Critical Review

Recommendations for the Design and Analysis of Treatment Trials for Alcohol Use Disorders

Katie Witkiewitz, John W. Finney, Alex H.S. Harris, Daniel R. Kivlahan, and Henry R. Kranzler

Background: Over the past 60 years, the view that “alcoholism” is a disease for which the only acceptable goal of treatment is abstinence has given way to the recognition that alcohol use disorders (AUDs) occur on a continuum of severity, for which a variety of treatment options are appropriate. However, because the available treatments for AUDs are not effective for everyone, more research is needed to develop novel and more efficacious treatments to address the range of AUD severity in diverse populations. Here we offer recommendations for the design and analysis of alcohol treatment trials, with a specific focus on the careful conduct of randomized clinical trials of medications and non-pharmacological interventions for AUDs.

Methods: This paper provides a narrative review of the quality of published clinical trials and recommendations for the optimal design and analysis of treatment trials for AUDs.

Results: Despite considerable improvements in the design of alcohol clinical trials over the past 2 decades, many studies of AUD treatments have used faulty design features and statistical methods that are known to produce biased estimates of treatment efficacy.

Conclusions: The published statistical and methodological literatures provide clear guidance on methods to improve clinical trial design and analysis. Consistent use of state-of-the-art design features and analytic approaches will enhance the internal and external validity of treatment trials for AUDs across the spectrum of severity. The ultimate result of this attention to methodological rigor is that better treatment options will be identified for patients with an AUD.

Key Words: Alcohol Randomized Trials, Research Design, Alcohol Use Disorder, Eligibility Criteria, Data Analysis, Missing Data Approaches.

The public health and societal costs of alcohol use disorders (AUDs) are considerable (U.S. Burden of Disease Collaborators, 2013). Alcohol use is the third leading global contributor to the burden of disease as measured in disability-adjusted life years (World Health Organization, 2009, 2011). The estimated cost of excessive alcohol consumption in the United States was $223.5 billion in 2006 dollars (Bouchery et al., 2011). Yet, despite social and economic costs of heavy drinking, many individuals with AUDs never receive treatment (Substance Abuse and Mental Health Services Administration, 2014). For individuals with an AUD who seek care, relatively few efficacious treatments are available (Jonas et al., 2014a; National Institute for Health and Care Excellence, 2011; Zindel and Kranzler, 2014).

Currently, only 3 AUD medications are approved by the Food and Drug Administration (FDA): disulfiram, acamprosate, and naltrexone (including oral and long-acting injectable formulations), and 1 additional AUD medication, nalmefene, is approved by the European Medicines Agency (Mann et al., 2013). For a handful of other drugs that are not approved to treat AUDs (e.g., gabapentin, ondansetron, varenicline, topiramate), there is a varying amount of evidence of efficacy (Blodgett et al., 2014; Jonas et al., 2014a,b; Zindel and Kranzler, 2014). The need for additional development of medications has been voiced by a collaboration of researchers, and representatives of the pharmaceutical industry, the FDA, and the National Institute on Alcohol Abuse and Alcoholism, who have collectively formed the Alcohol Clinical Trials Initiative (ACTIVE; Anton et al., 2012).

Nonpharmacological treatments for AUD are also limited,
with good evidence of efficacy for fewer than a dozen behavioral treatments (Miller and Wilbourne, 2002; National Institute for Health and Care Excellence, 2011). Most of the efficacious treatments yield small-to-moderate effect sizes, which may contribute to the fact that, in general, initiation and engagement in any of these treatment options is suboptimal (Brozson et al., 2013; Harris et al., 2009a, 2012). The low rates of utilization of treatments for AUDs and the modest efficacy of existing interventions underscore the need to develop and test new treatments that are more efficacious and attractive to patients and clinicians, including those that take into account the heterogeneity of AUDs (Anton et al., 2012; Lane and Sher, 2014; Litten et al., 2015).

To enhance the validity of future AUD treatment trials, we review key elements of research design and analysis in both pharmacological and nonpharmacological alcohol treatment studies. Unlike prior reviews that covered design and analysis considerations for a wide variety of clinical conditions (Berger, 2006; Chu et al., 2011; Freemantle, 2001; Harris et al., 2009b) or that focused on limited design questions in alcohol treatment studies (e.g., intent-to-treat approaches [Del Re et al., 2013]; randomization [Hedden et al., 2006]), the current review provides a broad review of recommendations that are comprehensive, but specific to alcohol treatment studies. We focus on 4 areas: (i) considerations for recruitment, randomization, and retention; (ii) measurement selection for outcome assessment, intervention fidelity, and adherence monitoring; (iii) timing of assessments and considerations for studying mechanisms of behavior change; and (iv) statistical methods used to assess treatment efficacy.

Improvements in these areas are critical to advancing the methodological quality of alcohol clinical trials and the assessment of the reliability, validity, and scope of inference of treatment effects. Also, no matter how well studies are designed and conducted, it is critical to improve the quality of reporting of findings from AUD treatment trials. Witkiewitz and colleagues (2015) make recommendations for improving the reporting of study features in 4 areas of relevance to alcohol treatment studies: (i) trial registration, (ii) procedures for recruitment and retention, (iii) procedures for randomization and intervention design, and (iv) statistical methods used to assess treatment efficacy.

Eligibility Criteria

The selection of participants for alcohol clinical trials is often biased by the exclusion of individuals with residential instability, low motivation, and co-occurring psychiatric, medical, or drug use disorders (Hoertel et al., 2014; Humphreys and Weisner, 2000; Humphreys et al., 2008; Rychtarik et al., 1998). Although exclusion criteria are often used in clinical trials to reduce within-group variance and thereby improve statistical power to detect clinically meaningful effects between groups (Food and Drug Administration, 1998), the application of exclusion criteria may also unintentionally impact the estimation of treatment effects intended to generalize to a less restricted population of treatment-seeking patients (Humphreys et al., 2008). The application of extensive exclusion criteria can reduce the external validity of clinical trial findings and contribute to the gap between clinical research and clinical practice (Kazdin, 2008; Newham and Page, 2010).

Humphreys and colleagues (2008) found that commonly used eligibility criteria substantially influenced estimates of treatment effects in 25 treatment programs across the United States. The direction (positive or negative) and degree (small to large) of influence differed across 2 samples, suggesting that the impact of eligibility criteria on treatment outcomes cannot be reliably predicted across patient populations. We recommend that investigators consider the impact of exclusion criteria and document the reasons for excluding research participants (see Witkiewitz et al., 2015). Recommendations for the design of pragmatic trials, which emphasize real-world applications to inform intervention delivery, could also be consulted to increase applicability of clinical trial findings to real-world settings (Loudon et al., 2015).

Randomization

Although the randomized clinical trial (RCT) is the “gold standard” design to evaluate treatment efficacy, it is not fail-safe. Thus, investigators need carefully to consider the choice of randomization procedures and assess whether randomization was successful. Hedden and colleagues (2006) reviewed several randomization options available for treatment trials for substance use disorders, the features of which are summarized in Table 1. The appropriate selection of a particular randomization scheme often depends on the size of the sample, potential imbalances that might occur across treatment and control groups, and a priori ideas about other variables (e.g., demographics, genotype) that might impact the effectiveness of the intervention (Kranzler et al., 2011).

The stratified permuted block randomization procedure is the most commonly used and recommended approach to adjust for covariate imbalances; however, it is critical that the covariates used in stratification are included in the analysis and considered when conducting power analyses (Matts and Lachin, 1988). The use of permuted block and other covariate-adjusted randomization schemes can help to avoid
differences in baseline characteristics across intervention groups, which can occur in the context of randomization failures that occur either systematically or by chance (Berger, 2005a). When baseline differences across groups are observed after randomization, propensity score or instrumental variable methods may be useful to estimate the degree of influence that the baseline characteristics have on treatment outcomes and to potentially reduce bias in treatment effect estimates (Berger, 2005b; Leyrat et al., 2013; Merkx et al., 2014). Propensity score or instrumental variable methods also can be useful when randomization is not feasible (Humphreys et al., 2014; Magura et al., 2013; Ye and Kaskutas, 2009). Methodological development in the application of instrumental variables and propensity score approaches is an area of active ongoing research (Austin, 2013; McCaffrey et al., 2013; Myers et al., 2011). As with any statistical technique, these methods can introduce bias if implemented incorrectly or if the instrumental variables are not well selected (Austin, 2008; Myers et al., 2011). Thus, investigators are encouraged to consult the most recent statistical literature when applying these methods. Sensitivity analyses, in which the outcomes are examined with and without the inclusion of covariates, and with and without the use of instrumental variables and/or propensity scores, can be used to gain a better understanding of the influence of these approaches on treatment effects.

**Study Retention and Procedures to Minimize Missing Data**

It is often difficult to retain participants in alcohol clinical trials (Kranzler et al., 1996). Despite numerous procedures for handling missing data, described below, it is far preferable to prevent the occurrence of missing data by devoting resources to ensuring study retention. Recommendations to improve retention in medication trials include flexible dosing schedules, the continued assessment of individuals who discontinue treatment, the addition of the study treatment to an existing effective treatment, and the use of rescue medications or psychotherapy for individuals who do not respond to treatment (National Research Council, 2010; Zweben et al., 2005). However, it is important to consider the unintended effects of these strategies on generalizability to real-world patients and procedures (e.g., if the same strategies are not available in real-world settings). Additional recommendations for reducing missing data include allowing follow-ups to be conducted via the phone or Internet (instead of only in-person), providing incentives/bonuses for completing assessments, using interim contacts with participants to maintain accurate contact information, asking participants to nominate “locators” who will know how to reach them should routine efforts prove unsuccessful, having and communicating to participants a clear project identity, and emphasizing to participants and staff the importance of engagement with the treatment (e.g., medication adherence), retention in the research, and the significance of the study (Davis et al., 2002; National Research Council, 2010). To gauge treatment effects accurately in RCTs, it is essential for investigators to attempt to follow-up all individuals at every time point regardless of whether they discontinued the intervention or did not receive the allocated intervention (National Research Council, 2010). Adaptive treatment protocols, such as those implemented in sequential multiple assignment randomized trial (SMART) designs, may also improve retention and generalizability by providing an individualized approach and by increasing treatment intensity for participants who need a higher level of care (Hinshaw et al., 2004; Lei et al., 2012; McKay, 2009; Murphy et al., 2007). Recommendations for the analysis of SMART designs should be carefully considered when such designs are employed (Chakraborty and Murphy, 2014; Nahum-Shani et al., 2012).

---

**Table 1. Summary of Methods for Randomization in Randomized Clinical Trials**

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Description</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>Each successive patient has an equal probability of assignment to all treatments</td>
<td>Ease of use</td>
<td>Very high potential for imbalances; no covariate stratification</td>
</tr>
<tr>
<td>Permutated block</td>
<td>Within each successive block, patients have a random probability of assignment to each treatment</td>
<td>Ease of use; stratification on covariates</td>
<td>Potential for imbalance; need to control for stratified covariates in analyses</td>
</tr>
<tr>
<td>Urn randomization</td>
<td>Adaptive randomization scheme where probability of assignment is adjusted based on degree of treatment of imbalance</td>
<td>Reduces imbalance; stratification on covariates</td>
<td>Difficulty in implementation; need to control for stratified covariates in analyses</td>
</tr>
<tr>
<td>Covariate adaptive</td>
<td>Adaptive randomization scheme where probability of assignment is adjusted based on degree of treatment of imbalance</td>
<td>Reduces treatment imbalance and covariate imbalance</td>
<td>Difficulty in implementation; need to control for covariates in analyses; assignment is not “random”</td>
</tr>
<tr>
<td>Two-stage approach</td>
<td>Combines urn randomization with covariate adaptive randomization</td>
<td>Reduces treatment imbalance and covariate imbalance</td>
<td>Not widely used; need to control for covariates in analyses; assignment is not “random”</td>
</tr>
</tbody>
</table>
Some of the efforts to reduce attrition or to increase follow-up with individuals who have not received or completed treatment may influence outcomes. Thus, if efforts are implemented differentially across treatment groups (e.g., devoting more effort to tracking active intervention participants), then the estimation of treatment effects may also be impacted (Clifford et al., 2007; Kaminer et al., 2008; Maisto et al., 2007). Adequate blinding of research staff, particularly those involved with retention and assessment procedures, can greatly reduce the impact of differential retention efforts on the estimation of treatment effects. Nonetheless, the potential impact and burden of assessment procedures should be considered when designing the study (Gastfriend et al., 2005) and when evaluating the efficacy of the treatment and comparator (including placebo) conditions (Litten et al., 2013; Weiss et al., 2008). Secondary analysis of treatment effects with the number of assessments completed as a covariate could provide an indication of the degree to which the amount of assessment enhanced or attenuated treatment effects (Clifford and Davis, 2012; Kurtz et al., 2013).

MEASUREMENT SELECTION

Outcomes Assessment

The most common primary outcome measures used in alcohol clinical trials are the frequency and/or intensity of alcohol consumption (Finney et al., 2003; Robinson et al., 2014), most often estimated using the timeline follow-back method (TLFB; Sobell and Sobell, 1992) or the Form-90 (Miller, 1996). These instruments provide daily alcohol consumption data, which are often aggregated to create the primary outcome measure. Commonly used aggregations include both continuous and binary outcomes. Continuous outcomes include the percentage of days abstinent (PDA), drinks per drinking day, drinks per day, and the percentage of heavy drinking days (PHDD; with “heavy” defined as 4 or more drinks in a day for women and 5 or more drinks for men) (Anton and Randall, 2005). Binary outcomes include any drinking and no heavy drinking (Falk et al., 2010) and composite clinical outcomes (Cisler and Zweben, 1999). It is important to note that continuous outcomes allow for more powerful statistical tests and will often require smaller samples to detect treatment effects than binary ones (Bakhshi et al., 2012; MacCallum et al., 2002; Yoo, 2010). The selection of the primary outcome for a clinical trial may often depend on the focus of the trial (e.g., the PDA outcome may be more appropriate for abstinence-oriented treatments, while PHDD may be preferable for a moderation-focused treatment). With regard to instrument selection, we highly recommend that investigators collect drinking data on a daily level using a calendar method, such that various aggregations and analytic methods can be considered, depending on the goals of the trial (Hallgren KA, Atkins D, Witkiewitz K under review).

An important consideration in outcome selection is the identification of measures with established psychometric properties. Numerous studies have examined the reliability and validity of the TLFB and Form-90 across multiple administration methods, including comparisons between telephone and computer (Sobell et al., 1996), online and in-person (Pedersen et al., 2012), individual and group (Pedersen and LaBrie, 2006), and telephone versus self-administration (Maisto et al., 2008). The general conclusion from these studies is that both assessments, regardless of the method of administration, typically produce reliable and valid estimates of alcohol consumption in the context of clinical research (Del Boca and Darkes, 2003).

Self-reported alcohol consumption can be verified using collateral informants (i.e., proxy reporters) or alcohol biomarkers, although both approaches have limitations and currently most researchers ultimately rely on self-reported alcohol consumption for primary analyses. In general, collateral reports, such as family members, roommates, or close friends, correspond moderately with self-reported alcohol use and related problems (Donovan et al., 2004; Whitford et al., 2009), but they may be skewed in the direction of less reporting (e.g., less alcohol use, higher functioning) than self-reports (Connors and Maisto, 2003). When collateral informants’ reports are used to validate self-reported drinking (Connors and Maisto, 2003; Donohue et al., 2004), it is important that the informants be confident in their knowledge of the patient’s drinking behavior (Sobell et al., 1997). Collateral reports have also been shown to be less helpful for confirmation of self-report in some populations (i.e., college students; Laforge et al., 2005).

Biomarkers of alcohol use are also commonly used to validate self-report data, and the correspondence between the two is often quite high (Anton et al., 2006; Litten et al., 2010; UKATT Research Team, 2005). The most commonly used biomarkers, carbohydrate-deficient transferrin (CDT), gamma-glutamyltransferase, and the ratio of alanine aminotransferase to aspartate aminotransferase, are relatively inexpensive, but may be less sensitive to the level of alcohol consumed. Liver enzyme tests also have lower specificity than newer tests (Hashimoto et al., 2013; Hock et al., 2005), such as CDT and those measuring direct ethanol metabolites, including ethyl glucuronide, ethyl sulfate, and phosphatidylethanol (Hashimoto et al., 2013; Jatlow et al., 2014). The newer biomarkers show good correspondence with self-reported alcohol consumption (Crunelle et al., 2014; Hartmann et al., 2007) and may be very useful to confirm recent self-reported abstinence, although they have limited sensitivity to detect the level of alcohol consumption or to confirm nonrecent drinking via self-report (Jatlow et al., 2014), reducing their utility as primary outcome measures for alcohol clinical trials. Transdermal alcohol monitors are a promising new approach to collect real-time drinking data without reliance on self-report and have been used successfully to detect alcohol use in contingency management trials (Barnett et al., 2011; Dougherty et al., 2014). However, the
available sensors may not be accurate in detecting lower levels of drinking (Barnett et al., 2014), are costly, and may not be acceptable to patients. A range of secondary outcome measures also may be of interest, including alcohol-related problems, quality of life, psychosocial functioning, healthcare utilization, and dependence severity (Cisler et al., 2005; LoCastro et al., 2009; Tiffany et al., 2012a). However, there is very little agreement regarding the “best” measures for assessing primary or secondary outcomes (Donovan, 2012; Tiffany et al., 2012b).

The most common secondary outcome measures used in alcohol clinical trials include the Drinker Inventory of Consequences (DrInC; Miller et al., 1995); the Short Index of Problems, a shortened version of the DrInC (Feinn et al., 2003; Kiluk et al., 2013); the Addiction Severity Index (McLellan et al., 1992); measures of craving (e.g., the Obsessive Compulsive Drinking Scale [Anton, 2000]; Impaired Control Scale [Heather et al., 1993]), and other measures of functioning (e.g., Short Form Health Surveys; Ruta et al., 1993; Ware et al., 1996) or quality of life (e.g., Brief Symptom Inventory [Derogatis, 1983]; World Health Organization Quality of Life scale [World Health Organization, 1998]). Evidence supports the reliability and validity of these measures (Cacciola et al., 2007, 2011; Forcehimes et al., 2007), the utility of including nonconsumption outcomes in large-scale alcohol clinical trials (Cisler et al., 2005; LoCastro et al., 2009), and their correspondence with alcohol consumption measures (Donovan et al., 2006; Witkiewitz, 2013).

Measures of Intervention Fidelity and Medication Adherence

During the course of an intervention, it is important to monitor the fidelity of its delivery. In nonpharmacological interventions, fidelity is often measured by the training and competence of the interventionists and the degree of adherence to the intervention protocol. Ongoing supervision of interventionists and remediation of the intervention if it is not being delivered competently are critical to ensuring that minimum levels of fidelity are maintained (Carroll and Rounsaville, 2007; Miller et al., 2005). Established measures of adherence and competence (e.g., the Yale Adherence and Competence Scale; Carroll et al., 2000) can be adapted for use with newer interventions (Chawla et al., 2010). In pharmacotherapy trials, fidelity is often indexed by adherence to medication, measured using self-report, pill counts, urinary riboflavin, electronic medication event monitoring systems, or monitoring of medication blood levels. The degree of medication adherence can have an impact on alcohol treatment outcomes (Baros et al., 2007; Stout et al., 2014), as can the quality of the adherence monitoring (Swift et al., 2011). Thus, adherence rates and monitoring method should always be reported (Witkiewitz et al., 2015). Each of the adherence monitoring techniques has strengths and weaknesses. For example, monitoring medication blood levels cannot be used to monitor adherence in the placebo arm of a study. Using observation of medication ingestion to monitor adherence is not always feasible and can be thwarted by subjects hiding the medication in their cheeks or under their tongues.

TIMING OF ASSESSMENTS AND MECHANISMS OF BEHAVIOR CHANGE

The Timing of Assessments

In many alcohol clinical trials, assessments coincide with treatment visits, with 1 or multiple post treatment assessments occurring over a follow-up period. Thus, the intervention duration, which ranges from a few weeks to several months, often determines the assessment schedule during treatment. Follow-up assessments, when conducted, are often performed at 2- or 3-month intervals, generally for 6 to 12 months post treatment. The choice of when to assess progress in treatment should take into account the expected process of change during treatment and the clinical course of AUD (Collins and Graham, 2002; Maisto et al., 2014). For example, in a recent alcohol clinical trial, the greatest reduction in alcohol consumption occurred in the weeks preceding the initiation of treatment (Stasiewicz et al., 2013). Thus, including an assessment of pretreatment drinking that is consistent with the primary drinking outcome (e.g., the PHDD) and over a sufficient period prior to treatment to be a stable estimate may be critical to understanding the process of change in an alcohol clinical trial. Researchers who have limited resources may also consider a planned missingness approach, whereby individuals are randomly assigned to complete a limited number of assessment time points (Enders, 2010; Mun et al., 2015; Schafer and Graham, 2002). This approach can reduce costs, as well as participant and research staff burden, while also providing a wealth of data that can be analyzed using a variety of missing data approaches (described below).

With the growing availability of smartphones, the potential for repeated contacts via ecological momentary assessment (EMA) could be more widely incorporated into alcohol clinical trial designs, both for assessment and treatment support (Cohn et al., 2011; Gurvich et al., 2013; Gustafson et al., 2014). EMA measures participants’ current self-reported behaviors, environmental exposures, and experiences in real time, in the subjects’ natural environments (Shiffman et al., 2008). The conduct of multiple assessments over time, either daily or multiple times within days, can be particularly useful for analyses involving time-ordered effects (e.g., mediation analyses; Kranzler et al., 2014; Miranda et al., 2014) and for examining dynamic processes of change (Witkiewitz and Marlatt, 2004). The decision whether to use EMA should balance the impact of participant burden, which can be substantial, with the specific research questions that can be addressed uniquely using EMA. For example, if the primary outcome measure is any drinking during a specific period, then EMA may not be necessary (Shiffman et al., 2008).
Mechanisms and Moderators of Behavior Change

One of the major advantages of multiple assessments is that they may allow an examination of the mechanisms of change that underlie treatment efficacy (Kazdin, 2007; Longabaugh, 2007; Longabaugh and Magill, 2011). However, examining mechanisms of change often requires larger samples and raises unique measurement considerations. For example, it is important to measure the hypothesized mechanism on multiple occasions prior to the anticipated change in behavior and continue measurement until after the behavior is anticipated to change. Behavioral and in vivo assessments, such as EMA and laboratory-based studies, may be particularly useful to study hypothesized mechanisms of behavior change in alcohol treatment trials (Miranda et al., 2014).

The examination of potential mechanisms of change also requires specific analytic approaches. Such studies often require multiple longitudinal assessments, formal tests of mediation (Cheong et al., 2003; MacKinnon, 2008), and approaches that yield stronger causal inferences regarding mediators’ effects on outcomes (MacKinnon and Pirlott, 2015).

Of growing interest in the context of personalized medicine is the effort to identify genetic variants that moderate the response to treatment (e.g., Johnson et al., 2011; Kranzler et al., 2014). This approach promises to identify a priori individuals who are most likely to respond to a particular medication. From a clinical trials perspective, it also offers the potential for enrichment trials, in which participants are selected based upon their genotype, for which there is evidence of moderation. For a detailed discussion of this issue, see Jones and colleagues (2015). In addition to identifying potential treatment responders, validated genetic moderators can help to elucidate the mechanism of a medication’s effects by implicating a specific neurotransmitter receptor or other medication target or pathway.

STATISTICAL ANALYSES

Data Transformation and Alternative Distributions

Often, data from alcohol and drug treatment studies are not normally distributed. For example, participants in alcohol treatment trials generally report an increasing number of abstinent days during the active intervention, yielding highly skewed data with a preponderance of zero values (also called “zero inflation”) or a preponderance of values at the highest end of the range of values (e.g., 100% days abstinent). Thus, it may be necessary to transform the data prior to analysis or use analytic approaches that accommodate non-normal data distributions. The most commonly applied data transformations in alcohol clinical trials, such as square root or arcsine transformations (Project MATCH Research Group, 1997), do not eliminate the problem of zero inflation because the square root and arcsine of zero are zero. Alternative statistical models, including generalized linear models (Atkins et al., 2013; DeSantis et al., 2013), growth mixture models (Gueorguieva et al., 2010; Witkiewitz and Masyn, 2008), and latent Markov models (Witkiewitz, 2008; Witkiewitz et al., 2010), may be useful to address such distributional problems. These alternative statistical models are also an area of active ongoing methodological research (Bauer and Curran, 2003; Morgan et al., 2014; Sher et al., 2011; Steinley and Brusco, 2011).

Incorporating Site, Therapist, and Group/Cohort Effects

Methods that incorporate treatment site, therapist, and group/cohort effects in analyses are important for all alcohol clinical trials. Ignoring systematic effects of these factors can lead to substantial bias in treatment effect estimates and standard errors, with a loss of statistical power (Chu et al., 2011; Feaster et al., 2011; Kraemer and Robinson, 2005). Although the inclusion of site, intervention condition, and site-by-treatment interaction effects as “fixed effects” (i.e., the inclusion of site indicator variables as a covariate in the analysis) has been recommended (Kraemer and Robinson, 2005), large site differences in sample size or treatment effects can nonetheless lead to biased standard errors and make interpretation of the findings difficult (Chu et al., 2011). Likewise, therapist effects often explain a considerable amount of the variance in psychosocial treatment outcomes (Imel et al., 2008; Moyers and Miller, 2013). Oftentimes in alcohol clinical trials research, the therapists’ effects on the outcomes are larger than the specific effects for each active intervention being delivered (Miller and Moyers, 2015). Under some circumstances, problems may occur when fixed effects models are used to examine site, therapist, and group effects, for example, when the sample size is unbalanced across sites, therapists, or groups. With sample size imbalance, the fixed effect estimate is weighted by the sample sizes within sites, therapists, or groups, that is, the fixed effects will be more strongly influenced by the largest samples within the respective category. Finally, this approach is problematic when there is insufficient power to detect a site-by-treatment interaction due to small samples within sites or a small number of sites.

Given these limitations and the results of numerous simulation studies, many authors have recommended that multisite, multitherapist, and multigroup studies be analyzed using mixed effects models (i.e., hierarchical linear models, multilevel models: Chu et al., 2011; Kahan, 2014; Kahan and Morris, 2013; Moerbeek et al., 2003). Such models incorporate site and/or therapist and/or group information as a random effect, allowing for an estimation of the variability in outcomes that are due to therapist or group differences within or between sites. Mixed effects models provide more efficient and less biased estimates of the treatment effect, less biased standard errors, and greater statistical power than models that ignore site, therapist, or group effects or that treat these effects as fixed (Chu et al., 2011; Kahan, 2014; Kahan and Morris, 2013; Moerbeek et al., 2003). The analysis of data from groups that allow for rolling admission (i.e.,
open groups) requires even greater care, as the nonindependence of individuals within groups becomes a moving target (Morgan-Lopez and Fals-Stewart, 2006, 2007). Likewise, we recognize that “site” and “therapist” can be moving targets in some treatment settings, so that it may not be feasible to incorporate site and therapist as random effects. For example, individuals within a trial may transfer between sites or be recruited from multiple diverse programs within a site (e.g., intensive outpatient, aftercare), and some individuals may see multiple therapists or be enrolled in multiple types of treatment with different therapists. In these situations, it may be preferable to measure characteristics of the treatment received (e.g., number of groups attended) and/or therapists (e.g., working alliance with all providers) that could be incorporated as fixed effects in predicting treatment outcomes.

Primary Outcome Analyses

The primary analysis should be prespecified in the clinical trial registration and based on the original allocation of participants to intervention conditions, such that all individuals initially allocated to an intervention group are included in the analyses as members of that group. This analysis, often referred to as an “intention-to-treat” (ITT) analysis, has been compromised in some alcohol clinical trials by the use of a “modified” ITT approach or the incorrect use of the ITT approach (Del Re et al., 2013). The modified ITT approach inappropriately eliminates certain individuals from analyses either because of missing data (see section below) or because a minimum level of treatment adherence is not achieved, resulting in potentially biased treatment estimates. When study analyses require some level of adherence (called “as treated”) or total adherence (called “per protocol”), estimates of intervention effects will be biased unless nonadherence with the research protocol is random (Ye et al., 2014). Excluding randomized patients from the primary analysis has been considered acceptable when errors are made in the implementation of eligibility criteria or when patients never received any of the interventions (Fergusson et al., 2002). In cases of differential noncompliance across intervention groups, alternative approaches to ITT analyses, such as complier average causal effects models (Tucker et al., 2012; Ye et al., 2014), may produce less biased estimates of the intervention effects.

When a stratified randomization procedure is used, the stratification variables need to be reflected in the analysis and the resulting study report (Witkiewitz et al., 2015). Failure to account for balancing variables can lead to nonindependence of subjects across treatment groups, reduced statistical power, biased standard errors, and increased type I error rates (Kahan and Morris, 2012).

Subgroup (i.e., Moderator) Analyses

Rigorous (vs. exploratory) analyses of alcohol clinical trials that examine potential subgroup differences (van den Brink et al., 2013) or personalized medicine approaches (Kranzler and McKay, 2012) require prespecification of the analyses and adequate power to support them. Measurement of treatment moderators may require additional assessments, such as in pharmacogenetic studies, which require the collection of blood, saliva, or other tissue samples from which DNA can be extracted for genotyping (Goldman et al., 2005; Johnson et al., 2011). Important elements in a rigorous subgroup analysis include the use of a valid randomization procedure, adjustment for covariates with correction for multiple comparisons, consideration of the potential for nonreplicability of subgroup differences when overall intervention effect sizes are small or absent (Grouin et al., 2005), and evaluation of the statistical significance of subgroup differences via tests of moderation effects.

Moderation analyses are critical for advancing therapeutic development and personalized medicine (Kranzler and McKay, 2012), but are particularly susceptible to bias, especially when subgroups are small (Brookes et al., 2004). Investigators are encouraged to consider methods to estimate effect sizes of treatment moderators (Kraemer, 2013) and methods to combine moderation analyses across studies to increase sample size and reduce bias (Brown et al., 2013; Wallace et al., 2013).

In general, any tests of moderation should ideally be planned a priori based on prior research or theoretical considerations and reflected in the trial registration. Analyses that are proposed after examining the collected data may be important for scientific reasons, but may yield spurious findings (Freemantle, 2001; Pocock et al., 1987). When reporting such findings, they should be identified as having resulted from analyses conducted after the data were collected (Witkiewitz et al., 2015).

Missing Data Analyses

Unfortunately, missing data are common in alcohol clinical trials. Even the most carefully designed and executed studies are likely to experience some participant dropout, which occurs for a variety of reasons (Ball et al., 2006; Brorson et al., 2013; Coulson et al., 2009; Palmer et al., 2009). Numerous causes of missing data have been identified, such as relocation, incarceration, changes in employment, and childcare needs, although oftentimes it is not possible to determine why some data are missing. The first critical step in dealing with missing data is to conduct attrition analyses to examine whether participants with missing data and those with complete data are systematically different on any available measures that were obtained prior to dropout. Importantly, many studies might not be powered to show effects on attrition, even when systematic influences on attrition do exist.

Variables most critical to examine in such analyses include demographic and other baseline variables that were identified a priori as potentially impacting outcomes, and any outcome measures that were collected prior to the time of
dropout from the study. Systematic attrition biases can also be assessed via sensitivity analyses that examine the impact of missing data on the outcomes of interest (Enders, 2010, 2011; Jackson et al., 2014; Schafer and Graham, 2002). Pretreatment variables that predict attrition should be accounted for in subsequent analyses (National Research Council, 2010). Investigators might also consider including auxiliary measures at baseline that have been shown to predict attrition in prior studies (e.g., conscientiousness, honesty, and humility; Satherley et al., 2015) or may also consider adding a question at each assessment regarding the participant’s intention to stay enrolled in the study (Schafer and Graham, 2002). The inclusion of auxiliary variables (defined as measures that may be associated with attrition and not necessarily associated with the outcomes of interest) can be critical in the missing data analytic models described next (Graham, 2009).

The second step in dealing with missing data is to use an analytic model that produces the least biased estimates of the intervention effect, even in the presence of missing data. Witkiewitz et al. recently published 2 studies that compared commonly used statistical approaches for handling missing data for continuous outcomes in alcohol clinical trials, using a simulation approach (Hallgren and Witkiewitz, 2013) and real-world data (Witkiewitz et al., 2014). The approaches included last observation carried forward, baseline observation carried forward, placebo mean imputation, poor outcomes (e.g., heavy drinking) imputation, full information maximum likelihood, and multiple imputation. Consistent with 30 years of prior research on missing data approaches, the findings from both studies clearly showed full information maximum likelihood and multiple imputation. Consistent with 30 years of prior research on missing data approaches, the findings from both studies clearly showed full information maximum likelihood and multiple imputation to generate the least biased estimates of intervention effects in alcohol clinical trials. Importantly, in both studies, 2 of the most commonly used methods in alcohol clinical trials research—last observation carried forward and poor outcome imputation (e.g., assuming drinking; Del Re et al., 2013)—tended to produce the most biased estimates of treatment effects in alcohol clinical trials (Hallgren and Witkiewitz, 2013; Witkiewitz et al., 2014). Although they are the preferred approaches, full information maximum likelihood estimation and multiple imputation can introduce bias if data are missing not at random. Sensitivity analyses to examine the effects of different missing data models can be useful in evaluating the impact of missing data on the analysis of primary outcomes (Enders, 2010, 2011; Jackson et al., 2014; Witkiewitz et al., 2012).

### CONCLUSIONS AND FUTURE DIRECTIONS

Calls to develop additional and more efficacious treatments for AUDs can be found in the peer-reviewed scientific literature (Anton et al., 2012; Davies et al., 2013; Litten et al., 2012; Witkiewitz and Marlatt, 2011), as well as the popular press (Beck, 2014). The need for more treatment options with greater efficacy is reflected in the substantial global public health impact and societal costs of heavy drinking (Bouchery et al., 2011; World Health Organization, 2011) and by the high proportion of individuals with AUDs who never receive treatment or who drop out from treatment before receiving benefits from it (Substance Abuse and Mental Health Services Administration, 2012).

A key factor in the development of more efficacious AUD treatments is the use of sound methods to evaluate treatment efficacy. Recent reviews have shown that many RCTs in alcohol treatment have produced potentially biased estimates of treatment effects due to faulty design and analysis decisions (Del Re et al., 2013; Hallgren and Witkiewitz, 2013;

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Selected relevant citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Careful consideration and reporting of exclusion criteria</td>
<td>Altman et al. (2001); Humphreys et al. (2008)</td>
</tr>
<tr>
<td>2. Selection of a randomization procedure that reduces imbalances across treatment groups</td>
<td>Hedden et al. (2006); Matts and Lachin (1988)</td>
</tr>
<tr>
<td>3. Consider use of propensity score or instrumental variable methods when group imbalances exist</td>
<td>Berger (2005b); Leyrat et al. (2013)</td>
</tr>
<tr>
<td>4. Use of well-validated self-report measures and either biomarkers or carefully selected collateral informants to validate self-reported consumption</td>
<td>Litten et al. (2010); Miller (1996); Sobell and Sobell (1992); Sobell et al. (1997)</td>
</tr>
<tr>
<td>5. Careful consideration of the timing of assessments to inform the durability of treatment effects and variation in outcomes over time</td>
<td>Collins and Graham (2002); Stasiewicz et al. (2013)</td>
</tr>
<tr>
<td>6. Investigation of data distributions and the use of either data transformation or alternative analytic techniques when data are non normal</td>
<td>Atkins et al. (2013); Gueorguieva et al. (2012); Witkiewitz et al. (2010)</td>
</tr>
<tr>
<td>7. An intention-to-treat (ITT) analytic approach that takes into account clustering of patients within sites, therapists, and/or groups and that controls for covariate adjustment in the randomization procedure</td>
<td>Del Re et al. (2013); Kahan and Morris (2013); Moerbeek et al. (2003)</td>
</tr>
<tr>
<td>8. An alternative approach to ITT, such as the complier average causal effect model, if there is unequal compliance across groups</td>
<td>Tucker et al. (2012); Ye et al. (2014)</td>
</tr>
<tr>
<td>9. Minimizing missing primary outcome data by continuing to assess all individuals (including those who drop out of treatment) and maximum likelihood estimation or multiple imputation to accommodate missing data</td>
<td>Enders (2010); Hallgren and Witkiewitz (2013); Witkiewitz et al. (2014)</td>
</tr>
<tr>
<td>10. Sensitivity analyses to evaluate the impact of missing data on the treatment effect estimates</td>
<td>Enders (2011); Jackson et al. (2014)</td>
</tr>
</tbody>
</table>
Humphreys et al., 2008; Jonas et al., 2014a). Some of these problems identified include excessive exclusion criteria, biased randomization, inappropriate primary outcome analyses, and the use of faulty approaches to handle missing data.

The current review focused on design and analysis considerations for RCTs and the evaluation of treatment efficacy, yet many of the recommendations are also suitable for comparative effectiveness RCTs. Comparative effectiveness studies and pragmatic trials are critical for informing clinical care, but a review of methods for comparative effectiveness research is beyond the scope of the current review. A recent review of the comparative effectiveness of pharmacotherapies for AUD provides some suggestions for future research (Jonas et al., 2014b). Guidelines for pragmatic trials have also been developed (Loudon et al., 2015).

Based on our systematic review of the literature, we provide 10 recommendations for future research in Table 2. Valid methods developed for use in treatment trials for a variety of disorders, including AUDs, are available for the design and implementation of alcohol treatment trials. The high cost of clinical trials (both in the resources expended and the exposure of participants to the risks of experimental treatments) argue forcefully for the use of state-of-the-art methods in treatment trials for the disorder. The availability of new, highly efficacious treatments that could result from the high cost of clinical trials (both in the resources expended and the exposure of participants to the risks of experimental treatments) argue forcefully for the use of state-of-the-art methods in treatment trials for the disorder. The availability of new, highly efficacious treatments that could result from the use of these methods could help to remedy the low rate of treatment utilization by individuals with an AUD.

ACKNOWLEDGMENTS

This research was supported by grants from the National Institute on Alcohol Abuse and Alcoholism (AA022328 to KW and AA021164 and AA023192 to HRK) and by the U.S. Department of Veterans Affairs (RCS-14-232 and the VISN 4 Mental Illness Research, Education, and Clinical Center). The views expressed do not reflect those of the Department of Veteran Affairs or other institutions.

DISCLOSURES

HRK has served as a consultant or advisory board member for Alkermes, Lilly, Lundbeck, Otsuka, Pfizer, and Roche. KW has served as a consultant for Alkermes. HRK and KW have received financial support from the American Society of Clinical Psychopharmacology’s Alcohol Clinical Trials Initiative (ACTIVE), which is supported by AbbVie, Ethypharm, Lilly, Lundbeck, and Pfizer. Some of the views on clinical trial design and methods reflect discussions that HRK and KW had during their participation in ACTIVE meetings.

REFERENCES


