Five Urban Legends of CNS Clinical Trial Methodology: Unsuccessful Solutions to the Problem of Failed Trials

Williams JBW [1,2], Popp D[1], Reines S[1], Detke M[1,3]

[1]MedAvante, Inc. [2]College of Physicians and Surgeons, Columbia University [3]Indiana University School of Medicine

INTRODUCTION

The problem of failed trials in CNS is well recognized (Khin et al. 2011). As the number of failed trials has grown, drug developers have attempted several strategies to improve signal detection and reduce the failure rate. We present five common strategies and evaluate the evidence for their effectiveness.

URBAN LEGEND #1
INCREASING SAMPLE SIZE WILL INCREASE STATISTICAL POWER

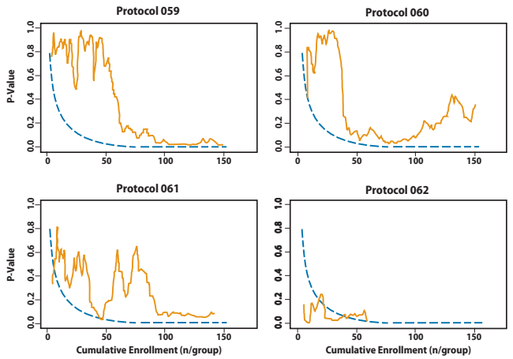
If one assumes statistical power increases with sample size and effect size is fixed, it appears reasonable to assume that increasing sample size will increase effect size in a given study.

Testing the legend:
Liu et al. (2008) examined four randomized, double-blind placebo controlled phase III depression trials with at least 150 subjects per treatment arm to investigate the effect of cumulative sample size on drug-placebo difference.

The results:
A positive treatment effect for paroxetine was observed before the first 100 patients per treatment arm were enrolled in each study. Continuing to enroll patients did not maintain the achieved level of significance in most cases, and in one study it actually appeared to turn a positive result into a negative one.

Figure 1: P-value for paroxetine v. placebo on the HAM-D17 as a function of cumulative patient enrollment.

(**Dashed line:** calculated P-value; **solid line:** observed P-value)



Combined data from all four trials showed later-enrolling patients were more likely to be placebo responders than earlier-enrolling patients.

Summary:
Treatment effect size in both US and non-US depression trials has been decreasing over time despite a steady increase in sample size per treatment arm (see also Khin et al., 2011). Larger sample sizes do not ensure positive treatment effects in psychiatric clinical trials. In addition, increasing sample size requires increasing the number of raters, sites, and countries, all of which contribute to increased variability which may in turn decrease effect size while adding time and cost.

URBAN LEGEND #2
CHOOSING THE "RIGHT" SITES WILL REDUCE RISK OF FAILURE

There is a belief that selecting investigative sites with proven results will continue to yield conclusive results.

Testing the legend:
Gelwicks et al. (2002) analyzed data from 21 clinical trial sites that participated in at least two different trials and randomized at least 30 subjects across trials.

The results:
Site performance across consecutive studies of fluoxetine was inconsistent, with <0.50 correlation within sites across studies on randomization rates, protocol completion percentage, percentage of placebo responders and drug-placebo difference.

Figure 2: Correlations of the same metrics for the same sites in consecutive trials.

Randomization Rate	Percent of Protocol Completers	Percent of Placebo Responders	P Values for Drug-Placebo Separation*
0.498	0.032	0.226	0.207

* Correlations of 1 – (P-value of drug-placebo treatment difference)

Summary:
Since most sites only enroll 5-10 subjects per study arm, they are underpowered to demonstrate drug-placebo differences. In addition, many sites experience considerable personnel turnover from year to year, which may affect their success rate. Good performance of a site in one clinical trial had a very low correlation with good performance in the next study at that same site.

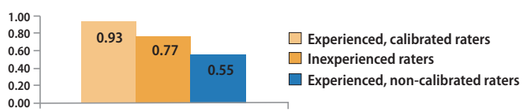
URBAN LEGEND #3
USING THE MOST EXPERIENCED RATERS WILL REDUCE RISK OF FAILURE

It seems logical that employing more experienced raters will minimize variability and improve signal detection.

Testing the legend:
Kobak et al. (2009) examined the relative impact of experience and calibration by calculating interrater agreement across three groups of raters: an experienced and calibrated cohort, an experienced but non-calibrated cohort, and an inexperienced cohort. Thirty subjects with MDD were assessed in independent interviews by two different raters on the same day using the Structured Interview Guide for the Hamilton Depression Scale (SIGH-D).

The results:
The highest interrater agreement was achieved by experienced and calibrated raters (r=0.93), followed by inexperienced raters (r=0.77). Experienced but non-calibrated raters achieved the lowest interrater agreement (r=0.55).

Figure 3: Intraclass correlation coefficient (ICC) values of interrater agreement of three rater cohorts.



Summary:
Clinical experience alone does not result in good interrater reliability within a cohort of raters. Experienced raters must be carefully calibrated with each other to achieve this.

URBAN LEGEND #4
INCREASING RATER TRAINING WILL REDUCE RISK OF FAILURE

A frequently cited cause of trial failure is inadequate rater training. There is the potential for huge variability across raters in a single trial, which negatively affects study power and signal detection. Some believe increasing the intensity of rater training will reduce variability.

Testing the legend:
Demitrack et al. (1998) trained 85 raters on the HAM-D in an intensive, six hour iterative training session with four videotapes and discussion between each tape. Training consisted of a lecture on the HAM-D, a detailed review of each individual item and how to rate it, and a review of a training manual designed for the particular study.

The results:
ICCs across the four training tapes ranged from 0.65-0.79 and did not improve across the six hours of reliability training.

Figure 4: Relation between baseline rating status (tape 1 ICC score) and trained rating status (tapes 2 through 4 ICC score average).

"Baseline" Performance ICC (Based on Tape 1 ICC Value)	"Trained" Performance ICC (Based on Tapes 2-4 ICC Value) N	Mean	SD	p*
ICC > 0.75				
Yes	65	.83	.06	.04
No	13	.76	.11	
ICC > 0.80				
Yes	55	.84	.06	<.01
No	23	.77	.10	
ICC > 0.85				
Yes	38	.85	.05	<.01
No	40	.78	.09	

NOTES: Seven raters were excluded from this analysis because of missing values; *p* values are from *t*-tests of the mean values in each subgroup.

*Significant difference in the comparisons of the standard deviations in each subgroup.

Summary:
Intensive group rater training at the beginning of a study does not improve interrater reliability significantly, even in the short term. Thus, the variability that results from combining raters from many different sites remains.

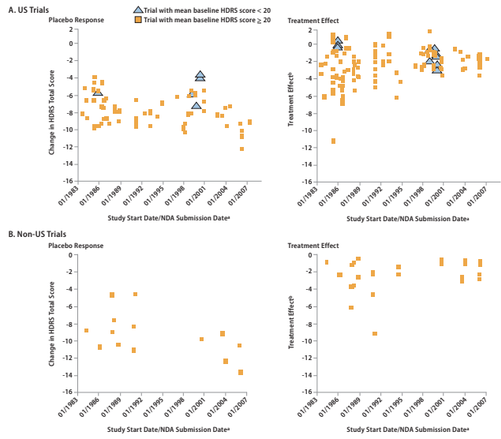
URBAN LEGEND #5
CERTAIN REGIONS OF THE WORLD HAVE BETTER SIGNAL DETECTION

Many researchers believe that greater signal detection can be obtained outside the US.

Testing the legend:
Khin et al. (2011) conducted a meta-analysis of 81 randomized double-blind clinical trials of antidepressants that were submitted to the FDA between 1983 and 2008, including both US and ex-US studies. Another meta-analysis (Khin et al, 2009) looked at 35 randomized double-blind clinical trials of schizophrenia.

The results:
Both meta-analyses documented increasing placebo response across both US and ex-US regions, and a decrease in treatment effect for US studies.

Figure 5: Placebo responses and treatment effects over time in US and non-US MDD trials.



*Horizontal axes denote the trial date, if known, or, if unknown, the NDA submission date. †Calculated as the drug-placebo difference, ie, mean change in HDRS total score for the antidepressant group minus mean change for the placebo group. Abbreviations: HDRS = Hamilton Depression Rating Scale, MDD = major depressive disorder, NDA = new drug application.

Summary:
Although there is regional variability, the failure rate of studies of depression and schizophrenia outside the US is increasing.

CONCLUSIONS

Strategies for improving signal detection in CNS clinical trials are often used without clear evidence of their efficacy. Increasing sample sizes, targeted site selection, using experienced but non-calibrated raters, increasing rater training, and conducting trials ex-US have not proven successful in increasing the success rate. These "urban legends" are widely touted, but evidence to support them is lacking.

References

Demitrack MA, Faries D, Herrera JM, Debrota DJ, Potter WZ. The problem of measurement error in multisite clinical trials. Psychopharmacol Bull, 1998; 34 (1): 19-24. Gelwicks S, Debrota DJ, Engelhardt N, Potter WZ. Analysis of site performance across multiple clinical trials. Presented at the 42nd annual meeting of the National Clinical Drug Evaluation Unit (NCDEU), June 2002. Khin NA, Chen YF, Yang Y, Yang P, Laughren TP. Exploratory analyses of efficacy data from major depressive disorder trials submitted to the US Food and Drug Administration in support for new drug applications. J Clin Psychiatry, 2011; 72 (4): 464-472. Khin NA. Update on regulatory and scientific issues regarding the reliance of efficacy data from foreign sites to support New Drug Applications (NDAs) and supplements. Presented at the 49th annual meeting of the National Clinical Drug Evaluation Unit (NCDEU), July 2009. Kobak KA, Brown B, Sharp I, Levy-Mack H, Wells K, Ockun F, Williams JBW. Sources of unreliability in depression ratings. J Clin Psychopharm 2009; 29 (1) ,82-85. Liu KS, Snavely DB, Ball WA, Lines CR, Reines SA, Potter WZ. Is bigger better for depression trials? J Psychiat Res, 2008; 42 (8): 622-630.

Disclosures

Popp, D, Williams, JBW, MedAvante Inc. Detke, M, MedAvante Inc., NIH, Denysias, Inc., Sonkei Inc., Insight Neuropharma, Inc., Jeevan Scientific, Inc., Pam Lab, Inc., Columbia NW Pharmaceuticals, Naurex, Inc.

MEDAVANTE® RESEARCH INSTITUTE
Calibrated for Certainty