						
1 至少进行两次测量	至少进行两次测量			只进行一次测量		初 版 COSMIN 清单
2 确保测量方法将独立实施	测量方法将独立实施	可以认为测量方法将是独立实施的	不清楚测量方法是否将独立实施	测量方法将不是独立实施的		新增条目
3 在测量间隔期，确保受试者的待测构念将是稳定的	有证据支持受试者的待测构念将是稳定的	可以认为受试者的待测构念将是稳定的，但没有明确提供证据	不清楚受试者的待测构念是否稳定	受试者的待测构念不稳定		偏倚风险清单框目 6/7（1）
4 确保两次测量之间的时间间隔长短适宜，既能够防止受试者回忆，且确保受试者稳定	时间间隔将是适当的		不清楚时间间隔是否将是适当的，或没有描述时间间隔	时间间隔不适当		偏倚风险清单框目 6/7（2）
5 确保测量前后的测量情境将相似，如测量方式、测量环境及指导语等	有证据支持测量情境将是相似的	可以认为测量情境将是相似的，但没有明确证据支持	不清楚测量情境是否将会相似	测量情境将是不相似的		偏倚风险清单框目 6/7（3）
6 分析中包含足够的的样本量（考虑到预估的样本流失和缺失值）	≥100 名受试者	50-99 名受试者	30-49 名受试者	＜30 名受试者		样本量

测量误差的统计方法						
7 对于定量数据:计算测量标准误(standard error of measurement, SEM)，最小可测变化值(smallest detectable change,SDC)或一致性限度(limits of agreement, LoA)	将计算 SEM、SDC 或 LoA，并明确描述对应的模型或公式*	将计算 SEM 或 SDC，但不会描述 SEM 或 SDC 的模型或公式，或这不是最佳的选择**		将基于 Cronbach's alpha 系数或基于另一人群的标准差计算 SEM	不适用	偏倚风险清单框目 7（4）
8 对于二分类/多分类/有序数据：计算一致性百分比	将计算阳性和阴性一致性百分比	将计算一致性百分比		将不会计算一致性百分比	不适用	偏倚风险清单框目 7（5）
9 明确说明如何处理缺失的数据、条目等信息	清晰描述缺失信息将如何处理		未清晰描述缺失信息的处理方式			初版 COSMIN 清单
稳定性的统计方法						
7 对于定量数据：计算组内相关系数（ICC）	将计算 ICC,并明确描述 ICC 的模型或公式*	将计算 ICC，但 ICC 的模型或公式没有被描述或不是最佳的**	将计算 Pearson 或 Spearman 相关系数	将不会计算 ICC 或 Pearson 或 Spearman 的相关关系	不适用	
8 对于二分类/多分类/有序数据：计算 Kappa 值	将计算 Kappa 值			将不会计算 Kappa 值	不适用	
9 对于有序数据：计算加权 Kappa 值	将计算加权 Kappa 值并描述其加权方式		将计算未加权的 Kappa 值,或不描述		不适用	
10 明确说明了将如何处理缺失的数据、条目等信息	清晰描述了缺失信息的处理方式		未清晰描述缺失信息的处理方式			初版 COSMIN 清单

*适当地选择和描述将要计算的 ICC 的模型（即单向随机效应模型或双向随机或混合效应模型）、类型（即用于单次或多次测量）和定义（即用于一致性或绝对一致）[11]； **ICC 公式与研究问题不一致

效标效度

由于 PROMs 测量的只能是由患者自我报告的结局，所以这些测量方法不存在严格意义上的金标准。唯一的区别是，比较同一 PROM 的不同测量方法或版本，也有优劣之分。

效标效度	很好	良好	模糊	不良	不适用	备注
设计要求						
1 说明拟定的标准是否可以被视为合理的金标准	有证据支持该标准是恰当的金标准	可以认为该标准是恰当的金标准，但没有明确提供证据	不清楚该标准是否是恰当的金标准	该金标准不恰当		初版 COSMIN 清单
2 分析中包含足够的的样本量（考虑到预估的样本流失和缺失值）	人数最少的组别中受试者≥50 名	人数最少的组别中受试者有 30-50 名	人数最多的组别中受试者＜30 名			样本量
3 合理制订并使用所关注的 PROM 和金标准的评估时间表	PROM 和金标准将在同一时间评估	PROM 和金标准将不会同时评估，但可以认为受试者在测量间隔期不会有变化	PROM 和金标准将不会同时评估，但不清楚受试者是否会有变化	PROM 和金标准将不会同时评估，且预计受试者会有变化		新增条目
统计方法						
4 对于定量数据：计算相关性或 AUC	将计算相关性或 AUC			未计算相关性或 AUC	不适用	偏倚风险清单框目 8（1）

5 对于二分类数据：计算敏感性和特异性	将计算敏感性和特异性			未计算敏感性 不适用和特异性	偏倚风险清单框目 8（2）
6 明确说明了将如何处理缺失的数据、条目等信息	清晰描述了缺失信息的处理方式		未清晰描述缺失信息的处理方式		初版 COSMIN 清单

AUC=受试者工作特征曲线下的面积

构念效度的假设检验

由于 PROMs 不存在金标准，检验 PROMs 效度的常用方法是测量以下假设：1）与其他高质量的结局测量工具的预期关系（A 部分），和/或 2）相关组别之间的预期差异（B 部分）。在评估 PROM 的构念效度时，提前定义假设非常重要，如此可以使研究者在数据收集和分析后得出无偏倚的结论。

构念效度的假设检验 **A. 与其他测量工具进行比较（聚合效度）** ***设计要求***	**很好**	**良好**	**模糊**	**不良**	**不适用**	**备注**
1 对研究所关注的 PROM 和其他结局测量工具之间的预期关系提出假设	提出的假设包括预期相关性的方向和程度。		提出的假设模糊或没有提出假设，但可以假定预期关系	不清楚预期关系		初版 COSMIN 清单
2 提供对被比较工具所测量构念的清晰描述	被比较工具测量的构念清晰			被比较工具测量的构念不清晰		偏倚风险清单框目 9a（1）
3 被比较工具的测量属性充分	在与目标人群相似的群体中，被比较工具的测量属性充分	被比较工具的测量属性充分，但是不清楚是否适用于该研究群体	在任意目标人群中，有被比较工具测量属性的部分信息（或有参考文献）	没有被比较工具的测量属性信息，或者有证据表明被比较工具测量属性不充分		偏倚风险清单框目 9a（2）
4 分析中包含足够的样本量（考虑	≥100 名受试者	50-99 名受试者	30-49 名受试者	＜30 名受试者		样本量

到预估的样本流失和缺失值）						
5 合理制订并使用所关注的 PROM 和被比较工具的评估时间表	PROM 和被比较工具将在同一时间评估	PROM 和被比较工具将不会同时评估，但可以认为受试者在测量间隔期不会发生改变	PROM 和被比较工具将不会同时评估，但不清楚受试者是否会发生改变	PROM 和被比较工具将不会同时评估，且预计受试者会发生改变		新增条目
统计方法						
6 验证假设的统计方法合适	统计方法是合适的	可以认为统计方法是合适的	统计方法不是最合适的	统计方法不合适		偏倚风险清单框目 9a（3）
7 明确说明了将如何处理缺失的数据、条目等信息	清晰描述了缺失信息的处理方式		未清晰描述缺失信息的处理方式			初版 COSMIN 清单

B. 亚组比较 （区分效度/已知组别效度）	很好	良好	模糊	不良	不适用	备注
设计要求						
1 提出关于亚组之间平均差的假设	提出的假设包括平均差的预期方向和大小		提出的假设模糊或没有提出假设，但可以假定预期关系	不清楚预期关系		初版 COSMIN 清单
2 提供对亚组重要特征的清晰描述（例如疾病或人口统计学特征）	充分描述了亚组的重要特征	充分描述了各亚组大部分重要特征	未充分描述或未描述各亚组的重要特征			偏倚风险清单框目 9b（5）
3 分析中包含足够的样本量（考虑到预估的样本流失和缺失值）	≥100 名受试者	50-99 名受试者	30-49 名受试者	＜30 名受试者		样本量
统计方法						
4 验证假设的统计方法合适	统计方法是合适的	可以认为统计方法是合适的	统计方法不是最合适的	统计方法不合适		偏倚风险清单框目 9b（6）
5 明确说明了将如何处理缺失的数据、条目等信息	清晰描述了缺失信息的处理方式		未清晰描述缺失信息的处理方式			初版 COSMIN 清单

反应度

反应度可用于表明纵向有效性。当有金标准时，可以使用本框中 A 部分的标准方法；当检验 PROMs 的分数变化量和被比较测量工具的分数变化量之间关系的假设时，可以使用 B 部分；当检验不同亚组分数变化量差异的假设时，可以使用 C 部分；当检验干预前后 PROMs 分数变化量差异的假设时，可以使用 D 部分。在评估 PROM 的反应度时，事先确定假设非常重要，这样可以使研究者在数据收集和分析后得出无偏倚的结论。

反应度 ***A. 效标方法（与金标准比较）***	**很好**	**良好**	**模糊**	**不良**	**不适用**	**备注**
设计要求						
1 拟定的标准能够被视为合理的金标准	有证据支持该标准是恰当的金标准	可以认为该标准是恰当的金标准，但没有明确提供证据	不清楚该标准是否是恰当的金标准	该金标准不恰当		初版 COSMIN 清单
2 合理制订并使用所关注的 PROM 和金标准的评估时间表	PROM 和金标准将在同一时间评估	PROM 和金标准将不会同时评估，但可以认为受试者在测量间隔期不会发生改变	PROM 和金标准将不会同时评估，但不清楚受试者是否会发生改变	PROM 和金标准将不会同时评估，且预计受试者会发生改变		新增条目
3 第一次和第二次测量之间的时间间隔适当	时间间隔适当			时间间隔不适当		新增条目
4 描述在测量间隔期可能发生的任何事件（如干预，疾病进展，其他相关事件）	充分描述了测量间隔期可能发生的任何事件（如治疗）		不清楚或没有描述在测量间隔期可能发生的事件			初版 COSMIN 清单

5 确保一定比例的受试者可能在所测量的构念上发生改变（改善或退化）	有证据支持部分受试者可能会发生改变	可以认为部分受试者会改变，但没有提供证据	不清楚部分受试者是否会改变	受试者可能没有改变		初版 COSMIN 清单
6 分析中包含足够的样本量（考虑到预估的样本流失和缺失值）	人数最少的组别中受试者≥50 名	人数最少的组别中受试者有 30-50 名	人数最多的组别中受试者＜30 名			样本量
统计方法						
7 对于定量数据：计算分数变化量之间的相关性或者受试者工作特征曲线下的面积（Area Under the Receiver Operating Curve, AUC）	将计算 AUC 或相关性			将不会计算 AUC 或相关性	不适用	偏倚风险清单框目 10a（1）
8 对于二分类数据：计算敏感性和特异性	将计算敏感性和特异性			将不会计算敏感性和特异性	不适用	偏倚风险清单框目 10a（2）
9 明确说明了将如何处理缺失的数据、条目等信息	清晰描述了缺失信息的处理方式		未清晰描述缺失信息的处理方式			初版 COSMIN 清单

B. 构念方法（假设检验：与其他测量工具进行比较）						
	很好	**良好**	**模糊**	**不良**	**不适用**	**备注**
设计要求						
1 对所研究的PROM的变化分数和其他结局测量工具（的变化分数）之间的预期关系提出假设	提出的假设将包括预期相关性的方向和程度。		提出的假设模糊或将不会提出假设，但可以假定预期关系	不清楚预期关系		初版 COSMIN清单
2 提供对被比较工具所测量构念的清晰描述	被比较工具测量的构念清楚			被比较工具测量的构念不清楚		偏倚风险清单框目 10b（4）
3 被比较工具的测量属性充分	在与目标人群相似的群体中，被比较工具的测量属性充分	被比较工具的测量属性充分，但是不清楚是否适用于该研究群体	在任意目标人群中，有被比较工具测量属性的部分信息（或有参考文献）	没有被比较工具的测量属性信息，或者有证据表明被比较工具测量属性不充分		偏倚风险清单框目 10b（5）

4 合理制订并使用所关注的 PROM 和被比较工具的评估时间表	PROM 和被比较工具将在同一时间评估	PROM 和被比较工具将不会同时评估，但可以认为受试者在测量间隔期不会发生改变	PROM 和被比较工具将不会同时评估，但不清楚受试者是否会发生改变	PROM 和被比较工具将不会同时评估，且预计受试者会发生改变		新增条目
5 第一次和第二次测量之间的时间间隔适当	时间间隔适当			时间间隔不适当		新增条目
6 描述在测量间隔期可能发生的任何事件（如干预，疾病进展，其他相关事件）	充分描述了测量间隔期可能发生的任何事件（如治疗）		不清楚或没有描述在测量间隔期可能发生的事件			初版 COSMIN 清单
7 确保一定比例的受试者可能在所测量的构念上发生改变（改善或退化）	有证据支持部分受试者可能会发生改变	可以认为部分受试者会发生改变，但没有提供证据	不清楚部分受试者是否会发生改变	受试者可能没有发生改变		初版 COSMIN 清单
8 分析中包含足够的样本量（考虑到预估的样本流失和缺失值）	≥100 名受试者	50-99 名受试者	30-49 名受试者	＜30 名受试者		样本量
统计方法						
9 验证假设的统计方法充分且合适	统计方法合适	可以认为统计方法合适	统计方法不是最合适的	统计方法不合适		偏倚风险清单框目 10b（6）
10 明确说明了将如何处理缺失的数据、条目等信息	清晰描述了缺失信息的处理方式		未清晰描述缺失信息的处理方式			初版 COSMIN 清单

C. 构念方法(假设检验：亚组比较)						
	很好	**良好**	**模糊**	**不良**	**不适用**	**备注**
设计要求						
1 预先（即在数据收集之前）提出关于亚组变化分数之间差异的假设	提出的假设包括变化分数之间的预期差异		提出的假设模糊或没有提出假设，但可以假定预期差异	不清楚预期差异		初 版 COSMIN 清单
2 提供对亚组重要特征的清晰描述（例如疾病或人口统计学特征）	充分描述了各亚组的重要特征	充分描述了各亚组大部分的重要特征	未充分描述或未描述各亚组的重要特征			偏倚风险清单 10c（8）
3 第一次和第二次测量之间的时间间隔适当	时间间隔适当			时间间隔不适当		新增条目
4 描述在测量间隔期可能发生的任何事件（如干预，疾病进展，其他相关事件）	充分描述了测量间隔期可能发生的任何事件（如治疗）		不清楚或没有描述在测量间隔期可能发生的事件			初 版 COSMIN 清单
5 确保一定比例的受试者可能在所测量的构念上发生改变（改善或退化）	有证据支持部分受试者可能会发生改变	可以认为部分受试者会改变，但没有提供证据	不清楚部分受试者是否会改变	受试者可能没有改变		初 版 COSMIN 清单
6 分析中包含足够的样本量（考虑到预估的样本流失和缺失值）	≥100 名受试者	50-99 名受试者	30-49 名受试者	＜30 名受试者		样本量

统计方法						
7 验证假设的统计方法充分且合适	统计方法合适	可以认为统计方法合适	统计方法不是最合适的	统计方法不合适		偏倚风险清单框目 10c（9）
8 明确说明了将如何处理缺失的数据、条目等信息	清晰描述了缺失信息的处理方式		未清晰描述缺失信息的处理方式			初 版 COSMIN 清单

D. 构念方法(假设检验：干预前后比较)						
	很好	**良好**	**模糊**	**不良**	**不适用**	**备注**
设计要求						
1 对于干预前后的预期变化，预先（即在数据收集前）提出具有挑战性的假设	提出的假设包括预期变化		提出的假设模糊或没有提出假设，但可以假定预期变化	不清楚预期变化		初 版 COSMIN 清单
2 对干预措施进行充分描述，以便重复实验，包括如何和何时实施这些干预措施	充分描述了干预措施		未充分描述干预措施	未描述干预措施		偏倚风险清单框目 10d（11）
3 第一次和第二次测量之间的时间间隔适当	时间间隔适当			时间间隔不适当		新增条目
4 描述在测量间隔期可能发生的任何事件（如干预，疾病进展，其他相关事件）	充分描述了测量间隔期可能发生的任何事件（如治疗）		不清楚或没有描述在测量间隔期可能发生的事件			初 版 COSMIN 清单

5 确保一定比例的受试者可能在所测量的构念上发生改变（改善或退化）	有证据支持部分受试者可能会发生改变	可以认为部分受试者会发生改变，但没有提供证据	不清楚部分受试者是否会发生改变	受试者可能没有发生改变		初版 COSMIN 清单
6 分析中包含足够的样本量（考虑到预估的样本流失和缺失值）	≥100 名受试者	50-99 名受试者	30-49 名受试者	＜30 名受试者		样本量
统计方法						
7 验证假设的统计方法合适	统计方法合适	可以认为统计方法合适	统计方法不是最合适的	统计方法不合适		偏倚风险清单框目 10d（12）
8 明确说明了将如何处理缺失的数据、条目等信息	清晰描述了缺失信息的处理方式		未清晰描述缺失信息的处理方式			初版 COSMIN 清单

跨文化调适流程

跨文化调适不是一种测量属性，而是 PROM 新版本开发阶段的一部分。高质量的跨文化调适将可能形成一个高质量的翻译版 PROM。该模块旨在提供用于评估跨文化调适流程质量的标准。当后续测量翻译后的 PROM 的跨文化效度时，请参考前文的“跨文化效度\测量不变性”板块，并使用“语言”作为分组变量。

跨文化调试流程	**很好**	**良好**	**模糊**	**不良**	**备注**
设计要求					
1 描述开发 PROM 时使用的原始语言（the original language）以及 PROM 翻译后的靶语言(the target language)。如若翻译所对照的源语言(the source language)并非原始语言（而是他国语言），则所使用的源语言（即区别于原始语言的他国语言版本）也应进行介绍。	将描述原始语言、源语言、靶语言（若源语言即为原始语言，则仅描述原始语言和靶语言）			将不会描述翻译时所对照的源语言（源语言可能是原始语言）	初版 COSMIN 清单
2 确保各条目将进行正向翻译及回向翻译	多次正向翻译及多次回向翻译	多次正向翻译，但仅一次回向翻译	正向翻译及回向翻译各一次	仅一次正向翻译	初版 COSMIN 清单
3 确保两名正向翻译者的母语均为靶语言	两名正向翻译者的母语均为靶语言		只有一名正向翻译者母语为靶语言	两名正向翻译者母语均不是靶语言	初版 COSMIN 清单
4 确保其中一名正向翻译者对 PROM 所涉及的疾病以及 PROM 所测量的构念有专业储备；同时，其他正向翻译者对 PROM 所测量的构念不了解。	其中一位正向翻译者熟悉该疾病和构念，其他翻译者不了解	不清楚两位正向翻译者对该疾病和构念方面的专业储备	两位正向翻译者都是该疾病或构念方面的专家；或都对该领域不了解		初版 COSMIN 清单

5	确保两名回向翻译者母语均为原始语言或源语言	两位回向翻译者的母语均为源语言		只有一名回向翻译者母语为源语言	两名回向翻译者母语均不是源语言	初版 COSMIN 清单
6	确保两名回向翻译者均不了解 PROM 中包含的疾病以及所测量的构念	两位回向翻译者均不了解所包含的疾病及所测量的构念	不清楚两位回向翻译者是否有所包含疾病及所测量构念的专业知识			新增条目
7	确保各翻译者的翻译工作是独立完成的	各翻译者将独立完成翻译	可以认为各翻译者独立完成翻译	不确定各翻译者是否将独立完成翻译	翻译者们将不会独立完成翻译	初版 COSMIN 清单
8	清晰描述对原始语言版本及所翻译的靶语言版本之间差异的解决方法	充分说明如何解决翻译者间差异	仅粗略说明或不说明如何解决翻译者间差异			初版 COSMIN 清单
9	确保跨文化调适将由一个（包含原始开发者的）委员会进行审核	跨文化调适将被委员会（包含除翻译者以外的人员，如原始开发者）审核	跨文化调适将不会被委员会审核			初版 COSMIN 清单
10	写一份阐述跨文化调适流程的反馈报告	将撰写一份反馈报告		将不会撰写反馈报告		新增条目

11 进行预实验（比如认知访谈）以检验： （4） 各条目与患者病情体验之间的相关性 （5） PROM 的全面性 （6） PROM 指导语、条目、选项和回忆期的可理解性	将使用被广泛认可或合理的质性研究方法来评估这三个方面	将评估三个方面。但仅使用量性（调查）研究方法；或可以认为使用的方法适当，但未进行清晰描述	不清楚受试者是否将会被询问： 每一项条目是否相关； **以及**可理解性； **以及**所有条目组合起来的全面性，或是否怀疑评估 PROM 全面性的方法不恰当	使用的方法不恰当，或受试者将不会被问及所有条目的相关性、全面性或可理解性	偏倚风险清单 框目 1
12 在能够代表目标人群的患者群体中开展预实验	研究将在能够代表目标人群的样本中开展	可以认为研究将在能够代表目标人群的样本中开展	不确定研究是否将在能够代表目标人群的样本中开展	研究将不会在能够代表目标人群的样本中开展	偏倚风险清单 框目 1

参考文献

1. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19(4):539-49. doi: 10.1007/s11136-010-9606-8 [doi]
2. Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res 2018;27(5):1171-79. doi: 10.1007/s11136-017-1765-4
3. Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21(4):651-57. doi: 10.1007/s11136-011-9960-1 [doi]
4. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27(5):1147-57. doi: 10.1007/s11136-018-1798-3
5. de Vet HC, Terwee CB, Mokkink L, et al. Measurement in Medicine: a practical guide: Cambridge University Press 2010.
6. Mokkink LB, Vet HC, Prinsen CA, et al. COSMIN methodology for systematic reviews of Patient- Reported Outcome Measures (PROMs) - user manual 2018. Available from: www.cosmin.nl.
7. Terwee CB, Prinsen CA, de Vet HCW, et al. COSMIN methodology for assessing the content validity of Patient-Reported Outcome Measures (PROMs). User manual., 2018. Available from: www.cosmin.nl.
8. Fayers PM, Hand DJ. Factor analysis, causal indicators and quality of life. Qual Life Res 1997;6(2):139-50.
9. Fayers PM, Hand DJ, Bjordal K, et al. Causal indicators in quality of life research. Qual Life Res 1997;6(5):393-406.
10. Streiner DL. Being inconsistent about consistency: when coefficient alpha does and doesn't matter. J Pers Assess 2007;80:217-22.
11. McGraw KOW, S.P. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1996;1:30-46.