GENERAL CONSIDERATIONS IN DEVELOPING AND EVALUATING OUTCOME MEASURES OF ABUSE LIABILITY

Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials
IMMPACT-XII: Outcome Measures for Human Experimental and Clinical Studies of the Abuse Liability of Analgesic Medications
Hilton Rockville
October 1-2, 2009

Laurie Burke
Study Endpoints and Labeling Development
Office of New Drugs
FDA/CDER

Overview

a) Treatment benefit

b) Substantial evidence

d) Content validity of outcome measures

e) Other measurement considerations

Treatment benefit

- *Treatment benefit* — The impact of treatment on how a patient *survives, feels, or functions*.
  - The impact of treatment on other concepts are surrogate measures of treatment benefit
- Treatment benefit may be measured as
  - Comparative efficacy (e.g., an improvement or delay in the development of symptoms or decrements in function)
  - Comparative safety (e.g., a reduction or delay in treatment-related toxicity or reduced drug abuse liability)

Treatment Benefit Claims

- May be found in…
  - Indications section of labeling
  - Other sections of labeling
    - Clinical Studies section
    - Clinical Pharmacology section
    - Other
  - Advertising and promotion

Substantial Evidence

- 21 CFR 314.126(a)
  - *"Reports of adequate and well-controlled investigations provide the primary basis for determining whether there is 'substantial evidence' to support the claims of effectiveness for new drugs. Therefore, the study report should provide sufficient details of study design, conduct and analysis to allow critical evaluation and a determination of whether the characteristics of an adequate and well-controlled study are present."*

Characteristics of an *Adequate and Well-Controlled* Study

- 21 CFR 314.126 (b)

(1) Clear statement of objectives
(2) Study design permits valid comparison (appropriate control)
(3) Select patients with disease/condition (treatment) or at risk of disease (prevention)
(4) Baseline comparability (randomization)
(5) Minimize bias (blinding, etc.)
(6) Appropriate methods for assessment of outcome
(7) Appropriate methods of analysis

Patient Selection

- 21 CFR 314.126(b)(3)
  - *The method of selection of subjects provides adequate assurance that they have the disease or condition being studied, or evidence of susceptibility and exposure to the condition against which prophylaxis is directed.*

Well-defined and Reliable Endpoints

- 21 CFR 314.126(b)(6)

  – *"The methods of assessment of subjects' response are well-defined and reliable. The protocol for the study and the report of results should explain the variables measured, the methods of observation, and the criteria used to assess response."*

FDA Initial Review of Measures to Support Claims

1. Do the study objectives state what we are measuring, i.e., the "concept"?
2. Is the instrument "fit for purpose"?

   Are the measurement properties documented and adequate?

   - Target concept
   - Target population
   - Target indication

Content Validity

The extent to which the score produced by an instrument:

- Measures the targeted concept
  - Contains the relevant and important aspects of that concept;
  - The items represent a sufficient sampling of content to represent the concept
- Matches the targeted objectives and claims
- Is meaningful/comprehensive in the targeted population
- Is interpretable in the planned study

What Is an Instrument/Measure?

- Items
- Response option
- Recall period
- Structure (e.g., subscales)
- Scoring
- Instructions for use
- Interpretation guidelines
- Measurement property documentation

Other Measurement Properties Are Important…But are evaluated after content validity is established

- Construct validity: Demonstrates expected relationships with other measures or with scores produced in patient groups known to be similar or diverse
- Reliability: Demonstrates: stability of scores over time; internal consistency; agreement between assessors
- Ability to detect change: Demonstrates how scores change over time in response to an intervention

Language and Culture

- Testing in targeted language and culture groups is an essential consideration for all reporting scales—PRO and ClinRO
- Should be taken into consideration early in instrument development so that content validity can be established concurrently across language and culture groups

For ClinROs: Inter-rater Reliability

- Degree of agreement among raters
- Influenced by many factors including:
  - Level of training/experience/specialty
  - Perspective (community vs tertiary care setting)
  - Health care system
  - Clinical culture

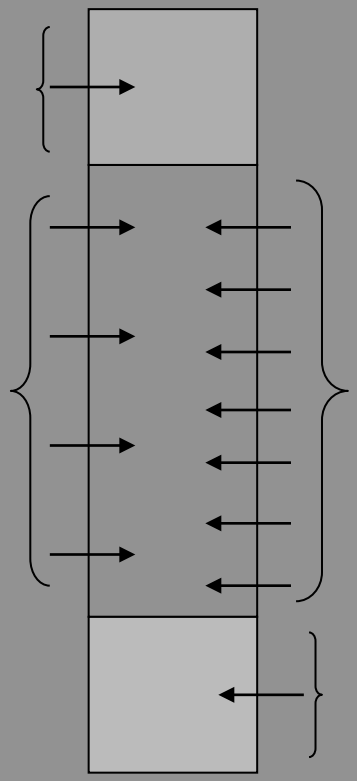
Establishing Content Validity for Reporting Scales

- **First step: Identify the targeted concepts**
- **Next steps (iterative):**
  - Literature review
  - Expert opinion
  - Qualitative research with a sample of those who will be reporting (patients, caregivers, clinicians) similar to the targeted trial participants
  - Quantitative testing
    - Factor analysis
    - Rasch/IRT

Content Validity of a PRO: Intersection of Disease Attributes & Patient Experience

Disease Attributes

Patient Experience



Observed (Clinician or Lab)

Signs & symptoms of disease

Patient-reported disease experience

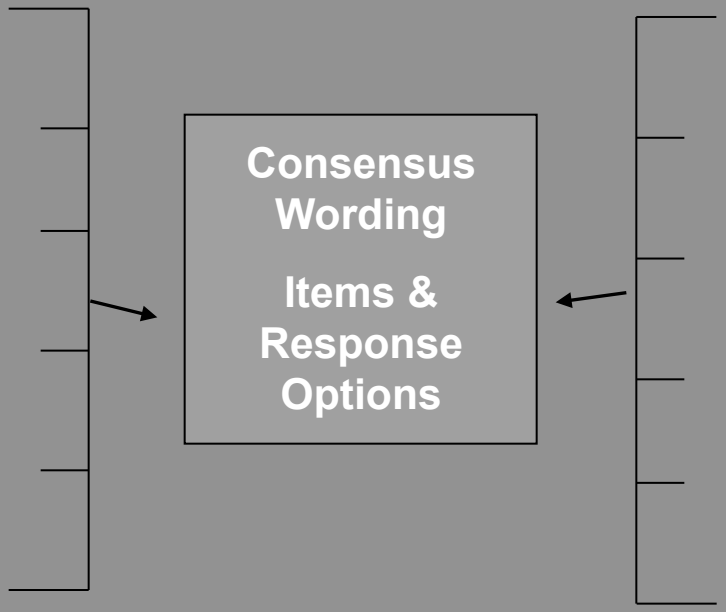
Patient-reported experiences UNRELATED to disease attributes

ISPOR Task Force on Content Validity of Existing Instruments, 2009

Content Validity:
Content Consensus through Qualitative Research

Concept Elicitation
(Focus Groups & Interviews)

Instrument Evaluation
(Cognitive Interviews)



ISPOR Task Force on Content Validity of Existing Instruments, 2009

Content Validity Review

Based on the process used to develop and select the structure and items
  - Confidence in methodological rigor; Not face validity

Is a targeted measurement concept identified?

Does the instrument sufficiently measure all of the important concerns related to the targeted concept in the targeted study population?

Is general vague language avoided? (This is generally used when the targeted concept is not defined and the respondent does not report the basis for the rating.)

Is the rating scale written in language that is familiar and meaningful to the respondent?

Q: In the past 7 days, have you had adequate relief of your irritable bowel symptoms?

- **"Relief"**

  Refers to a comparison of present to some unspecified time in past
- **Binary response option: yes/no**

  Does not quantify response or absence of symptoms
- **Recall Period**

  Is 7-day recall appropriate?

Q: In the past 7 days, have you had adequate relief of your irritable bowel symptoms?

- Huge variability (what's "adequate"?)
- Fails to quantify baseline severity (mild, moderate, severe)
- Fails to quantify treatment effect (minimal improvement vs. complete resolution)
- Fails to capture worsening symptoms

"Misuse/Abuse Liability" Measurement: Similar to "Satisfaction with Treatment"?

- Determined by the relative discrepancy between the expectations held before treatment compared with the perceived performance of the treatment
- Expectations are based on many attributes of the treatment including effectiveness, onset of action, side effects, confidence in the treatment, disruption in life, dosing and convenience/ease of use
- Importance weightings for these attributes are highly variable between patients

"Misuse/Abuse Liability" Measurement: Similar to "Health Related Quality of Life"?

- A multidomain concept that represents the patient's overall perception of the effect of illness and treatment on physical, psychological, and social aspects of life.
- Claiming a statistical and meaningful improvement in HRQL implies:
  - (1) that all HRQL domains that are important to interpreting change in how the clinical trial's population feels or functions as a result of the targeted disease and its treatment were measured;
  - (2) that an overall improvement was demonstrated
  - (3) that no decrement was demonstrated in any important domain.

What is the CONCEPT measured?
Does the CONCEPT match the targeted CLAIM?

PRO:

- Liking ratings
- Symptoms (e.g., blurred vision, spaced out, euphoria)
- Take again/street value/ monetary worth
- Categorization of effect to be like known drug class
- Drug strength assessment

ClinRO:

- Muscle-relaxation, posture
- Impaired speech
- Observed confusion
- Overall strength of drug effect
- Global based on DSM-IV-TR

Performance Measures:

- Motor speed, coordination, and reaction time
- Cognitive performance
- Memory

Physiologic Measures:

- Pupillary dilation/constriction
- Heart rate, blood pressure
- Skin temperature
- Urine drug screens

Composites:

- Drug attractiveness

Concept Clarification

- **What is the target population of concern?**
- **What is the expected problem?**
- **How can this problem be addressed?**
- **What is the intended outcome/concept/claim?**
  - Improve something? Stabilize something? Prevent something?
- **How is this concept currently defined?**
  - Empirically? Clinically?
- **How is this concept currently measured?**
  - Instrument? Other?
- **Is this approach or instrument appropriate?**
  - Fit for purpose? Sufficiently sensitive? Sufficiently meaningful?

Summary

- Any conclusion of treatment benefit needs to be based on substantial evidence
- Substantial evidence demands adequate and well-controlled studies
- Adequate and well-controlled studies must include well-defined and reliable assessments
- Well-defined and reliable assessments are not simple to develop and should always begin with the establishment of content validity

Extra slides

"Clinically significant" change

- Small randomized intervention study needed to determine a responder definition in the target population
- Alternatively, for study results using continuous variables, FDA recommends presentation of the entire distribution of responses for treatment compared with control groups.
  - Avoids the need to pick a responder criterion
  - The cumulative distribution function displays a continuous plot of the percent change from baseline on the X-axis and the percent of patients experiencing that change on the Y-axis.

Figure 1: Percentage of Patients Achieving Various Levels of Pain Relief as Measured by Pain Severity at 48 Hours Compared to Baseline- Post Operative Bunionectomy

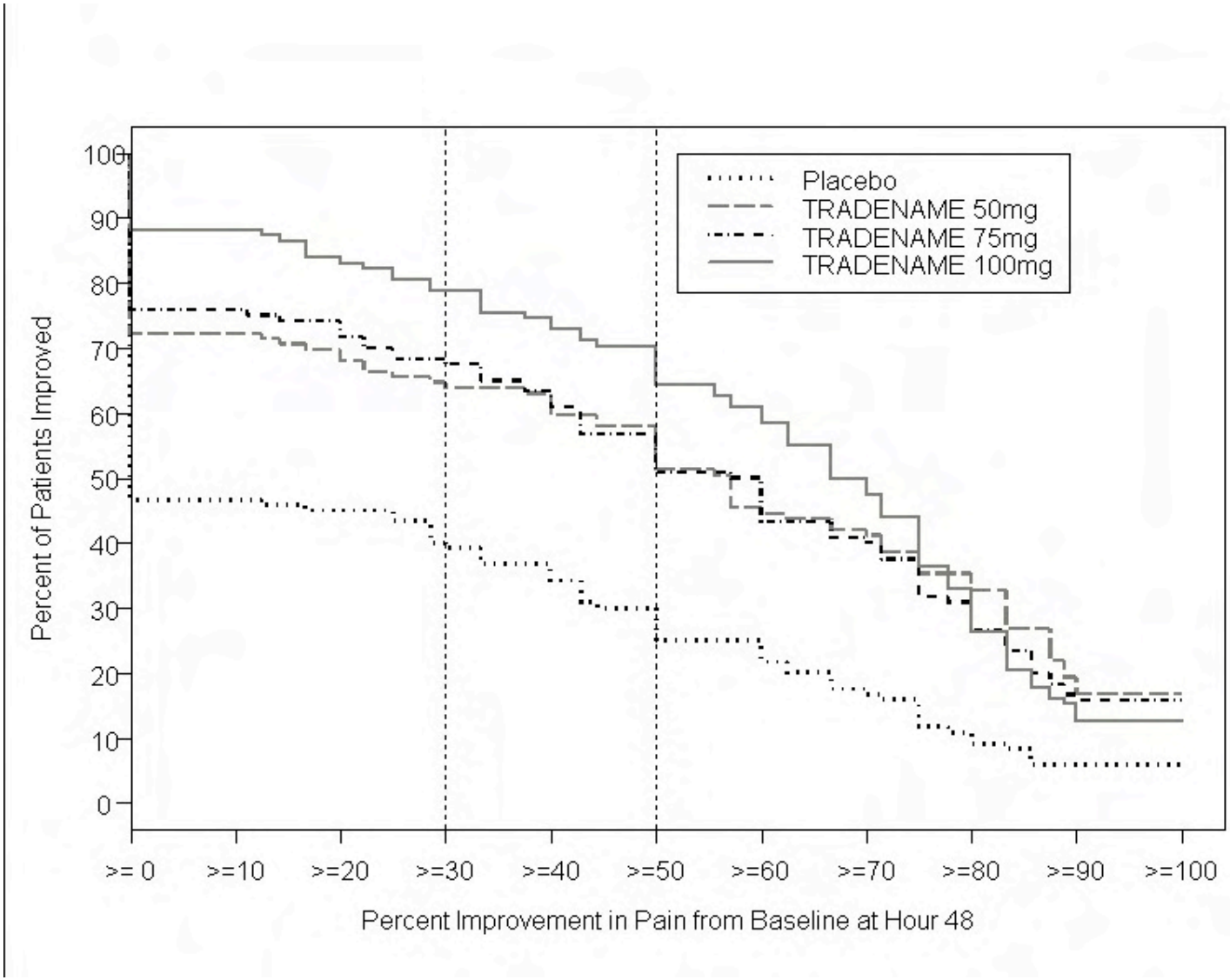
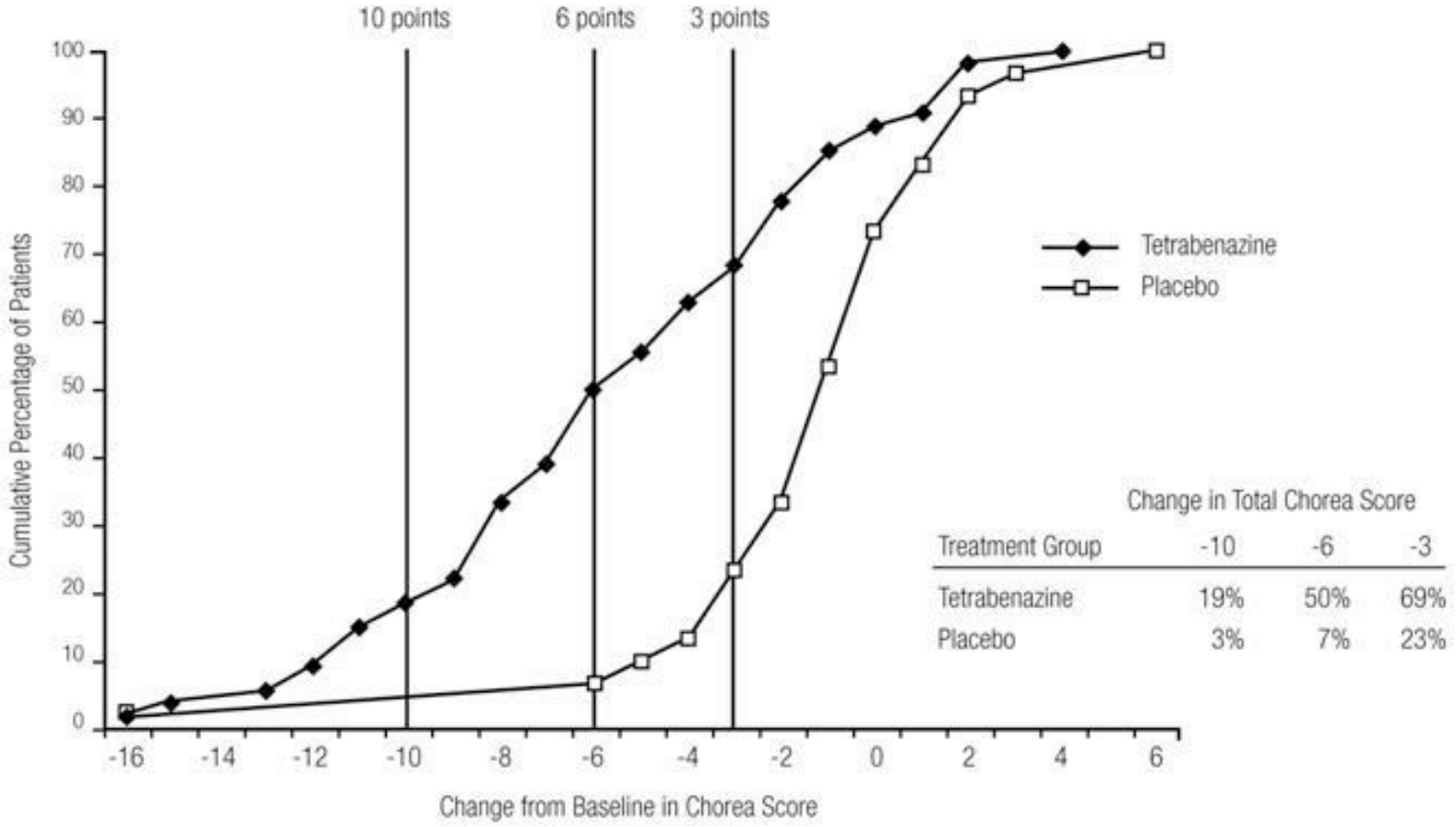


Figure 2. Cumulative Percentage of Patients with Specified Changes from Baseline in Total Chorea Score.



Treatment Group	Change in Total Chorea Score -10	-6	-3
Tetrabenazine	19%	50%	69%
Placebo	3%	7%	23%

Assay sensitivity in studies to show an absence of an adverse event

- To show that a drug does not have a particular adverse effect by showing similar rates of the event in drug-treated and placebo-treated patients, placebo-controlled trials have the same assay sensitivity problem as any equivalence or non-inferiority trial (see ICH E10, section 1.5.1).
- To interpret the result, one must know that if the study drug had caused an adverse event, the event would have been observed. Ordinarily, such a study should include an active control treatment that does cause the adverse event in question

Reporting of negative findings

- A negative finding can be reported if the absence of the adverse reaction is convincingly demonstrated in a trial of adequate design and power
- A concept must be convincingly measured before it can be reported as a negative finding
- A concept must be identified before the measurement can be deemed adequate