

# Criteria for Gauging Response to Sodium Oxybate for Narcolepsy

Alana D. Steffen<sup>1</sup>, PhD; Chinglin Lai<sup>2</sup>, PhD; Terri E. Weaver<sup>1</sup>, PhD

<sup>1</sup>College of Nursing, University of Illinois at Chicago

<sup>2</sup>Jazz Pharmaceuticals

## **Institution(s) at which the work was performed:**

University of Illinois at Chicago

## **Financial Support** (Disclosure of the presence OR absence of financial support)

The clinical trials were funded by Jazz Pharmaceuticals, Inc. This manuscript reports a secondary analysis of the clinical trials data. There was no financial support for analyses and manuscript preparation beyond usual salary from the authors' respective institutions.

## **Conflicts of interest, and off-label or investigational use:**

Dr. Lai was an employee of Jazz Pharmaceuticals, Inc. who, in the course of this employment, received stock options exercisable for, and other stock awards of, ordinary shares of Jazz Pharmaceuticals. He is now retired. Jazz Pharmaceuticals pays Dr. Weaver a royalty fee for the use of the Functional Outcomes of Sleep Questionnaire in other studies. This secondary data analysis involved data from clinical trials of an investigational drug.

## **Corresponding author's full address, phone and fax numbers and e-mail address:**

Alana D. Steffen  
845 S. Damen Avenue (MC 802), Room 616  
Chicago, Illinois 60612-7350  
Office phone: (312) 413-5632  
Email: steffena@uic.edu

Running head: Gauging response to sodium oxybate

Word count: 3801

Number of references: 19

Contribution of authors: Dr. Steffen conducted the analyses and wrote the manuscript. Dr. Lai contributed to the analysis plan, provided background for use of these data, and critically reviewed drafts of the work. Dr. Weaver contributed to the conceptualization of the paper, contributed to the writing, and provided critical review of the work.

## Abstract

Our objective was to define responder criteria using an anchor-based approach for frequency of cataplexy attacks and excessive daytime sleepiness in narcolepsy patients undergoing sodium oxybate treatment. We used pooled data from two randomized, placebo-controlled, double blind, multi-center 4-week and 8-week trials of sodium oxybate for narcolepsy with cataplexy and analyzed using receiver operator characteristics analysis. Percent change in frequency of weekly cataplexy attacks and the Epworth Sleepiness Scale outcomes were compared to Clinical Global Impression of Change ratings, used as the anchor to define true response. Participants (n=336) were 39% male, 89% white, with a mean age of 41.5 (15.3), reporting a median of 20.5 cataplexy attacks per week and a mean Epworth Sleepiness score of 17.5 at baseline. A majority (51%) were much or very much improved based on Clinical Global Impression of Change ratings, considered a true response to treatment. Area under the curve values for % reduction in cataplexy attacks (77%) and % change in sleepiness score (78%) supported response definition thresholds of 46% and 12%, respectively. Classification using either response definition agreed with the anchor for approximately 71% of participants. Cataplexy response definition was more sensitive (Cataplexy=0.77, ESS=0.69) while sleepiness was more specific (Cataplexy=0.66, ESS=0.75). Both responder definitions showed a dose response relationship with sodium oxybate demonstrating their validity using an external criterion. Weekly cataplexy attacks and ESS can be used to help document clinical response to narcolepsy treatment using criteria of 46% and 12% reductions, respectively.

Keywords: Patient reported outcome, responder analysis, clinical trials.

## Introduction

Narcolepsy patients experience debilitating symptoms including excessive daytime sleepiness and cataplexy attacks. The former is commonly measured via self-report using the Epworth Sleepiness Scale (ESS) and the latter may be tracked daily by use of a cataplexy diary. Patients' report of symptoms and their improvement are increasingly recognized as important for clinical trials that establish the efficacy of new drugs. However, it is not always clear what constitutes a clinically relevant response to treatment in research or clinical settings. There has been evolving discourse on the best method for defining meaningful change. While clinical trials are often analyzed by comparing group means from a treatment versus control group, the difference in change between groups is not necessarily the relevant level of change for an individual.

In 2009 the Food and Drug Administration (FDA) published a guidance to industry for using patient reported outcomes (PROs) in clinical trials that test drugs and medical devices for efficacy (US Department of Health and Human Services, 2009) and others have also contributed to this discussion (Cappelleri and Bushmakina, 2014, Wyrwich et al., 2013, McLeod et al., 2011, Snapinn and Jiang, 2007). The FDA emphasized a need for each PRO to have a corresponding responder definition (RD), that is, an empirically determined threshold that identifies individual-level meaningful change over a specified time period. Such RDs should be established *a priori* using anchor-based, empirical methods for use in clinical trials to describe the proportions of individuals within the treatment conditions showing a meaningful response. RD criteria may also be useful to clinicians to document treatment efficacy, inform dose titration, and help counsel patients regarding symptom management and safety concerns. To our knowledge, there have been no published RDs for narcolepsy PROs.

Our goal in this analysis was to use empirical anchor-based methods to establish responder definitions (RDs) for excessive daytime sleepiness and frequency of cataplexy attacks using pooled data from two Phase III randomized double blind controlled trials of sodium oxybate for the treatment of

narcolepsy. We assessed if either PRO was more reliably associated with an anchor measure of clinical improvement, the Clinical Global Impression of Change (CGI-c), and if the RDs based on these PROs show agreement with each other. Finally, we validated the PRO RDs by showing that the proportions of responders based on the PRO RDs differ by dose of sodium oxybate.

## Methods

This secondary analysis used data from two double-blinded randomized controlled trials testing sodium oxybate at a range of doses versus placebo to assess potential RD thresholds for symptoms of excessive daytime sleepiness and frequency of cataplexy attacks. These multicenter trials were conducted at 42 sites from 1997 to 1998 and 2000 to 2004 and have been described in detail elsewhere. Briefly, participants in the two trials were 16 years old or older, had a positive history of narcolepsy with current symptoms of excessive daytime sleepiness, cataplexy, and recurrent sleep attacks. Reflecting real world clinical practice, those taking stable doses of stimulants, approximately 80%, were permitted to continue the medication during their trials. Hypnotic and anticataplectic medications were gradually discontinued with an additional washout period. In the 1997-1998 4-week trial (n=136) the total nightly doses investigated were 3, 6, and 9 g. In the 2000-2004 8-week trial (n=228) the total nightly doses were 4.5, 6, and 9 g, with the first four weeks incorporating a titration procedure for the two higher dose groups. Participants who completed the measures of excessive daytime sleepiness or frequency of cataplexy attacks at baseline and endpoint who were also assessed by physicians at endpoint are included in these analyses (n=322-327; 88.5-89.8% of the enrolled participants depending on the outcome).

## Measures

The Clinical Global Impression of Severity (CGI-s) and Clinical Global Impression of Change (CGI-c) measures were rated by the blinded clinician at baseline and at the study endpoint, respectively. The response options for the CGI-s ranged from normal to extremely ill. The CGI-c rating, made at the study

endpoint, referenced the baseline CGI-s rating, and used response options of Very Much Worse, Much Worse, Minimally Worse, No Change, Minimally Improved, Much Improved, Very Much Improved.

The CGI-c ratings of Much Improved and Very Much Improved were considered a favorable response to treatment while the ratings of Minimally Improved, No Change, Very Much Worse, Much Worse, or Minimally Worse were considered a non-favorable response. The dichotomized version of the CGI-c served as the anchor measure of treatment response for assessing the responder definition thresholds for excessive daytime sleepiness and reduction in the frequency of cataplexy attacks.

The Epworth Sleepiness Scale (ESS)(Johns, 1991) was used to measure excessive daytime sleepiness. Conceptually, this self-administered scale measures sleep propensity by asking individuals to rate their likelihood of falling asleep in 8 soporific situations, shows acceptable internal consistency, and has been used in clinical trials with Narcolepsy patients detecting a range of response to treatment effect sizes (Weaver, 2001). Previous research has suggested a score of 11 or more has sensitivity of 93.5% and specificity of 100% for distinguishing excessive daytime sleepiness from normal daytime sleepiness (Johns, 2000). Coefficient alpha was 0.76 for the baseline administration using the current data.

Daily patient diaries included detailed descriptions of the symptoms to be recorded including frequency of cataplexy events, nocturnal awakenings, total sleep time, hypnagogic hallucinations, and sleep paralysis as well as any adverse events. Participants were trained to record symptoms if they occurred and were instructed to make diary entries every morning and evening. Cataplexy attacks were defined as having sudden onset, precipitated by emotion, localizable to a specific part of the body, and did not occur during a sleep attack (i.e., patient remained lucid and aware) (Xyrem<sup>(R)</sup> International Study Group, 2005). Baseline and endpoint frequencies of cataplexy attacks were based on 2 weeks of diary entries and calculated as number of cataplexy attacks per week. The intraclass correlation (ICC) of daily attacks was 0.739.

## Statistical Analyses

A preliminary step was to consider the formula for calculating our outcomes, for example, simple change scores or % improvement from baseline. The distributions of baseline and follow-up measures of our two outcomes, ESS and cataplexy frequency, as well as their change scores, were examined. ESS scores were limited in range by the nature of the measure with a possible range of 0 to 24 and the distributions were approximately normal. However, the frequency of cataplexy had great variability with a range of 0 to 912 cataplexy events per week; an extreme range in the difference between baseline and endpoint was also evident (-310 to 97). Due to this variability, we chose to compute change scores as the % change from the baseline measure for both the outcomes. A second preliminary task was to assess if the correlations between the anchor measures and outcomes were substantial and similar across the two studies contributing to the pooled dataset. We examined if the Spearman correlations between the ordinal version of CGI-c, our anchor measure, and each outcome were significantly different between the 4- and 8-week trials using Fisher's z transformation.(Snedecor, 1989) We described the sample and compared responders and non-responders, as defined by the dichotomized anchor, CGI-c, using t-tests and chi-squared statistics.

Our primary approach for identifying responder definition (RD) thresholds was an anchor based receiver operator characteristic (ROC) analysis. The dichotomized anchor measure of response, the CGI-c, was regressed onto the percent change scores for each measure in separate logistic models with the ROC graph plotting sensitivity, the percent classified as improved out of the total improved, versus 1-specificity, the false positive rate, for each value in the range observed. We compared the Liu, Youden, and (0,1) optimization criteria for choosing cutpoints that maximized the joint sensitivity and specificity for identifying responders based on the product of sensitivity and specificity, their sum, and the nearest distance to perfect sensitivity and specificity, respectively (Fluss, 2005, Liu, 2012, Youden, 1950, Clayton, 2013). We compared the ESS and cataplexy frequency percent change from baseline to assess if either

outcome was more or less associated with CGI-c. The area under the curve (AUC) statistics were calculated for the ESS and cataplexy frequency percent change scores, along with 95% confidence intervals, and were tested for a statistically significant difference using a nonparametric approach for measures assessed on the same sample (DeLong, 1988). In the results, we present agreement and kappa statistics based on our selected cutoffs as well as sensitivity and specificity.

Distribution methods were used as a secondary approach to identifying responder thresholds. We used the baseline standard deviations (SD) and calculated 0.5 SD as well as a standard error of measurement,  $SEM = SD\sqrt{1 - \alpha}$ , where alpha represents a reliability estimate (McLeod et al., 2011). The SD for cataplexy frequency is calculated excluding the extreme upper and lower 2.5%, due to the extreme outliers.

Finally, we include PRO RD results by treatment dose as validation of our approach. In preparation, we compared the outcomes for groups receiving low total nightly doses, 3 g (used only in the 4-week trial) and 4.5 g (used only in the 8-week trial) using t-tests [ESS % change:  $t(87)=0.91$ ,  $p=0.3677$ ; Cataplexy % change:  $t(89)=-0.19$ ,  $p=0.8519$ ] and thus combined these lowest dose groups. Other doses were consistent across the 4- and 8-week trials. Cumulative distribution plots of each outcome by the resulting four dose groups are shown to display all possible RD thresholds as well our selected RD thresholds indicated with a vertical line. We tested the RD outcomes by treatment groups using logistic regression testing for dose trend as well as the four levels of dose using placebo as the reference group. All analyses were performed with Stata, version 12.1 (StataCorp, 2011).

## Ethics

The studies contributing data to these analyses were conducted in eight countries in accordance with the Helsinki Declaration of 1975, revised in 1997. Each participating trial centers' institutional review board approved the studies and participants gave written informed consent (International, 2005, The U.S. Xyrem<sup>(R)</sup> Multicenter Study Group, 2002).

## Results

We began by computing the Spearman correlations of our anchor measure of response, the CGI-c, with our PROs, percent change measures of ESS and frequency of cataplexy attacks, and found moderate associations that were not significantly different across the 4- and 8- week trials (ESS 4-week  $\rho = 0.57$ , 8-week  $\rho = 0.52$ , Fisher's  $z = 0.61$ ,  $p = .5419$  test comparing difference between correlations; Cataplexy 4-week  $\rho = 0.61$ , 8-week  $\rho = 0.50$ , Fisher's  $z = 1.39$ ,  $p = .1645$ ). These moderate correlations support the CGI-c as an appropriate anchor. Data were pooled because the relationships of the anchor (CGI-c) with the ESS and cataplexy outcomes were similar over the two trials (Spearman  $\rho=0.53$ ,  $p<.001$  and  $0.55$ ,  $p<.001$  for CGI-c with ESS and cataplexy, respectively, for pooled data). A multiple linear model regressing CGI-c on both ESS and cataplexy, with extreme negative cataplexy values truncated to the 5<sup>th</sup> percentile, showed an adjusted  $R^2 = 0.34$  with both ESS and cataplexy significantly related to CGI-c [ $p<.001$ , 95% confidence intervals for b coefficients ESS (0.01, 0.02), cataplexy (0.006, 0.01), standardized beta=0.34 for both]. A description of the aggregated sample is presented in Table 1. The sample was 39% male and 89% white; responders and non-responders had similar distributions. However, responders were slightly younger than non-responders [mean (SD) age responders=39.7 (14.9), non-responders=43.3 (15.6),  $p=0.03$ ]. Overall severity did not differ between responders and non-responders with a vast majority of participants assessed as moderately to markedly ill. Similarly, the baseline measures of the PROs did not differ based on responder status. The mean (SD) ESS score at baseline is 17.5 (3.8) points which is indicative of severe daytime sleepiness. The median frequency of cataplexy attacks per week was 20.5 with 50% of the sample having between 11.5 and 38 episodes per week. At follow-up, the groups classified by CGI-c were significantly different on percent change measures of ESS and frequency of cataplexy attacks; findings that are consistent with the moderate correlations reported above and the area under the curve results reported below.

Insert Table 1 about here



## Response definition threshold

The results of the ROC models are shown in Figure 1 and Table 2. Figure 1 displays ROC curves for both ESS and Cataplexy percent change scores. These outcomes were similarly related to the dichotomous anchor measure indicating treatment response, the CGI-c, with no significant difference noted between the AUCs,  $\chi^2(1) = 0.10$ ,  $p = 0.75$ . The AUCs are indicative of the reliability of the association of each PRO with the anchor measure and indicate a strong relationship with the anchor response definition. The empirical cutpoints estimated using all three optimization criteria (i.e., Liu, Youden, nearest to 0,1) yielded the same estimates, shown in Table 2. We examined the empirical cutpoints separately by trial and found lower thresholds for improvement in the 4-week trial compared to the 8-week trial. The ESS cutpoint for the 4-week trial was similar to the pooled data whereas the cataplexy cutpoint for the 8-week trial was closer to the pooled data. We chose our RD thresholds by rounding to the next largest whole number based upon the pooled data. Based on these thresholds we created dichotomized indicators of response for ESS (ESS\_RD) and cataplexy (cataplexy\_RD). Note that cataplexy\_RD is slightly more sensitive while the ESS\_RD is more specific. Agreement and kappa statistics were comparable for the measures compared to the CGI-c anchor and slightly lower when compared to each other (agreement=69.28%, kappa=0.39).

The distribution-based thresholds are based on the original scale of the ESS and cataplexy frequency measures. For ESS, the 0.5 SD is 1.9 and the standard error of measurement (SEM) is the same. Cataplexy frequency SD=24.6 was based on a truncated range of 5.25-123 (95% of sample) due to extreme values. The distribution based thresholds for cataplexy frequency are 0.5 SD=12.3 and SEM=12.6. As a point of comparison, the ESS\_RD at the sample mean would indicate a 2.1 point decrease in the ESS score. Similarly, the cataplexy\_RD at the median would be equivalent to a decrease of 9.4 cataplexy attacks per week and a decrease of 17.9 attacks for those at the 75<sup>th</sup> percentile.

## Validation

While our anchor measure of clinical improvement and RD thresholds were derived without regard for treatment condition or dose, we present these findings here as a validation of our approach. Both PROs showed dose response relationships in the pooled data. Figure 2 displays the cumulative distribution functions of placebo and other dose levels used in these studies for all possible thresholds. The points at which each curve is dissected by the red line can be used to assess how the proportion of responders differs by treatment dose. For example, the placebo group shows that approximately 30% were responders based on ESS\_RD whereas almost 80% of patients in the highest dose group were responders (see Figure 2a). A logistic model regressing ESS\_RD onto treatment groups showed significant odds ratios (ORs) indicating a higher likelihood of clinical response compared to placebo for the 6 gram (g) group [95% confidence interval (CI) for OR: 1.40, 5.00], and the 9 g group (3.65, 15.35) but not the lowest dose group (3-4.5 g, 0.82, 2.89). A model of cataplexy\_RD regressed on the treatment group showed similar findings (95% CI for OR for lowest dose group: 1.17, 3.87; 6g: 1.31, 4.52, 9g: 2.44, 9.47 reference group: placebo). Models for dose trend were significant for both PRO response definitions (ESS\_RD 95% CI for dose trend OR: 1.15, 1.33; cataplexy\_RD 95% CI for dose trend OR: 1.10, 1.27).

## Discussion

In this secondary analysis of clinical trial data, we have empirically established responder definitions (RDs) that correspond to a Clinical Global Impression of Change (CGI-c) assessment of much or very much improved with fair and consistent accuracy for two debilitating symptoms of narcolepsy. A 12% improvement in the Epworth Sleepiness Scale (ESS) and 46% improvement in weekly cataplexy frequency are the thresholds indicating meaningful improvement over a 4 to 8 week period of treatment among a sample of mostly moderate to markedly ill narcolepsy patients concurrently treated with stimulant therapy. Using these RD cutoffs, we showed a dose response relationship with sodium

oxybate demonstrating the validity of these thresholds. These RDs may be useful in describing patient-centered outcomes in future clinical trials or may be useful criteria to consider for clinical management.

We found only one recent study reporting outcome measurement properties for response to narcolepsy treatment (van der Heide et al., 2015). Similar to the present study, response was classified using Much or Very Much Improved ratings on the CGI-c and reported the Epworth Sleepiness Scale (ESS) as an outcome. In contrast, their sample included patients with and without cataplexy and the context was a trial evaluating modafinil, used to improve wakefulness, in which patients were allowed to continue sodium oxybate if already taking it. They compared outcomes for responders and non-responders, based on the CGI-c, but their primary focus was on performance measures of sustained attention and wakefulness. Thus, RD thresholds were not reported. The authors confirmed the test-retest reliability of the ESS in narcolepsy patients as well as its sensitivity to detect change and showed an effect size of 1.2 between responders and non-responders over an 8-week trial. A corresponding effect size using our data was 0.52 between responders and non-responders (data not shown). This substantial difference illustrates the need to consider sample and agent characteristics when evaluating treatment response. We agree with van der Heide and colleagues who noted that treatment response might be best assessed by using complementary measures of different aspects of narcolepsy burden (van der Heide et al., 2015). In our study a reduction in cataplexy of at least 46% was a more sensitive measure for identifying response and failure to show a 12% decrease or greater in sleepiness was slightly better at identifying non-response (see Table 2, sensitivity and specificity). Using both outcomes provides a more complete assessment of patient functioning.

This study's strengths include a large sample, an anchor-based, statistical approach to identifying RD thresholds, and evaluation of two patient reported outcomes. However, the findings would have been stronger if there had been an additional anchor representing the patient's perspective, such as a patient global rating of change. Unfortunately, a patient global rating of change was not

captured in the two studies from which these data came. Similarly, Multiple Sleep Latency Tests would have been helpful for validating the RD thresholds but were also not collected. Another limitation was that our CGI-c anchor was dichotomized for the ROC analysis, leading to some loss of information. We did, however, present the moderate Spearman correlations using the continuous CGI-c measure to support its use as the anchor. Use of a dichotomized outcome is a criticism of responder analysis in general (Snapinn and Jiang, 2007), however, we felt the development of the RD thresholds using statistical methods would be useful for future therapeutic trials as well as a simple standard applicable to individuals for clinical management of patients. Finally, it should be noted that the CGI-c is a global assessment that may have been informed by change in any symptom or side effect and was not specific to excessive daytime sleepiness and/or cataplexy frequency. Thus, these findings are most generalizable to a similar population of narcolepsy patients considering treatment using similar agent properties, dose, and duration.

In conclusion, we established two thresholds that can be applied to categorize individuals as having meaningful improvement within a 4-8 week trial of sodium oxybate. To our knowledge, we are the first to present responder definitions for the treatment of narcolepsy. While both RD thresholds were comparable for predicting response to treatment, excessive daytime sleepiness and cataplexy frequency are both important aspects of disease burden that should be considered in assessing patient response.

**Acknowledgements:** The clinical trials were funded by Jazz Pharmaceuticals, Inc. We would like to thank Trudy Vanhove for her contribution to the analysis design.

### **Abbreviations**

AUC, area under the curve  
CGI-c, Clinical Global Impression of Change  
CI, confidence interval  
ESS, Epworth Sleepiness Scale  
FDA, Food and Drug Administration

g, gram

ICC, intraclass correlation

OR, odds ratio

PRO, patient reported outcome

RD, response definition

ROC, receiver operator curve

SD, standard deviation

SEM, standard error of measurement

Table 1. Demographic and clinical characteristics of participants by treatment responder classification				
	Total (n=336)	Responder <sup>a</sup> (n=171)	Non-Responder (n=165)	p-value
Gender, n (%)				0.8809
Male	131	66 (50%)	65 (50%)	
Female	205	105 (51%)	100 (49%)	
Race, n (%)				0.4287
White	298	149 (50%)	149 (50%)	
Black	29	18 (62%)	11 (38%)	
Other	9	4 (44%)	5 (56%)	
Age, yr [mean (SD)]	41.5 (15.3)	39.7 (14.9)	43.3 (15.6)	0.0349
CGI-S, n (%)				0.2969
Normal	3	1 (33%)	2 (67%)	
Borderline	22	9 (41%)	13 (59%)	
Slightly III	83	37 (45%)	46 (55%)	
Moderately III	138	78 (57%)	60 (43%)	
Markedly III	67	37 (55%)	30 (45%)	
Extremely III	23	9 (39%)	14 (61%)	
Baseline ESS [mean (SD), range 6-24]	17.5 (3.8)	17.3 (3.7)	17.7 (3.9)	0.3677
Baseline Cataplexy [Median (IQR), range .5-874]	20.5 (27.5)	21.5 (30.3)	18.5 (23.8)	0.2928
Trial				0.755
4-week	125	65 (52%)	60 (48%)	
8-week	211	106 (50%)	105 (50%)	
Sodium Oxybate Dose				<.0001
placebo	92	24 (26%)	68 (74%)	
3 g	30	14 (47%)	16 (53%)	
4.5 g	62	32 (52%)	30 (48%)	
6 g	85	47 (55%)	38 (45%)	
9 g	67	54 (81%)	13 (19%)	
ESS % change [mean (SD), range -71.4-90.9]	14.8 (25.3)	26.5 (25.4)	2.6 (18.5)	<.0001
Cataplexy % change [mean (SD), range -831.8-100]	35.1 (82.4)	56.7 (83.0)	12.3 (75.6)	<.0001
<sup>a</sup> Classified by anchor measure, CGI-c, as Much or Very Much Improved				

Table 2. Receiver operating characteristic results and responder definition threshold performance for patient reported outcomes		
	ESS	Cataplexy
ROC AUC	0.78	0.77
(95% CI)	(0.73 - 0.83)	(0.72 - 0.82)
Empirical cut point	11.44	45.76
(95% CI)	(6.70 - 16.17)	(30.07 - 61.45)
AUC at cut point	0.73	0.72
RD Threshold	12	46
Sensitivity	0.69	0.77
Specificity	0.75	0.66
Agreement CGI-c	71.69%	71.39%
Kappa CGI-c	0.43	0.43

Abbreviations: ROC receiver operating characteristic; AUC area under the curve; CI confidence interval;

RD response definition, CGI-c clinical global impression of change.

Figure 1. Receiver operator characteristic (ROC) curves for patient reported outcomes as predictors of treatment response

---

Insert Figure 1 here

---

Figure 2a. Cumulative distribution function of excessive daytime sleepiness by treatment dose

---

Insert Figure 2a here

---

Note: Vertical red line denotes response definition threshold of 12%

Figure 2b. Cumulative distribution function of cataplexy frequency by treatment dose

---

Insert Figure 2b here

---

Note: Vertical red line denotes response definition threshold of 46%

Cappelleri, J. C. and Bushmakina, A. G. Interpretation of patient-reported outcomes. *Stat. Methods Med. Res.*, 2014, 23: 460-83.

Clayton, P. cutpt-Empirical estimation of cutpoint for a diagnostic test. In, <http://fmwww.bc.edu/RePEc/bocode/c/cutpt.sthlp>, 2013.

DeLong, E., DeLong, Dm, Clarke-Pearson, DI Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 1988, 44: 837-45.



- Fluss, R., Faraggi, D., Reiser, B. Estimation of the Youden index and its associated cutoff point. *Biometrical J.*, 2005, 47: 458-72.
- International, X. A double-blind, placebo-controlled study demonstrates sodium oxybate is effective for the treatment of excessive daytime sleepiness in narcolepsy. *J. Clin. Sleep Med.*, 2005, 1: 391-97.
- Johns, M. A new method for measuring daytime sleepiness: the Epworth Sleepiness Scale. *Sleep*, 1991, 14: 540-45.
- Johns, M. Sensitivity and specificity of the multiple sleep latency test (MLST), the maintenance of wakefulness test and the Epworth sleepiness scale: failure of the MSLT as a gold standard. *J. Sleep Res.*, 2000, 9: 5-11.
- Liu, X. Classification accuracy and cut point selection. *Stat. Med.*, 2012, 31: 2676-86.
- McLeod, L. D., Coon, C. D., Martin, S. A., Fehnel, S. E. and Hays, R. D. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. *Expert review of pharmacoeconomics & outcomes research*, 2011, 11: 163-9.
- Snapinn, S. M. and Jiang, Q. Responder analyses and the assessment of a clinically relevant treatment effect. *Trials*, 2007, 8: 31.
- Snedecor, G., Cochran, W. G., *Statistical Methods*. Iowa State University Press, Ames, Iowa, 1989 (Eighth edition).
- StataCorp Stata Statistical Software: Release 12. In. StataCorp LP, College Station, TX, 2011.
- The U.S. Xyrem<sup>(R)</sup> Multicenter Study Group A Randomized, Double Blind, Placebo-Controlled Multicenter Trial Comparing the Effects of Three Doses of Orally Administered Sodium Oxybate with Placebo for the Treatment of Narcolepsy. *Sleep*, 2002, 25: 42-49.
- US Department of Health and Human Services Guidance for industry on patient-reported outcome measures: Use in medical product development to support labeling claims. In: F. A. D. ADMINISTRATION (Ed, <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf>, 2009.
- Van Der Heide, A., Van Schie, M. K., Lammers, G. J. *et al.* Comparing Treatment Effect Measurements in Narcolepsy: The Sustained Attention to Response Task, Epworth Sleepiness Scale and Maintenance of Wakefulness Test. *Sleep*, 2015, 38: 1051-8.
- Weaver, T. E. Outcome measurement in sleep medicine practice and research. Part 1: assessment of symptoms, subjective and objective daytime sleepiness, health-related quality of life and functional status. *Sleep Med. Rev.*, 2001, 5: 103-28.
- Wyrwich, K. W., Norquist, J. M., Lenderking, W. R. and Acaster, S. Methods for interpreting change over time in patient-reported outcome measures. *Qual. Life Res.*, 2013, 22: 475-83.
- Xyrem<sup>(R)</sup> International Study Group Further evidence supporting the use of sodium oxybate for the treatment of cataplexy: A double-blind, placebo-controlled study in 228 patients. *Sleep Med.*, 2005, 6: 415-21.
- Youden, W. Index for rating diagnostic tests. *Cancer*, 1950, 3: 32-35.