

Improving the Analysis of Randomized Controlled Trials: a Posterior Simulation Approach

Jeffrey A. Mills*

University of Cincinnati, Cincinnati, USA

Gary J. Cornwall

Bureau of Economic Analysis, Washington D.C., USA

Beau A. Sauley

University of Cincinnati, Cincinnati, USA

Jeffrey R. Strawn

University of Cincinnati & Cincinnati Children's Hospital, Cincinnati, USA

Key Words: Randomized controlled trial, Hypothesis testing, Bayesian inference, Bayes factor, Posterior simulation, Monte Carlo, anxiety, placebo, treatment

Disclosures: While the authors do not have relationships that would impact the topic or represent a conflict or reasonable source of bias, the following are provided in the spirit of full disclosure. Dr. Strawn has received research support from the National Institutes of Health, Edgemont, Allergan, Lundbeck, Neuronetics. He has received material support from Assurex/Genesight. The remaining authors reported no conflicts of interests. Any views expressed here are those of the authors and not necessarily those of the Bureau of Economic Analysis or U.S. Department of Commerce.

***Correspondence:** Jeffrey A. Mills, Department of Economics, ML0371, Lindner College of Business, University of Cincinnati, Cincinnati OH 45221, Tel: 513.556.2619, Fax: 513.556.6278, email: millsjf@ucmail.uc.edu

Summary. The randomized controlled trial (RCT) is the standard for establishing efficacy and tolerability of treatments. However, the statistical evaluation of treatment effects in RCTs has remained largely unchanged for several decades. A new approach to Bayesian hypothesis testing for RCTs that leverages posterior simulation methods is developed. This approach (1) employs Monte Carlo simulation to obtain exact posterior distributions with fewer restrictive assumptions than required by current standard methods, allowing for a relatively simple procedure for inference with analytically intractable models, and (2) utilizes a novel approach to Bayesian hypothesis testing.

1. Introduction

During the last century, the randomized controlled trial (RCT) became the “gold standard” for establishing the efficacy of treatment interventions in medicine, social science, education and beyond (Bothwell *et al.*, 2016). By the 1980s, the US Food and Drug Administration and the European Medicines Agency required positive RCTs (*i.e.*, trials in which an experimental intervention is superior to a control intervention) for the approval of new medications or for the approval of previously available interventions in new diseases. Beyond satisfying regulatory requirements, RCTs provide estimates of average treatment effect (ATE) (Bothwell *et al.*, 2016) as well as comparative efficacy and tolerability data among interventions (Lumley, 2002). There has been considerable effort to improve the design of these trials; for example, a large literature has developed on adaptive designs (Berry, 2006; Collins *et al.*, 2012). However, the statistical methodology for evaluating treatment-control differences from these trials, with the exception of a handful of theoretical innovations, *e.g.*, clinical trial simulations (Holford, Ma, & Ploeger, 2010), have remained largely unchanged over the past forty years or more.

In clinical RCTs, the traditional approach to comparing treatment effects for quantitative measures (*e.g.*, change in dimensional symptom severity measure) utilizes a Welch t-test or, if covariates are available (*e.g.* study site, study visit, interaction terms), Mixed Model for Repeated Measures (MMRM) analysis is employed. For binary outcomes and in small samples, Pearson’s χ^2 test and Fisher’s exact test are commonly used (March *et al.* 2004; Strawn *et al.* 2015; Walkup *et al.* 2008). Pearson’s test is an asymptotic test, so unreliable with small samples, while Fisher’s exact test is a conditional frequentist test that increases discreteness and thus the conservatism of the test. Because of this conservatism, the observed rejection rate is often below the nominal significance level (Williamson *et al.*, 2017) The limitations of these approaches (*e.g.*, distributional assumptions, difficulty addressing heterogeneity of variance) have been extensively discussed (Lee & Chu, 2012). While Bayesian approaches to analysis have been suggested (Spiegelhalter, *et al.*, 1994, Spiegelhalter *et al.*, 2004, Lee & Chu, 2012), these have not addressed both the inability to test hypotheses without employing unrealistic informative priors, and the difficulty obtaining analytically intractable posterior distributions.

Modern Monte Carlo (MC) posterior simulation methods have largely addressed the problem of analytical intractability, but have yet to be widely applied to the analysis of RCTs. Problems with the Bayesian approach to hypothesis testing may be the main reason preventing widespread adoption: the common sentiment appears to be that, while Bayesian inference would be the preferred approach, and despite the many problems with the frequentist approach to hypothesis testing, (Gelman & Carlin, 2017; McShane & Gal, 2017) the problems with Bayesian hypothesis testing are considered more serious (Bernardo, 1999; Cousins, 2017). For example, this sentiment has been expressed as: “we could have followed a Bayesian inference procedure. However, in a clinical trial context, a traditional hypothesis test is expected (due to both this being a common practice and because of regulatory requirements)” (Williamson *et al.*, 2017).

With these considerations in mind, and recognizing the potential advantages of Bayesian methods in this context (Lee and Chu 2012; Spiegelhalter, Abrams, and Myles 2004), we sought to develop and validate a new approach to Bayesian hypothesis testing for RCTs that leverages posterior simulation methods developed over the last few decades. This approach provides a fully Bayesian inference and decision framework for analysis of RCTs. Specifically, this approach (1) employs modern MC simulation to obtain exact posterior distributions with fewer restrictive assumptions than required by current standard methods, allowing for a relatively simple procedure for inference with analytically intractable models, and (2) employs a novel approach to Bayesian hypothesis testing that allows the use of the same priors used for inference (including uninformative priors) and does not suffer from the Jeffreys-Lindley-Bartlett paradox (Mills, 2018).

Categorical and quantitative data from a federally-funded NIH trial of pediatric patients with anxiety disorders (Walkup *et al.*, 2008) were analyzed to validate the proposed methodology and to assess the impact of relaxing restrictive assumptions regarding variance-covariance structure and treatment response in this RCT.

2. Objective Posterior Odds for Bayesian hypothesis testing

The lack of a widely accepted method for precise null hypothesis testing represents a major hurdle to applying Bayesian inferential methods to the analysis of RCTs. Despite the many flaws and misgivings with null hypothesis significance testing (NHST) and the use of p -values, many researchers do not

consider current Bayesian testing methods a viable alternative (Cousins, 2017; Gelman & Carlin, 2017; McShane & Gal, 2017). This section applies a new Bayesian testing procedure (Mills, 2018) to evaluate RCTs; this procedure addresses several of the shortcomings of both NHST and Bayes factors.

Suppose we wish to evaluate the hypotheses $H_0: \delta = 0$ vs. $H_1: \delta \neq 0$, where δ is the difference in average treatment effect (ATE) between treatment and control, or the difference in efficacy for two different treatments. The new testing procedure is derived by replacing the null and alternative hypotheses. For any ε , the null hypothesis is defined as $H_0: |\delta| < \varepsilon$, and the alternative hypothesis is replaced with a set of hypotheses, $H_z: |\delta - \delta_z| < \varepsilon$, $\delta_z \in \{\Theta: \delta_z \neq 0, \delta_z = \delta_z + 2\varepsilon, z \in \mathbb{Z}\}$, where $z \in \mathbb{Z}$ is the set of integers, Θ is the parameter space for δ , and $\delta_z \in \Theta$ for all $z \in \mathbb{Z}$ defines a partition P_z of Θ such that $\delta_z = \delta_{z-1} + 2\varepsilon$. The null hypothesis can then be compared to each alternative hypothesis. Minimizing expected loss over the set of hypotheses defined by the partition leads to the decision rule: reject H_0 if,

$$O_z = \frac{p(\delta - \delta < \varepsilon | D)}{p(|\delta| < \varepsilon | D)} > \frac{c_0}{c_1}, \quad (1)$$

where the posterior odds ratio, O_z , provides the maximum odds against the null hypothesis given data D , $\delta = \operatorname{argmax}_{\delta} p(\delta | D)$ is the maximum a posteriori (MAP) estimate, c_0 is the loss associated with incorrectly rejected the null hypothesis (type I error), and c_1 is the loss associated with incorrectly failing to reject the null hypothesis (type II error). As $\varepsilon \rightarrow 0$, $O_z \rightarrow O$, where

$$O = \frac{p(\delta = \delta | D)}{p(\delta = 0 | D)}. \quad (2)$$

The evidence against the null hypothesis is then evaluated by computing the objective posterior odds (2). This testing procedure does not suffer from the Jeffreys-Lindley-Bartlett paradox and allows the use of the same priors used for posterior inference, so that scientific objectivity can be maintained (Mills, 2018). The outcome of the test is determined by the evidence from the data and any background information incorporated in the likelihood and prior. With a relatively uninformative prior, the prior has little to no influence on the test result.

For precise testing for an unknown mean, critical odds values for O that approximately match 10%, 5% and 1% significance levels are $c_0/c_1 \approx 4:1, 7:1$ and $30:1$ respectively (Mills, 2018). While these rules of thumb may be useful when no other information is available, the derivation of the posterior odds from decision theoretic considerations allows more careful consideration of an appropriate value for c_0/c_1 . This is advisable because it is difficult, and possibly misguided, to provide generic guidelines if the magnitude of effect will vary depending on the situation; a 10% difference may be a substantial improvement in some settings, but trivial in others. There are merits to keeping the determination of statistical significance separate from the determination of magnitudinal significance, and consideration of the relative costs of the type I and II errors, c_0/c_1 , for a particular situation can provide more appropriate guidelines. For example, experimental physicists often follow the '5-sigma' rule, so for a Student- t posterior and 50 observations, $O > 270,000:1$ is required. In psychiatry on the other hand, something close to 5% or even 10% significance may be sufficient to indicate that a therapy with no negative side effects is worthy of consideration by practitioners, especially if failure to treat will likely have serious consequences for the patient. In this case $O \geq 4:1$ may be sufficient to warrant treatment, or at least further study.

The comparison of means for control vs. treatment effects is analytically difficult, resulting in the Behrens-Fisher distribution for quantitative variables, and a complicated distribution for binary variables (Pham-Gia & Turkkan, 1993; Pham-gia, Thin, & Doan, 2017). This becomes intractable without restrictive assumptions for differences in differences comparisons. These difficulties are avoided by drawing a pseudo-sample from the marginal posterior from each group, and subtracting the samples from one another to obtain a sample from the posterior of the difference in means (Lancaster, 2004). For example, for a treatment group sample, x_1 , with unknown population ATE, μ_1 , and variance, σ_1^2 , and a control group sample, x_0 , with unknown ATE, μ_0 , and variance, σ_0^2 , the posterior density for $\delta = \mu_1 - \mu_0$ can be computed by pseudo-random sampling from the posterior density for each sample mean (or proportion if the outcome variables are categorical). This same algorithm can then be used to carry this procedure further and compute differences of the differences, obtaining the posterior density of the difference in efficacy of treatment vs. control from two different treatments, or from one-time period to the next. This allows comparison of ATE differences for two

different treatments or in a treatment at two different time periods, which readily extends to $G > 2$ different treatment groups and $T > 2$ time periods. These advantages of using posterior simulation have not, to our knowledge, previously been fully exploited for analysis of RCTs.

The testing procedure above can be implemented by drawing a MC pseudo-sample from the posterior of δ , evaluating the kernel density at 0 and δ , and computing the posterior odds given by equation (2). This circumvents problems due to analytical intractability and allows computations of the posterior density and odds with far fewer restrictions than are necessary when deriving the posterior analytically. The law of large numbers assures that the expected value of any function of the MC sample converges to its true value, i.e. for a sample of M draws for z , as $M \rightarrow \infty$,

$$\frac{1}{M} \sum_{r=1}^M f(z^r) \rightarrow E(f(z)), \quad (3)$$

where z^r is the r th pseudo-sample value. The accuracy of the simulated posterior density can be increased by increasing the pseudo-sample size, M . Chen (2005) provides guidance on alternative, more efficient posterior density evaluation algorithms when more analytical structure can be placed on the posterior densities, which have been examined in the context of ANOVA testing (Mills & Namavari, 2018).

3. Testing ATE differences with categorical outcome data

Categorical treatment outcomes are common in RCTs and may reflect events (*e.g.*, complication, stroke, etc. or categorical clinical status, such as remission or response). The goal is to evaluate ATE for a particular treatment relative to control group, then to compare ATEs for different treatments or for a particular treatment outcome at different time periods. This section utilizes the hypothesis testing and posterior simulation approach presented in section 2 to develop a procedure for comparison of treated vs. control categorical outcomes in RCTs. This is then extended to allow comparison of average treatment effect (ATE) of different treatments relative to control.

When the outcome of an RCT is either success or not (such as remission or not), this naturally leads to a Binomial likelihood. Standard Bayesian inference using a Beta(a, b) prior with $a = b = 1/2$ or 1

is noncontroversial, resulting in a posterior, $\text{Beta}(s + a, n - s + b)$, for s observed successes in n . When there are $G > 2$ treatment groups (and similarly for $T > 2$ time periods), a Dirichlet, $\text{Dir}(\alpha_1, \dots, \alpha_G)$ is the natural conjugate prior for the multinomial likelihood, leading to a Dirichlet, $\text{Dir}(\bar{\alpha}_1, \dots, \bar{\alpha}_G | s_1, \dots, s_n, n_1, \dots, n_G)$ posterior, where $\bar{\alpha}_g = \alpha_g + s_g$, and s_g is the number of successful outcomes in n_g trials (Gelman *et al.*, 2014).

The observed data for each group are represented by $y_{igt} = 1$, if treatment is effective, 0 otherwise, for individual i in treatment group g at time period t . The number successfully treated out of n_{gt} subjects in group g at time period t is $s_{gt} = \sum_{i=1}^{n_{gt}} y_{igt}$, and the probability of effective treatment for group g at time period t , $p(y_{igt} = 1 | \theta_{gt}) = \theta_{gt}$. With prior $\theta_{gt} \sim \text{Beta}(a, b)$, and likelihood $s_{gt} | \theta_{gt}, n_{gt} \sim \text{Bin}(s_{gt} | \theta_{gt}, n_{gt})$, the posterior is $\theta_{gt} \sim \text{Beta}(s_{gt} + a, n_{gt} - s_{gt} + b)$, with $a = b = 1$ for a uniform prior for θ_{gt} . This results in a treatment and a placebo group posterior density for each of the two treatments. To compare posterior means for θ_{gt} , we proceed numerically using MC simulation. This avoids Behrens-Fisher type problems and allows for unequal and unknown variances across samples. The algorithms are as follows.

Algorithm 1: Two categorical treatment outcomes ATE.

1. Draw $r = 1, \dots, M$ from $\theta_{gt}^{(r)} \sim \text{Beta}(s_{gt} + a, n_{gt} - s_{gt} + b)$,
2. Compute the posterior for ATE in time period t , $\delta_t = \theta_{1t} - \theta_{0t}$ from the MC sample $\delta_t^{(r)} = \theta_{1t}^{(r)} - \theta_{0t}^{(r)}$.
3. Compute posterior moments, highest posterior density (HPD) intervals and posterior density plots from the sample $\delta_t^{(r)}$ to evaluate the ATE for θ_{1t} , vs. θ_{0t} .
4. Test $H_0: \delta_t = 0$, vs. $H_1: \delta_t \neq 0$ using the posterior odds from the MC sample,

$$O_t = \frac{p(\delta_t = \delta_t | n_{1t} s_{1t}, n_{0t}, s_{0t})}{p(\delta_t = 0 | n_{1t} s_{1t}, n_{0t}, s_{0t})}, \quad (4)$$

where δ_t is the MAP estimate of δ_t .

The posterior odds, O_t , provide evidence against the null hypothesis of no difference in ATE. For $G > 2$ categorical treatment outcomes ATE, replace step 1. of Algorithm 1 with drawing from $\theta_{gt} \sim \text{Dir}(\bar{\alpha}_1, \dots, \bar{\alpha}_G | s_1, \dots, s_n, n)$, then compute $\delta_{gt} = \theta_{1gt} - \theta_{0gt}$ for each outcome relative to the control group.

Algorithm 2: Categorical outcome ATE comparison of two treatments.

1. Perform steps 1. and 2. of Algorithm 1 for each of the two treatments to obtain MC samples $\delta_{1t}^{(r)}, \delta_{2t}^{(r)}, r = 1, \dots, M$.
2. Compute the posterior density for the difference in efficacy between treatments from the MC samples, $\Delta_{12t}^{(r)} = \delta_{1t}^{(r)} - \delta_{2t}^{(r)}$.
3. Compute posterior moments, HPD intervals and posterior density plots from the sample $\Delta_{12t}^{(r)}$ to evaluate difference in efficacy between the two treatments.
4. Test $H_0: \Delta_{12t} = 0$, vs. $H_1: \Delta_{12t} \neq 0$ using the objective posterior odds computed from the MC sample,

$$O_t = \frac{p(\Delta_{12t} = \bar{\Delta}_{12t} | n_{11t}, n_{12t}, s_{11t}, s_{12t}, n_{01t}, n_{02t}, s_{01t}, s_{02t})}{p(\Delta_{12t} = 0 | n_{11t}, n_{12t}, s_{11t}, s_{12t}, n_{01t}, n_{02t}, s_{01t}, s_{02t})}, \quad (5)$$

where $\bar{\Delta}_{12t}$ is the MAP estimate of Δ_{12t} .

4. Testing ATE differences with quantitative data

The Bayesian machinery described above can be leveraged for posterior inference and testing with quantitative outcomes. There is not always so obvious a choice for likelihood specification as with Bernoulli trials when the outcome of an experimental trial is a quantitative measure, so some discussion of initial distributional assumption is warranted. An advantage of the proposed approach is that any distribution can be adopted to represent outcomes from the different groups, with possibly different distributional assumptions for different groups. Testing for heterogeneity can be implemented (Ding, Feller, & Miratrix, 2018) with no assumption of homogeneity required with regard to any of the parameters of the different sample distributions. Further, by employing grid approximations to the empirical sample distribution, a nonparametric approach is also possible.

However, experience with data from many RCTs, along with central limit theorem (CLT) and information theoretic (maximum entropy) justifications, indicate that an assumption of normally distributed outcomes is generally reasonable for the standard comparison of means analysis that is typically needed with RCTs, unless further model structure is to be imposed.

Given two samples from an RCT for a treatment group, x_1 , and a control group, x_0 , and assuming the validity of the CLT and the maximum entropy principle to this case, under fairly general regularity conditions, regardless of the distribution of each sample, $x_g = (x_{g1}, x_{g2}, \dots, x_{gn})$, where x_{gi} is the observed outcome for individual i in treatment group g , the likelihood function for the mean of each sample will be approximately Gaussian, $p(x_g | \mu_g, \sigma_g^2) \sim N(\mu_g, \sigma_g^2/n_g)$, $x_g = \sum_i x_{ig}/n_g$, with unknown mean, μ_g , and variance, σ_g^2 . In this case, the sample mean, sample variance, and number of observations are sufficient statistics to fully determine the likelihood. The Jeffreys prior, $p(\mu, \sigma^2) \propto 1/\sigma^2$, $-\infty < \mu < \infty$, $\sigma^2 > 0$, is a standard choice to represent uninformative prior beliefs concerning μ and σ^2 , and results in a Normal-Inverse Gamma (NIG) joint posterior density. An NIG prior can also be used which results in the same NIG posterior functional form. The resulting marginal posterior density for μ_g is,

$$p(\mu_g | x_g) \propto \left(1 + \frac{n_g(\mu_g - x_g)^2}{\nu_g s_g^2} \right)^{-\frac{\nu_g}{2}}, \quad 6$$

which is in the form of a Student- t density with $\nu_g = n_g - 1$, and $s_g^2 = \sum_{i=1}^{n_g} (x_{gi} - x_g)^2 / \nu_g$. The marginal posterior density for σ_g^2 is an Inverted-Gamma, $IG(\nu_g/2, \nu_g s_g^2/2)$ (Gelman *et al.*, 2004). A secondary goal can be to determine if the variances are equal across samples. In a frequentist approach, this is often conducted as a pre-test to decide whether equality of variances can be assumed when testing for equality of means. This pre-testing is unnecessary with the Bayesian approach as posterior simulation allows for inference and testing without any restrictions on the variances in each sample, and without conditioning on particular point estimate values for each variance.

The marginal density for μ_g is generally the posterior of interest. For each of the two samples, x_1 and x_0 , a large Monte Carlo (MC) sample for each mean is obtained by pseudo-random draws from this Student- t distribution, (6). The posterior density of the difference in means for the two samples, $p(\delta|x_1, x_0)$, $\delta = \mu_1 - \mu_0$, is then computed from the difference of the two MC samples using (3). The algorithm for comparison of ATE across treatments or time periods is then similar to **Algorithm 1**, except the initial distribution for posterior simulation is a Student- t instead of a Beta, and the parameter of interest is the mean difference in ATE rather than the mean difference in proportion of successes. For a particular time period, t (with the time period subscript removed to simplify the notation), the algorithm is as follows.

Algorithm 3: Quantitative treatment outcome ATE comparison.

1. Given sample means, x_1 and x_0 , sample standard deviations s_1 and s_0 , and samples sizes n_1 and n_0 , obtain a pseudo-random sample, $\mu_g^{(r)}$, $r = 1, 2, \dots, M$, $g = 0, 1$, from each of the marginal posterior distributions $p(\mu_g|x_g)$, which are Student- t distributions, as in (6).
2. Compute M differences in means for treatment vs. control, $\delta^{(r)} = \mu_1^{(r)} - \mu_0^{(r)}$.
3. Compute the posterior density for the difference in ATE between treatments from the MC samples of differences in means to evaluate ATE for a particular treatment.
4. Test $H_0: \delta = 0$, vs. $H_1: \delta \neq 0$ using the posterior odds computed from the MC sample,

$$O = \frac{p(\delta = \hat{\delta} | x_1, x_0)}{p(\delta = 0 | x_1, x_0)}, \quad (7)$$

where $\hat{\delta}$ is the MAP estimate of δ .

Algorithm 4: Quantitative treatment outcome ATE comparison of two treatments.

1. Perform steps 1. and 2. of Algorithm 3 for each of the two treatments to obtain MC samples $\delta_1^{(r)}, \delta_2^{(r)}, r = 1, \dots, M$.
2. Compute the posterior density for the difference in ATE between treatments from the MC samples, $\Delta_{12}^{(r)} = \delta_1^{(r)} - \delta_2^{(r)}$.

3. Compute posterior moments, HPD intervals and posterior density plots from the sample $\Delta_{12}^{(r)}$ to evaluate difference in efficacy between the two treatments.
4. Test $H_0: \Delta_{12} = 0$, vs. $H_1: \Delta_{12} \neq 0$ using the objective posterior odds computed from the MC sample,

$$O = \frac{p(\Delta_{12} = \bar{\Delta}_{12} | x_{11}, x_{12}, x_{01}, x_{02})}{p(\Delta_{12} = 0 | x_{11}, x_{12}, x_{01}, x_{02})}, \quad (8)$$

where x_{ij} represents the sample from treatment group g (treatment or control) for treatment j (e.g. drug j), and $\bar{\Delta}_{12}$ is the MAP estimate of Δ_{12} .

It is important to emphasize that the posterior distribution obtained in this way is the small sample distribution, so no asymptotic convergence assumptions are required, and no restrictive assumptions on the variances or number of observations of the different samples is imposed. The algorithm extends in an obvious manner to comparing more than two treatments by conducting pairwise comparisons. Mills and Namavari (2018) perform a simulation study for multiple testing with the above algorithm using an ANOVA framework, and find that the procedure performs well in comparison to the standard Welch t -test and a frequentist seemingly unrelated regression (SUR).

5. Sequential Bayesian updating

The Bayesian inferential machinery naturally allows for sequential updating as new data become available, with no requirement to specify a stopping rule before a randomized trial begins, and no need for Bonferroni type multiple testing bias corrections (Berry *et al.*, 2010). This allows for both early stopping if evidence suggests efficacy or not, and for continuing a study if the evidence to date is only suggestive of possible efficacy. Given the costs, both clinical and financial, of conducting randomized controlled studies in many settings, this flexibility is an important feature of Bayesian inference that researchers wish to take advantage of. By combining the Bayesian inferential updating process with the novel hypothesis testing methodology, and fully leveraging posterior simulation methods to allow fewer distributional assumptions, sequential testing and analysis for RCTs becomes viable.

For categorical data, since the Beta distribution is the natural conjugate prior for the binomial likelihood, the posterior density remains Beta with the parameters updated as new data become

available. As in section 3, with a $\text{Beta}(a, b)$ prior, the posterior for one sample, n_1, s_1 , is $\theta \sim \text{Beta } s_1 + a, n_1 - s_1 + b$. For a second sample from the same population, n_2, s_2 , the updated posterior is $\theta \sim \text{Beta } s_1 + s_2 + a, n_1 + n_2 - s_1 - s_2 + b$ (Gelman *et al.*, 2004).

For quantitative data, suppose we have an initial sample of data, x (omitting the group and time period subscript for notational simplicity), consisting of n observations. As in section 4, since there is no compelling reason to think that a Gaussian distribution for the sample mean is inappropriate, CLTs and the maximum entropy principle strongly suggest adopting a Gaussian likelihood, $x \sim N \mu, \sigma^2/n$. Adopting either an uninformative prior distribution, $p(\mu, \sigma^2) \propto \sigma^{-2}$ or if additional prior information is available an NIG prior, leads to a Student- t marginal posterior for μ , $\mu \sim t x, s^2, \nu$, as in equation (6). Suppose a second sample of m observations is obtained from the same distribution, $y \sim N \mu, \sigma^2$. Combining with the likelihood for y , it can be shown that the marginal posterior for μ based on the combined sample $z = x, y$ is $\mu \sim t z, s_z^2, \nu_z$, where, with an uninformative prior, $z = \frac{n\bar{x} + my}{n+m}$, $s_z^2 = \frac{ns^2 + ms_y^2}{n+m}$, $\nu_z = \nu + m$, and $s_y^2 = \sum_{i=1}^m (y_i - \bar{y})^2 / m$ (Gelman *et al.*, 2004). These posterior summary statistics provide sufficient information to completely determine the posterior distribution, so these updating equations can be used as more observations become available and the functional forms of the posteriors remain the same.

This flexibility to examine the evidence from data as it becomes available, with no prior stopping rule required, is an important advantage of Bayesian inferential methods. This advantage carries over to the Bayesian hypothesis testing procedure, so if evidence becomes conclusive earlier than expected (as when a new treatment performs much better than a control or current treatment), the clinical, financial and time savings can be substantial. Along with the evaluation algorithms developed in previous sections, the sequential procedures herein are applied to a large clinical trial of anxious youth in the next section.

6. Application: The Child/Adolescent Anxiety Multimodal Study

In pediatric populations and in certain diseases (*e.g.*, psychiatric disorders), a confluence of psychological, neurobiological, developmental, genetic and social determinants of both symptom

expression and treatment response contribute to high levels of within-sample variance. This heterogeneity further influences drug-placebo separation (Dobson & Strawn, 2016). Additionally, placebo-controlled RCTs may not reflect “highly individualized interventions” (Bothwell *et al.*, 2016), may fail to capture intervention-nonspecific aspects of treatment (Ablon & Jones, 2002) and could delay treatment for some individuals (Kennard *et al.*, 2009). In child and adolescent psychiatry, where the evidence-base for treatment often lags the evidence base for the same interventions in adults, RCTs are urgently needed. Yet, RCTs in children and adolescents are expensive and, in some cases, cost-prohibitive and intensely resource demanding—a particularly important consideration given a near crisis-level shortage of child and adolescent psychiatrists and given that there are few child and adolescent psychiatrists with clinical trials expertise (Walkup, 2017).

To address these limitations, which are often compounded in pediatric clinical trials, strategies to: (1) decrease placebo-response rates (Kowatch *et al.*, 1999; Stein *et al.*, 2006; Nakonezny *et al.*, 2015); (2) increase sample homogeneity; (3) optimize randomization (Lee & Chu, 2012) and (4) mirror clinical practice (*e.g.*, adaptive trial designs) (Almirall *et al.*, 2012) have been proposed. However, the statistical methodology for evaluating drug-placebo differences has remained largely unchanged. The contribution herein provides a new methodology that employs recent advances in statistical technique for an improved analysis, in particular, modern MC simulation methods and a new approach to Bayesian hypothesis testing providing posterior odds ratios in place of frequentist *p*-values.

Antidepressant medications are commonly used to treat anxiety disorders in children and adolescents (Wehry *et al.*, 2015), conditions associated with significant morbidity and mortality. Nearly a dozen RCTs and three meta-analyses (Locher *et al.*, 2017; Strawn, Welge, Wehry, Keeshin, & Rynn, 2015; Wang *et al.* 2017) support the efficacy and tolerability of these medications. However, non-medication-related factors frequently influence drug-placebo differences, and individual clinical trials are often small. Moreover, the type of anxiety disorder under study, age distribution of the participating patients, randomization strategy (*e.g.*, balanced or unbalanced), the number of study sites and inclusion criteria substantially contribute to heterogeneity (Dobson & Strawn, 2016; Nakonezny *et al.*, 2015; Strawn *et al.*, 2017; Varigonda, Jakubovski, & Bloch, 2016). These limitations of current clinical trial design, particularly in pediatric patients, results in attempts to attenuate the influence of these factors by increasing the sample size. In turn, this approach makes these clinical trials more expensive, increases

placebo exposure, prolongs the duration of the clinical trial, and delays the delivery of results to clinicians who ultimately rely on these data to improve treatments.

The largest trial of an antidepressant in pediatric patients with anxiety disorders, the Child/Adolescent Anxiety Multimodal Study (CAMS), evaluated children and adolescents aged 7-17 years with anxiety disorders and compared (1) cognitive behavioral therapy (CBT); (2) the antidepressant, sertraline; (3) CBT + sertraline and (4) placebo. This 5-year study randomized children ($N=488$) (2:2:2:1) to these interventions and was conducted at 6 centers across the United States. The study methods have been extensively described in prior publications (Compton *et al.*, 2010) as have baseline characteristics of the patients (Kendall *et al.*, 2010), acute (Walkup *et al.*, 2008) and long-term outcomes (Piacentini *et al.*, 2014). Response was measured by using categorical and dimensional measures. As described in the initial efficacy study (Walkup *et al.* 2008), categorical response was defined by a score of 1 (very much improved) or 2 (much improved) on the Clinical Global Impression–Improvement Scale (CGI-I) (Guy, 1976) which ranges from 1 to 7 (lower scores reflect greater improvement, as compared with baseline), whereas the Pediatric Anxiety Rating Scale (PARS) score (RUPP, 2002) was the primary dimensional outcome measure for anxiety symptom severity.

The CAMS dataset permits validation of our approach and allows us to assess the impact of relaxing restrictive assumptions regarding variance-covariance structure and medication response. Additionally, the impact of decreasing the sample size can be evaluated in CAMS. Specifically, by leveraging sequential analysis to optimize sample size, we are able to assess the feasibility of conducting the study with a smaller sample.

For the past four decades, successful treatment of psychiatric disorders has been operationalized as a patient having minimal symptoms or impairment and is defined by a CGI-I score ≤ 2 (Guy, 1976). The CGI-I, a clinician-administered instrument, is anchored with the question: “Compared to his or her condition at baseline, how much has he or she changed?” is rated on a seven-point scale. Scores of 1 reflect patients who are “very much improved;” scores of 2 reflect patients who are “much improved;” scores of 3 describe patients who are “minimally improved;” scores of 4 reflect patients who have had “no change;” scores of 5 reflect patients who are “minimally worse;” and scores of 6 and 7 are associated with the descriptions “much worse” and “very much worse,” respectively. The

rating - which is performed by a clinician - is based both on observed and reported symptoms, functional impairment and behavior over a 7 day period.

Figure 1 presents the posterior densities for the ATE for each treatment and placebo group, with the posterior densities for the differences and difference in differences given in the bottom panel. This provides evidence of difference in efficacy between just sertraline relative to placebo compared to sertraline combined with CBT relative to placebo. The posterior odds, Bayesian density tail areas (p -values) and 0.95 HPD intervals in Table 1 indicate that for the sertraline vs. placebo comparison the posterior odds are 16.7:1 against no difference. Posterior odds for treatment that includes sertraline + CBT compared to placebo are greater than 5000:1, providing strong evidence of efficacy for these active treatments (vs. placebo). Finally, posterior odds are 72.8:1 against no difference between sertraline with CBT vs. sertraline monotherapy, controlling for placebo effect.

FIGURE 1: Posterior density functions for means and differences in CGI

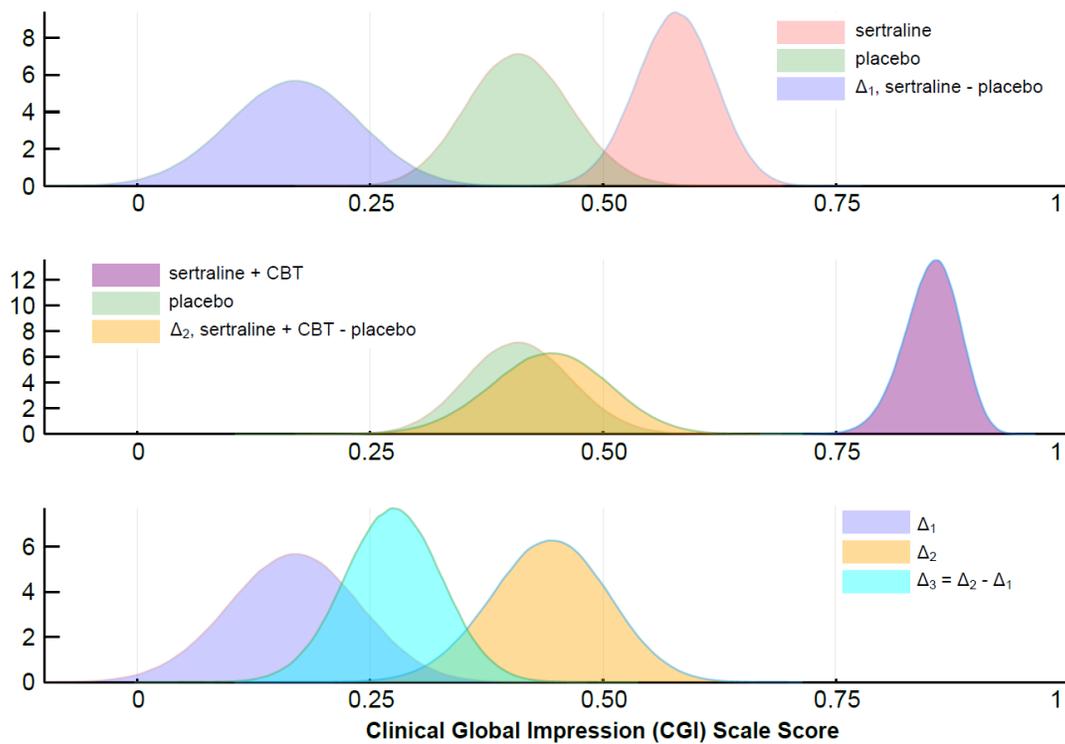


Table 1: Posterior Odds of Difference in Categorical ATE

	Odds against H_0	Bayesian p -value	0.95 HPD
$\Delta_1 = \text{sertraline} - \text{placebo}$	16.69	0.0180	0.03, 0.30
$\Delta_2 = \text{sertraline+CBT} - \text{placebo}$	>50000.0	<0.0001	0.32, 0.56
$\Delta_3 = \Delta_2 - \Delta_1$	72.80	0.0033	0.09, 0.46

The PARS score provides a quantitative measure of improvement in anxiety symptom severity. It consists of a 50-item clinician-rated checklist of anxiety symptoms in children/adolescents in addition to 7 dimensional questions related to anxiety symptom severity (i.e., number of symptoms, severity of symptom distress, behavioral avoidance, interference at home and outside of home) that are rated on a 6-point scale (0 = none to 5 = extreme). Higher scores represent higher levels of distress and anxiety.

FIGURE 2: Posterior density functions for means and differences in PARS

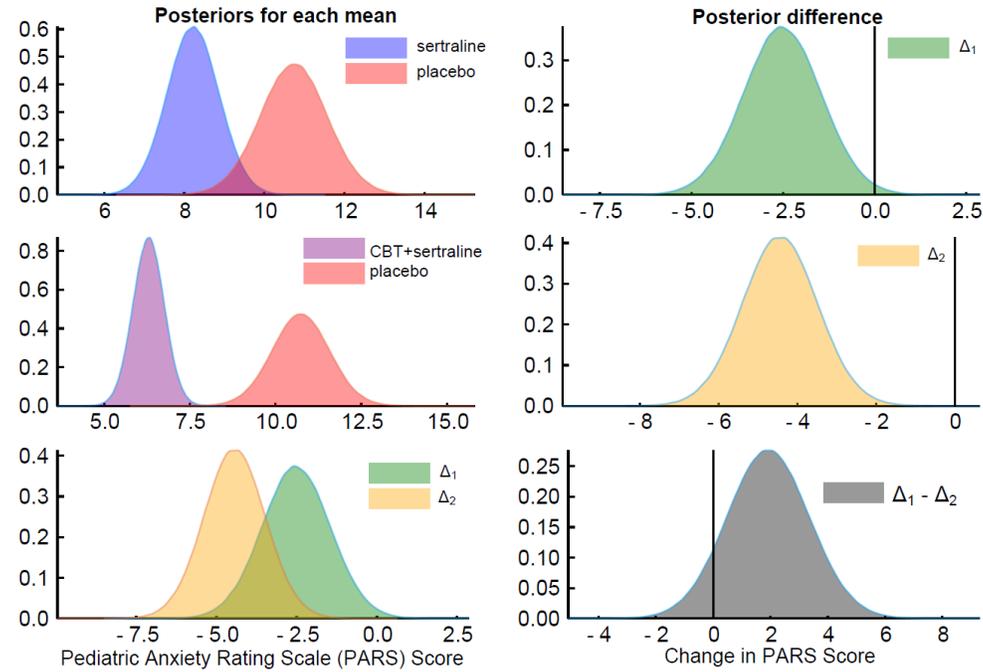


Table 2: Posterior Odds of Difference in Quantitative ATE

	Odds against H_0	Bayesian p -value	0.95 HPD
$\Delta_1 = \text{Sertraline} - \text{placebo}$	15.73	0.0194	-4.63, -0.41
$\Delta_2 = \text{Sertraline} + \text{CBT} - \text{placebo}$	24400.8	<0.0001	-6.94, -1.93
$\Delta_3 = \Delta_2 - \Delta_1$	2.45	0.1832	-0.91, 4.75

Figure 2 presents posterior density functions for mean PARS score and posterior densities of the difference between treatment (sertraline and CBT) groups and control (placebo) groups in CAMS. Posterior odds, Bayesian density tail areas (p -values) and 0.95 HPD intervals are provided in Tables 2. For the quantitative PARS data, there is clear evidence that both sertraline, and the combination of sertraline and CBT provide significant ATE improvements over placebo, whereas there is no statistically significant evidence of a difference in efficacy between the two treatments.

It is of interest that the categorical analysis reveals a statistically significant difference between sertraline + CBT and sertraline monotherapy whereas the quantitative analysis does not indicate a significant difference (odds = 2.45, $p = 0.183$). This difference potentially relates to differences in the instruments. For example, the CGI-I better reflects symptom severity as well as functional impairment, whereas the PARS (quantitative measure) primarily reflects symptom severity. This also contrasts with the original study (Walkup *et al.*, 2008) which leveraged trajectory trend model assumptions, resulting in a likelihood with smaller variance. Adopting the same trend based likelihood would lead to a similar reduction in posterior variance for the methodology proposed herein.

A Sequential analysis comparing difference in efficacy, as measured by PARS for sertraline + CBT vs. placebo, and sertraline vs. placebo provides evidence on potential advantages of the approach. The study used a 2:1 ratio of treatment to placebo. The sequential analysis was therefore conducted starting with 8 treated and 4 placebo subjects and adding an additional 8 treated and 4 placebo receiving subjects each round of the analysis. To illustrate the results, posterior densities are given in **Figure 3** up to the first 192 observations, maintaining the 2:1 treated to placebo ratio.

FIGURE 3: Sequential posterior densities

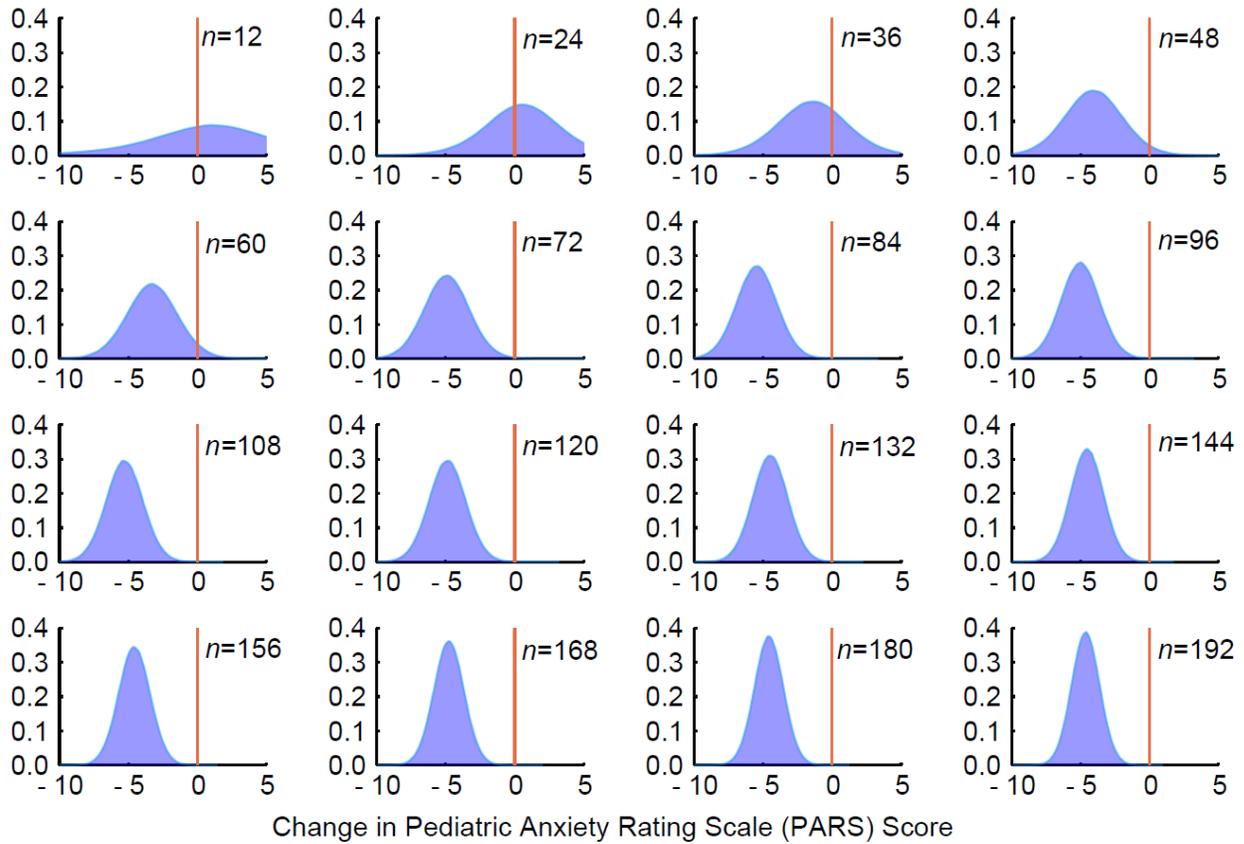


Table 3: Sequential odds against no difference

N	Odds	p-value	N	Odds	p-value
12	1.0	0.7982	108	904.2	0.0003
24	1.0	0.8337	120	402.3	0.0006
36	1.2	0.5753	132	307.6	0.0008
48	6.3	0.0628	144	680.9	0.0003
60	4.8	0.0823	156	1333.6	0.0001
72	64.3	0.0045	168	3252.2	4.8e-5
84	522.5	0.0006	180	4895.7	2.4e-5
96	295.1	0.0009	192	8006.1	2.0e-5

As evident from Table 3, with less than half the sample (84 subjects, 56 treated, 28 placebo), the posterior odds are already over 500:1 against no difference in ATE. Thus, for the detection of the primary outcomes, the study could have ended earlier (1) decreasing placebo exposure, (2) reducing financial cost and (3) resulting in earlier dissemination of the study findings.

7. Conclusion

A new procedure for evaluating the evidence from RCTs has been presented that has several advantages over the prevailing approach. The new procedure exploits posterior simulation methods to allow ease of use with fewer distributional assumptions than previously possible. In particular, there is no requirement to impose any restrictive assumptions about unknown variances from different samples, and by marginalizing with respect to all nuisance parameters rather than conditioning on ‘plug-in’ estimators, the uncertainty due to the unknown parameters is accounted for. Moreover, this approach allows for exact inference with regard to the relative comparative efficacy and tolerability of treatments as opposed to reliance on asymptotic approximations.

A new testing procedure is also introduced that has a number of important advantages over both frequentist testing methods and previously available Bayesian methods. The presentation of odds, as opposed to standard effect sizes and p -values, decreases the possibility of misinterpretation of p -values by practitioners (McShane & Gal, 2017), and does not require a choice between reporting a p -value for one or two-tail areas. Lastly, sequential updating of evidence is easily conducted allowing for more flexible adaptive trial designs.

This approach to analyzing RCTs is applied to the CAMS data and suggests that, had these approaches been utilized, a smaller sample size may have been required and so the study could possibly have been conducted over a shorter period of time. Finally, this procedure has also been used for evaluation of prior psychopharmacologic treatments and comparative efficacy studies (Strawn *et al.*, 2018).

REFERENCES

- Ablon, J.S. & Jones, E.E. (2002). Validity of controlled clinical trials of psychotherapy: Findings from the NIMH treatment of depression collaborative research program, *Am. J. Psychiatry* 159: 775–783.
- Almirall, D., Compton, S.N., Rynn, M. a., Walkup, J.T., & Murphy, S. a. (2012). SMARTer Discontinuation Trial Designs for Developing an Adaptive Treatment Strategy, *J. Child*

- Adolesc. Psychopharmacol. 22: 364–374.
- Bernardo, M. (1999). Nested Hypothesis Testing : The Bayesian Reference Criterion, In: Bayesian Stat., Vol. 6, p. 101–130.
- Berry, D.A. (2006). Bayesian clinical trials, *Nat. Rev. Drug Discov.* 5: 27–36.
- Berry, S.M., Carlin, B.P., Lee, J., & Muller, P. (2010). *Bayesian Adaptive Methods for Clinical Trials*. CRC Press, Boca Raton, FL.
- Bothwell, L.E., Greene, J.A., Podolsky, S.H., & Jones, D.S. (2016). Assessing the Gold Standard — Lessons from the History of RCTs, *N. Engl. J. Med.* 374: 2175–2181.
- Collins, S.P., Lindsell, C.J., Pang, P.S., Storrow, A.B., Peacock, W.F., Levy, P., Rahbar, M.H., Del Junco, D., Gheorghiu, M., & Berry, D.A. (2012). Bayesian adaptive trial design in acute heart failure syndromes: Moving beyond the mega trial, *Am. Heart J.* 164: 138–145.
- Compton, S.N., Walkup, J.T., Albano, A.M., Piacentini, J.C., Birmaher, B., Sherrill, J.T., Ginsburg, G.S., Rynn, M.A., McCracken, J.T., Waslick, B.D., Iyengar, S., Kendall, P.C., & March, J.S. (2010). Child/Adolescent Anxiety Multimodal Study (CAMS): rationale, design, and methods, *Child Adolesc Psychiatry Ment Heal.* 4: 1.
- Cousins, R.D. (2017). The Jeffreys–Lindley paradox and discovery criteria in high energy physics, *Synthese* 194: 395–432.
- Ding, P., Feller, A., & Miratrix, L. (2018). Decomposing Treatment Effect Variation, *J. Am. Stat. Assoc.* 0–0.
- Dobson, E.T. & Strawn, J.R. (2016). Placebo Response in Pediatric Anxiety Disorders: Implications for Clinical Trial Design and Interpretation, *J. Child Adolesc. Psychopharmacol.* 26: 686–693.
- Gelman, A. & Carlin, J. (2017). Some Natural Solutions to the p-Value Communication Problem—and Why They Won’t Work, *J. Am. Stat. Assoc.*
- Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian Data Analysis*, Chapman Texts Stat. Sci. Ser.
- Guy, W. (1976). CGI Clinical Global Impressions, In: *ECDEU Assess. Man.*, p. 217–222.
- Holford, N., Ma, S.C., & Ploeger, B.A. (2010). Clinical Trial Simulation: A Review, *Clin. Pharmacol. Ther.* 88: 166–182.
- Jack Lee, J. & Chu, C.T. (2012). Bayesian clinical trials in action, *Stat. Med.* 31: 2955–2972.

- Kendall, P.C., Compton, S.N., Walkup, J.T., Birmaher, B., Albano, A.M., Sherrill, J., Ginsburg, G., Rynn, M., McCracken, J., Gosch, E., Keeton, C., Bergman, L., Sakolsky, D., Suveg, C., Iyengar, S., March, J., & Piacentini, J. (2010). Clinical characteristics of anxiety disordered youth, *J. Anxiety Disord.* 24: 360–365.
- Kennard, B.D., Silva, S.G., Mayes, T.L., Rohde, P., Hughes, J.L., Vitiello, B., Kratochvil, C.J., Curry, J.F., Emslie, G.J., Reinecke, M.A., & March, J.S. (2009). Assessment of safety and long-term outcomes of initial treatment with Placebo in TADS, *Am. J. Psychiatry* 166: 337–344.
- Kowatch, R.A., Carmody, T.J., Emslie, G.J., Rintelmann, J.W., Hughes, C.W., & Rush, A.J. (1999). Prediction of response to fluoxetine and placebo in children and adolescents with major depression: A hypothesis generating study, *J. Affect. Disord.* 54: 269–276.
- Lancaster, T. (2004). *Introduction to Modern Bayesian Econometrics*. Blackwell, Oxford.
- Lee, J.J. & Chu, C.T. (2012). Bayesian clinical trials in action., *Stat. Med.* 31: 2955–72.
- Locher, C., Koechlin, H., Zion, S.R., Werner, C., Pine, D.S., Kirsch, I., Kessler, R.C., & Kossowsky, J. (2017). Efficacy and Safety of Selective Serotonin Reuptake Inhibitors, Serotonin-Norepinephrine Reuptake Inhibitors, and Placebo in Common Psychiatric Disorders A Meta-analysis in Children and Adolescents, *Jama Psychiatry* 02115: 1–10.
- Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons, *Stat. Med.* 21: 2313–2324.
- March, J., Silva, S., Petrycki, S., Curry, J., Wells, K., Fairbank, J., Burns, B., Domino, M., McNulty, S., Vitiello, B., Severe, J., & Treatment for Adolescents With Depression Study (TADS) Team (2004). Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: Treatment for Adolescents With Depression Study (TADS) randomized controlled trial, *JAMA J. Am. Med. Assoc.* 292: 807–820.
- McShane, B.B. & Gal, D. (2017). Statistical significance and the dichotimization of evidence, *J. Am. Stat. Assoc.* 112: 885–895.
- Mills, J.A. (2018). *Objective Bayesian Precise Hypothesis Testing*, University of Cincinnati.
- Mills, J.A. & Namavari, H. (2018). *Objective Bayesian ANOVA Testing*, University of Cincinnati.
- Nakonezny, P.A., Mayes, T.L., Byerly, M.J., & Emslie, G.J. (2015). Predicting placebo response in adolescents with major depressive disorder: The Adolescent Placebo Impact Composite Score (APICS), *J. Psychiatr. Res.* 68: 346–353.
- Pham-gia, T., Thin, N. Van, & Doan, P.P. (2017). *Inferences on the Difference of Two*

- Proportions : A Bayesian Approach, *Open J. Stat.* 7: 1–15.
- Pham-Gia, T. & Turkkan, N. (1993). Bayesian analysis of the difference of two proportions, *Commun. Stat. - Theory Methods* 22: 1755–1771.
- Piacentini, J., Bennett, S., Compton, S.N., Kendall, P.C., Birmaher, B., Albano, A.M., March, J., Sherrill, J., Sakolsky, D., Ginsburg, G., Rynn, M., Bergman, R.L., Gosch, E., Waslick, B., Iyengar, S., McCracken, J., & Walkup, J. (2014). 24- and 36-week outcomes for the child/adolescent anxiety multimodal study (CAMS), *J. Am. Acad. Child Adolesc. Psychiatry* 53: 297–310.
- Spiegelhalter, D.J., Abrams, K.R., & Myles, J.P. (2004). Bayesian approaches to clinical trials and health care evaluation, *Stat. Pract.*
- Stein, D.J., Baldwin, D.S., Dolberg, O.T., Despiegel, N., & Bandelow, B. (2006). Which factors predict placebo response in anxiety disorders and major depression? An analysis of placebo-controlled studies of escitalopram, *J. Clin. Psychiatry* 67: 1741–1746.
- Strawn, J.R., Dobson, E.T., Mills, J.A., Cornwall, G.J., Sakolsky, D., Birmaher, B., Compton, S.N., Piacentini, J., McCracken, J.T., Ginsburg, G.S., Kendall, P.C., Walkup, J.T., Albano, A.M., & Rynn, M.A. (2017). Placebo response in pediatric anxiety disorders: results from the child/adolescent anxiety multimodal study, *J. Child Adolesc. Psychopharmacol.* 27: 501–508.
- Strawn, J.R., Mills, J.A., Sauley, B.A., & Welge, J.A. (2018). The Impact of Antidepressant Dose and Class on Treatment Response in Pediatric Anxiety Disorders: A Meta-Analysis, *J. Am. Acad. Child Adolesc. Psychiatry* 57: 235–244.e2.
- Strawn, J.R., Prakash, A., Zhang, Q., Pangallo, B.A., Stroud, C.E., Cai, N., & Findling, R.L. (2015). A randomized, placebo-controlled study of duloxetine for the treatment of children and adolescents with generalized anxiety disorder, *J. Am. Acad. Child Adolesc. Psychiatry* 54.
- Strawn, J.R., Welge, J.A., Wehry, A.M., Keeshin, B., & Rynn, M.A. (2015). Efficacy and tolerability of antidepressants in pediatric anxiety disorders: a systematic review and meta-analysis, *Depress Anxiety* 32: 149–157.
- RUPP. The Pediatric Anxiety Rating Scale (PARS): development and psychometric properties. (2002), *J. Am. Acad. Child Adolesc. Psychiatry* 41: 1061–9.
- Varigonda, A.L., Jakubovski, E., & Bloch, M.H. (2016). Systematic review and meta-analysis: Early treatment responses of selective serotonin reuptake inhibitors and clomipramine in pediatric obsessive-compulsive disorder, *J. Am. Acad. Child Adolesc. Psychiatry* 55: 851–859.e2.
- Walkup, J.T. (2017). Antidepressant Efficacy for Depression in Children and Adolescents: Industry-

and NIMH-Funded Studies, *Am. J. Psychiatry* 1–8.

Walkup, J.T., Albano, A.M., Piacentini, J., Birmaher, B., Compton, S.N., Sherrill, J.T., Ginsburg, G.S., Rynn, M.A., McCracken, J., Waslick, B., Iyengar, S., March, J.S., & Kendall, P.C. (2008). Cognitive behavioral therapy, sertraline, or a combination in childhood anxiety. *N. Engl. J. Med.* 359: 2753–2766.

Wang, Z., SH, W., Sim, L., et al (2017). Comparative effectiveness and safety of cognitive behavioral therapy and pharmacotherapy for childhood anxiety disorders: A systematic review and meta-analysis, *JAMA Pediatr.*

Wehry, A.M., Beesdo-Baum, K., Hennelly, M.M., Connolly, S.D., & Strawn, J.R. (2015). Assessment and treatment of anxiety disorders in children and adolescents., *Curr. Psychiatry Rep.* 17: 591.

Williamson, S.F., Jacko, P., Villar, S.S., & Jaki, T. (2017). A Bayesian adaptive design for clinical trials in rare diseases, *Comput. Stat. Data Anal.* 113: 136–153.