CRITICAL PATH INSTITUTE

Expanding the toolbox to interpret meaningfulness of treatment effects

Clinical Outcome Assessment Program Annual Meeting

April 16-17, 2026

Washington, D.C.

Disclaimer

The views and opinions expressed in the following slides are those of the individual presenters and should not be attributed to their respective organizations/companies, the U.S. Food and Drug Administration, or the Critical Path Institute.

Session Participants

Moderator:

- *Fraser Bocell, MEd, PhD* – Senior Clinical Outcome Assessment Scientist, Clinical Outcome Assessment Program, Critical Path Institute

Presenter:

- *Kevin P. Weinfurt, PhD* – James B. Duke Distinguished Professor of Population Health Sciences, Duke University School of Medicine

Panelists:

- *Kim Cocks, PhD* – Senior Director, Patient-Centered Outcomes, Adelphi Values

- *Andrew Trigg, MSc* – Statistical Innovation Lead, Bayer

Session Agenda

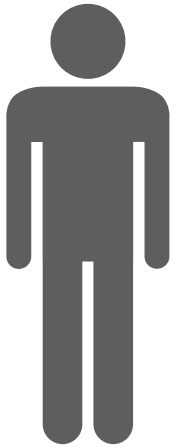
- Introduction (10 minutes)
- Panel Discussion on Approaching with NUANCE (40 minutes)
- Audience Q&A (10 minutes)

Meaningfulness of Endpoints Based on Clinical Outcomes Assessments:

A Historical, High-Altitude Introduction

Kevin P. Weinfurt, PhD
Duke University School of Medicine

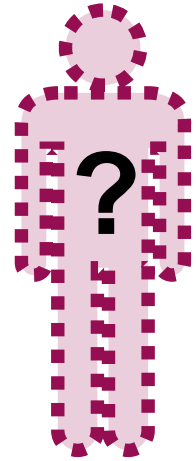
Want to assess a meaningful aspect of how patient feels/functions (*Meaningful Aspect of Health*)



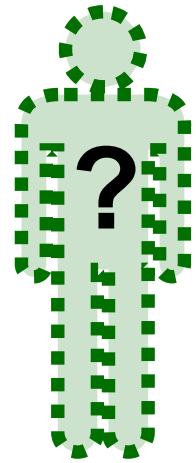
Feeling/ functioning in daily life

Because we eventually want to know whether a product has a favorable effect on how the patient will feel or function (relative to some comparator)

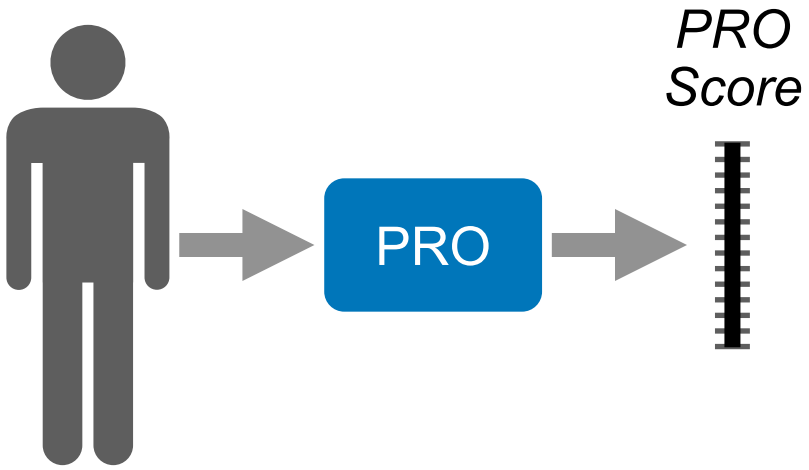
Treatment

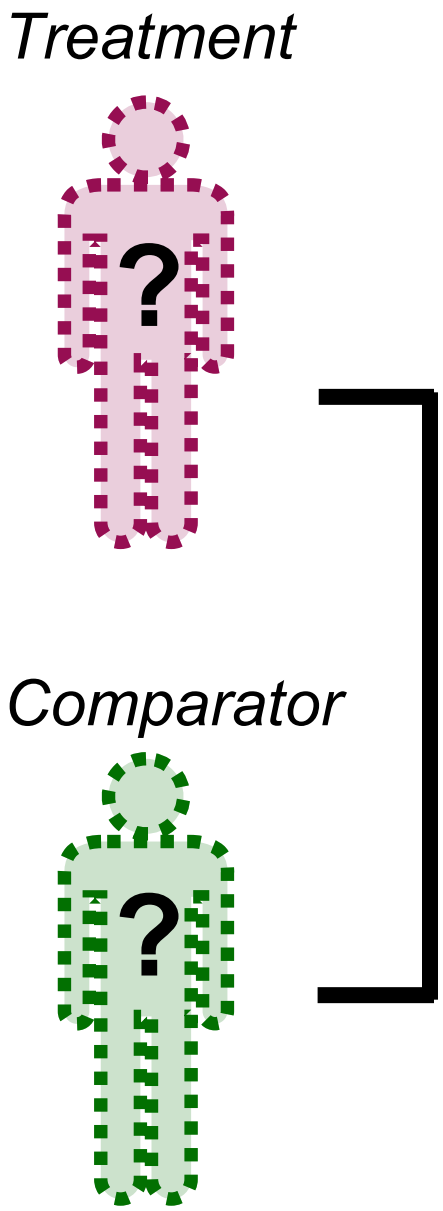


Comparator

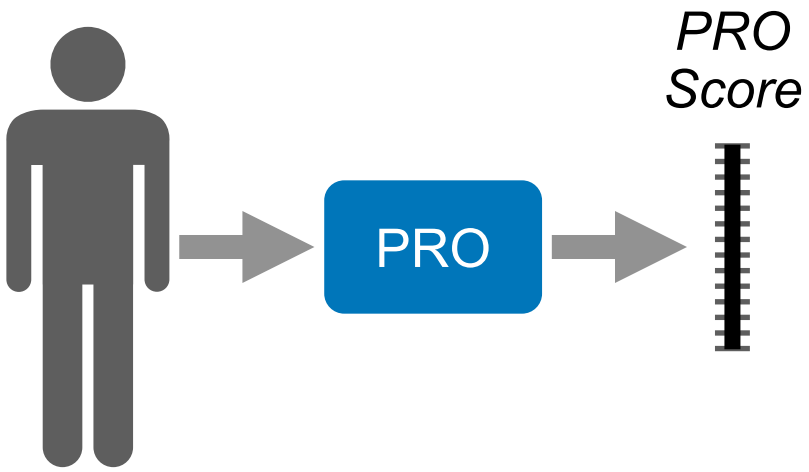


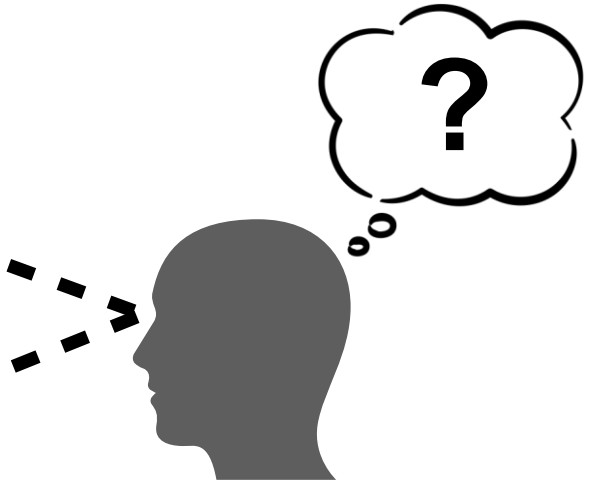
Assess patient with a Patient-Reported Outcome (PRO) measure, which generates a score

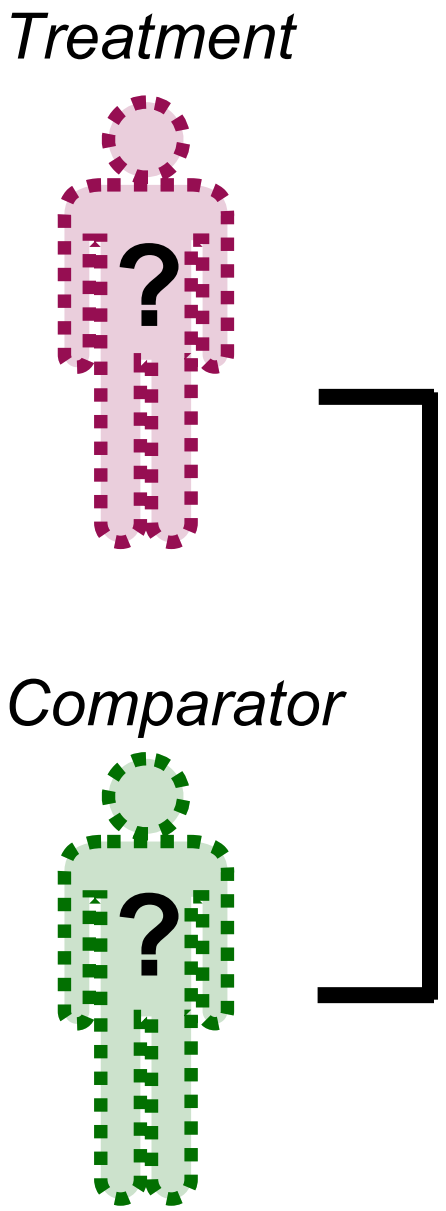




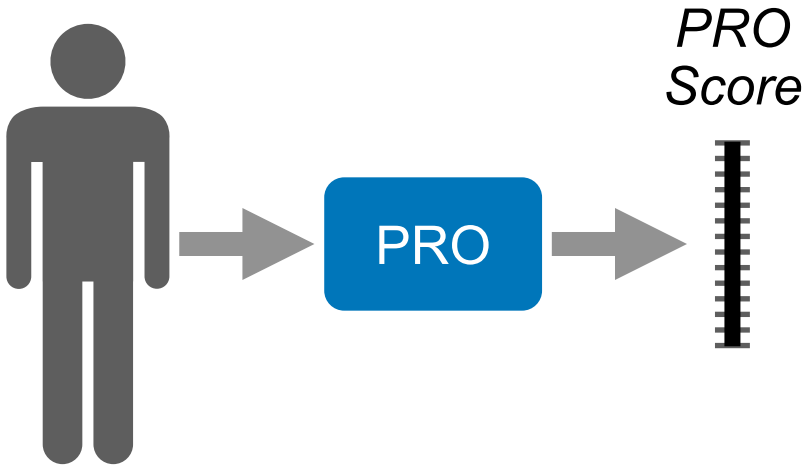
The PRO score is an unfamiliar metric, difficult to translate into expected patient experiences

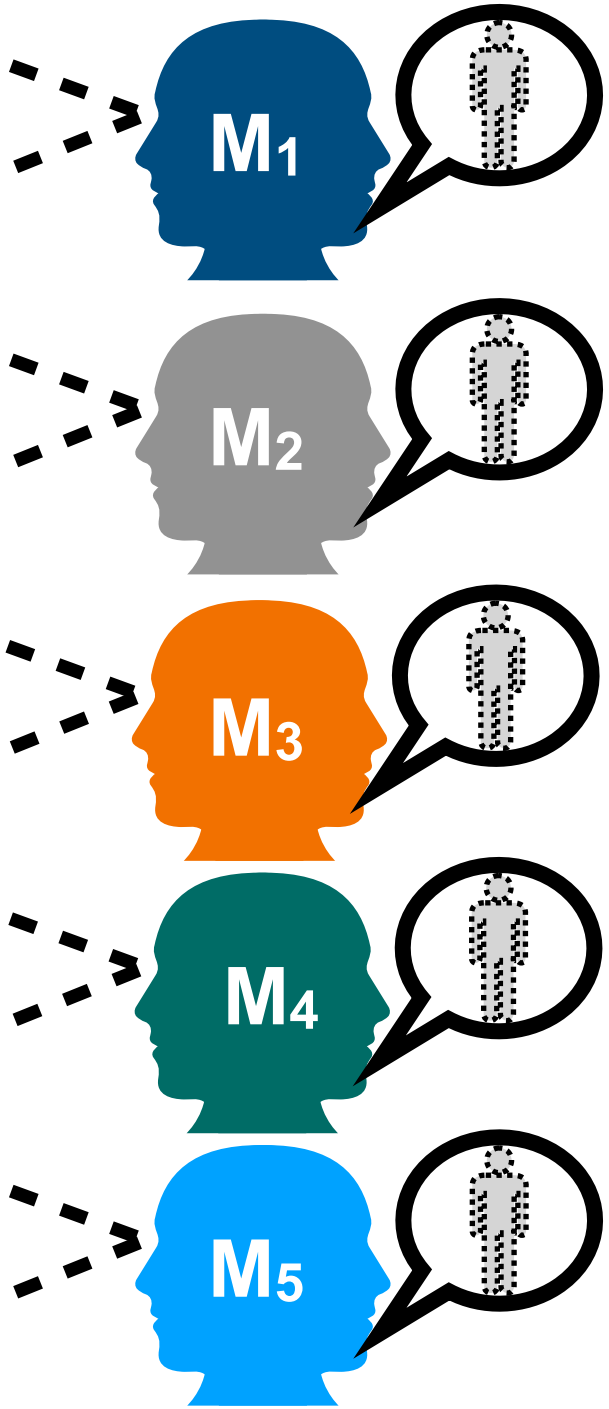


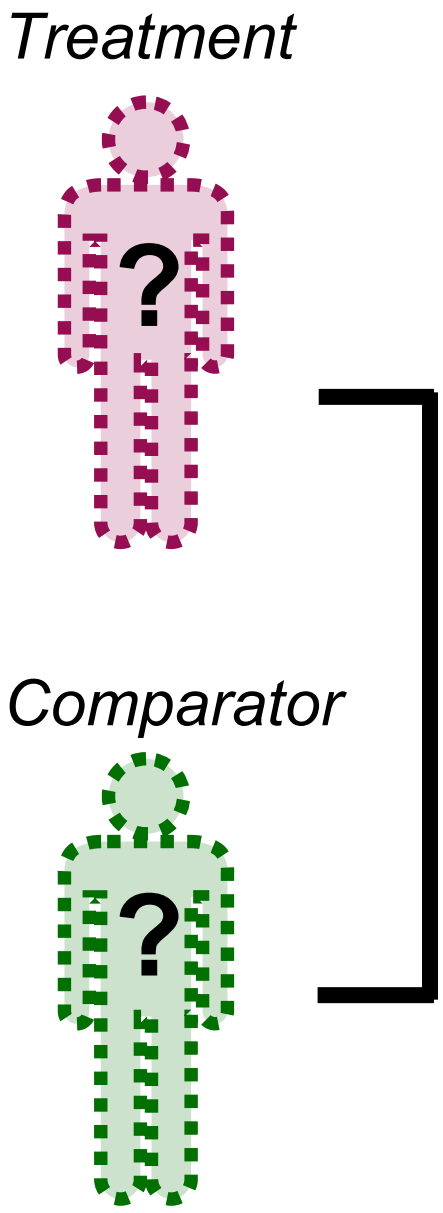




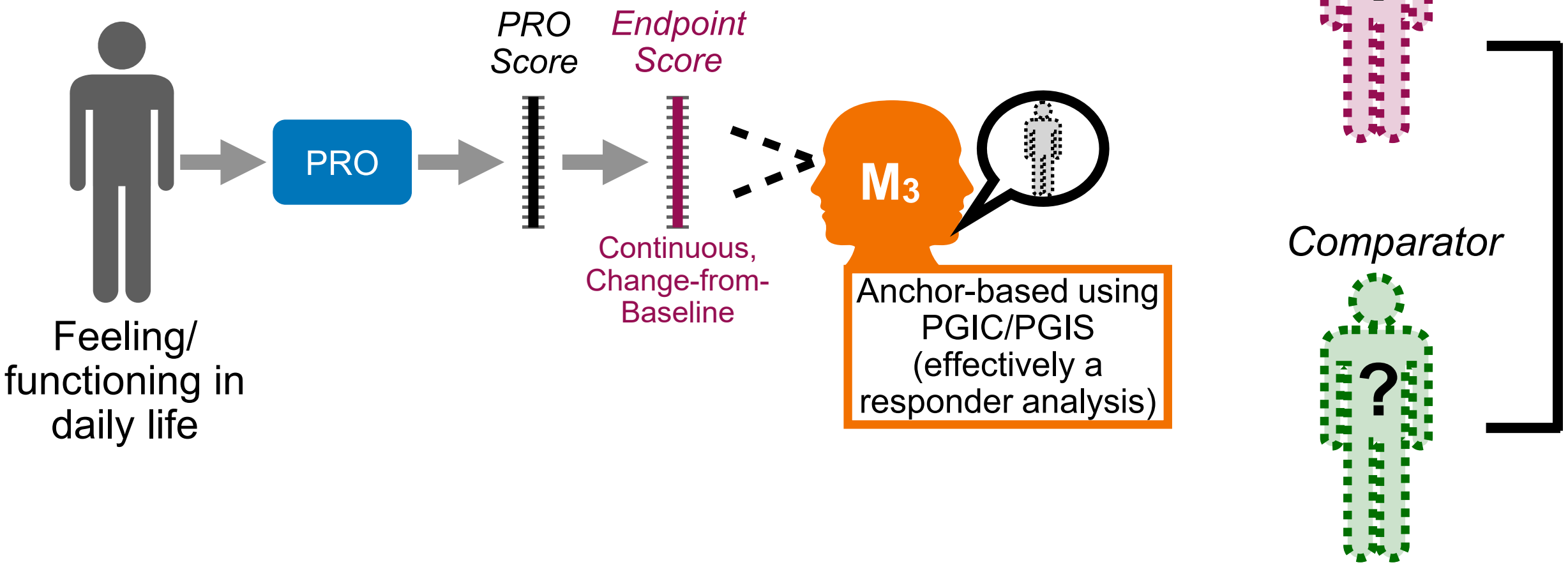
1990s/early 2000s, multiple methods emerged to help translate PRO scores into something more relatable to patients' daily lives



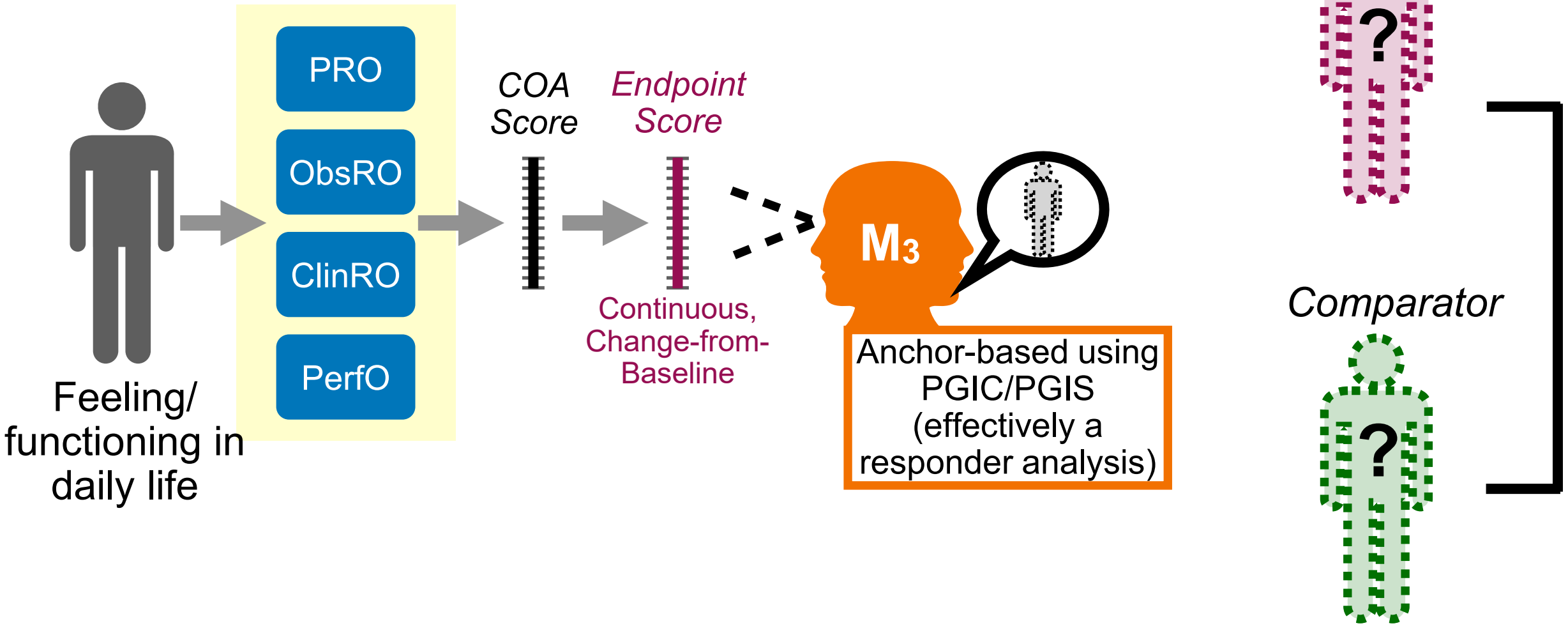




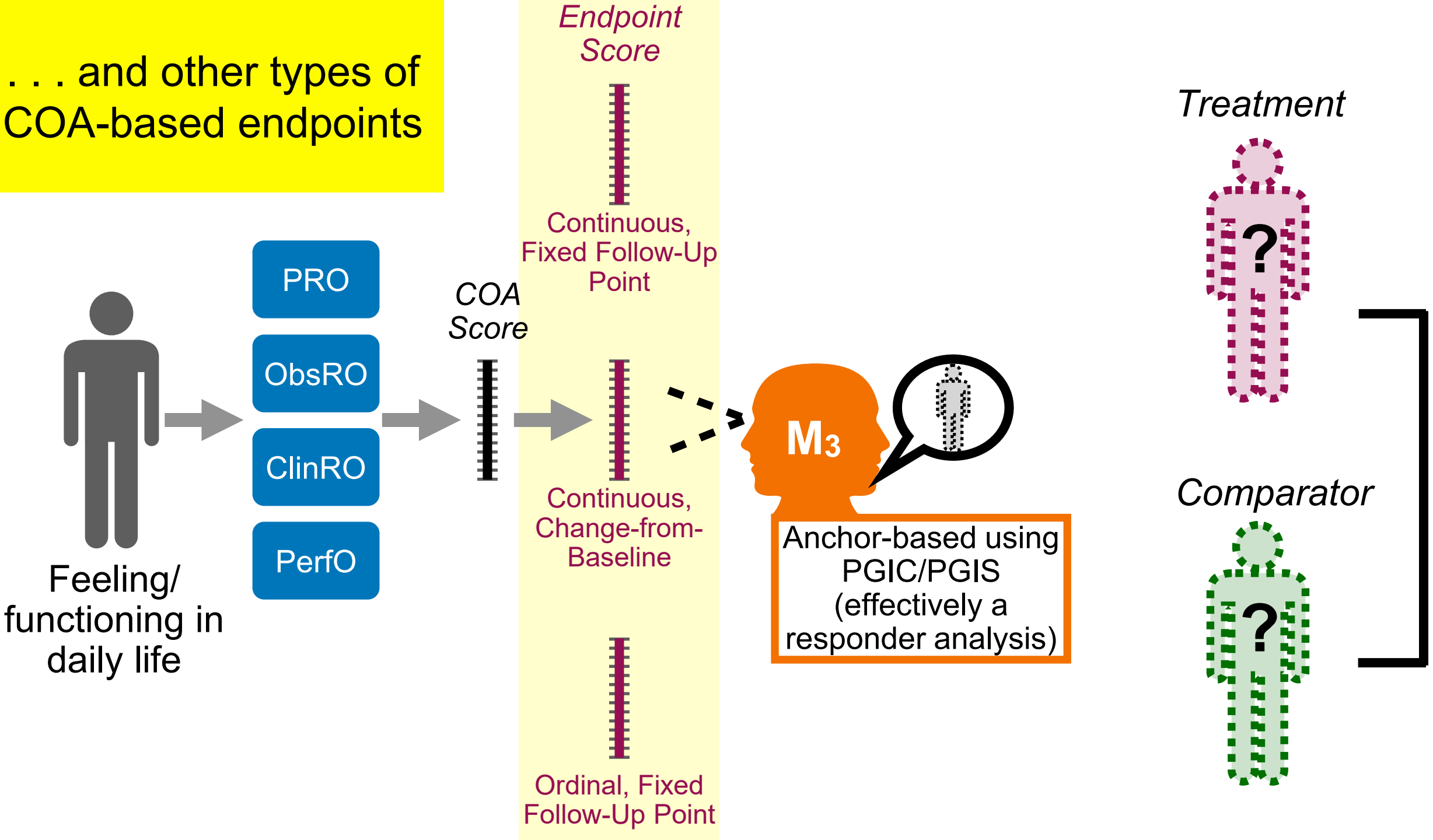
After that, the field narrowed to focus on one main method applied to one type of PRO-based endpoint



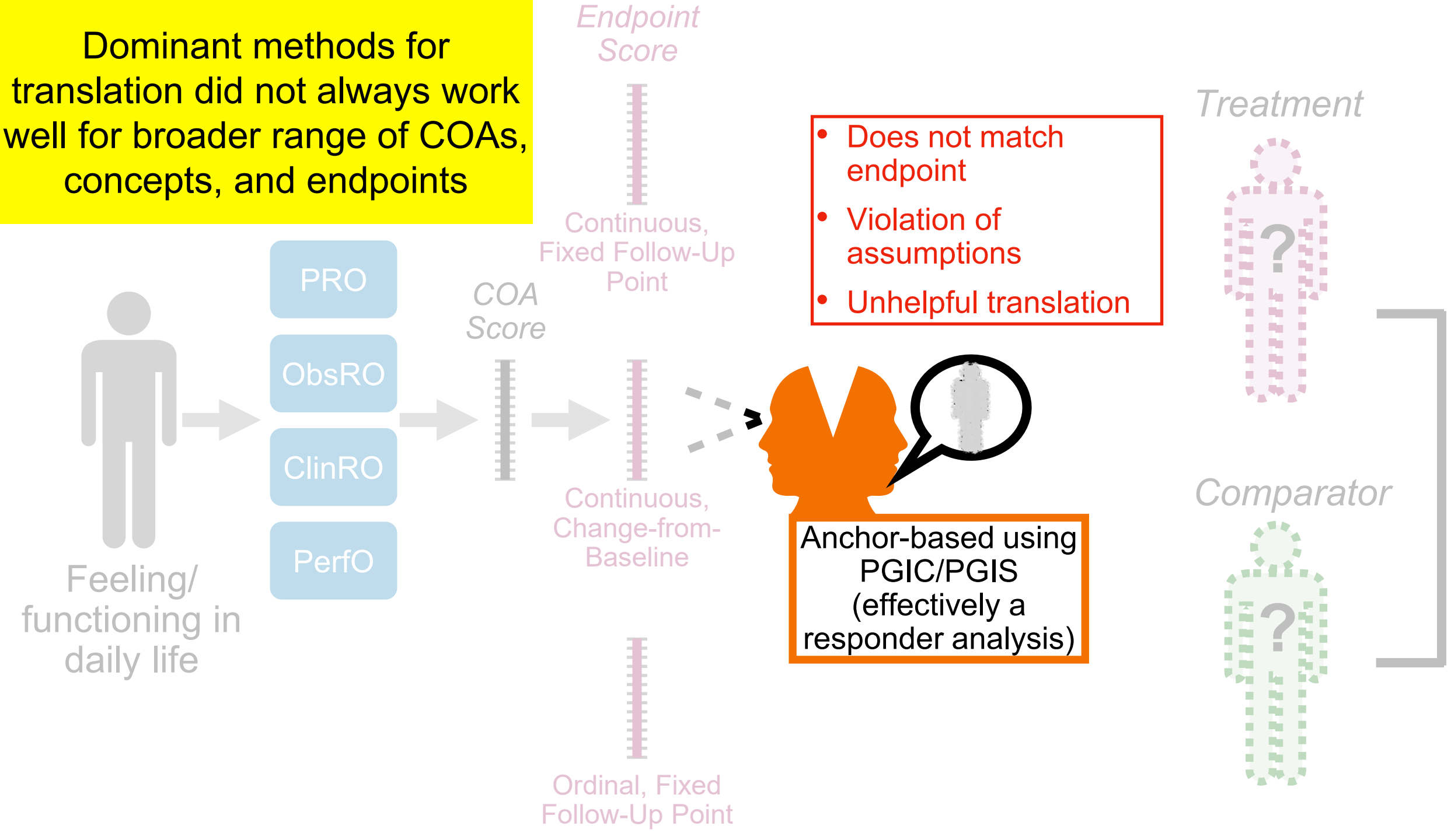
Over time, focus broadened to include other types of Clinical Outcome Assessments (COAs) . . .



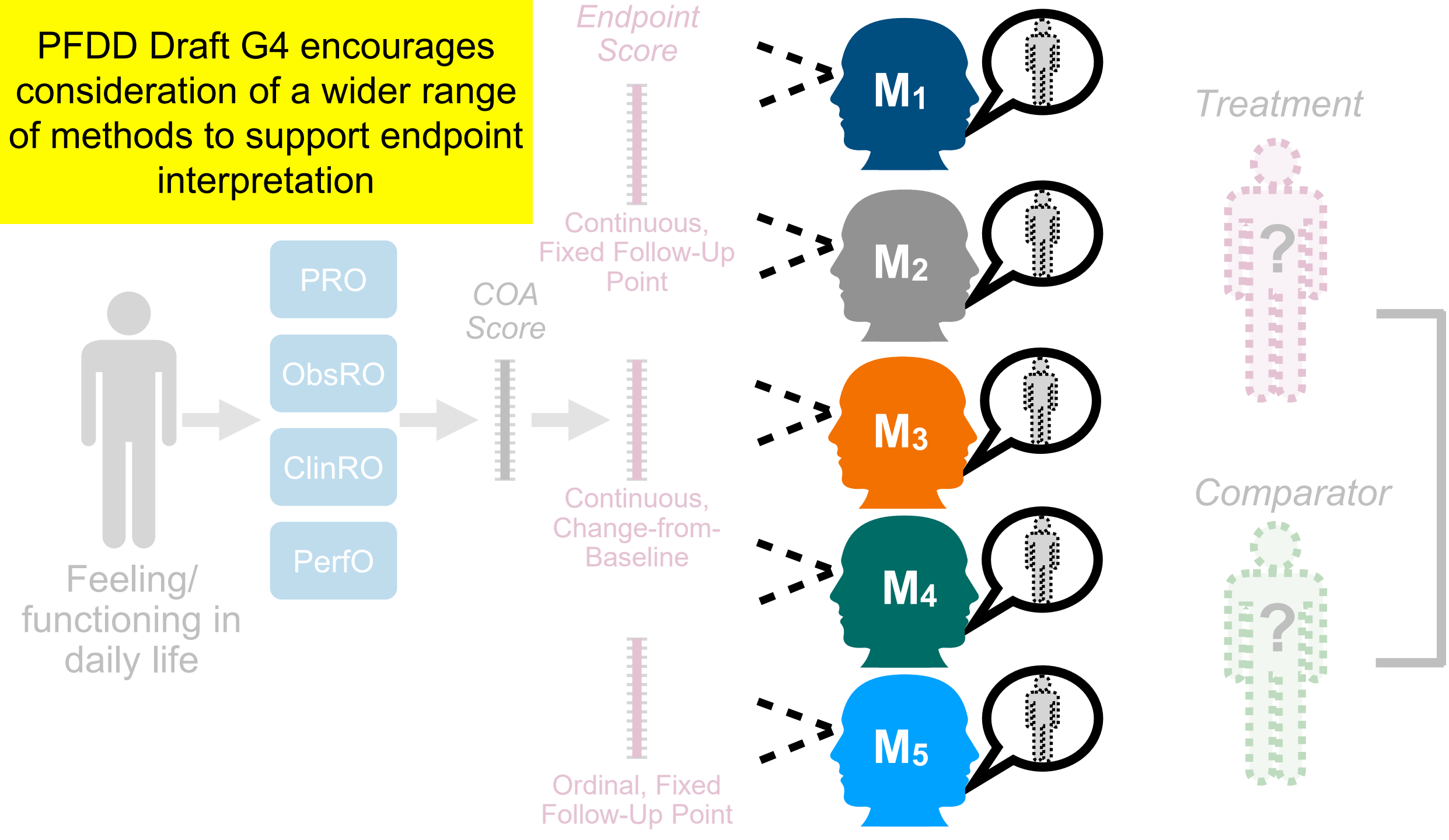
. . . and other types of COA-based endpoints



Dominant methods for translation did not always work well for broader range of COAs, concepts, and endpoints



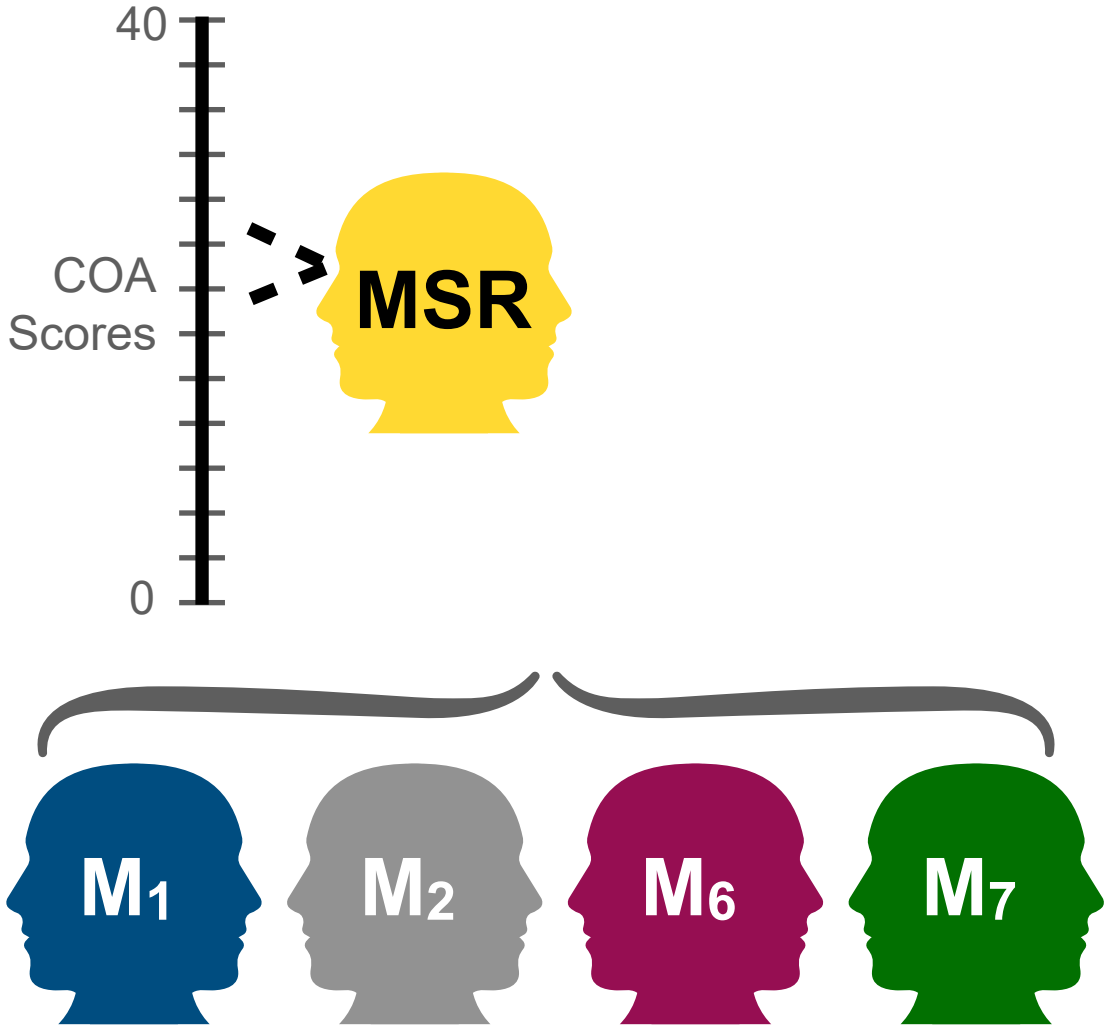
PFDD Draft G4 encourages consideration of a wider range of methods to support endpoint interpretation



All methods can be categorized by their approach to translation

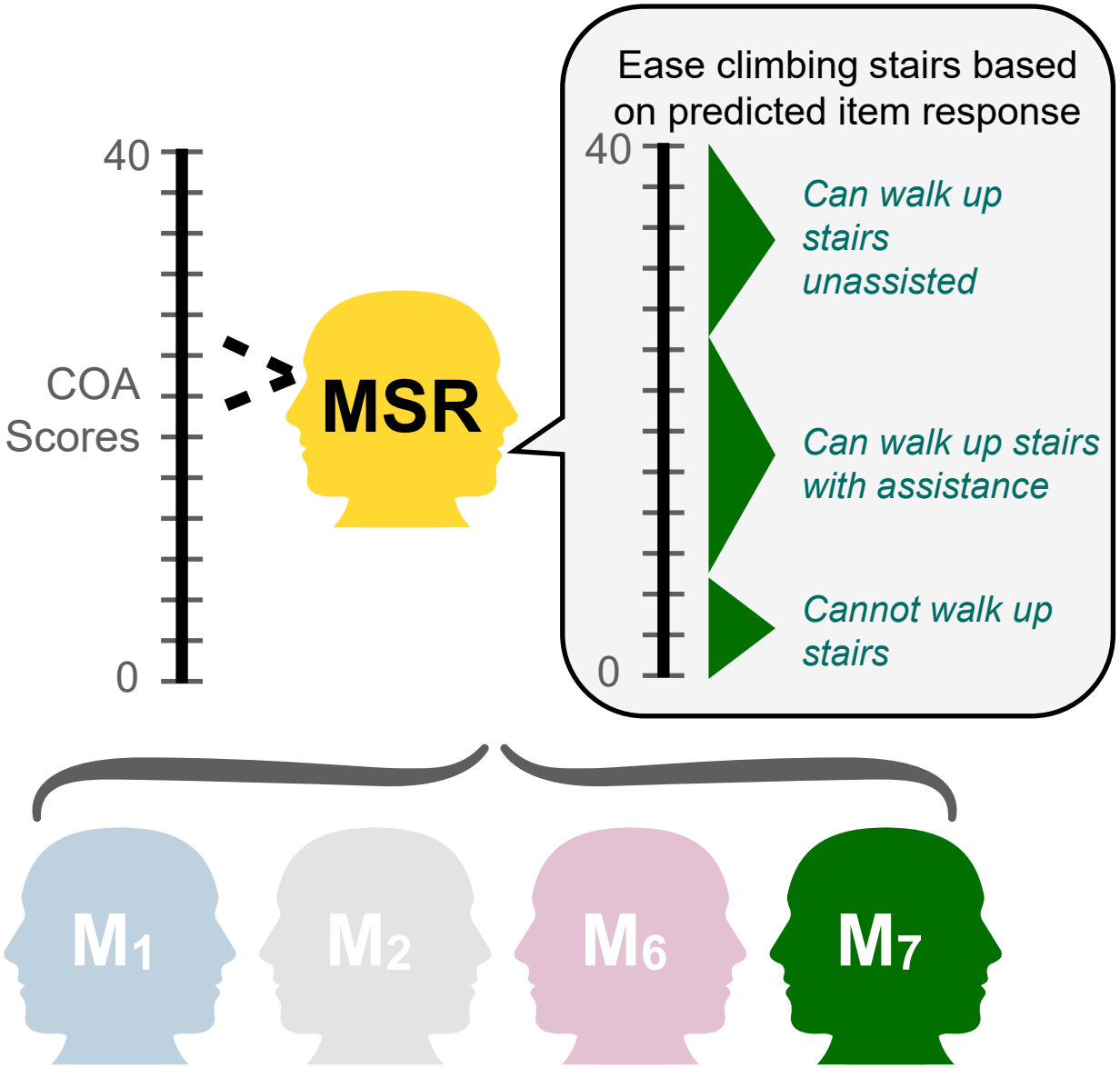
Meaningful Score Region (MSR) Approaches

Translate scores into something I understand better



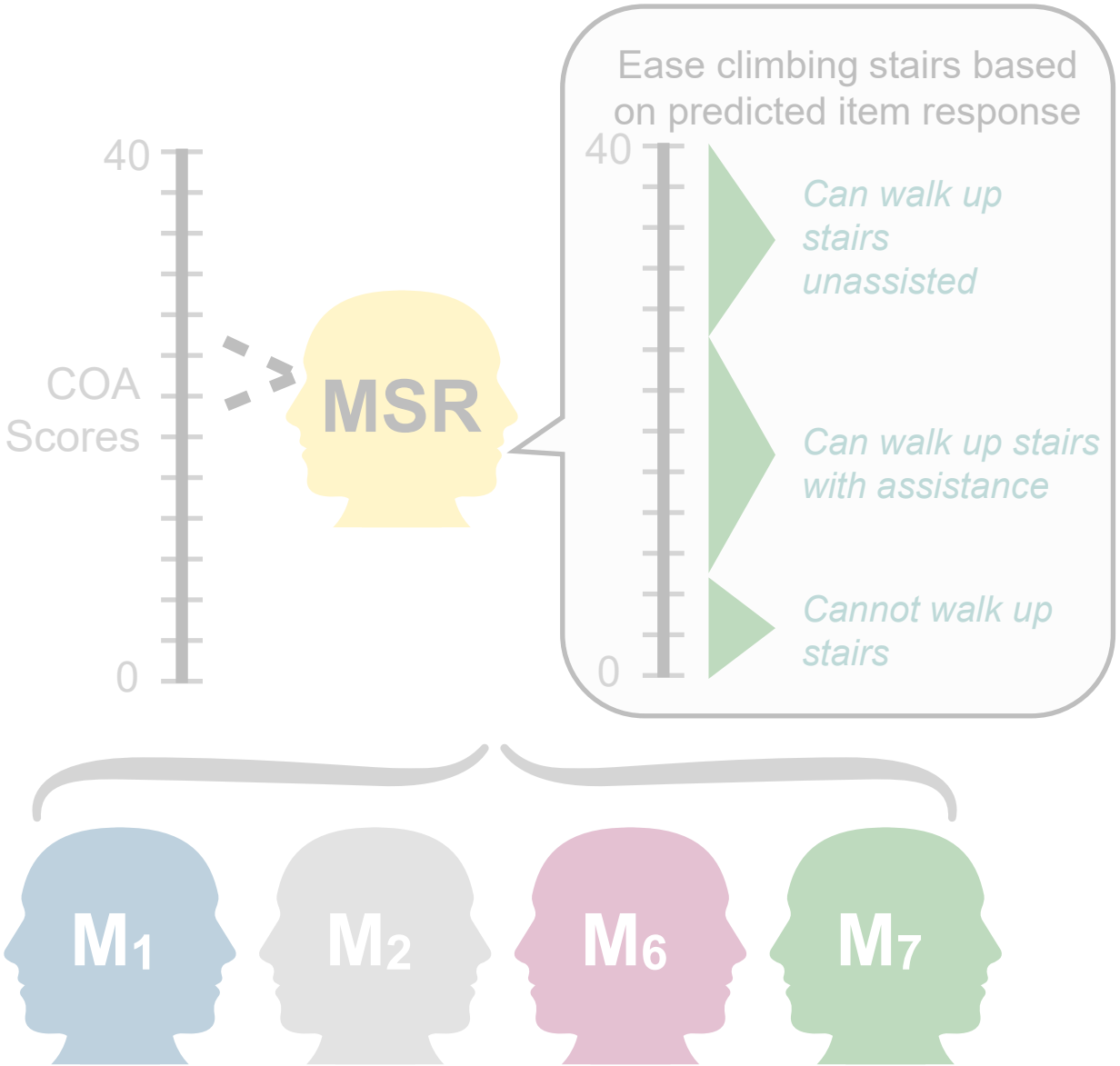
Meaningful Score Region (MSR) Approaches

Translate scores into something I understand better



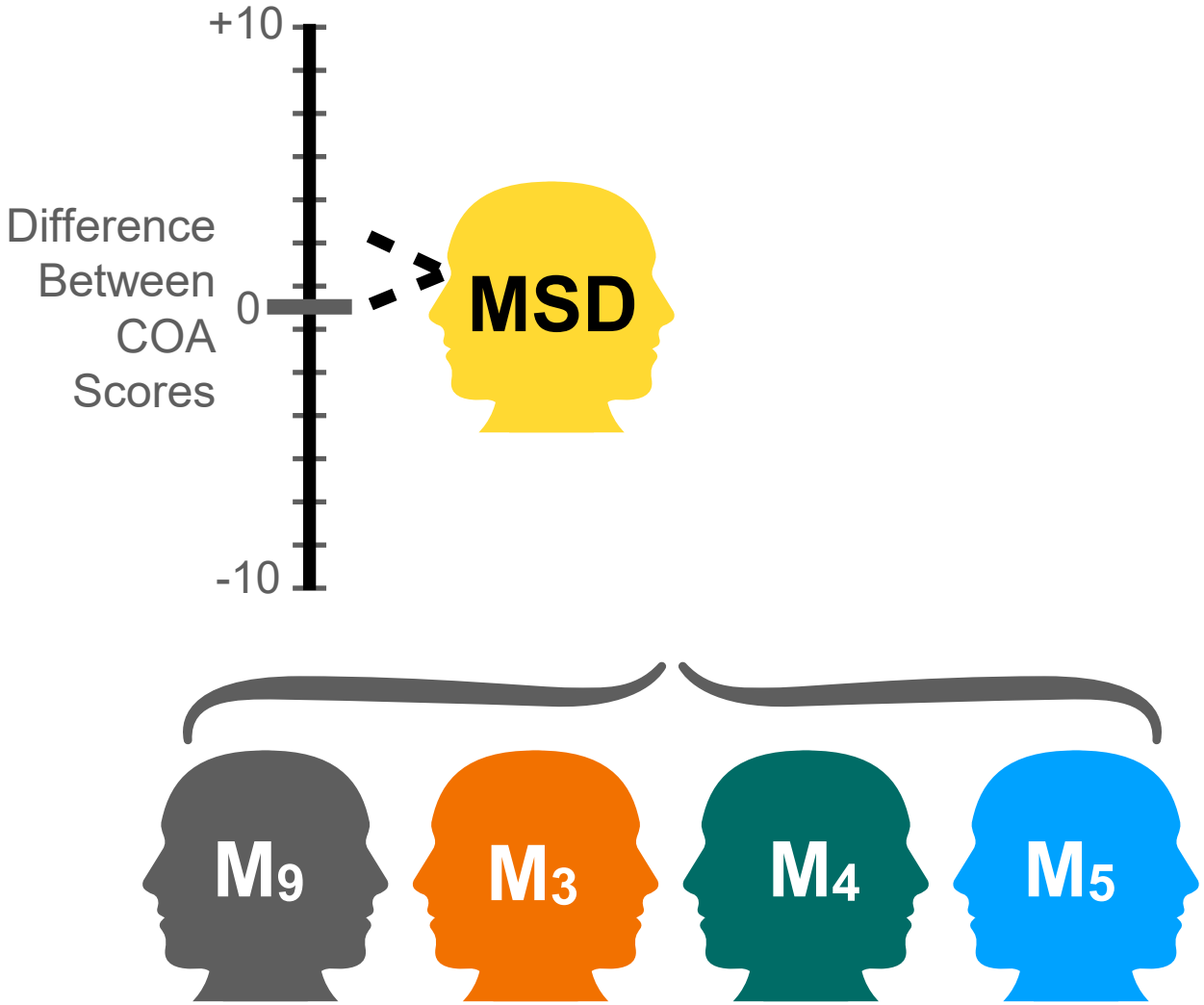
Meaningful Score Region (MSR) Approaches

Translate scores into something I understand better



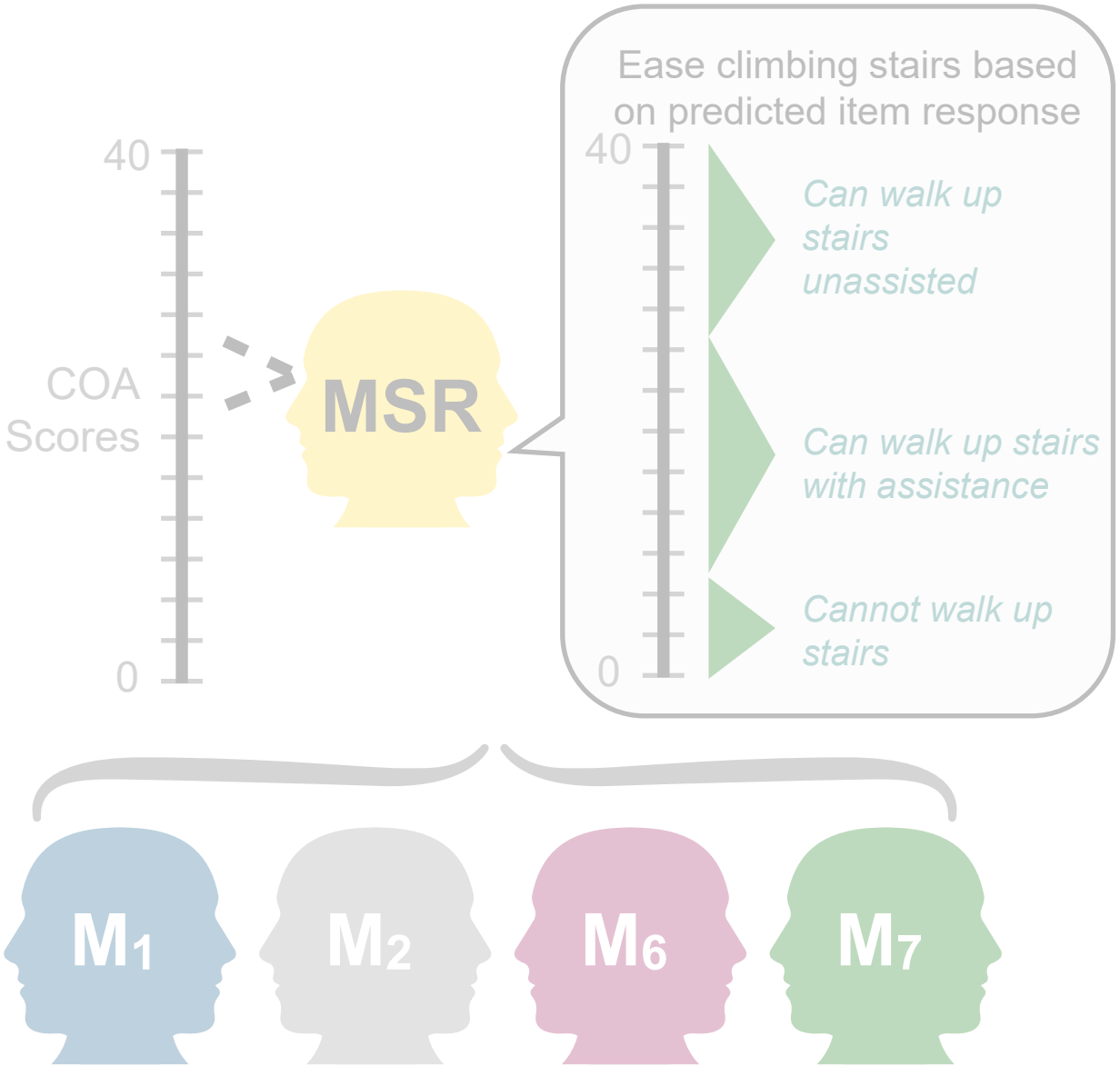
Meaningful Score Difference (MSD) Approaches

Translate score differences to some other difference I understand better



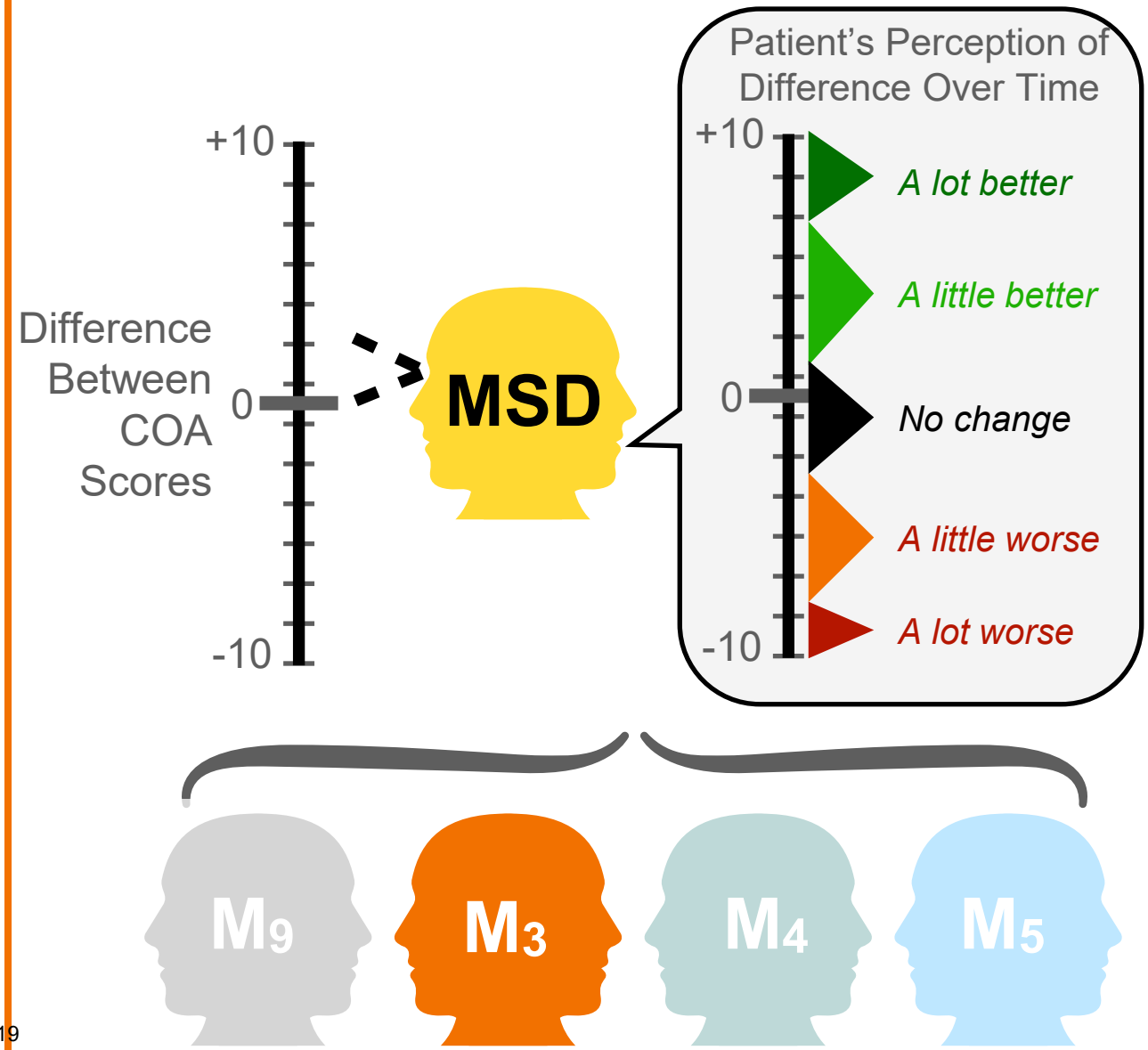
Meaningful Score Region (MSR) Approaches

Translate scores into something I understand better



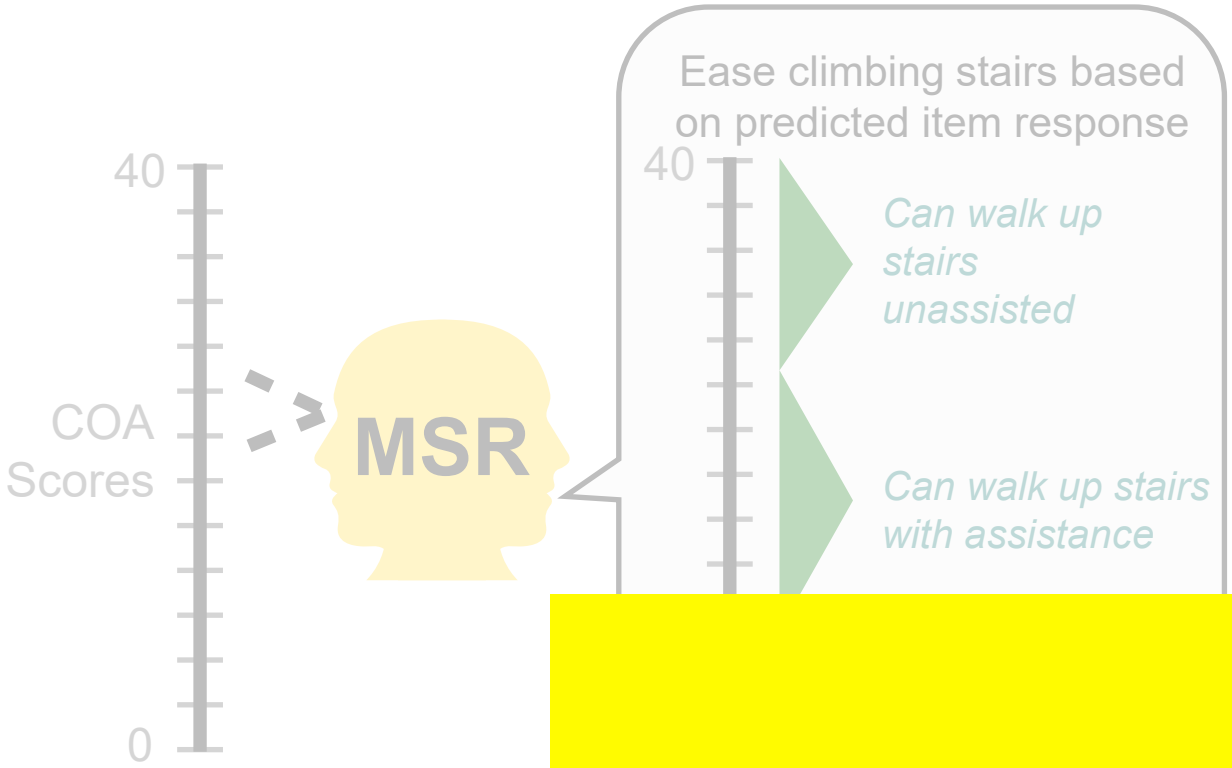
Meaningful Score Difference (MSD) Approaches

Translate score differences to some other difference I understand better



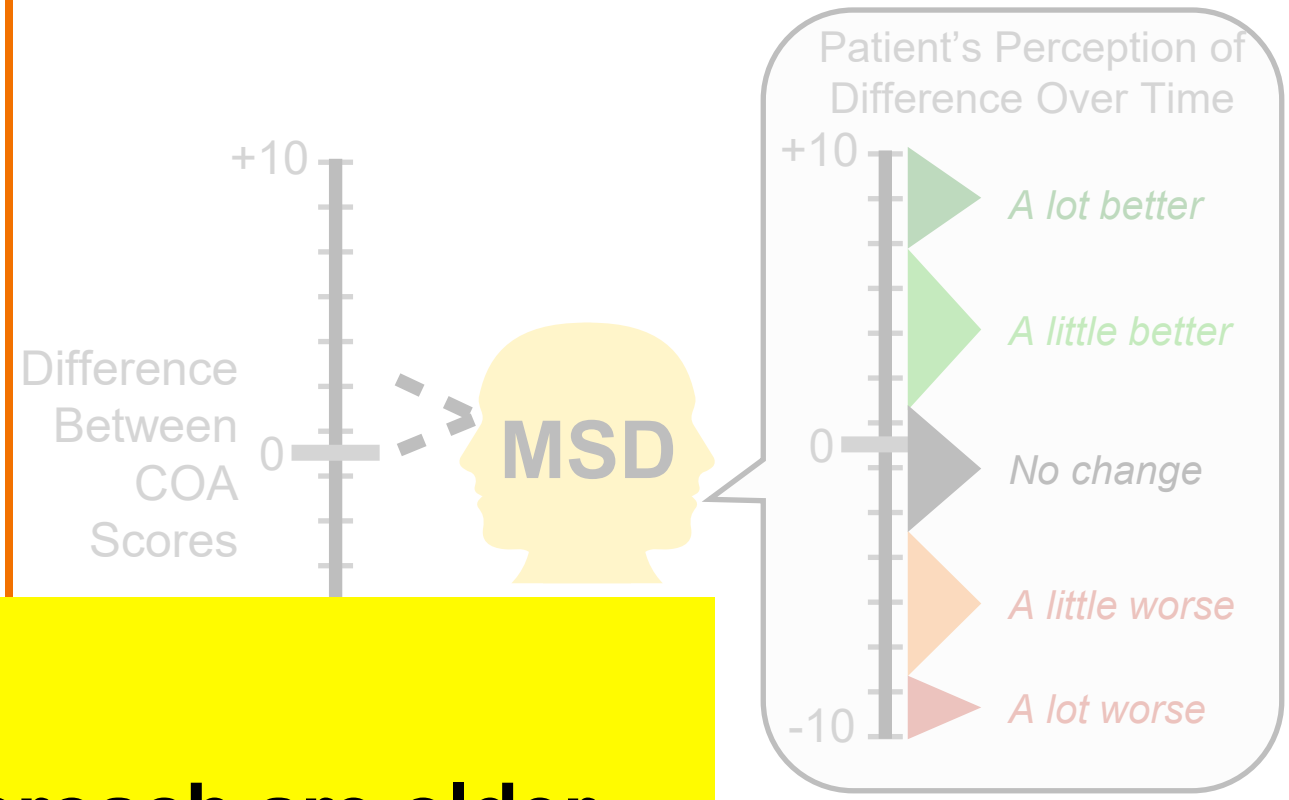
Meaningful Score Region (MSR) Approaches

Translate scores into something I understand better



Meaningful Score Difference (MSD) Approaches

Translate score differences to some other difference I understand better

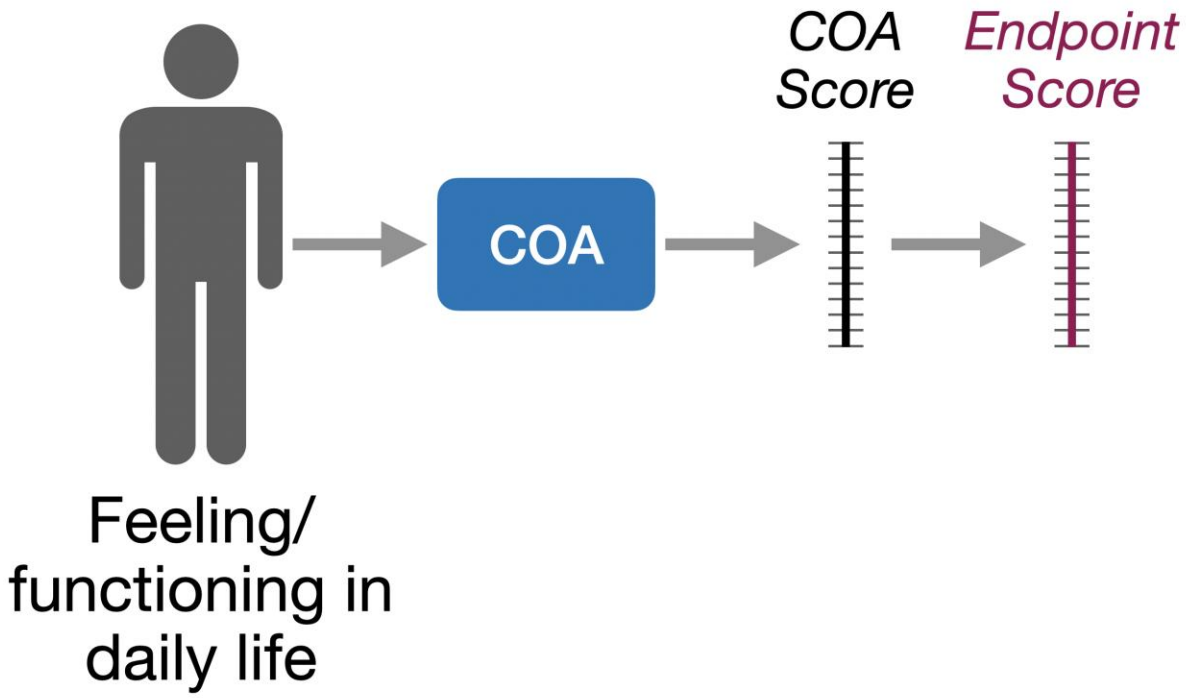


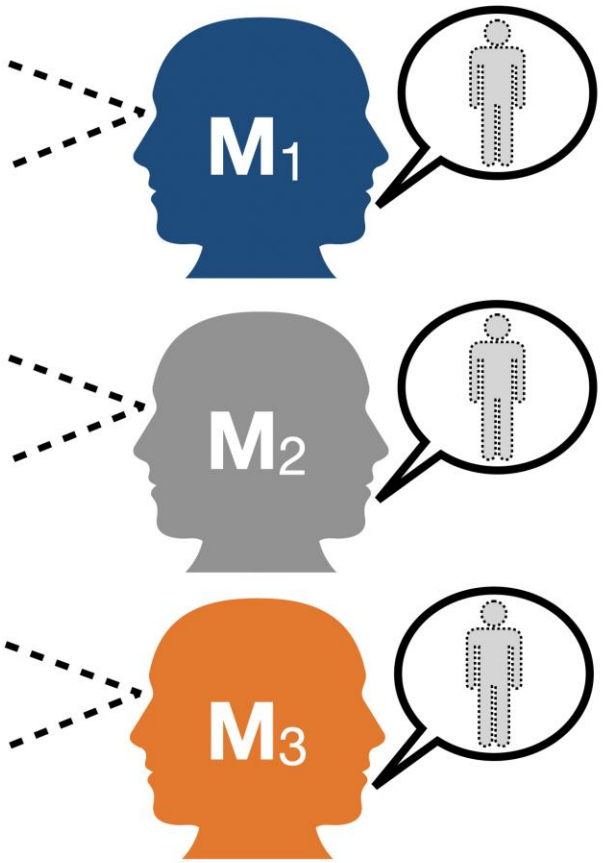
Within each approach are older, newer, and yet-to-be-developed methods.

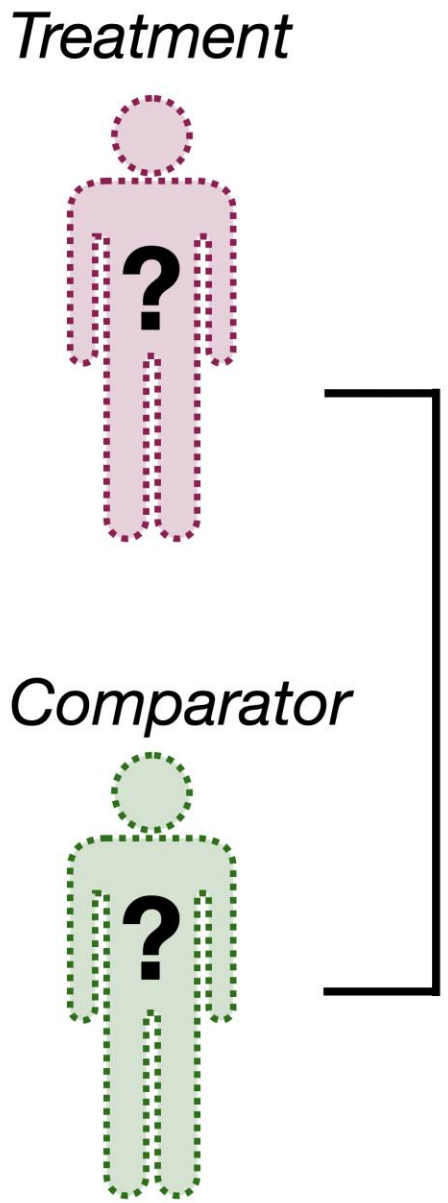
Criteria for the quality/rigor of a method in any context of use include:

- How well its assumptions are met in the context of use
- How well it reflects the uncertainty in its results
- How easily decision-makers can use the results to better understand how a patient's life might differ with vs. without treatment.

Multiple translators (methods) may provide a richer context for interpreting the treatment effect.

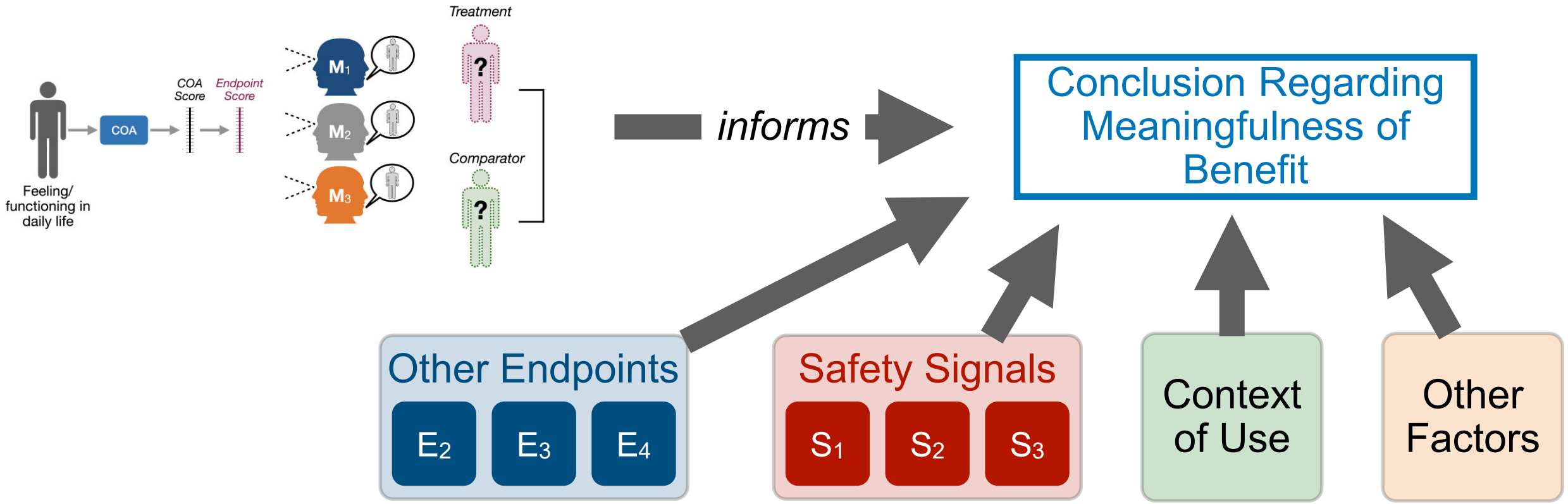






The result of applying any method will inform our belief that the treatment is doing something meaningful for patients.

But that belief is also informed by other things, such as results from other endpoints, safety signals, context of use, etc.

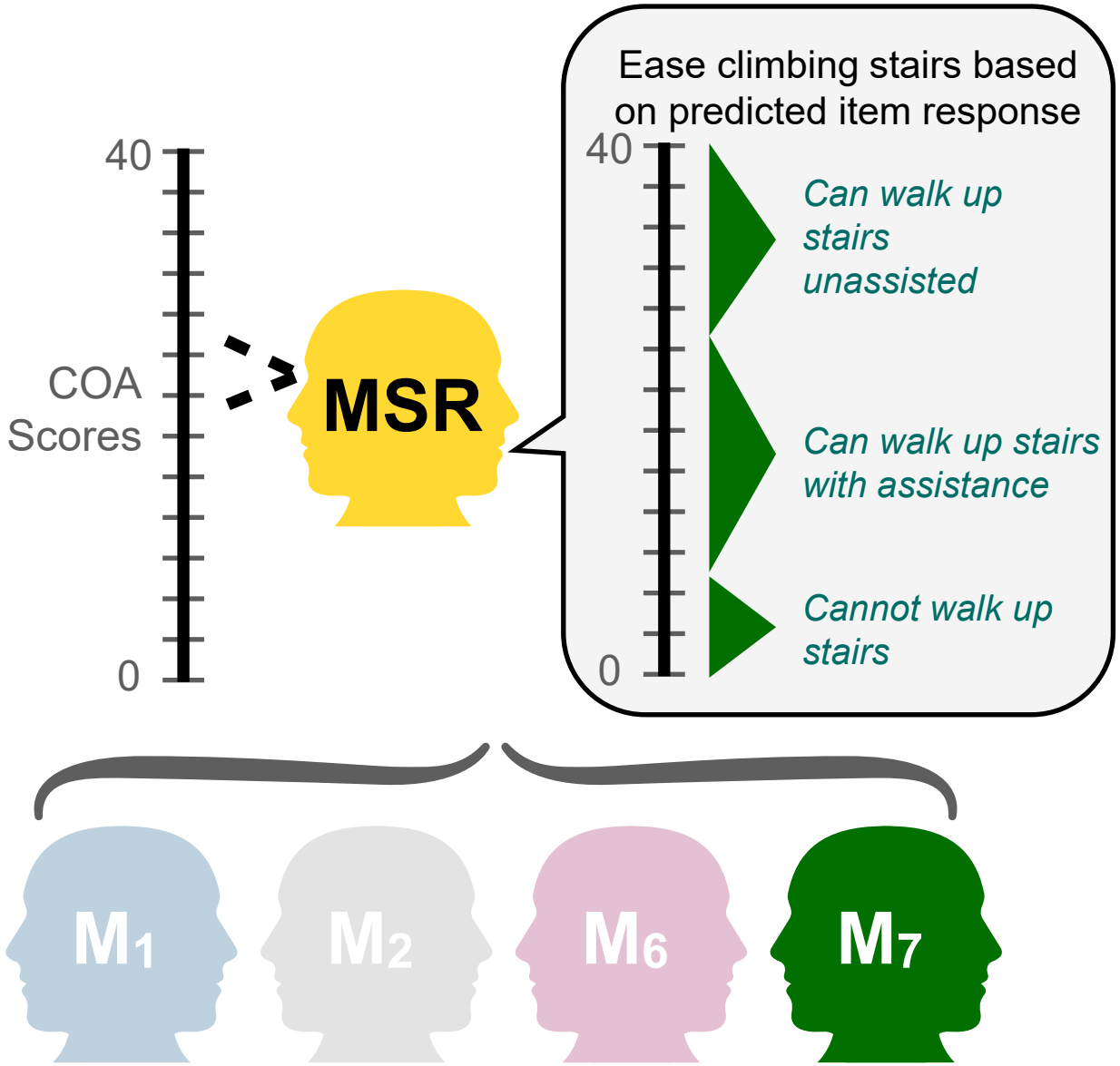


Key Points

- Interpreting treatment effects on COA-based endpoints often requires methods for translation of scores into something more relatable
- Over time, the field narrowed its choice of COA-based endpoint and methods of translation, but these approaches are not always the best fit
- PFDD Draft Guidance 4 invites greater flexibility in selecting, developing, and applying methods to support COA-based endpoint interpretation
- Different methods can be categorized by the approach they take to translation
  - **Translate scores into something more relatable (*Meaningful Score Region*)**
  - **Translate differences in scores into something more relatable (*Meaningful Score Difference*)**
- Results from these methods are just one part of the evidence used to evaluate the meaningfulness of a treatment benefit

Meaningful Score Region (MSR) Approaches

Translate scores into something I understand better



Rest of session will focus on Meaningful Score Region (MSR) approaches

Approaching MSR with “N.U.A.N.C.E”

- **N**ot always required.
- **U**nderstand and explain your rationale.
- **A**nchor selection matters.
- **N**on-Transferability: MSRs may not be universal.
- **C**ontent-based interpretation has value.
- **E**valuate uncertainty.

Approaching MSR with “N.U.A.N.C.E”

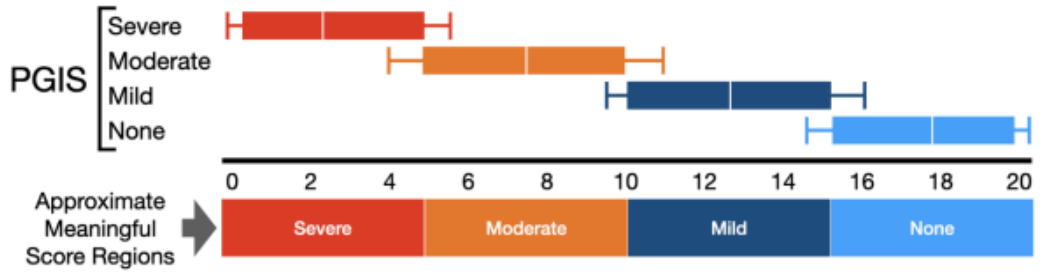
- **Not always required.**
- Understand and explain your rationale.
- Anchor selection matters.
- Non-Transferability: MSRs may not be universal.
- Content-based interpretation has value.
- Evaluate uncertainty.

Not always required

- Creation of approximate meaningful score regions based on the distribution of scores across severity levels

Once regions are defined, we can superimpose treatment effects to assess if, on average, treatment results in less severe disease/symptoms than the comparator

Figure 1. Example of Approach for Interpreting COA Scores in Terms of Meaningful Score Regions Corresponding to Patient Global Impression of Severity (PGIS).



Example MSR plots have no overlap between the severity categories so boundaries can be selected which separate the bulk of the distribution for scores within each severity

Figure 4. Least Squares (LS) Means Scores (With 95% Confidence Interval) on Functioning Measure Scores at Follow-up Time Point for Products A and B Relative to Meaningful Regions of Scores Based on Patient Global Impression of Severity.

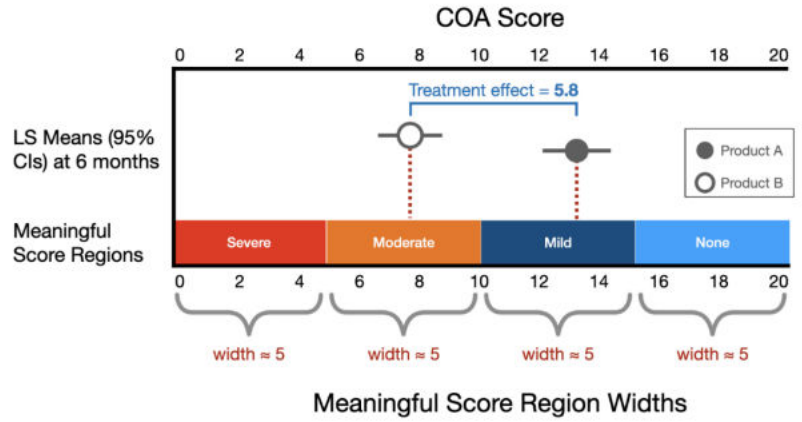
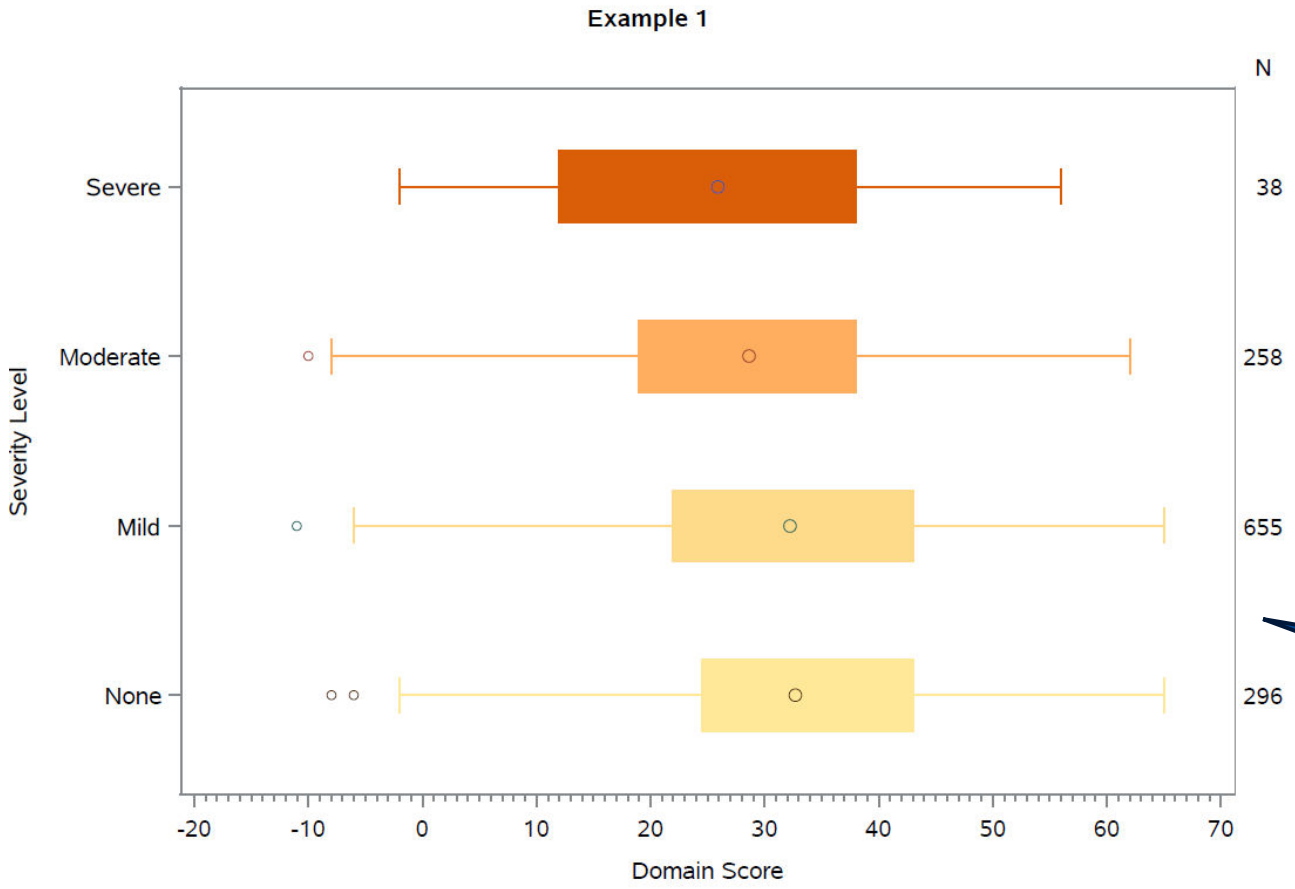


Figure Source: PFDD4 https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-incorporating-clinical-outcome-assessments-endpoints-regulatory

Three example MSR plots from the same trial

Example 1: Performance-based test vs patient rating of severity

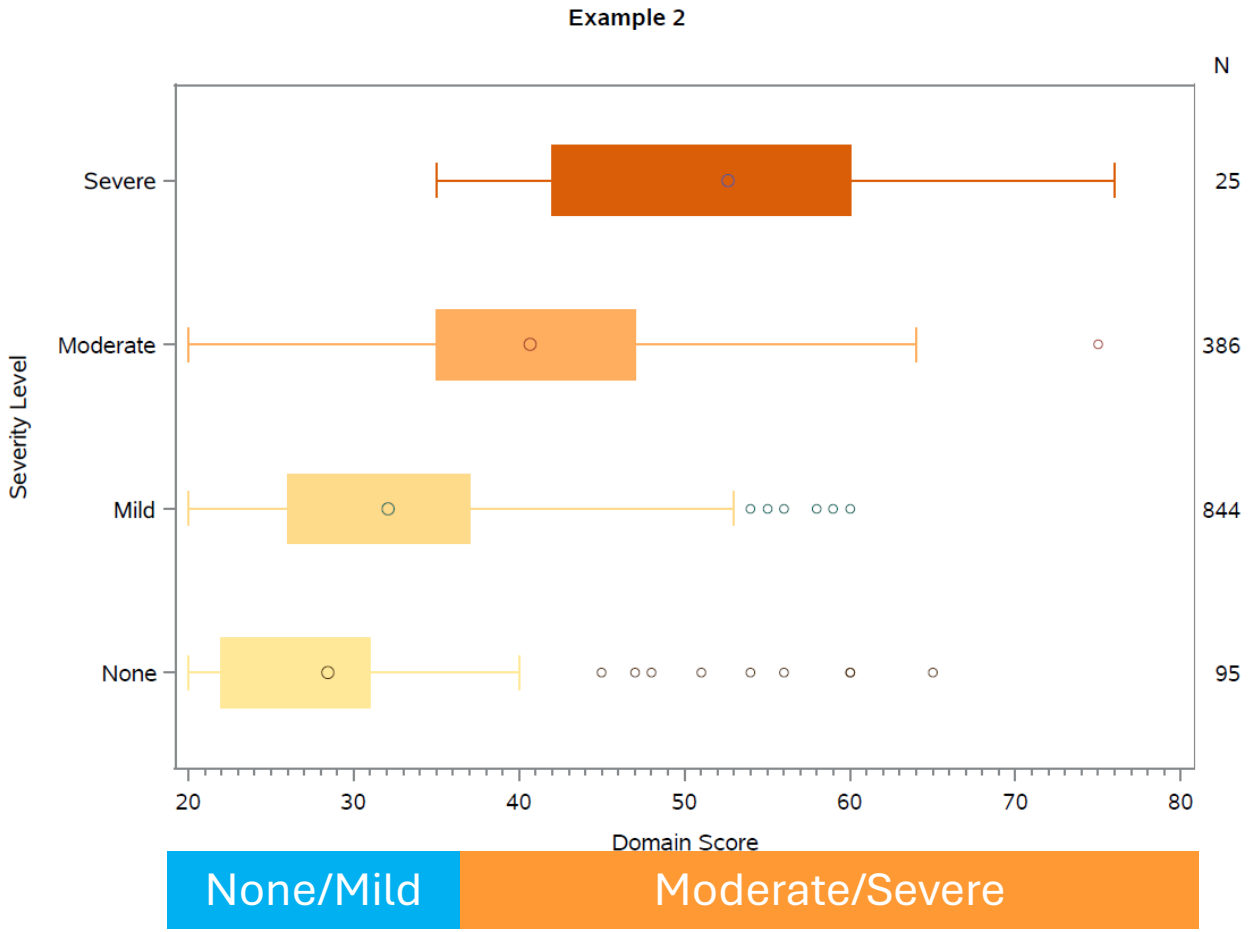


- Poor correlation of the anchor with the score (0.12)
- Lots of overlap between the severity categories
- Lots of variation in scores within severity categories
- Due to trial entry criteria and the aims of the treatment, patients likely to have milder disease (small sample size in severe rating)

Regions cannot be defined that clearly distinguish between severity levels

A more workable example…

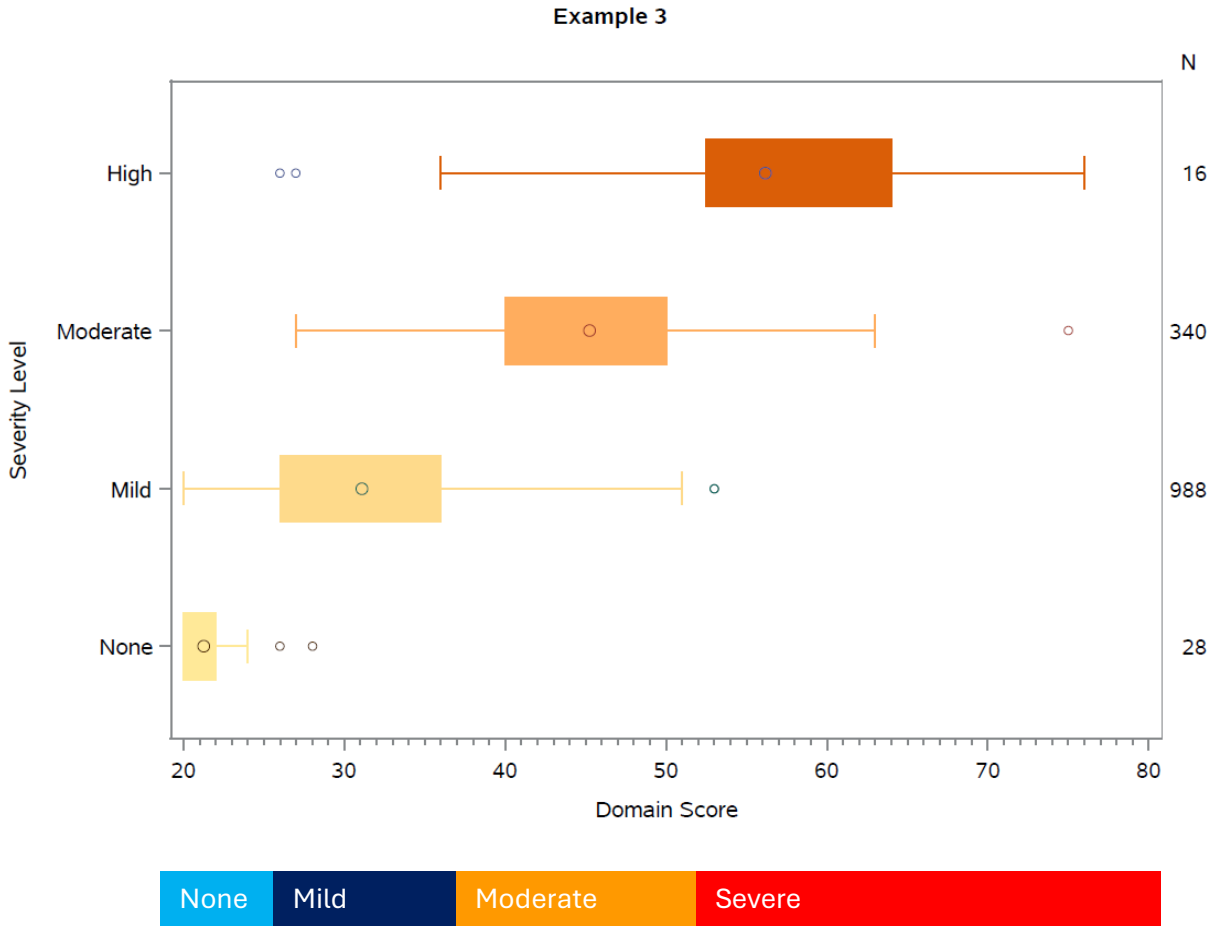
Example 2: Interviewer-based assessment vs a separate global rating of severity



- Sufficient correlation as an anchor (0.54)
- Still some overlap despite the strong correlation but could probably define boundaries
- May only be able to define None/Mild and Moderate/Severe regions due to overlap

A ‘textbook’ example

Example 3: Interviewer-based assessment with a global severity scale contained in the same



- The anchor is highly correlated at 0.82 (but also measured within the same instrument…circular?)
- Unrealistic for most anchors to be this highly correlated
- Unequal widths, does it matter?
- Narrow population (mainly mild or moderate)

When does it work, when is it not appropriate?

Are MSRs appropriate where treatments are expected to slow decline rather than improve the condition?

Even if there is a strong anchor and regions can be defined, are they useful?

Super-imposed group means in the same region may be a ‘good’ result interpreted as not meaningful

Do MSRs make more sense where we expect a treatment to improve a symptom or condition?

Knee surgery after injury → may expect a lack of function returning to normal function within a relatively short time frame

Asthma → may expect uncontrolled asthma to be controlled while on medication

Eczema → may expect a flare to return to something resembling normal skin

Are application of MSRs limited where we have narrow patient populations?

In trials we may focus on only a subgroup of severities entering the study

Can we expect anchors to be strong enough to define regions?

Do we need stronger than typically see for PGI-S and therefore different methods required to define the regions?

Approaching MSR with “N.U.A.N.C.E”

- **N**ot always required.
- **U**nderstand and explain your rationale.
- **A**nchor selection matters.
- **N**on-Transferability: MSRs may not be universal.
- **C**ontent-based interpretation has value.
- **E**valuate uncertainty.

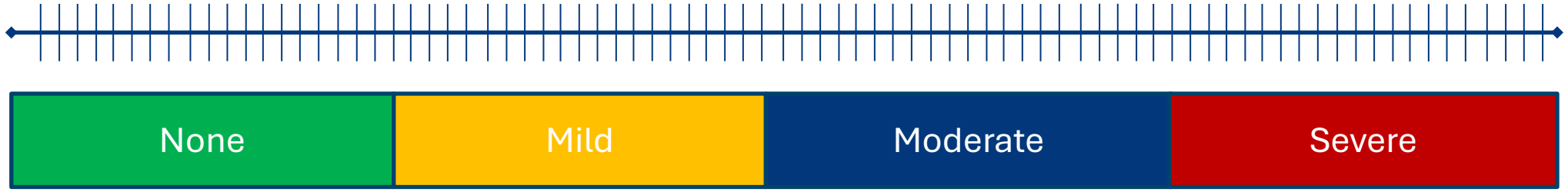
Understand and explain your rationale

- Rationale for inclusion of MSRs to aid interpretation of the treatment effect
  - How are patients talking about their disease/treatment?
  - What are their expectations of treatment?
- Rationale for NOT including MSRs may also be important
  - Upfront decision based on expectation of disease course/treatment impact on the disease course
  - After looking at whether there are suitable anchors
  - Once it is clear from MSR plots that regions cannot be defined
- Justify choice of anchor or multiple anchors to define MSRs
  - Are there accepted cut offs used to define severity in practice?
  - Patient-centred vs accepted clinical cut offs?
  - Is there qualitative data to support the anchor and meaning of the categories?

Approaching MSR with “N.U.A.N.C.E”

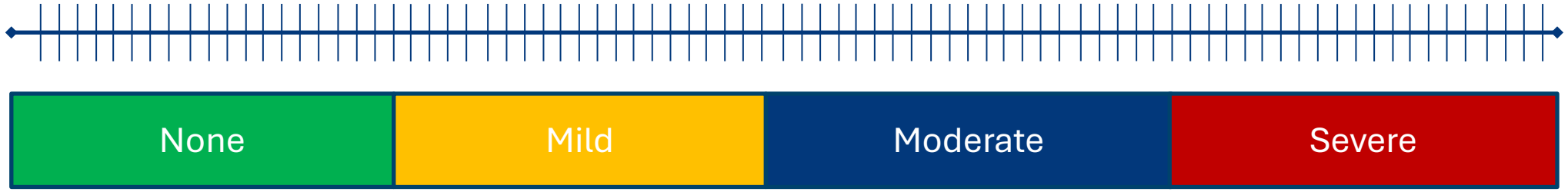
- **N**ot always required.
- **U**nderstand and explain your rationale.
- **A**nchor selection matters.
- **N**on-Transferability: MSRs may not be universal.
- **C**ontent-based interpretation has value.
- **E**valuate uncertainty.

Anchors to estimate MSRs



Patient global impression of severity (PGI-S)

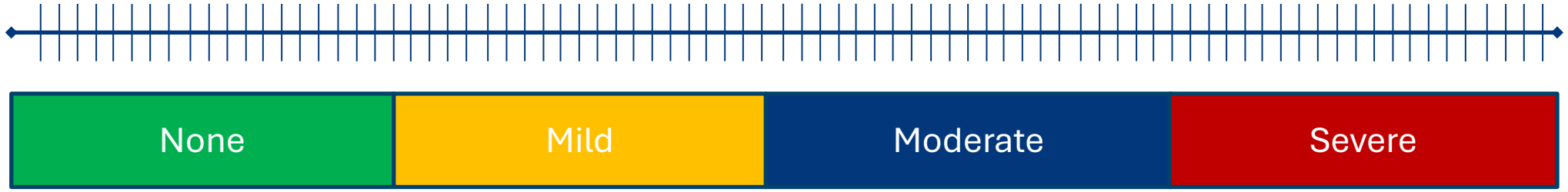
Anchors to estimate MSRs



Patient global impression of severity (PGI-S)

"if the treatment effect is larger than the width of the widest score region, this suggests that the treatment effect reflects a meaningful difference to patients and/or caregivers " Draft PFDD #4, lines 1212-14

Anchors to estimate MSRs



Patient global impression of severity (PGI-S)

"if the treatment effect is larger than the width of the widest score region, this suggests that the treatment effect reflects a meaningful difference to patients and/or caregivers " Draft PFDD #4, lines 1212-14

Anchors to estimate MSRs

$$Max\ MSR\ width \geq \frac{Scale\ range}{Number\ of\ anchor\ categories}$$

For a 4-category PGI-S, that is:

- At least 2.5 points for a 0-10 scale (e.g. Pain, Itch)
- At least 25 points for a 0-100 scale (e.g. EORTC, KOOS, SF-36)

Very rare to see such large treatment effects

CRITICAL PATH INSTITUTE

Anchors are not the only option to estimate MSRs

Approaching MSR with “N.U.A.N.C.E”

- **N**ot always required.
- **U**nderstand and explain your rationale.
- **A**nchor selection matters.
- **N**on-Transferability: MSRs may not be universal.
- **C**ontent-based interpretation has value.
- **E**valuate uncertainty.

Approaching MSR with “N.U.A.N.C.E”

- **N**ot always required.
- **U**nderstand and explain your rationale.
- **A**nchor selection matters.
- **N**on-Transferability: MSRs may not be universal.
- **C**ontent-based interpretation has value.
- **E**valuate uncertainty.

Approaching MSR with “N.U.A.N.C.E”

- **N**ot always required.
- **U**nderstand and explain your rationale.
- **A**nchor selection matters.
- **N**on-Transferability: MSRs may not be universal.
- **C**ontent-based interpretation has value.
- **E**valuate **uncertainty.**

Evaluate uncertainty

Treatment Response vs Meaningful Score Regions

Mean Placebo:

15

95% CI Placebo (Width):

2

Mean Treatment:

18

95% CI Treatment (Width):

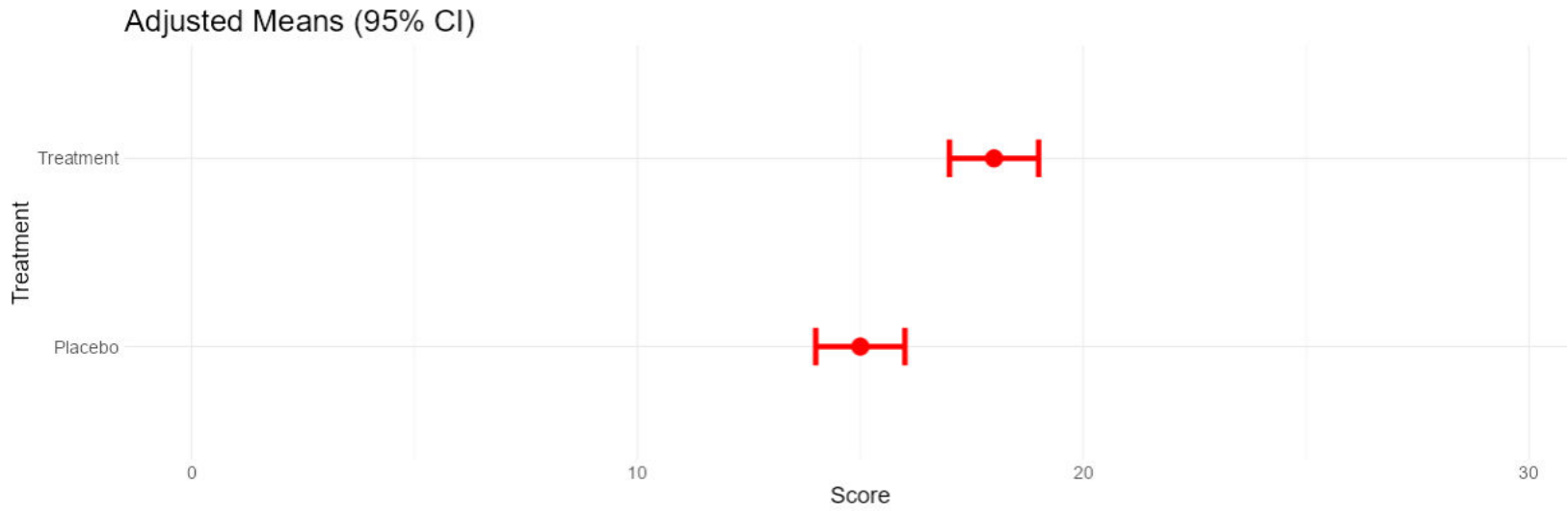
2

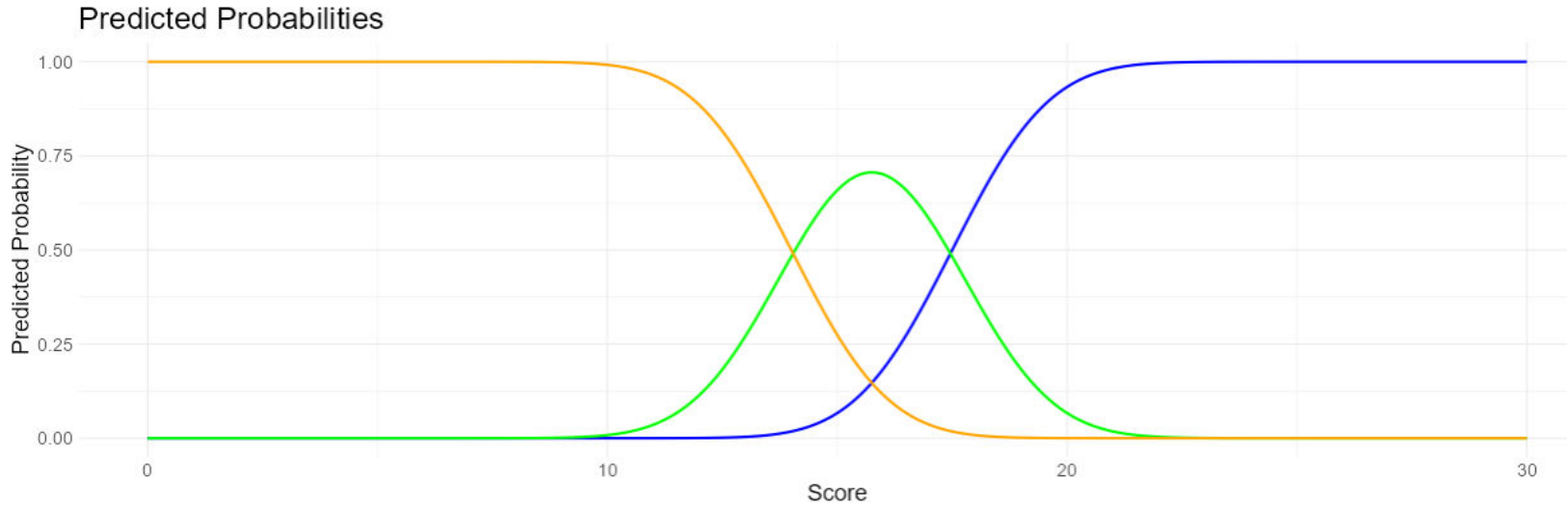
Threshold for Region 1:

17.5

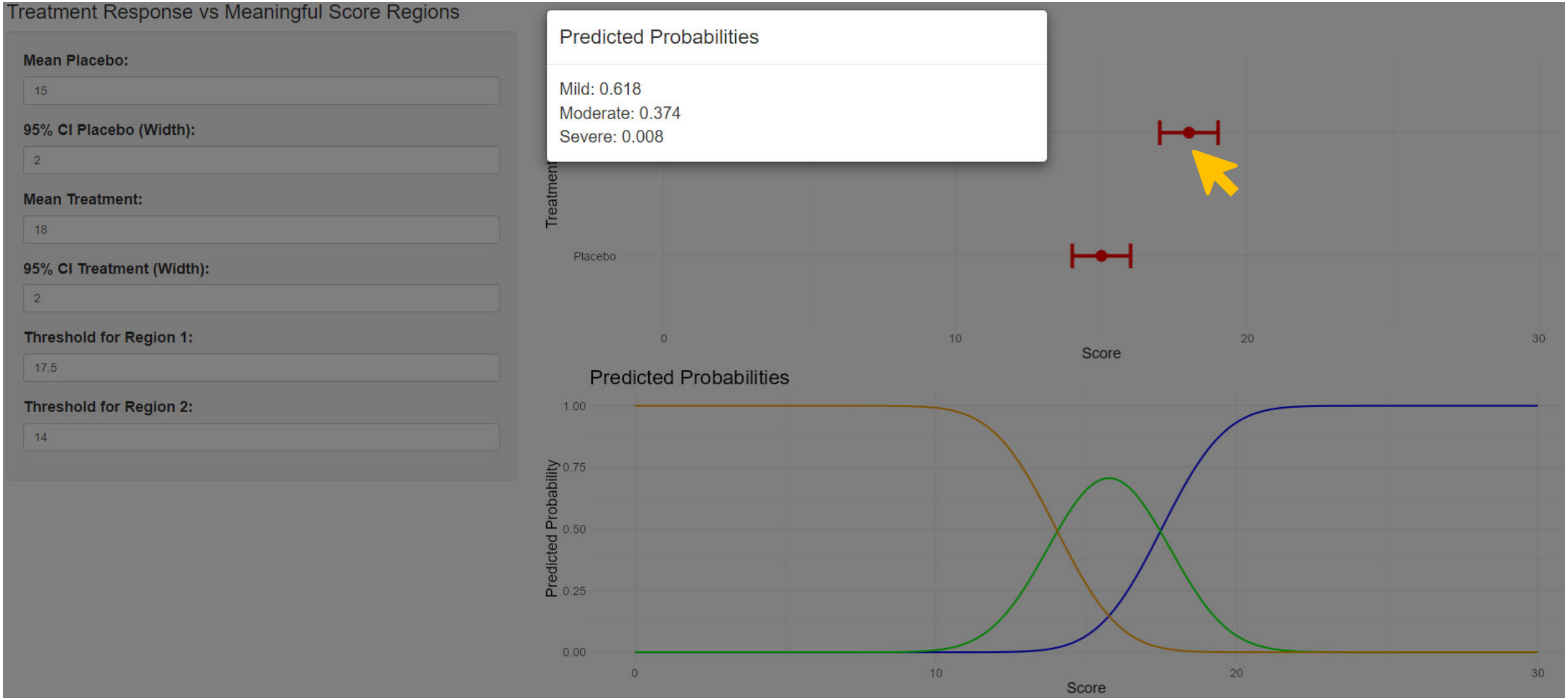
Threshold for Region 2:

14





Evaluate uncertainty



Approaching MSR with "N.U.A.N.C.E"

- **N**ot always required.
- **U**nderstand and explain your rationale.
- **A**nchor selection matters.
- **N**on-Transferability: MSRs may not be universal.
- **C**ontent-based interpretation has value.
- **E**valuate uncertainty.

Session Participants and Q&A

Moderator:

- *Fraser Bocell, MEd, PhD* – Critical Path Institute

Presenter:

- *Kevin P. Weinfurt, PhD* – Duke University School of Medicine

Panelists:

- *Kim Cocks, PhD* – Adelphi Values

- *Andrew Trigg, MSc* – Bayer

Thank You!

Advancing Drug Development. Improving Lives. Together