

**患者报告结局在药物临床研究中应用的
指导原则
(征求意见稿)**

2021年9月

目 录

一、 引言	1
二、 患者报告结局的定义	1
三、 患者报告结局测量量表的研发、翻译、改进	2
(一) 患者报告结局测量量表的研发	3
(二) 用于患者报告结局测量量表的翻译和/或文化调适	7
(三) 患者报告结局测量量表的改进	9
四、 患者报告结局测量量表的选择与评价	9
五、 临床研究中患者报告结局的考虑	11
(一) 估计目标框架	11
(二) 选择患者报告结局作为临床研究终点	11
(三) 研究方案和研究报告中有关量表的阐述	12
(四) 量表的有效应答	12
(五) 缺失数据	13
(六) 多重性问题	14
(七) 结果的解释	14
(八) PRO/ePRO 的质量控制	15
(九) 真实世界研究中 PRO/ePRO 的使用	16
六、 电子化患者报告结局	16
(一) ePRO 测量	16
(二) 使用 ePRO 的一般考虑	17
七、 与审评机构的沟通交流	19
参考文献	20
附录 1: 词汇表	22
附录 2: 中英文词汇对照	26

1 患者报告结局在药物临床研究中应用的指导原则

2 (征求意见稿)

3 4 一、 引言

5 临床结局是评价药物治疗获益与风险的核心依据，如何
6 准确、可靠、完整地观测临床结局至关重要。患者报告结局
7 (patient-reported outcome, PRO)是临床结局的形式之一，在药
8 物注册临床研究中得到越来越广泛的使用。另外，随着以患
9 者为中心的药物治疗的理念和实践的不断发展，在药物全生
10 命周期中获取患者体验数据并将其有效地融入到药物的研
11 发和评价中日益受到重视，而 PRO 也是其中的一个重要组
12 成部分。

13 本指导原则旨在阐明 PRO 的定义以及在药物注册研究
14 中的适用范围，PRO 测量特别是量表研发和使用的一般原则，
15 PRO 数据采集的质量控制，数据分析和解释需要注意的事项，
16 以及与监管部门的沟通等，为申办者提供药物注册研究中合
17 理使用 PRO 数据提供指导性意见。

18 本指导原则适用于使用 PRO 作为终点指标支持药品注
19 册的临床研究，包括临床试验和真实世界研究。本指导原则
20 并不特别指定某一测量工具（如量表）的适用场合，也不针
21 对特定疾病。

22 二、 患者报告结局的定义

23 患者报告结局定义为任何来自患者直接报告且不被他
24 人修改或解读的对自身疾病和相应治疗感受的评估结局。

25 量表是 PRO 测量使用较多的工具，主要用于主观测量，
26 如疼痛、生存质量等，但现有量表并不能解决所有的主观测
27 量问题，如某些症状（如恶心）或症状群。

28 PRO 强调患者自己报告结局，需要指出，在医疗环境下
29 医生根据观察或判断所记录的患者症状信息不属于 PRO，而
30 属于医生报告结局。一般而言，在试验纳入排除标准中，应
31 排除因年龄或疾病原因无法应答量表的患者。对于只有患者
32 知道的自身症状或感受的结局不鼓励采用代理报告结局。

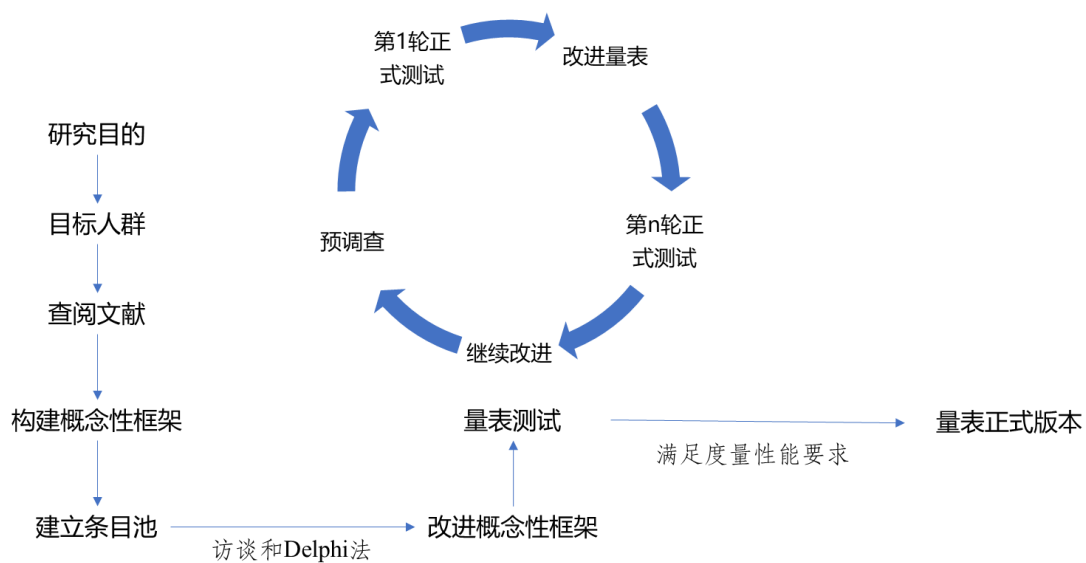
33 PRO 的数据采集有纸质记录和电子化载体两种手段。使
34 用电子化手段记录 PRO 称为电子化患者报告结局（electronic
35 patient-reported outcome, ePRO）。

36 **三、 患者报告结局测量量表的研发、翻译、改进**

37 临床研究中，一旦确定使用量表测量 PRO，如果尚无适
38 合研究项目的量表，需专门针对研究目的进行研发；如果已
39 有公认的适合研究项目的中文量表，可直接使用；如果已有
40 公认的适合研究项目的外文量表，需经过研发形成正式的中文
41 文版本后使用；如果已有量表并不完全适合研究项目时，需
42 改进后使用。在已有的成熟量表中如何选择更适合于拟开展
43 的研究项目，需要考虑其科学性和可操作性。

44 (一) 患者报告结局测量量表的研发

45 量表的研发过程大致为：根据研究目的定义目标人群→
46 查阅文献→构建概念性框架→建立条目池→通过访谈和
47 Delphi 法改进概念性框架→形成量表测试版→预调查→改
48 进量表测试版→第 1 轮正式测试→继续改进测试版→第 2
49 (3...) 轮正式测试（如需要，直至满足对量表的度量性能要
50 求）→形成正式版本（如图 1 所示）。量表的研发通常用于
51 有效性评价，也可以针对重要的安全性事件进行研发，其原
52 理和过程是一样的。



53

54 图 1 量表研发过程示意图

55 1. 构建概念性框架

56 量表的结构有一级结构、二级结构和三级结构，临床研
57 究中以一级和二级结构较常用。一级结构的量表有单条目量

58 表（如视觉模拟疼痛量表）和多条目量表（如简化版口腔干
59 燥量表）。下面以二级结构量表为例进行阐述。

60 二级结构量表的第一级是维度，第二级是条目。量表概
61 念性框架的初步成型一般基于研发者查阅文献、专家知识和
62 经验以及必要的调研。维度的数量和命名依据对研究内容的
63 理解设定，每个维度下的条目数和条目内容用以体现其所属
64 维度的内涵和重要程度，例如每个条目等权时，维度下的条
65 目数量就体现了维度的重要性。

66 2. 建立条目池

67 量表的底层结构是条目，体现具体的设问内容，而维度
68 则是概念性的。为了后续的条目设计，需建立尽可能丰富的
69 条目池，条目的来源可以是所有可能的途径，包括文献、访
70 谈、相关领域的量表研发平台、研发报告、研究者设计，等
71 等。

72 3. 条目设计

73 条目设计是量表的核心内容之一。如果条目池足够丰富
74 和成熟，绝大多数的条目一般从条目池中获取，但也会有个
75 别条目由研发者设计。在问题的陈述中，应尽可能采用封闭
76 式问题，避免含混不清的词语、具有双重含义或倾向性引导
77 的问题、双重否定的陈述、以及负面陈述和患者不情愿回答
78 的问题；同时应避免应答的天花板或地板效应，以及一个条

79 目同时问两个以上问题等。在阅读理解方面，尽量使用常用
80 语，对文化水平的要求不宜太高（如具备小学毕业文化程度
81 的阅读能力即可）。

82 4. 标度方法

83 条目的标度有二分类标度、等级标度（如 Likert 标度）、
84 连续标度（如视觉模拟标度）、图形标度等方法，其中以 5
85 级 Likert 标度法最常用。具体采用几级 Likert 标度要以量表
86 的度量性能达到最佳为标准。

87 5. 访谈和 Delphi 法实施

88 在研发者初步形成了量表的概念性框架后，首先需要进
89 行专家访谈和/或专家调查，根据专家反馈意见调整概念性框
90 架。专家调查的主要目的是考证结构的合理性、条目表述的
91 准确性、应答的可行性和维度及条目的赋权。维度和条目的
92 赋权是量表研发最为关键的环节。专家调查法（Delphi 法）
93 的实施通常不止一轮，具体多少轮以达到专家意见相对统一
94 为止，特别是在赋权方面。

95 6. 预调查和正式调查

96 在综合专家意见改进初始的概念性框架后形成量表的
97 初始测试版，继而需要在少量的目标人群中进行测试，然后
98 根据测试结果改进，形成正式测试版。使用正式测试版在目
99 标人群中展开调查，其样本量需根据预调查的参数（如

100 Cronbach α 系数) 进行估计, 正式测试版的改进也是根据相
101 应的测试结果进行调整, 其测试的轮次取决于量表度量性能
102 的满意程度。

103 7. 验证概念性框架

104 预调查和正式调查都是验证概念性框架的过程。评价概
105 念性框架的适用性主要基于其度量性能, 包括信度和效度。

106 (1) 信度: 信度是指在相似条件下所获得的测量结果的
107 一致性, 用于评价测量工具的可靠性。PRO 量表常用的信度
108 指标有重测信度、内部一致性信度、测试者间信度和测试者
109 内信度。重测信度用于评价量表的可重复性, 初测和再测之
110 间的相关系数不宜太低(如不低于 0.4)。内部一致性信度用
111 于评价量表的内在一致性, 常用 Cronbach α 系数评价(通常
112 不低于 0.7 为宜)。测试者间信度和测试者内信度分别用于
113 评价测试者间和测试者内的一致性, 通常用组内相关系数
114 (intraclass correlation coefficient, ICC) 评价, 有文献报道认
115 为 ICC 一致性可划分成 0.4 为差, 0.4-0.75 为尚可, >0.75 为
116 很好。

117 (2) 效度: 效度是指测量在多大程度上反映了想要测量
118 的内容, 用于评价测量工具的有效性。一个好的量表应该既
119 可靠又有效。信度高并不代表效度也高(例如重度抑郁症症

120 量表用于测量重度抑郁症有较高的信度和效度，而用于测
121 量躁狂症时效度就不高)，但信度低的话，效度必然不会高。

122 量表效度的评价方法很多，以 3C 方法较为常用，即内
123 容效度（content validity）、标准效度（criterion validity）和
124 结构效度（construct validity）。内容效度主要基于专家知识
125 和经验主观上判断量表的维度和条目的内容是否合理，是否
126 能正确反映想要测量的内容。标准效度表示研发量表与所谓
127 “金标准”量表的相关程度。由于金标准通常不存在，且如果
128 存在则研发意义有限（仅在研发量表具有极大的便利性等情
129 况下），因此应用较少。结构效度常通过确证性因子分析方
130 法评估观测数据产生的结构与概念性框架的一致性。

131 除了上述 3C 概念外，效度的另一个重要指标是检测变
132 化的能力，又称反应度，即能够灵敏地反映患者结局变化（如
133 干预前后的变化，给予不同干预的反应等）的能力。

134 8. 撰写量表说明书

135 为了确保量表的正确使用，应撰写量表使用说明书。量
136 表说明书包括但不限于：目标人群，含引导语在内的完整量
137 表结构，维度和条目的赋值以及量表的计分规则，度量性能，
138 有效应答的规定，缺失数据的处理，回忆期限（如涉及）等。

139 （二）用于患者报告结局测量量表的翻译和/或文化调
140 适

141 临床研究中用于 PRO 测量的原研量表如果为外语，通
142 常需要翻译成中文后才能应用。原研量表的某个或某几个条
143 目如果因为文化方面的差异而无法被患者理解或难以获得
144 有效配合时，还会涉及文化调适问题。量表的翻译和/或文化
145 调适是否恰当，要以翻译和/或文化调适后的量表与原研量表
146 的度量性能是否相近为衡量标准。量表的翻译和/或文化调适
147 可按以下步骤进行：

148 1. 准备阶段。查阅量表研发的所有相关资料；组建多
149 学科翻译团队（如英译中、中译英、医学等专业人员）；建
150 立与量表研发者的沟通渠道，除获得使用该量表最新版本的
151 授权许可外，通过交流更好地理解量表的含义，以使翻译更
152 为准确。

153 2. 正向翻译。两个或多个翻译人员独立将原语言版本
154 的量表译成中文版本，然后综合各个翻译文稿形成中文初稿。

155 3. 回译。由母语为原语言且又熟悉中文的翻译人员将
156 中文初稿翻译回原语言，将回译版本与原文进行比较，如有
157 较大差异，需进一步修改中文译稿，直至回译稿与原文的差
158 异达到可接受的程度，形成中文版初版。

159 4. 量表的文化调适。如果量表中有个别条目不适于当
160 地文化，需对其进行调适，调适结果是否满意应以调适后量
161 表与原研版的度量性能相近为判断原则。

162 5. 中文版初版测试。在目标人群中使用中文版对抽取
163 样本进行测试，如果量表的度量性能与原研版相近，中文版
164 可定稿；如果相差较大，则需要进一步完善中文版，直至度
165 量性能达到要求为止，形成中文版终版。

166 6. 中文版研发报告。中文版终版形成后，撰写完成研
167 发报告，记录整个研发过程，报告度量性能，编写量表说明
168 书，必要时申报中文版软著权。

169 (三) 患者报告结局测量量表的改进

170 当已有量表并不完全适合研究项目时，应改进后使用。
171 例如，经早期（如 II 期）临床试验数据分析，所用量表不满
172 足研究所需的信度和/或效度，则在 III 期开展之前需对量表
173 进行改进或研发新的量表，以确保 III 期试验所用量表具有
174 足够的信度和效度。

175 四、 患者报告结局测量量表的选择与评价

176 量表作为 PRO 测量工具应具有良好的度量性能，应既可
177 靠又有效。正确选择适用于拟开展的研究项目的用于 PRO 测
178 量的量表甚为关键，结合科学性和可操作性，建议重点关注
179 以下要点：

180 1.量表的适用性：考察量表的构建，关注其整体概念是否
181 满足量表研发的目的和符合适用人群，研究目标人群应与原
182 研量表的适用人群一致。

183 2.规范文件或系统: 是否有规范的量表相关文件或系统,
184 包括但不限于说明文件(特别是量表得分的解释)、用户使
185 用手册、数据收集的标准格式、重要的参考数据(用于设计
186 时的样本量估计)等。

187 3.研制过程: 量表的使用目的是否明确定义, 研制过程是
188 否严格规范, 量表的结构(维度和条目及其赋权)是否合理,
189 发表的结果是否详尽。

190 4.权威性: 研发成果是否在同行评议期刊公开发布, 是否
191 得到较广泛的引用和应用, 是否被指南推荐。

192 5.语言和文化: 量表的有效性验证是否考虑了不同的教
193 育、文化和种族背景; 新的语种版本是否经过规范的翻译和
194 回译以及验证。经翻译和/或文化调适后的量表其度量性能应
195 与原研量表相近。

196 6.验证: 是否通过足够大的样本量进行验证, 条目设计和
197 赋值是否合理, 是否有足够的信度和效度。

198 7.可行性: 量表在使用时的可行性, 包括但不限于实施过
199 程的可操作性、使用多个量表时的条目重叠问题等。患者的
200 应答负担过重可以导致缺失和拒绝应答现象增多, 降低 PRO
201 数据的质量。增加患者应答负担的因素包括: 量表内容太多,
202 内容重复性高, 同时选择多个量表且其中某个/些量表意义不

203 大，量表界面设计不便阅读，条目涉及不便回答的隐私，条
204 目设计的不合理等。

205 **五、 临床研究中使用患者报告结局的考虑**

206 (一) 估计目标框架

207 ICH E9 (R1)中提出的估计目标框架构建的准则和方法
208 对于以 **PRO** 为试验终点的临床研究同样适用。估计目标框
209 架需在方案中明确定义。

210 (二) 选择患者报告结局作为临床研究终点

211 临床研究如未选择已有公认的客观指标，而选择患者报
212 告结局作为主要终点应说明理由及依据。将患者报告结局设
213 为临床研究中的主要终点或次要终点应结合目标适应症的
214 疾病机制、药物作用机理及临床定位等因素综合考虑。对于
215 将 **PRO** 作为主要或关键次要终点，应注意以下问题：①需要
216 有充分的依据，且与研究目的相一致；②如未能采用双盲设
217 计，会产生较大的主观评价偏倚风险，应极为慎重；③观察
218 期应足够长以体现 **PRO** 具有临床意义的变化；④应控制整
219 体 I 类错误；⑤样本量确定应充分考虑差异的预期至少应具
220 有临床意义。建议与监管部门在试验设计阶段进行沟通。

221 选择的患者报告结局应能反映出患者对药物作用的感
222 受，药物作用不仅限于有效性，也可为安全性或生存质量的

223 变化，合理选择患者报告结局有助于让研究更好地反映患者
224 体验，使药物研发回归以患者为中心的理念。

225 (三) 研究方案和研究报告中有关量表的阐述

226 使用量表测量的患者报告结局作为关键指标时，应在研
227 究方案中对其进行说明，包括但不限于：选择和使用量表的
228 合理性；必要时简要介绍量表的研发和应用情况，特别是针
229 对某些应用较少的量表；量表度量性能的评价方法和指标；
230 量表数据的采集与质量控制；量表数据的分析方法；量表使
231 用的详细说明和培训计划等。

232 临床研究报告中应包括但不限于：量表数据的收集情况
233 （有效应答、缺失等）；报告所使用量表的度量性能（如信
234 度、效度），并与原研量表比较，当差别比较大时，应分析
235 具体原因和评价对研究结论的潜在影响；量表数据的详尽分
236 析结果以及相应的合理解释。

237 (四) 量表的有效应答

238 患者在填报量表时可能会出现缺失、消极应答（如在 5
239 级 Likert 条目的应答中固定勾选某一级）等现象，从而使得
240 量表的数据失真，因此，量表的使用都应设定有效应答的标
241 准，并在量表使用说明书中规定。例如，某量表规定超过 15%
242 的条目未应答，或所有条目都勾选某一级（如“非常满意”）
243 被视为该研究对象的无效应答。在研究方案和统计分析计划

244 中需要详细阐明判断有效应答的标准并阐述理由。如果最终
245 判断为无效应答则与无应答一样视为缺失值处理。有些情况
246 下，除了考虑整个量表是否有效应答外，量表的某一维度可
247 能被视为关键变量，此时可能会对维度的应答是否有效事先
248 做出规定。

249 (五) 缺失数据

250 PRO 数据特别是量表测量的数据，出现缺失较为常见。
251 因此，研究的实施过程中加强质量控制，尽可能减少缺失十
252 分必要。对于提前停止用药的患者，建议在条件允许的情况
253 下继续收集 PRO 数据以减少信息的丢失。对于多维度量表
254 中的条目数据，通常会采用填补方法，具体方法优先采用原
255 研量表说明书提供的方法，其次采用文献报道中的主流方法，
256 再次通过当前研究数据的探索性分析确定（通常在探索性研
257 究中完成）。某些情况下，有可能采用医生评估和患者自评
258 两种量表，如某研究同时采用医生评估的汉密尔顿抑郁量表
259 和患者自评的患者全局改善量表，此时，如果不是两种量表
260 出现相同的缺失，用其中的一种量表的观测数据估计另一种
261 量表的缺失数据或许是更可取的填补方法。如果不做填补，
262 除了缺失太多被视为无效应答外，需要根据原研量表的规定
263 或事先在方案中定义当条目的分值缺失时的规则处理整个
264 量表和各维度的分值。

265 (六) 多重性问题

266 当 PRO 被列为主要终点之一或关键次要终点时，会涉
267 及多重性问题，其一般处理原则参见《药物临床试验多重性
268 问题指导原则》。申办者需要在临床研究方案和统计分析计
269 划中事先规定针对多重性问题所采用的决策策略和多重性
270 调整方法。PRO 使用的量表通常包括多个维度，如果其中某
271 个或某几个维度具有重要临床意义，并在方案中被称为关键
272 次要指标（申报方拟在说明书中声称该特定获益），亦会涉
273 及多重性问题，设计时需考虑整体 I 类错误的控制。

274 由于量表的多维度和多条目特性，除了侧重于量表整体
275 得分的分析外，各个维度和条目的分析也是必要的，从广义
276 上讲是涉及多重性问题的，但只要它们未被列为主要终点或
277 关键次要终点、或者不在说明书中声称特定的获益，无需进
278 行多重性调整。

279 (七) 结果的解释

280 基于量表的 PRO 的结果解释与其它用来评估治疗获益
281 的终点指标相同，阳性结果需同时具有临床意义和统计意义。

282 最小临床意义差别 (**minimum clinical important difference**,
283 **MCID**) 通常用于界定临床意义的阈值，例如，使用 10 分制视
284 觉模拟疼痛量表测量疼痛程度时，干预前后平均分下降多
285 少才有临床意义，或较基线平均下降分两组的差值大于多少

286 才有临床意义。在确定 MCID 时，应首选相关指南、专家共
287 识等公认的标准；如果没有公认的标准，则需与监管机构及
288 时沟通交流并达成共识，统计方法可能为其提供一定的依据。

289 采用统计方法估计 MCID，常用的有基于分布的方法和
290 基于锚定的方法。基于分布的方法较为简单，有文献报道可
291 以将 0.2、0.5、0.8 分别对应低、中、高效应量进行划分，在
292 不同的研究中选择 0.2、1/3 或 0.5 的效应量为阈值。该方法
293 的主要缺陷是未考虑患者的观点。基于锚定的方法是根据患
294 者对临床意义的感受设置一个外部的全局性指标（如无改善、
295 轻微改善、显著改善），然后看对应的量表分值的变化量。
296 通常，全局性指标（等级变量）与量表分值变化量的相关系
297 数至少达到 0.3 以上才有意义，有研究认为相关系数 0.3 为
298 低度相关，0.5 为高度相关。估计 MCID 还有其它统计方法，
299 如基于混合线性模型的方法等，可进一步参阅相关文献。

300 （八） PRO/ePRO 的质量控制

301 应保证研究实施过程中不同研究中心、患者、观察者数
302 据采集的一致性，从而提升临床研究质量。在方案中至少需
303 要明确但不限于：

- 304 • 建立质量控制标准操作规程；
- 305 • PRO/ePRO 数据采集的时间点和实施顺序；

306 • 针对相关人员使用 PRO/ePRO 测量工具的培训和指导，
307 包括判断量表完整性的方法和标准，数据填写、存储、传输
308 的时间和方式等，使之充分理解使用量表的目的、说明书中
309 的具体内容以及在量表数据收集过程的质量控制环节；

310 • PRO/ePRO 的数据管理计划。

311 另外，使用 PRO/ePRO 的临床研究需要更持续主动地现
312 场监查，保证 PRO/ePRO 数据收集的准确性。

313 (九) 真实世界研究中 PRO/ePRO 的使用

314 真实世界研究中，PRO/ePRO 的使用通常限于前瞻性研
315 究，如前瞻性观察性研究或实效临床试验。采集的 PRO/ePRO
316 数据管理或治理的具体方法参见《真实世界数据用于支持药
317 物研发与审评的指导原则》。

318 **六、 电子化患者报告结局**

319 (一) ePRO 测量

320 与纸质 PRO 相比，ePRO 在数据收集的准确性、完整性、
321 高效性、实时性、灵活性、依从性、安全性和患者隐私保护
322 等方面具有明显优势。ePRO 的不足主要体现在某些患者可
323 能会在操作电子设备方面遇到困难，特别是年老、年幼、以
324 及因疾病限制了动手操作能力的患者人群。

325 目前 ePRO 数据的采集大致有基于电话的交互式语音应
326 答系统和基于屏幕的报告系统两种类型。基于电话的交互式

327 语音应答系统以自动呼叫为特色，运用预先录制的问题以及
328 回答选项脚本，并允许患者使用按键记录应答，数据直接存
329 储到中央数据库。基于屏幕的报告系统可安装在患者自己的
330 电子设备上，如智能手机、平板电脑、计算机，甚至是可穿
331 戴医疗设备，又称为自带设备，病人可访问设备上的网站或
332 软件，根据自身情况选择答案并被记录保存。

333 ePRO 系统可与电子病历系统或电子数据采集系统对接，
334 形成个体水平的完整数据流；其时间记录功能可有效防止和
335 识别应答回填或提前应答等影响数据可靠性的行为；其远程
336 监控功能有助于研究者、数据管理人员实时进行在线数据管
337 理和远程数据监测，对有疑问的数据进行质疑标注，及时对
338 受试者进行回访。

339 （二）使用 ePRO 的一般考虑

340 以药物注册为目的的临床研究中，ePRO测量工具及数据
341 采集和数据管理等，应遵循药物临床试验数据管理、电子数
342 据采集、真实世界数据治理相关指导原则的基本要求。

343 基于网络化平台的ePRO测量方式不同于纸质PRO测量
344 工具，其数据通常上传至在线数据收集中心，供用户综合管
345 理，实现数据存储、监查和导出。因此，为确保研究者具备
346 电子源数据维护和保存权限，研究机构有原始文档支持，以
347 供申办者稽查及监管部门核查，使用ePRO测量工具应遵循以

348 下原则：

349 1. 研究者应具有维护和确认 ePRO 源数据准确性、真实
350 性的管理权限。研究者通过稽查轨迹捕捉 ePRO 数据通过测
351 量设备上传后任何数据变化和修改，避免申办者或第三方机
352 构独自控制原始 ePRO 数据的采集/管理系统。

353 2. 数据安全管理体系和访问控制机制。采用加密技术保
354 证数据在收集、提取、传输和存储过程中的完整性、保密性、
355 可追溯性，防止任何个人或机构修改原始数据，删除患者报
356 告的不良事件、高危预警等数据，建立相对应的访问控制机
357 制，避免计划外揭盲风险。

358 3. 数据备份。避免试验过程中存在数据损坏或丢失、无
359 法对源数据进行重建或验证的风险。

360 4. 数据保存。研究机构和研究者应保存有电子源数据或
361 电子文档，使监管部门核查人员可以在临床研究现场检查、
362 核实和复制原始数据。

363 如果经对研究数据的分析发现ePRO量表度量性能与原
364 研量表有较大差距，应考虑ePRO量表在实施中存在的潜在问
365 题，并予以纠正。此外，基于项目反应理论的ePRO测量工具，
366 通过计算机自适应测试技术，根据前项条目的答案选择后项
367 条目，可在保障测量精度的同时减少条目数量，降低患者应
368 答负担，申办方使用此类ePRO测量工具，需提交概念性框架

369 构建、条目库设计筛选流程、程序构建规则以及结果分析解
370 读等相关资料。

371 **七、 与审评机构的沟通交流**

372 当申请人计划采用PRO/ePRO作为确证性研究主要或关
373 键次要终点时，应与监管机构及时沟通。沟通的问题包括但
374 不限于目标适应症疾病背景、选择PRO作为主要或关键次要
375 研究终点的理由及依据、研究设计类型、PRO/ePRO改进和/
376 或文化调适（如有）依据、信度与效度验证、最小临床意义
377 差别及依据、实施方面的质量控制等问题。进行沟通前，申
378 请人应该向监管机构预先提供包含PRO/ePRO统计分析考虑
379 的试验方案和PRO/ePRO的相关资料。在试验过程中，如果因
380 为更改PRO/ePRO而使临床试验方案做出重大调整，应与监
381 管机构及时沟通。

382

参考文献

- 383 [1] Acquadro C, Berzon R, Dubois D, et al. Incorporating the patient's perspective into drug
384 development and communication: an ad hoc task force report of the Patient-Reported
385 Outcomes (PRO) Harmonization Group meeting at the Food and Drug Administration,
386 February 16, 2001. *Value Health*. 2003; 6(5): 522-531.
- 387 [2] Bukhari M. "PROMs vs. PREMs (Patient-Reported Experience Measures)."; *Patient*
388 *Reported Outcome Measures in Rheumatic Diseases*. Ed. Miedany YE. London: Springer,
389 2016; 405-417
- 390 [3] Byrom B, Watson C, Doll H, et al. Selection of and Evidentiary Considerations for Wearable
391 Devices and Their Measurements for Use in Regulatory Decision Making: Recommendations
392 from the ePRO Consortium. *Value Health*. 2018; 21(6): 631-639.
- 393 [4] Calvert M, Blazeby J, Altman DG, et al, CONSORT PRO Group. Reporting of patient-
394 reported outcomes in randomized trials: the CONSORT PRO extension. *JAMA* 2013; 27;
395 309(8): 814-822.
- 396 [5] Cohen J. A power primer. *Psychological Bulletin* 1992; 112(1): 155–159
- 397 [6] Coons SJ. ePRO systems validation: clearly defining the roles of clinical trial teams and
398 ePRO system providers. *Value Health*. 2013; 16(4): 457-458.
- 399 [7] Coons SJ, Gwaltney CJ, Hays RD, et al. Recommendations on evidence needed to support
400 measurement equivalence between electronic and paper-based patient-reported outcome
401 (PRO) measures: ISPOR ePRO Good Research Practices Task Force report. *Value Health*.
402 2009; 12(4): 419-29.
- 403 [8] Copay AG, Subach BR, Glassman SD, et al. Understanding the minimum clinically important
404 difference: a review of concepts and methods. *Spine J Off J North Am Spine Soc*. 2007; 7:
405 541–546.
- 406 [9] Doward LC, Gnanasakthy A, Baker MG. Patient reported outcomes: looking beyond the
407 label claim. *Health Qual Life Outcomes*. 2010; 8: 89.
- 408 [10] EMA. Reflection paper on the regulatory guidance for the use of health relate quality of life
409 (HRQL) measures in the evaluation of medicinal products. 2005.
- 410 [11] EMA. Reflection paper on expectations for electronic source data. 2010.
- 411 [12] EMA. Reflection paper on the use of patient reported outcome (PRO) measures in oncology
412 studies. 2014.
- 413 [13] Fayers PM, Machin D. *Quality of Life: The assessment, analysis and reporting of patient-*
414 *reported outcomes* (3rd Edit). John Wiley & Sons, Ltd. 2016.
- 415 [14] Ferreira ML, Herbert RD, Ferreira PH, et al. A critical review of methods used to determine
416 the smallest worthwhile effect of interventions for low back pain. *J Clin Epidemiol*. 2012; 65:
417 253–261.
- 418 [15] FDA. Clinical outcome assessment (COA) qualification program [https://www.fda.gov](https://www.fda.gov/drugs/drug-development-tool-ddt-qualification-programs/clinical-outcome-assessment-coa-qualification-program)
419 [/drugs/drug-development-tool-ddt-qualification-programs/clinical-outcome-assessment-](https://www.fda.gov/drugs/drug-development-tool-ddt-qualification-programs/clinical-outcome-assessment-coa-qualification-program)
420 [coa-qualification-program](https://www.fda.gov/drugs/drug-development-tool-ddt-qualification-programs/clinical-outcome-assessment-coa-qualification-program)
- 421 [16] FDA. Guidance for industry: Assessing COVID-19-Related Symptoms in Outpatient Adult
422 and Adolescent Subjects in Clinical Trials of Drugs and Biological Products for COVID-19

423 Prevention or Treatment. 2020

424 [17] FDA. Guidance for industry: Electronic source data in clinical investigations. 2013.

425 [18] FDA. Guidance for industry: Patient-Reported Outcome Measures: use in medical product
426 development to support labeling claims. 2009.

427 [19] FDA. Patient-Focused Drug Development: collecting comprehensive and representative
428 input. 2020.

429 [20] FDA. Roadmap to patient-focused outcome measurement in clinical trials. 2015;
430 <https://www.fda.gov/media/87004/download>

431 [21] FDA: Plan for issuance of patient-focused drug development guidance. 2017.

432 [22] FDA. Upper facial lines: developing botulinum toxin drug products. 2014.

433 [23] Fiero MH, Pe M, Weinstock C, et al. Demystifying the estimand framework: a case study
434 using patient-reported outcomes in oncology. *Lancet Oncol* 2020; 21: e488–94

435 [24] Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait.
436 *Biometrics*, 1975; 31: 651-659.

437 [25] Fox MW, Onofrio BM, Onofrio BM, et al. Clinical outcomes and radiological instability
438 following decompressive lumbar laminectomy for degenerative spinal stenosis: a comparison
439 of patients undergoing concomitant arthrodesis versus decompression alone. *J*
440 *Neurosurg*. 1996; 85(5):793-802.

441 [26] Hong K, Majercak KR, Villalonga-Olives E, et al. Patient-reported outcomes in breast cancer
442 FDA drug labels and review documents. *J Patient Rep Outcomes*. 2021; 5(1):36. doi:
443 10.1186/s41687-021-00308-y.

444 [27] Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal
445 clinically important difference. *Control Clin Trials*. 1989;10:407–415.

446 [28] Lawrance R, Degtyarev E, Griffiths P, et al. What is an estimand & how does it relate to
447 quantifying the effect of treatment on patient-reported quality of life outcomes in clinical
448 trials? *Journal of Patient-Reported Outcomes*. 2020; 4(1):68.

449 [29] Ly JJ, Crescioni M, Eremenco S, et al. Training on the use of technology to collect patient
450 reported outcome data electronically in clinical trials: best practice recommendations from
451 the ePRO Consortium. *Ther Innov Regul Sci*. 2019; 53(4): 431-440.

452 [30] Mokkink LB, Terwee CB, Knol DL, et al. Protocol of the COSMIN study: COnsensus-based
453 Standards for the selection of health Measurement Instruments. *BMC Med Res*
454 *Methodol*. 2006; 6: 2.

455 [31] Walters S. Quality of life outcomes in clinical trials and health-care evaluation: A practical
456 guide to analysis and interpretation. John Wiley & Sons, Ltd. 2009.

457 [32] Wild D, Grove A, Martin M, et al. Principles of good practice for the translation and cultural
458 adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR task
459 force for translation and cultural adaptation. *Value Health*. 2005; 8(2): 94-104.

460 [33] ICH guideline: Addendum on estimands and sensitivity analysis in clinical trials to the
461 guideline on statistical principles for clinical trials E9 (R1)

462

附录1：词汇表

463

标准效度 (Criterion Validity): 又称校标效度, 指研发的PRO量表的评分与已知所谓“金标准”量表对同一概念的度量的相关程度。大多数PRO量表因无金标准而无法衡量其标准效度。

464

465

466

测量工具 (Instrument): 一种获取数据以及支持其使用的所有信息和文档的工具(如量表), 通常包括实施方面的详尽指引、数据收集的标准格式、用以说明计分和分析方法以及目标疾病人群的结果解释的规范文件等。

467

468

469

470

代理报告结局 (Proxy-reported Outcome): 一种基于代理报告人的测量, 而不是患者自己报告的结果。代理报告不等同于医生报告和观察者报告, 代理人的报告可能基于其观察结果解读或增加个人观点。不鼓励使用代理报告结局度量只有患者自己才能感受的症状。

471

472

473

474

475

度量性能 (Measurement Properties): 所有与PRO量表应用相关的属性, 包括内容效度、结构效度和信度等。

476

477

概念 (Concept): 又称感兴趣的概念 (concept of interest, COI)。在监管层面, 概念是PRO量表所捕捉或反映的个体在临床、生物、生理、功能的状态或体验。在PRO层面, 概念代表了患者对其健康状况或治疗相关的功能或感觉。

478

479

480

481

患者报告结局 (Patient-reported Outcome, PRO): 为任何来自患者直接报告且不被他人修改或解读的对自身疾病和相

482

483 应治疗感受的评估结局。

484 **患者为中心的药物治疗 (Patient-focused Drug Development,**
485 **PFDD)**: 指一套系统的方法, 在药物全生命周期中, 该法有
486 助于确保患者的体验、观点、需求和优先顺序能够被捕获并
487 有效地融入到药物的研发和评估中。

488 **回忆期限 (Recall Period)**: 患者在应答当时距所要回答的
489 PRO 条目或问题的时间。回忆可以是瞬间的(实时的), 或前
490 一段时间的追溯。回忆期不宜太长, 如一般不超过一周。

491 **检测变化的能力 (Ability to Detect Change)**: 测量工具检测
492 PRO测量得分随测量条件(干预前后不同时间点、不同干预、
493 不同人群等)变化而呈现差异的能力。

494 **结构效度 (Construct Validity)**: 又称构建效度, 指由观测数
495 据所呈现的PRO量表的条目、维度和所要表达的概念之间的
496 结构关系是否与量表研发的理论构想相符。

497 **Cronbach α 系数**: 用于评价量表内在一致性的信度指标。

498 **量表的概念性框架 (Conceptual Framework of a Scale)**: 基
499 于既往研究(文献)、专家知识和经验以及必要的调研所构
500 建的量表的维度和条目的框架。维度的数量和命名依据对研
501 究内容的理解设定, 每个维度下的条目数和条目内容用以体
502 现其所属维度的内涵和重要程度(例如, 每个条目等权时,
503 维度下的条目数量就体现了维度的重要性)。

504 **内容效度 (Content Validity):** 基于专家知识的定性研究, 验
505 证量表是否能够测定其所希望测定的内容。

506 **生存质量 (Quality of Life, QoL):** 又称生命质量, 生活质量,
507 用于评估生活各个方面所体现的整体健康状况。

508 **条目 (Item):** 某一问题、陈述或任务 (以及标准化的应答选
509 项), 被用于患者对特定概念的评估。

510 **调适 (Adaptation):** 基于种族间语言和文化差异的考虑对量
511 表进行的任何更改。调适不会改变PRO量表的结构, 但会调
512 整小部分内容以适用于另一种模式、语言或人群。调适研究
513 是为了验证PRO量表在新环境或新语言下的度量性能。

514 **维度 (Domains/Dimensions/Factors):** 构成量表的第一级结
515 构 (二级结构量表) 或第一和第二级结构 (三级结构量表),
516 用于表达组成量表的某一方面内容 (概念)。一个维度由一个
517 或多个条目构成。

518 **效度 (Validity):** 指测量在多大程度上反映了想要测量的内
519 容, 用于评价PRO量表的有效性。

520 **信度 (Reliability):** 指在相似条件下所获得的测量结果的一
521 致性和可重复性, 用于评价PRO量表的可靠性。

522 **I类错误 (Type I Error):** 指原假设 (或称无效假设) 正确但
523 检验结果拒绝了原假设的错误, 相当于把实际上无效的药物
524 经统计推断得出有效结论的错误。其概率需控制在某一水平,

525 该水平称为检验水准，或称显著性水准，用 α 表示。

526 **医生报告结局 (Clinician-reported Outcome, ClinRO):** 由医
527 生基于对患者的观察所报告的观测结果。

528 **症状 (Symptom):** 只能由患者察觉和感知的疾病、健康状况
529 或治疗效果的任何主观证据。

530 **治疗获益 (Treatment Benefit):** 治疗对患者生存、感觉或功
531 能的影响。治疗获益可以通过有效性或安全优势来证明。例
532 如，治疗效果可通过症状进展方面的改善或延缓来测量，也
533 可通过减少或延缓治疗相关毒性来测量。不能直接获取治疗
534 对患者生存、感觉或功能的效果的测量是治疗获益的替代测
535 量。

536 **终点 (Endpoint):** 符合临床试验的目标、设计和数据分析的
537 测量指标，用于治疗组间的统计学比较。例如，治疗某种疼
538 痛的临床试验，治疗结束时相对基线的疼痛评分该变量。

539 **最小临床意义差别 (Minimum Clinical Important Difference, MCID):**
540 通常用于界定临床意义的阈值，例如，使用10分制
541 视觉模拟疼痛量表 (VAS) 测量疼痛程度时，干预前后平均
542 分值下降多少才有临床意义，或较基线平均下降分两组的差
543 值大于多少才有临床意义。

中文	英文
标准效度	Criterion Validity
测量工具	Instrument
测试者间信度	Inter-rater Reliability
测试者内信度	Intra-rater Reliability
重测信度	Test-retest Reliability
代理报告结局	Proxy-reported Outcome
电子化患者报告结局	Electronic Patient-Reported Outcome (ePRO)
电子数据采集	Electronic Data Capture (EDC)
度量性能	Measurement Properties
概念性框架	Conceptual Framework
汉密尔顿抑郁量表	Hamilton Depression Scale (HAMD)
患者报告结局	Patient-reported Outcome (PRO)
患者自评的患者全局改善量表	Patient Global Impression of Improvement (PGIIMP)
回忆期限	Recall Period
量表	Scale

中文	英文
简化版口腔干燥量表	Summated Xerostomia Inventory, SXI
交互式语音应答系统	Interactive Voice Response Systems (IVRS)
基于屏幕的报告系统	Screen-based Reporting Devices
基于锚定的方法	Anchor-based Method
检测变化的能力	Ability to Detect Change
结构效度	Construct Validity
内部一致性信度	Internal Consistency Reliability
内容效度	Content Validity
生存质量	Quality of Life
视觉模拟标度	Visual Analog Scale (VAS)
条目	Items
条目池	Item Pool
调试	Adaptation
维度	Domains/Dimensions/Factors
项目反应理论	Item Response Theory (IRT)
效应量	Effect Size
效度	Validity
信度	Reliability

中文	英文
医生报告结局	Clinician-reported Outcome (ClinRO)
以患者为中心的 药物研发	Patient-focused Drug Development (PFDD)
应答负担	Respondent Burden
组内相关系数	Intraclass Correlation Coefficient (ICC)
I类错误	Type I Error
终点	Endpoint
症状	Symptom
重度抑郁症 症状量表	Symptoms of Major Depressive Disorder Scale (SMDDS)
治疗获益	Treatment Benefit
自带设备	Bring-Your-Own-Device (BYOD)
最小临床意义 差别	Minimum Clinical Important Difference (MCID)