Bayesian Analysis of Randomized Controlled Trials

Julian Bautista*, Alex Pavlakis*, Advait Rajagopal*

January 6, 2018

Abstract

*The New School for Social Research, 6 E 16 St, New York, NY, 10003.

Contents

1 Introduction

Bayesian methods are gaining popularity in many fields because they allow researchers to incorporate all available information into flexible and transparent statistical models. Many researchers have begun to incorporate Bayesian methods into medical, pharmaceutical, and social-science research. The purpose of this paper is to show how Bayesian methods can be the standard in applied research. We provide a general overview of the approach and an example analysis of Randomized Control Trial (RCT) on the impact of a smartphone application on eating disorder behavior.

- Andrew Gelman, chapter 9, RCTs and why Bayes?
- Refer other people who did this
- Tom, MC stuff, accounts for the excess of zeros [mention both Tom's papers here]

  Randomized controlled trials in the context of treating binge-eating disorder and bulimia nervosa have been studied previously. It is important to understand the nature of the variable of interest, objective binge eating (OBE). Grotzinger, Hildebrandt and Yu [2016] have a very clear exposition of the problem. Moreover they address some of the modeling issues that arise due to specific properties of binge eating data. OBE refers to the number of binge eating episodes captured at different stages in the treatment process. That implies that OBE data is count data and therefore discrete. Such data is also characterized by a large number of zeros which indicates remission. This makes the data highly positively skewed with an excess of zeros as many participants show remission even early in the treatment process. In Grotzinger, Hildebrandt and Yu [2016] they explore a semi-continuous treatment of the data and discuss a latent variable growth curve model to estimate a growth factor over time. Zero inflated Poisson (ZIP) and zero inflated negative binomial models (ZINB) are also possibilities suggested by them for an appropriate treatment of the positively skewed data with excess zeros. While these approaches are correct for the problem, we believe that using Bayesian hierarchical models to capture treatment effects in RCT's is the best way forward and that is the primary contribution this paper makes to the literature. We use multilevel models, with carefully selected prior distributions to set-up a modeling framework that can be easily adopted for the class of problems with RCT's and clinical trials. We use the dataset studied by Hildebrandt et al. [2017] to test the effectiveness of a smartphone app on eating behavior (explained fully in Section 4).
- Our improvement, is that we are rigorous, reproducible, open to checking, criticism
- Bayesian mehtods allow us to be transparent with model specs and assumptions (see Section 2)
- heterogenous treatment effects (flexibility and benefits of hierarchical structures)
- Mention benefits of priors and how past studies help with prior info, this is perfect for Bayes

- What does each section do? end.

2 Bayesian Data Analysis

Bayesian data analysis (BDA) and inference is the process of developing and fitting a probability model to data. The result is learning the probability distribution of the parameters of the model and being able to evaluate the fit of the model to the data as well as being able to make predictions for new observations [Gelman et. al 2013; Gelman and Hill 2007].

There are three main steps to BDA, which are listed below and explained in detail in section 2.1, 2.2 and 2.3 respectively.

1. Set up a probability model.
2. Calculate the distribution of the model parameters, given the observed data. This is called the *posterior distribution* of the parameters.
3. Check the fit of the model, whether the conclusions are sensible and how sensitive they are to modeling assumptions. Expand or alter the model if needed and repeat the steps.

2.1 Model development

Model development involves setting up a joint probability distribution that accounts for all observed data and unobserved parameters. The model should include all knowledge of the experiment or data collection process and should be logically consistent with scientific nature of the problem at hand. We approach model development in three steps.

1. **Exploratory Data Analysis**: Look at your data. What distributions do different variables appear to take? What seems important? What variables are correlated? Answers to these questions are essential to specifying a the following two model components. For most scientific problems there is no one-size-fits all model or reliable automated model choosing program; you have to get your hands dirty. The process of exploratory data analysis and plotting informs the choice of likelihood function and gives insight into what priors are appropriate for the problem. It allows the researcher to explore and solidify intuition about the problem at hand and the nature of the data.
2. **Setting up a Likelihood Model**: Choose a model that represents the relationship between your data and parameters of interest. A clear understanding of the type of data is an important prerequisite for picking a likelihood function. Data can be binary, categorical, ordinal, count, or continuous and each of these types of data requires a different kind of treatment. For instance, if your outcome variable is binary (0 or 1), a logistic regression framework may be appropriate. If data is continuous and defined on the entire real line, a normal likelihood function might fit the bill better. In this paper, the dependent variable is count data and thus we employ a Poisson likelihood. Again there may not be one perfect or “correct” choice of a function and it is common

practice in Bayesian analysis to set up different models and compare them and see which performs better.

3. **Choosing a prior distribution**: Bayesian analysis requires the assignment of a prior distribution. A prior distribution serves three functions. First, it regularizes the parameter space by specifying what values the parameters can take and more importantly what values it definitely cannot assume. Second, it makes assumptions about the underlying scientific nature of the problem explicit. Third, it facilitates the calculation of a posterior distribution and makes it possible to have posterior simulations or "draws" from the posterior. Prior choice depends on the parameter or coefficient of interest. We could assign a completely "noninformative" or flat prior to our coefficient which is equivalent to saying that the coefficient is a draw from a uniform distribution on the whole real line. This boils down to a maximum likelihood estimate of the coefficient. But this is rarely a useful approach. Statisticians have information about the parameter and can use "weakly informative" or "specific informative priors" to reflect the researcher's knowledge. For instance, if a coefficient of interest *must* be between 0 and 1, we can assign a *Beta* prior distribution to that coefficient. If we believe that a coefficient is close to zero, but may be positive or negative, we could assign a more informative *Normal*(0, 1) prior distribution to it.

   Get citations here
   `https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations`

To make this concrete, let θ be a parameter of interest (or a vector of parameters) and y be data from which we want to estimate θ. Then the prior distribution is represented as $p(\theta)$ and the likelihood is $p(y|\theta)$. Our goal is to estimate the *posterior distribution* of the parameter conditional on the data and prior information, denoted by $p(\theta|y)$. Using Bayes' rule we can say that;

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{\int_\theta p(\theta)p(y|\theta)d\theta}$$

The numerator is a product of the prior and likelihood distributions. The denominator is what we call the "evidence" and is a constant value when integrated over the range of θ. Thus using proportionality we can approximate the posterior distribution up to a normalizing constant in the following manner.

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

We use this approximation to the posterior because calculating the denominator as a closed form integral is very difficult and may not be possible. We address how to deal with these difficulties in section 2.2. The Bayesian approach combines data and prior information in a way that can yield more precise inferences that either on its own [**cite Gelman here**].

2.2 Model estimation

The goal of model estimation is to calculate the *posterior distribution* of the parameters. This is the conditional distribution of the parameter given the observed data. The posterior distribution is obtained by simply multiplying the prior and likelihood to get joint probabilities and then normalizing to get posterior probabilities. Calculating the posterior analytically is often difficult and leads to intractable integrals with no closed-form solution. So the standard practice is to use Markov Chain Monte Carlo sampling methods to approximate the posterior up to a normalizing constant and sample from it. There are some other approaches to calculating the posterior distribution, but those are beyond the scope of this paper. We recommend the use the Bayesian probabilistic modeling language Stan. Stan uses a Hamiltonian Monte Carlo sampling algorithm (from a broader class of MCMC sampling methods) to approximate the posterior distribution. Betancourt [2017] has a clear exposition of how the algorithm works. It is sufficient that Stan returns the joint posterior of all parameters conditional on data from which we can obtain the marginal posteriors of desired parameters. Stan also has the ability to generate posterior simulations while calculating the posterior and these can be used for model checking and validation. Using these posterior distributions we can now predict and simulate replicated data from our model, which leads us to section 2.3.

2.3 Model checking, comparison and expansion

We begin evaluating the performance of our model by ensuring that it converges on robust parameter estimates and that the results make sense given what we know about the observed data and the problem itself. We also need to see how sensitive our results are to the modeling assumptions and the priors. Sometimes using stronger priors can help with convergence and more stable results. We use our posterior distributions of the parameters to check our model by carrying out *posterior predictive checks.* This could be graphical checks where we simulate "fake" data from our model and compare it graphically to the raw data. We could also use certain test statistics and compare the values of those statistics across models and see how close they are to the ground truth obtained from the observed data itself. Based on the fit of the model, we can either expand the model to include more parameters, predictors or change the model and reevaluate performance. We are not looking to explicitly choose the "right" model but are interested to learn where each of our approaches and models might be lacking and make these shortcomings explicit while trying to improve them.

3 Comparison of Bayesian Data Analysis to Other Methods

There are two primary benefits to a Bayesian approach for the analysis of Randomized Controlled Trials. First is the improvement of predictions through the use of a greater amount of information within the models. Second is a greater level of transparency through explicit assumptions. This section will end with a discussion of some of the risks of Bayesian approaches.

3.1 Better Prediction

As discussed earlier, the choice of a prior distribution is a critical assumption that directly affects the results of the analysis. Bayesian analysis lends itself well to the infinitely many distributions that can be used. But beyond the choice of a distribution, the specifications of hyper-paramaters of these distributions allow for the usage of data that exists beyond that which was collected within the study.

Data does not live only on the spreadsheet. It is contained in the researcher's knowledge of the literature, the data generating process, and anything else that is relevant. This can include common sense knowledge such as the understanding that estimates on ratios will be between 0 and 1, or it could be an expected estimate based on previous studies done. Regardless of the source of the information, knowledge about an estimator can be translated into the prior to improve accuracy.

One of the greatest advantages of specifying prior distributions in a hierarchical or multilevel model is to allow for partial pooling. Hierarchical data structures include any data that can be represented in categories. In a non-Bayesian scenario, there are two choices a researcher has: a complete pooling model, and a no pooling model. With complete pooling, the categories in data are completely interchangeable, thus ignoring the uniqueness of categories. With no pooling, each category is treated as independent of the others, thus ignoring the interrelatedness of each of the categories. Setting up a prior distribution allows us to say that our relevant parameters have a common mean, but are each different from one another. This partial pooling is often more accurate as it takes into account both the uniqueness and interrelatedness of categories within a hierarchical model.

Unlike traditional approaches, the Bayesian approach does not output a point estimate with confidence intervals. Instead it estimates a joint posterior distribution (correct up to a normalizing constant) for the parameters conditional on data. This distribution contains all the information from the prior and the model. From this joint posterior we can obtain the marginal distribution of each parameter with credible posterior intervals. This provides an accurate account of the inherent uncertainty in the estimation of the parameters. Generating a posterior also allows for simulation using posterior draws from this distribution. This is helpful for model checking, expansion and further analysis.

3.2 Transparency

Bayesian analysis also lends itself to better scientific scrutiny on models overall. This occurs because non-Bayesian techniques and traditional frequentist models make tacit assumptions. Bayesian methods make underlying modeling assumptions explicit in the form of a prior. For example, a simple linear regression tacitly assumes that the parameter is normally distributed and has a uniform prior. While this assumption may be appropriate for some problems, in many occasions it does not get scrutinized closely because it is an implied assumption. Meanwhile, the power of a prior to influence estimates puts a magnifying glass on all of these issues that were easily glossed over previously. Close scrutiny is a critical component for rich academic discourse. It puts the onus on researchers to justify the assumptions of their model and allows for vibrant discussion.

3.3 Risks of a Bayesian Approach

Bayesian analysis is not without its downsides, the obvious being the use of bad priors to generate biased estimates. However, this risk is mitigated substantially by close scrutiny on priors. A more problematic issue is the interpretability of posteriors. While it's possible to generate point estimates and confidence intervals, the posterior itself is difficult to interpret. Aside from this, Bayesian inference inherently stops being an encapsulation of a single dataset, as we begin to input information from other sources including the researcher. What excess information is and is not valid is up for open debate.

4 Impact of Smartphone App on Eating Behavior

Hildebrandt et all (2017) conducted an experiment to test whether the Noom Monitor, a smartphone application, could augment the effect of in-person therapeutic treatment on binge eating behavior. The treatment, known as *guided self-help treatments based on cognitive-behavior therapy* (CBT-GSH), had been shown in previous research to reduce binge eating behavior by 10-50%. The Noom Monitor application was designed to facilitate CBT-GSH. For this example, we consider two research questions from the experiment:

1. Is CBT-GSH more effective at reducing binge eating behavior when facilitated by the Noom Monitor?
2. Does the effect of the Noom Monitor vary over time?

4.1 Experimental design

66 men and women with Bulimia Nervosa (BN) or Binge Eating Disorder (BED) were randomly assigned into two treatment conditions: CBT-GSH (N= 33) or CBT-GSH + Noom (N=33). Therapy lasted for 12 weeks. Assessments were conducted at weeks 0, 4, 8, 12, 24, and 36. The primary outcome was Objective Bulimic Episodes (OBE).

4.2 Exploratory data analysis

Figure 1 displayes OBEs per week for each individual in both treatment conditions. A few aspects of the data immediately stand out, which suggest that any model should account for individual-level effects and time-level effects, and should let treatment effects vary over time.

- The number of OBEs decreases over the course of the treatment for almost all subjects
- The biggest decreases in OBEs appear to occur in the early stages of treatment

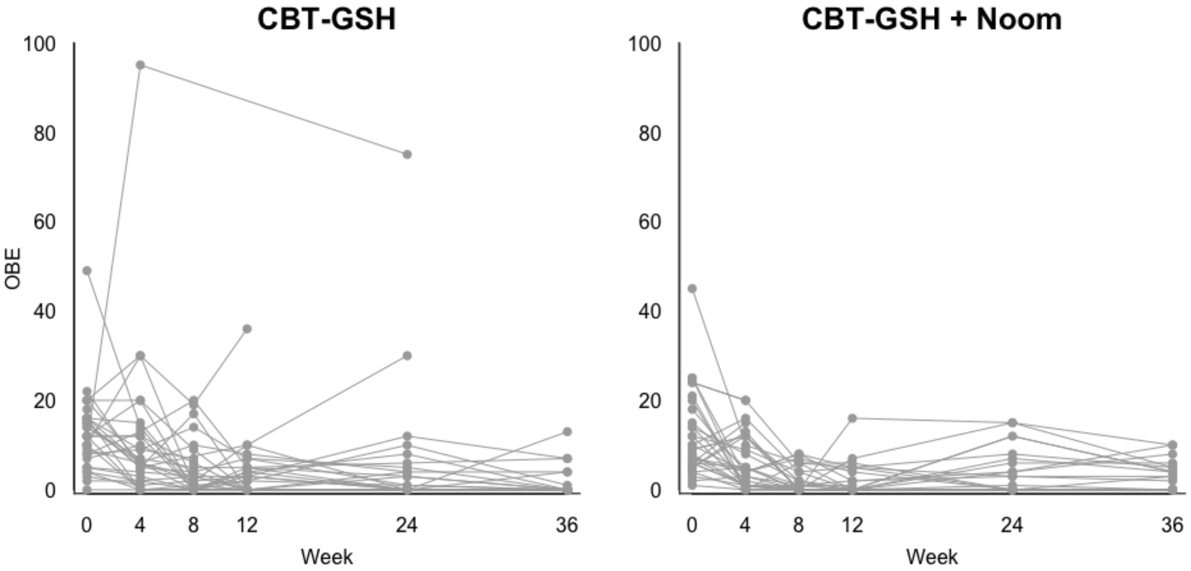

Figure 1: *Plots display OBEs over time for each individual in the CBT-GHS groups (left panel) and CBT-GHS + Noom group (right panel).*

Source of Prior Information	
Experimental Design	Outcome variable is nonnegative integers
Literature	Effect size is close to zero
Exploratory Data Analysis	There is variation in OBEs at the individual level There is variation in OBEs over time Treatment effects may vary over time

Table 1: *Sources of prior information.*

- The primary sources of variation in OBE appear to be *between people* and *over time.*

Figure 2 displays the distribution of OBEs in each condition in each week. We notice three characteristics of the data from these histograms.

1. The distributions appear to condense around zero for both conditions over time
2. The distributions in the CBT-GSH condition appear to have longer tails than those in the CBB-GSH+Noom condition
3. OBEs are count data; they must be nonnegative integers.

These three characteristics suggest that the appropriate model for OBEs is the Poisson distribution, because it is restricted to nonnegative integers and can concentrate its density around low numbers with a long tail.

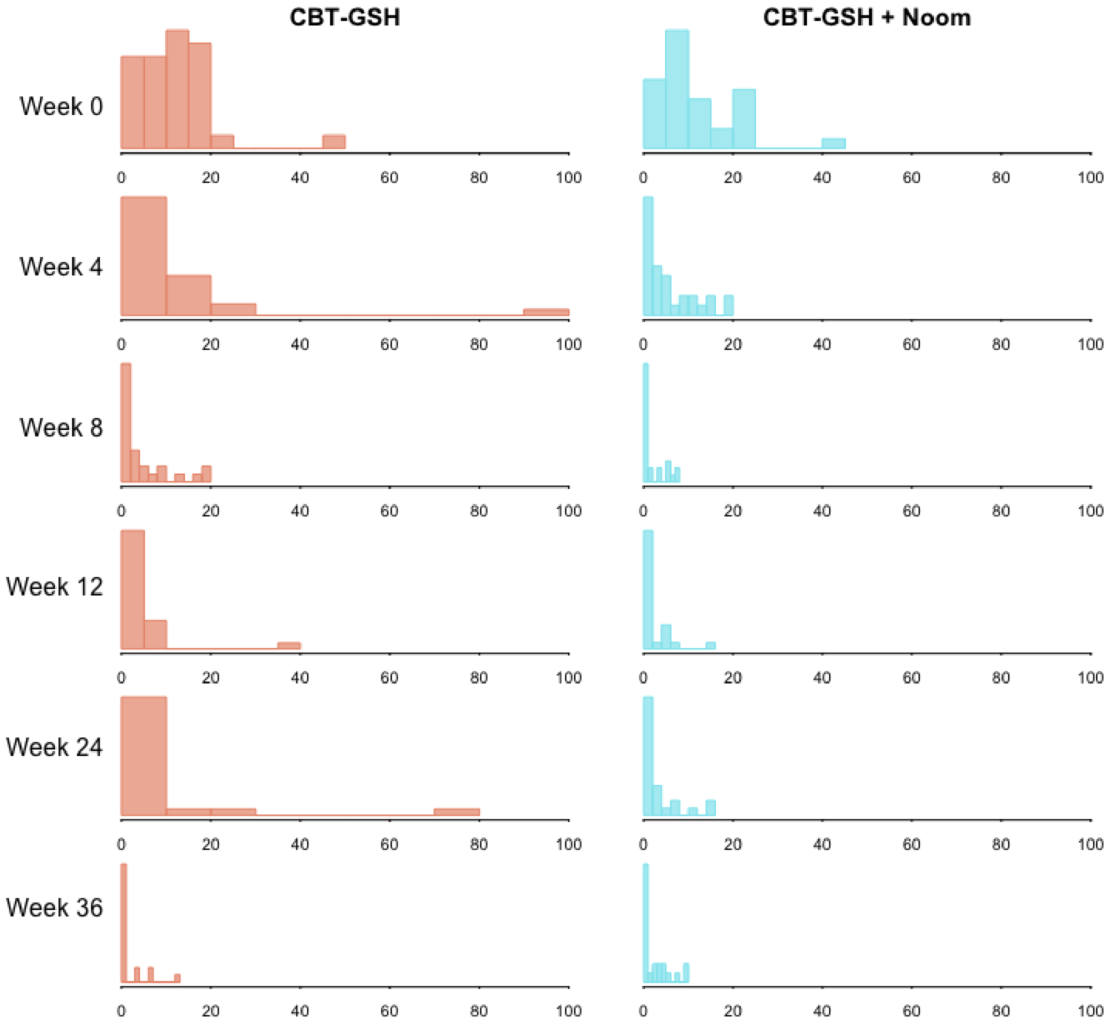

Figure 2: *Histograms display the distribution of OBEs in each condition in each week.*

4.3 Model development

We analyze RCTs by modeling the outcome of interest (in this case OBE) as a function of the treatment and all available pre-treatment covariates. The coefficients associated with the treatment reveal the average treatment effects. Inclusion of all available pre-treatment covariates accounts for variation in the outcome variable, decreasing uncertainly around treatment effects and providing the model with more predictive power. We conduct *intent-to-treat* analysis, meaning that our inferences will be based on initial treatment assignment, and will not account for mid-experiment dropouts.

The outcome variable is restricted to be nonnegative integers, so we fit a Poisson regression model, with hierarchies on individuals, time periods, and treatment effects. For each individual in each time period, the number of OBEs follows a Poisson distribution, with a mean dependent on the characteristics of the individual and the time period.

$$OBE_{i,t} \sim Poisson(\lambda_{i,t}) \tag{1}$$

$$\lambda_{i,t} = exp(\alpha_i + \beta_t + \gamma_t T_i + X_i \theta) \tag{2}$$

$$T_i = \begin{cases} 0, & \text{if } CBT - GSH \\ 1, & \text{if } CBT - GSH + Noom \end{cases} \tag{3}$$

α is an individual-specific intercept, β is a time-specific intercept, γ is a time-specific treatment effect, T is a treatment indicator, X is a matrix of individual level covariates (age, sex, race, etc), and θ is a vector of effects. Subscripts $i = 1, ..., 66$ indicate individuals and subscripts $t = 0, 4, 8, 12, 24, 36$ indicate time periods.

We believe that individual-level intercepts are simultaneously unique to the individual and common to the population; that is, each individual has their own baseline predilection to engage in eating disorder behavior, but their baseline predilections are not drastically different from each other. We operationalize this concept by modeling all individual-level intercepts as coming from a common distribution, with *hyperparameters* μ_α and τ_α.

$$\alpha_i \sim Normal(\mu_\alpha, \tau_\alpha) \ \forall \ i \in 1, ..., 66 \tag{4}$$

Similarly, we believe that time-specific treatment effects may be unique to each period but similar over time. We operationalize this concept by modeling all time-specific treatment effects γ as coming from a common distribution, with *hyperparameters* μ_γ and τ_γ.

$$\gamma_t \sim Normal(\mu_\gamma, \tau_\gamma) \ \forall \ t \in 0, 4, 8, 12, 24, 36 \tag{5}$$

μ_γ is the *grand mean*, the overall treatment effect; τ_γ is the variation in treatment effects over time; and each individual γ_t is a time-period specific treatment effect. This approach has a natural smoothing effect: any extreme estimates of γ_t will be partially-pooled back toward the grand mean μ_γ.

We assign the following prior and hyperprior distributions:

$$\mu_\alpha \sim Normal(5, 5) \tag{6}$$

$$\tau_\alpha \sim Cauchy^+(0, 30) \tag{7}$$

$$\mu_\gamma \sim Normal(0, 5) \tag{8}$$

$$\tau_\gamma \sim Cauchy^+(0, 30) \tag{9}$$

$$\theta \sim Normal(0, 1) \tag{10}$$

The normal distributions around the individual and treatment effects allow us to guide the model to the appropriate range of parameter values, but with wide enough variance (5 in each case) to let the model find its own way in that range. Half cauchy priors on the variance parameters are weakly informative, with much of their mass around zero but gentle slopes in their tails, which have been shown to be effective prior distributions for variance parameters (Gelman, 2006).

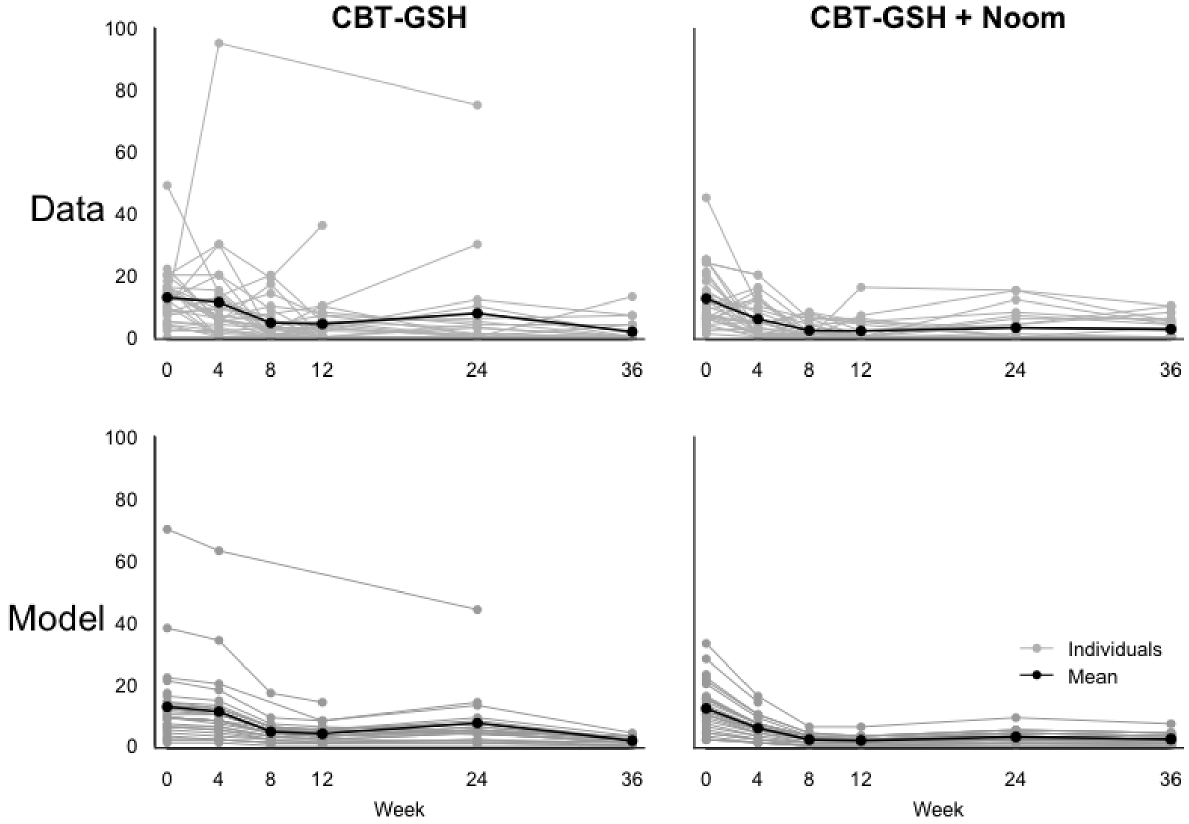

Figure 3: *Upper plots show OBEs in each period for each individual in both treatment group, with black lines representing means. Lower plots show modeled OBEs for each individual in each treatment group, with black lines representing modeled means. The model appears to be able to recover OBEs over time fairly well for both treatment conditions.*

4.4 Model estimation

We estimate this model with *Hamilton Monte-Carlo* in Stan. Model code is appended to this document. We find that the model converges with four chains of 2000 iterations each (see table M).

4.5 Posterior predictive checking

Before using our model to make inferences about time-specific treatment effects, we check its fit by comparing model-simulated OBE to data OBE. If model simulations do not track the data well, we may want to revisit our model's assumptions before trusting its inferences. If the model's simulations recover patterns in the data, we are more inclined to trust its inferences.

Figure 3 displays OBEs in each period for each individual in each treatment group, from raw data (upper plots) and model simulations (lowers plots). Black lines display means for each. This suggests that the model is broadly able to pick

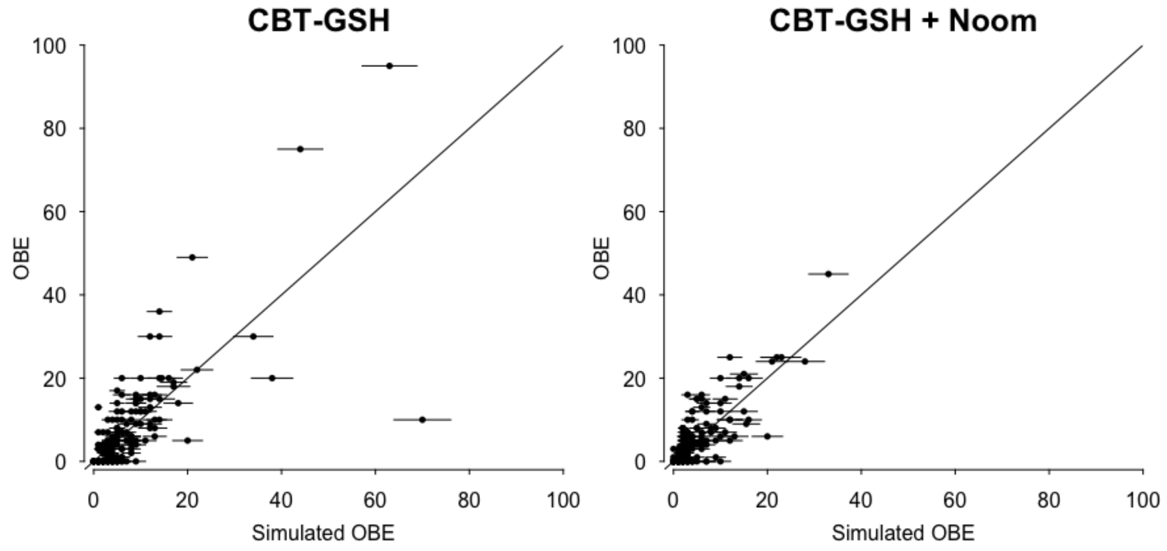

Figure 4: *Plots display siulated OBEs to raw data for both treatment conditions. Model simulations appear to match the raw data well, particularly for the Noom condition, which had fewer outliers.*

up on the key variables that determine OBE over time for the duration of this experiment.

Another way to check the fit of the model is by comparing simulated data directly against the raw data. Figure 4 shows this for both treatment conditions. Simulated data for the Noom condition appears to better track the raw data than simulated data for the no Noom condition. This is unsurprising, since the no Noom condition tended to have more outliers, which we would not expect (or want) our model to pick up perfectly from such a small sample.

We conclude our model evaluations by mapping modeled density curves for each condition in each time period over the histograms in figure 2 Figure 5 shows that our model is able to broadly pick up on the patterns in the data over time and between treatment conditions.

4.6 Results

Model results are displayed in table 2. Results suggest that using the Noom Monitor smartphone application during CBT-GSH may slightly decrease OBEs. There some evidence that the treatment effect varies over time, with the Noom effect being slightly more pronounced during the later stages of therapy.

Figure 6 displays modeled OBE for both treatment groups (upper plot) and smoothed treatment effects (lower plot). In each measurement period, simulated OBE are higher for the No Noom condition than for the Noom condition, with

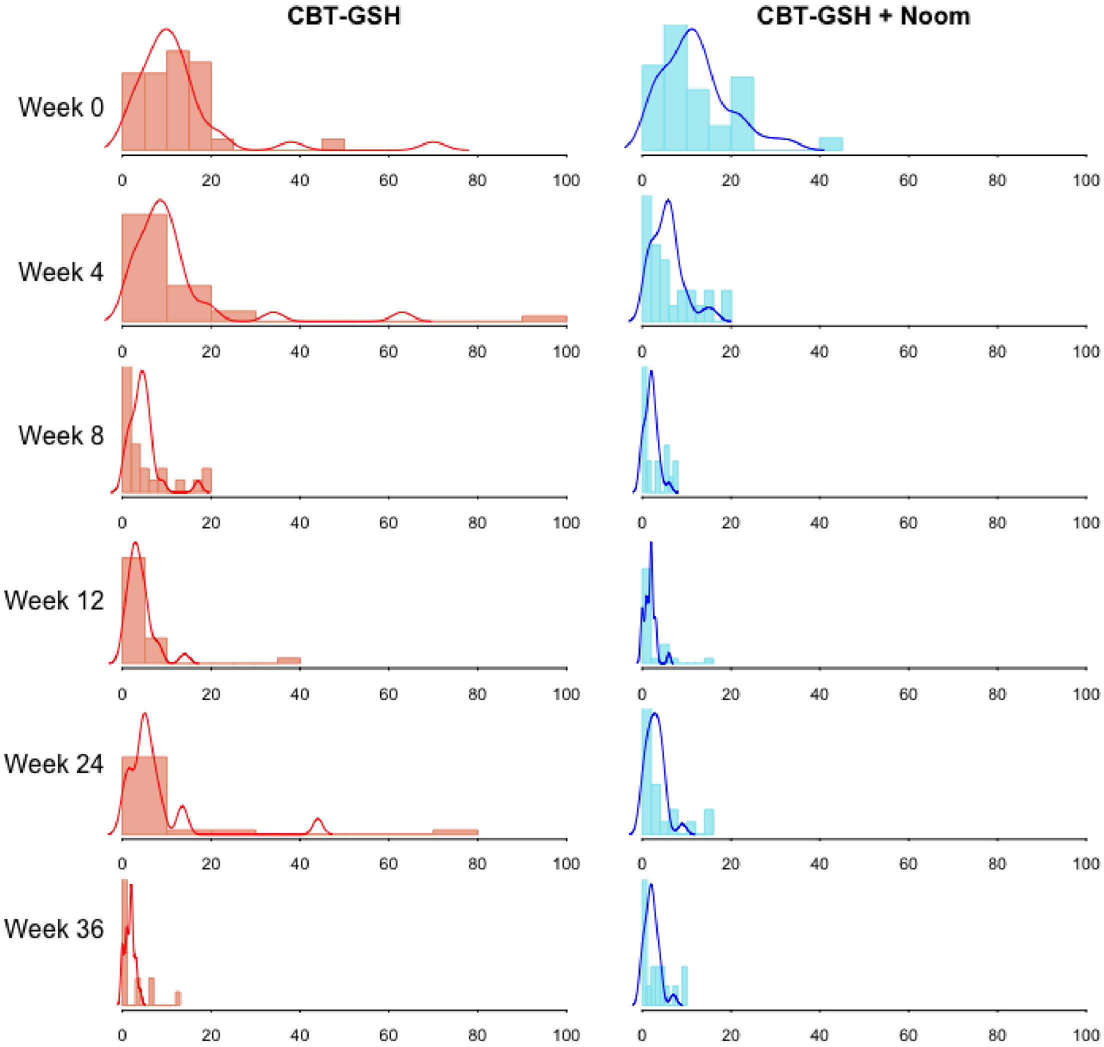

Figure 5: *Histograms display the distribution of OBEs in each condition in each week.*

some of the difference likely attributable to use of the Noom Monitor smartphone app.

5 Conclusion

6 Appendix

	mean	25%	50%	75%
γ_0	0.18	-0.45	0.15	0.78
γ_4	-0.43	-1.05	-0.46	0.16
γ_8	-0.70	-1.33	-0.71	-0.10
γ_{12}	-0.65	-1.28	-0.68	-0.04
γ_{24}	-0.72	-1.34	-0.75	-0.11
γ_{36}	0.21	-0.42	0.19	0.82
μ_γ	-0.34	-0.98	-0.36	0.26
τ_γ	0.64	0.43	0.56	0.77

Table 2: *Table displays model results for Noom effects in all six time periods and grand mean and variance parameters.*

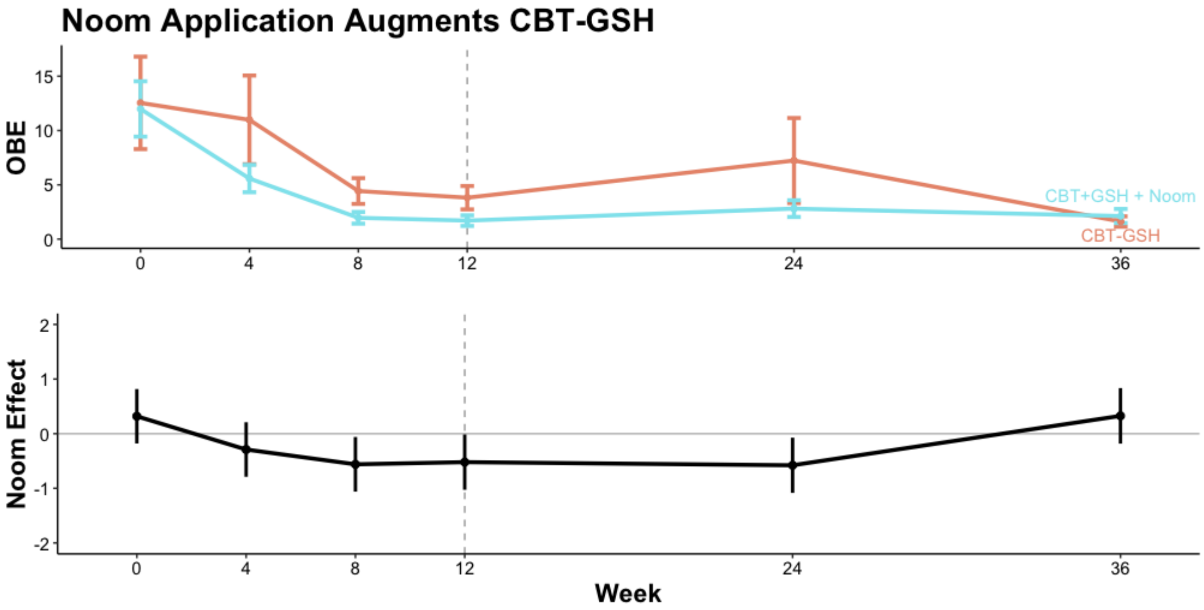

Figure 6: *Upper plot displays modeled OBEs in each time period for the Noom (blue) and no Noom (orange) conditions with 95% intervals. Lower plot displays modeled treatment effects in each period,with 50% intervals.*