Interpretation of Human Abuse Potential Studies and Clinically Important Responses to ADFs

Kerri A Schoedel, PhD
Director and Principal
Altreos Research Partners, Inc.

Abuse Deterrent Formulation Science Meeting
Sep 30-Oct 1, 2013

Disclosure

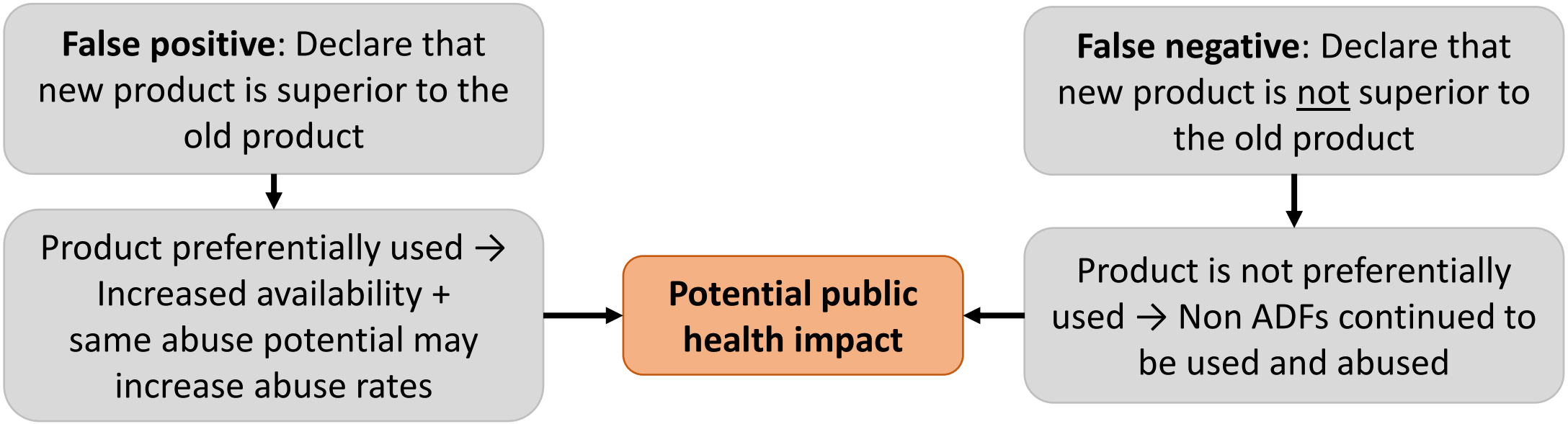
- *Consultant to pharmaceutical and biotech companies*

An Evolution in Study Design

- NCE human abuse potential studies: 2 primary purposes
  - **1) Rule out potential for abuse (unscheduled vs. scheduled – PBO vs Test)**
  - 2) Aid in scheduling placement (*relative* abuse potential – Test vs. Control)
- Design of human abuse potential studies have evolved to answer Question 1 really well
  - Studies are *well designed* to rule out **false negative** → Low Type II error
    - Relatively large sample size (N=30-40)
    - Sensitive population + qualification/enrichment
    - Multiple, redundant highly sensitive measures
  - At expense of higher potential for false positives → Higher Type I error
    - Acceptable based on public health risk

ADF Study Implications

- ADF study design co-opted from NCE approach but has one primary purpose
  → Determine ***relative*** abuse potential to aid in ***product labeling***
- **Current** potential public health implications of ADF study results:



- Current situation (new ADF vs. old non-ADF): Studies may require a more balanced approach between Type I and Type II error → The only way to minimize both types of error:
  - **Improve the test**
  - Increase the sample size

“Improving the test”: Use of standardized instruments

- Use of bipolar Drug Liking (“at this moment”) scale as primary endpoint is justified
  - Used for many years, high face validity and relevance to concept of abuse (Griffiths et al., 2003)
  - Sensitive, specific, reliable, good construct validity as shown by meta-analysis data (19 studies; CPDD 2012a; 2012b)
- Overall Drug Liking and willingness to Take Drug Again
  - Performance characteristics similar to Drug Liking → Slightly less sensitive but more specificity
  - May be ideal for a “balanced” approach to Type I and II error
- Use of measures of “High” (historically also ARCI MBG or drug value/choice)
  - Unacceptable performance characteristics, including higher variability and poor construct validity (convergent and discriminative)
    - As shown by Factor Analysis (CPDD 2013): **redundant to** other measures with better measurement properties (i.e., Drug Liking, Overall Drug Liking/Take Drug Again)
  - “High” is ambiguous and based on drug abuser vernacular → likely to change over time

Other Measures in Guidance

POMS?

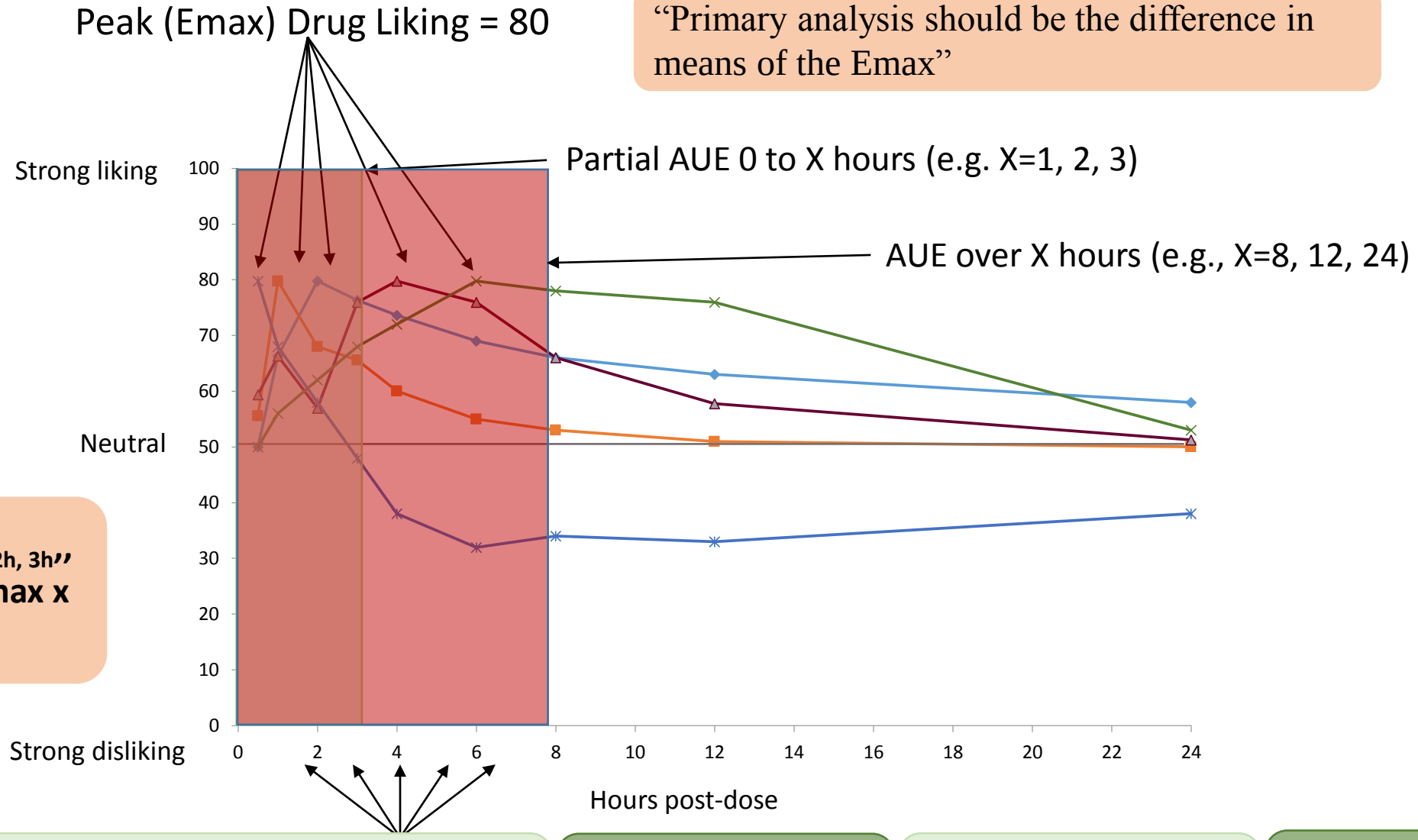
- Assesses mood states with 6 derived **Scales**:
  - Anger-Hostility; Confusion-Bewilderment; Depression-Dejection; Fatigue-Inertia; Tension-Anxiety; Vigor-Activity; Friendliness
  - Which scale is considered predictive?
  - No literature data to suggest that POMS is predictive of abuse of opioids or any other drug class

Subject-Rated Nasal Scales

- Not systematically validated/evaluated but probably useful for interpretation
  - More directly relevant to testing hypotheses around ADF features

… attention should be given to the profile of subjective effects … in terms of **onset, peak, duration of activity**, and **offset.**

"Primary analysis should be the difference in means of the Emax"



Emax , $AUE_{0\text{-}1h,\ 2h,\ 3h}$,, AUE, Emax/TEmax x AUE…???

Preferable to rely on *subject's* rating of the overall experience (i.e., end-of-day/next-day measures)

Overall Conclusions

Emax or partial AUE + TEmax (+ AUE)

Interpretation (the Why?)

Future Directions for the ADF Measures

- Common questions in interpreting scale responses:
  - How important is onset vs. duration vs. peak?
  - Do the bad/aversive effects actually outweigh the good?
  - WHY???
    - Why does the subject want to take the drug again? (Abuse? Relief of withdrawal? Diversion?)
    - What specifically does the subject like or not like about the drug?
  - Value of open-ended feedback or at least follow-up to some questions
    - Approach is not “validated” but may help relieve some of the burden of (mis)interpretation from Sponsors and the Agency
      - Do the intended/hypothesized effects of the ADF match what subjects are actually saying?

Bipolar End of Day/Next-Day Measures (e.g, Overall Drug Liking)

Interpretation of Results in Context of Statistical Approach

"Substantial" decreases in the responses for the potentially ADF compared to the positive control are evidence of deterrence.

H_0: $\mu_{Control} - \mu_{Test} \leq \delta_1$ vs H_a: $\mu_{Control} - \mu_{Test} > \delta_1$

$\boldsymbol{\delta_1 > 0}$

Validation test (between PBO and Comparator) also needs a margin (δ_2)

H_0: $\mu_{Control} - \mu_{Placebo} \leq \delta_2$ vs H_a: $\mu_{Control} - \mu_{Placebo} > \delta_2$

$\boldsymbol{\delta_2 > 0}$

…the sponsor **should review the literature** and consult with appropriate experts, and then **propose the values of δ_1 and δ_2 to the FDA**

Data currently available in the Public Domain to Determine CID δ_1 and δ_2

	Drug Liking VAS $\mu_c - \mu_T =$	95% CI of LS Mean Difference Available?	Tier 3 Labeling	Post-Market Data Available?
IR Opioid (oral; High dose) vs. Placebo (δ_2) (Multiple sources; meta-analysis)	35-40	✓	N/A	✓
OXECTA® (intranasal) (Schoedel et al., 2012)	22.7	✗	✗ ✓	✗
EMBEDA® (intranasal) (Setnik et al., 2013)	15.6	✗	✗	✗
EMBEDA® (oral crushed) (Stauffer et al., 2011)	21.4	✗	✗ ✓	✗
OxyContin® (intranasal) (product label)	13.6	✗	✓	✓

Super Superiority?

- Type of hypothesis testing proposed sometimes referred to as "super superiority", i.e., statistically significant result implies both statistical significance and clinical relevance.
- ***However:***
  - Drug Liking and other subjective scales are proxy measures not direct outcomes
    - i.e., cannot ask prescribers or abusers what "magnitude" would be clinically meaningful to them
    - Many different perspectives on how to anchor clinical significance (Deaths? Addiction? Intoxication?)
  - Super superiority approach based on 1 primary endpoint contradicts other parts of guidance:

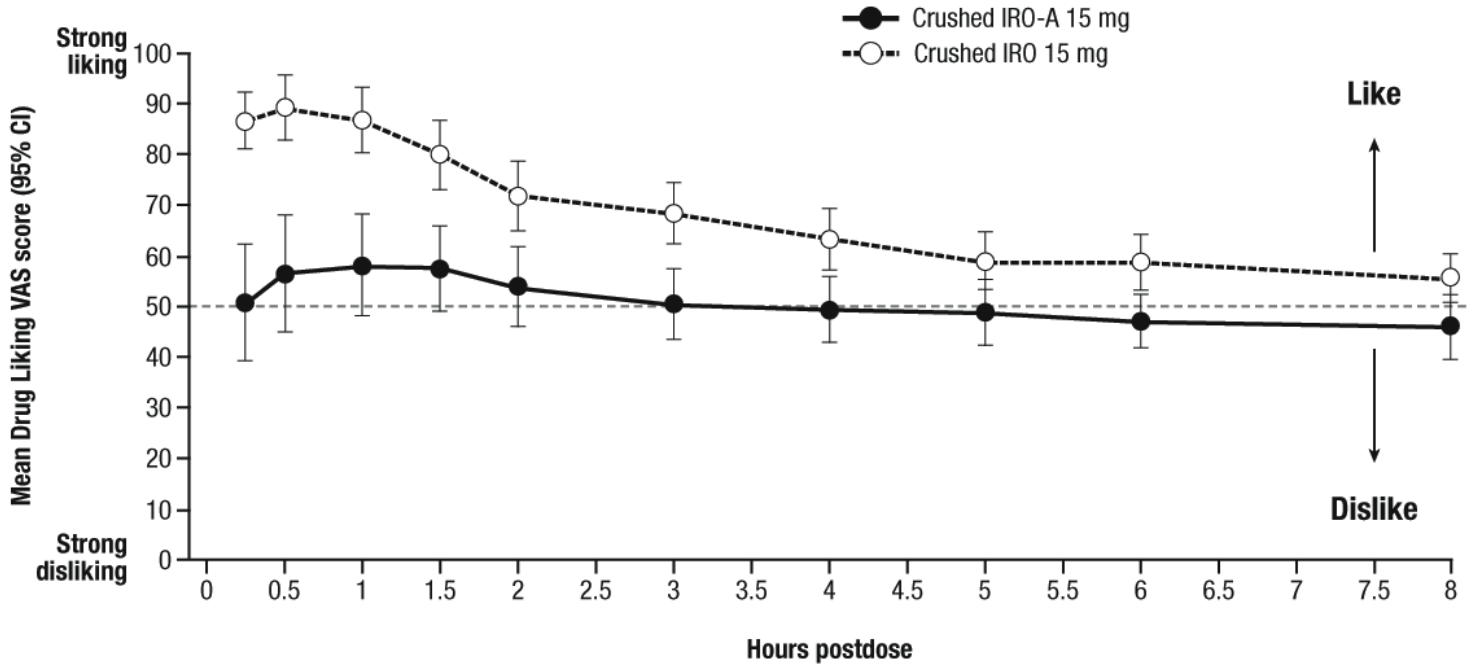
> **Importance of rate of onset, duration**, etc. + other effects (negative)
> The overall assessment of abuse potential … based on the **pattern of findings across all measures**
> What constitutes a clinically significant difference in drug liking…**area requiring further research**

- Insufficient data available to Sponsors to determine a predictive clinically meaningful margin (especially δ_1) or to support a super superiority design **at this time**
  - With potentially inaccurate estimate of margins, high risk of making a Type I (if margin is too low) or Type II error (if margin is high)
    - Potential public health consequences

Interpretation of Data

Useful graphs include mean time course profiles, heat-maps, and continuous responder profiles.

Time course profiles

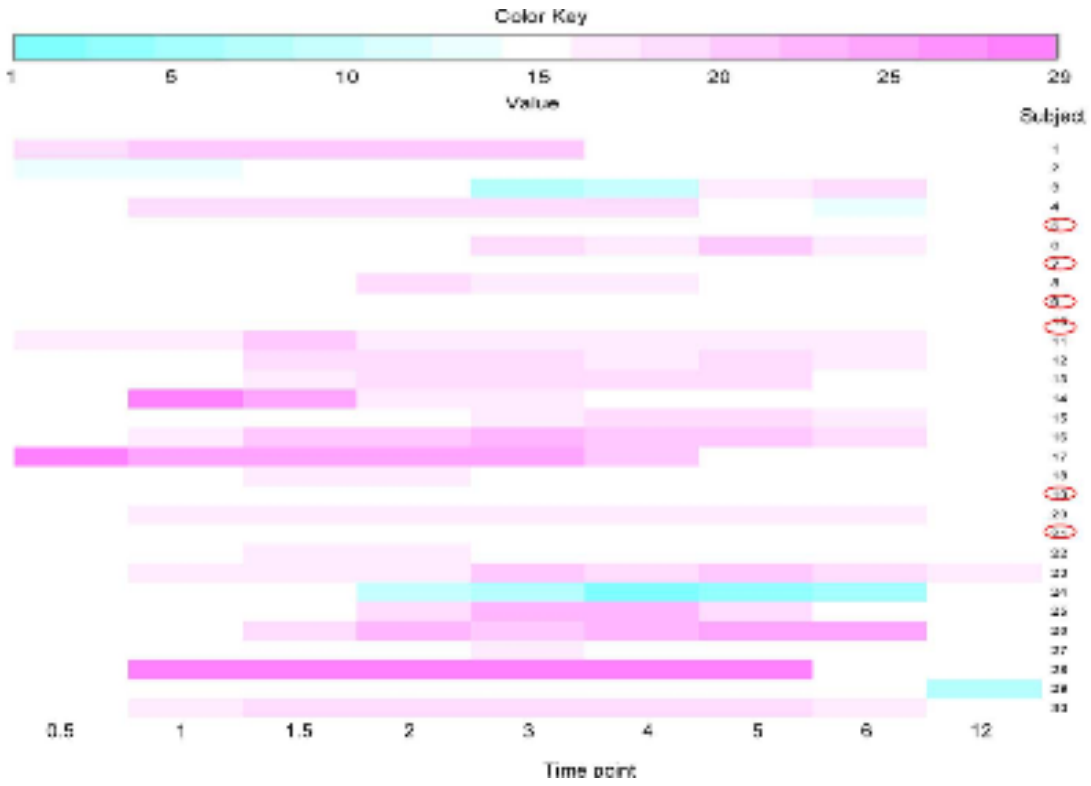


Useful

Visual evaluation of onset, duration and offset
Used for many years in labels to show other pharmacologic data (i.e., PK)

Schoedel et al., 2012

Heat Maps

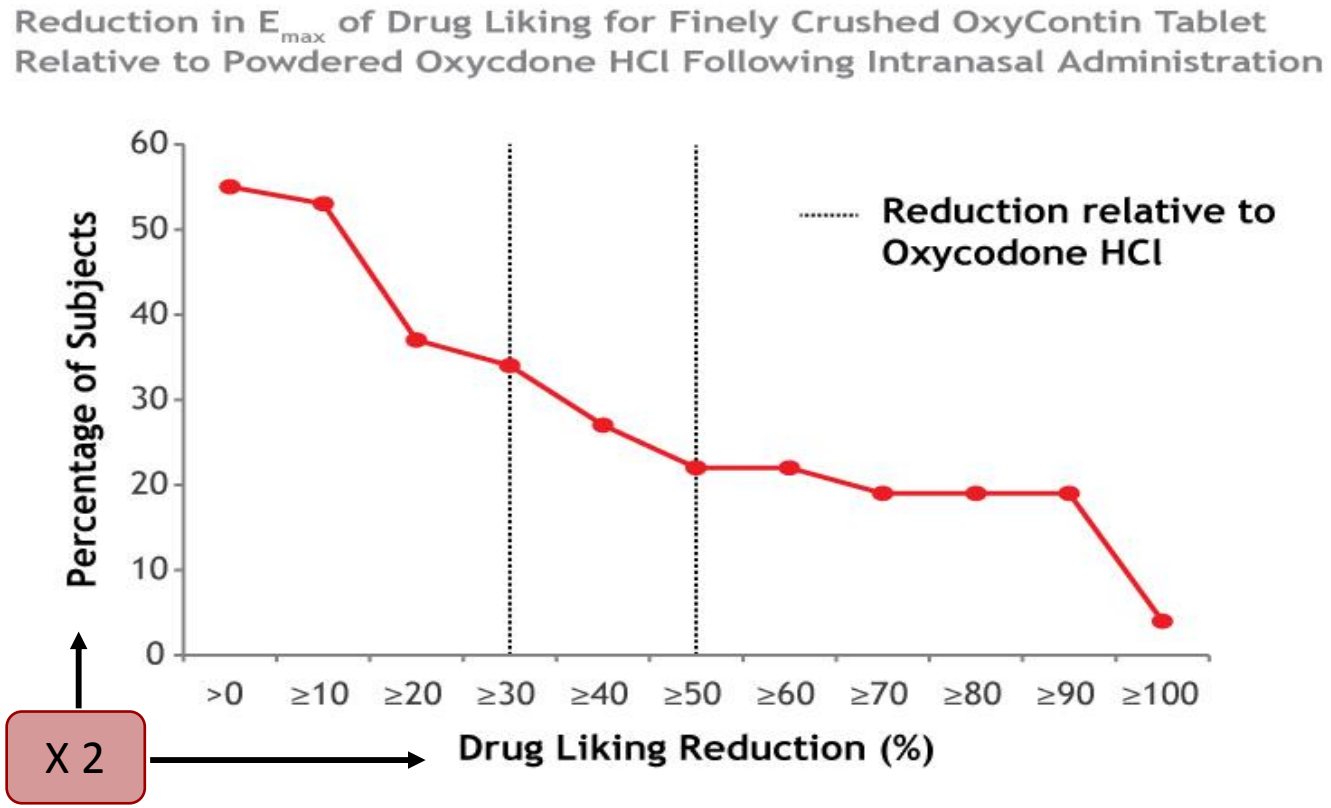


Somewhat Useful

Visual evaluation of individual responses
How do these data inform conclusions?
May be difficult for end-user (prescriber) to interpret in label

Chen & Wang., 2012

Continuous Responder Profiles



Predictive validity?
~53% subjects ≥10% reduction in Drug Liking
71% reduction in actual IN (or IV) abuse??

Arbitrary margins?

Difficult to interpret >
Consider end user (prescriber)

Useful?

OxyContin® Label

Summary of Interpretation and CID

- The Bad News:
  - Insufficient data to determine appropriate δ_1 super-superiority margin
    - Determining appropriate margins would require a consensus approach
    - But, by the time sufficient data generated, context may change (i.e., non-inferiority to existing ADFs)
- The Good News:
  - Many of the scales used are inherently "interpretable"
    - High face validity of "liking" and "take drug again"
    - Subject can tell you when they feel "neutral" or dislike a product (bipolar scales)
    - Subjects can synthesize all factors (magnitude, onset, duration, negative, etc.) and directly tell you what they think and how they will behave (end of session measures)
  - We have a placebo comparator → can perform analysis of $\mu\text{Test} - \mu\text{Placebo}$
  - We could use additional controls (e.g., tramadol? Oral arm in nasal study/Intact arm in oral study)
  - Studies have been successfully used to make higher stakes decisions for decades (relative scheduling)
- Other considerations for interpretability
  - Remember conceptual framework: flexible, adaptive approach and public health implications
  - Consider the purpose of ADF studies (label) and end users (prescribers)