Rethinking the Statistical Approach to Randomized Controlled Trials: Beyond the Mixed Models for Repeated Measures

Vikram Suresh, Jeffrey A. Mills, Jeffrey R. Strawn

March 2023

1 Abstract

2 Introduction

Randomized controlled trails (RCT) are the gold standard for evidence based evaluation of treatment effects. While the successes of clinical trials over the last several decades justify their high costs and resource intensiveness, it is imperative that the best available methods are employed to analyze the data generated by a clinical trial. The MMRM has been the proven work horse for analysis of the evidence from clinical trials across medical, surgical, and psychotherapeutic interventions. The measured clinical variables generally correlate during a trial. Put simply, these observations are not independent. While the MMRM represents a family of models, the standard approach usually includes a design matrix with fixed effects and a covariance pattern to describe the correlations among the measured variables.

However, advances in statistical theory and computational power over the last few decades have provided researchers with new tools and technology. These two improvements have shown promise of potentially being superior to the standard MMRM approach. First, the time dependent patient observations are better characterized explicitly in the model specification rather than in the model error covariance [1] [2]. Second, the increasing popularity of Bayesian Inference has allowed for a different approach to modeling clinical trial data. Bayesian Hierarchical Models (BHM) have several characteristic advantages [3], they allow

for and estimate degree of heterogeneity in the sample. The Bayesian Inference is valid for small samples using dynamic models with simple error structures [4]. The Bayesian Inference can be powerful in situations where standard models and other Machine Learning models are sub-optimal such as:

- When we only have a small amount of data.
- When the data is very noisy.
- When we need to quantify confidence.
- When we want to incorporate prior beliefs into the model.
- When model explainability is important.

With these considerations in mind, we examine the performance of different model specifications to assess treatment effects in RCTs. The dynamic correlations within the data can be modeled through either an explicit specification or an error variance structure. Appropriate models must be able to parse the total variation in the data into systematic variation across time, the signal, and an approximation error, the noise. Moreover, when appropriate models sufficiently capture this systematic variation, the approximation error is independent and identically distributed. When models fail to capture these systematic variations, they have failed to capture the signal in the data. Modeling systematic variations by way of the approximation error through complex error structures is a dynamic limitation of the model specification [1].

While preference among statisticians is for large samples so as to increase the precision and reliability of estimates, clinical investigators prefer to limit the patient risk and exposure to experimental conditions. For clinicians and those conducting clinical trials, having 'large' trials is often untenable. Larger trials require more time, are more difficult to implement, and potentially delay the arrival of new interventions to patients in the clinic.

Some of the most commonly used MMRM model covariance structures are not diagonally dominant and so are inconsistent with time series theory [5]. The MMRM is a static model built on large sample assumptions and estimating in-sample heterogeneity is difficult. Additionally, the standard methods have sometimes struggled to capture existing treatment effects [4]. We take an agnostic view and perform the model comparison to evaluate the performance of these procedures. A simulation study is performed to evaluate the predictive accuracy of the models in-sample and out-of-sample, and their ability to identify

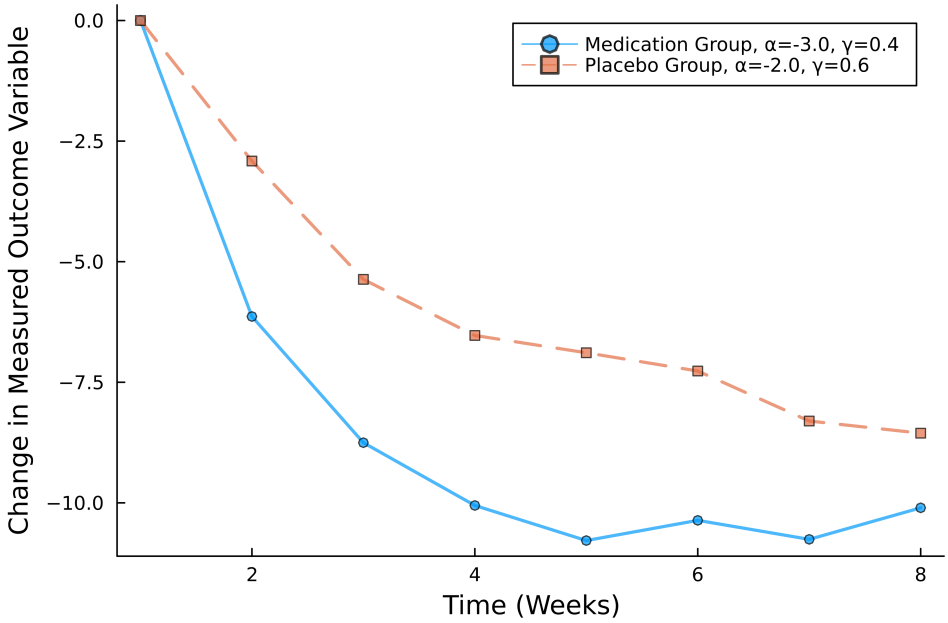

Figure 1: Data generated from an Autoregressive order 1 (AR) process with different long run patient outcomes. The smaller the rate parameter γ, the quicker the decay.

significant treatment effects in the data through estimation and hypothesis testing. These performance metrics ensure the models accurately predict patient outcomes, while estimating treatment effects. The superior models from the simulation study are then applied to pseudo data generated using the estimates from two randomized, placebo-controlled trials of antidepressant medications in children and adolescents, Child/Adolescent Anxiety Multimodal Study (CAMS) [6] and Neurofunctional Predictors of Escitalopram Treatment Response in Adolescents With Anxiety (FiESTAA) [7]. These studies were selected given the importance of anxiety disorders in children and adolescents [7], as well as their prevalence, substantial morbidity and their tendency to increase the likelihood of developing depressive disorders.

The accurate identification of treatment effects is even more important than in-sample fits. Models that capture the treatment effect under small sample conditions decrease patient risk by not requiring a large number of patients subject to experimental conditions, and also minimizing trial costs. The definition of treatment effect can include the *rate* (γ) at which patients receiving treatment

Table 1: Data Generating Processes and their Covariance Patterns

Data Generating Process	Fixed Effects	Covariance Pattern
Time Effects	Time Indicators, Treatment and Treatment-Time	Homoskedastic IID
Time Effects - CS	Time Indicators, Treatment and Treatment-Time	Compound Symmetry
Time Effects - AR	Time Indicators, Treatment and Treatment-Time	Autoregressive Order 1
Log Trend	Logarithmic Trend, Treatment-Trend	Homoskedastic IID
Log Trend - CS	Logarithmic Trend, Treatment-Trend	Compound Symmetry
Log Trend - AR	Logarithmic Trend, Treatment-Trend	Autoregressive Order 1
AR	Treatment, AR1 Coefficient, Treatment-AR1	Homoskedastic IID

improve, i.e., treated patients improve much quicker to an eventual stable end point. They could improve to different stable end point condition (α) compared to those not receiving treatment (i.e., being free of treatment). This is best demonstrated in Figure 1, where it can be noted the patients receiving medication reach a stable end point outcome that is different than patients receiving placebo and the *rate* at which they arrive at these outcomes is also different.

3 Methods

Pseudo data are generated from several theoretical dynamic panel data processes so as to best approximate repeated measures data from trials such as CAMS and FiESTAA that assess and measure symptoms across time. All of the data generated processes fit into the following framework,

$$
\begin{aligned}
y &= X\beta + \epsilon \\
\epsilon &\sim \text{MvNormal}(\mathbf{0}, \Sigma) \\
\mathbf{0} &= \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix}'_{1, N \times T}
\end{aligned}
$$

Where X is the design matrix corresponding to the fixed effects in a model, Σ is the appropriate covariance pattern, autoregressive order 1 (AR), compound

symmetry (CS) or homoskedastic, independent and identically distributed (IID). When the covariance pattern is IID, Σ takes the simple form, $\Sigma = \sigma^2 I$, I is $(N \times T) \times (N \times T)$ identity matrix, sample of N patients, and T measurements made per patient across time. The design matrices and covariance patterns for the data generating processes are described in Table 1.

The correlations among the repeated measurements for each patient i can be generalized to a variety of structures, including heteroscedastic variances.

$$R_i = \begin{bmatrix} \sigma_1^2 & \theta_{1,2} & \theta_{1,3} & \cdots & \theta_{1,T} \\ \theta_{2,1} & \sigma_2^2 & \theta_{2,3} & \cdots & \theta_{2,T} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \theta_{T,1} & \cdots & \theta_{T-2,T} & \theta_{T-1,T} & \sigma_T^2 \end{bmatrix}_{T \times T}$$

The covariance pattern is therefore a block diagonal matrix with entries for each patient across the diagonal. While such a pattern is theoretically possible to be specified, it is infesible to estimate the large number of unknown parameters. For the same reason, the frequently used Toeplitz structure was not included as the increasing measurements across time implies $(T-1)$ off diagonal covariance parameters to be estimated. As recommended by Brown and Prescott 2015, we choose the two most popular structures, the autoregressive order 1 (AR) and compound symmetry (CS) [8].

For an AR(1) covariance pattern is a block diagonal matrix of the form,

$$\Sigma_{AR} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{T-1} & \cdots & \rho^2 & \rho & 1 \end{bmatrix}_{T \times T} \otimes I_{N \times N}$$

Similarly, a compound symmetry (CS) covariance pattern is also a block diagonal matrix,

$$\Sigma_{CS} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho & \cdots & \rho & \rho & 1 \end{bmatrix}_{T \times T} \otimes I_{N \times N}$$

It can be proven theoretically that the variance-covariance matrices need to

be diagonally dominant, which is violated in the case of a compound symmetry (CS) structure [5].

3.1 Data Generating Processes

For each patient during each study visit (i.e., consistent) some outcome variable $y_{i,t}$ is measured. The trial consists of $i = [1, 2, \cdots, N]$ patients and measurements made for each patient at each visit $t = [1, 2, \cdots, T]$. Patient treatment indicator x_i having been randomized into a treated or untreated group.

Time Effects (TE)

$$y_{i,t} = \alpha + \beta x_i + \sum_{j=2}^{T} \delta_j D_j + \sum_{j=2}^{T} \gamma_j x_i D_j + \epsilon_{i,t}$$

$$\epsilon_{i,t} \sim \text{MvNormal}(\mathbf{0}, \Sigma)$$

$$\begin{cases} D_j = 1 & \text{if j = t} \\ D_j = 0 & \text{otherwise} \end{cases}$$

Time Effects - Compound Symmetry (TE-CS)

$$y_{i,t} = \alpha + \beta x_i + \sum_{j=2}^{T} \delta_j D_j + \sum_{j=2}^{T} \gamma_j x_i D_j + \epsilon_{i,t}$$

$$\epsilon_{i,t} \sim \text{MvNormal}(\mathbf{0}, \Sigma_{CS})$$

Time Effects - Autoregressive Order 1 (TE-AR)

$$y_{i,t} = \alpha + \beta x_i + \sum_{j=2}^{T} \delta_j D_j + \sum_{j=2}^{T} \gamma_j x_i D_j + \epsilon_{i,t}$$

$$\epsilon_{i,t} \sim \text{MvNormal}(\mathbf{0}, \Sigma_{AR})$$

Logarithmic Trend (LogT)

$$\Delta y_{i,t} = \delta log(t) + \gamma x_i log(t) + \epsilon_{i,t}$$

$$\epsilon_{i,t} \sim \text{MvNormal}(\mathbf{0}, \Sigma)$$

Where, $\Delta y_{i,t}$ represents change from baseline measurement, i.e.,

$$\Delta y_{i,t} = y_{i,t} - y_{i,1}$$

Logarithmic Trend - Compound Symmetry (LT-CS)

$$\Delta y_{i,t} = \delta log(t) + \gamma x_i log(t) + \epsilon_{i,t}$$
$$\epsilon_{i,t} \sim \text{MvNormal}(\mathbf{0}, \Sigma_{CS})$$

Logarithmic Trend - Autoregressive Order 1 (LT-AR)

$$\Delta y_{i,t} = \delta log(t) + \gamma x_i log(t) + \epsilon_{i,t}$$
$$\epsilon_{i,t} \sim \text{MvNormal}(\mathbf{0}, \Sigma_{AR})$$

Autoregressive Order 1 (AR)

$$y_{i,t} = \alpha + \beta x_i + \gamma y_{i,t-1} + \delta x_i y_{i,t-1} + \epsilon_{i,t}$$
$$\epsilon_{i,t} \sim \text{MvNormal}(\mathbf{0}, \Sigma)$$

The simulations include different sample sizes, i.e., the number of patients N=[25,50,100,150] and differing number of measurements per patient T=[4,6,8]. The patients were randomized into medication or placebo group through a *Bernoulli Process* with probability $p = 0.5$. We are able to identify the best performing models in relatively small samples to significantly larger ones. The data are generated for each (N, T) pair from each of these processes. We choose the parameters for the simulations such that the end point treatment effect is approximately 20%, model variance $\sigma^2 = 1.0$ and correlations in the error structures $\rho = 0.65$. For each pseudo data sample that is generated, we fit the seven different specifications to the generated sample. Since the true process is often unknown, fitting various models to data from a known process allows identification of the model or set of models that best fit data from any process that resembles repeated measures from a clinical trial. We compare the fitted models by their prediction errors within training sample and testing sample. All the models were fit using an uninformative prior in a Bayesian Hierarchical setup using Julia v1.8 [9].

4 Results

The results from the Monte Carlo simulations are described in this section. Across multiple sample sizes, 7 different data generating processes, 7 fitted models to each process, the autoregressive arder 1 (AR) model performs best in and

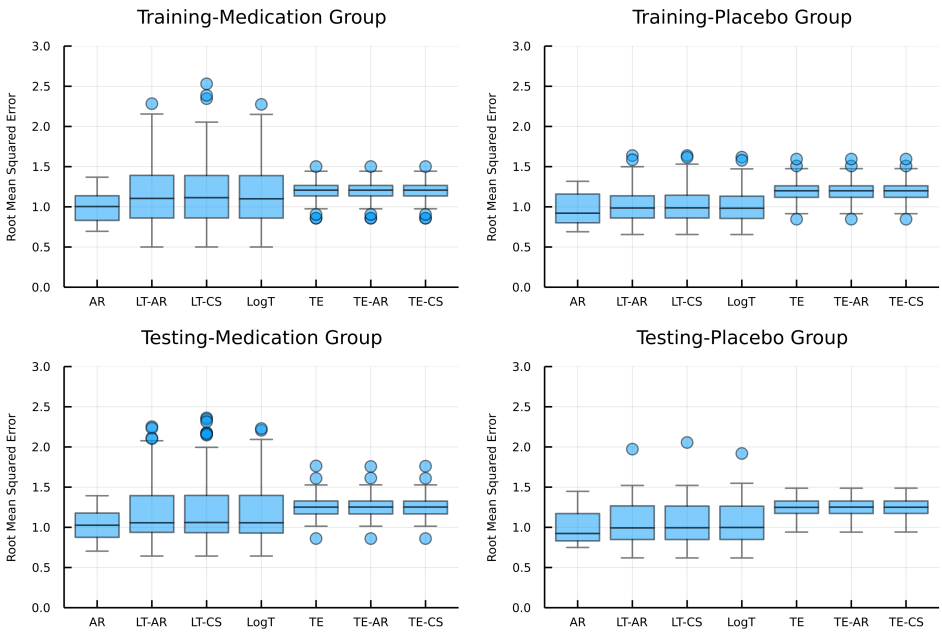

Figure 2: Root mean squared error boxplot aggregated across all N, T training and testing samples. Showing the variation in prediction error by each model categorized appropriately.

out-of-sample as shown in Figures 2 and 3. We compare the RMSE for treatment and placebo group within training samples and then using the estimates against a separately generated testing sample of size $N' = 50$. The boxplot for the root mean squared error (RMSE) values for each fitted specification across all the generated pseudo samples by treatment category are shown in Figure 2. The total number of times across all categories and samples each specification outperforms other specifications in predictive RMSE is shown in Figure 3. Even though Vallejo et. al. find misspecifying the covariance structure can affect estimates [10], there is no significant gain in predictive accuracy. We find that parsimonious models with few variance parameters tend to perform best by RMSE across various sample sizes. These results are consistent under various conditions, high correlation coefficient in the covariance pattern ($\rho = 0.85$), no treatment effect, and low model variance ($\sigma^2 = 0.01$).

While prediction errors offer one comparative metric among models, identifying treatment effects is a necessary imperative. The interpretation of the treatment effect depends on the chosen model and how it influences the out-

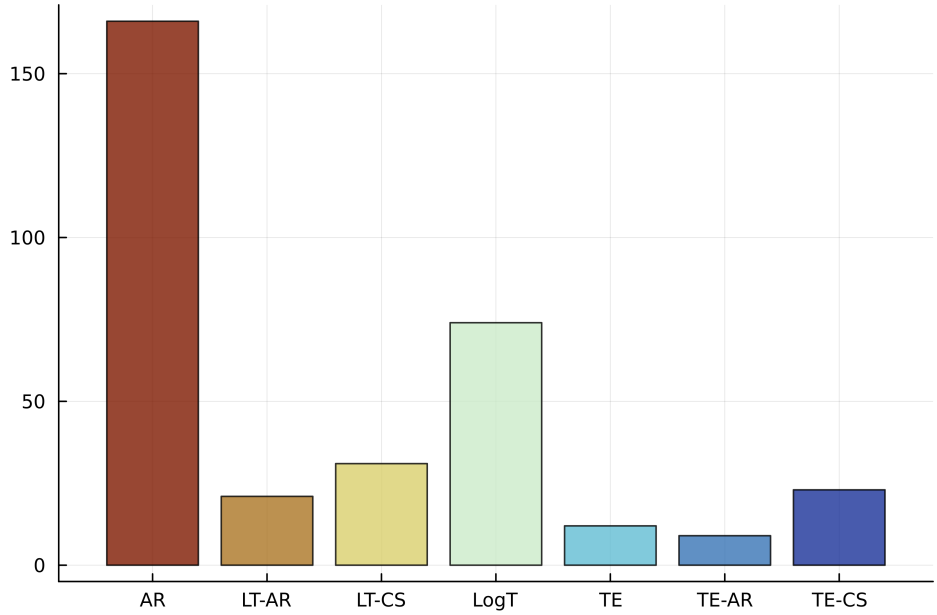

Figure 3: Total number of times each model predicts with least root mean squared error.

come variable. The logarithmic trend model perhaps offers a more concise and easier to interpret results than the other models. It can be noted that when a treatment effect is significant and identified by the logarithmic model, there is a shift in the slope of the trajectory of the outcome variable. The AR model as defined above includes the possibility of shifts in the intercept and the slope for the medication treated patients. The treatment effect estimate from this model should therefore include both estimates as each represents a different shift in trajectory of the outcome variable. To determine the ideal AR specification, we included an AR specification with only the treatment-slope, and one with only the treatment-intercept in the subsequent sections.

Autoregressive Order 1 (AR) - Slope

$$y_{i,t} = \alpha + \gamma y_{i,t-1} + \delta x_i y_{i,t-1} + \epsilon_{i,t}$$
$$\epsilon_{i,t} \sim \text{MvNormal}(\mathbf{0}, \Sigma)$$

Autoregressive Order 1 (AR) - Intercept

$$y_{i,t} = \alpha + \beta x_i + \gamma y_{i,t-1} + \epsilon_{i,t}$$
$$\epsilon_{i,t} \sim \text{MvNormal}(\mathbf{0}, \Sigma)$$

Since the the logarithmic trend and AR models fit the data best in-and-out of sample 3, we evaluate their ability to estimate significant treatment effects in pseudo data simulated using estimates from two Federally funded double-blind placebo-controlled trials in patients with anxiety disorders Walkup et al. [6] and Strawn et al. [7]. The primary outcomes have been previously described, but briefly, the CAMS trial involved treatment response of children and adolescents to combined treatment with a selective serotonin reuptake inhibitor (SSRI) and cognitive behavior therapy (CBT). The FiESTAA trial, involved determining escitalopram treatment response in adolescent with generalized anxiety disorder relative to placebo. The primary outcomes in these studies were clinical global impression severity scale (CGI-S) and pediatric anxiety rating scale (PARS). The AR specification including both treatment slope and treatment intercept struggles to identify a significant treatment effect perhaps due to confounding by including both. The Logarithmic Trend, AR-Slope and AR-Intercept model identify the existing treatment effect in the data.

Additionally, using the estimates from the original studies using a TE-AR specification, we generate data to get different levels of endpoint treatment effect, ranging from 0% to 20% in increments of 5%. We perform Monte Carlo simulations to find the model that best identifies existing treatment effect using this simulated data. The power curves from these simulations for different sample sizes are shown in Figures 4 and 5. Figure 4 shows the logarithmic trend specification estimates significant treatment effects better than the other specifications for FiESTAA like pseudo data of different sample sizes as the true treatment effect signal steadily increases. Figure 5 shows similar power curves that demonstrate the comparable if not superior performance of the logarithmic trend specification using the CAMS like pseudo data. The parsimonious AR specification either using a treatment-slope or treatment-intercept performs better in estimating significant treatment effects than AR specification that includes both.

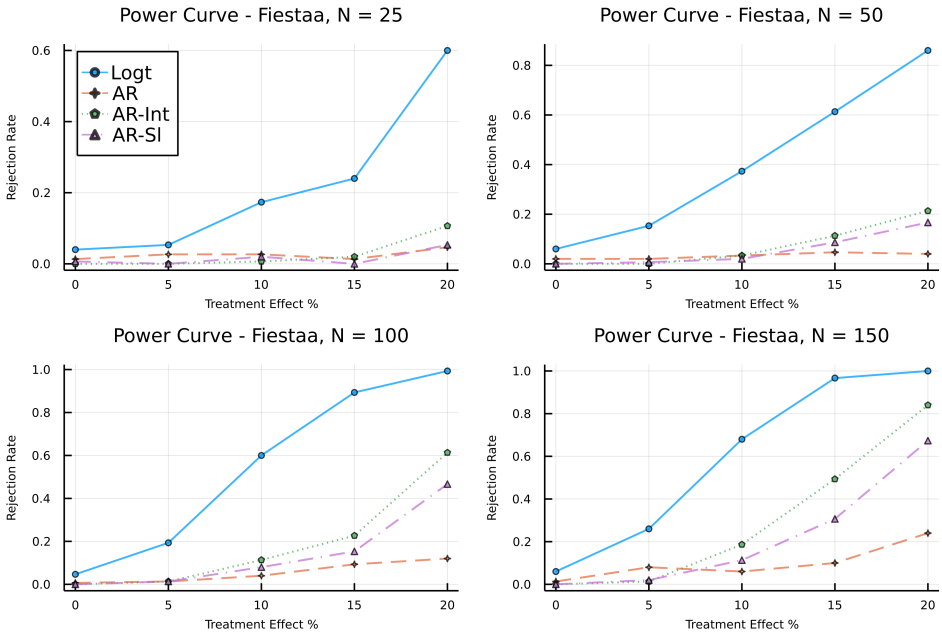

Figure 4: Power curves for FiESTAA like data showing logarithmic trend model estimating true treatment effects in the data better than other specifications.

5 Discussion

To compare models without unnecessary complications, we make a few simplifying assumptions about the data generated; all the patients have the same number of measurements across time (balanced panel), there are no measurements missing, i.e., patients do not drop out of the study at random. While this is not a comprehensive comparison across all the possible Mixed Effects Models, further studies can explore additional models. The time effects model, while relatively flexible in fitting a variety of time trajectories without assuming a definite functional form, requires a parameter for each time point. This leads to ambiguity in how best to test for significant treatment differences over time and potentially less statistical precision. However, this additional flexibility is useful for identifying specific time paths of improvement, when separation between two treatment effects occurs.

BHM multilevel models allows for unobserved heterogeneity in the treatment response by modeling each individual's treatment trajectory [4]. This unobserved heterogeneity, while not considered in this simulation study, can be

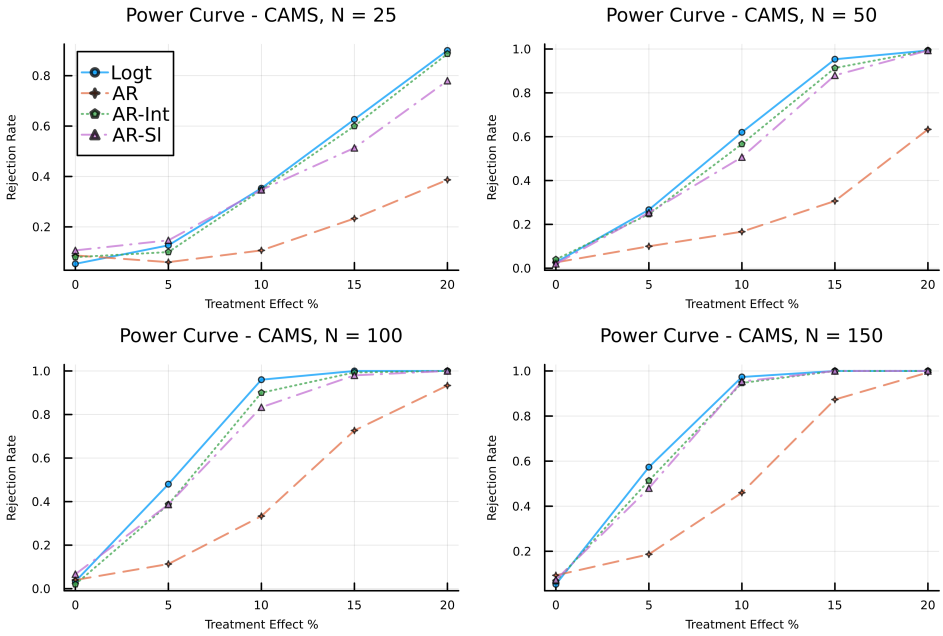

Figure 5: Power curves for CAMS like data showing logarithmic trend model estimating true treatment effects in the data as good as the AR specifications.

implemented with relative ease using the AR and logarithmic trend models.

The results of the simulation studies herein reveal the superior performance of the AR and logarithmic trend models in small samples. Further, these models have a more parsimonious specification of time dependency, and therefore a simpler error structure.This leads to hypothesis tests for treatment effects that are simpler and statistically more powerful.

References

[1] Mizon GE. A simple message for autocorrelation correctors: Don't. Journal of Econometrics. 1995;69(1):267-88. Annals of Econometrics: Bayesian and Classical Econometric Modelling of Time Series. Available from: `https://www.sciencedirect.com/science/article/pii/030440769401671L`.

[2] Enders W. Applied Econometric Time Series. Wiley; 2018. Available from: `https://books.google.com/books?id=s9QZyAEACAAJ`.

[3] Lewis RJ, Angus DC. Time for Clinicians to Embrace Their Inner Bayesian?: Reanalysis of Results of a Clinical Trial of Extracorporeal Membrane Oxygenation. JAMA. 2018 12;320(21):2208-10. Available from: https://doi.org/10.1001/jama.2018.16916.

[4] Suresh V, Mills JA, Croarkin PE, Strawn JR. What next? A Bayesian hierarchical modeling re-examination of treatments for adolescents with selective serotonin reuptake inhibitor-resistant depression. Depression and Anxiety. 2020 Jun;37(9):926-34. Available from: https://doi.org/10.1002/da.23064.

[5] Hamilton JD. Time Series Analysis. Princeton University Press; 1994. Available from: https://doi.org/10.1515/9780691218632.

[6] Walkup JT, Albano AM, Piacentini J, Birmaher B, Compton SN, Sherrill JT, et al. Cognitive Behavioral Therapy, Sertraline, or a Combination in Childhood Anxiety. New England Journal of Medicine. 2008 Dec;359(26):2753-66. Available from: https://doi.org/10.1056/nejmoa0804633.

[7] Strawn JR, Mills JA, Schroeder H, Mossman SA, Varney ST, Ramsey LB, et al. Escitalopram in Adolescents With Generalized Anxiety Disorder. The Journal of Clinical Psychiatry. 2020 Aug;81(5). Available from: https://doi.org/10.4088/jcp.20m13396.

[8] Brown H, Prescott R. Applied Mixed Models in Medicine. Statistics in Practice. Wiley; 2015. Available from: https://books.google.com/books?id=3_veBQAAQBAJ.

[9] Bezanson J, Edelman A, Karpinski S, Shah VB. Julia: A fresh approach to numerical computing. SIAM review. 2017;59(1):65-98. Available from: https://doi.org/10.1137/141000671.

[10] Vallejo G, Ato M, Valdés T. Consequences of Misspecifying the Error Covariance Structure in Linear Mixed Models for Longitudinal Data. Methodology. 2008 Jan;4(1):10-21. Available from: https://doi.org/10.1027/1614-2241.4.1.10.