Measuring outcomes in randomized clinical trials of insomnia treatments

Charles M. Morin

Université Laval, Québec, Canada

Summary Significant efforts have been made in the last decade to develop evidence-based guidelines for the treatment of insomnia and other sleep disorders. Despite such progress, there are still no standard assessment methods to document outcome and no accepted criteria to define what should be a successful outcome in the treatment of insomnia. This paper reviews methodological and conceptual issues related to the measurement of outcomes in clinical trials of insomnia. Selected studies of behavioral and pharmacological therapies of insomnia are summarized to illustrate the types of dependent variables and assessment instruments used to document treatment efficacy. Additional outcome variables and assessment methods of potential interest are discussed, and criteria for interpreting and reporting outcomes are summarized. As most studies have relied on fairly narrow criteria to define an effective treatment, the need to broaden the scope of outcome assessment is highlighted. For instance, it is essential to document treatment efficacy beyond the simple reduction of insomnia symptoms and to incorporate additional indicators of success. Given that insomnia is associated with significant morbidity, an effective treatment should not only improve sleep parameters, but it should also produce clinically meaningful changes in daytime functioning, fatigue, mood, and quality of life. The need to evaluate outcome from multiple perspectives and to develop a core-assessment battery that would consider efficacy, clinical significance, and cost-effectiveness are discussed. © 2003 Elsevier Science Ltd. All rights reserved.

INTRODUCTION

Insomnia is a prevalent and costly health complaint that may present as a symptom or a clinical syndrome [1–4]. With increasing demand for accountability and cost-containment from the government, third-party payers, and consumers, there is also an increasing need for documenting the efficacy, clinical utility, and cost-effectiveness of treatments for sleep disorders. Despite recent efforts to develop evidence-based guidelines for the evaluation [5] and treatment of insomnia [6], there is still no consensus on how we should measure outcome and no accepted standards on what should be an optimal outcome after therapy. Should treatment outcome be defined strictly in terms of sleep improvements, or should we consider daytime functioning, subjective well being, and quality of life? Should outcome evaluation rely exclusively on polysomnography, or incorporate patients’ and clinicians’ measures? And, how should we judge the clinical significance of outcomes? Clinical studies of hypnotic medications and behavioral therapies have used fairly different methods and criteria to evaluate outcome,
and those may not always have been optimal for informing practitioners about what to expect in their everyday clinical practice.

Evidence-based guidelines and empirically supported therapies [7] rely heavily on efficacy rather than on effectiveness criteria, with a predominant emphasis on experimental control and statistical significance rather than on clinical utility and applicability of treatment. Those criteria, although necessary, may yield a narrow definition of “effective treatment” as they fail to take into account effect size, clinical significance, and treatment utility and acceptability. Evidence that a treatment is statistically superior to a control condition on a single sleep parameter is not very informative and should not be an end in itself. The magnitude of improvement, the impact of treatment on daytime functioning and subjective well being, and the proportion of patients who are likely to benefit from therapy represent additional clinical indicators of success potentially useful to inform both clinicians and researchers.

The present paper addresses conceptual and methodological issues related to the measurement of treatment outcome in randomized clinical trials of insomnia therapies. The first part summarizes outcome variables and assessment methods used in selected clinical trials of behavioral and pharmacological therapies for insomnia. Criteria and guidelines for selecting outcome variables and assessment instruments are discussed. Different methods for interpreting and reporting outcomes are examined, including the use of effect sizes, improvement rates, the proportion of patients achieving remission, and the durability and generalizability of treatment effects. The paper concludes with some recommendations for clinical practice and for future research.

SELECTING OUTCOME VARIABLES

The strength of the conclusions that can be drawn from treatment studies is largely dependent upon the use of rigorous methodology and a selection of outcome variables that are conceptually sound, statistically reliable, and clinically relevant. This selection should be comprehensive enough to capture not only the presenting sleep complaint, but also its associated correlates and consequences such as daytime functioning and fatigue, mood, quality of life, and functional status.

Table 1 summarizes the main outcome measures and assessment instruments used in 16 clinical studies of behavioral [8–13], pharmacological [14–20], and combined/comparative behavioral and pharmacological therapies for insomnia [21–23]. Those studies were selected to illustrate a variety of measures and instruments used by different investigators around the world and with different forms of insomnia (i.e. primary and secondary). Their selection was intended to be neither exhaustive nor based on any particular criteria except to provide a sample of current methodological practices and outcome measures used in clinical trials of insomnia. Although the majority of the selected studies addressed the treatment of primary insomnia, a few of them focused on insomnia associated with medical [9] and psychiatric conditions [15].

As the chief complaint among insomnia patients involves difficulties falling or staying asleep or poor sleep quality, it is only logical to select sleep parameters reflecting on such complaints as primary outcome variables. As shown in Table 1, the most frequently used dependent variables include sleep onset latency, the number and duration of awakenings, total sleep time, sleep efficiency, and sleep quality. When polysomnography (PSG) is available, percentages of sleep time spent in each stage are also reported in addition to sleep continuity parameters. Although it is common practice to report multiple dependent sleep variables, the primary dependent variable is usually based on whether the treatment under evaluation is intended for initial, middle, or mixed insomnia. Sleep onset latency and number and duration of nocturnal awakenings are the most frequently reported dependent variables with sleep onset and sleep maintenance insomnia, respectively [24]. There is, however, no standard definition of early morning awakening and no study that have focused specifically on this subtype of insomnia. As the majority of insomnia patients present with mixed insomnia [25], sleep efficiency may represent the single best end point to capture the extent of sleep disturbances at baseline and to evaluate changes with treatment.

Sleep disturbances are associated with functional impairments, reduced quality of life, and increased health-care utilization [1, 4, 26–28]. Therefore, it is imperative to document the impact of treatment on those collateral variables. For instance, daytime fatigue is probably the most frequent complaint associated with insomnia in clinical practice [29, 30] and, to be considered clinically meaningful, sleep improvements should also lead to a reduction of daytime fatigue. Nonetheless, very few studies have actually targeted
### Table 1  Selected treatment outcome studies of behavioral and pharmacological therapies for insomnia

<table>
<thead>
<tr>
<th>Authors, year</th>
<th>Diagnosis</th>
<th>Treatment conditions (length of treatments and follow-up assessments)</th>
<th>Primary outcomes</th>
<th>Secondary outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backhaus et al., 2001</td>
<td>Primary chronic insomnia $n = 20$</td>
<td>1. Multicomponent CBT (6 wk; 12 mo)</td>
<td>1. Pittsburgh sleep quality index</td>
<td>1. Beck depression inventory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Polysomnography (TST, SOL, SE)</td>
<td></td>
<td>2. State-trait anxiety inventory</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3. Questionnaire about insomnia-related cognitions</td>
</tr>
<tr>
<td>Cluydts et al., 1998</td>
<td>Chronic insomnia $n = 160$</td>
<td>1. Continuous (zolpidem 10 mg every nights) 2. Intermittent (zolpidem 10 mg for 5 nights followed by placebo for 2 nights) (2 wk; n/a)</td>
<td>1. Sleep diaries (TST)</td>
<td>1. Sleep diaries (SQ, global evaluation of impairment, SOL, NA, early awakenings)</td>
</tr>
<tr>
<td>Currie et al., 2000</td>
<td>Insomnia secondary to chronic pain $n = 60$</td>
<td>1. Multicomponent CBT 2. Waiting-list control (7 wk; 3 mo)</td>
<td>1. Sleep diary (TST, SOL, NA, SE, WASO, SQ)</td>
<td>1. Multidimensional pain inventory scale</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Pittsburgh sleep quality index</td>
<td></td>
<td>2. Beck depression inventory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Actigraphy (activity level)</td>
<td></td>
<td>3. Medication quantification scale</td>
</tr>
<tr>
<td>Edinger et al., 2001</td>
<td>Primary chronic sleep-maintenance insomnia $n = 75$</td>
<td>1. Multicomponent CBT 2. Progressive muscle relaxation</td>
<td>1. Sleep diaries (TST, WASO, SE, SQ)</td>
<td>1. Self-efficacy scale</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Polysomnography (TST, WASO, SE)</td>
<td>2. Beck depression inventory</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Psychological placebo (6 wk; 6 mo)</td>
<td>3. Insomnia symptoms questionnaire</td>
</tr>
<tr>
<td>Elie et al., 1999</td>
<td>Primary insomnia or insomnia associated with mild non psychotic psychiatric disorders $n = 574$</td>
<td>1. Zaleplon 5 mg 2. Zaleplon 10 mg 3. Zaleplon 20 mg 4. Zolpidem 10 mg 5. Placebo (28 days; 3 days)</td>
<td>1. Sleep diaries (SOL, TST, NA, SQ)</td>
<td>1. Benzodiazepine withdrawal symptom questionnaire</td>
</tr>
<tr>
<td>Espie et al., 2001</td>
<td>Primary chronic insomnia $n = 139$</td>
<td>1. Multicomponent CBT 2. Waiting-list control (6 wk; 12 mo)</td>
<td>1. Sleep diaries (SOL, WASO, TST)</td>
<td>1. Pittsburgh sleep quality index</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2. Sleep diaries (NA, SQ, use of sleep medication)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3. Beck depression inventory</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4. State-trait anxiety inventory</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5. Penn state worry questionnaire</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6. Epworth sleepiness scale</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7. Actigraphic data</td>
</tr>
<tr>
<td>Authors, year</td>
<td>Diagnosis</td>
<td>Treatment conditions (length of treatments and follow-up assessments)</td>
<td>Primary outcomes</td>
<td>Secondary outcomes</td>
</tr>
<tr>
<td>-----------------</td>
<td>--------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Hajak et al., 2001 | Primary insomnia $n = 47$     | 1. Doxepin 25 mg  
2. Doxepin 50 mg  
3. Placebo (4 wk; 2 wk)                                                                 | 1. Polysomnography (TST, SE, WASO, SOL, REM latency, % of sleep time in each stage, sleep period time) | 1. Fisher somatic undesired effects check-list  
2. Investigator ratings (severity of illness, global improvement) |
| Hauri, 1997     | Primary insomnia $n = 26$     | 1. Sleep hygiene and relaxation plus triazolam 0.25 mg  
2. Sleep hygiene and relaxation  
3. Waiting list control (6 wk; 10 mo)                                   | 1. Polysomnography (SOL, TST, WASO, SE, NA per hour)                                                   | 1. Performance batteries (platform balance test, reaction time measures, digit span, divided attention test, digit symbol substitution task, auditory-verbal memory test, vigilance test)  
2. Multiple sleep latency test |
| Jacobs et al., 1993 | Chronic sleep-onset insomnia $n = 26$ | 1. Multicomponent CBT (10 wk; 6 mo)                                                                                             | 1. Polysomnography (SOL, TST, SE)                                                                    | 1. EEG power spectra  
2. Beck depression inventory  
3. Center for epidemiologic studies depression scale  
4. Spielberger state-trait anxiety inventory |
| Kripke et al., 1990 | Chronic insomnia $n = 107$   | 1. Flurazepam 30 mg  
2. Flurazepam 15 mg  
3. Midazolam 15 mg  
4. Placebo (2 wk; n/a)                                                        | 1. Polysomnography (SOL, WASO, SE)                                                                   | 1. Insomnia impact scale  
2. Beliefs and attitudes about sleep scale  
3. Fatigue severity scale  
4. Epworth sleepiness scale |
| Lichstein et al., 2001 | Primary insomnia $n = 74$     | 1. Sleep hygiene + relaxation  
2. Sleep hygiene + sleep compression  
3. Sleep hygiene + psychological placebo (6 wk; 1 yr)                           | 1. Polysomnography (SOL, TST, WASO, NA, SE, % of sleep time in each stage)                            | 1. Insomnia impact scale  
2. Beliefs and attitudes about sleep scale  
3. Fatigue severity scale  
4. Epworth sleepiness scale |
<table>
<thead>
<tr>
<th>Authors, year</th>
<th>Diagnosis</th>
<th>Treatment conditions</th>
<th>Primary outcomes</th>
<th>Secondary outcomes</th>
</tr>
</thead>
</table>
| McClusky et al., 1991 | Primary chronic insomnia $n = 30$ | 1. Stimulus control and relaxation training  
2. Triazolam 0.5 mg (3 wk; 5 wk) | 1. Sleep diaries  
(NA, WASO, SOL, TST, Morning restudies, difficulty to fall asleep) | 1. Beliefs and attitudes about sleep scale  
2. Beck depression inventory  
3. Beck anxiety inventory  
4. Neuropsychological measures |
| Morin et al., 1999 | Primary chronic insomnia $n = 78$ | 1. Multicomponent CBT  
2. Temazepam 30 mg  
3. Combination of both  
4. Placebo (8 wk; 3 mo, 12 mo, 24 mo) | 1. Sleep diaries  
(WASO, TST, SE)  
2. Polysomnography  
(WASO, TST, SE)  
3. Sleep impairment index | 1. Digit symbol substitution test  
2. Digit symbol copying test  
3. Subjective morning sleepiness |
| Scharf et al., 1994 | Primary chronic insomnia $n = 75$ | 1. Zolpidem 10 mg  
2. Zolpidem 15 mg  
3. Placebo (5 wk; 3 days) | 1. Polysomnography  
(SOL, SE, NA, WASO, % of sleep time in each stage)  
2. Morning questionnaire  
(SOL, TST, NA, SQ, ease of falling asleep, refreshing quality of sleep) | 1. Subjects’ ratings of daytime mood and behavior  
2. Somatic symptoms and side effects  
3. Evaluation of daytime status by a general practitioner  
4. Respondents’ preference for a particular medication |
| van der Kleijn, 1989 | Primary insomnia $n = 53$ | 1. Zopiclone 7.5 mg  
2. Temazepam 20 mg  
3. Placebo (3 wk; n/a) | 1. Morning questionnaire  
(SQ, SOL, status after awakening) | 1. Adverse events |
| Walsh et al., 2000 | Primary insomnia $n = 113$ | 1. Zaleplon 10 mg  
2. Placebo (5 wk; 2 days) | 1. Polysomnography (SOL, TST, wake time in the fourth quarter of the night, % of sleep time in each stage, REM latency)  
2. Sleep diaries (SOL, TST, NA) | |

Note: SOL = Sleep-onset latency; NA = Number of awakenings; WASO = Wake after sleep onset; TST = Total sleep time; TWT = Total wake time; SE = Sleep efficiency; SQ = Sleep quality.
Likewise, the subjective insomnia complaint is often accompanied by reports of neurobehavioral impairments (e.g. attention, concentration, memory, information processing, coordination); yet, only rarely have investigators incorporated such measures in clinical trials [18, 21, 22]. Despite discrepancies between subjective and objective deficits [31, 32], which are due partly to a faulty appraisal mechanism and to the lack of measurement sensitivity, such complaints must be targeted for outcome assessment because it is often the perception of those daytime deficits that prompt patients to seek treatment.

There is a high rate of comorbidity between insomnia and psychopathology, particularly with affective and anxiety disorders [1, 33–35]. Even among individuals who do not meet threshold for a psychiatric diagnosis of major depression or generalized anxiety disorder, psychological symptoms of anxiety and depression, as well as other mood disturbances (e.g. dysphoria, irritability), are extremely frequent in chronic insomnia [25, 30, 36, 37]. Although it is often difficult to determine whether such clinical features represent a cause or a consequence of insomnia, psychological symptoms tend to covary with sleep disturbances. The presence of higher emotional distress is also an important feature distinguishing patients who do or do not seek treatment for insomnia [38]. These psychological features, with the perceived daytime impairments, represent ideal targets to document the clinical significance of treatment outcome. Less than 50% of the selected studies in Table 1, mostly behavioral studies [8–12, 22], have included such outcome measures of psychological symptomatology (i.e. anxiety, depression). As most of those clinical trials have focused on primary insomnia, and typically excluded patients with severe psychopathology, only modest changes in symptom scores have been associated with treatment.

Quality of life (QOL) is very relevant to sleep [27, 28]. It is a fairly broad and multidimensional construct that may incorporate any of the following domains: psychological well being, health status, social and occupational functioning, leisure activities, life satisfaction, standard of living, etc. Despite the extensive variability in defining and measuring this construct [39–41], there is increasing attention to quality of life assessment as a means of broadening the evaluation of treatment outcome in sleep disorders research [42, 43]. Although none of the selected studies in Table 1 used such measures, several ongoing clinical trials by our group and other investigators are using QOL measures as part of the outcome assessment battery.

In addition to the psychological and emotional aspects of insomnia, it is important to consider also the physiological parameters of this condition. Indeed, there is increasing evidence that chronic primary insomnia is characterized by elevated CNS and physiological arousal (e.g. temperature, evoked potentials, metabolic rate, cortisol secretion) [44–46]. Unfortunately, such indices have not been used as outcome measures in clinical trials of insomnia.

Finally, with the increasing evidence that individuals with chronic sleep disturbances use health care services more than good sleepers [1, 26], the demonstration that sleep improvements is associated with long-term reduction of health services utilization would provide solid evidence of the cost-effectiveness of treatment. Such evidence, in turn, might increase reimbursement from third-party payers for insomnia therapies.

In summary, it is essential to select dependent variables that will reflect on outcome beyond the simple reduction of insomnia symptoms. As insomnia is associated with significant morbidity, an effective treatment should not only improve sleep, but also produce clinically meaningful changes in daytime functioning, mood, and quality of life.

SELECTING ASSESSMENT INSTRUMENTS TO MEASURE OUTCOMES

Once relevant outcomes are identified, the next step involves selecting adequate instruments to measure those outcome variables. Several psychometric and practical considerations should guide the selection of assessment instruments [47]. The first criterion should be the basic psychometric properties of an instrument, such as its reliability (internal consistency, temporal stability) and validity (e.g. construct, discriminant, convergent, predictive). The construct validity is particularly important to ensure that an instrument truly measures what it is intended to measure. Another important feature is its sensitivity to detect changes produced by an intervention. The instrument must be sensitive enough to detect small, yet clinically meaningful changes. The range of possible scores should be large enough to avoid a “ceiling effect” or a “floor effect”. It is also important to distinguish instruments that are primarily designed to measure outcome
from diagnostic instruments (e.g. Structured Sleep Disorders Interview) or from those intended for measuring specific features of insomnia (e.g. Pre-Sleep Arousal scale). Finally, there is also a very practical consideration that assessment instruments should be as brief as possible to minimize burden on patients.

Unlike several other medical (e.g. hypertension) or psychological disorders (e.g. depression), which can be evaluated with only one assessment modality, insomnia can be defined according to multiple measurement systems (i.e. subjective, behavioral, physiological). The evaluation of treatment outcome can then rely on subjective (self-report questionnaires, sleep diary evaluation of treatment outcome) and objective measurements (polysomnography). Despite the inevitable divergence among those assessment modalities, they have the significant advantage of providing multiple perspectives on treatment outcome.

Polysomnography (PSG). PSG recording is the gold standard for assessing sleep in clinical trials of hypnotic drugs [16–18, 20] and it is increasingly used in behavioral studies as well [10, 13, 21, 22]. The use of PSG assessment for insomnia is still controversial in clinical practice. Although it is not indicated for the routine clinical evaluation of insomnia [5, 48], PSG is essential in clinical trials, at least to screen for other sleep disorders and, ideally, as a repeated measure before and after treatment. Aside from its costs and burden on subjects, two important issues must be considered when using PSG, the number of recording nights and the recording site (laboratory versus home). There is currently no accepted research standard on how many nights of recordings are necessary to obtain data that are reliable and representative of a typical night’s sleep at home. This issue is particularly difficult to resolve because of the extensive variability that characterizes the sleep patterns of individuals with insomnia. Although it may take up to one week of recording to achieve some stability in sleep patterns [49], it is simply not feasible in most clinical trials to have that many PSG recordings given its cost, inconvenience, and burden on subjects. For practical purpose, it appears that two consecutive nights of recording (three at baseline to allow for an adaptation night) represent a cost-effective compromise to evaluate outcome. Pooling data over two consecutive nights of recording may reduce extreme data due to variability, and attenuate the “first night effect” or, the more common “reverse first night effect” among insomniacs. Another important consideration is whether PSG recording should be conducted in the laboratory or at home. While the sleep laboratory enhances standardization of assessment procedures, the use of ambulatory, home-based, PSG monitoring is more convenient for subjects and can attenuate their reactivity to laboratory PSG [50]. Despite its significant advantage of providing objective measures of sleep parameters, PSG does not seize all the different facets of clinical insomnia. For this reason, additional subjective measures are essential to capture those more subjective and experiential dimensions.

Daily sleep diaries. A daily sleep diary is the most widely used instrument to document outcome in insomnia treatment studies. A typical sleep diary includes entries for bedtime, arising time, quantitative estimates of sleep latency, number and duration of awakenings, total sleep time, and sleep quality ratings. Despite significant discrepancies between subjective and PSG measures of sleep, the diary reflects on the subjective complaint and perception of sleep and allows for prospective self-monitoring over extensive periods of time in the patient’s home. Daily morning estimates of sleep onset latency and time awake after sleep onset yield a reliable and valid index of insomnia, even though they do not reflect the absolute PSG values. A critical issue with sleep diary assessment is to ensure that the monitoring period is long enough to establish baseline severity and to estimate accurately changes with treatment. Because of significant nighttime-to-night variability in the sleep patterns of chronic insomniacs, it is recommended to obtain baseline data for at least one week, preferably two, and to continue monitoring sleep throughout treatment and during periodic follow-ups. When combining sleep diary and PSG assessment, it is important not to rely exclusively on morning questionnaires completed in the laboratory for the subjective data as those data may not be very representative of a typical night’s sleep at home. Several types of sleep diaries are available including one version with quantitative estimates of sleep parameters [29], one with visual analog scales [51], one with nominal scales [52], one that has been validated with pain patients [53], and one computerized diary [54].

Wrist actigraphy. The use of actigraphy as an outcome measure in treatment studies [9, 11] is appealing because of its objective and unobtrusive nature. This device provides reasonably accurate estimates of global sleep parameters (e.g. total sleep time, total wake time, and time in bed) relative to polysomnography [55–57]. However, there are important discrepancies both between actigraphy and PSG, and between actigraphy and sleep diary data, particularly when estimating more discrete-specific sleep parameters (e.g. sleep latency, time awake after sleep onset).
More troublesome, those discrepancies are not always in the same direction. Some investigators [58] have reported that actigraphy underestimate time awake and overestimate time asleep and sleep efficiency relative to polysomnography, whereas others have reported opposite relationships [57]. Thus, actigraphy should be considered as a complementary measure of treatment outcome, not as a substitute for either daily sleep diaries or polysomnography.

Symptom questionnaires. Because of their ease of administration and scoring, self-report questionnaires are widely used to document outcome in insomnia treatment studies (see Table 1). Although there are numerous questionnaires to measure various clinical features of insomnia, only a few instruments have been specifically designed and validated for measuring treatment outcome. The Pittsburgh Sleep Quality Index (PSQI; [59]) is a self-rating scale intended to measure general sleep disturbances. It is composed of 19 self-rated items assessing sleep quality and disturbances. The PSQI has excellent psychometric properties, with a cut-off score of five achieving maximum sensitivity/specificity for primary insomnia. Although this instrument was not designed originally for insomnia patients specifically, it has been used as a primary or secondary end point in several clinical trials of insomnia [8, 9, 11].

The Insomnia Severity Index (ISI; [29, 60]) is a brief 7-item scale assessing, on a 0–4 point scale, the perceived severity of insomnia (initial, middle, late), and the degree of dissatisfaction with sleep, interference with daily functioning, noticeability of impairment, and distress caused by the sleep disturbances. The total ISI score ranges from 0–28, with a score of 14 providing the best cut-off to optimize sensitivity and specificity [60, 61]. Parallel versions can also be completed by clinicians and significant others to provide collateral validation of treatment outcome [22].

In the Athens Insomnia Scale [62] was recently developed in accordance with the ICD-10 diagnostic criteria for insomnia. It is an 8-item scale evaluating the severity of insomnia as well as its impact on daytime well being and functional capacity. Several additional instruments including the Insomnia Symptom Questionnaire [63], Leeds Sleep Evaluation Questionnaire [64], the St Mary’s Hospital Sleep Questionnaire [65], and the Sleep Problems Scale [66] have been used in selected treatment studies. In addition, the Clinical Global Impression scale is a generic outcome measure to assess global clinical status. It contains a single item with four possible responses: 0 (Unchanged or worse), 1 (Minimal — slight improvement), 2 (Moderate — decided improvement), and 3 (Marked, vast improvement). Frequently used in hypnotic trials, this instrument is useful to obtain a global assessment at repeated intervals, but its validity in open clinical trials is limited because it is completed by the treating clinician rather than by an independent assessor.

There are several additional self-report measures of insomnia, including the Sleep Disturbance Questionnaire (SDQ; [67, 68]), Dysfunctional Beliefs and Attitudes about Sleep (DBAS; [29]), Pre-Sleep Arousal Scale [69], and the Sleep Hygiene Awareness and Practice Scale [70]. Those measures are better suited to examine process-outcome relationships than to document outcome per se. For example, the DBAS is a 30-item self-report scale assessing sleep-related beliefs and attitudes that are instrumental in maintaining sleep difficulties. Two recent studies have reported significant correlations between sleep improvements and reductions of dysfunctional sleep cognitions on the DBAS [10, 71]. Likewise, the SDQ contains 12 items related to etiological factors of insomnia (e.g. intrusive cognitions, physical tension, sleep anticipatory anxiety, and sleep incompatible activities), and can be used to document the relationships between changes in those factors and sleep improvements.

Psychological and fatigue measures. Most behavioral treatment studies of insomnia have incorporated measures of psychological symptoms (e.g. anxiety and depression), with the Beck Depression Inventory [72] and the State-Trait Anxiety Inventory [73] being the two instruments most frequently used (see Table 1). The Profile of Mood States [74] has also been used in hypnotic trials. Those instruments provide valuable data about the intensity of psychological symptoms and are potentially useful for clinical correlation with improvements of sleep patterns [5, 29]. As mentioned earlier, however, one problem is that patients with significant symptomatology of anxiety and depression are usually excluded from studies of primary insomnia, leaving little room for detecting improvements on those scales. The Fatigue Severity Scale [75] and the Multidimensional Fatigue Scale [76] have been used in recent treatment studies of primary [13] and secondary insomnia [77]. It is surprising that such measures have not been used more frequently given that fatigue is one of the most frequent functional impairments associated with insomnia. A few studies [11, 13] have also incorporated measures of sleepiness, but with the limited evidence of objective daytime sleepiness in primary insomnia [78], it is difficult to detect meaningful changes associated with treatment. Measures of sleepiness might be more relevant to demonstrate the
absence of residual daytime sleepiness associated with hypnotic medications or with sleep restriction therapy.

Psychological measures. Insomnia patients usually complain that sleep disturbances impair their abilities to think and to function during the day. Thus, it is relevant to include measures of attention and concentration, information processing speed, reaction time, and motor coordination. Instruments assessing those functions include the Digit Span, Wechsler Memory Scale, Simple Reaction Time, Continuous Performance Test, Digit Symbol Substitution, Trail Making Test, etc. [5]. Although such measures have traditionally been used to show that hypnotic drugs do not produce daytime residual effects [18, 20], recent behavioral studies have also incorporated those measures to evaluate whether non-drug treatment for insomnia improved daytime functioning [6, 21]. A difficult issue, however, is that current evidence suggests mild and selective deficits associated with insomnia at baseline [31, 79, 80], thus leaving little room for showing improvements with treatment. It will be important in the future to select measures that are not only reliable and valid, but that are also sensitive to detect subtle neurobehavioral deficits associated with insomnia.

Quality of life and functional status. Although there is increasing attention to evaluating quality of life as part of outcome assessment, the definition of QOL varies extensively across disorders and studies [39]. Not surprisingly, there are literally dozens of different QOL measures, including illness-specific and more generic QOL measures designed to be used across diseases/populations. While it is beyond the scope of this paper to review those scales (for a review, see [39, 40]), it is worth mentioning that the most frequently used QOL measures in sleep disorders research (mostly with sleep apnea and narcolepsy) include the Short-Form Health Survey of the Medical Outcomes Study Questionnaire (SF-36; [81]), the Sickness Impact Profile scale (SIP; [82]), and the World Health Organization Quality of Life (WHOQOL; [83]) (see [43]). The SF-36 is divided into eight domains representing different aspects of health status; a score for each domain can be derived as well as a global score for physical and mental health functioning. Significant differences on the SF-36 have been reported between poor and good sleepers [27, 28], but the scale has not yet been used to document changes over time with treatment. As for most QOL instruments, the SF-36 was originally designed and validated for more severe health problems (chronic pain, COPD) and it is unclear whether it would be sensitive enough to detect changes related to the more subtle effects of insomnia. Another instrument, combining sleep and psychological well-being questionnaires, has been used in one multicenter study of zopiclone as a measure of quality of life [84]. There is a definite need to develop prospectively a QOL measure that would be specific to insomnia.

In summary, no single assessment instrument can capture all dimensions of insomnia treatment outcome. There are well-documented discrepancies between subjective and objective measures of sleep and daytime functions and such discrepancies are not unique to the insomnia literature. Indeed, measures of different psychological constructs (e.g., anxiety and depression) obtained from the same source (subjective report) are more strongly correlated with each other than are measures of the same construct (i.e., anxiety) obtained from different assessment modalities (self-report versus clinician evaluations). For this reason, it is essential to document treatment efficacy with multiple outcomes and multiple assessment modalities. Investigators should select instruments that do not overlap too much with each other and that provide complementary rather than repetitive information.

INTERPRETING AND REPORTING TREATMENT OUTCOMES

In addition to reporting statistical tests for group differences and effect sizes, several clinical indicators can be used to interpret and document treatment outcomes. Those criteria may include improvement rates on the primary dependent variable, the proportion of patients achieving remission or reaching some pre-determined criteria, as well as several other indices of the clinical significance or practical impact of sleep improvements in the patient’s everyday life.

Statistical significance and effect size. Statistical significance is still the gold standard criterion to define treatment efficacy. The most classic method for measuring outcome is to compare means or proportions on the primary dependent variable for an experimental and a control condition and to conduct standard tests of statistical significance. When treatment produces a change from baseline to posttreatment that is statistically greater (P < 0.05 or 0.01) than that obtained in a control condition, it is declared significantly effective. Because such statistical tests provide no information about the magnitude of
change, the effect size is another index that is increasingly reported in clinical trials. There are several methods for generating an effect size but the most common one is based on the difference between the posttreatment means of the experimental and control groups divided by the pooled standard deviation of both groups. The resulting $d$ statistic is expressed as a standard deviation unit. By convention, an effect size of 0.2 is considered small, one of 0.5 is medium, and one of 0.8 is large [85]. An effect size of 0.5 would indicate that the average treated patient has moved one half standard deviation above (or below) the mean of the untreated control subject on the dependent variable of interest. Results from meta-analyses indicate that behavioral and pharmacological therapies produce effect sizes falling in the medium to large range for sleep continuity and sleep quality parameters, with an advantage for behavioral therapies on sleep onset latency and for pharmacotherapy on total sleep time [24, 86–88]. Although those effect sizes are considered reliable and statistically significant, they provide no information about the absolute change on the dependent variable of interest (i.e. sleep onset latency or wake after sleep onset). As such, they are not very useful to inform clinicians and their patients about what to expect in terms of treatment outcome. For this reason, it is essential for investigators to report also absolute changes in means from baseline to posttreatment. For example, evidence from meta-analyses indicate that insomnia therapies reduce sleep onset latency from an average of 60–65 min at baseline to about 30–35 min at posttreatment, for a global improvement rate of about 50%. Likewise, the average gain in total sleep time, although statistically significant, rarely exceeds 30–45 min with either behavioral or drug therapy. Another issue is that it may be useful to consider also changes in standard deviations of key sleep parameters rather than focusing only on group means. Given the extensive night-to-night variability that characterizes insomnia, such changes would represent a clinically useful marker of sleep improvements [67].

Clinical significance. Although insomnia treatments may produce statistically significant and reliable changes on sleep parameters, a more critical issue is whether such changes make a real difference in the patient’s everyday life. The clinical significance of outcome refers to the extent to which changes in the patient’s status, though statistically significant, are clinically meaningful. Several indicators can be used to gauge clinical significance [22, 25, 89–91]. For instance, does sleep improvement (a 30-min reduction of SOL or a 45-min increase of TST) produce meaningful changes in daytime functioning and subjective well being? Does treatment return a patient within the normal population (i.e. good sleepers)? For example, a hypnotic drug may produce a 20-min reduction of sleep onset latency and be statistically greater ($P < 0.01$) than a 5-min reduction obtained by a placebo control condition. Such improvement, however, may be of little practical or clinical value for the patient. Likewise, behavior therapy may reduce the amount of time awake after sleep onset from 90 min at baseline to 45 min after treatment. Although this 50% improvement rate may be statistically significant, the average treated patient would still meet criteria for insomnia after treatment. Finally, combined therapy might increase sleep efficiency from 75 to 85%, but does it also improve daytime energy, mood, and mental abilities? It is not enough to prove that a treatment is statistically superior to a control condition, or that it produces an improvement rate greater than 50%, it is also essential to consider both the clinical importance of those changes and the end-state functioning of patients as additional indicators of a successful outcome.

Despite the lack of consensus on what should be considered an optimal outcome in insomnia treatment studies, several methods have proved useful to document the clinical significance of outcomes in the psychological treatment literature (see [92, 93]). Those methods are based on normative comparisons, the computation of a reliable change index, collateral evaluations from significant others or clinicians, and measures of treatment acceptability and patients’ satisfaction. The use of normative comparisons is the most appealing method to assess clinical significance. The extent to which treatment can bring an insomnia patient within the distribution of a good sleepers group would provide convincing evidence that outcome was clinically meaningful. Several criteria can be used to evaluate clinical significance depending on the availability of normative data. One requires that the posttreatment score of a given patient on the primary dependent variable moves by some pre-determined criteria (usually two standard deviations) away from the baseline mean of the dysfunctional (i.e. insomniac) group, or comes closer to the mean of the normative group than to the mean of the poor sleepers group at the end of treatment. Although these indices of change have been used extensively in psychological studies [94], they often require normative data or knowledge about a group mean towards which a patient’s score is expected to regress. Unfortunately, normative data
from healthy comparison groups drawn from a comparable population are frequently unavailable.

In the absence of normative data, the end-state functioning or clinical status of a patient after treatment is another useful indicator of clinical significance. This method consists of reporting the proportion of patients who fall below a critical threshold used to define clinical insomnia. This threshold may be based on the presence/absence of an insomnia diagnosis (according to DSM-IV or ICSD criteria), on standard criteria used for patient selection purpose (i.e. SOL and WASO < 30 min, SE > 80–85%), or on cut-off scores of validated instruments to detect insomnia (e.g. score below 14 on the ISI). Although such dichotomous end-points provide useful complementary outcome data, a difficult issue that arises is whether those proportions should be based on the number of patients who initiate (intent-to-treat) or those who complete treatment (completers). For example, of 100 patients enrolled in a clinical trial, 60 patients achieved adequate treatment response, 20 patients failed to respond, and 20 patients dropped out before the end of treatment for various reasons (side effects, initiate another treatment, lost interest). Should the reported ratio of treatment responders be 60% (60/100) or 75% (60/80)? Although the intent-to-treat approach is currently the preferred method for reporting outcome [95], it may be too conservative for estimating the true treatment effects because, even in clinical practice, some patients discontinue treatment prematurely for a variety of reasons that may or may not be related to the intervention.

Other indices of a clinically meaningful treatment response may involve collateral evaluations from significant others or clinicians [22]. The assumption is that changes noticeable by a significant other (usually a spouse) provide additional evidence of treatment efficacy. Other factors that could be used to validate the clinical significance of outcome include treatment acceptability or preference [96–98], patient satisfaction with treatment, duration of change, the perceived need for additional treatment, the degree of emotional distress, and (for behavioral treatment studies) the use of hypnotic medications [99]. Regardless of how effective treatment is, if it isn’t acceptable by prospective patients it is unlikely to be very useful. From a public health perspective, there is also a need to document the cost-benefits and cost-effectiveness of different insomnia interventions; despite some preliminary retrospective data on this issue, it has not been prospectively evaluated in insomnia clinical trials.

WHEN AND HOW OFTEN SHOULD WE MEASURE OUTCOMES?

Repeated assessment of symptoms and clinical status is a sine qua non condition of randomized controlled trials. The issue of when and how frequently outcome should be measured, however, is not always straightforward [100]. Current practices vary extensively across clinical trials of behavioral and drug therapies (see Table 1). Hypnotic studies focus on the evaluation of acute treatment effects and withdrawal effects (or lack thereof) upon drug discontinuation. Outcome evaluation is often restricted to very brief assessment intervals, including 2–3 baseline nights, followed by a few additional nights while on medication, and a few more upon drug discontinuation. Such practices are essentially dictated by FDA regulations and the fact that hypnotics are approved only for short-term use; there may be little commercial incentive at examining their long-term impact although, from a clinical and scientific perspective, such studies are very much needed. In a recent review of clinical trials of benzodiazepines and zolpidem [87], the median treatment duration was only 1 week (range: 4–35 days), and follow-ups were virtually absent. By comparison, the average duration of behavioral therapies is 5 weeks [24] and, in about 50% of the studies, additional follow-ups are conducted on average six months after treatment. Interim assessment during the course of treatment would be important to study the speed and trajectory of therapeutic response.

The frequency, timing, and intervals between outcome assessments should be based on several factors including the research questions, the nature of the treatments under investigation, as well as the expected speed of therapeutic response and durability of clinical benefits over time. Ideally, assessment should be conducted before initiating treatment, during the course of therapy, and on several occasions after treatment completion [78]. Baseline assessment is essential to establish initial insomnia severity and posttreatment evaluation is essential to evaluate changes with the intervention. Monitoring patient’s status throughout the course of therapy is also useful to determine the speed of therapeutic response, a critical factor influencing whether a patient will comply with or drop out of treatment. Repeated assessment is also important to document process-outcome relationships. Despite evidence of treatment efficacy, there is little information about the mechanisms of changes and
about why treatment works for some patients but not for others. For instance, the role of hyperarousal in insomnia is well documented [44, 45], yet there is still little evidence that reduction of arousal with relaxation therapy or hypnotics lead to sleep improvements. On the other hand, there is evidence that changes in beliefs and attitudes about sleep is an important mediating factor of successful outcome in insomnia therapy [10, 71]. Such evidence is based on correlations between pre to posttreatment change scores in sleep parameters and in beliefs and attitudes. Repeated assessment of those variables throughout the course of therapy would allow for using more sophisticated statistical time series analyses and gain a better understanding of mechanisms of changes.

Follow-up assessment is essential to evaluate whether initial treatment benefits are sustained over time [11, 22]. Considering that patients entering treatment have often experienced sleep difficulties for more than 10 years, it is not enough to demonstrate that treatment is effective for one night, one week, or even one month. Also, because insomnia is a fairly recurrent condition, even those who benefit from initial treatment may remain vulnerable to episodes of insomnia in the long-term. What should be the ideal number and timing of follow-up assessments? Although repeated and long-term follow-ups are desirable, there is always an increasing risk of attrition over time. Life events, unrelated to a study, may confound interpretation of long-term outcomes. For example, the longer the interval is between the end of treatment and follow-up assessment, and the more likely a patient may have experienced major life events (e.g., divorce, death in the family), initiated another treatment (e.g., herbal supplements), or simply no longer be interested or available to complete assessment.

To summarize, in addition to evaluating the magnitude of sleep improvements, it is equally important to document the trajectory of changes with repeated assessments over time. Some measures (sleep diaries) should be collected on an ongoing basis, others can be administered at weekly intervals (self-report questionnaires), and still others (PSG) are needed at strategic points such as baseline, posttreatment, and follow-up.

**RECOMMENDATIONS AND CONCLUSIONS**

Significant advances have been made in the evaluation and treatment of insomnia in the past decade, yet there is still no standard method to document outcomes or criteria to define what is a successful outcome. Further progress in the development of evidence-based guidelines will continue to depend upon the conduct of high-quality randomized clinical trials, and a critical component of such studies will remain the selection of reliable, valid, conceptually sound, and clinically relevant outcome measures. There is a need to develop a core battery of standardized instruments, including a common core of assessment instruments (e.g., PSG and sleep diaries) for all clinical trials and optional modules for use with specific treatments and populations. Those measures must meet minimal psychometric requirements (reliability and validity), they must be sensitive to changes, and practical to use in research studies and in clinical practice. Table 2 provides a summary of outcome variables and assessment instruments that could be used for such purposes. It is not an exhaustive list but rather a sample of measures that have been used in treatment outcome studies.

There is a need to move beyond the exclusive focus on insomnia symptoms and to broaden the scope of outcome assessment. It is essential to target multiple areas/domains including sleep (symptoms, diagnosis), functional impairment, psychological and biological variables, quality of life/subjective well being, and cost-benefit/cost-effectiveness. Investigators need to incorporate assessments from multiple sources including patients, clinicians, and significant others (e.g., spouses). It is necessary to collect both subjective and objective data on sleep, as well as on daytime functioning. Further research is needed to identify reliable markers of CNS hyperarousal in insomnia such that researchers can use such biological indices as outcome measures. It is also critical to evaluate outcome beyond the initial treatment phase. More attention should be paid to the trajectory of change over time both in terms of speed of recovery and durability of clinical benefits. It would be useful to include not only outcome but also process measures in order to study the mechanisms and attribution of changes, and to refine our treatment protocols. These process measures may be quite different for behavioral and drug therapies. Some standardization across studies would also be highly desirable in the timing of assessment at baseline, posttreatment, and follow-ups. Finally, it would be useful to incorporate measures of treatment preference/acceptability and patients' satisfaction as such measures may play an important role in determining who will seek, initiate, and complete treatment.
### Table 2. Potential outcome variables and assessment instruments for insomnia treatment studies

<table>
<thead>
<tr>
<th>Outcome variables</th>
<th>Assessment instruments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sleep/wake parameters</td>
<td>Polysomnography</td>
</tr>
<tr>
<td></td>
<td>Sleep diary</td>
</tr>
<tr>
<td></td>
<td>Actigraphy</td>
</tr>
<tr>
<td>Insomnia symptoms</td>
<td>Insomnia severity index [29, 60]</td>
</tr>
<tr>
<td></td>
<td>Pittsburgh sleep quality index [59]</td>
</tr>
<tr>
<td></td>
<td>Insomnia Athens scale [62]</td>
</tr>
<tr>
<td>Daytime functioning</td>
<td>Fatigue severity scale [75]</td>
</tr>
<tr>
<td>Fatigue</td>
<td>Multidimensional fatigue inventory [76]</td>
</tr>
<tr>
<td>Performance</td>
<td>Reaction time (simple/choice reaction time; continuous performance test)</td>
</tr>
<tr>
<td></td>
<td>Attention (digit span; divided attention test)</td>
</tr>
<tr>
<td></td>
<td>Coordination (digit symbol substitution, finger tapping)</td>
</tr>
<tr>
<td></td>
<td>Auditory/visual vigilance</td>
</tr>
<tr>
<td></td>
<td>Memory (word list, figures)</td>
</tr>
<tr>
<td>Psychological symptoms</td>
<td>Beck depression inventory [72]</td>
</tr>
<tr>
<td></td>
<td>Beck anxiety inventory</td>
</tr>
<tr>
<td></td>
<td>State-trait anxiety inventory [73]</td>
</tr>
<tr>
<td></td>
<td>Profile of mood states [74]</td>
</tr>
<tr>
<td>QOL</td>
<td>SF-36 [81], Sickness impact profile [82], WHOQOL [83]</td>
</tr>
<tr>
<td>Clinical utility</td>
<td>Treatment satisfaction/preference</td>
</tr>
<tr>
<td>Cost/effectiveness</td>
<td>Health care utilization</td>
</tr>
<tr>
<td>Global assessment</td>
<td>Clinical global impression scale</td>
</tr>
<tr>
<td>Process outcome measures</td>
<td>Beliefs and attitudes about sleep scale [29]</td>
</tr>
<tr>
<td></td>
<td>Insomnia symptom questionnaire [63]</td>
</tr>
<tr>
<td></td>
<td>Sleep disturbance questionnaire [67, 68]</td>
</tr>
<tr>
<td></td>
<td>Pre-sleep arousal scale [69]</td>
</tr>
</tbody>
</table>

### Practice Points

- Clinicians should evaluate treatment outcome with daily sleep diaries and selected self-report questionnaires targeting sleep/insomnia symptoms, psychological and fatigue symptoms, and more global measures of treatment satisfaction/acceptability.
- Outcome measures must be reliable, valid, sensitive to changes, conceptually sound and clinical relevant. It is essential to select a broad range of variables including sleep, daytime functioning, mood, and quality of life.
- Normative data on sleep and other outcome variables would be extremely useful to determine the degree of sleep impairments at baseline and the end state functioning or clinical status of patients who have completed insomnia treatment.
- Despite the critical need for outcome evaluation, care should be taken to limit the burden on research participants enrolled in clinical trials and on patients seen in clinical practice.

### Research Agenda

- A consensus conference is needed among insomnia investigators and practitioners to define what should be an optimal outcome when treating insomnia and how we should measure such outcome.
- Treatment outcome evaluation needs to move beyond the exclusive focus on insomnia symptoms. It is essential to broaden the scope of outcome assessment with additional indicators of
success taking into consideration the practical as well as statistical significance of changes.

- It is necessary to target multiple areas/domains including sleep, functional impairment, psychological/subjective well being, and quality of life. Evaluation of the cost-benefits and cost-effectiveness of insomnia therapies should also be incorporated to the evaluation of treatment outcome.
- It is critical to evaluate outcome beyond the acute treatment phase. More attention should focus on the trajectory of changes over time both in terms of speed of recovery and durability of therapeutic benefits. For this, there should be a standard protocol that would define the timing of baseline, posttreatment, and follow-up assessments.
- There is a need for further prospective research to develop and validate a core battery of standardized instruments to document outcomes in randomized clinical trials of behavioral and pharmacological therapies of insomnia. This battery needs to incorporate assessment from multiple sources including patients and clinicians.

ACKNOWLEDGEMENTS

Preparation of this article was supported in part by grants from the National Institute of Mental Health (Grant #60413) and the Canadian Institutes of Health Research (Grant#MT42504). The author is grateful to Colin A. Espie for his comments and to Geneviève Belleville for her assistance in preparing an earlier draft of this manuscript.

REFERENCES


88. Smith MT, Perlis ML, Park A, Smith MS, Pennington J, Giles DE, Buysse DJ. Comparative meta-analysis of


