



June 29, 2023

Dockets Management Staff (HFA-305)  
Food and Drug Administration  
5630 Fishers Lane, Rm. 1061  
Rockville, MD 20852

**Re: Docket No. FDA-2023-D-0026 Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints for Regulatory Decision-Making**

Dear Recipient:

The Biotechnology Innovation Organization (BIO) thanks the Food and Drug Administration (FDA or Agency) for the opportunity to submit comments regarding the Draft Guidance Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments.

BIO is the world's largest trade association representing biotechnology companies, academic institutions, state biotechnology centers and related organizations across the United States and in more than 30 other nations. BIO's members develop medical products and technologies to treat patients afflicted with serious diseases, to delay the onset of these diseases, or to prevent them in the first place.

BIO and its members appreciate the opportunity to work with the Agency to develop and align on approaches that are robust, practical, and expedite patient-focused drug development. To further enhance the Guidance, we believe that a few areas would benefit from more explanation, examples, and references. We have identified through our comments areas where these additions would be beneficial.

### **General Comments**

We commend the Agency for developing guidances where there is alignment across CDER, CBER and CDRH. The PFDD guidance series represents input from all divisions; however, we have experienced that the use of patient experience data (PED) and recommendations for clinical outcome assessment (COA) endpoints have not been implemented consistently. We suggest FDA increase COA and statistical methods expertise across review divisions and/or utilize the Division of Clinical Outcome Assessment (DCOA) and other biostatistics teams with psychometric expertise in the review of COA data to ensure consistency of feedback with the



recent guidance documents and across divisions. We also recommend FDA have processes in place to share knowledge relevant to the guidance series across Centers as well as more broadly with interested stakeholders.

Regarding the guidance, we think it could benefit from including more specific examples, ensuring more consistency throughout the document, and communicating concepts in a way that can be understood by all stakeholders, particularly patients. We also note that the guidances in this series contain very technical language that may make it hard to understand, especially for patients. We suggest that FDA provide a summary version of the guidance that provides main highlights and visuals, which are more friendly to patients and other stakeholders. We believe this recommendation aligns with FDA's CDER guidance snapshot pilot.<sup>1</sup>

Overall, we appreciate the statement that this series of guidance documents, once finalized, will replace the 2009 PRO Guidance; it provides much-needed clarity. The draft guidance from December 2009 clearly specified that the principles presented there were relevant for labeling claims. It is currently unclear whether the principles included in Guidance 4 are also applicable only to COA endpoint intended for labeling claims, or whether it is the intent for this guidance to be applicable more broadly. To ensure clarity for all Stakeholders, the Agency should include specific language stating that this guidance applies to all regulatory decision-making and not specifically for labeling.

1. COAs can be used as endpoints to assess safety and tolerability. The guidance speaks to additional considerations needed when the COA is to inform risk, but these considerations are not listed, or a reference is not provided on where these considerations can be found. We recommend the guidance specifically state whether a COA to inform risk is in scope of this guidance. If in scope, we suggest providing some reference on what these considerations are and evidentiary needs for inclusion in label. and/or regulatory decision-making. It is important that sponsors have guidance on how to incorporate clinical outcome assessments into endpoints that cover the full range of patient experience, including safety and tolerability.
2. As clinical assessments can be used as screening tools, we suggest the Agency include further guidance on approaches to identify thresholds on clinical assessments for patient inclusion for trial enrichment including discussion of maintaining a study sample that generates efficacy findings that can be generalizable to the larger disease population.

---

<sup>1</sup> See <https://www.fda.gov/drugs/guidances-drugs/guidance-snapshot-pilot> and <https://www.fda.gov/drugs/news-events-human-drugs/using-innovative-communication-methods-increase-awareness-and-understanding-cder-guidance-documents>



3. We recommend that the guidance be expanded to address COA endpoints for use in single-arm trials, including open-label trials, and provide recommendations on how to address inter-patient variability in trials without randomization or where blinding is infeasible. We note that the FDA has published a retrospective analysis assessing the impact of knowledge of treatment assignment in multiple myeloma in which they found no evidence of a meaningful impact on how patients reported symptoms, function or health status.<sup>2</sup> Given this, FDA's Oncology Center of Excellence has recommended that PROs be included throughout the drug development continuum, regardless of the stage of development or trial design.<sup>3</sup> Therefore, we request FDA to explicitly recognize in this guidance that COAs can be considered valid and meaningful even in the absence of placebo-controlled, blinded trials.

Lastly, we appreciate the discussion on endpoint strategies when a disease affects multiple aspects of feeling and functioning, in particular the discussion on multi-component endpoints and personalized endpoints. This is critical to allow sponsors to design COAs and appropriate COA measurement strategies for diseases that have heterogeneous manifestations.

## **FDA – Sponsor Engagement on Clinical Outcomes Assessment Development**

We appreciate that the guidance recommends meeting early with FDA to discuss collection of patient experience data generally and COAs specifically. However, we note that it can often be difficult to obtain a meeting in a timely manner or it is unclear which meeting type is appropriate to share developments and evidence generation plans and data with the Agency. We also appreciate that during the May 4th, 2023, webinar on this guidance, FDA mentioned the different meeting types that could be utilized to obtain advice on a new COA. We recommend that the guidance expand on this topic by including the information from the webinar as well as specific recommendations on how each formal meeting type could be utilized for COA discussions.<sup>4</sup> This type of information could help sponsors navigate which meeting and what time points are most appropriate for the discussion they wish to have with the Agency, potentially reducing the number of meeting requests FDA receives that are out of scope or deemed inappropriate. We also suggest that the “Formal Meetings Between the FDA and Sponsors or Applicants of PDUFA Products” be referenced in this guidance.

---

<sup>2</sup> Roydhouse JK, Mishra-Kalyani PS, Bhatnagar V, et al. “Does Knowledge of Treatment Assignment Affect Patient Report of Symptoms, Function, and Health Status? An Evaluation Using Multiple Myeloma Trials.” *Value Health*. 2021 Jun;24(6):822-829.

<sup>3</sup> Bhatnagar V, Kluetz PG. “Encouraging Rigorous Patient-Generated Data All along the Drug Development Continuum.” *J Natl Cancer Inst*. 2022 Jul 28.

<sup>4</sup> BIO FDA-Sponsor Engagement Framework for COA Development

[https://www.bio.org/sites/default/files/2023-](https://www.bio.org/sites/default/files/2023-06/BIO_FDA_Sponsor_Engagement_Framework_for_COA_Development.pdf)

[06/BIO\\_FDA\\_Sponsor\\_Engagement\\_Framework\\_for\\_COA\\_Development.pdf](https://www.bio.org/sites/default/files/2023-06/BIO_FDA_Sponsor_Engagement_Framework_for_COA_Development.pdf)



## Harmonization Across Guidances

As there are many existing FDA guidances that are relevant to the concepts in this guidance, we suggest stronger linkages across guidance documents including, but not limited to:

- Qualitative patient input was extensively described in Guidances 1 and 2 and Guidance 4 should tie how qualitative patient input can inform interpretation of COA endpoints.
- Item Response Theory (IRT) and Computerized Adaptive Testing (CAT) was discussed in draft Guidance 3 and would be applicable to some of the discussions in Guidance 4.
- DHT guidance - While we understand that the FDA is, understandably, reticent to introduce new classifications such as adding digital health technologies (DHTs) to the list of COA types, it is confusing that the DHT-related and COA-related guidance documents are being released around the same time, address many of the same validation topics, but do not explicitly reference each other. It would be helpful if this guidance document was more explicit in the discussion of DHT-enabled endpoints, the related tools, and how (or if) fitness-for-purpose of those tools will be evaluated from a COA perspective. Additionally, we recommend that the Agency discuss how DHTs can complement COAs within the guidance by referencing FDA's [Framework for the Use of DHTs in Drug and Biological Product Development \(fda.gov\)](#)
- There seems to be a lack of alignment with the estimand framework throughout the draft guidance. It does not really follow the thinking process laid out in the ICH E9 (R1) estimands addendum and primarily focuses on estimation and technical details, without requiring to first specify a proper estimand. Furthermore, there is no discussion on the importance of patient input into intercurrent events.

## Terminology

We note that there is no mention of the COA dossier in the draft guidance. We request that the Agency please comment if sponsors should continue to submit one for any primary and key secondary COA-based endpoint. If the answer is yes, it would be helpful if the Agency could confirm the structure and content-related expectations for the COA dossier.

We also recommend that FDA update the [PFDD glossary](#) to ensure consistency in use of terms across FDA's PFDD-focused guidances and resources, and additionally that FDA cross-reference the glossary, as appropriate, in the current draft guidance. We have noted in the table below certain terms that would benefit from clarification and perhaps this can be done in conjunction with updates to the PFDD glossary and with reference to the [BEST \(Biomarkers, Endpoints, and other Tools\) Resource](#).

We appreciate that FDA appears increasingly open to a number of strategies for constructing COA based endpoints and assessing meaningful change. That said, we found that the draft



guidance introduced many new methodologies and terminologies rather than leverage existing literature and consensus documents. We note that general alignment on meaningful change methodology has been emerging amongst experts through groups such as ISOQOL, ISPOR, and the C-Path PRO Consortium. Given the complexities in this field, we believe it is important that FDA elaborate on how any new concepts introduced in the guidance relate to previously used terminology such as ‘within-patient meaningful change’ and ‘between-groups meaningful difference’.

In addition, we appreciate the thoughtful discussion on how meaningful score differences (MSD) can be estimated and applied. Since MSD does not refer to “change”, we believe this will help enable a broader understanding of what constitutes a meaningful treatment effect (e.g., in some cases, prevention of worsening may be identified as a meaningful treatment effect). We recognize, however, that introducing new terminology may create confusion and recommend that FDA elaborate on how the MSD concept can be used for a more comprehensive characterization of meaningful treatment effect.

We strongly suggest changing the terms so that they aligned with the terms in FDA Guidance 4 Discussion by replacing “MSD” by “Meaningful Within-Patient Change (MWPC).” MSR may pertain to group differences and more distinctly referred to a meaningful group-level difference (MGLD), which pertains to within-group difference over time or between-group difference at a given time. MWPC (or MSD) is the threshold for responder analysis or time-to-event (e.g., deterioration), and it should not be used to interpret the group-level difference such as treatment-effect between two arms. This threshold on a COA is a way for interpreting the magnitude of within-patient change score over two separate times that an individual patient would consider a meaningful improvement or deterioration.

In contrast, MGLD could be used to interpret the treatment effect between two arms or between two time points within the same arm. The MGLD threshold is a way for interpreting the magnitude of a between-group difference in mean scores at a given time, or a within-group difference in mean scores at two times, that patients would consider a meaningful improvement or deterioration.



BIO recommends revising Section III.C to apply the MSD to examine responder percentages per treatment group or perhaps it can be more informative to provide the CDFs so that readers of the label can see the comparison at all levels of change<sup>567</sup> and not just the at the MSD.

## Statistical Analyses of COA scores

We appreciate the openness to new methods and willingness to consider different types of evidence to support COA score interpretation. In the context of medical product development, efficiency is crucial to meeting unmet medical needs. If there is a way to be more specific about which analyses reviewers would prefer to see for certain types of data, that would be welcomed in the final guidance document; this specificity would allow for more precise planning for data analysis.

This draft guidance is focused on modern statistical methods and increased statistical rigor, which we feel is a step in the right direction. In the entire series of PFDD guidance documents, the agency cited many references from the social/behavioral science and education literature. In social/behavioral sciences and education (and statistics and medicine), there has been a movement towards placing more emphasis on confidence intervals (focusing on range of scores) and less emphasis on statistical significance testing (i.e., p values) (e.g., APA, 2019; Cohen, 1994; Kline, 2004; Sterne et al., 2001; Wasserstein et al., 2019). As the guidance series seems to be quite heavily influenced by the social/behavioral science and education research methodology and statistical methods, it would be helpful to understand whether it is reasonable to expect the agency to focus more on range of scores and confidence intervals (and less on p values) in their review of endpoints and COA evidence for regulatory decision making. We recommend Guidance 4 include more information on the agency's view on this topic. Additionally,

---

<sup>5</sup> Coon CD, Cappelleri JC. Interpreting change in scores on patient-reported outcome instruments. *Therapeutic Innovation & Regulatory Science*. 2016; 50:22-29.

<sup>6</sup> Griffiths P, Sims J, Williams A, Williamson N, Cella D, Brohan E, Cocks K. How strong should my anchor be for estimating group and individual level meaningful change? A simulation study assessing anchor correlation strength and the impact of sample size, distribution of change scores and methodology on establishing a true meaningful change threshold. *Quality of Life Research*. 2022. Online. (Note: the term minimal or minimally or minimum is fraught with danger and should not be used for the reason given in the McLeod et al., Figure 4.)

<sup>7</sup> McLeod LD, Cappelleri JC, Hays RD. Best (but of forgotten) practices: Expressing and interpreting meaning and effect sizes in clinical outcome assessments." *The American Journal of Clinical Nutrition*. 2016; 103:685-693. <https://doi.org/10.3945/ajcn.115.120378>. (Erratum: *The American Journal of Clinical Nutrition*. 2017; 105:241. <https://doi.org/10.3945/ajcn.116.148593><https://doi.org/10.3945/ajcn.116.148593><https://doi.org/10.3945/ajcn.115.120378><https://doi.org/10.3945/ajcn.116.148593>.)



clarification for when p-values will be important (i.e., hypothesis testing for endpoints) versus evidence generated for the interpretation of change in a COA would be helpful.

Broadly, there seems to be a push towards greater statistical rigor in FDA Guidance 4. This draft version refers to well-established statistical principles and literature and relies upon that statistical literature to make recommendations. Because COA analysis involves fitting statistical models and interpreting statistical model output, it is imperative that FDA COA guidance be developed in conjunction with psychometricians to ensure alignment with psychometric principles/literature. Additionally, psychometricians specially trained on these methods should provide guidance to statisticians and mathematicians in the FDA who may be involved in the review of COA evidence either through DDT Qualification process or through the IND/BLA review. However, there is a need to further incorporate statistical methods into the final guidance document. For example, the FDA noted several types or approaches to “personalized” COA assessments but provided limited detail on appropriate statistical modeling for each approach. Additionally, we note that the Guidance 4 “personalized COA approach” examples did not include CAT testing, as noted in the Agency’s Duke Margolis Center for Health Policy 2017 reference: We ask the Agency to clarify that the exclusion of CAT, as a “personalized approach” reflects a necessity for document brevity or changed/updated guidance as to CAT relevance within the personalized framework and interpretation of findings across the group. Thus, Guidance 4 could be amended to further incorporate statistical methods in a way that would help achieve the research objectives.

The methods to aid in the interpretation of treatment effect between clinical trial treatment groups described in Section III are new and less focused on meaningful within-patient change. Conceptually this is a change to the way that within patient change has been used for quite some time. It would be helpful for the Agency to explain this shift to the new methods more clearly in the guidance. For instance, additional clarity is needed regarding whether apply the draft guidance is suggesting meaningful change assessments be applied to between-groups efficacy assessments or are within-patient change assessments still relevant evidence to generate, including supportive references if this is deemed appropriate by the Agency. The Agency provides a variety of approaches to identify meaningful change thresholds, including estimating ranges and regions. While we appreciate the flexibility to utilize various methods, it would be helpful to clarify under which circumstances each approach would be recommended by the Agency. The acknowledgment that “any choice of threshold MSD that attempts to distinguish between meaningful and non-meaningful differences will not correspond to some patients’ experiences” (lines 921-922) is highly appreciated, as is the consideration of applying a range of MSD values as opposed to a single value. Nevertheless, the guidance indicates that an MSD must be established for each COA for each specific combination of patient population, baseline status, and direction of change. This work is highly resource intensive, particularly if the MSD must be specific to a disease, patient population, background standard of care, location, calendar time, COA version, endpoint, and length of follow-up (lines 1051-1053). The resources required



to re-establish an existing instrument's measurement properties (including, but not limited to, MSD) in each specific context of use may have the negative unintended consequence of discouraging the inclusion of patient-centered outcomes in clinical trials. We recommend the guidance discuss which measurement properties must be re-established for a specific context of use in Phase 2 trials before being considered fit for purpose in Phase 3 trials after a COA tool has undergone rigorous development in the general population or a relatively broad clinical population (e.g., cancer).

The MSD and MSR approaches to determining meaningful change may be appropriate in idealized settings (i.e., consistency across baseline or similarly sized regions). However, we have concerns that they may not be feasible in other settings such as neurodegeneration where often the aim is to slow progression. More traditional approaches based on within-patient meaningful change may be needed in such settings.

In addition, the meaningful change sections, as written, could be confusing to sponsors and other stakeholders. We recommend that a public hearing on the topic be convened to foster an improved understanding and inform the development of the final guidance document.

### **Conclusion**

BIO appreciates this opportunity to submit comments regarding the draft guidance Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints for Regulatory Decision-Making. Specific, detailed comments are included in the following chart. We would be pleased to provide further input or clarification of our comments, as needed, and we look forward to future opportunities to collaborate with the Agency on this critical topic.

Sincerely,

/s/

Neil Ichiro Laruan

Manager, Science & Regulatory Affairs  
Biotechnology Innovation Organization





## BIO Comment Table of Proposed Changes

LINE/ SECTION	ISSUE	PROPOSED CHANGES
25-26	Type of assessment(s) made (e.g., Patient-Reported Outcome (PRO) measures, Observer Reported Outcome (ObsRO) measures, Clinician-Reported Outcome (ClinRO) measures	PRO and ClinRO acronyms used prior
Line 39 and Line 145	<p>“Methods, standards, and technologies for collecting and analyzing COA data for regulatory decision-making.”</p> <p>“maximum value of the daily 200 mobile sensor assessments for 7 days prior to the week 36 study visit”</p>	<p>Appreciate the agency calling out the use of technologies and including mobile sensor assessments as an example. However, there is a lack of discussion or reference to other FDA guidances about considerations for using such modalities and assessments in drug development, or how sensor assessments fit into the spectrum of COAs. May also be helpful to discuss how these can complement COAs.</p>
49	<p>We appreciate FDA’s efforts to provide guidance that will allow sponsors to better serve patients by integrating the patient experience into all aspects of medical product development. While methodological guidance is helpful, it remains unclear how FDA will communicate Agency thinking about patient experience data submitted in support of a development program.</p>	<p>We appreciate FDA’s encouragement to interact early with FDA to obtain feedback, but we strongly urge FDA to explain which transparency measures will be implemented to provide insights for sponsors, patients, and other appropriate subject matter experts about how these data were considered to support regulatory decision-making. Such transparency will support industry and regulators alike to advance and increase regulatory acceptance of PFDD generally.</p> <p>We also recommend that the document indicate how early in the development process sponsors should engage with the agency, and with which part of FDA sponsors should engage (e.g., the review division).</p>



<p>51-55</p>	<p>Original text:</p> <p>“FDA recommends that stakeholders engage with patients and other appropriate subject matter experts (e.g., clinical and disease experts, qualitative researchers, survey methodologists, statisticians, psychometricians, patient preference researchers) when designing and implementing studies to evaluate the burden of disease and treatment, and perspectives on treatment benefits and risks.”</p> <p>Consistent with the rest of the guidance where “patients and/or caregivers” is used, we recommend using that language here as well.</p>	<p>BIO recommends the following proposed change:</p> <p>“FDA recommends that stakeholders engage with patients (<u>and/or caregivers if appropriate</u>) and other appropriate subject matter experts (e.g., clinical and disease experts, qualitative researchers, survey methodologists, statisticians, psychometricians, patient preference researchers) when designing and implementing studies to evaluate the burden of disease and treatment, and perspectives on treatment benefits and risks.”</p>
<p>64</p>	<p>Provide clarification regarding the term “survive”</p>	<p>Guidance 3 includes a clarifying note Suggestion to include the same note in Guidance 4</p>
<p>77</p>	<p>“Section III of this guidance describes methods to aid in the interpretation of treatment effects on COA-based endpoints in terms of patients’ views on the effect of a medical product.”</p>	<p>Recommend ending the sentence at "endpoints" the guidance and section III applies to all COAs. It may be misleading to describe COAs as “patients' views” since most are used to compare self-reported experiences at different timepoints, but not views (exceptions are COAs such as PGIC).</p>
<p>85-86</p>	<p>Though the text and examples in this guidance focus mostly on treatment benefit (e.g., improvement in disease-related symptoms or impaired functions),</p>	<p>Benefit is not always improvement but may be delay or prevention of worsening. Suggest expanding example to “(e.g., improvement or delay of worsening in disease-related symptoms or impaired functions).</p>
<p>86</p>	<p>It is important to emphasize that in addition to improvements and treatment risks, COA</p>	<p>Please consider adding language:</p> <p><u>In addition, COAs can be used to demonstrate that treatments do not</u></p>



	<p>may focus on no deterioration of symptoms or quality of life impacts.</p>	<p><u>compromise quality of life or that two treatment arms are equivalent for certain symptoms</u></p>
88-90	<p>The document should contain more details on establishing benefit-risk and additional consideration to inform treatment risk.</p> <p>Clarification is requested on the following point: The COA should help assess AEs regardless of whether they are symptomatic (namely, AEs based on laboratory abnormalities should be included).</p>	<p>Recommend that the document provide details on how considerations for establishing benefit-risk differ from those for establishing benefit. Also, recommend clarification of any connection with outcomes like “worsening or deterioration.”</p>
117-118	<p>Please define “meaningful” and clarify whether this is intended to be “clinically meaningful” or broader, as meaningful to patients as determined by qualitative research, etc.</p> <p>Item 1 requires selection of endpoints that should reflect an aspect of the patient’s health that is “meaningful”. Meaningful is broad in the context of selecting endpoints.</p>	<p>“Generally, endpoints that are based on COAs should (1) reflect an aspect of the patient’s health that is <u>clinically meaningful, or qualitatively meaningful to a patient’s HRQoL</u>”</p> <p>Recommend considering specifics or providing an example for a meaningful vs not-meaningful endpoint</p>
120	<p>The early mention of “mean score at 12 weeks” wording will have readers and sponsors confused about why the FDA isn’t concerned about how much the score has changed.</p> <p>While it is admirable that FDA is attempting to create an understanding of the difference between COA measures and COA endpoints, this example, which is a</p>	<p>Recommend that the guidance either introduce the new perspective on mean change earlier in the document or that the text be excluded here.</p>



	<p>confusing change from “mean change from baseline at 12 weeks,” is placed too early in the document.</p>	
123	<p>Please specify where, i.e., in which document(s), sponsors should describe the COA-based endpoint per the bullet points outlined in lines 125 and on.</p> <p>In addition, please clarify whether this refers to the study protocol(s) or also the COA dossier. As noted under “General Comments” above, there is no mention of the COA dossier. It would be helpful if the Agency clarify whether sponsors should continue to submit one for any primary and key secondary COA-based endpoint.</p>	<p>“<u>In the study protocol’s and COA dossier’s relevant sections</u>, sponsors should clearly describe the COA-based endpoint, including: ...”</p>
125 - 127	<p>“Type of assessment(s) made (e.g., Patient-Reported Outcome (PRO) measures, Observer-Reported Outcome (ObsRO) measures, Clinician-Reported Outcome (ClinRO) measures, Performance Outcome (PerfO) measures).”</p>	<p>We note that some passive monitoring DHT-derived measures (e.g., change in real-world walking speed, sleep duration) may fit the definition of a COA, by providing information into how patients function. However, these measures would not fit into any existing COA category as outlined by the agency. We therefore reiterate the need for a 169-173clearly defined fifth COA category and suggest the term “Passive Monitoring COA”.</p>
138-139	<p>Justification of rules for handling missing item responses is recommended. Often, this choice is purely based on aligning with the user manual or original development publication of an existing COA so that this is performed consistently across studies. Please confirm that aligning with these</p>	<p>Additional text:          “Handling missing item responses or task results in accordance with the original developer’s instructions is deemed sufficient justification.”</p>



	original rules is deemed sufficient justification.	
155	The terminology of “concept(s) of interest” might have different interpretation for different sponsors	Request that a standardized terminology or example be included to provide clarity
157-159	Clinical trial objective or hypothesis corresponding to the endpoint, ensuring that the objective/hypothesis is specific (e.g., “To compare the patient-reported physical functioning between arms at 24 weeks” rather than “To compare the patient-reported outcomes of product X vs. Y”).	Suggest revising the example to “To compare the CHANGE (from baseline) of the patient-reported physical functioning between arms at 24 weeks...” Should the example specify the direction of change in PRO endpoint in the objective/hypothesis?
165	The use of “indication” might be construed to mean “the disease to which the selected endpoints is intended for.”	Recommend clarification regarding use of “indication” in this context.
166	Explanation for why the selected COA is fit-for-purpose in the planned trial.	Suggest revising it to “Evidence for the selected COA...” or “Explanation for why the selected COA is fit-for-purpose in the planned trial, including the following: [list of items / reasons / pieces of evidence or measurement properties to be addressed].”
169-173	“In some cases, for endpoints based on a COA that measures a concept of interest that is indirectly related to some meaningful aspect of health for the patient (e.g., based on a neurological functioning test that is thought to be indicative of the patients’ cognitive functioning), it might be sufficient to provide support for the adequacy of the endpoint for measuring this aspect of health.”	We encourage FDA to elaborate on what type(s) of evidence could be considered to provide support for an endpoint based on COAs that measure concepts indirectly related to meaningful aspects of health. For example, we suggest that FDA discuss how correlation analyses between “direct” measures (e.g., PROs or activities of daily living) and the “indirect” measure could be used.  Change to:



		<p>In some cases, for endpoints based on a COA that measures a concept of interest that is indirectly related to some meaningful aspect of health for the patient (e.g., based on a neurological functioning test that is thought to be indicative of the patients’ cognitive functioning), it is sufficient to provide support for the adequacy of the endpoint for measuring this aspect of health.</p> <p>Please also consider adding examples of the situation(s) in which it might be sufficient and what kind of support is required for the adequacy of the endpoint.</p> <p>It is not clear why we would only need to demonstrate that a measure captures neurological functioning, without making a connection to patient cognitive functioning in this example. Isn’t it always necessary to demonstrate links between indirect measures and patient experience?</p>
175-176	If a multi-component endpoint, justification for the components included and the algorithm for combining them into the endpoint.	<p>Suggest revising to:</p> <p>“If a multi-component endpoint, justification for the components included and the algorithm for combining them into the endpoint is needed.”</p>
178	Please specify what type of data, information or evidence sponsors should list to provide support for the strength of the proposed endpoint. Please provide examples. In addition, please define “limitations” in “limitations of the proposed endpoint” and clarify what would be acceptable limitations of a COA-related endpoint that would not risk the endpoint’s acceptability by the Agency as fit for purpose. Please also provide examples.	<p>Suggest adding: “<u>Examples of data, information or evidence that may provide support for the strength of the proposed endpoint include...</u>”</p> <p>Also suggest adding language to clarify the meaning of “limitations of the proposed endpoint,” including examples.</p>



195	<p>Some COAs may require minimization of learning effect before the study starts running. This recommendation is missing here. Thus, the first COA administration might be recommended before the baseline. If it is mentioned on line 206 then, please clarify.</p>	<p>Suggest adding recommendations regarding minimization of learning effect or provide further clarification on determination of patient’s baseline value.</p>
198 - 202	<p>Some diseases, conditions, or clinical trial designs may necessitate more than one baseline assessment or longer/shorter baseline periods.</p> <p>When multiple baseline measurements are taken, the protocol should define how the baseline value will be calculated from the multiple measurements.</p>	<p>Inconsistent use of “assessment” and “measurements”. Suggest using “assessment”</p>
204 - 209	<p>The guidance recommends using a separate COA visit for severity screening versus baseline value. While this generally sounds sensible, we note that for example in the disease-specific draft guidance “Estrogen and Estrogen/Progestin Drug Products to Treat Vasomotor Symptoms and Vulvar and Vaginal Atrophy Symptoms — Recommendations for Clinical Evaluation”, it recommends inclusion criteria based on moderate to severe hot flashes <b>during the baseline period</b> (lines 101-3 in <a href="https://www.fda.gov/media/71359/download">https://www.fda.gov/media/71359/download</a>)</p>	<p>Clarify here, or at lines 182-3, whether recommendations within disease-specific guidance documents should take priority over the recommendations in this current guidance, when contradictions exist. Including if the disease-specific guidance is draft. Also, request that the document specify what actions the sponsor should take if the value from the screening assessment is very different from the value from later pre-randomization assessment or if the later assessment falls outside of the inclusion criteria.</p>



	Please confirm if disease-specific guidance takes priority in scenarios such as this.	
206	When screening and baseline scores differ significantly, there is risk for high variability of the COA.	<p>Please consider explaining FDA’s expectations for the reproducibility of baseline values, e.g., screening and baseline values should be within 15% of each other.</p> <p>Further, we suggest elaborating on and clarifying the reason for this recommendation. In addition, please comment on whether COA scores obtained post-randomization at first study visit (prior to any study procedures or treatments) are acceptable as baseline values.</p>
206-209	The guidance recommends not to use the patient’s baseline value, “Rather, a separate, later pre-randomization assessment should be used as the patient’s value”. Given the potential for patient burden on consecutive assessment closer to randomization, it may be possible that screening value is close enough to randomization and there is no need to reassess the outcome.	Consider relaxing requirement on the need to recommend two assessment pre-randomizations if timing of collection is negligible or a burden to patient.
208 - 209	“Rather, a separate, later pre-randomization assessment should be used as the patient’s baseline value.”	<p>We acknowledge that pre-randomization assessment is ideal; however, in some cases the time from randomization to first dose is long. This can limit the value of pre-randomization assessment while a pre-first dose assessment may provide a more appropriate baseline value. Therefore, we request that FDA clarify circumstances in which a pre-first dose assessment may be acceptable to determine baseline values. We suggest this may be appropriate in double-blind trials.</p>
211 - 213	“If the trial includes a run-in period during which the patient’s score from the COA might be expected to change (e.g., medication washout, patient behavior modification), then this should be	<p>Provide clarification and examples on how to address the run-in period situations.</p> <p>Please clarify and provide an example on how to address placebo run-in period</p>





	<p>considered when planning the timing of assessments.”</p> <p>The draft guidance includes more detailed text on the timing of assessments (line 544) and guidance on related to ‘run-in’ when discussing ‘practice effects’ (line 1314); however, it would be helpful to provide clarification and examples on how to address the run-in period situations.</p>	
<p>215-240; 268-319</p>	<p>“Endpoints based on COA scores at a fixed time point or a summary of COA scores over time...”</p> <p>And</p> <p>“Endpoints constructed by computing change from baseline or percent change from baseline COA scores...”</p>	<p>We note that using change from baseline scores while controlling for baseline score in an appropriate model is a common approach for analyzing COAs.</p> <p>Therefore, we recommend that Section 2b be expanded to include discussion of models (e.g., linear and proportional odds models) for controlling for the baseline score when using either follow-up or change from baseline as the dependent variable.</p> <p>We also recommend that the current discussion in Section II.A.2.d be incorporated into Section II.A.2b as an example of change from baseline without controlling for baseline, acknowledging that this is not the recommended approach.</p>
<p>218, 626-627 and general</p>	<p>The guidance refers to ordinal scores throughout. In practice, the majority of PRO scores are technically on an ordinal scale, for example:</p> <ul style="list-style-type: none"> <li>• A single item 0-10 numerical rating scale for pain</li> </ul>	<p>Provide guidance on when it may become appropriate to treat ordinal scales as continuous for analysis. Based on past experience and publications, I recommend treating ordinal scales as continuous when there are greater than 7 categories. But even a cut-off of &gt;5 may be acceptable in certain cases. Suggested text is:</p> <p><i>“In practice, many COA scores are technically on an ordinal scale but with many categories (e.g., a 0-10 numerical rating scale, or an average of multiple ordinal items). Generally, when ordinal scales</i></p>



	<ul style="list-style-type: none"> <li>• The Oxford Knee Score: average of 12 ordinal 5-point items, yielding a score from 0-48</li> </ul> <p>For both above examples, we cannot guarantee equal spacing between each possible score, so they are not interval or ratio scales but ordinal. However, it would be unusual to model the above example scores as ordinal (e.g., proportional odds model) and in fact such a model is unlikely to converge with many categories.</p> <p>Please provide guidance as to when a score can be considered essentially continuous for analysis.</p> <p>We also recommend stating that scoring based on latent variable models (e.g., item response theory, factor analysis) can produce truly continuous scores at the interval level).</p>	<p><i>have greater than 7 categories, it is appropriate to treat them as continuous for analysis (Bollen 1981; Rhemtulla 2012)."</i></p> <p>Bollen, K. A., &amp; Barb, K. H. (1981). Pearson's R and Coarsely Categorized Measures. <i>American Sociological Review</i>, 46(2), 232–239.</p> <p>Rhemtulla M, Brosseau-Liard PÉ, Savalei V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. <i>Psychol Methods</i>. 2012 Sep;17(3):354-73.</p>
218	<p>The statement "In most situations in which a COA produces ordinal or continuous (interval or ratio scale) scores" currently reads as if 'interval' and 'ratio' are alternate terms for 'ordinal' and 'continuous' when this is not the case and presumably not the intention.</p> <p>Ordinal variables should not be used as covariates in ANCOVA, MANCOVA, or regression analysis because such variables</p>	<p>Recommend rewording to: "In most situations in which a COA produces ordinal scores, or continuous (interval or ratio scale) scores"</p>



	are not linear. All covariates must be linear and possess an interval level of measurement.	
218-222	<p>This text deserves a follow-up explanation of how this is the same as change from baseline with covariate adjustment and the FDA’s reasoning for expressing this EP concept in the words used.</p> <p>It is critical that all important stakeholders, especially patients, understand these words versus “adjusted change from baseline.</p>	Recommend clarifying wording and testing the new wording to ensure it is understandable to a layperson. Also recommend FDA’s reasoning is included and expressed in a way that is accessible to a layperson.
221	The scope of the document appears to be for randomized clinical trials, except the introduction of consideration for single arm trials in 1362	Recommend the document specify upfront the scope and direct potential consideration for non-randomized single arm trials
228-229	<p>Original text:</p> <p>“Justification of the fixed time point should also take the recall period of the COA (where applicable) into consideration.”</p> <p>The draft guidance is currently unclear about situations where the recall period would influence the fixed time point.</p>	BIO recommends that FDA provide specific examples to illustrate how the recall period should inform the time point.
235-236	Text on repeated measures does not specify what types of summaries are allowable.	Recommend revising text to provide more specific directives on when using repeated measures is an appropriate methodology.
255-261	For example, FDA recommends that the rationale include evidence that patients and/or their caregivers view health states above the threshold to be meaningfully	1. Shall the guidance use the term MWPC when discussing the “threshold to be meaningfully different”?



	<p>different from health states below the threshold. This recommendation also applies to the use of ordinal or continuous COA data to define an event for a time-to-event endpoint. Of note, data used to derive a score threshold(s) should be different than that used to demonstrate effectiveness (e.g., data from registration trial(s)). In addition to prespecifying a single score threshold, sponsors should also conduct analyses to explore treatment effects over a range of thresholds.</p>	<ol style="list-style-type: none"> <li>2. “health states” can be confusing with health states generated from preference-based measure. We request FDA include a more specific definition for this term.</li> <li>3. Unclear what “data used to derive a score threshold(s) should be different than that used to demonstrate effectiveness” means. The example of “data from registration trial(s)” does not help explain this sentence. We request clarification as to whether the FDA wants different methodologies (distribution-based versus anchor-based) or different datasets to derive the threshold and then demonstrate effectiveness.</li> </ol> <p>Suggest changing the last sentence to “In addition to prespecifying a single score threshold, sponsors should also conduct sensitivity analyses using different thresholds to examine robustness of the primary analysis.”</p> <p>Further suggest adding: “<u>Examples of the types of evidence that may be accepted to support patient and/or caregiver views of meaningful differences from the prespecified threshold include ...</u>”</p>
258-260	<p>Original text:</p> <p>“Of note, data used to derive a score threshold(s) should be different than that used to demonstrate effectiveness (e.g., data from registration trial(s)).”</p> <p>This may not always be feasible in some disease areas due to population size and development timelines. We believe that there should be mention of situations where thresholds can only be derived from a partial set of the registrational trial.</p>	<p>BIO recommends the following revision:</p> <p>“Of note, data used to derive a score threshold(s) should be different than that used to demonstrate effectiveness (e.g., data from registration trial(s)). <u>However, in the case of small patient populations (e.g., rare diseases) and/or expedited development programs, using a partial set (e.g., 1/3 of the total sample size could be used for estimation of MSDs) of the data to derive the threshold could be an acceptable option.</u>”</p>



258-259	Please address the use of adaptive trial designs in rare diseases and whether FDA would be able and willing to exercise flexibility in accepting data from registrational trials to support meaningful change thresholds.	Suggest addressing use of adaptive trial designs and the Agency’s exercise of regulatory flexibility in rare disease drug development.
263-266	The first sentence has been important for dermatology endpoints, and the value of this approach has been well-recognized in the field of dermatology. The Meaningfulness of Treatment Benefit Section, however, seems a confusing way to approach dermatology endpoints, and the current practice that has led to approved medications that are key breakthroughs in that field (e.g., alopecia areata, atopic dermatitis).	Recommend softening the language in these sentences to recognize the value of dichotomization of continuous COAs to interpret for all stakeholders.
271-273	<p>This text lacks an important distinction that in practice a model using COA score change-from-baseline would also include the baseline score as a covariate. Currently it reads as if there are only two options:</p> <ul style="list-style-type: none"> <li>• Model follow-up score with baseline score as covariate</li> <li>• Model change from baseline score, without baseline score as covariate</li> </ul> <p>This is open to misinterpretation and seems to contrast what is provided in FDA draft guidance “Adjusting for Covariates in</p>	<p>Suggest rewrite text more in line with draft guidance “Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products” – and cite this resource.</p> <p>Add text “As discussed in Section II.A.2.a, in comparative trials, the preferred method for adjusting for baseline status is to do so in the context of a statistical model. For continuous scores, the choice of modelling follow-up score or change-from-baseline score is purely an issue of interpretability, as long as the baseline score is a covariate in the model. Modelling the change-from-baseline COA score as an outcome, without also adjusting for baseline score as a covariate, is less preferable (FDA 2021, EMA 2015)”</p> <p><a href="https://www.fda.gov/media/148910/download">https://www.fda.gov/media/148910/download</a></p>



	Randomized Clinical Trials for Drugs and Biological Products”.	<a href="https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf">https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf</a> <a href="https://www.fda.gov/media/148910/download">https://www.fda.gov/media/148910/download</a> <a href="https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf">https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf</a> <a href="https://www.fda.gov/media/148910/download">https://www.fda.gov/media/148910/download</a> <a href="https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf">https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf</a>
275 - 277	“COA scores that are ordinal are challenging to interpret in terms of change from baseline because the difference between two ordinal scores cannot be assumed to have the same meaning across scores”	<p>We note that this only applies if a linear model for analysis is chosen. However, other models such as proportional odds models can be used to properly account for the ordinal nature of the response. We encourage FDA to consider how other models can be integrated to support interpretation of change from baseline scores. Further BIO would appreciate additional context on whether it is reasonable to describe changes on ordinal scales in the context of proportion of patients moving across categories along the scale (i.e., shift tables)</p> <p>Lastly, we suggest this text concerning statistical modelling of ordinal endpoints belongs better within the section “Analyzing ordinal data” at line 622. We recommend moving the text to within the section “Analyzing ordinal data” at line 622.</p>
283-287	This approach is over-complicated and likely to be met with suspicion/confusion when presenting to clinicians and patients, due to the use of predicted scores rather than observed data.	We recommend deleting this text, because using change from baseline as an outcome, but also making sure to control for baseline score as a covariate, is an acceptable approach for continuous scores as detailed in FDA draft guidances “Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products” and EMA guidance “Guideline on adjustment for baseline covariates in clinical trials.”
289	The meaning of the parenthetical, “(e.g., to evaluate some devices)”, is not clear and	Please clarify the parenthetical, “(e.g., to evaluate some devices)”, or suggest deleting if additional context or information is not provided.



	could be misleading if additional context or information is not provided.	
290-292	<p>“For situations in which it is not possible to conduct a randomized, controlled trial and a single arm trial is done instead (e.g., to evaluate some devices), a change-from-baseline endpoint might be the best available option.”</p> <p>Sponsors and other readers will question why change from baseline works in a single arm study and is not the recommended approach for randomized studies.</p> <p>Please expand the text to include studies of rare diseases, oncology, and other instances where it is not ethical or feasible to conduct a randomized controlled trial.</p>	<p>Please consider expanding the text as follows:</p> <p>“For situations in which it is not possible to conduct a randomized, controlled trial and a single arm trial is done instead (e.g., to evaluate some devices, <b><u>studies of rare diseases, oncology, and other instances where it is not ethical or feasible to conduct a randomized controlled trial</u></b>), a change-from-baseline endpoint might be the best available option.”</p> <p>Recommend clarifying issue of continuous COA scores analyses by considering whether the “scores at 12 weeks”-like language is needed versus “mean changes compared across groups and mean group differences at timepoint <i>t</i> controlling for baseline scores.”</p>
293	<p>An additional advantage of using percent change-from-baseline is to define a responder definition, where the threshold for meaningful within-patient change varies according to baseline severity.</p> <p>For example, on a 0-10 pain scale (10 is worst), if severe patients (scoring 7-10) required a larger change for it to be considered meaningful than moderate patients (scoring 4-6), then the use of percentage change as a responder threshold can reflect this.</p>	<p>Recommend adding text:</p> <p>“Another advantage of using percent change-from-baseline is to define a responder definition where the threshold for meaningful within-patient change varies according to baseline severity.”</p>
293-319	There are additional concerns with “percent change”.	Consider taking all or part of this rationale to buttress the points made in the draft guidance.

Unlike simple change, percent change based on transformed scores and original scores can give different p-values in the comparison of cumulative distribution function (CDF) plots between treatments. When the minimum possible score exceeds zero (typically, a minimum possible score of at least one), a CDF analysis using percent change from baseline based on transformed 0-100 scale scores is different from an analysis using percent change from baseline based on original scores, leading to different results and possibly different conclusions. Transformed scores can give more extreme values on percent change than untransformed scores when baseline values are close to zero.

Which score to use: transformed score or original score? Which metric to use: percent change or simple change? The choice of which type of score (original or transformed) and type of metric (percent or simple) to use should be based on how a patient-reported outcome was developed, analyzed, interpreted, and reported before CDF plots are later considered for the purpose to enhance interpretation on a patient-reported outcome. After all, the purpose of the CDF plot in this context is simply to enrich the interpretation of scores as they were intended to be scored and were

Reference (and attached):

Bushmakin AG, Cappelleri JC. A note on cumulative distribution functions for patient-reported outcomes. Patient Reported Outcomes (PRO) Newsletter. 2011; 45 (spring issue):11-12.



Bushmakin\_PRO  
Newsletter\_2011\_pag





	applied in the primary analysis that preceded consideration of a CDF plot.	
308-309	Percent change-from-baseline is undefined if the baseline score on a COA is zero, and some kind of imputation is required to include the observation in the analysis. The document does not acknowledge that zero could be the actual score.	Unclear what “some kind of imputation” can be used if COA score at baseline is zero. Recommend changing the terminology from “imputation” to “transformation.”  Recommend document include discussion of case in which the baseline score is actually zero.
315-319	If the reason for considering percent change-from-baseline is that the treatment effect is expected to be multiplicative rather than additive (e.g., treatment improves a patient’s symptom severity by 20% of the patient’s severity level without treatment), then a logarithmic or similar transformation could be applied to continuously distributed COA scores prior to comparing groups (Senn 2007)	Using log or transformation of scores can be problematic and create biases in COA score interpretation especially if the treatment effect is expected to be multiplicative. Unsure if this suggestion should be included in the guidance without clear evidence about the suggested rationale for COA.
333-334 562-564 1456	<p>“When possible, sponsors can evaluate multiple endpoints in earlier phase trials to inform the selection of endpoints for later trials.”</p> <p>“...when the COA is complex and potentially burdensome, sponsors might consider seeking input from members of the patient community to ensure that the planned length of the trial and timing of COA assessments is feasible and as convenient as possible for the patients and/or caregivers.”</p>	Please include best practices for how patient groups-FDA-Sponsors can expeditiously align when patient group and FDA expectations don’t align regarding patient burden.



	<p>Sponsors must balance various considerations when developing a COA strategy. Input from patient communities and FDA are important when developing a COA strategy, but sponsors need to reconcile scientific, regulatory, and practical considerations (e.g., communities may want to see some COAs measured that are not feasible to include or may want to exclude COAs which are necessary per FDA).</p> <p>Therefore, any additional guidance that the Agency can provide regarding how to balance patient burden, willingness to participate, and FDA expectations, particularly with regard to 1) valuating multiple endpoints in early phases and 2) assessment frequency, especially when the patient and FDA perspectives do not align, would be very helpful.</p>	
346-351	<p>The language proposed here leads us to interpret that the agency is suggesting that only one personalized endpoint can be considered as a primary endpoint.</p>	<p>We request that the agency clarify if the agency is suggesting if all other endpoints will be considered secondary.</p>
356-359	<p>Co-primary endpoints. This option may be appropriate when there are multiple aspects of health that are critically important to the disease being studied, such that a treatment benefit can only be concluded if the medical</p>	<p>Would this also include a co-primary endpoint that involves PRO and ClinRO assessments of the same aspect of health? (e.g., appearance of acne, cellulite, psoriasis). Where does this example fall? This does not seem to be a multi-component endpoint because it is the same aspect of health.</p>



	product has an effect on each of the designated endpoints.	
377-381	Additional guidance is needed about which aspects of the scoring algorithm are expected to be specified and the related timing. If a sponsor is developing a new scale, then the dimensionality and factor structure will not be known when it is first used in studies.	Please provide additional guidance about which aspects of the scoring algorithm are expected to be specified and the desired timing. Additionally, some of this information may be available in previous guidance documents. Links to previous guidance where this information is already detailed (such as draft guidance 3) will be helpful.
391-392	Please provide an example approach to estimate weightings.	Please provide an example approach to estimate weightings.
399-405	It is unclear whether FDA would allow a COA to be used for a randomized withdrawal study design.  Other multi-component endpoints are constructed with the objective of demonstrating the absence of all symptoms.	Please consider adding an example such as, “Time to return of symptom(s) from withdrawal timepoint.”  Further, recognizing this is not intended to be an exhaustive list, another example that may be helpful could be “Improvement on one or more subscales with no worsening on other subscales (from an instrument with multiple concepts)”.
419-420	The use of within-patient multi-component endpoints can be efficient if the treatment effects on the different components are generally concordant.  Additionally, the meaning of the term concordant is not clear in this context.	May be helpful to provide examples here of multi-component endpoints and in particular, within-patient multi-component endpoint.  Suggest adding a parenthetical after “concordant” to clarify the meaning of that term (perhaps with an “i.e.”).
450	When deconstructing multi-component endpoints, risk is introduced for competing events, e.g., a patient may be counted only once for a composite endpoint with their first event and cannot contribute to other endpoint components, causing proportions	Please provide examples for when the examination of individual components of a multi-component endpoint would or would not be appropriate to complement the conceptual discussion in the October 2022 final guidance, “Multiple Endpoints in Clinical Trials”



	of individual endpoints not to reflect true outcomes	
463-465	There is the potential for bias when those completing or administering the COA are aware of the thresholds for being considered a meaningful improvement (or worsening).	Sentence is oddly written, suggest the following revision:  There is the potential for bias when those completing or administering the COA are aware of the defined meaningful improvement (or worsening) threshold. Propose that the guidance contain recommendations for addressing such cases.
463-468	In some situations, it is not possible to blind/mask the treatment (e.g., in cases of very large treatment effects), and COA administrators are aware about the threshold effect.	Add a recommendation to consider what approach should be taken when no blinding/masking is possible
515	An additional concern of personalized endpoints is what to do when treatment arms are significantly unbalanced in terms of which symptoms/goals are identified by patients.  One solution is to stratify randomization by selected symptoms.	Add suggested text:  “Treatment arms may be imbalanced in terms of which symptoms/goals are selected by patients. One solution to this issue is to employ stratified randomization based on this factor.”
520-522	It is not clear why it is a concern that changes might occur during the trial in what patients regard as their “most bothersome” or “most severe” symptom. The fact that a symptom that was considered by a patient at baseline as the “most bothersome” or “most severe” is no longer so suggests improvement, possibly due to the treatment.	Please clarify why this is a concern.
538-542	Please note that some subjects will choose endpoints that are easier to treat than others.	Suggest adding to the second noted sentence as follows: “...it is important to measure all relevant symptoms and areas of functioning in






		addition to those identified as most important to the individual patients, <b>which will vary with respect to how easy they are to treat.</b>
559	Regarding frequency of assessments, should another consideration regarding frequency be to understand within-patient variability? The challenge with many COAs is the higher within-patient variability (signal vs noise). A minor limitation is the covariance assumption particularly when assessments are not conducted in equal intervals/spaces.	Recommend document provide information on whether there should be consideration of frequency to understand within-patient variability.
560 - 565	“In many instances, such as when a COA is planned to be frequently measured (e.g., event-triggered data collection) or when the COA is complex and potentially burdensome, sponsors might consider seeking input from members of the patient community to ensure that the planned length of the trial and timing of COA assessments is feasible and as convenient as possible for the patients and/or caregivers.”	<p>We urge the FDA to use stronger language to ensure that the patient community is consulted, and that representative patient input is obtained. We also note that input from the patient community is especially important for event-triggered data collection and potentially burdensome COAs.</p> <p>We recommend that the text be revised as followed:  <i>“In many instances, such as when a COA is planned to be frequently measured (e.g., event-triggered data collection) or when the COA is complex and potentially burdensome, sponsors <b>should seek representative</b> input from members of the patient community...”</i></p> <p>We also recommend that FDA expand this section to encourage patient input into how the COA is collected (e.g., paper, electronic, at home, in clinic) to support adherence.</p>
570	It is difficult to decide an appropriate timing of assessment when treatment arms differ in their cycle lengths (For example, chemotherapy based on a 21-day versus 28-day cycle). If administering COAs every 14 days, patients will be reporting on their experiences at different times relative to	Please provide recommendations for scenarios where treatments have different cycle lengths.



	dosing, where side effects are likely to be worse shortly after receiving medication. Please provide recommendations for scenarios where treatments have different cycle lengths.	
572	Event-triggered data collection	Event-triggered assessments may be a good use case for passive monitoring using sensor-based DHTs. If the agency is open to such uses, it may be helpful to add an example such as “Another example could be the use of passive digital sensors to detect triggering events, such as acoustic recognition of a coughing fit used to prompt a patient report of respiratory symptoms.”
597	“It will typically be of interest to understand treatment effects regardless of adherence to treatment, such that the protocol should include plans to continue to follow patients and administer the COA after discontinuation of treatment.”	<p>We recommend that FDA clarify whether “discontinuation of treatment” includes discontinuation of treatment due to safety reasons, participant decision to leave the clinical trial, and/or discontinuation of treatment at the end of the clinical trial.</p> <p>Assuming that each of the above scenarios are in scope, we also recommend that this section includes a discussion on the plan for missing data in cases where the clinical trial participant chooses to leave the clinical trial.</p>
597-599	<p>This recommendation needs to also consider cost and other logistical, operational considerations that may or may not outweigh measurements taken after discontinuation of treatment.</p> <p>Consideration should also be given to the patient burden and feasibility of collecting the data once a study participant decides to discontinue the treatment.</p>	<p>Please consider expanding to highlight the voluntary nature of collecting and providing the following data, and also provide insight into how FDA could use this data to inform regulatory decision making.</p> <p><b>“While voluntary,</b> <i>It will typically be of interest to understand treatment effects regardless of adherence to treatment, such that the protocol should include plans to continue to follow patients and administer the COA after discontinuation of treatment.”</i></p> <p>Additional comment from PCOA: Consideration should also be given to the patient burden and feasibility of collecting the data once a study participant decides to discontinue</p>



	It isn't clear how FDA would use this data, and what value it would provide for regulatory decision making.	the treatment. What is the value of collecting this data from the Agency's perspective? How would this data be used by the Agency?
597-599	We appreciate that the guidance provides recommendations to ensure patient follow up and COA administration in cases of adherence.  However, the document does not discuss application of Bayesian methods.	Consider reference ICH E9 R1 and handling of intercurrent events.  Consider including information on general current thinking if sponsors consider using Bayesian approaches in the design and analysis of COAs.
601	Section B. Estimation and Missing Data	Request consideration of inclusion of a possible ceiling or floor effects on estimation, in the document.
618	Aren't all time points fixed (prespecified)? Do we mean designated primary time point?	Clarify what is meant by "the fixed time point."
629-643	Conventional parametric statistical methods are generally robust enough to approximate correct inference on quantitative data. Although there may be concern that the assignment of integers to the categories is somewhat arbitrary and that the distances between adjacent scores do not represent equal gradations, moderate differences among various scoring systems seldom produce marked changes in conclusions (Baker et al., 1966; Snedecor and Cochran, 1980).  For example, it has been recommended to assign integer scores to ordinal categories and conduct parametric methods when the	Consider revising these lines in accordance with the following insightful literature warning of misconceptions and mis-undertakings on this topic:  Baker BO, Hardyck CD, Petrinovich LF. 1966. Weak measurements vs. strong statistics: An empirical critique of S. S. Steven's procriptions on statistics. Educational and Psychological Measurement 26:291–309.  Cappelleri JC, Zou KH, Bushmakin AG, Alvir JMJ, Alemayehu D, Symonds T. 2013. Patient-Reported Outcomes: Measurement, Implementation and Interpretation. Boca Raton, Florida: Chapman & Hall/CRC Press.  Carifio, L, Perla R. 2008. Resolving the 50-year debate around using and misusing Likert scales. Medical Education, 42:1150–1152.

	<p>data have an underlying continuous scale (Snedecor and Cochran, 1980).</p> <p>The central limit theorem for means, one of the most celebrated results in statistics, makes the assumption of normality appropriate on the sampling distribution of the mean even if the individual data are not normally distributed, provided that the sample size is large enough. Therefore, parametric statistical tests for means are generally appropriate for quantitative data because of the central limit theorem (whether or not the individual data are normally distributed). Further support for parametric methods is based on the assignment of consecutive integers being viewed as just a monotonic transformation that is analogous to other types of transformations such as log and square root transformations, which are commonly employed to help correct departures from the usual assumptions. Thus, from a pragmatic perspective, under most circumstances (unless the distribution of scores is severely skewed), data from ordinal rating scales can be analyzed as if they were based on interval-level measurements without introducing severe bias.</p>	<p>Gaito J. 1980. Measurement scales and statistics: Resurgence of an old misconception. <i>Psychological Bulletin</i>, 87, 564–567.</p> <p>Norman G. 2010. Likert scales, levels of measurement and the “laws” of statistics. <i>Advances in Health Science Education</i>, 15, 625-632.</p> <p>Snedecor GW, Cochran WG. 1980. <i>Statistical Methods</i>. 7th edition. Ames, IA: The Iowa State University Press.</p> <p>        Gaito_Psychological        Bulletin_1980.pdf</p> <p>        Norman_Advances in        Health Sciences Educa</p> <p>        Carifio_Medical        Education_2008.pdf</p>
--	--	---





	Reference: Cappelleri JC, Zou KH, Bushmakin AG, Alvir MJ, Alemayehu D, Symonds T. 2013. Patient-Reported Outcomes: Measurement, Implementation and Interpretation. Boca Raton, Florida: Chapman & Hall/CRC Press.	
641 - 642	“The key point when choosing an analytic approach is that the results are interpretable and address the appropriate clinical question.”	We agree with this statement but must emphasize that it is important to first formulate the clinical question within the estimand framework. We request that FDA revise as follows:  <i>“The key point when choosing an analytic approach is that the results are interpretable and address the appropriate clinical question <b>as defined within the estimand framework.</b>”</i>
661 - 663	“Missing data are problematic because they may lead to reduced power and potential bias in the estimated treatment effect when missingness is related to treatment effectiveness or to adverse events from the treatment.”	We note that per the ICH E9(R1) addendum, what is considered missing data and to what extent it is problematic will depend on the estimand, the intercurrent events and the strategy chosen for them. Therefore, we suggest that FDA revise this text as follows:  <i>“<b>Data that were intended to be collected per protocol but are missing</b> are problematic because they may lead to reduced power and potential bias in the estimated treatment effect when missingness is related to treatment effectiveness or to adverse events from the treatment.”</i>
Footnote 21	“Potential intercurrent events and methods to handle intercurrent events should be addressed in the statistical analysis plan.”	We note that intercurrent events are part of the estimand which defines the clinical question of interest. Specification of the clinical question should be done either before trial design or early in the design stage and described in the protocol prior to describing the analysis plan in the SAP. Therefore, we urge FDA to revise this statement accordingly to align with ICH E9(R1).
671-674	“When a person does not complete a COA at a given time point, the site should be notified so that research staff can contact	We note that PROs and other COAs are point in time assessments and should not be collected retrospectively if the assessment period has passed. Retrospective data collection is error prone (e.g., due to recall



	<p>the appropriate person (patient, caregiver, study, or site staff) to obtain the needed assessment.”</p>	<p>bias). Therefore, this recommendation is not appropriate and should be removed.</p> <p>We suggest that instead of the current recommendation, FDA revise this text to encourage the use of reminders (e.g., app-based reminders) when feasible to improve completion of protocol-specified assessments.</p>
667-670	<p>This begins with collecting only those COAs necessary to assess the endpoint (e.g., for efficacy, safety, tolerability) and designing a data collection plan that is least burdensome and as easy as possible for patients and/or caregivers.</p>	<p>Request to provide recommendations regarding the acceptable level of missing data and acceptable and not acceptable reasons for missing data. Request to provide examples of reasons for missing data to be included in the trail data collection.</p> <p>Suggest also including mention of clinic/study site staff because they need to potentially manage PRO completion (e.g., administering eCOA devices), and clinicians since they may be completing ClinROs</p>
689-1526	<p>Patient/Observer/Clinician global impression scales are discussed as possible anchors. In Coon &amp; Cook (2017), which is cited as a general reference for MSD estimation methodology, it is stated that “In some therapeutic areas (e.g., schizophrenia), the clinician global impression of change (CGIC) scale may be substituted for the PGIC to obtain a clinical judgment of the patient’s condition. However, unless there is impairment associated with the condition that would likely render the patient’s feedback unreliable, a (suitable) patient-reported anchor is always preferable.”. It is not clear under what circumstances the agency would accept Clinician/Observer</p>	<p>Include guidance on CoUs in which patient-reported anchors can be omitted (or not).</p> <p>Recommend review the full reasoning for the re-focus available in the FDA records of comments submitted for the 2006 Draft Guidance to ensure that the current state of COA research is reflected, and that comparison of meaningful change threshold(s) at the individual level is not conflated with treatment effects comparing group mean changes.</p>



	<p>scales as anchors, without patient-reported scales.</p> <p>Laurie Burke and Donald Patrick attempted to translate the MSD (a 2023 term applied backwards in time) to the interpretation of a treatment effects in the 2006 Draft PRO Guidance. This attempt was not successful and led to re-examination of this issue with updates for interpretation at the individual patient change level in the 2009 Final PRO Guidance.</p>	
699-712	<p>It appears that interpretability of COA scores is referring to within-patient change, within-group difference, and between-group difference. Is interpretation at the group level or individual level or both?</p>	<p>Please clarify and confirm what type(s) of interpretation is being referred to: within-patient change, within-group difference, between-group difference.</p>
Line 701-709	<p>The statements and examples try to use, in part, patient-level measurement to interpret the treatment effect at the group level. This does not seem appropriate because a “2-point difference of treatment effect” is at the group level (which mixes patients with improvement, worsening, and stable status), which is NOT something that individual patients would notice as important in their daily lives.</p>	<p>Suggest to clearly distinguish patient-level meaningful change from group-level difference. It is not appropriate to use patient-level meaningful change to interpret group-level difference.</p>
704-705	<p>Regarding the statement “For example, if a treatment is shown to reduce scores on a performance outcome measure by an average of 2 points on a 15-point scale, it</p>	<p>This statement conflates the interpretation of within-group change and within-individual change. BIO Recommends deleting this example.</p>




	<p>would be helpful to know whether a 2-point difference corresponds to something that patients would notice as important in their daily lives”.</p>	
<p>731-732</p>	<p>“For these types of measures, it may be more challenging to infer how different scores on the measure correspond to different experiences the patients might have; this means that additional empirical support is needed to translate scores on the measures to corresponding patient experiences in their daily lives.”</p>	<p>We agree that it may be challenging for patients to comment on the relevance of “indirect” measures (i.e., those that measure concepts of interest more indirectly related to meaningful aspects of health) in their daily lives. Nonetheless we believe such indirect measures can be critical for a more comprehensive assessment of the patient’s experience with disease and treatment. For example, many performance outcomes and DHT-derived clinical outcome assessments are indirect measures of meaningful clinical outcomes. Therefore, we commend the FDA for recognizing their role as part of this guidance series.</p> <p>We request that FDA elaborate on the empirical support that could be provided to demonstrate the relationship between an indirect measure and a clinical outcome of interest. We suggest that quantitative evidence to describe this relationship and/or expert interviews to assess how an indirect measure relates to a meaningful concept of interest could be considered. Further public discussions on this topic could be beneficial to develop consensus on appropriate methodologies.</p>
<p>746-748</p>	<p>“...the scores themselves are directly interpretable in terms of patients’ experiences, and therefore, additional supporting evidence may not be necessary for interpretation.”</p> <p>Please confirm that anchor measures would not be required in these situations and additional data analysis would not be</p>	<p>We request that the FDA specifies what methodology would be appropriate to generate additional evidence demonstrating how these transformed scores relate to patient experiences.</p> <p>In addition, there may be situations where both the raw score and the t-transformed scores might be available for a particular COA. It would be appreciated if FDA could provide guidance on what type of score should be used, whether both or whether a case could be made for the use of one type of score over the other. This question links to the</p>



750-755	<p>required. If anchor methods are not required, what level of change would be deemed meaningful (e.g., 1 or 2 levels?)</p> <p>“Other COAs produce scores that are more difficult to interpret on their own because they use a metric that is unfamiliar and/or abstract, such as a COA measure that produces transformed scores (e.g., linear transformation of a 0-4 raw score scale to a 0-100 score scale). There might be very good reasons to generate a score on such a metric, but it increases the complexity of describing the endpoint in labeling. In this case, FDA recommends additional evidence to justify how scores relate to meaningful patient experiences.”</p>	similar question posed under lines 293-319 above.
758	<p>Please interpret the MSD in relation to the most recently used term “Meaningful Within Person Change (MWPC)” and meaningful change thresholds. In order to link these new terms with previous terms, it would help to link these terms so readers can adjust previously used terms/methods to the current proposal to use MSD and MSR terminology instead.</p>	<p>Line 771: “... meaningful score differences (III.B.1), <u>previously referred to as clinically meaningful within-patient change, and in terms of meaningful score regions (III.B.2) [provide some text to refer to previous terms, such as between-group differences]</u>.</p> <p>Line 778: Suggest adding the acronym “(MWPC)”.</p>
758-1053	<p>Approaches for Collecting Evidence to Support Interpretability of COA-Based Endpoints</p>	<p>There is a <b>need</b> to show how the new concepts of Meaningful Score Differences (MSDs) and Meaningful Score Regions (MSRs) relate to the widely used concepts of meaningful-within patient change (MWPC) and meaningful clinically important differences (MCIDs).</p>



		Are the MSD and MSR now replacing those? Or are there additional options to help interpret meaningful change? I'm struggling to understand how they tie in.
762-763	“If the body of evidence supporting the interpretability of COA scores (e.g., from existing literature) is not sufficient...”, FDA recommends conducting empirical studies to support interpretability of COA scores prior to conducting a registration trial. ...”	What level of evidence would be considered sufficient by the Agency? Please provide examples.  In addition, FDA should recognize that in rare disease it might be extremely difficult to conduct separate empirical study to support interpretability of COA scores prior to conducting a registration trial. Moreover, even if such a study was possible, FDA should specify how similar a population in that study should be to the population included in a clinical trial? Whether that study should also be interventional (with the same intervention of interest), or a natural history study would suffice.
767-769	“Based on such empirical studies, sponsors should prespecify the range of estimates that will be used to interpret the treatment effect(s) in a registration trial.”  We think it is important to clarify that in practice there would be a single value pre-specified as a starting assumption, with ranges as sensitivity analysis. This is more in line with other parts of this guidance e.g., lines 260-261.	BIO recommends editing the text to state: “Based on such empirical studies, sponsors should prespecify estimates that will be used to interpret the treatment effect(s) in a registration trial, plus ranges around these estimates as sensitivity analysis.”
801-802	The value of MSD is the same for improvement and deterioration (Crosby et al. 2003). If this assumption is not true, then it is possible to use different values for MSD depending on the direction of change	If there is evidence of a linear relationship between the target COA and anchor predictor, then the amount of meaningful score difference is the same for both improvement and deterioration.  References:

		<p>Cappelleri JC, Bushmakin AG. Interpretation of patient-reported outcomes. <i>Statistical Methods in Medical Research</i>. 2014; 23:460-483. (attached)</p> <p>Cappelleri JC, Zou KH, Bushmakin AG, Alvir JMJ, Alemayehu D, Symonds T. <i>Patient-Reported Outcomes: Measurement, Implementation and Interpretation</i>. Boca Raton, Florida: Chapman &amp; Hall/CRC Press. 2013.</p> <p>Bushmakin AG, Cappelleri JC. <i>A Practical Approach to Quantitative Validation of Patient-Reported Outcomes: A Simulation-Based Guide Using SAS</i>. Hoboken, New Jersey: John Wiley &amp; Sons. 2022.</p> <div style="text-align: center;">         Cappelleri_SMMR_2014.pdf     </div>
786	<p><i>“Expected treatment effect for the average patient”</i></p>	<p>This suggests that MSD is an expected treatment effect for an average patient. These are not necessarily the same (though they are correlated). Suggest changing to <b>“(1) to contextualize the observed treatment effect for the average patient in some target population”</b></p>
785-788	<p>The following statement is questionable: “Regardless of the approach used to determine the MSD, the MSD can be used in at least two ways: (1) to evaluate the expected treatment effect for the average patient in some target population; or (2) to use as a threshold in descriptive analyses that identify individual patients who might have changed by a meaningful amount.”</p>	<p>Remove the suggestion that thresholds for group-level and individual-level interpretation can be used interchangeably and estimated in the same way.</p> <p>Focus the presentation of meaningful score difference on meaningful within-patient change, to align with current emphasis in the field.</p> <p>Add additional citations providing more detailed coverage of anchor-based methods (suitable papers provided in comment).</p>



It is widely recognized in the field of meaningful change on COAs that thresholds for interpreting average treatment effects (i.e., group-level interpretations) versus individual-level change should be distinct and are targeted by different methods. See for example:

- Trigg, A., Lenderking, W.R. & Boehnke, J.R. Introduction to the special section: “Methodologies and considerations for meaningful change”. *Qual Life Res* 32, 1223–1230 (2023).
- Coon, C.D., Cappelleri, J.C. Interpreting Change in Scores on Patient-Reported Outcome Instruments. *Ther Innov Regul Sci* 50, 22–29 (2016).
- Terwee, C.B., Peipert, J.D., Chapman, R. et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. *Qual Life Res* 30, 2729–2754 (2021).
- Sabah SA, Alvand A, Beard DJ, Price AJ. Minimal important changes and differences were estimated for Oxford hip and knee scores following primary and revision arthroplasty. *J Clin Epidemiol.* 2022 Mar;143:159-168.





	<p>In addition, the Coon &amp; Cook 2018 paper cited throughout this section focuses on thresholds for within-individual change: “Also recognized is the distinction between individual- and group-level guidelines for what constitutes meaningful differences. The distinction has real-world consequences ... The purpose of this paper is to present current methods for setting thresholds for use in interpreting change in individual-level COA scores.”</p> <p>Therefore, the statement on lines 785-8 is not in line with current thinking in the field of meaningful change.</p> <p>Given that methods to estimate thresholds for within-individual change are more advanced and provide more easily interpretable summaries of meaningful treatment benefit, I would recommend focusing on within-patient change thresholds.</p>	
786 - 787	MSD is defined as a threshold for within-patient meaningful change. It is unclear whether it can be used to interpret the treatment effect if the COA endpoint is (as recommended above) a group-level measure (e.g., adjusted mean change from baseline).	Please discuss the difference between group-level and individual-level thresholds and specify how the mean-change-from-baseline can be interpreted.



Section B	Introduction of terminology ‘meaningful score differences’ and ‘meaningful score regions’	Unclear how this terminology relates to previously used terminology such as ‘within-patient meaningful change’ and ‘between-groups meaningful change’. It seems MSD is still referring to within-patient, but MSR is less clear and the term ‘meaningful differences or scores’ throughout could be confused as referring to between-group differences. Is it possible to clarify?
Section B	“Meaningful score differences” and “Meaningful score regions”	For those uninitiated in measurement science, multiple terms used to describe meaningful or clinically important differences in score and thresholds gets very confusing. Would be helpful to include a glossary to explain what the differences are.
795-799	<p>We do not recommend modifying the MWPC of COA according to the baseline values of the COA in part because of regression to the mean.</p> <p>Baseline severity on the target COA of interest assumes that the requirement of greater change for the more severely impaired is entirely a clinical phenomenon. However, this requirement may be influenced in part by a statistical phenomenon: regression to the mean.</p> <p>Regression to the mean is an error-based artifact describing the statistical tendency of extreme scores to become less extreme at follow-up. Individuals with extreme scores at baseline would require a greater change to be considered clinically meaningful than those with less extreme scores.</p>	<p>BIO believes that an assumption that the value of MSD is the same regardless of baseline score diverges from the FDA Guidance 4 Discussion document and should not state otherwise in this guidance, for consistency.</p> <p>Please clarify the implications if a treatment produces a meaningful change among one severity subgroup but not the other.</p>



Thus, individuals with the greatest impairments in COA at baseline would have the greatest opportunity to change (improve) than individuals with less extreme scores, leading to concluding erroneously that those with severe impairments have shown clinically meaningful improvement when much of that change can be attributed to regression to the mean.

There are methodological concerns, and no consensus exists on how to adjust for regression to the mean in this particular context from a purely anchor-based approach.

Moreover, having different MWPCs for different values of baseline COA scores would introduce unnecessary confusion, obfuscation, and complexity in the implementation and interpretation of the MWPC.

The estimate MWPC for a given COA that we have provided is taken as the average value of MWPC across all patients regardless of baseline severity (for example, the average value across all patients whose CGI-S rating improved by one



	<p>category). Results and conclusions, therefore, apply to the typical patient on average.</p> <p>References:</p> <p>Campbell DT, Kenny DA. 1999. A Primer on Regression Artifacts. New York, NY: The Guilford Press.</p> <p>Crosby RD, Kolotkin RL, Williams GR. 2003. Defining clinically meaningful change in health-related quality of life. <i>Journal of Clinical Epidemiology</i> 56;395-407.</p> <p>Crosby RD, Kolotkin RL, Williams GR. 2004. An integrated method to determine meaningful changes in health-related quality of life. <i>Journal of Clinical Epidemiology</i> 57;1153-1160.</p>	
<p><b>805-818</b></p>	<p>The FDA 2009 PRO Guidance had it right: “... anchor-based approach to defining responders makes use of patient ratings of change administered at different periods of time or upon exit from a clinical trial. These numerical ratings range from <i>worse</i> to <i>the same</i> and <i>better</i>. The difference in the PRO score for persons who rate their condition <i>the same</i> and <i>better</i> or <i>worse</i> can be used to define responders to treatment. Patient</p>	<p>An adjustment is needed by subtracting the mean COA score from no-change group on the anchor measure (that is, subtracting this mean COA score from the mean COA score associated with meaningful change on the anchor, for the reasons cited in the adjoining column.)</p>



ratings of change are less useful as anchors when patients are not blinded to treatment assignment.”

For more on this crucial point, please see Chapter 7 in Bushmakin AG, Cappelleri JC. A Practical Approach to Quantitative Validation of Patient-Reported Outcomes: A Simulation-Based Guide Using SAS. Hoboken, New Jersey: John Wiley & Sons. 2022.

“The point is that it is the *difference in the scores* for persons who rate their condition *the same* and *better* (or *worse*) should be used as a meaningful threshold. Such a calibration (by *the same* category) is analogous to adjusting for placebo in an active intervention study, where it is the relative or placebo-adjusted treatment effect that is important, rather than the unadjusted or absolute effect. Thus, this approach calibrates the relationship between change in target PRO measure and change in anchor external measure.

By contrast, a non-calibrated approach will make no such correction and therefore does not adjust for *the same* category on the anchor by subtracting out its corresponding mean change score on the target PRO. This non-calibrated approach, in effect, forces no change on the anchor to correspond to no



	<p>change on the target PRO and therefore assumes a perfect harmony exists in the relationship between these two measures, which is not generally the case (as the two measures are expected to measure similar but distinct concepts).”</p>	
822-823	<p>FDA recommends that sponsors use multiple anchor measures to inform decisions about a plausible range of MSD values.</p> <p>Differences in COA scores should be related to differences documented by one or more anchors. The stronger the relationship, the more confidence in translating differences in the anchor to differences in COA scores</p>	<p>Further guidance on how to triangulate MSD threshold from multiple anchors will be very helpful, especially using weighted approach based on relationship between COA score and the anchor.</p>
837-839	<p>The draft guidance states, “An anchor should be plainly understood by respondents in the context of use,” and recommends to formally test anchors in cognitive interviews. In many cases, the anchors are PGI-S and PGI-C based on a widely tested and used model, with short questions using simple wording. These global scales are generally customized to include the name of the disease and the concept of interest, which both are generally being widely used across the study and become familiar to the patient. The recommendation made here mean that</p>	<p>We suggest removing the recommendation to conduct formal testing of anchors. We believe this adds significant cost and possibly delays the conduct of the trial, with very limited if any benefit in terms of informing regulators, prescribers and patients’ decisions. Rather, flag that anchors should be inspired from existing simple, widely used models, and be kept concise in the wording, specific, and easily understandable by the target population.</p>



	<p>sponsors need to conduct additional research work, to validate a methodology (anchors) used to support empirical supportive information to interpret results from existing, extensively validated COAs.</p>	
<p>850-852, 936-942</p>	<p>"Differences in COA scores should be related to differences documented by one or more anchors. The stronger the relationship, the more confidence in translating differences in the anchor to differences in COA scores."</p> <ol style="list-style-type: none"> <li>1. It would be helpful if the word "related" were replaced with a more precise term or terms. Presumably this means correlated.</li> <li>2. It would be helpful to have more explicit &amp; quantitative guidance on this point, ideally considering the quantitative impact of the strength of relationship to estimation of MSDs, and whether threshold adjustments should be made in cases of weak COA-anchor correlation (Coon &amp; Cook, 2017).</li> </ol>	<p>Make language more explicit; add quantitative guidance</p>
<p>851-852, 945-947</p>	<p>The statements “The stronger the relationship, the more confidence in translating differences in the anchor to differences in COA scores” and “threshold estimates from some anchors can be weighted more heavily than those estimates from other anchors based on the quality of</p>	<p>Recommend the use of triangulation methodology to derive a single threshold or small range, based on weighting estimates by anchor quality as suggested in current text. Add appropriate citation for this described approach: Trigg, A., Griffiths, P. Triangulation of multiple meaningful change thresholds for patient-reported outcome scores. Qual Life Res 30, 2755–2764 (2021).</p>



	<p>the anchor” point towards the practice of triangulation. There is a method to triangulate different threshold estimates using a weighted average, where the weights are driven by the correlation between anchor and COA score (Trigg &amp; Griffiths 2021). The confidence interval around the weighted average can guide the range. This method seems appropriate given the above statements.</p>	
854-856	<p>Original text:</p> <p>“Selected anchors should be assessed at comparable time points to the target COA. Sponsors should also ensure that, where applicable, the recall period of the anchor measure is consistent with the period covered by the COA-based endpoint.”</p> <p>We believe that additional guidance regarding a recall period should be referenced here.</p>	<p>This guidance is based on “should”. Suggested stating: Selected anchors should be assessed at time points consistent with the target COA.</p> <p>BIO further recommends that Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders (2022) describing the selection criteria for an appropriate recall period be referenced here.</p>
878-879	<p>“When a suitable anchor cannot be found, sponsors can consider other methods to inform the choice of MSD, such as Idio Scale Judgment (Cook et al. 2017).”</p>	<p>Further information on the methods and analysis expectations for the Idio Scale Judgment would be helpful.</p>
883-994	<p>Section III.B.1.b "Analyses of anchors to inform choice of meaningful score difference”.</p> <p>It is unfortunate that there is no specific endorsement of a specific analytical</p>	<p>Endorse specific methodology for MSD estimation or include/cite detailed example.</p>





	methodology (i.e., algorithm) to calculate MSD. If the agency is not prepared to endorse one or more specific algorithms, perhaps a detailed example could be included (or explicitly referenced) to highlight an acceptable use of a particular method or combination of methods.	
893-894	Recommend clarifying that the correlation coefficient between the COA score differences and the anchor (e.g., PGIC) or change in the anchor score (e.g., PGIS) is a useful measure of the association.	Add text “A correlation coefficient should be used to quantify the strength of this association.”
901	For ordinally-scaled anchors measured at two time points	“For ordinally-scaled anchors measured at two or more time points”
908	Different COA change scores within a target anchor category are often observed by patient baseline severity groups. For example, patients who are classified as mild at baseline may need a larger improvement in scores to reach the target anchor category.	Please clarify what acceptable mean/median score differences would be across baseline categories within the target anchor category.
913-916	It is important to consider that even in the case of a single true MSD, the values in the table would show some variability.	It would be helpful if the Agency provided additional guidance on how Table 1 will be used to determine whether there is more than one MSD.
919	Using the magnitude of a <i>state change</i> ( <a href="#">Wyrwich, Spertus et al. 2004</a> ) can greatly aid in selecting the defined MSD. Including this in Line 1027 is too late for this discussion to flow.	Recommend adding <i>state change</i> to the suggested methods for selecting an MSD or range of thresholds early on in the MSD discussion. For example, the May 4, 2023, FDA Workshop began with an MSD of 7.5 vs. 8.5 in the early steps. Knowing the state change for the COA in question could have provided an easy path to the conclusion on the appropriate MSD level.



921	By definition, use of the median or means to derive meaningful within-patient change will misclassify approximately half of the target anchor group. It would be helpful if the Agency clarify what an acceptable method is for balancing the errors referenced here.	Please clarify what an acceptable method is for balancing the errors referred to by the Agency.
929-954	Generally, a wider range of thresholds should be selected when there is greater uncertainty about what patients would regard as an impactful difference....	No guidance for a wider range of threshold based on the correlation of the COA and anchor. Is there a minimum correlation coefficient that is acceptable?
956-980	Other methods, such as a model Rasch, can be used to define the region of the scoring range associated with each PGIS category and should be considered.	Please consider including other scoring methods in this discussion.
956-1018	<p>III.B.2 "Interpreting in Terms of Meaningful Score Regions"</p> <p>MSRs appear not to require longitudinal studies to calculate MSD based on change in anchor, which is welcome. But we should seek clarity: Assuming that the criteria for a strong anchor have been met (see III.B.1.a), can meaningful-score regions be established using cross-sectional studies alone?</p>	Clarify: "Assuming that the criteria for a strong anchor have been met, can meaningful score regions be established using cross-sectional studies alone?"
956-1018 and 1172-228	<p>The use of meaningful score regions to interpret clinical trial endpoints is questionable.</p> <p>In practice, these regions have been used to help patients and clinicians understand a</p>	Please provide further justification for the use of meaningful score regions as applied to interpretation of treatment effect, including appropriate citations.



	<p>single patient’s health at a single point in time. This can be described as a cross-sectional, individual-level interpretation. The notion that these regions can be applied to interpret treatment differences in change from baseline, or within-individual change over time, requires further justification including appropriate citations.</p> <p>Another criticism is that one patient could report a 10-point improvement within a single region, versus another reporting a 3-point improvement spanning two regions. It seems unintuitive to conclude that the second patient experienced a meaningful change but the first did not. So, this criticism should be addressed.</p>	<p>Address the criticism that larger score changes could be considered less meaningful than lower score changes.</p>
986-990	<p>Figure 1. Box-whisker plots. This figure uses an unusually straightforward example of MSRs. “Bookmarking or similar methods in which patients, caregivers, and/or clinicians make judgments to sort patient experiences into a small number of ordinal categories (e.g., none, mild, moderate, or severe) (Cook et al. 2019).”</p>	<p>We recommend that the guidance informs the reader that this is an unusually straightforward example of MSRs or alternatively recommend replacing the example with one that is more likely to be encountered by a sponsor. Also recommend that the guidance inform the reader that other psychometric methods can better inform appropriate cross-sectional cut points than box plots.</p> <p>Bookmarking is an interesting approach, however likely limited by the fact that patients tend to struggle differentiating/describing health states with medium severity. They often tend to agree on descriptions/experiences associated with mild severity or maximum severity but have more difficulty recognizing and categorizing moderately severe health states. As a result, it may be difficult to obtain</p>



		distinctly meaningful score regions using this method, especially in rare diseases with heterogenous presentations.
992-1011	It is not clear on what is the metric. Is it individual change or group difference?	It is not clear on what is the metric. Is it individual change or group difference?  Assume a simple linear regression. What is the outcome variable and what is the predictor variable? Is the target COA measure the outcome or predictor? Is the anchor measure the outcome or predictor?
1005-1009	It would be helpful if the Agency would clarify if an item within a multi-item PRO measure could be used as an anchor to define an MSR for the score of that same multi-item measure?	It would be helpful if the Agency provided suggestions on how a sponsor might select the item to use to define the MSR and what evidence is needed to support such item selection? When would such a method be acceptable?
1013-1018	This approach is more promising than the one in the previous lines, but more details would be helpful.  A Rasch approach could also be used. Specifically, the Wright-Andrich item-step map could be generated to illustrate item-step severity in comparison to patient severity. (See, for example: Gugiu et al, 2019, <i>Journal of Child and Family Studies</i> .) Ideally, one would then superimpose the group means for the baseline PGIS and there would then be regions defined by item-steps. In absence of an anchor, one can conduct qualitative interviews with patients to determine meaningful regions based on the item steps.	We suggest adding additional details on the use of the Item Response Theory (IRT).  Suggest adding the citations for the roots or original ideas of IRT, such as: Rasch G. (1960). Probabilistic models for some intelligence and achievement tests. Copenhagen: Danish Institute for Educational Research; Chicago: MESA Press.  Lord, F. M. (1980). Applications of item response theory to practical testing problems. Mahwah, NJ: Erlbaum.  Masters, G. N. (1982). A Rasch model for partial credit scoring. <i>Psychometrika</i> , 47, 149-174.  Thurstone, L. L. (1925). A method of scaling psychological and educational tests. <i>Journal of Educational Psychology</i> , 16, 433-451.



1043-1045	<p>Original text:</p> <p>“Sponsors who are considering conducting exit interviews or surveys should submit a study protocol and interview guide to FDA for review as early as possible, ideally prior to beginning the registration trial.”</p> <p>We believe exit interviews or surveys are not the only options to inform meaningful differences or scores. We believe that the option of deriving meaningful differences or scores from a partial set of the registrational trial should also be mentioned.</p>	<p>BIO recommends the following revision:</p> <p>“Sponsors who are considering conducting exit interviews or surveys should submit a study protocol and interview guide to FDA for review as early as possible, ideally prior to beginning the registration trial. <u>In the case of small patient populations and expedited development programs, deriving meaningful differences or scores using a partial set of the registrational trial data could be another option and should be discussed in advance with the relevant review division.</u>”</p>
1094-154	<p>There are two main issues with the proposed use of ‘meaningful score difference’ thresholds to interpret average treatment differences at a group-level.</p> <ol style="list-style-type: none"><li>1. Methods to estimate meaningful between-group thresholds from a patient perspective are currently lacking. While the currently recommended anchor-based method (calculating mean difference between patients who reported to be “a little better” and those who reported to be “about the same” – see for example Terwee 2021, Trigg 2023), this makes a strong assumption that the between-group comparison is based on one group where nobody improves and another where everyone does. In practice we</li></ol>	<p>Remove the suggestion that thresholds for group-level and individual-level interpretation can be used interchangeably and estimated in the same way.</p> <p>Focus the presentation of meaningful score difference on meaningful within-patient change, to align with current emphasis in the field.</p> <p>If maintaining that group-level thresholds should be used and can vary by baseline score, please provide further justification including appropriate citations where possible.</p>



	<p>would assume not all patients receiving active treatment improve, and some patients in control arm improve. The current approach is therefore likely to overestimate thresholds for between-group differences and result in false conclusions that no meaningful difference is observed (when in fact it has at the group-level).</p> <p>2. The notion of a between-group threshold varying by baseline severity is questionable and is an example of conflating interpretation at the within-patient versus between-group levels. It is plausible to imagine within-patient thresholds varying by baseline score (e.g., patients with more severe pain require a larger improvement to consider it meaningful). But the notion of a between-group threshold varying by baseline score does not have a strong theoretical rationale and requires further justification.</p> <p>Terwee, C.B., Peipert, J.D., Chapman, R. et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. <i>Qual Life Res</i> 30, 2729–2754 (2021).</p>	
--	--	--



	<p>Trigg, A., Lenderking, W.R. &amp; Boehnke, J.R. Introduction to the special section: “Methodologies and considerations for meaningful change”. Qual Life Res 32, 1223–1230 (2023).</p>	
<p>1099-1120; 1127-1129, Figure 2; 1139-1141, Figure 3</p>	<p>The approach described here (represented by figure 2 and 3) is inconsistent with previous FDA documents and also with how MSD is defined in this document.</p> <p>The following is stated earlier (page 20; lines 777-778): “...Often, MSD is determined based on what patients would regard as a clinically meaningful within-patient change (i.e., improvement or deterioration from the patient’s perspective) ...”.</p> <p>From the above quote it is clear that when we state MSD, we mean “meaningful within-patient change (MWPC)” (and, conversely, when we say, “meaningful within-patient change,” then we call it MSD now).</p> <p>The example given by FDA is “... clinical trial comparing a new product A to a current product B, scores (0-20) on a PRO measure of functioning were analyzed using an ANCOVA with baseline functioning scores as the covariate...”</p>	



	<p>We know that the treatment difference will be the same in the above model whether PRO changes from baseline were used or where “direct” (i.e., absolute) PRO scores were used as the outcome (dependent) variable. Then the question arises on why we use MWPC (or MSD) for the interpretation of the differences of LSmeans “at 12 weeks post-randomization.”</p> <p>The FDA suggests using MSR analysis to create severity regions. In our opinion this should be done using <b>lengths</b> of score regions (based on MSR analysis). Generally, it -- length of MSR region -- will be simply a difference in the COA scores corresponding to 1-point difference in the PGIS.</p>	
1106-1108	<p>“Based on three different anchor-based analyses conducted using an independent sample of patients the sponsor prespecified a range of MSD for the PRO functioning measure of 3 to 5 points.”</p> <p>Consider adding “Based on three different anchor-based analyses conducted using an independent sample (where feasible), the sponsor...”</p>	





1110-1120	<p>This section appears to indicate that differences between treatments (e.g., treatment and placebo) should be compared against the MSD. This is a departure from the within-patient analyses (e.g., responder analyses) that are advocated elsewhere in the guidance. It is not clear whether the intention is that MSDs derived from within-patient differences can be applied to between-group differences. Arguably, it would be more consistent to apply a between-group difference to interpret treatment differences.</p>	<p>We suggest that the Agency clarify whether the intention is that MSDs derived from within-patient differences can be applied to between-group differences.</p>
1127	<p>Figure 2 does not incorporate the FDA’s understanding that a few “super responders” can lead to high mean change difference but few patients who actually got better in the treatment group.</p> <p>This figure serves as a key reason to investigate the MSD for patient in the context of use and then examine for the percentage of responders. Point estimates of the difference in means between two groups may mask important changes for individual patients or types of patients in each group. The MSD will thus not reveal whether some groups within the trial obtain a large benefit while other groups do not benefit at all.</p>	<p>Recommend revising Section III.C to apply the MSD to examine responder percentages per treatment group. Moreover, analysis of the cumulative distribution of patients’ response to the experimental treatment within each group compared to responses of the control groups can help in evaluating the consistency of effects across the entire distribution. This distribution curve will reveal the extent to which overall results are driven by outliers who improve or worsen more than others. A cumulative distribution curve provides information on what type of responses contributed to the mean group response and provides more useful data than a simple point estimate of the difference between group mean changes.</p>
1139	<p>Figure 3 has errors, including (1) the mean and CI provide are inconsistent with the figure and the text, (2) even if correctly</p>	<p>Recommend the same changes as noted for line 1127, namely: (<i>see line above</i>)</p>



	displayed to match the text, this figure serves as a key reason to investigate the MSD for patients in the context of use and then examine for the percentage of responders and the CDF. (See comments for line 1127 above.)	
1157-1162	Please provide a suggestion of how the proportion of patients experiencing meaningful within-patient change could be summarized, if the MSD varies by baseline score.	Please provide a suggestion of how the proportion of patients experiencing meaningful within-patient change could be summarized, if the MSD varies by baseline score.  Suggest that using thresholds based on percentage change is one way to address this (e.g., 30% reduction in pain).
1207-1210	<p>This depiction (represented by figure 4), alongside modified figures 2 and 3 (i.e., using MSRs in figures 2 and 3) represents a good approach to assess clinical relevance of the treatment effect.</p> <p>Note that for this interpretation (figure 4) the model should use "direct/original" (absolute) COA scores (not the changes from baseline!)</p> <p>The huge treatment effect (5.8) depicted in Figure 4 will demonstrate much more interpretable data about the treatment if we understood the MSD (not MSR) and could understand how many patients achieved the MSD. This figure serves as a key reason to investigate the MSD for patient in the context of use and then examine for the</p>	Consider refining these lines as noted.



	percentage of responders and the CDFs reported.	
1220-1228	Please provide example plots to match these descriptions.	Please provide example plots to match these descriptions.
1262-1266	Patients', clinicians', and/or caregivers' knowledge of treatment assignment (e.g., in single arm trials, open label trials, open-label treatment extension periods) is likely to influence how they report information on a PRO, ClinRO, or ObsRO measure, or how they engage with PerfO tasks (e.g., amount of encouragement given to patients when measuring walking distance), which will bias estimates of treatment effect.	<p>The presumption that a lack of masking is likely to bias COA data is over-general and not supported by research. While double-blind randomized designs are preferred, many situations exist in which they are unfeasible or unethical (e.g., indications in which no SoC or ethical alternative treatment is available) and we should not dismiss COA data as biased in these situations. Examples of publications supporting this position include Atkinson et al 2016 "Trustworthiness of Patient-Reported Outcomes in Unblinded Cancer Clinical Trials" and Lord-Bessen et al. 2023 "Assessing the impact of open-label designs in patient-reported outcomes: investigation in oncology clinical trials". also refer to 2022 FDA OCE workshop on the topic <a href="#">FDA Workshop: 7th Annual Clinical Outcome Assessment in Cancer Clinical Trials Workshop - 06/29/2022   FDA</a> also refer to 2022 FDA OCE workshop on the topic <a href="#">FDA Workshop: 7th Annual Clinical Outcome Assessment in Cancer Clinical Trials Workshop - 06/29/2022   FDA</a><a href="#">FDA Workshop: 7th Annual Clinical Outcome Assessment in Cancer Clinical Trials Workshop - 06/29/2022   FDA</a></p> <p>An alternative may be to soften the bias declaration language to something like the following: "Patients', clinicians', and/or caregivers' knowledge of treatment assignment (e.g., in single arm trials, open label trials, open-label treatment extension periods) <del>is likely to</del> may influence how they report information on a PRO, ClinRO, or ObsRO measure, or how they engage with PerfO tasks (e.g., amount of encouragement given to patients when measuring walking distance); <del>which will</del> may bias estimates of treatment effect).</p>



		I think that softens the FDA’s declaration of bias w/o sparking a larger debate.
1272-1274	A practice effect (sometimes also called a learning effect) is any change that results from....	<p>We note that learning effects and practice effects are not synonymous. A learning effect is a short-term practice effect.</p> <p>Further, a practice effect does not have to be a “change”, a practice effect is a bias. Practice and learning could prevent a score from worsening, therefore the bias being no change. For example, though Parkinson’s symptoms are progressing the progression is not reflected in the PerfO because the participant has learned how to complete the task.</p> <p>Suggest revising to state: A practice effect (sometimes also called a learning effect) is a bias that results from...</p>
1281	The language about practice effects is unclear.	<p>Suggested revision:</p> <p>“Practice effects may be problematic for studies conducted to support a <del>medical product</del> regulatory application <u>for a therapeutic drug, biologic, or device.</u>”</p>
1295-1320	Some of the approaches listed for attenuating practice effects, specifically “Increase the length and number of assessments for the run-in period” (line 1314), may add considerable burden to the study participant and be in contradiction to section IV.A.7 (Minimizing Patient Burden).	<p>BIO recommends the following revision on line 1297:</p> <p>“Some general strategies for mitigating practice effects are summarized below. <u>These strategies should be considered in relation to the corresponding additional patient burden introduced during the trial. See Section IV.A 7 Minimizing Patient Burden.</u>”</p>
1329-1358	As we seek to make clinical trials more inclusive, there may be participants enrolled in trials who use assistive devices not	Consider adding a line addressing the use of other assistive devices (and use them from the start of the trial) to support activities of daily living that do not impact the actual COA of interest in the trial



	related to the reason for the trial (such as a blind participant who uses a white can for assistance in navigating the world, but not for stability).	
1352-1353	If the use of the assistive device could be influenced by treatment and altering the need for the assistive device is not a primary goal of treatment, construct a supportive endpoint based on whether an assistive device is used.	May be helpful to provide some examples here.
1763	The Vickers 2001 paper is not cited within the main text. We assume it was intended to belong in the section at 293 but note that the comparisons in this paper compare percent change-from-baseline as an outcome without controlling for baseline as a covariate. In practice, statisticians would control for baseline scores in such an analysis, so the results of this study are not directly relevant to common practice.	Delete Vickers 2001 reference.