Aggregating and Analyzing Daily Drinking Data in Clinical Trials: A Comparison of Type I Errors, Power, and Bias

KEVIN A. HALLGREN, PH.D.,^{a,*} DAVID C. ATKINS, PH.D.,^a & KATIE WITKIEWITZ, PH.D.^b

^a*Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington*

^b*Department of Psychology, Center on Alcoholism, Substance Abuse, and Addictions, University of New Mexico, Albuquerque, New Mexico*

ABSTRACT. Objective: Statistical analyses in alcohol clinical trials often use longitudinal daily drinking data (e.g., percentage of drinking days) to test treatment efficacy. Such data can be aggregated and analyzed in many ways. To assess how statistical analytic decisions may influence substantive results, the current report compares different aggregation methods (e.g., computing percentages of drinking days vs. using daily binary indicators of drinking) and statistical methods (i.e., least squares regression, linear mixed-effects models [LMM], generalized linear mixed models [GLMM], and generalized estimating equations [GEE]) for testing the effects of treatment on drinking outcomes in clinical trials. **Method:** A simulation study repeatedly resampled daily drinking data from the treatment period of the Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence (COMBINE) Study at different sample sizes. Treatment effects in each data set were modeled using different aggregation and statistical methods. **Results:**

Type I error rates were near the expected rate for most models but on occasion were mildly elevated when disaggregated daily drinking data were analyzed using GLMM or GEE with an exchangeable correlation structure. Most methods yielded similar statistical power, although power decreased when modeling disaggregated daily drinking with GLMM and had mixed increases and decreases when the longitudinal nature of data was ignored by using fully aggregated data with independent samples *t* tests. **Conclusions:** When testing treatment main effects, relatively simpler statistical methods with fewer repeated measures may perform equally well or better than more complicated methods. Patterns of significance and treatment effect size estimates are likely comparable across most studies that use different aggregation and statistical methods, but differences between these methods may occasionally have an important impact on conclusions in clinical trials. (*J. Stud. Alcohol Drugs*, 77, 986–991, 2016)

THE PRIMARY OUTCOMES IN MOST clinical trials for alcohol use disorder (AUD) are indices of alcohol consumption, typically measured via Timeline Followback (TLFB; Sobell & Sobell, 1992) or other calendar methods (e.g., Form 90; Miller & Del Boca, 1994). Such methods yield daily drinking data, but there is no consensus on how to aggregate these data into summary measures, such as percentage of drinking days (PDD) and percentage of heavy drinking days (PHDD). For example, TLFB data could be represented as a single PDD or PHDD estimate over the full period of interest (e.g., O'Malley et al., 2015; Ponizovsky et al., 2015), aggregated into repeated measures over weekly or monthly intervals (e.g., Anton et al., 2006; Project MATCH Research Group, 1997), or left completely disaggregated as daily binary indicators of drinking and heavy drinking (e.g., DeSantis et al., 2013; McCrady et al., 2016).

Different aggregation methods produce different scales of measurement and also introduce limits on how the data

can be analyzed statistically. For example, full aggregation (with a single PDD or PHDD observation per subject) represents average drinking over the entire period and may be analyzed using ordinary least squares (OLS) regression, analysis of variance (ANOVA), or *t* tests. Aggregation across discrete periods introduces repeated-measures data (e.g., repeated measurement of monthly PDD) that must be accounted for statistically, for example, by using repeated-measures ANOVA or a linear mixed-effects model (LMM, also called multilevel model). Completely disaggregated daily data (e.g., daily indicators of drinking and heavy drinking) yield daily binary indices of drinking, and statistical analyses must account for repeated measures and binary outcomes (e.g., with generalized linear mixed models [GLMM] or generalized estimating equations [GEE]). Importantly, each method of aggregation yields a different measure of drinking and requires a different statistical model.

The choice of methods to analyze any data set should be guided by theoretical considerations and the structure of the data. However, in many cases different methods are used but the substantive goal is the same: to compare how one or more treatments affect alcohol consumption. There has been little work demonstrating how different aggregation and statistical methods influence treatment effect estimates, power, and type I errors. The present study empirically compares common aggregation and statistical methods for modeling treatment effects in AUD clinical trials that use PDD and PHDD as outcome measures.

Received: February 11, 2016. Revision: July 12, 2016.

Support for this research was provided by National Institute on Alcohol Abuse and Alcoholism (NIAAA) Grants T32AA007455 and R01AA022328. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAAA or the National Institutes of Health.

*Correspondence may be sent to Kevin A. Hallgren at the Department of Psychiatry and Behavioral Sciences, University of Washington, 1959 NE Pacific Street, Box 356560, Seattle, WA, 98195, or via email at: khallgre@uw.edu.

Method

Data source

Within-treatment data were analyzed from the Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence (COMBINE) Study (Anton et al., 2006), a multisite randomized clinical trial that examined combinations of pharmacotherapy and behavioral interventions for AUD. The study used a $2 \times 2 \times 2$ design in which participants ($N = 1,383$) were randomized to one of eight conditions that received combinations of naltrexone or placebo-naltrexone, acamprosate or placebo-acamprosate, and medication management (MM) or combined behavioral intervention (CBI) with MM. A ninth condition received only CBI and no medication or placebo. The treatment period was 16 weeks. Eighty-four percent of participants ($n = 1,160$) provided complete within-treatment drinking data and were included in the present analysis to minimize the effect of missing data on the results (see Hallgren & Witkiewitz, 2013, and Witkiewitz et al., 2014, for tests of methods for handling missing data in alcohol clinical trials).

Daily alcohol consumption was assessed every 4 weeks during treatment using the Form 90 (Miller & Del Boca, 1994). Each day was classified as a drinking day if any drinking occurred and as a heavy drinking day if four/five standard drinks were exceeded within a day for women/men.

Resampling procedure

The effects of treatment on PDD and PHDD were examined for two comparisons that were significant in the original COMBINE Study: (a) naltrexone versus placebo (without CBI, with or without acamprosate) and (b) CBI with MM versus MM-only (without naltrexone, with or without acamprosate). A quasi-simulation procedure (Hallgren, 2013) randomly sampled subsets of 50, 100, 200, or 400 participants from each pair of contrasted conditions. Half of each sample was in the active treatment condition (naltrexone or CBI), and half was in the comparison condition (placebo or MM-only). Daily drinking data from each sample were aggregated and analyzed using each of the methods described below. This sampling, aggregation, and analysis procedure was repeated 3,000 times for each combination of treatment contrasts and sample sizes. Statistical power was empirically estimated from the proportion of models with statistically significant treatment effects ($\alpha = .05$, two-tailed).

Type I error rates were tested by simulating fabricated treatment conditions from the same COMBINE condition. Participants were drawn from the placebo-only condition at three sample sizes (50, 100, and 200); half were designated (wrongly) as receiving treatment, and the other half were designated (correctly) as receiving placebo. Significant treat-

ment effects would incorrectly imply different outcomes in two groups from the same population (i.e., all participants were in the same placebo-only condition), which is a type I error. This procedure was repeated 3,000 times, and type I error rates were estimated empirically as the proportion of models with type I errors.

Aggregation and statistical methods

For each resampled data set, we computed (a) fully aggregated PDD and PHDD (single values for each participant) for the full 4-month treatment period, (b) repeated measures of monthly (28-day) PDD and PHDD over the same period, and (c) disaggregated daily binary indicators of drinking and heavy drinking. Differences between treatments were then tested with plausible statistical models based on the type of aggregation. Fully aggregated PDD and PHDD were tested using OLS regression, which was computationally identical to ANOVA and an independent samples t test for comparing two treatment groups. Monthly aggregated data were analyzed using LMMs (Snijders & Bosker, 2012) to account for non-independence in repeated measurements.¹ LMMs were structured as multilevel growth-curve models with fixed effects for treatment condition, linear time, and Condition $\times$ Time interactions and random subject-level intercepts and slopes over time. LMMs with identical specifications using arcsine square-root-transformed PDD and PHDD (LMM-asin) were also tested (e.g., Project MATCH Research Group, 1997). Disaggregated daily binary indicators of drinking and heavy drinking were modeled several ways. GLMMs (Breslow & Clayton, 1993) were structured as growth-curve models with fixed effects for treatment condition, time, and Treatment $\times$ Time interactions, subject-level random intercepts and slopes over time, and a logit link from the prediction model to the binary daily indicators. Three other approaches used GEEs, in which the correlation due to repeated measures is handled by a correlation model of the residuals and binary outcomes are modeled with a logit link. The GEEs modeled residuals with no correlation or independence (GEE-ind), a lag-1 autoregressive correlation structure (GEE-ar1) with stronger correlations between repeated measures that were temporally closer, or an exchangeable (GEE-ex) or equal correlation among all repeated measures. All GEE models used Huber–White standard errors.

Each of the analytic approaches can plausibly be used to test similar substantive hypotheses of within-treatment differences in PDD and PHDD. Nonetheless, each method has different assumptions, interpretations, and practical considerations. For example, the OLS model used here assumes no effect of time, and therefore tests whether treatments

¹Repeated-measures ANOVA is another alternative that was not tested here because of more stringent assumptions (e.g., sphericity, no missing data) that are often not met in alcohol clinical trials (see Quené & van den Bergh, 2004).

differ when their effects are averaged over the full treatment period. Other models explicitly included time effects, and their main effects therefore indicate whether treatments differ at the point where time was centered (equal to zero); alternatively, a Treatment $\times$ Time interaction would indicate whether conditions differ in their rates of change over time. GEE models have marginal interpretations in which treatment effects reflect differences across a population, whereas GLMMs have conditional interpretations in which treatment effects reflect differences within a specific individual, holding all other variables constant. LMMs have both marginal and conditional interpretations. The analytic approaches also differ in terms of estimation algorithms, methods for handling missing data, and other assumptions (e.g., outcome distributions). We do not discuss these differences at length here but refer readers to Diggle et al. (2002) for an approachable overview.

Time variables were centered at the last day of treatment, and the main effects of treatment at that time were the focus of all analyses that included time effects. Separate analyses that controlled for gender, age, and baseline alcohol dependence yielded similar patterns of results that are not reported here. R software (R Core Team, 2014) was used with the lme4 (Bates et al., 2014) and geeM packages (McDaniel & Henderson, 2015) for fitting mixed-effects models and GEEs, respectively.

Results

Based on the raw COMBINE data, naltrexone reduced PDD and PHDD by a mean difference of 6.9 and 6.5 percentage units relative to placebo across the full treatment period, and CBI reduced PDD and PHDD by 4.9 and 5.0 percentage units relative to MM-only. In the resampled data sets, OLS, LMM, GLMM, GEE-ind, and GEE-ar1 provided unbiased treatment effect estimates relative to these differences in means. However, GEE-ex overestimated treatment differences by up to 0.5 percentage units (up to 9% relative error), and LMM-asin underestimated treatment differences by 1.3 to 3.7 percentage units (up to 57% relative error).

The upper and middle panels of Figure 1 display statistical power for each method. Power estimates from each method are represented as bar graphs, and methods yielding significantly higher or lower power than LMM² (referenced as vertical dashed lines) are indicated by asterisks. OLS regression with fully aggregated data yielded higher power than LMM in three PDD models of naltrexone versus placebo but lower power in models of CBI versus MM-only with $n \geq 100$. The mixed results for OLS are consistent with the patterns of change in drinking over time for naltrexone (relative to placebo) and CBI (relative to MM-only). That is, in

LMM-based models with the complete COMBINE data set, the effect of naltrexone (vs. placebo) on PDD did not differ over time (i.e., no Treatment $\times$ Time interaction; $b = -0.63$, $SE = 0.77$, $p = .42$). However, the effect of CBI (vs. MM-only) increased gradually over time for PHDD ($b = -1.49$, $SE = 0.74$, $p = .04$), with a similar, marginally significant pattern for PDD ($b = -1.46$, $SE = 0.76$, $p = .06$), and full aggregation across this period likely reduced power by ignoring this Treatment $\times$ Time interaction. GLMMs with daily data had significantly lower power in 5 of the 16 conditions tested, particularly when $n \geq 200$ and when modeling CBI versus MM-only. LMM-asin, GEE-ind, GEE-ar1, and GEE-ex had no significant differences in statistical power compared with LMM.

The lower portion of Figure 1 displays type I error rates, which were generally near their expected alpha levels (dashed lines). GLMM and GEE-ex each produced significantly higher type I error rates than the LMM when modeling heavy drinking and $n = 50$.

Discussion

When testing treatment main effects, relatively simpler statistical methods with fewer repeated measurements (e.g., LMM with monthly PDD and PHDD) may perform equally well or better than more complicated methods (e.g., GLMM, GEE) that use more frequent measurements and disaggregated daily data. Most methods yielded similar type I error rates, although type I errors were elevated for heavy drinking at the smallest sample size in the GLMM and GEE-ex analyses. No method clearly outperforms the others in terms of statistical power, although GLMM may often reduce power. GEE-ex may slightly overestimate treatment effects, suggesting the potential for bias, whereas arcsine transforming PDD and PHDD data before using LMM may yield substantially lower model-implied marginal treatment effect estimates.

These results are promising in suggesting that the patterns of significance and treatment effect size estimates should be mostly comparable across studies that use different aggregation and statistical methods. This finding is similar to other work showing only small empirical differences between methods (e.g., Gardiner et al., 2009) in the context of longitudinal depression research.

Aggregation and analytic strategies should be informed by theoretical, methodological, and practical reasoning. For example, for pharmacotherapies, such as naltrexone, that are hypothesized to affect drinking outcomes throughout the course of treatment, a simple OLS-based t test may be sufficient or even optimal for detecting main effects of treatment. In contrast, for behavioral therapies that may cause gradual improvements over time (i.e., "sleeper effects," Carroll et al., 2006), an LMM-based model with repeated measures would allow treatment effects to be pinpointed to a specific time when the effect may be largest, which may improve power.

²LMM was chosen as reference group because of its frequent application in clinical trials and because its power estimates were often similar to other tested models.

FIGURE 1. Power (top and middle panels) and type I error rates (bottom panels) for different aggregation and statistical methods. Dashed vertical lines indicate statistical power obtained from LMMs (top and middle panels) and the expected type I error rate for two-tailed alpha of .05 (bottom panels). Models that yielded significantly different power or type I error rates compared to the LMMs are marked with asterisks. PDD = percentage of drinking days; PHDD = percentage of heavy drinking days; LMM = linear mixed-effects model; OLS = ordinary least squares model; LMM-asin = linear mixed-effects model with arcsine square-root transformation of the outcome variable; GLMM = generalized linear mixed model; GEE-ind = generalized estimating equations with independence correlation structure; GEE-ar1 = generalized estimating equations with autoregressive correlation structure; GEE-ex = generalized estimating equations with exchangeable correlation structure; CBI = combined behavioral intervention, MM = medication management. *Methods yielding significantly higher or lower power than LMM (referenced as vertical dashed lines).

Disaggregating daily drinking data and using more complex statistical models (e.g., GLMM, GEE) may make analyses more complex while offering little benefit for analyzing main effects of treatment. However, disaggregation may be beneficial for secondary purposes, such as modeling how treatment affects daily process variables like mood, treatment adherence, or craving (e.g., Cooney et al., 2015; Richardson et al., 2008).

Arcsine transformation did not improve power or type I errors and made models more difficult to interpret. The arcsine transformation also biased the model-implied marginal effect estimates; thus, we echo other researchers (Warton & Hui, 2011) in cautioning against using the arcsine transformation.

Limitations

The present study has several limitations. Longitudinal models assumed linear time effects and did not include higher-order polynomials (e.g., quadratic growth). Not all models could include identical effects (e.g., no random slopes in GEE, no time effects in OLS). Recent studies have found that mixture models (i.e., representing unmeasured subgroups) and zero-inflated models (i.e., representing non-drinkers as a specific subgroup) improve model fit (Atkins et al., 2013; Witkiewitz, 2008; Witkiewitz et al., 2010); however, these models add computational complexity, require large samples, and are less commonly used in treatment research. Testing the above methods with missing data was, unfortunately, beyond the scope of the present study, and participants with missing data were excluded; therefore, the results should be interpreted as best-case scenarios if missing data are accounted for in an ideal manner. It is also possible that results may differ in other studies, particularly those unlike the COMBINE Study in the patterns of change over time, analytic aims, or types of outcome measures used.

The present study also has several strengths. Data came from a large, well-known, and well-controlled clinical trial. Despite the considerable variability in the statistical methods used in clinical alcohol research, this is, to our knowledge, the first head-to-head comparison of aggregation and statistical methods using calendar-based daily drinking data from an AUD clinical trial.

Conclusions

The choice of methods for aggregating and analyzing drinking data is not trivial. Different aggregation and statistical methods require different assumptions, interpretations, and practical considerations that should be addressed within the context of a specific study. Nonetheless, different methods are often used to test similar or identical substantive hypotheses.

Statistical analysis in clinical trials should characterize

what happened in the study, and the tools for doing so will usually vary depending on the study design and data. For the specific goal of evaluating treatment main effects, the results of the present study may provide empirical guidance for selecting an analytic approach. This may be particularly relevant given that data analytic plans must be specified a priori in many clinical trial registries. Different aggregation and statistical methods may yield small differences, but these differences can occasionally affect substantive interpretations.

References

- Anton, R. F., O'Malley, S. S., Ciraulo, D. A., Cisler, R. A., Couper, D., Donovan, D. M., . . . Zweben, A., & the COMBINE Study Research Group. (2006). Combined pharmacotherapies and behavioral interventions for alcohol dependence: The COMBINE study: A randomized controlled trial. *JAMA*, 295, 2003–2017. doi:10.1001/jama.295.17.2003
- Atkins, D. C., Baldwin, S. A., Zheng, C., Gallop, R. J., & Neighbors, C. (2013). A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors*, 27, 166–177. doi:10.1037/a0029508
- Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2014). lme4: Linear mixed-effects models using eigen and S4, version 1.1-8 [Computer software]. Retrieved from <http://arxiv.org/abs/1406.5823>
- Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9–25. doi:10.2307/2290687
- Carroll, K. M., Easton, C. J., Nich, C., Hunkeler, K. A., Neavins, T. M., Sinha, R., . . . Rounsaville, B. J. (2006). The use of contingency management and motivational/skills-building therapy to treat young adults with marijuana dependence. *Journal of Consulting and Clinical Psychology*, 74, 955–966. doi:10.1037/0022-006X.74.5.955
- Cooney, N. L., Litt, M. D., Sevarino, K. A., Levy, L., Kranitz, L. S., Sackler, H., & Cooney, J. L. (2015). Concurrent alcohol and tobacco treatment: Effect on daily process measures of alcohol relapse risk. *Journal of Consulting and Clinical Psychology*, 83, 346–358. doi:10.1037/a0038633
- DeSantis, S. M., Bandyopadhyay, D., Baker, N. L., Randall, P. K., Anton, R. F., & Prisciandaro, J. J. (2013). Modeling longitudinal drinking data in clinical trials: An application to the COMBINE study. *Drug and Alcohol Dependence*, 132, 244–250. doi:10.1016/j.drugalcdep.2013.02.013
- Diggle, P. J., Heagerty, P., Liang, K., & Zeger, S. L. (2002). *Analysis of longitudinal data*. Oxford, England: Oxford Press.
- Gardiner, J. C., Luo, Z., & Roman, L. A. (2009). Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine*, 28, 221–239. doi:10.1002/sim.3478
- Hallgren, K. A. (2013). Conducting simulation studies in the R programming environment. *Tutorials in Quantitative Methods for Psychology*, 9, 43–60.
- Hallgren, K. A., & Witkiewitz, K. (2013). Missing data in alcohol clinical trials: A comparison of methods. *Alcoholism: Clinical and Experimental Research*, 37, 2152–2160. doi:10.1111/acer.12205
- McCrary, B. S., Epstein, E. E., Hallgren, K. A., Cook, S., & Jensen, N. K. (2016). Women with alcohol dependence: A randomized trial of couple versus individual plus couple therapy. *Psychology of Addictive Behaviors*, 30, 287–299. doi:10.1037/adb0000158
- McDaniel, L. S., & Henderson, N. (2015). geeM: Solve generalized estimating equations, version 0.7.4 [Computer software]. Retrieved from <http://CRAN.R-project.org/package=geeM>
- Miller, W. R., & Del Boca, F. K. (1994). Measurement of drinking behavior using the Form 90 family of instruments. *Journal of Studies on Alcohol, Supplement 12*, 112–118. doi:10.15288/jsas.1994.s12.112
- O'Malley, S. S., Corbin, W. R., Leeman, R. F., DeMartini, K. S., Fucito, L.

- M., Ikomi, J., . . . Kranzler, H. R. (2015). Reduction of alcohol drinking in young adults by naltrexone: A double-blind, placebo-controlled, randomized clinical trial of efficacy and safety. *Journal of Clinical Psychiatry*, 76, e207–e213. doi:10.4088/JCP.13m08934
- Ponizovsky, A. M., Rosca, P., Aronovich, E., Weizman, A., & Grinshpoon, A. (2015). Baclofen as add-on to standard psychosocial treatment for alcohol dependence: A randomized, double-blind, placebo-controlled trial with 1 year follow-up. *Journal of Substance Abuse Treatment*, 52, 24–30. doi:10.1016/j.jsat.2014.11.007
- Project MATCH Research Group. (1997). Matching Alcoholism Treatments to Client Heterogeneity: Project MATCH posttreatment drinking outcomes. *Journal of Studies on Alcohol*, 58, 7–29. doi:10.15288/jsa.1997.58.7
- R Core Team. (2014). R: A language and environment for statistical computing [Computer software]. Version 3.1.0. Vienna, Austria.
- Richardson, K., Baillie, A., Reid, S., Morley, K., Teesson, M., Sannibale, C., . . . Haber, P. (2008). Do acamprosate or naltrexone have an effect on daily drinking by reducing craving for alcohol? *Addiction*, 103, 953–959. doi:10.1111/j.1360-0443.2008.02215.x
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, England: Sage.
- Sobell, L. C., & Sobell, M. B. (1992). Timeline followback: A technique for assessing self-reported alcohol consumption. In A. Z. Litten & J. Allen (Eds.), *Measuring alcohol consumption: Psychosocial and biological methods* (pp. 41–72). Totowa, NJ: Humana.
- Warton, D. I., & Hui, F. K. (2011). The arcsine is asinine: The analysis of proportions in ecology. *Ecology*, 92, 3–10. doi:10.1890/10-0340.1
- Witkiewitz, K. (2008). Lapses following alcohol treatment: Modeling the falls from the wagon. *Journal of Studies on Alcohol and Drugs*, 69, 594–604. doi:10.15288/jsad.2008.69.594
- Witkiewitz, K., Falk, D. E., Kranzler, H. R., Litten, R. Z., Hallgren, K. A., O'Malley, S. S., & Anton, R. F., & the Alcohol Clinical Trials Initiative (ACTIVE) Workgroup. (2014). Methods to analyze treatment effects in the presence of missing data for a continuous heavy drinking outcome measure when participants drop out from treatment in alcohol clinical trials. *Alcoholism: Clinical and Experimental Research*, 38, 2826–2834. doi:10.1111/acer.12543
- Witkiewitz, K., Maisto, S. A., & Donovan, D. M. (2010). A comparison of methods for estimating change in drinking following alcohol treatment. *Alcoholism: Clinical and Experimental Research*, 34, 2116–2125. doi:10.1111/j.1530-0277.2010.01308.x