Industry perspective

Jim Shaw
Lead, Patient-Reported Outcomes Assessment
Worldwide Health Economics and Outcomes Research
Bristol Myers Squibb

Disclosures

The content herein represents the presenter’s own views and not necessarily those of his employer or anyone else.

Summary of Key Concerns

- Provenance of the 15% rule
- Difficulties had in applying rule and interpreting results
- Conflict with other PRO guidances
- Increased complexity and communication challenges
- Possible spillover to other markets

Provenance of 15% rule

- It is not patient centered (no patient input to validate)
- Data source/methodology has been questioned
  - VFA reviewed all meaningful change thresholds accepted through AMNOG process and concluded they ranged from 2-40% with a central value of 9%
- VFA review highlighted variance in change thresholds across patient populations an PRO measures
- One size does not fit all

(VFA 2020)

Difficulties in applying 15% rule

- There are PROs for which scores are theoretically unbounded
  - SF-36v2 and PROMIS measures scored on T-score metric

- No scientific consensus on how to apply 15% of scale rule for these measures
  - Use maximum and minimum calculable scores to define range?
  - Use external (normative) values to define range?
  - Define range based on set number of SDs from the mean?

Difficulties in interpreting results

- Application of the 15% rule can yield quantities that are difficult to explain or interpret

- Consider the following scenario for the SF-36v2
  - Scales and component summaries scored on T-score metric using 1998 US population norms
  - For conceptually related scales (“physical” or “mental”), assume the minimum possible score change needed to meet or exceed 15% of calculable score range threshold
  - For other scales, assume no change in score
  - Assess whether PCS and MCS score changes meet or exceed threshold based on 15% of calculable score range

(Ware et al. 2000)

Difficulties in interpreting results

"Physical" Scales	Score Change	Range	15% Cutoff
PF	6.3	14.9 – 57.0	6.3
RP	7.3	17.7 – 56.9	5.9
BP	8.5	19.9 – 62.1	6.3
GH	7.2	16.2 – 63.9	7.2
PCS	9.7	0.7 – 80.7	12.0

"Mental" Scales	Score Change	Range	15% Cutoff
VT	9.4	20.9 – 70.8	7.5
SF	10.9	13.2 – 56.8	6.5
RE	7.8	9.2 – 55.9	7.0
MH	8.4	7.8 – 64.1	8.4
MCS	12.6	-8.8 – 81.6	13.6

Abbreviations: PF, physical functioning; RP, role-physical; BP, bodily pain; GH, general health; VT, vitality; SF, social functioning; RE, role-emotional; MH, mental health; PCS, physical component summary; MCS, mental component summary.

- Changes in conceptually related scale scores meet or exceed 15% of calculable score range threshold, but PCS and MCS scores fail to do so
- Meaningful changes in components do not translate into meaningful changes in summaries

(Ware et al. 2000)

Conflicts with other PRO guidances

- Benchmarks for meaningful change used in analyses supporting regulatory and other HTA communications will differ from those used in Germany

- In general, FDA sets evidentiary bar for interpretation of meaningful change
  - Anchor-based approaches supplemented with distribution-based methods
  - CDF plot(s)
  - Qualitative support for meaningful change
  - Different values for different scales as well as for improvement and deterioration

(FDA 2009, 2018)

Other stakeholder perspectives – EMA

- Strategies for EMA are often benchmarked against FDA requirements since specific guidance is limited
  - "recommended that HRQL instrument be previously validated for the condition studied (e.g. … responsiveness and interpretability for the specific condition/setting)"
  - "…the determination of MID should be based upon a combination of statistical reasoning and *clinical judgment* and none of them on its own is sufficient"
  - "there is [no] single value of change of relevance for a PRO instrument across all applications and patient samples."
  - "…magnitude of relevance of change should be based primarily on relevant patient-based and clinical anchors"

(EMA 2005, 2016)

Other stakeholder perspectives – HTA bodies and consortia

Agency	Perspective
NICE	• "…does the instrument fail to reflect known changes in health?" • "…Effect sizes do not indicate the value or importance of a change" (i.e., distributional statistics insufficient)
HAS	• "psychometric qualities must be demonstrated, in terms of…sensitivity to change" • "…the objective and clinical relevance threshold [should be] pre-specified…" • "…challenge lies in developing tools that can assist with interpreting quality of life scores by assessing the clinical relevance of quality of life differences observed, such as the minimal important difference"
CADTH	• "Decision-makers are generally concerned with the impact of interventions on patients… [This entails] the need for clinically meaningful outcomes to inform the duration and quality of life."
EUnetHTA	• "…clinical endpoints should be interpreted in terms of…statistical and clinical relevance." • "Documentation of the…responsiveness…of the HRQoL instruments…should be provided." • "Clinical endpoints should be…sensitive (responsive to change); and *recognised/used by physicians.*" • "Some HRQoL measures have been shown to be unresponsive…hence it is important to establish what constitutes a clinically meaningful difference in scores" • "Generic HRQoL instruments are believed to be less responsive than disease-specific instruments" • "…generic HRQoL measures [should be] considered in the assessment. The generic instrument should include all HRQoL dimensions on which improvement is *considered meaningful from a societal point of view*."

(Brazier and Longworth 2011; CADTH 2017; EUnetHTA 2015a, 2015b; HAS 2018)

Other stakeholder perspectives – developers/scientific community

- Researchers are moving directionally toward FDA recommendations and in ways that conflict with IQWiG guidance

(Bottomley et al. 2021; FDA 2019)

Other stakeholder perspectives – developers/scientific community

- Researchers are moving directionally toward FDA recommendations and in ways that conflict with IQWiG guidance

"We do not agree with the proposed deterioration threshold (i.e., 10-points) for the time-to-deterioration analyses for the EORTC QLQ-C30 domains…as there is insufficient evidence to support this threshold. We acknowledge that this 10-point threshold is cited in the Osoba et al. (1998) publication; however, this study has several limitations…"

(Bottomley et al. 2021; FDA 2019)

Other stakeholder perspectives – developers/scientific community

- Researchers are moving directionally toward FDA recommendations and in ways that conflict with IQWiG guidance

"We do not agree with the proposed deterioration threshold (i.e., 10-points) for the time-to-deteriorat
support th
(1998) pub

"Studies [have] observed that thresholds for some scales could be much lower or even much higher, thus the 10-points change might not be relevant for all the scales. However, this threshold has recently been revised and a difference as low as 5 points might still be considered clinically meaningful."

(Bottomley et al. 2021; FDA 2019)

Other stakeholder perspectives – developers/scientific community

- Researchers are moving directionally toward FDA recommendations and in ways that conflict with IQWiG guidance

- There have been recent efforts to confirm meaningful within-subject change thresholds for several EORTC measures

(Bottomley et al. 2021; FDA 2019)

Within-subject meaningful change thresholds for select EORTC subscales

Measure/Subscale	Items	Improved	Worsened	Minimum Change
QLQ-C30				
Physical function	5	5	5	6.7
Emotional function	4	5	5	8.3
Role function	2	15	15	16.7
Cognitive function	2	15	15	16.7
Social function	2	15	15	16.7
Global health	2	5	5	8.3

(Cocks et al. 2015)

Within-subject meaningful change thresholds for select EORTC subscales

Measure/Subscale	Items	Improved	Worsened	Minimum Change
QLQ-MY20				
Body image	1	33	33	33.3
Disease symptoms	6	16	11	5.6
Future perspective	3	11	11	11.1
Side effects	10	6	9	3.3
QLQ-PAN26				
Body image	2	5.4	4.9	16.7
Digestive symptoms	2	13.8	4.5	16.7
Pancreatic pain	4	6.3	3.1	8.3
Sexual dysfunction	2	7.4	3.6	16.7

(Sully et al. 2019; Reni et al. 2021)

Within-subject meaningful change thresholds for select EORTC subscales

- For subscales of several EORTC measures, responder definitions have been estimated as changes of <5 to 33 points
- Minimum observable change is a function of the number of items within a subscale as well as the number of response options per item
- Application of the 15% rule is overly conservative for some multi-item scales (e.g., physical function, global health) but perhaps too liberal for single-item symptom scales
- Given inter-scale variation, “choosing a global responder definition across subscales is not recommended”
- Expect ongoing EORTC MID project to be extended to systematically estimate within-subject change thresholds across measures

(Cocks et al. 2015; Musoro et al. 2018; Sully et al. 2019; Reni et al. 2021)

Increased complexity and communication challenges

- AMNOG submissions will include results based on multiple thresholds for meaningful change
  - Best practice thresholds should serve as base case with 15% of scale rule applied in sensitivity analyses
- Additional analyses will translate into additional resources (time, human capital, money) at both local and global/HQ level
- Added work will not resolve underlying misalignment between regulatory and HTA stakeholders
- 15% rule limits comparability with earlier assessments based on different change thresholds
- Lack of consistent agreement between IQWiG and G-BA
  - G-BA acceptance of 15% rule for measures like EQ-5D VAS but not the SF-36 may result in variance in added benefit decisions

Potential for spillover

- 15% of scale rule may raise bar for demonstrating additional clinical benefit based on some PRO measures (e.g., EORTC QLQ-C30, EQ-5D VAS)

- Worsened benefit assessments would impact price directly in Germany and indirectly in markets that reference Germany

- Potential implications for EU-wide HTA assessment
  - EC has proposed mandatory joint European HTA assessment
  - Components of German system could be adopted into EU HTA

(EC 2018)

A possible side benefit

- The 15% of scale rule may provide a means to demonstrate meaningful change in cases where no accepted threshold exists
  - G-BA not accepting of distributional statistics (e.g., 0.5 SD)

- Goal should be acceptance of meaningful change thresholds derived using robust methods, but in cases where there is no alternative, the 15% rule could prove useful

Conclusions

- New guidance has added complexity, will make it more challenging to interpret and communicate results, and is inconsistent with other stakeholder guidances and evolving best practice

- Need to work with authorities to ensure that evidence-based thresholds for meaningful change play a role in the assessment process and that any role for the 15% rule is clearly delineated

- One size does not fit all

Backup


6%
7%
2,000.0
1,500.0
1,000.0
500.0
0.0
1 2 3 4 5 6
16.64
15.65
6.61
7.84
10.97
15.81
0.7
0.323044069
0.283720087
0.825218477
0.142935278
0.180160946
0.172184709
0.016982048
0.151785714
0.393072768
0.128864544

References

1. Amendment to IQWiG Methods Paper: Patient Perspective in AMNOG at Risk. Berlin, Germany: Verband Forschender Arzneimittelhersteller; 2020.
2. Appendix 2 to the Guideline on the Evaluation of Anticancer Medicinal Products in Man. The Use of Patient-Reported Outcome Measures in Oncology Studies. London, UK: European Medicines Agency; 2016.
3. Bottomley A, Coens C, Mierzynska J et al. Lancet Oncology. 2021;22(5):655-664.
4. Brazier J, Longworth L. NICE DSU Technical Support Document 8: An Introduction to the Measurement and Valuation of Health for NICE Submissions. Sheffield, England: Decision Support Unit, School of Health and Related Research (ScHARR), University of Sheffield; 2011.
5. Cocks K, Buchanan J. Defining Responders on the EORTC QLQ-C30 Subscales. 22nd Annual Meeting of the International Society for Quality of Life Research, Vancouver, Canada, October 21-24, 2015.
6. Comments received from US Food and Drug Administration in response to advice/information request, 2019.
7. Endpoints used for Relative Effectiveness Assessment: Clinical Endpoints. Diemen, The Netherlands: European Network for Health Technology Assessment; 2015a.
8. Endpoints used for Relative Effectiveness Assessment: Health-Related Quality of Life and Utility Measures. Diemen, The Netherlands: European Network for Health Technology Assessment; 2015b.
9. Evaluation of Health Technologies at HAS: Role of Quality of Life. Saint-Denis, France: Haute Autorité de Santé; 2018.

References

10. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Silver Spring (MD): US Food and Drug Administration; 2009.
11. Guidelines for the Economic Evaluation of Health Technologies: Canada. 4th ed. Ottawa, Canada: Canadian Agency for Drugs and Technologies in Health; 2017.
12. Musoro JZ, Hamel J-F, Ediebah DE et al. BMJ Open. 2018;8(1):e019117.
13. Patient-Focused Drug Development Guidance Public Workshop. Methods to Identify What is Important to Patients & Select, Develop or Modify Fit-for-Purpose Clinical Outcomes Assessments. Silver Spring (MD): US Food and Drug Administration; 2018.
14. Proposal for a Regulation of the European Parliament and of the Council on Health Technology Assessment and Amending Directive 2011/24/EU. Brussels, Belgium: European Commission; 2018.
15. Reflection Paper on the Regulatory Guidance for the Use of Health-Related Quality of Life (HRQL) Measures in the Evaluation of Medicinal Products. London, UK: European Medicines Agency; 2005.
16. Reni M, Braverman J, Hendifar A et al. Annals of Surgical Oncology. Online ahead of print April 3, 2021.
17. Sully K, Trigg A, Bonner N et al. European Journal of Haematology. 2019;103(5):500-509.
18. Ware JE, Kosinski M, Dewey JE. How to Score Version 2 of the SF-36 Health Survey (Standard & Acute Forms). Lincoln (RI): QualityMetric Inc.; 2000.