

基于 COSMIN 指南的健康老龄化测评工具的系统评价

金锦珍¹,王岩¹,陈冉²,王子恒²,李春玉²

摘要:**目的** 评价健康老龄化测评工具,为老龄化研究提供参考。**方法** 在中国知网、万方、维普、SinoMed、Web of Science、PubMed 数据库中检索,基于 COSMIN 指南对纳入研究进行评价。**结果** 纳入的 14 项研究共包含 14 种健康老龄化测评工具,存在一定的偏倚风险,推荐等级均为 B 级;9 项研究缺少比较拟合指数、塔克-刘易斯指数等涉及结构效度的类似指标;7 项研究未进行多组因素分析或者项目功能差异分析;12 项研究的组内相关系数或加权 *Kappa* 值未报告;12 项研究未定义假设。**结论** 应继续完善健康老龄化测评工具,依据测评工具质量评价的指南,规范研究设计和报告,为筛选出优质的测量工具以更准确地评估老龄化水平提供基础。

关键词: 老年人; 健康老龄化; COSMIN; 指南; 测量学性能; 评估工具; 质量评价

中图分类号: R47; R161.7 **DOI:** 10.3870/j.issn.1001-4152.2023.11.097

Systematic review of healthy aging measurement tools based on COSMIN guideline Jin Jinzhen, Wang Yan, Chen Ran, Wang Ziheng, Li Chunyu. School of Medicine, Yanbian University, Yanji 133002, China

Abstract: **Objective** To evaluate health aging measurement tools and provide a reference for aging research. **Methods** Such databases as CNKI, Wanfang, VIP, SinoMed, Web of Science, and PubMed, were searched to retrieve publications on healthy aging measurement tools. The included tools were evaluated based on Consensus Based Standards for the Selection of Health Measurement Instruments (COSMIN) guidelines. **Results** A total of 14 studies totaling 14 healthy aging measurement tools were included; all studies were flawed with a certain bias and graded level B; 9 studies lacked comparative fit index, Tucker-Lewis index and other similar indicators related to structure validity; 7 studies did not conduct multi-group factor analysis or differential item functioning analysis; intraclass correlation coefficient (*ICC*) or weighted *Kappa* values were not reported in 12 studies; 12 studies did not define hypotheses. **Conclusion** The measurement tools of healthy aging should be improved, and the research design and reporting should be standardized according to the quality evaluation guidelines of the measurement instruments, so as to provide the basis for selection of high-quality assessment tools to assess the level of aging more accurately.

Key words: the aged; healthy aging; COSMIN; guideline; measurement properties; assessment tools; quality appraisal

我国老年人口规模长期位居世界首位,预计我国 65 岁及以上老年人口规模和比例将在 2040 年前超过所有发达国家总和,2050 年我国老年人口比例将达到约 37.8%^[1]。“十四五”时期是我国人口老龄化从相对慢速转为相对快速的重要变轨期^[2],有效应对我国人口老龄化,事关群众福祉、社会和谐与国家发展,对于全面建设社会主义现代化国家具有重要意义^[3]。健康老龄化是全球应对老龄化的重要理念和战略思想^[4],也是我国应对老龄化的长期战略任务^[5]。为明确健康老龄化水平,从而为老龄事业发展提供依据,目前国内外研究者已经开发出多种测评健康老龄化的工具。但相关量表的评分方法、条目内容等各异的,对健康老龄化测评工具的测量性能评价仍然缺乏,这可能会影响健康老龄化评估工作的开展。本研究采用选择健康测量工具的共识标准(Consensus-Based Standards for the Selection of Health Measurement Instruments, COSMIN)对健康老龄化测评工具进行评价,旨在筛选出质量较高的健康老龄化测

评工具,为有效评估健康老龄化提供循证依据。

1 资料与方法

1.1 纳入与排除标准 纳入标准:①研究对象为成年人;②研究内容涉及健康老龄化测评工具的测量学性能评价;③研究类型为测评工具的开发、验证、调适等;④至少对测评工具的任意 1 项测量属性(反应度、组内相关系数、结构效度、内容效度、内部一致性等)进行评价;⑤中、英文文献。排除标准:①将测评工具仅应用于结局指标的测量;②将健康老龄化测评工具作为其他患者报告结局测量的效度检验标准的研究;③文献类型为综述、会议、评论、社论等;④不能获取全文或重复发表。

1.2 检索策略 采用 COSMIN 推荐的研究内容、研究对象、工具类型及量表属性相结合的检索策略,以健康老龄化;工具,问卷,量表,效度,信度,反应度,克朗巴赫系数,组内相关系数,结构效度,内容效度,内部一致性为中文检索词。healthy aging; tool, screening, questionnaire, scale, validity, reliability, responsiveness, Cronbach, ICC, structure validity, content validity, internal consistency 为英文检索词。选定时限为自建库至 2022 年 8 月,在中国知网、万方数据、维普网、SinoMed、Web of Science、PubMed 数据库中

作者单位:延边大学 1. 医学院 2. 护理学院(吉林 延吉,133002)

金锦珍:女,博士在读,副教授

通信作者:李春玉,chyli@ybu.edu.cn

收稿:2023-01-08;修回:2023-03-18

检索。并追溯检索纳入研究的参考文献,以更全面补充有关研究。

1.3 文献筛选与资料提取 由2名研究者独立筛选文献、提取资料并交叉核对。通过阅读标题、摘要及全文,剔除不符合纳入标准的文献。资料提取主要内容包括:第一作者、发表时间、量表名称、地区、样本量、研究对象年龄及来源、量表维度、维度数与条目数、评分方法、完成时间、重测时间、量表测量学性能(内容效度、结构效度、效标效度、跨文化效度、假设检验、测量误差、稳定性、反应性、内部一致性)。结果不一致时,请第3位评价者进行讨论,所有作者均参与会议并达成最终共识。

1.4 评价步骤 由2名评价者采用COSMIN系统综述指南(2018版)^[6]对量表的测量特性及研究的方法学质量进行独立评价,并交叉核对结果,如遇分歧则请第3位评价者协商。

1.4.1 偏倚风险评价 COSMIN偏倚风险评价清单^[7]的译本,评价纳入研究的偏倚风险。此清单对患者自我报告结局测量工具(Patient Reported Outcome Measures,PROMs)的开发、内容效度、结构效度、内部一致性、跨文化效度/测量等同性、信度、测量误差、效标效度、结构效度的假设检验、反应度共10项内容进行评价。以“最低分计数”为原则,采用4点评分法(非常好、合格、有问题的、不合格)对每项内容中的所有条目进行评价,最低得分即为该项内容的最终得分。对于存在“不适用”选项的特定研究,该标准不被考虑在最低分计数。

1.4.2 测量属性评价 根据www.cosmin.nl提供的《更新版良好测量学属性的标准》,评价纳入研究量表的测量属性,并整合各量表测量特性的评价结果。

该标准用于评价量表的结构效度、内部一致性、信度、测量误差、结构效度的假设检验、跨文化效度/测量等同性、效标效度、反应度。评价选项设为:足够(+)、不足(-)及不确定(?)3个等级。若某测量特性在各研究中均为足够(+)、不足(-)及不确定(?),则此特性的总体等级也分别为足够(+)、不足(-)及不确定(?)。若某测量特性在各研究中获得评价不一致,且无法解释造成评价不一致的原因,则此测量特性的总体评级为不一致(±)。

1.4.3 证据等级评价 运用COSMIN改良版的推荐分级、评估、发展与评估(The Grading of Recommendations Assessment Development and Evaluation,GRADE)^[8]综合评定量表测量特性。首先假定各个测量特性均为高质量,再根据偏倚风险、不一致性、间接性、不精确性4个因素,对各个测量特性的质量进行降级。最后,依据证据客观评价结果,最终确定高、中、低、极低4个级别的推荐等级。若量表的内容效度为“足够”,且内部一致性证据等级不低于“低”,则量表可推荐使用,推荐强度为A级;若量表具有应用的潜力,但仍需要开展更多研究,以进一步评估其测量特性,则其推荐强度为B级;若有“高质量”证据证明量表内容效度“不充分”,则量表不建议使用,其推荐强度为C级。

2 结果

最终纳入的14篇文献包括6个英文版量表^[9-14]和8个中文版量表^[15-22]。中文量表包含我国学者开发的本土量表6个^[15,17-18,20-22]和汉化及文化调适量表2个^[16,19]。基本情况见表1。纳入健康老龄化测评工具的偏倚风险评价结果见表2,测量属性评价结果见表3。

表1 纳入健康老龄化测评工具的基本情况

文献	量表名称	样本来源地	样本量	研究对象 年龄(岁)	条目/ 维度数	维度或领域	评分方法
Sanchez-Niubo等 ^[9] 2021	全球老龄化常用量表	全球	343915	18~114	41/7	认知;心理症状;活力;感觉功能;运动/灵活性; 日常生活活动;工具性日常生活活动	未知
Lee等 ^[10] 2017	成功老龄化简表	美国	550	≥65	16/4	主动参与;健康资源;积极心态;重要的人际关系	Likert 4级
Jaspers等 ^[11] 2017	健康老龄化评分	荷兰	3527	65~85	未知/7	慢性疾病;心理健康(CES-D);认知功能(MMSE); 身体功能(bADL及iADL);疼痛;社会支持;生 活质量	各领域不同
Thanakwang等 ^[12] 2014	与文化相关的泰国成年 人积极老龄化综合量表	泰国	500	60~96	35/7	自力更生;积极参与社会;发展灵性智慧;建立经 济保障;保持健康的生活方式;积极学习;加强家 庭关系,确保晚年生活得到照顾	Likert 4级
Cyarto等 ^[13] 2013	健康老龄化测试	澳大利亚	419	49~100	22/4	活动;体质量与营养;社会支持;乐观	未报告
Thiamwong等 ^[14] 2008	健康老龄化工具	泰国	403	60~95	35/9	自力更生;积极参与社会;发展灵性智慧;建立经 济保障;保持健康的生活方式;积极学习;加强家 庭关系,确保晚年生活得到照顾	Likert 5级
李嘉佳 ^[15] 2020	积极老龄化测量工具指 标体系评分表	甘肃	596	>60	23/4	健康;参与;安全;支持环境	Likert 5级
吴凡等 ^[16] 2019	汉化版健康老龄化量表	北京	254	≥60	35/9	生活充足且简单;接受年老;压力管理;拥有社会 关系及支持;行善助人;自我照护;躯体功能正 常;认知功能正常;社交参与	Likert 5级
童佩等 ^[17] 2019	成功老龄化量表	广东	600	60~100	53/6	健康自主;经济保障;生活适应;学习层面;社会 和亲友关系;家庭层面	Likert 4级
Du ^[18] 2019	中国老年人生活方式评 价系统测试	陕西	482	69~80	62/3	身体健康指标;心理健康指标;社会福祉指标	Likert 5级

续表 1 纳入健康老龄化测评工具的基本情况

文献	量表名称	国家/地区	样本量	研究对象 年龄(岁)	条目/ 维度数	维度或领域	评分方法
程彦伶 ^[19] 2014	中文版成功老龄量表	山东	181	60~94	20/5	内心和功能性应对机制;生命存在;内省的老年卓越;怀旧的老年卓越;精神性	Likert 5 级
任立山 ^[20] 2014	成功老龄化他评量表	浙江	300	65~93	25/5	目前慢性病情况;躯体功能;认知功能;社会功能;主观幸福感	Likert 5 级
胡敏 ^[21] 2012	积极老龄化测评问卷	江西	200	>60	21/4	身体活力;生活满意;家庭支持;积极参与	Likert 5 级
刘恒 ^[22] 2010	老年人口健康评价量表	江苏	849	≥60	39/5	躯体机能;生活自理功能;情绪性格;记忆功能;社会适应性	Likert 5 级

注:CES-D,流调中心抑郁量表;bADL,基础性日常生活活动能力;iADL,工具性日常生活活动能力;MMSE,简易精神状态量表。

表 2 纳入健康老龄化测评工具的偏倚风险评价

文献	PROMs 的开发	内容 效度	结构 效度	内部 一致性	跨文化效度/ 测量等同性	信度	测量 误差	效标 效度	反应度
Sanchez-Niubo 等 ^[9]	I	D	V	I	D	D	D	V	A
Lee 等 ^[10]	D	D	V	—	A	I	D	I	I
Jaspers 等 ^[11]	D	D	V	V	A	I	A	V	A
Thanakwang 等 ^[12]	D	D	A	V	A	D	A	V	A
Cyarto 等 ^[13]	D	D	A	V	A	I	A	V	A
Thiamwong 等 ^[14]	D	D	V	V	A	D	A	V	A
李嘉佳 ^[15]	D	D	V	V	A	I	D	I	I
吴凡等 ^[16]	D	D	A	V	A	D	A	V	A
童佩等 ^[17]	D	D	A	V	A	D	I	V	A
Du ^[18]	D	D	V	V	A	I	A	I	I
程彦伶 ^[19]	D	D	V	V	A	D	A	V	A
任立山 ^[20]	D	D	A	V	A	D	A	V	A
胡敏 ^[21]	D	D	A	V	A	I	A	V	A
刘恒 ^[22]	D	D	A	V	A	I	A	V	A

注:V,非常好;A,合格;D,有问题的;I,不合格;—,不适用。

表 3 纳入健康老龄化测评工具的测量属性评价与推荐等级

文献	结构 效度	内部 一致性	跨文化效度/ 测量等同性	信度	测量 误差	效标 效度	假设 检验	反应度	总体	推荐 等级
Sanchez-Niubo 等 ^[9]	+	+	+	+	?	+	?	?	±	B
Lee 等 ^[10]	?	?	+	?	?	?	?	?	±	B
Jaspers 等 ^[11]	—	+	+	?	?	+	+	+	±	B
Thanakwang 等 ^[12]	?	+	+	?	?	+	?	?	±	B
Cyarto 等 ^[13]	?	?	?	?	?	+	?	?	±	B
Thiamwong 等 ^[14]	?	?	?	?	?	+	?	?	±	B
李嘉佳 ^[15]	?	+	?	?	?	?	?	?	±	B
吴凡等 ^[16]	+	+	?	+	?	+	?	?	±	B
童佩等 ^[17]	?	+	?	?	?	+	?	?	±	B
Du ^[18]	—	+	?	?	?	?	+	+	±	B
程彦伶 ^[19]	?	+	?	?	?	+	?	?	±	B
任立山 ^[20]	?	+	+	?	?	+	?	?	±	B
胡敏 ^[21]	+	+	+	?	?	+	?	?	±	B
刘恒 ^[22]	?	+	+	?	?	+	?	?	±	B

注:+,足够;—,不足;?,不确定;±,不一致。

3 讨论

3.1 纳入工具的偏倚风险分析 本研究结果显示,所纳入的国内外健康老龄化测评工具虽然丰富多样,但是测量工具的开发存在一定偏倚风险。首先,本研究所纳入文献大多没有采用质性研究方法收集数据,或者不能清晰描述所采用的质性研究过程。14 项研究在 PROMs 的开发方面存在问题或不合格,主要体现在:未阐述访谈者是否接受培训;不清楚是否被录

音并逐字转录,或者只有录音而没有直接转录,或者只在访谈时做了笔记。PROMs 的开发环节需要采用访谈等质性研究方法收集数据,以挖掘、获取、解释真实世界的原始信息,保证概念引出的全面性、相关性、可理解性。因此测评工具开发研究中的访谈者应接受过专门的培训,具有丰富的数据收集经验、掌握访谈技巧,并事先拟定合适的话题或访谈提纲。建议今后开发工具类的研究应加入质性研究方法,并清晰报告访谈者资质、访谈提纲、对患者与专业人员的提问等质性研究的具体内容。

其次,本研究所纳入文献多缺乏对测量稳定性的检验,这在一定程度上使工具的信度与测量误差存在一定程度的模糊性。信度反映测量中随机误差的大小,效度反映真实值与测量值之间的距离^[23]。所有研究在信度方面存在问题或不合格,4项研究^[9-10,15,17]在测量误差方面存在问题或不合格。因此,工具开发的有关研究应测量信度与测量误差,报告测量工具的稳定性,明确测量数据的可靠程度,从而为其他研究提供借鉴。最后,本研究所纳入的14项研究在内容效度测量方面存在问题,测量工具的条目是否恰当反映所要测量的内容存疑。为此,建议今后的工具开发研究应注重组织相关领域的专家或利益相关者进行内容效度评价。

3.2 纳入工具的测量属性分析 本研究结果显示,所纳入的健康老龄化测评工具测量属性不一致:①9项研究^[10,12-15,17,19-20,22]缺少比较拟合指数(CFI)、塔克-刘易斯指数(TLI)等涉及结构效度相关指标。结构效度能够考察测量的数据结构与目标测量事物特征是否相符,客观评价量表的整体架构,在量表开发中具有重要地位^[24]。结构效度的评价通常采用探索性因子分析和验证性因子分析,二者目的均是对反映事物特征的多个变量进行聚类,然后生成潜变量以解释每一类的变异,但是二者的使用条件与理论依据不同,分别适用于量表评价的开发阶段与验证阶段^[24]。因此,建议有关研究在应用量表做前期调查时采用探索性因子分析,依据资料数据找出影响因子及因子与观测变量的相关程度,来推断数据的因子结构。针对调整后量表的调查采用验证性因子分析,根据先前的理论和已有的知识,经过推论和假设,形成关于一组变量之间关系的模型,同时可明确描述理论模型中的细节^[25]。②7项研究^[13-19]未进行多组因素分析或者项目功能差异(Differential Item Functioning, DIF)分析。多组因素分析与DIF分析能够发现各组别之间(如年龄、性别等)是否存在重要差异,从而验证工具的跨文化效度或测量等同性。鉴于此,在测量工具的开发过程中应进行多组因素分析或者DIF分析,以检验工具的有效普适性。③12项研究^[10-15,17-22]的组内相关系数(ICC)或加权Kappa值未报道。ICC既

可评价定量资料,又可评价分类资料信度^[26]。加权Kappa值是用加权的方法量化2个评价结果,是简单Kappa系数的推广,主要用于评价等级资料^[27]。ICC与加权Kappa值均是一致性检验的常用方法^[27],应清晰汇报有关数值以判定测量工具的一致性、稳定性与可靠性。④12项研究^[9-10,12-17,19-22]未定义假设。假设检验是推断统计的重要组成部分,建议按照已有信息对研究对象总体特征提出假设,再依据小概率原理和某一样本信息、统计量的分布规律,判断假设是否成立^[28]。

4 小结

纳入研究中各测量工具具有一定的应用潜力,同时也存在一定的偏倚风险,且测量属性评价并不一致。偏倚风险评估以评估每个研究的方法学质量为目标,测量属性评价以评估测量属性的优劣为目的,二者的侧重点不同,评价所用的条目不同,因此在信度、结构效度、内部一致性、跨文化效度/测量等同性、反应度等方面的评价等级存在的差异是被允许的。由于健康老龄化的界定与评判标准随着人口老龄化的发展而不断变化,因此未来仍需要设计规范、方法合理、报告严谨的高质量研究,以期准确评估健康老龄化,进而为老龄化研究提供科学的评估标准。建议今后的研究可依据COSMIN指南完善研究设计,规范指标报告,为准确选择适合我国人群的健康老龄化测量工具提供参考。纳入研究所报告的统计值有限,可能影响测评工具的评价,且本研究只纳入中英文文献,因此应谨慎解释评价结果。

参考文献:

- [1] 杜鹏,李龙. 新时代中国人口老龄化长期趋势预测[J]. 中国人民大学学报,2021,35(1):96-109.
- [2] 黄石松,伍小兰,李龙. 积极老龄观健康老龄化[J]. 前线,2021(2):82-85.
- [3] 中共中央,国务院. 中共中央 国务院关于加强新时代老龄工作的意见[EB/OL]. (2021-11-26)[2022-12-23]. <http://lgj.mofcom.gov.cn/article/zcgz/202207/20220703323672.shtml>.
- [4] 穆光宗,胡刚,林进龙. 康养中国:健康老龄化视域下养老体系之重构[J]. 杭州师范大学学报(社会科学版),2022,44(2):64-73.
- [5] 杜鹏,陈民强. 积极应对人口老龄化:政策演进与国家战略实施[J]. 新疆师范大学学报(哲学社会科学版),2022,43(3):91-99,2.
- [6] Prinsen C, Mokkink L B, Bouter L M, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures[J]. Qual Life Res, 2018, 27(5): 1147-1157.
- [7] Mokkink L B, de Vet H C W, Prinsen C A C, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures[J]. Qual Life Res, 2018, 27(5): 1171-1179.