The International Society for CNS Clinical Trials and Methodology, Washington, DC (February, 2011).

Using the GAF to assess functional impairment and change in Shift Work Disorder (SWD)

Authors: Brian Rothman, PhD¹, Ryan Dammerman, MD, PhD², Sofija Jovic, PhD¹, Christian Yavorsky, PhD³, Mark Opler, MPH, PhD¹,³

Affiliations: ¹ProPhase, LLC, ²Cephalon, ³New York University

Abstract

Background: Circadian rhythm sleep disorder (shift work type), more commonly referred to as shift work disorder (SWD) is characterized by persistent sleep and arousal problems, stemming from the mismatch between a person’s endogenous sleep-wake cycle and the pattern of sleep and wakefulness required by nocturnal shift work. Individuals with this disorder may report insomnia at certain times of the day and excessive drowsiness at other times, resulting in marked subjective distress. Anxiety, depression, and cognitive symptoms such as inattention are also commonly exhibited (Theodorou & Rice, 2007). Additionally, because circadian rhythm sleep-disorders are associated with functional impairment in social, occupational, and psychological realms (DSM-IV-TR), we hypothesized that broad-based psychiatric scales such as the Global Assessment of Functioning (GAF) would be able to encapsulate meaningful aspects of symptom course and severity. SWD is rarely examined in psychiatric research despite the often debilitating symptoms it produces. The GAF, a cross-sectional measure of a patient’s overall level of functioning and itself Axis V of the standard multi-axial psychiatric assessment (DSM-IV-TR), has been widely used in prior research as an outcome measure. Though some researchers have questioned the GAF’s clinical utility due to observed low inter-rater reliability estimates (e.g., Aas, 2010; Goldman et al., 1992; Rey et al., 1994), these problems have been largely mitigated in research settings by formal training and monitoring (Bates et al., 2002; Hilsenroth et al., 2000; Vatnaland et al., 2007). In accordance with the above hypothesis, we sought to determine whether the GAF could be used in clinical trials as a reliable and valid way to detect functional impairment and change in a group of SWD patients. Our secondary aim was to contribute to existing research on the GAF by assessing whether training method (e.g., video vs. written vignette) impacts reliability and rater accuracy.

Methods: Participants were principal investigators and site raters being trained at an investigator meeting (IM) for participation in a large, multi-center Phase IV clinical trial of SWD. The standardized rater training consisted of both online pre-IM and didactic components. Raters were also provided with interview guides addressing rating strategy and descriptions of GAF anchors. Post-training, raters assigned GAF scores (ranging from 0-100, with higher ratings indicating higher levels of functioning) after reviewing two videotaped interviews and a series of three vignettes depicting patients with sleep-disorder. Two patients were portrayed serially. One patient appeared in the videos, and the other appeared in the vignettes. Both patients met DSM-IV criteria for Circadian Rhythm Sleep Disorder – Shift Work Type. The videotaped patient was interviewed using a semi-structured psychiatric interview administered by a trained clinician, and the average length of each video was 23 minutes. The average length of each vignette was 300 words.

Paired sample t-tests were used to examine whether GAF ratings could detect differences in level of functioning at each time point. Additionally, to measure concordance with expert ratings, GAF change scores were computed and compared to gold-standard GAF ratings using one-sample mean comparisons. These change scores were also used to examine rater accuracy according to training modality (video vs. vignette). Expert “gold-standard scores” were obtained by a group of trained and calibrated clinicians.

Results: The GAF was sensitive to changes in functioning level in SWD patients. Mean comparisons between the two videotaped interviews were highly significant ((t(42) = -10.49, p<.001). Additionally, 81.4% of participants correctly observed an improvement in the patient’s functioning level between time 1 and time 2 (i.e., an increase in 10-20 GAF points). On average, this change score was statistically equivalent to the expert score.
Participant GAF ratings also differed significantly between vignette 1 and vignette 2 ($t(96) = -15.02, p<.001$) and between vignette 2 and vignette 3 ($t(57) = 17.47, p<.001$). However, concordance with expert ratings for the vignettes was mixed. While statistical agreement was achieved between time 1 and time 2 ($M_{diff} = 9.90$), differences in GAF scores assigned by expert vs. participant raters between time 2 and time 3 were significant ($t(57) = 6.56, p<.001$). Specifically, though raters and experts both observed a decline in the patient’s functioning between vignette 2 and vignette 3, raters assigned significantly higher GAF ratings than experts on average ($M_{diff} = -13.28\) vs. -20 for raters vs. experts, respectively).

**Conclusions:** Results suggest that the GAF can assess functional changes in SWD patients, and that this population exhibits social, occupational, and psychological problems that are observable when using broad-based psychiatric scales. Consistent with prior research (e.g., Hilsenroth, 2000), results also indicate that raters are able to score the GAF accurately following training that addresses rating strategy and consistent conceptualization of GAF anchors. In this study, video training materials were superior to vignettes for reaching concordance with expert ratings. While this finding may suggest that video-based trainings are preferable for rater accuracy, it may also be the case that data from the vignettes was subject to expectancy bias, a phenomenon shown to impact raters’ perceptions of treatment effects (e.g., Michelson, et al., 1985). The fact that raters showed higher agreement with expert scores when ratings changed in the expected direction (improvement in the second visit) rather than an unexpected one (worsening after baseline) supports this explanation. Thus, training on the GAF should address expectancy issues in addition to rater error and other inter-rater reliability problems. This approach will help to improve the GAF’s construct validity when applied to ratings of broader populations, as well as to this SWD cohort.

Cephalon sponsored the IM referenced in this abstract and provided a review of this analysis.

**References:**


