Panel Discussion 2

Identifying Optimal Recall Period for Measuring Specific Concepts

SECOND ANNUAL
PATIENT-REPORTED OUTCOME (PRO) CONSORTIUM WORKSHOP

March 15, 2011 ■ Silver Spring, MD

CRITICAL PATH INSTITUTE
collaborate · innovate · accelerate

FDA

Identifying Optimal Recall Period for Measuring Specific Concepts

Moderator:
Josephine M. Norquist, MS

Panelists:
Arthur A. Stone, PhD
Dennis A. Revicki, PhD
Elektra Papadopoulas, MD, MPH
Joseph G. Toerner, MD, MPH

CRITICAL PATH INSTITUTE
collaborate · innovate · accelerate

FDA

Opening Remarks and Introductions

Moderator:

Josephine M. Norquist, MS

Merck Sharp & Dohme, Corp.

Panel Overview

- The selection of the recall period is an important decision in the development of a PRO measure
- **Recall period** — *"The period of time patients are asked to consider in responding to a PRO item or question. Recall can be momentary (real time) or retrospective of varying lengths. "* (FDA PRO Guidance)
- *"PRO instruments that call for patients to rely on memory, especially if they must recall over a long period of time, compare their current state with an earlier period, or average their response over a period of time, are likely to undermine content validity. …., **items with short recall periods or items that ask patients to describe their current or recent state are usually preferable."*** (FDA PRO Guidance)
- The choice of an adequate recall period may depend on
  - **Intent of the PRO measure**
  - **Nature of disease or condition**
  - **Design and length of study**
  - **Ability to recall and patient burden**

Are daily diaries the best way to measure PROs?

Arthur A. Stone, Ph.D.
Department of Psychiatry and Behavioral Science
Stony Brook University

SECOND ANNUAL
PATIENT-REPORTED OUTCOME (PRO) CONSORTIUM WORKSHOP

March 15, 2011 ■ Silver Spring, MD

Co-sponsored by

CRITICAL PATH INSTITUTE
collaborate · innovate · accelerate

FDA

Disclosures & Support

Senior Scientist, Gallup Organization

Senior Consultant, PRO Consulting

Scientific Advisory Board member,

Wellness & Prevention, inc.

Supported by grants from the National Institute for Arthritis and Musculoskeletal Diseases (NIAMS) and the National Cancer Institute (NCI)

Overview

- Common terms
- Conceptual Arguments for Brief Reporting Periods
- Conceptual Arguments for Longer Reporting Periods
- Empirical Results of studies using Real-time assessments as standard for evaluating longer recall assessments
- Discussion of findings

Common terms

- OUTCOME PERIOD: the period of time that we are trying to characterize
  - Depending on the purpose of the research, the Outcome Period could be an hour, a day, a week, or weeks.
- REPORTING PERIOD: the period of time asked about about in an assessment
  - “How intense was your pain over the last 7 days?”
  - “Did you have an asthma attack today?”
  - “How bad is your fatigue right now?”
- Outcome Period need not equal Reporting Period

Overriding Goal

- Valid outcomes

- We want to be sure we are measuring exactly what we set out to measure
  - If respondents are attending to a specific instance in time, then they were not providing information about the entire reporting period
  - If respondents are basing their answers on "surrogate" information, such as reduced activity level, then they were not providing the requested information

Conceptual Arguments for Brief Recall Periods

- Our ability to remember symptoms is limited by memory capacity
  - Switch from episodic to semantic memory
- Retrospective judgments can be a *reconstructive*
  - This can lead to biased judgments
  - Heuristics: “peak-end” “immediate states” “duration neglect”
- Cognitive Interview results
  - Context: Pain intensity rating for last week
  - Result: No systematic review of the week

Cognitive Interview Results

167	"Well, I was thinking about how many days I was in pain, and whether I was in pain the whole day or not. "
188	"In the past I played golf one day, so I had pain in my left arm and some numbness. And two days 'cause the weather's been bad my hands were swollen, I had a time grasping clubs. So I didn't have no pain, but again it wasn't the worst."
153	"In my particular condition it comes in spurts. Sometimes it will be horrible for an hour or two, but then if I take more medicine or if I put ice packs, or if I do any of the things I previously described to you for a little while, I will get some relief. It won't go away, but it'll be less and somewhat tolerable. Last week there were three days consecutive that were horrific. It was constant pain, constant burning, constant throbbing, constant stabbing. Constant for three days. That to me is the worst possible and then some."
133	"I was trying to…actually I had that day in mind when I had gone shopping and that day was extremely difficult. But it's hard to say, because there's days that I'm so used to it that I don't even think about it. So, its very hard to rate. That's the best I can do with that question."

Conceptual Arguments for Brief Recall Periods

- Participants answer questions very quickly, so how can they be reviewing experiences during the reporting period?
  - If they are not providing information about the reporting period, then we have mislabeled the outcome

Conceptual Arguments for Longer Recall Periods

- Some content/domains may be better "suited" to longer recall
  - Opinions
  - Slow-changing states
- Improved practicalities
  - Participant burden
  - Expense
  - Infrequent events

Empirical Support

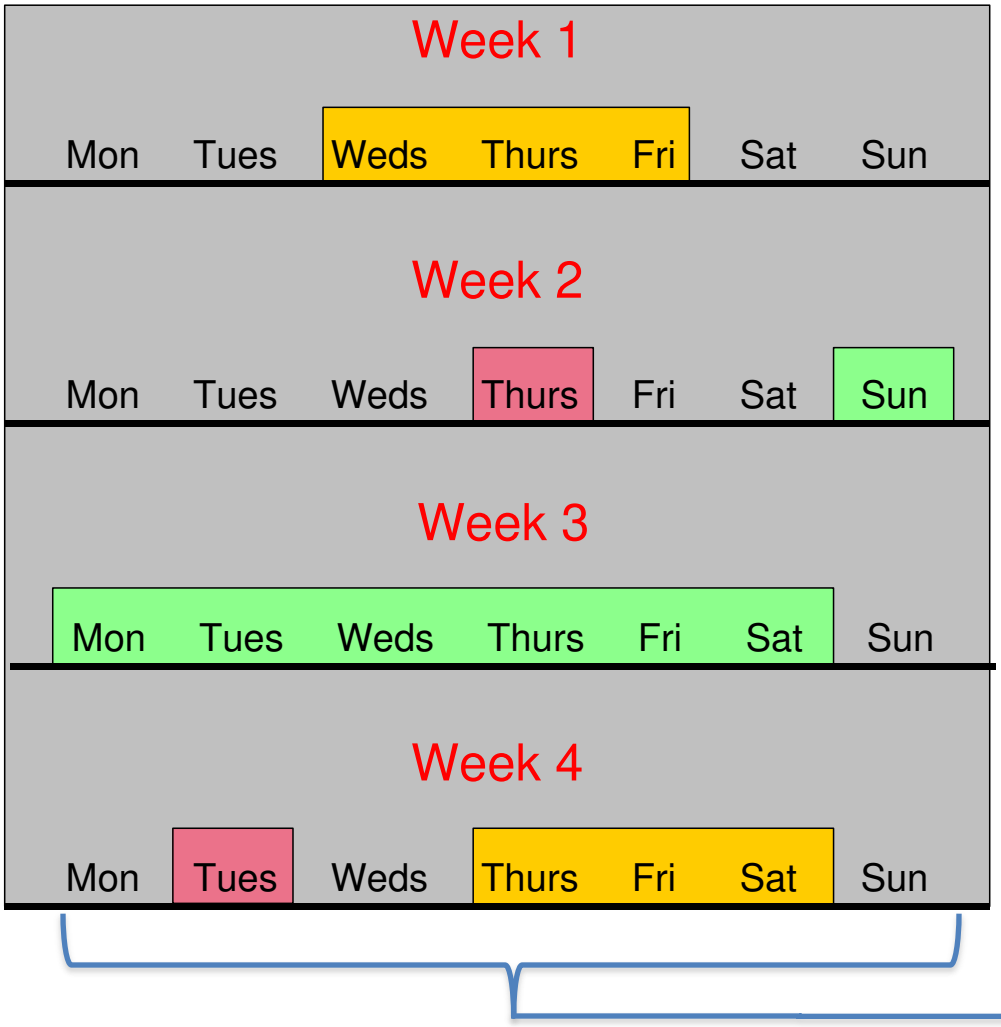
- Using real-time methods as a standard for evaluating recall methods

- End-of-day, Experience Sampling (ESM), and Ecological Momentary Assessment (EMA) techniques used to capture real-time data

Empirical Support

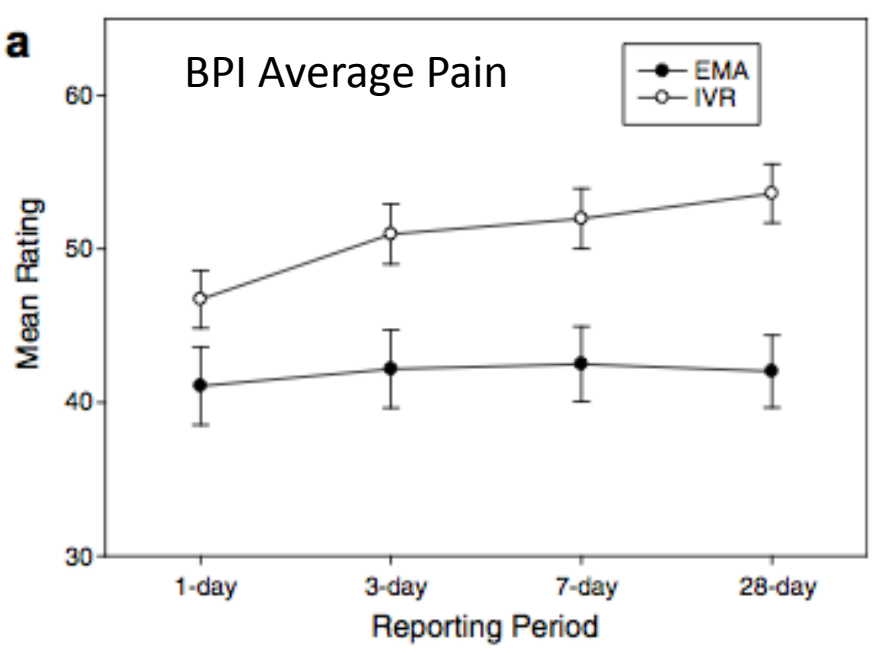
- Results from comparisons of real-time data capture studies (ESM, EMA) have shown:
  - Level differences: Sometimes recall measures indicate higher levels (eg, pain and fatigue)
  - Correspondence: About 50-65% of the variance is shared between 1-week recall ratings and the average of many (often 42 or more) momentary ratings of pain intensity throughout the same period
- Is this reasonable overlap?
  - Is the glass ½ empty or ½ full? More later
- Should real-time data be considered the gold standard?
  - It is a sampling technique, so not perfect

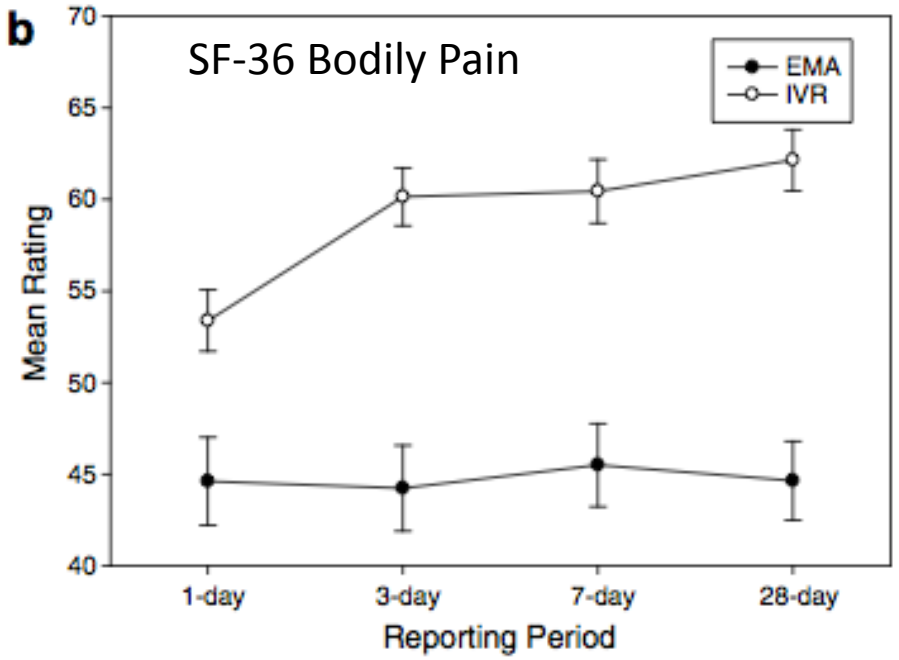
General Design of Studies



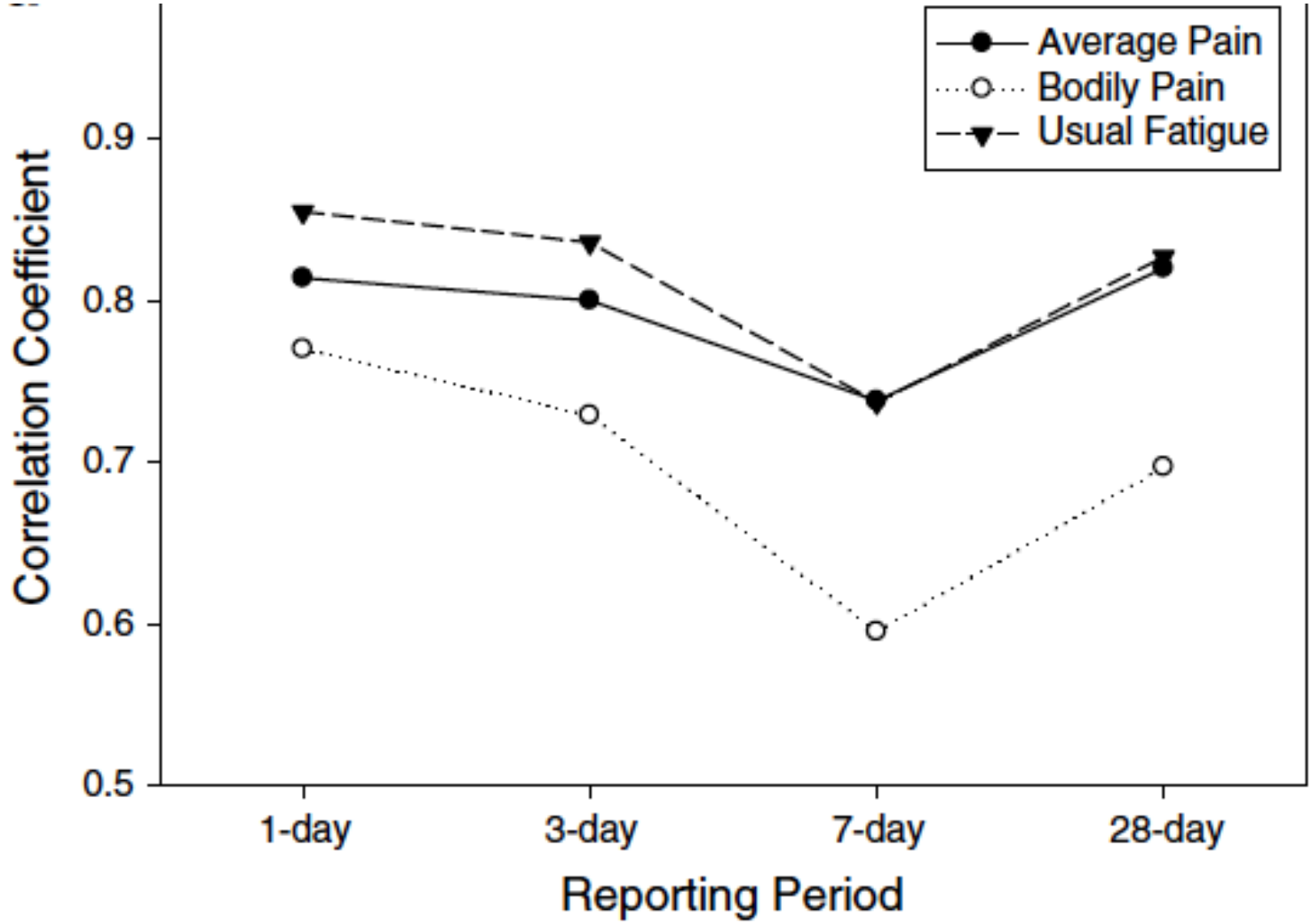
- Some studies have multiple assessments per day (EMA)
- Some studies have end-of-day assessments
- Rheumatology patients
- Pain, fatigue, sleep

Level Differences





Correspondence Differences



Recall of Specific Days

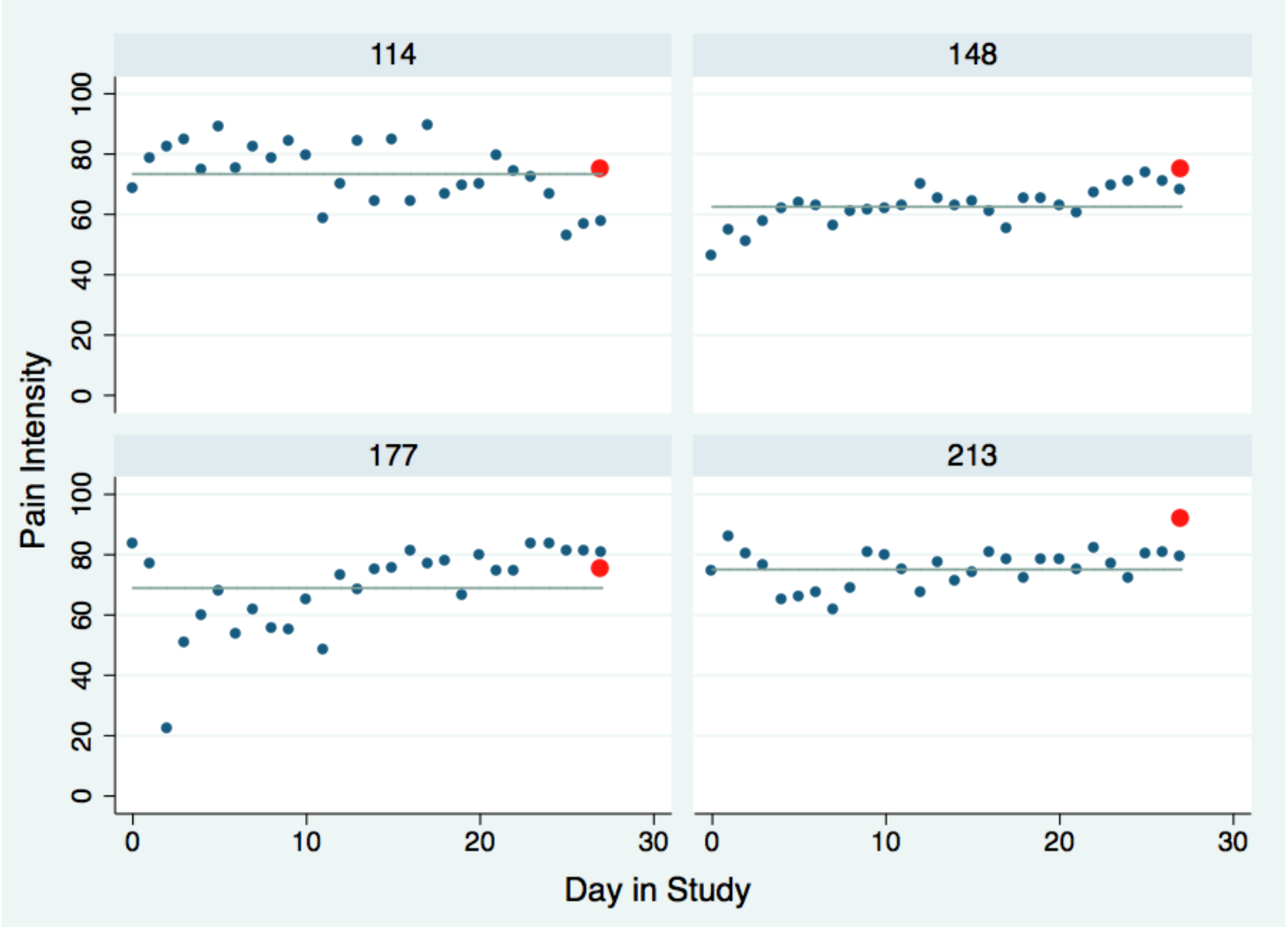
- Analytic Strategy
  - We also conducted a post 28-day recall task
    - Correspondence compares relative rank-ordering (correlation) of responses for Recall and Real-time across respondents
    - Prediction that shorter recall periods would be associated with higher correspondence

A Closer Look at 28-day Recall

- The 28-day average of momentary reports is 44.9
- The 28-day average of 28-recall is 62.2
- Difference is 17.3 points on 100-point scale
- Taking a closer look
  - For exposition, examining 28-day recall
  - Plot of daily average of momentary assessments
  - Defined a Match if Recall and Average daily assessments were close
  - Defined Non-match if there was a considerable difference, at least of 20 points on 100-point scale
  - IVR was rated with Verbal Descriptors
    - None, V Mild, Mild, Moderate, Severe, V Severe
  - Real-time pain with 0-100 VAS

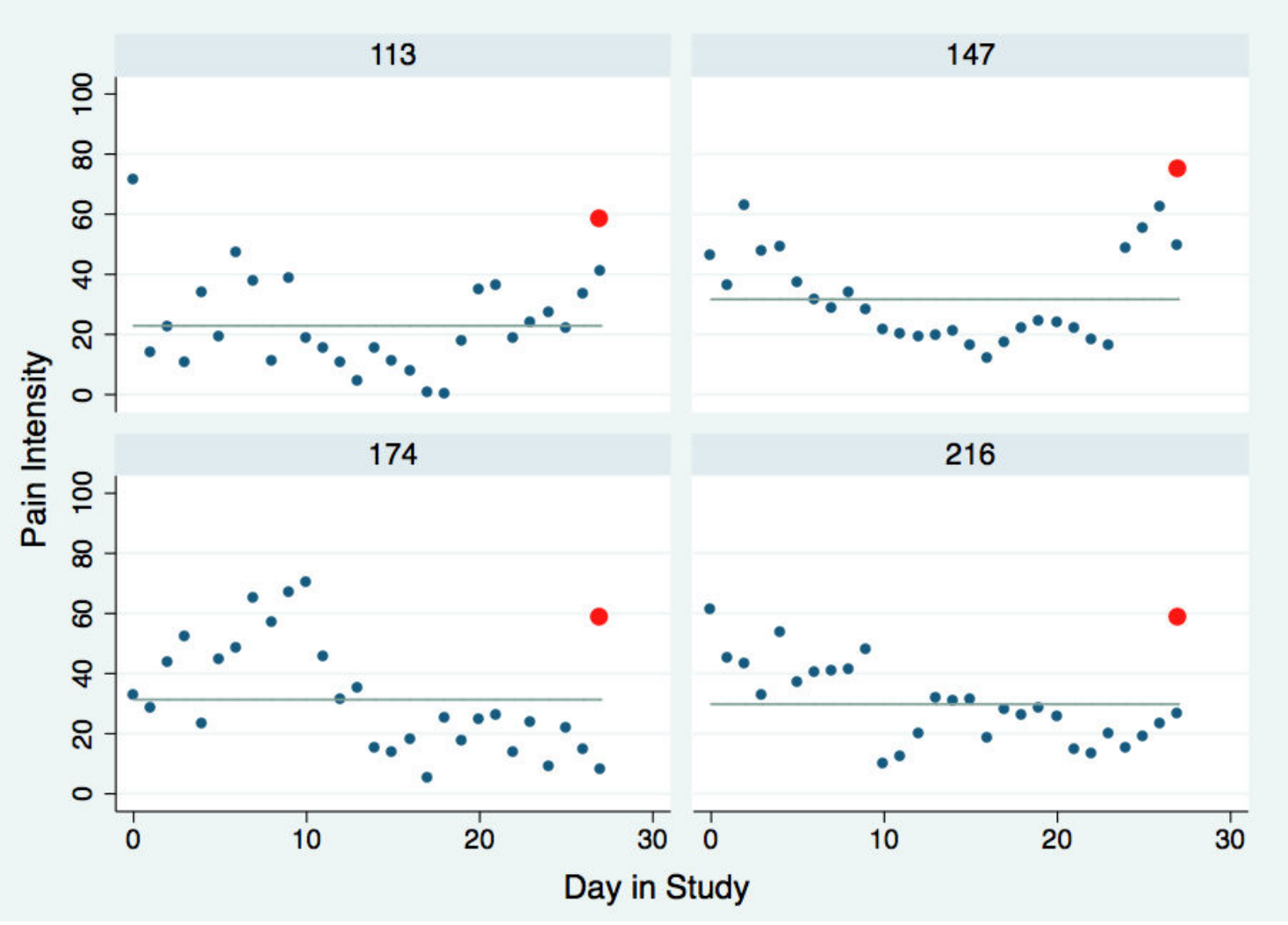
A Closer Look at 28-day Recall

- Examples of the Recall – Real-time Matches (59%)



A Closer Look at 28-day Recall

- Examples of the Recall–Real-time Non-matches (41%)



A Closer Look at 28-day Recall

- Those with close correspondence between Recall and Average Real-time:
  - Had *higher* mean levels of pain
  - Had *lower* day-to-day variability
- The variability result is consistent with our prior work showing greater discrepancies with higher variability

- Demographic differences:
  - Those with close correspondence:
    - » Low neuroticism score on NEO
    - » Less likely to be female
  - No age, income, marital differences

Discussion

- Using real-time as criterion, what is “good enough”?
  - Level differences
  - Correspondence differences

- There is *safety* in using a diary approach in terms of recall error and bias
  - But there are costs

Discussion

- Almost all research has examined Cross-sectional associations – We know little about CHANGE scores based on Recall versus Real-time

- How can respondents be yielding high correspondences *if they are not systematically reviewing and summarizing over the recall period?*

Alternatives to Systematic Review

- Given the answer speediness, respondents may be:
  - Guessing – but that would yield low correlations
  - Using their current symptom levels
  - Using their recent symptom levels
  - Using their beliefs about their symptom levels
- The last three options could yield fairly high correspondence
  - As the 28-day graphics showed, even correlations of .7 and .8 may have error that is unacceptable
  - However, at this point limited to pain and fatigue content

Alternatives to Systematic Review

- If this is the case, then we need to be quite cautious using “long” recall periods, since we are not measuring what we think we are measuring
  - Necessary to validate with real-time or other sources of information
  - Showing treatment effects is important, but does not fully validate the outcome

- This goes to the concept of “Construct Validity” and FDA’s view of “Content Validity” – that is, actually measuring what we think and say we are measuring

Identifying Recall Periods for Patient-Reported Outcomes: Is a Daily Diary the Best Approach?

Dennis A. Revicki, PhD
United BioSource Corporation, Bethesda, MD

SECOND ANNUAL
PATIENT-REPORTED OUTCOME (PRO) CONSORTIUM WORKSHOP

March 15, 2011 ■ Silver Spring, MD

Co-sponsored by

CRITICAL PATH INSTITUTE
collaborate · innovate · accelerate

FDA

Objectives

- Selecting recall periods: best practices
- Symptom versus functioning outcomes
- Evaluating different recall periods
- Recall period and treatment effects

Selecting Recall Periods: Best Practices

- Key concept: recall period needs to match the target PRO concept
- Daily recall/assessment may be best for symptoms (e.g., pain intensity, heartburn, etc.)
- Longer recall periods (1-4 weeks) may be best for measures of functioning or activities that do not occur daily (e.g., sexual dysfunction, physical functioning, emotional functioning, etc.)
- Shorter recall periods may be best for episodic events (e.g., heartburn episodes, headache pain) while longer recall periods for chronic concepts (e.g., sexual dysfunction, activity limitations)
- 7-day recall may be a good compromise between limits of recall and providing sufficient time for activities/behavior to occur
- Longer recall periods are associated with recall bias and increased measurement error

Selecting Recall Periods: Mixed Methods

- Qualitative Research
  - Focus groups/individual interviews with patients about their symptom or other problem experience
  - Variations in symptom/problem over time (daily, weekly, etc.)
  - Match recall period to reported variation in patient experience
- Quantitative Research
  - Evaluate empirically different recall periods
  - Examine correlation between different recall periods (i.e., daily versus weekly)
  - Evaluate reliability, validity, and responsiveness to change in clinical status
  - Select recall period that best fits concept and responsiveness

Recall Period and Treatment Effects

- Longer recall periods are associated with increased measurement error
- Increased noise (measurement error) will make it more difficult to detect a treatment effect (for a treatment that is effective)
- Increased noise will not make an ineffective treatment appear effective
- Treatment effects may vary by daily versus longer recall periods

Case Study: Daily versus 4-Week Recall of Sexual Desire in HSDD

- Qualitative research with pre- and post-menopausal women with HSDD demonstrated (Revicki et al. 2011):
  - Daily assessment of sexual desire was problematic
  - Women suggested longer recall periods (1-4 weeks)
  - Longer recall period allowed for sexual activity to occur
- Clinical trials of flibanserin endpoints
  - Sexually satisfying event (daily diary)
  - Sexual desire intensity (daily diary)
    - 0 (none) to 3 (severe) scale
  - FSFI sexual desire scale (4-week recall)
    - 0 to 5 intensity and frequency response scales (2 items)

Treatment Effects Comparing Daily Diary versus 4-Week Recall

- Study 71

Endpoints	Treatment Groups				Diff	P-Value
	Placebo		Flibanserin 100mg			
	N	Change	N	Change		
SSE	285	0.83	275	1.58	0.75	0.0047
Diary sexual desire	285	6.90	275	9.14	2.24	0.1320
FSFI sexual desire	290	0.55	280	0.90	0.35	0.0002

- Study 75

Endpoints	Treatment Groups				Diff	P-Value
	Placebo		Flibanserin 100mg			
	N	Change	N	Change		
SSE	381	1.11	371	1.86	0.75	0.0244
Diary sexual desire	381	6.77	371	8.48	1.71	0.3461
FSFI sexual desire	388	0.56	379	0.89	0.35	<0.0001

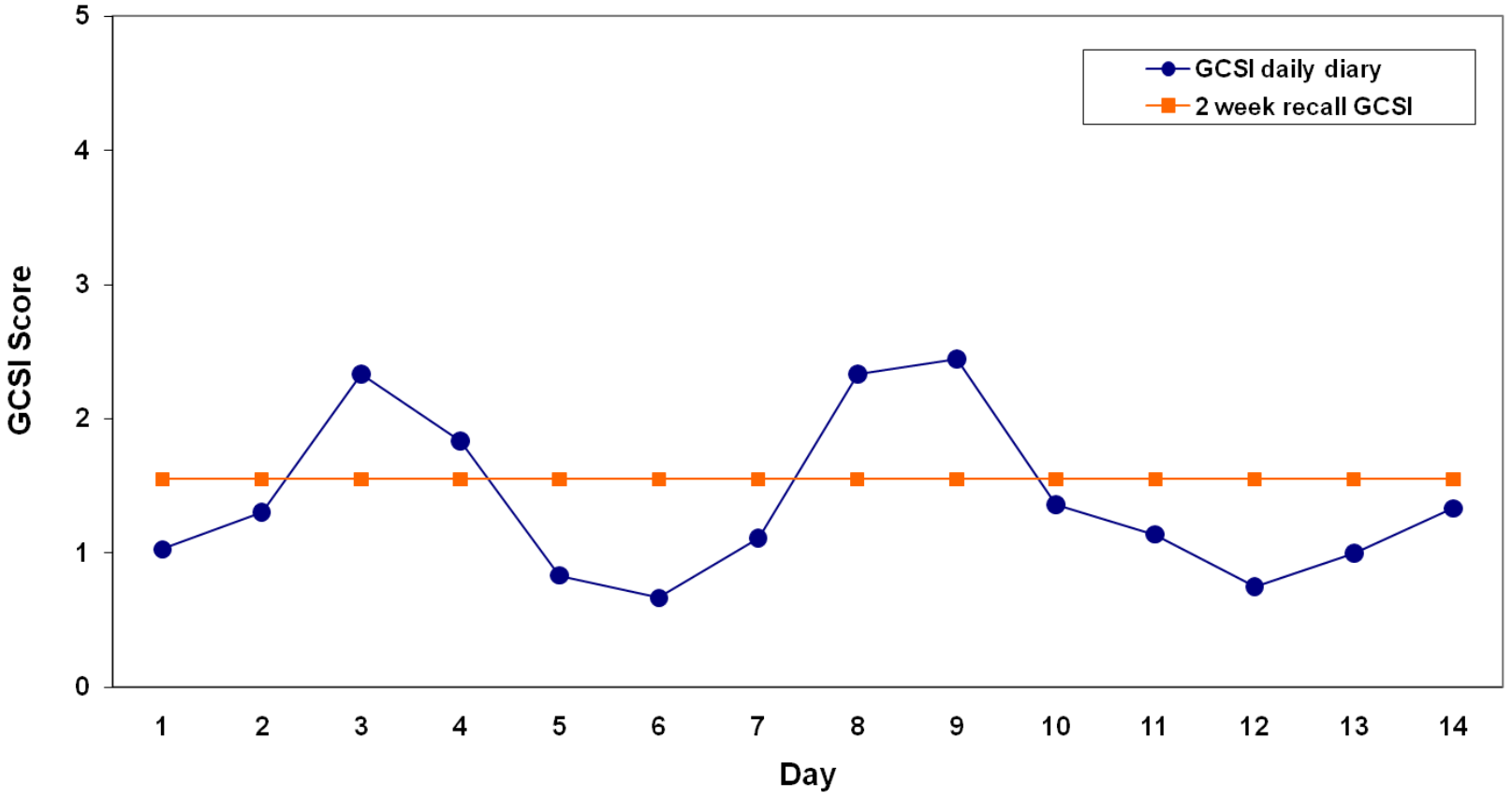
Sexual Desire:
Daily Diary versus Longer Recall

- Daily diary compliance decreased over study
  - Baseline: 75% completed 26/28 days and 96% completed 21/28 days
  - Weeks 21-24: 44% completed 26/28 days and 74% completed 21/28 days
- Qualitative supportive evidence for longer recall period (Revicki et al. 2011)
  - >90% thought daily assessment was not relevant
  - >90% thought 1-, 2-, or 4-week recall was best
- Qualitative and quantitative research in sexual dysfunction supports longer recall periods (K Weinfurt, personal communication, December 2010)
- Qualitative research in other areas of sexual dysfunction also supports longer recall period (1-2 weeks)

Case Study: Patient-Reported Symptoms in Gastroparesis

- Pilot study of the Gastroparesis Cardinal Symptom Index-Daily Diary (GCSI-DD) (Revicki et al. 2009)
- 12 gastroparesis patients followed for 2 weeks
- Compared daily diary versus 2-week recall versions

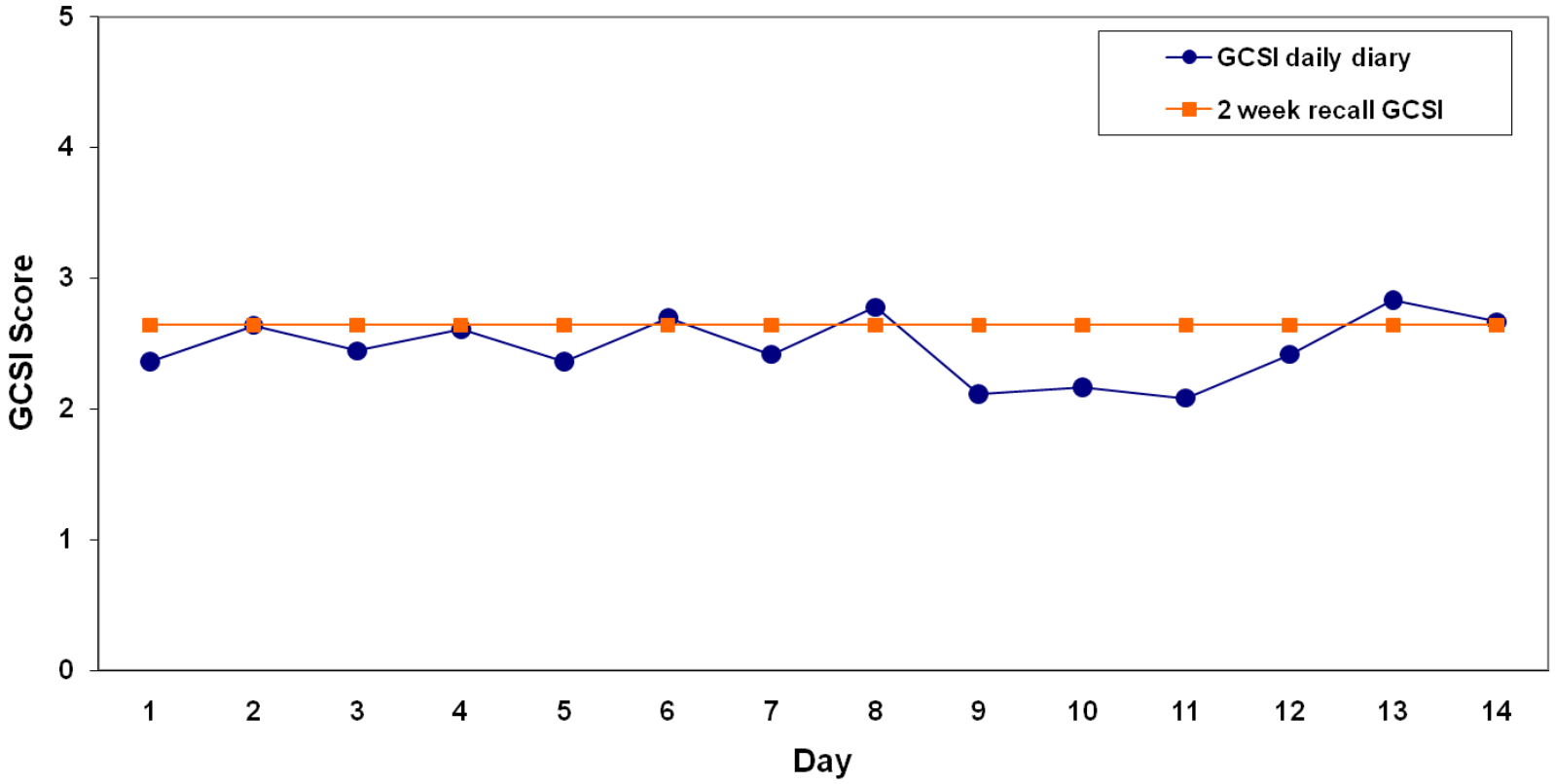
GCSI Daily Diary versus 2-Week Recall GCSI (Patient 6)



Symptom Variability: 6 patients

Parkman HP. Gastroparesis Clinical Endpoints: The ANMS PRO Initiative.

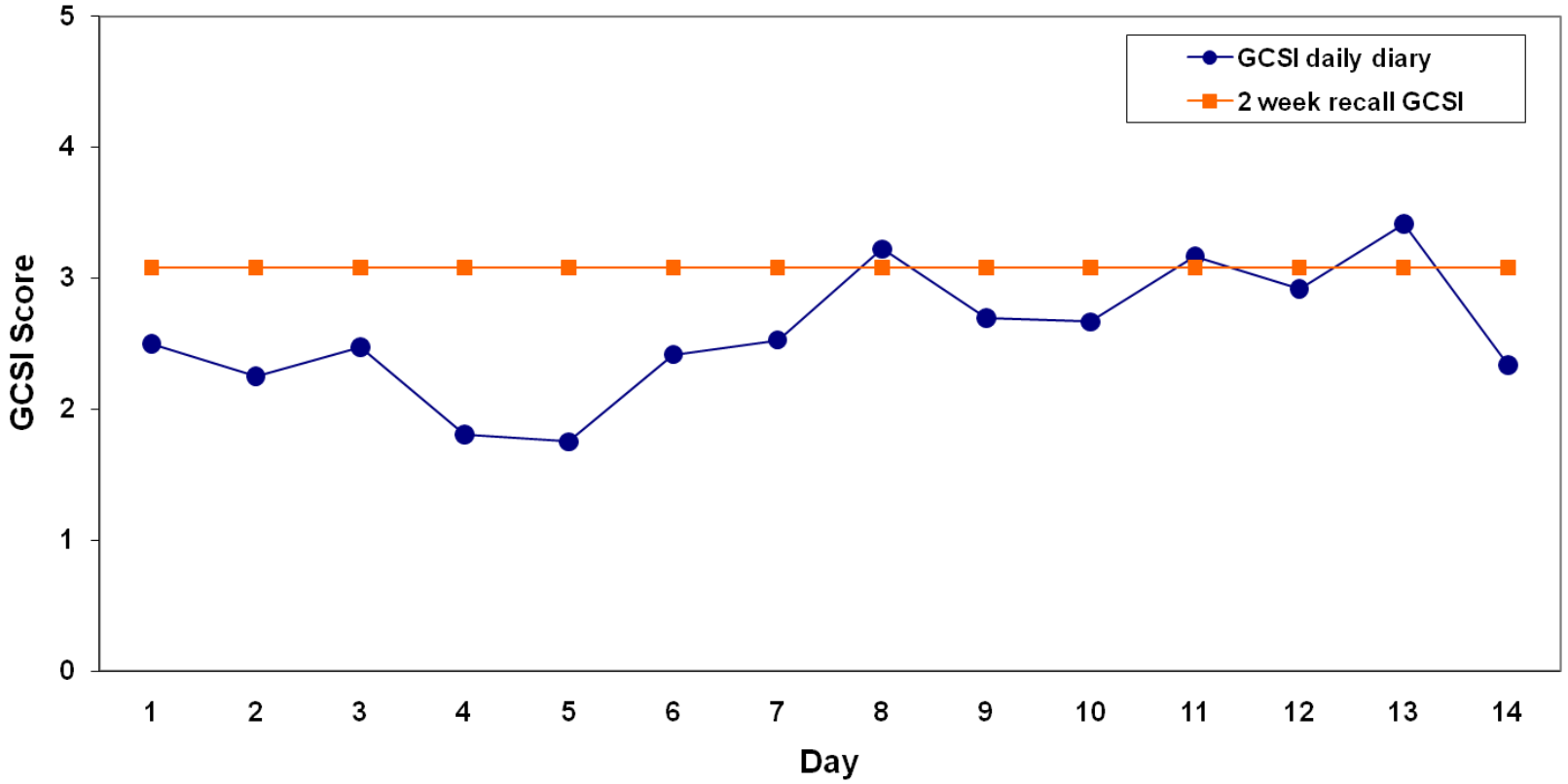
GCSI Daily Diary versus 2-Week Recall GCSI (Patient 4)



Relatively Constant Symptoms: 6 patients

Parkman HP. Gastroparesis Clinical Endpoints: The ANMS PRO Initiative.

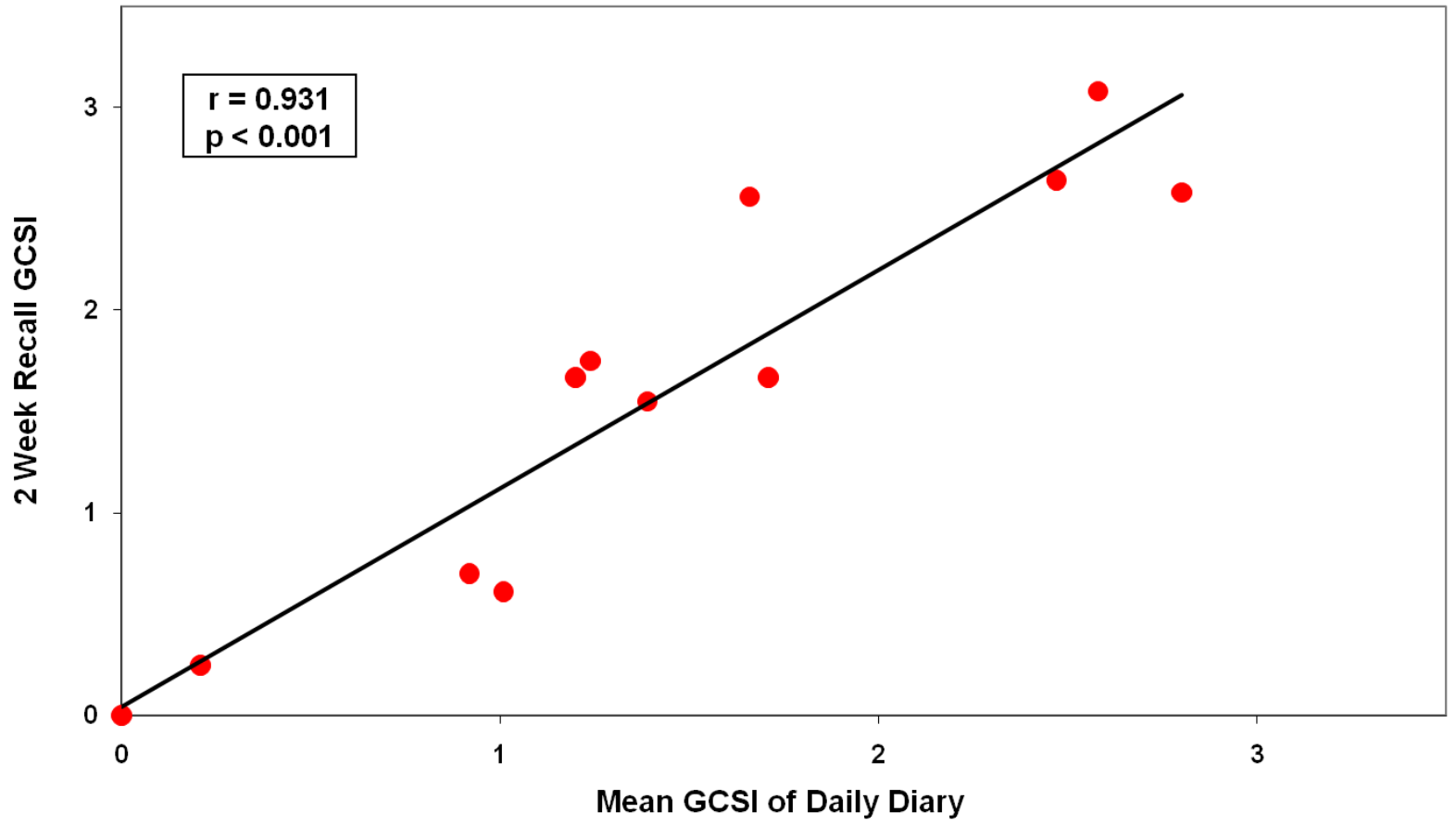
GCSI Daily Diary versus 2-Week Recall GCSI (Patient 3)



2-week recall GCSI may be influenced by recent symptoms: 3 patients

Parkman HP. Gastroparesis Clinical Endpoints: The ANMS PRO Initiative.

Correlation between Mean GCSI of Daily Diary and 2-Week Recall GCSI



Parkman HP. Gastroparesis Clinical Endpoints: The ANMS PRO Initiative.

Correlation between GCSI-D and GCSI at Visit 2

	GCSI (2-Week Recall)				
GCSI-D Average Days 1-14	**Nausea/ Vomiting**	**Fullness/ Early Satiety**	**Bloating**	**Total GCSI**	**Abdominal Pain/Discomfort**
Nausea/Vomiting	**0.96**	0.58	0.38	0.75	0.67
Fullness/Early Satiety	0.59	**0.91**	0.20	0.68	0.73
Bloating	0.51	0.32	**0.91**	0.76	0.70
Total GCSI	0.81	0.71	0.71	**0.93**	0.92
Abdominal Pain/Discomfort	0.66	0.78	0.45	0.83	**0.93**

Parkman HP. Gastroparesis Clinical Endpoints: The ANMS PRO Initiative.

Case Study: Patient-Reported Symptoms in Gastroparesis

- Psychometric evaluation study of the Gastroparesis Cardinal Symptom Index-Daily Diary (GCSI-DD) (Revicki et al. 2011)
- Four symptoms: nausea, bloating, excessive fullness, postprandial fullness
- Daily diary and 2-week recall versions
- Observational study of 62 gastroparesis patients starting new treatment and followed for 4 weeks

Comparison of Daily and 2-Week Recall GCSI Versions: Responsiveness at 4 Weeks

Score	N	Mean Change	P-Value	Effect Size
Nausea				
Daily Diary	25	-0.64	<0.0001	0.42
2-Week Recall	27	-1.41	<0.0001	1.00
Bloating				
Daily Diary	25	-0.62	<0.0001	0.34
2-Week Recall	27	-0.93	0.0006	0.50
Excessive Fullness				
Daily Diary	25	-1.08	<0.0001	0.83
2-Week Recall	27	-1.67	<0.0001	1.18
Postprandial Fullness				
Daily Diary	25	-0.83	<0.0001	0.53
2-Week Recall	27	-1.56	<0.0001	1.45
Summary Score				
Daily Diary	25	-0.63	<0.0001	0.61
2-Week Recall	27	-1.39	<0.0001	1.48

Responder based on patient-rated global change in overall gastroparesis symptoms

Summary

- Daily diaries and event assessments provide more accurate recall, but PROs with longer recall periods are often strongly correlated with daily assessments
- Cannot rely only on patient's perspective based on qualitative research
- In clinical studies, daily and longer assessments of the same PRO concepts may detect comparable treatment effects
  - Responsiveness may vary by concept assessed and recall period
  - Responsiveness also depends on item content and response scale

Summary

- Do we really need to worry about recall periods for PROs in clinical trials?
- For PROs, recall period should match concept assessed and variations in concept over time
- Longer recall periods increase measurement error and variability, which makes it more difficult to demonstrate a treatment effect (even for effective treatments)
- Psychometric evaluations comparing different recall periods are needed to determine whether recall period is “good enough” to capture concept of interest

Recall period: FDA Review Considerations

Elektra Papadopoulos, MD, MPH
Study Endpoints and Labeling Development
CDER/FDA

SECOND ANNUAL
PATIENT-REPORTED OUTCOME (PRO) CONSORTIUM WORKSHOP

March 15, 2011 ■ Silver Spring, MD

Co-sponsored by

CRITICAL PATH INSTITUTE
collaborate · innovate · accelerate

FDA

Disclaimer

This presentation represents the view of the presenter and not the FDA position.

Recall Period

- FDA guidance recommends that the recall period of an instrument should be adequate for the PRO instrument application
  - What do we mean by that?
  - Why are we concerned?

Recall Period: FDA Concerns

- Variability introduced by an inappropriate recall period may undermine accurate measure of the treatment effect
- Risks include
  - Difficulty in PRO data interpretation
  - Inability to accurately report findings in labeling

Content Validity and Recall Period

- ***Content validity*** — Evidence from qualitative research demonstrating that the instrument measures the concept of interest including evidence that the items and domains of an instrument are appropriate and comprehensive relative to its intended measurement concept, population, and use
- ***Recall period*** — The period of time patients are asked to consider in responding to a PRO item or question

Content Validity and Recall Period

- An instrument's recall period is a content validity consideration
- The time period considered by the respondent should reflect the reporting period that the instrument purports to measure

FDA PRO Guidance (2009)

PRO instruments that call for patients to rely on memory, especially if they must recall over a long period of time, compare their current state with an earlier period, or average their response over a period of time, are likely to undermine content validity.

Recall Period:
FDA Review Considerations (1)

- Population (e.g., children versus adults)
- Disease or condition
- Measurement concept (e.g., pain, itch, etc.)
- Symptom Characteristics (Frequency; Duration; Intensity; Variability; Saliency; Chronicity)

Recall Period: FDA Review Considerations (2)

- Aspect of the concept recalled
  - Average; Worst; How often; How intense
- Study design consideration examples
  - Time-to-improvement of symptoms would require more frequent assessments and shorter recall period
  - Heterogeneity in event frequency may suggest the use of an event log (e.g., sexual dysfunction)

Recall Period: Evidence Reviewed

- Concept measured
- Context of use
- Qualitative research
  - Concept elicitation interviews
  - Cognitive interviews
- Construct validity
  - Confirms what we learn through qualitative research
  - Will not replace or rectify evidence obtained through qualitative research

Summary

- Recall period is a content validity consideration
  - Important for interpretation and labeling
  - Should be appropriate for the concept measured, patient population and general context of use
  - Should reflect the reporting period that patients actually consider (qualitative research)
- No single recall period is suitable for all uses