Probabilistic Biostatistics with Julia

Jeffrey A. Mills[1] and Jeffrey R. Strawn[1, 2]

[1]University of Cincinnati
[2]Cincinnati Children's Hospital

ABSTRACT

Medical research requires statistical tools that are both sophisticated and powerful enough to address complex inferential problems, as well as intuitive and user-friendly enough to not require advanced statistical and programming expertise. Combining Julia with the Bayesian MCMC machinery addresses this need. Examples of Bayesian probabilistic biostatistics in Julia, including Bayesian adaptive trial design, sequential analysis of randomized controlled trials, and Bayesian hierarchical modeling for a meta-analysis are presented. These examples illustrate the application of a sophisticatedly simple approach to statistical analysis and modeling for clinical research.

Keywords

Julia, Bayesian methods, Meta-analyses, Biostatistics, Randomized Controlled Trials

1. Introduction

Julia is widely recognized as a solution to "the two-language problem": it is both fast and efficient (performance), and easy to use (user friendliness)[3]. Additionally, using the Bayesian MCMC machinery in Julia solves "the two field problem": clinical researchers often need expertise in both medicine and in statistical computing. As those conducting and funding clinical randomized controlled trials (RCTs) recognize the high costs of these studies (e.g., medication expense, time required, and potential exposure of patients to ineffective treatments), there has been greater enthusiasm for (1) improving statistical analytic methods for RCTs, and 2) using evidence-based methods to examine existing naturalistically collected clinical data to inform clinical practice without the need for RCTs. Thus, there is an urgent need to provide statistical tools to clinician-researchers that are intuitive and easy to use, yet sophisticated and powerful enough to answer questions that simpler methods cannot.

The Bayesian machinery of posterior simulation and Markov chain Monte Carlo (MCMC) methods together with Julia offer a solution to this "two-field problem". This enables exact small sample inference and hypothesis testing for complex models without requiring the restrictive assumptions necessary to obtain analytical tractability, and facilitates the analysis of complex models with basic statistical concepts: frequency distributions, density plots, means, medians, modes, standard deviations, quantiles, and posterior density ratios[8].

This paper presents examples from our research developing and applying Bayesian probabilistic approaches to inference and testing in RCTs[8, 11, 12, 13, 14]. Our approach leverages posterior simulation and MCMC methods which, being recursive, require efficient looping, so most available R packages rely on C++ code. This leads to the need to either code in C++ or something similar, or rely on a black box package. Julia offers a user-friendly coding experience similar to python, matlab or R, and many packages that, being written in Julia, allow one to examine and learn from the source code. We are currently developing our own package, BayesTesting.jl[7], along with taking advantage of several Julia packages (e.g. Distributions.jl, DataFrames.jl, CSV.jl and Turing.jl) to apply MCMC and hierarchical models to conduct analyses and meta-analyses of data from RCTs[11].

2. Moving Hypothesis Testing Forward

> At some future time trials may be evaluated using fully Bayesian notions of utilities and decisions ... which would enable designs to be built that do not violate the Likelihood Principle or Bayesian notions. But currently, the regulatory structure is such that confirmatory trials are usually judged and evaluated using Type I error[2], p.220-221.

Despite widespread criticism of p-values[1], the "$p \leq 0.05$ is 'statistically significant', $p > 0.05$ is not" is an iron law for publishing in leading journals in many fields[15]. In general, you cannot successfully publish applied science without precise hypothesis testing. Statisticians, especially Bayesians, argue against that approach, but there is too much institutional inertia towards the use of p-values as 'proper' applied science[15]. At the very least, a bridge from 'pure' hypothesis testing to a more complete Bayesian analysis is needed.

The comparison of means from two samples provides a canonical example. Suppose we have samples from two different treatments, x_1 and x_2, and wish to evaluate whether there is any evidence of a difference in average treatment effect (ATE). This is equivalent to evaluating the precise vs. composite hypotheses,

$$H_0 : \delta = 0, \quad H_1 : \delta \neq 0. \tag{1}$$

where $\delta = \mu_1 - \mu_2$, and μ_j is the ATE for treatment j.
Appealing to the Central Limit Theorem and the Principle of Maximum Entropy provides strong justification for assuming that the distribution of the sample mean, $\bar{x}_j$, for each sample is Gaussian,

$$\bar{x}_j \sim N(\mu_j, \sigma_j^2/n_j), \quad j = 1, 2, \tag{2}$$

where μ_j and σ_j^2 are the unknown population mean and variance of x_j, and n_j is the number of observations in sample x_j. Adopting uninformative priors for the mean and variance leads to conditional

posterior densities,

$$\mu_j | \sigma_j, x_j \sim N(\bar{x}_j, \sigma_j^2/n), \qquad (3)$$

$$\sigma_j^2 | x_j \sim IG((n_j-1)/2, (n_j s_j^2)/2) = Inv - \chi^2(n_j-1, s_j^2). \quad (4)$$

where s_j^2 is the sample variance [4]. This is as far as we need to go analytically. The rest of the problem can be solved numerically (though one more step is easily taken in this case, leading to a Student-t marginal posterior density for μ_j with posterior mean $\bar{x}_j$, variance s_j^2/n, and degrees of freedom $n_j - 1$).

Obtaining M draws from these conditional distributions provides pseudo-samples from the marginal posteriors for μ and σ^2 (or drawing directly from the Student-t marginal posterior for μ). We then obtain an MCMC pseudo-sample of size M from the posterior distribution of δ, $\delta^{(m)} = \mu_1^{(m)} - \mu_2^{(m)}$, $m = 1, ..., M$. In this way, posteriors for any function of the parameters are available, such as differences in differences and ratios. For example, we often wish to compare two treatments to placebo, then to each other, which is accomplished by obtaining $\Delta^{(m)} = \delta_1^{(m)} - \delta_2^{(m)}$. While the analytical distribution of Δ is unknown and asymptotic approximations require unrealistic assumptions, MCMC sampling allows the exact small sample posterior density to be approximated arbitrarily closely as M is increased. The rest of the analysis requires only knowledge of basic statistics: plotting the density of the MCMC draws, computing summary statistics, and testing hypotheses. Note that this addresses the Behrens-Fisher problem of testing the difference between the means of two populations when the variances of the two populations are not assumed either known or equal[10]. Differing sample sizes for each group are also allowed. Correlation across samples and other model extensions are straightforward.

To test the hypotheses in (1) we evaluate the posterior density ratio, PDR, (or 'posterior odds') against H_0, by evaluating the MCMC posterior density for δ at the value in the null hypothesis and at the mode. For treatment samples x_1 and x_2 the PDR is,

$$PDR(\delta = 0 | x_1, x_2) = \frac{p(\delta = \delta_{MAP} | x_1, x_2)}{p(\delta = 0 | x_1, x_2)}, \qquad (5)$$

where δ_{MAP} is the maximum a posteriori estimate of δ. We can also compute posterior tail probabilities (one-sided 'Bayesian p-values'), from the MCMC sample,

$$p\text{-value} = \min\left[P(\delta \leq 0 | x_1, x_2), P(\delta \geq 0 | x_1, x_2)\right], \qquad (6)$$

$$P(\delta \leq 0 | x_1, x_2) = \frac{\sum_{m=1}^{M} I(\delta^{(m)} \leq 0)}{M},$$

$$P(\delta \geq 0 | x_1, x_2) = \frac{\sum_{m=1}^{M} I(\delta^{(m)} \geq 0)}{M},$$

where the indicator function $I(.) = 1$ if the condition is true, 0 otherwise. These same formulas can be used to evaluate hypotheses concerning other quantities of interest, such as Δ. The PDR for joint hypotheses (suppose δ is a vector of parameters) can readily be evaluated using Rao-Blackwellization[7, 8].

3. Bayesian adaptive trial design and sequential analysis

The Bayesian machinery described above can be employed to perform adaptive trial design and sequential analysis of RCTs. In this
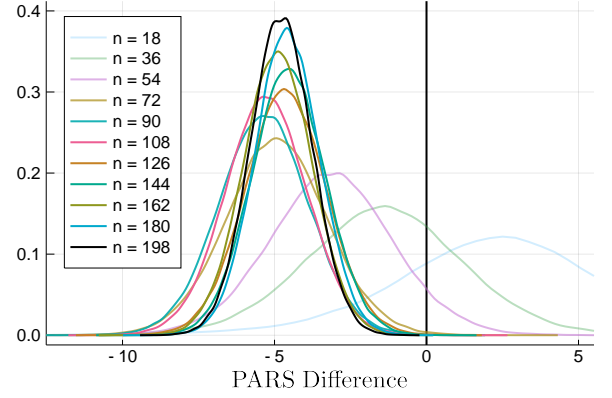

Fig. 1. Difference in ATE between treatment and placebo groups.

context, the Bayesian approach offers the advantage of no statistical requirement for a stopping rule. Though other reasons for a stopping rule are important, such as funding and avoiding bias due to stopping when a test critical value is just reached, a sequential analysis minimizes costs, time and the number of patients exposed to inferior treatment. Further, there is no satisfactory frequentist differences of differences analysis without restrictive assumptions. The analytical sampling distributions are unknown and intractable; asymptotic assumptions are invalid with small samples; and there may be other constraints (e.g., boundary conditions). Bootstrapping is also inferior due to the small sample sizes. The commonly used Welch t-test represents an ad hoc attempt to deal with different variances across samples that does not perform well in simulations[8]. Importantly, these issues are resolved by the Bayesian machinery.

In previous work[12, 8], data from a federally-funded NIH trial of pediatric anxiety disorders (Child/Adolescent Anxiety Multimodal Study [CAMS], N=488) were analyzed to validate the proposed methodology and to examine treatment and placebo responses. In CAMS, youth, aged 7-17 (mean age: 10.7 years) with generalized, separation and/or social anxiety disorders, were randomized (2:2:2:1) to cognitive behavioral therapy (CBT, n=139), a selective serotonin reuptake inhibitor (SSRI), sertraline (SRT, n=133), SRT+CBT (n=140), or pill placebo (PBO, n=76).

The following example sequentially analyzes the data for the combined treatment SRT+CBT relative to PBO using Bayesian updating. To evaluate the hypothesis in (1) where μ_1 is the ATE of SRT+CBT and μ_2 is the ATE of PBO. Starting with a sample of 12 (8 treated, 4 placebo), the posterior density was updated with each additional 6 observations (4 treated:2 placebo). The sequence of posteriors as n is increased for the difference in ATE between SRT+CBT and PBO is shown in Figure 1, and the PDR giving posterior odds against the null hypothesis (5), posterior probability that the difference is greater than zero (6), and two sided p-value ($\approx 2 \times p(\delta \geq 0)$) are given in Table 1.

There is clear evidence of a difference between SRT+CBT and PBO, δ, by $n = 72$ (posterior odds against H_0:$\delta = 0.0$ are 61.4:1, $P(\delta \geq 0 | s, n) = 0.0025$,), suggesting that for this treatment comparison the clinical trial could have been completed with less than half of the original sample with no change in the conclusions.
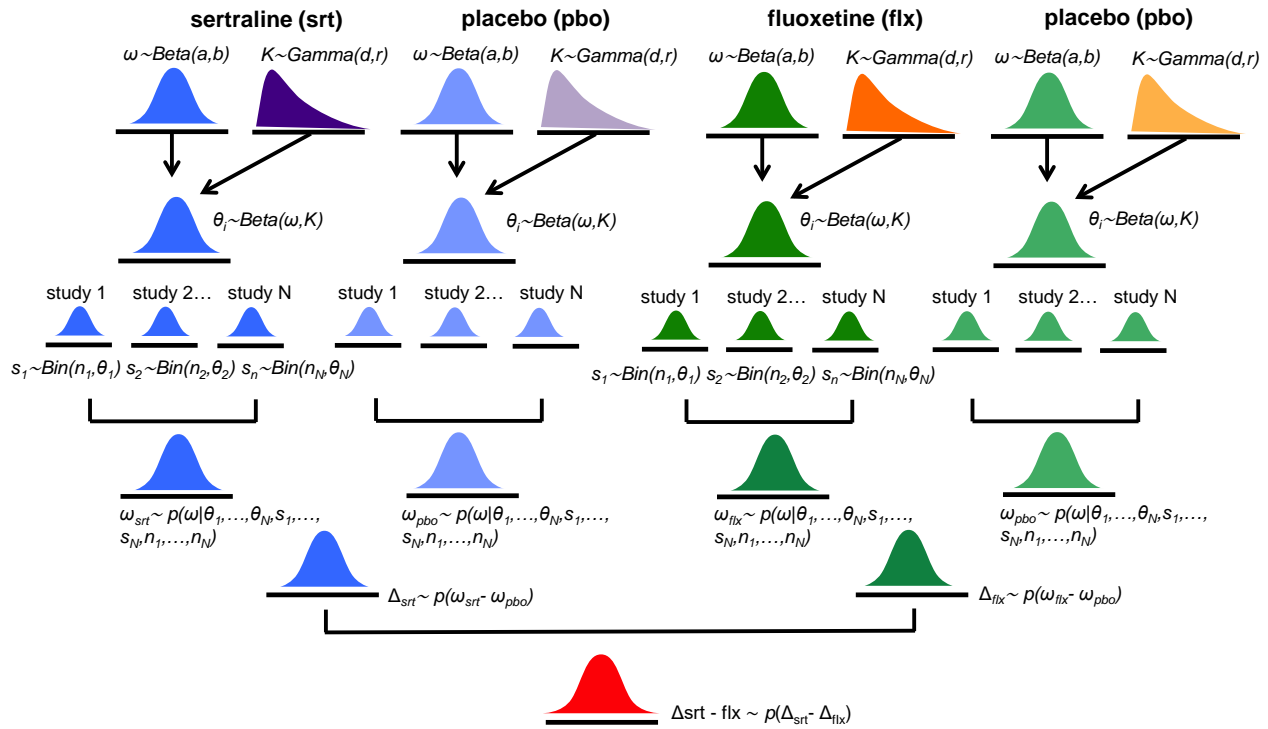
Fig. 2. Bayesian Hierarchical Model for Treatment Comparisons.

Table 1. Evidence Against $H_0 : \delta = 0$

n	$PDR(\delta = 0)$	$P(\delta \geq 0)$	p-value
18	1.37	0.2382	0.4763
36	1.19	0.2887	0.5775
54	3.49	0.0625	0.1250
72	61.35	0.0025	0.0051
90	572.96	0.0003	0.0005
108	1083.5	0.0001	0.0002
126	329.24	0.0005	0.0010
144	610.67	0.0002	0.0004
162	4596.8	0.0000	0.0000
180	2394.6	0.0000	0.0000
198	46452.9	0.0000	0.0000

4. Bayesian hierarchical modeling

A Bayesian hierarchical model (BHM) provides a flexible setting for modeling data that allows for heterogeneity across individuals or groups. If homogeneity across studies is assumed, the data can be combined via Bayesian updating as in the sequential analysis in the previous section. For categorical data, all that is needed are the number of occurrences, s_i, in n_i subjects in each study, i, of N studies. The posterior distribution is then

$$p(\theta | s, n, a, b) = \text{Beta}(s + a, n - s + b), \qquad (7)$$

where $s = \sum_{i=1}^{N} s_i$ and $n = \sum_{i=1}^{N} n_i$, with $a = b = 1$ for a uniform prior for θ.

Alternatively, BHMs can allow for heterogeneity across studies due to different trial sites, different treatments, such as different

medications, etc., and the same Bayesian machinery can be employed to compare a variety of treatments relative to placebo. For the categorical data examined herein there are two common BHM specifications: the Beta-Gamma[5], and the Logistic-Normal[6]. In the following, we adopted the Beta-Gamma BHM specification as it directly provides intuitively understandable parameter estimates, whereas the logistic specification is more difficult to interpret (though is better suited when covariates are included).

The complete Beta-Gamma BHM modeling framework is illustrated in Figure 2 for the case of comparing two sets of RCTs for different treatments (sertraline vs. fluoxetine) with placebo groups in each RCT. Each RCT in Figure 2 has a Binomial likelihood for s_i successes in n_i trials with probability of success θ_i (row 3). A common Beta prior with mode ω and precision parameter K is assigned to each θ_i (row 2), and a Beta and Gamma hierarchical prior are assigned to ω and K respectively (row 1). The hyperparameters a, b, d and r are selected to represent a priori knowledge about the distributions of ω and K (relatively uninformative priors are usually assigned unless other information is available).

The package Turing.jl provides Hamiltonian Monte Carlo (HMC) and MCMC algorithms to obtain posterior density MCMC chains from an large variety of models. The Julia macros provide a flexible and intuitive modeling and inference framework. The model specification in Turing.jl closely matches how one would write the model mathematically, e.g. for the Beta-Gamma BHM (Turing code examples for the Logistic-Normal are available at github.com/TuringLang/Turing.jl and github.com/StatisticalRethinkingJulia):

```
using Turing, MCMCChains
@model binomial_trials(s,n) = begin
    g = length(n)  # number of groups
    # hierarchical priors
    ω ~ Beta(2,3)
    K ~ Gamma(10,1/0.05)
    a = ω*K + 1.0
    b = (1.0 - ω)*K + 1.0
    # priors for each occurrence rate
    θ = Array{Real}(undef, g)
    for k in 1:g
        θ[k] ~ Beta(a,b)
    end
    # likelihood
    for i in 1:g
        s[i] ~ Binomial(n[i],θ[i])
    end
end
```

A few more lines of code provide posterior samples from the No U-Turn (NUTS) sampler in Turing.jl, allowing evaluation of the no mean difference between two groups hypothesis:

```
using BayesTesting

# treatment
ctrt = mapreduce(c->sample(binomial_trials(s2, n2),
    NUTS(5000,1000,0.65)),chainscat,  1:5)
# placebo
cpbo = mapreduce(c->sample(binomial_trials(s1, n1),
    NUTS(5000,1000,0.65)), chainscat,  1:3)

ωt = Array(ctrt["ω"][1001:end])  # treatment ω
ωp = Array(cpbo["ω"][1001:end])  # placebo ω
δ = ωt - ωp               # difference
plot(δ,st=:density,label="Difference")

# compute summary stats, PDR odds, tail prob.
[mean(δ) std(δ)]
quantile(δ,[0.025,0.5,0.975])
[mcodds(δ, h0=0.0) bayespval(δ)]
```

In recent work[9], a meta-analysis of results from several RCTs was conducted to obtain more complete evidence on the expected side effects of SSRIs. Side effect or adverse event (AE) occurrence in an RCT is generally recorded as a categorical variable, leading to binomial data consisting of number of AEs, s_i, in n_i subjects for trial i. The results below are for the "activation" (or restlessness) AE. Five studies included data for activation. Posteriors for each of the five studies for treatment (with SSRI) and placebo were obtained using Turing.jl (the third row of distributions in Figure 2), along with the distribution of the mode, ω, of the posterior for the estimated rate of success for each study, θ_i (the fourth row in the Figure 2). Given the MCMC pseudo-samples for each ω (e.g., SSRI-related side effects and placebo-associated side effects), outcome differences between treatment and placebo, then differences in differences between the two treatments relative to placebo (e.g. SSRI vs. SNRI-related side effects) can be determined (rows 5 and 6 in the Figure 2).

Using the NUTS sampler in Turing.jl, the BHM was estimated for each treatment arm of the study data. Run times to generate 5 chains using the above code were approx. 7 seconds or less on a laptop with CPU: Intel(R) Core(TM) i7-6600U CPU @ 2.60GHz running Julia Version 1.1.0. The stability of the chains and similarity of each chain density suggests convergence to the marginal posteriors has been attained, as illustrated in Figure 3.

The resulting posterior densities for each study probability of occurrence rate (risk), θ_j and the hierarchical probability of occur-
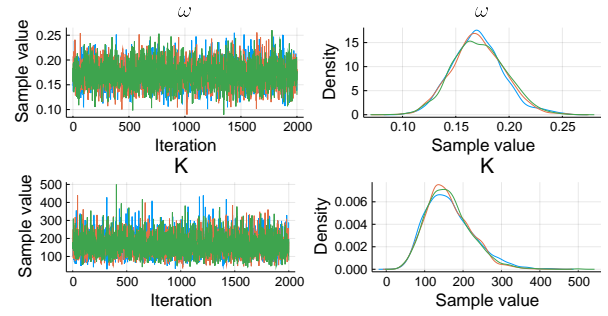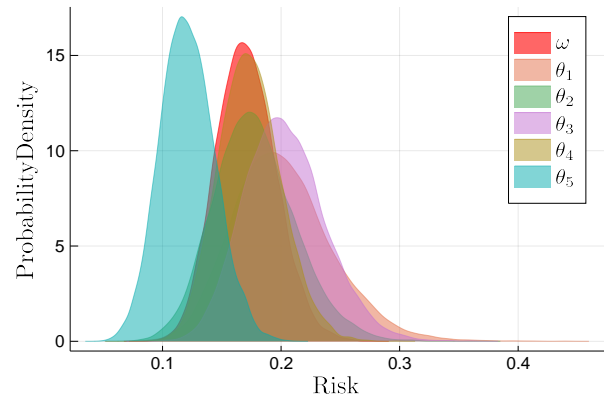

Fig. 3. HMC chains from No U-Turn Sampler.
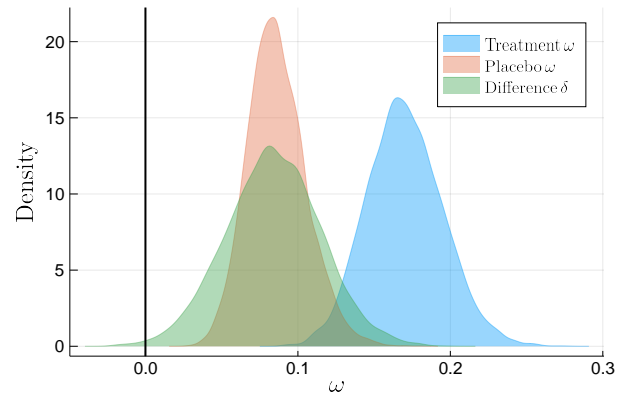

Fig. 4. Posterior Densities for Activation

Fig. 5. Treatment-Placebo Difference

rence across groups, ω are illustrated in Figure 4. Results using the logistic-normal specification gave very similar results. The results for activation due to SSRI (mean difference between treatment and placebo, δ), given in Figure 5 and Table 2, provide statistical evidence of an increased likelihood of activation with the treatment. To examine the impact of across study heterogeneity, Figure 6 presents the BHM posterior density for ω in comparison to combining all studies ignoring across study heterogeneity via equation (7).

Table 2. Posterior difference in treatment and placebo, δ

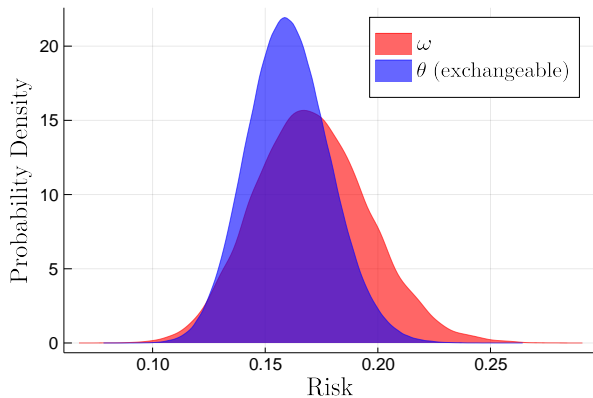| mean	SD	$PDR(\delta = 0)$	$P(\delta \leq 0)$	0.95 CrI	
0.084	0.031	35.75	0.003	0.023	0.145

Fig. 6.   ATE posterior density with and without heterogeneity

5.   Summary and Conclusions

This paper presented two examples of application of Bayesian probabilistic modeling from our research that illustrate our experiences with Bayesian inferential methods for clinical research using Julia. We have used this approach in several studies including: reevaluating the evidence from previously conducted RCTs; analysis of abandoned trials; joint evaluation of tolerability and efficacy in RCTs; and BHMs for meta-analysis evaluating adverse events ("side effects") in trial participants.

The approach presented provides a solution to the strong institutional inertia of '$\leq 5\%$ = statistical significant' through provision of PDRs (or 'posterior odds') as well as posterior tail probabilities ('Bayesian p-values'), posterior density intervals, and visualization of posterior densities. This allows for more flexibility in the choice of critical value for a particular test, i.e. the cut-off point for rejecting vs. failing to reject the null hypothesis. For example, researchers at CERN attempting to detect gravitational waves would wish to employ a much larger critical odds ratio (akin to the '5-sigma rule'), whereas researchers comparing the efficacy of two relatively harmless psychiatric treatments for anxiety or depression would undoubtedly find posterior odds that pass a much lower critical threshold convincing enough to recommend one treatment over another.

One of the major advantages of the contemporaneous Bayesian approach is its ability to utilize MC and MCMC methods. This approach is computationally intensive, but less mathematically and analytically burdensome. Moreover, it typically requires fewer restrictions than needed for analytical tractability. We find Julia to be ideal for scientific programming of this nature; it reduces the need for the researcher to wear multiple expertise 'hats'. The resulting conservation of clinical researchers' time and energy by using Julia is substantial, and allows greater focus on the scientific and clinical problems. Ultimately, this potentially hastens the arrival of effective treatments and findings to the clinic.

6.   References

[1] Donald Berry. A p-Value to Die For. *Journal of the American Statistical Association*, 112(519):985–897, 2017.

[2] Scott Berry, Bradley Carlin, Jack Lee, and Peter Muller. *Bayesian Adaptive Methods for Clinical Trials*. CRC Press, 2011.

[3] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

[4] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald S. Rubin. *Bayesian Data Analysis*. CRC Press, 2004.

[5] John Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan 2nd Edition*. CRC Press, 2011.

[6] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press, 2015.

[7] Jeffrey A. Mills. BayesTesting.jl: Bayesian Hypothesis Testing without Tears, doi:0.13140/RG.2.2.18951.70564. *University of Cincinnati*, 2018.

[8] Jeffrey A Mills, Gary C Cornwall, Beau A Sauley, and Jeffrey R Strawn. Bayesian Inference and Testing for Randomized Controlled Trials: a Posterior Simulation Approach. *University of Cincinnati*, 2019.

[9] Jeffrey A Mills and Jeffrey R Strawn. Antidepressant Tolerability in Pediatric Anxiety and Obsessive Compulsive Disorders: A Bayesian Hierarchical Modeling Meta-Analysis. *Journal of the American Academy of Child and Adolescent Psychiatry, (*under revision*)*, 2019.

[10] Patrícia de Siqueira Ramos and Daniel Furtado Ferreira. A Bayesian solution to the multivariate Behrens-Fisher problem. *Computational Statistics and Data Analysis*, 54(6):1622–1633, 2010.

[11] Jeffrey R Strawn, Jeffrey A Mills, and Paul E Croarkin. Switching Selective Serotonin Reuptake Inhibitors in Adolescents with Selective Serotonin Reuptake Inhibitor-Resistant Major Depressive Disorder: Balancing Tolerability and Efficacy. *Journal of child and adolescent psychopharmacology*, 29(4):250–255, 2019.

[12] Jeffrey R Strawn, Jeffrey A Mills, Beau A Sauley, and Jeffrey A Welge. The Impact of Antidepressant Dose and Class on Treatment Response in Pediatric Anxiety Disorders: A Meta-Analysis. *Journal of the American Academy of Child and Adolescent Psychiatry*, 57(4):235–244, 2018.

[13] J.R. Strawn, E.T. Dobson, J.A. Mills, G.J. Cornwall, D. Sakolsky, B. Birmaher, S.N. Compton, J. Piacentini, J.T. McCracken, G.S. Ginsburg, P.C. Kendall, J.T. Walkup, A.M. Albano, and M.A. Rynn. Placebo Response in Pediatric Anxiety Disorders: Results from the Child/Adolescent Anxiety Multimodal Study. *Journal of Child and Adolescent Psychopharmacology*, 27(6), 2017.

[14] J.R. Strawn, J.A. Mills, G.J. Cornwall, S.A. Mossman, S.T. Varney, B.R. Keeshin, and P.E. Croarkin. Buspirone in Children and Adolescents with Anxiety: A Review and Bayesian Analysis of Abandoned Randomized Controlled Trials. *Journal of Child and Adolescent Psychopharmacology*, 28(1), 2018.

[15] Ronald L Wasserstein, Allen L Schirm, and Nicole A Lazar. Moving to a World Beyond "p < 0.05". *American Statistician*, 73(sup1):1–19, 2019.