# COMPARISON OF BAYESIAN NETWORKS, G-ESTIMATION AND LINEAR MODELS TO ESTIMATE CAUSAL TREATMENT EFFECTS IN AGGREGATED N-OF-1 TRIALS

A PREPRINT

**Thomas Gärtner**
Hasso Plattner Institute
Prof.-Dr.Helmert-Straße 2-3
Potsdam, Germany
thomas.gaertner@hpi.de

**Juliana Schneider**
Hasso Plattner Institute
Prof.-Dr.Helmert-Straße 2-3
Potsdam, Germany

**Bert Arnrich**
Hasso Plattner Institute
Prof.-Dr.Helmert-Straße 2-3
Potsdam, Germany

**Stefan Konigorski**
Hasso Plattner Institute
Prof.-Dr.Helmert-Straße 2-3
Potsdam, Germany
stefan.konigorski@hpi.de

July 21, 2022

**ABSTRACT**

The aggregation of a series of N-of-1 trials presents an innovative and efficient study design, as an alternative to traditional randomized clinical trials. Challenges for the statistical analysis arise when there are carry-over effects or confounding of the treatment effect of interest.

In this study, we evaluate and compare methods for the analysis of aggregated N-of-1 trials in different scenarios with carry-over and confounding effects. For this, we simulate data of a series of N-of-1 trials for chronic nonspecific low back pain based on assumed causal relationships parameterized by directed acyclic graphs. In addition to existing statistical methods such as regression models, Bayesian Networks and G-estimation, we introduce a linear model adjusted for time-varying treatment carry-over effects.

The results show that all evaluated existing models have a good performance when there is no carry-over and confounding. When there are carry-over effects, our proposed method yields unbiased estimates while all methods show some bias in the estimation. When there is known confounding, all approaches that specify the confounding correctly yield unbiased estimates. Finally, the efficiency of all methods decreases slightly when there are missing values, and the bias in the estimates can also increase.

This study presents a systematic evaluation of existing and novel approaches for the statistical analysis of a series of N-of-1 trials. We derive practical recommendations which methods may be best in which scenarios.

***Keywords*** Causal Inferences · Time-varying Treatment · Data Simulation

## 1 Introduction

Within the last decade, personalized medicine has been on the rise. Treating patients on an individual level has been improved by the numerous possibilities to measure health outcomes with smart devices and application of novel data science approaches. In order to evaluate the effectiveness of health interventions on an individual level, N-of-1 trials have been established as the gold standard Nikles and Mitchell [2015], Davidson et al. [2021]. N-of-1 trials are

multi-crossover controlled trials, where each patient is his own control group. In addition to individual-level analyses for personalized treatment, series of N-of-1 trials can be analyzed jointly Stunnenberg et al. [2018], or also combined with results from standard randomized controlled trials (RCTs) to obtain population-level estimates on the effectiveness of treatments with equal or superior efficiency compared to non-crossover RCTs Punja et al. [2016], Blackston et al. [2019]. For the aggregate statistical analysis of series of N-of-1 trials, popularly used methods include parameter free methods like the Wilcoxon signed-rank test Green et al. [2004], Sierra-Arango et al. [2019], two-sample mean tests Schmid CH, Duan N., the DEcIDE Methods Center N-of-1 Guidance Panel [2014], methods that allow for covariate adjustments like linear models Odineal et al. [2019], Vrinten et al. [2015], linear mixed models Herrett et al. [2021] and Bayesian approaches Chen and Chen [2014], Samuel et al. [2019]. Also, methods for time adjustments such as autoregressive models have been proposed for the analysis Zhou et al. [2017].

Some studies have evaluated and compared different methods for the analysis. For example, Stunnenberg et al. Stunnenberg et al. [2018] applied both frequentist linear mixed models as well as Bayesian models and compared the approaches in a study on the effect of mexiletine on muscle stiffness in patients with nondystrophic myotonia. Zucker et al. Zucker et al. [2010] compared repeated-measure models, Bayesian hierarchical models, and simpler single-period, single-pair, and averaged outcome crossover models in the analysis of a published series of N-of-1 trials on rheumatolic treatments. Their results showed that depending on the assumptions, different mixed models yielded the best fit and that Bayesian models were sensitive to the specification of the priors. Chen & Chen Chen and Chen [2014] compared t-tests and mixed models in a simulation study when no carry-over was present, and found t-tests to yield highest power under this assumption. Finally, Araujo et al. Araujo et al. [2016] extended the work of Chen & Chen and considered t-tests and linear mixed models under different model assumptions on the study design, with a focus on how the study design incorporated randomization.

In this study, we focus on two particular challenges for the analysis of aggregated N-of-1 trials: carry-over effects and confounding Blackston et al. [2019]. First, as a patient cannot receive two treatments at the same time, the treatment is time-varying. This may introduce carry-over effects - i.e., that the effect of one treatment is still active when the other treatment is applied - complicating the analysis of the trials and the interpretation of the results. As one solution, wash-out periods can be introduced in the study design, where the patient does not receive any of the treatments. However, this is not always possible, so statistical methods have to be investigated regarding their robustness against known or unknown carry-over effects. Second, for the aggregated analysis of N-of-1 trials, the role of confounding variables has to be considered to ensure unbiased estimates of the causal treatment effects. Both challenges need to be considered in the statistical analysis, but best-practise recommendations are not available Gamble et al. [2017].

## 2 Methods

In the following, we investigate different statistical methods for the analysis of aggregated N-of-1 trials on simulated data sets that contain different levels of carry-over effects and confounding. Additionally, we compared the methods on data sets with and without missing data due to participants' dropout. As traditional methods, we included a sample mean comparison and linear regression models in the analysis Schmid CH, Duan N., the DEcIDE Methods Center N-of-1 Guidance Panel [2014], Nikles and Mitchell [2015]. Further, we introduced a linear model that specifically models and adjusts the treatment effect for carry-over. Finally, we considered Bayesian Networks Pearl et al. [2016] and G-estimation Hernan and Robins [2019]. The results of our simulation study showed that simple methods can provide unbiased and efficient treatment estimates. Among the considered approaches, our proposed simple adjusted linear model yielded the best performance across all scenarios. More specifically, modeling carry-over effects in linear models yields unbiased estimates even if a strong carry-over is present, but requires complete data. Linear models and 2-sample t-tests are robust against missing values, but yield biased effect estimates when strong carry-over is present. Furthermore, the sample mean yields biased effect estimates when confounding is present. Bayesian Networks and G-estimation showed a good overall performance, but Bayesian Networks yielded wider confidence intervals of effect estimates especially for small sample sizes. These results provide some guidelines which methods should be used in practical applications, and we provide the R package cinof1 (available from github.com/HIAlab/cinof1) with an implementation of all investigated methods.

To evaluate and compare the different statistical methods, we performed a Monte Carlo simulation study. In the following, we describe the simulation study set-up including the data generation model, a specific application of the data generation model to generate synthetic data of a series of N-of-1 trials on chronic nonspecific low back pain, and the different evaluated statistical methods. The data generation model is available through the Python package sinot (github.com/HIAlab/sinot), and the statistical methods are implemented in the R package cinof1 (github.com/HIAlab/cinof1).

### 2.1 Data Generation

#### 2.1.1 Data Model

In the simulation model, we combined data generation based on stochastic processes, time-varying treatment effects, and effects onto the outcome variable embedded in a causal graph. For the notation in the following, let $Z$ denote any of the variables in our model including the outcome $O$, treatment $T$ or other variables $C$.

First, we embedded the outcome variable $O$ and treatment variable $T$ in a directed acyclic diagram (DAG) with further variables $C$, which may be constant or time-varying, and can be simulated from several common distributions like Bernoulli, Gaussian, Poisson, or uniform distribution. Time-invariant variables would not change over time and describe for instance demographics or baseline conditions like previous diagnosis. Time-varying variables may change on each observation and could be measurements like the number of steps per day.

We included linear effects from variables $Z_j$ on $Z_i$ at time point $t$:

$$z_i^t = \sum_j (w_{j,i} \cdot z_j^t) + \varepsilon_i, \tag{1}$$

where $w_{j,i}$ denotes the linear causal effect of $Z_j$ on $Z_i$ and $\varepsilon_i$ denotes some random noise with mean $\mu_i$ and variance $\sigma_i^2$, $\varepsilon_i \sim N(\mu_i, \sigma_i^2)$.

To simulate binary variables, we defined a threshold $\lambda_i \in \mathbb{R}$ and applied a step function $f$ defined as:

$$f(z_i^t, \lambda_i) = \begin{cases} 0 & \text{if} \quad (z_i^t < \lambda_i) \\ 1 & \text{if} \quad (z_i^t \geq \lambda_i) \end{cases}. \tag{2}$$

Time dependencies can be added to the data simulation by letting the variable $Z_i^t$ depend on the weighted values of $Z_j$ at time points $(t - l)$, i.e. adding lags $l$:

$$z_i^t = \sum_j \sum_{l \in L} (w_{j,i}^l \cdot z_j^{t-l}) + \varepsilon_i, \tag{3}$$

where $L$ is a nonempty set of integers smaller than or equal to $t$.

The treatment variables $T$ have an exponential decay defined through wash-in effects $\gamma$ and wash-out effects $\tau$ to simulate carry-over, similar to Percha et al. Percha et al. [2019] (see Supplementary Figure FS1 for an illustration). After drawing the exogenous variables from pre-specified distributions, the endogenous variables were generated based on assumed weights $w_{j,i}$ and the DAG.

To simulate the outcome $O$, we modeled an underlying state $U$ with a baseline drift as a discrete-time stochastic process (Wiener process), see Supplementary Text 1 and Supplementary Figure FS2 for more details. Then, $O$ at time $t$ is a linear combination of the causal effects of the other variables and the underlying state:

$$o^t = z_{i=o}^t + u^t + \varepsilon_o, \tag{4}$$

where $\varepsilon_o \sim N(\mu_o, \sigma_o^2)$, $u^t$ denotes the underlying state at time point $t$ and $Z_{i=o}^t$ denotes the linear causal effects defined in the DAG on the outcome variable $O$ at time point $t$.

#### 2.1.2 Series of simulated N-of-1 trials of chronic nonspecific low back pain

For the simulation study, our aim was to generate a realistic synthetic data set from a series of N-of-1 trials, comparing the effect of daily exercises for back strength (*Treatment 1*) with the effect of daily exercises for back stretching (*Treatment 2*) on reducing the outcome variable *Chronic Nonspecific Low Back Pain*. Pain was measured daily. The study included two blocks of each 2 treatment periods. Each treatment period had a length of 4 weeks. The treatment order was randomly selected within each block to ensure that the patient was exposed to all treatments within each block. Additionally, the study contained a baseline assessment of medication use and different sociodemographic variables. These variables were identified based on a literature review including Burdorf and Sorock [1997] and expert interviews. *Demographics* encompasses variables gender and age, which were modeled as constant variables across all time points. *Education* was included by assessing if the patient had an academic degree or was enrolled in an academic

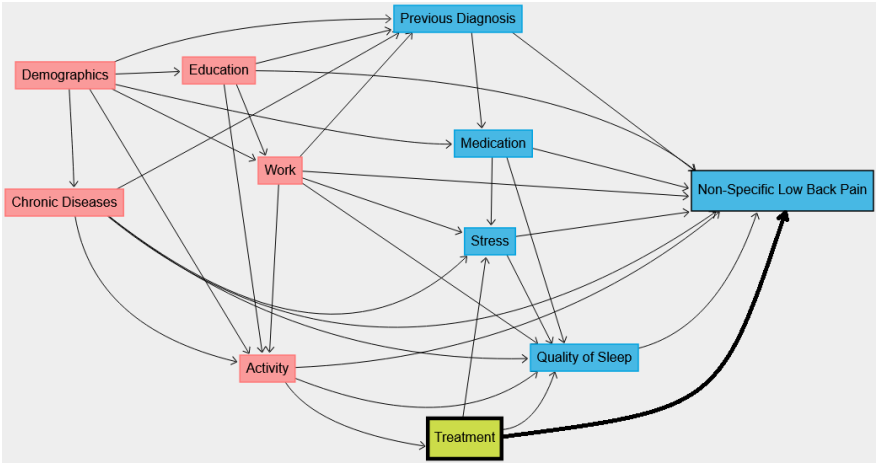


Figure 1: DAG of assumed causal effects in the simulated study. Variables in red boxes have a direct or indirect effect on treatment and the outcome, pain,

program. The variable *Work* identified whether the patient was working or not. Both variables were assumed as constant across all time points. The health status was assessed including daily measurements of *Medication*, which identified if a patient was using painkillers, *Previous Diagnosis* of nonspecific low back pain and *Chronic Diseases* indicating whether a patient had been diagnosed with related chronic diseases, for instance scoliosis or muscularly disorders. Besides the health status, lifestyle factors were tracked on a daily basis, including physical *Activity*, *Stress* levels, and *Quality of Sleep*.

We created a DAG with the assumed causal relationships between all the identified variables, see Figure 1, based on a literature review and expert knowledge. We assumed effects from Demographics on Education, Activity, Work, Previous Diagnosis, Medication and Chronic Diseases. Furthermore, we assume that there is no direct causal effect of Demographics on Nonspecific Low Back Pain, but an effect mediated by proxy variables, which leads to indirect causal paths from Demographics to nonspecific low back pain through e.g. Activity. For the treatment variables, we assumed that they have effects on Activity, Stress, Quality of Sleep and Nonspecific Low Back Pain. There is no effect on Demographics, Education, Work, Medication and Previous Diagnosis as they were assumed to be constant over time.

The DAG shown in Figure 1 was used to generate the data. Additionally to the dependencies shown in the graph, time dependencies were specified: Quality of Sleep as well as Activity depend on the previous values for Nonspecific Low Back Pain.

### 2.1.3 Generated data sets

Based on the data generation model, the low back pain study design, the DAG shown in Figure 1, and the time dependencies between variables described above, we generated data sets under four different scenarios.

Table 1: Overview of the different scenarios. In the scenarios with carry-over, the parameters for wash-in $\tau$ and wash-out $\gamma$ simulating the exponential decay were changed. $w_{A,T}$ denotes the effect from activity on treatment to introduce activity as a confounding variable.

| | **No Confounding** | **Confounding** |
|---|---|---|
| **No carry-over** | *Scenario 1*<br>$\gamma = \tau = 1$<br>$w_{A,T} = 0$ | *Scenario 3*<br>$\gamma = \tau = 1$<br>$w_{A,T} \neq 0$ |
| **With carry-over** | *Scenario 2*<br>$\gamma_1 = 3; \tau_1 = 6$<br>$\gamma_2 = 4; \tau_2 = 5$<br>$w_{A,T} = 0$ | *Scenario 4*<br>$\gamma_1 = 3; \tau_1 = 6$<br>$\gamma_2 = 4; \tau_2 = 5$<br>$w_{A,T} \neq 0$ |

All scenarios included covariate effects following the DAG in Figure 1, with the exception of the effect of activity on treatment, which was only included in some scenarios as a confounding effect. In scenario 1, the data was generated as a baseline for all methods and did not include carry-over effects, time dependencies, or confounding effect from activity

on treatment. In scenario 2, the data set was simulated with carry-over effects so that the treatment effect was heavily time-dependent. In scenario 3, strong confounding was modeled through an effect of physical activity on the treatment. In scenario 4, we combined both and generated a data set with carry-over and confounding effects.

If the edge from a variable $j$ on a variable $i$ was present in the DAG, the effect was set to $w_{ji} \neq 0$. If the edge was not present in the DAG, we assumed that there is no effect from $Z_j$ on $Z_i$ which results in $w_{ji} = 0$. Time dependencies were simulated in the same way, where we set $w_{ji}^t \neq 0$ when a time dependency was present. For all scenarios, the effects were identical except for the specifications mentioned in Table 1. The effect from treatment 1 on the outcome was set to -2 and from treatment 2 on the outcome to -4 resulting in a difference of 2 between the treatments (see Supplementary Text 2 for more details).

In addition to these 4 scenarios, we further investigated how the methods perform on data sets with missing values by replicating the 4 scenarios with missing values. For this, we used the same parameters and introduced missing values through two mechanisms. First, we deleted $10\%$ of the data points randomly with increased probability over time to mimic random drop-out. Second, we added a block of 10 consecutive missing days that were drawn randomly for each patient to simulate vacation (see Supplementary Text 3 for more details).

## 2.2 Statistical Methods

### 2.2.1 Overview

As described in the previous section, we generated datasets from 4 different scenarios, each of a series of N-of-1 trials on chronic nonspecific low back pain with 1000 participants. For the evaluation of the statistical models, in each scenario, we drew 100 samples of each 5, 10, 25, 50 and 100 participants, to also investigate the influence of the sample size in aggregated N-of-1 trials. Then we applied the different statistical models and evaluated their bias and efficiency in estimating the difference in the treatment effects of treatment 1 and treatment 2. Regarding the statistical methods, we compared standard statistical models for the analysis of aggregated N-of-1 trials, linear models that adjust for carry-over effects, G-estimation and Bayesian Networks.

### 2.2.2 Standard Statistical Models

First, we computed the sample means of both treatment groups and the naive estimate of the treatment effect difference. Its standard error was estimated as the empirical standard deviation of the estimate of the treatment effect difference across the 100 samples.

Second, we fitted a standard multiple linear regression model with pain as response variable and treatment and covariates as predictors. Hence, $\hat{\beta}_1$ in model (5) is an estimate of the effect of the treatment $T$ on the pain outcome $O$ adjusted for all covariates $C$ in Figure 1:

$$O = \beta_0 + \beta_1 T + \sum_j \beta_{2j} C_j + \varepsilon, \tag{5}$$

where $\varepsilon$ follows a normal distribution. For the implementation, we used the `lm` function in `R` from the `base` package with default settings, assuming independence between observations, computed non-weighted ordinary least squares estimates, extracted standard error estimates and performed standard Wald tests of the regression coefficients. The effect estimates and standard error estimates were then averaged across the 100 samples.

### 2.2.3 Linear Models for Wash-In and Wash-Out effects

To reduce the bias due to carry-over effect, we adjusted the multiple linear regression model for wash-in effects $\tau_j$ and wash-out effects $\gamma_j$. For that, we included a continuous time-dependent treatment effect variable instead of the binary treatment assignment variable.

Let $T_j^t$ indicate whether the patient was exposed to treatment $j$ at time point $t$ and let $\Theta_j^t$ denote its treatment effect, which we parameterized given $\tau_j$ and $\gamma_j$:

$$\Theta_j^t(T_j^t, \gamma_j, \tau_j) = \Theta_j^{t-1} + \frac{1 - \Theta_t^{t-1}}{\tau_j} \cdot T_j^t - \frac{\Theta_j^{t-1}}{\gamma_j} \cdot (1 - T_j^t). \tag{6}$$

We initialized $\Theta_0 = 0$, as we assume no treatment effect at the start point of the study. We then estimated the average effect of each treatment over time in a linear regression model:

$$O = \beta_0 + \beta_1 \Theta_1^t + \beta_2 \Theta_2^t + \sum_j \beta_{3j} C_j^t + \varepsilon. \tag{7}$$

With that, $\beta_1$ estimates the average effect of treatment 1 and $\beta_2$ the average effect of treatment 2, so that $\hat{\beta}_1 - \hat{\beta}_2$ is an estimate of the treatment effect difference. As $\Theta_j$ is a function on $\gamma_j$ and $\tau_j$ which are unknown, it was approximated through a grid search. In more detail, we iterated over several combinations of $\gamma$ and $\tau$ and fitted a linear model for each combination. Afterwards, we estimated $\gamma$ and $\tau$ by selecting the model with highest $R^2$ value. Estimates of $\beta_1$ and its standard error were obtained from the final model with highest $R^2$, using the `lm` function in `R` with default settings. The effect estimates and standard error estimates were then averaged across the 100 samples.

For an illustration, Figure 2 shows the treatment effects for 2 treatments with carry-over and the resulting overall treatment effect, which can be computed as the sum of the two treatment effects $\Theta_j^t(T_j, \tau_j, \gamma_j)$. As the observed overall treatment effect contains the effects from both treatments, it over- or underestimates the treatment effects.

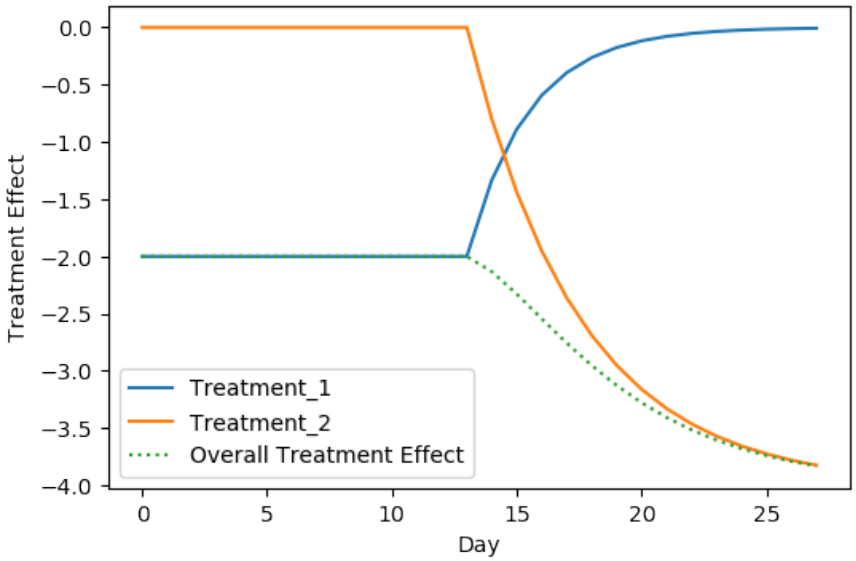


Figure 2: Illustration of the treatment effects with carry-over for a simulated patient.

#### 2.2.4 Bayesian Networks with Time Depending Variables

Bayesian Networks are graphical networks in which the joint and conditional probability distributions given by the assumed DAG are estimated and can be used for inference. To consider time dependencies, we specified lags of size 1 for the treatment variable and for nonspecific low back pain (see Figure 3). These lags were created during preprocessing and were included in the analysis as additional variables.

For the implementation, the `bnlearn` package was used. In the first step, we implemented an interface to convert the `DAGitty` graph to a Bayesian Network and defined the appropriate scale of each variable. The Bayesian Network was then fitted to the data with the `bnlearn::bn.fit` function with default settings. We estimated the parameters by the empirical mean of their posterior distribution using the `method = "bayes"` argument. For estimating the average treatment effect difference, we used the `bnlearn::cpdist` function of the fitted network Scutari et al. [2017] and first generated two random samples of patients under treatment 1 and under treatment 2, each of size 1000 through likelihood weighting given the treatment and confounding variables to have equal confounding distributions among the treatment groups. Then, we estimated the average treatment effect difference by the mean outcome difference between the two random samples. The standard error estimate of the treatment effect difference estimate was obtained by the standard deviation of the treatment effect difference across the 100 samples. In the analysis, we applied two Bayesian Networks. The first model was fitted to the DAG without any time dependencies. The second model additionally included the lags to model time-dependencies as described above.

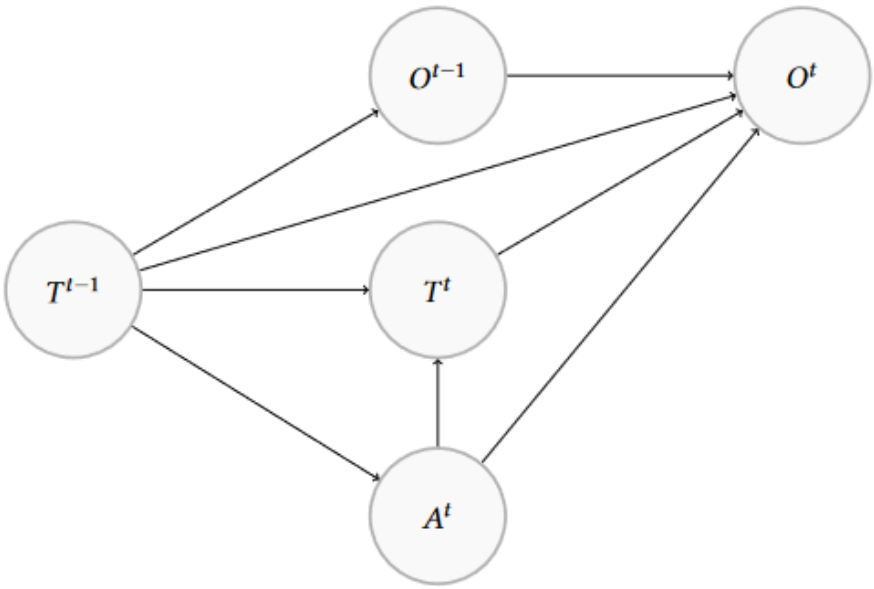


Figure 3: The DAG shows the assumed network structure between the nodes activity ($A$), treatment ($T$) and low back pain ($O$) at time point $t$ that is modeled in the Bayesian Network with lags. This is a snippet of the relationships between these three variables and embedded in the bigger network in Figure 1.

#### 2.2.5 G-estimation

G-estimation estimates average treatment effects - in our case average treatment effect differences - in a structural nested mean model Naimi et al. [2016], Hernan and Robins [2019] and can be applied to both time-varying and time-invariant treatment variables. In the structural nested mean model, we model the expected conditional difference between the outcome under treatment $T$, $O^T$, and the potential outcome under the first treatment, $O^{T=0}$, as

$$E[O^T - O^{T=0}|T, C] = \beta_1 T + \beta_2 C, \tag{8}$$

where $\beta_1$ quantifies the average causal treatment effect, $C$ includes all variables observed in the data set, and $T$ indicates whether treatment 1 ($T = 0$) or treatment 2 ($T = 1$) was given. For estimating the average treatment effect difference using G-estimation, we search for $\psi$ which minimizes $|\theta_1|$ in the following equation:

$$logit(P(T|H(\psi), C)) = \theta_0 + \theta_1 H(\psi) + \sum_j \theta_{2j} C_j, \tag{9}$$

where $\psi$ is the individual causal effect induced by the corresponding assumed rank-preserving model and $H(\psi)$ is defined as $H(\psi) = O^T - \psi T$. Minimizing for $|\theta_1|$ is based on the conditional exchangeability assumption which implies that $\theta_1$ should be 0 for the true $\psi$.

Compared to the Bayesian Networks, we do not use lags in this model. Here, we used generalized estimating equations with *independence* and *ar1* correlation structure from the R-package `geepack` Højsgaard et al. [2005] to fit equation 9. In the model based on the ar1 correlation matrix in the generalized estimating equations, the value for the estimated effect difference estimates the difference between the treatment effects with respect to the introduced wash-in effect. Standard error estimates were obtained by the standard deviation of the treatment effect differences across the 100 samples.

## 3 Results

As described in section 2.1.3, we considered four different scenarios in the simulation study. In each scenario, all methods were evaluated on 100 samples of 5, 10, 25, 50 and 100 patients, respectively. Figure 4 shows the mean estimates of the treatment effect difference for all models in all scenarios with and without missing values and their respective standard error estimates. Supplementary Text 4 provides more details.

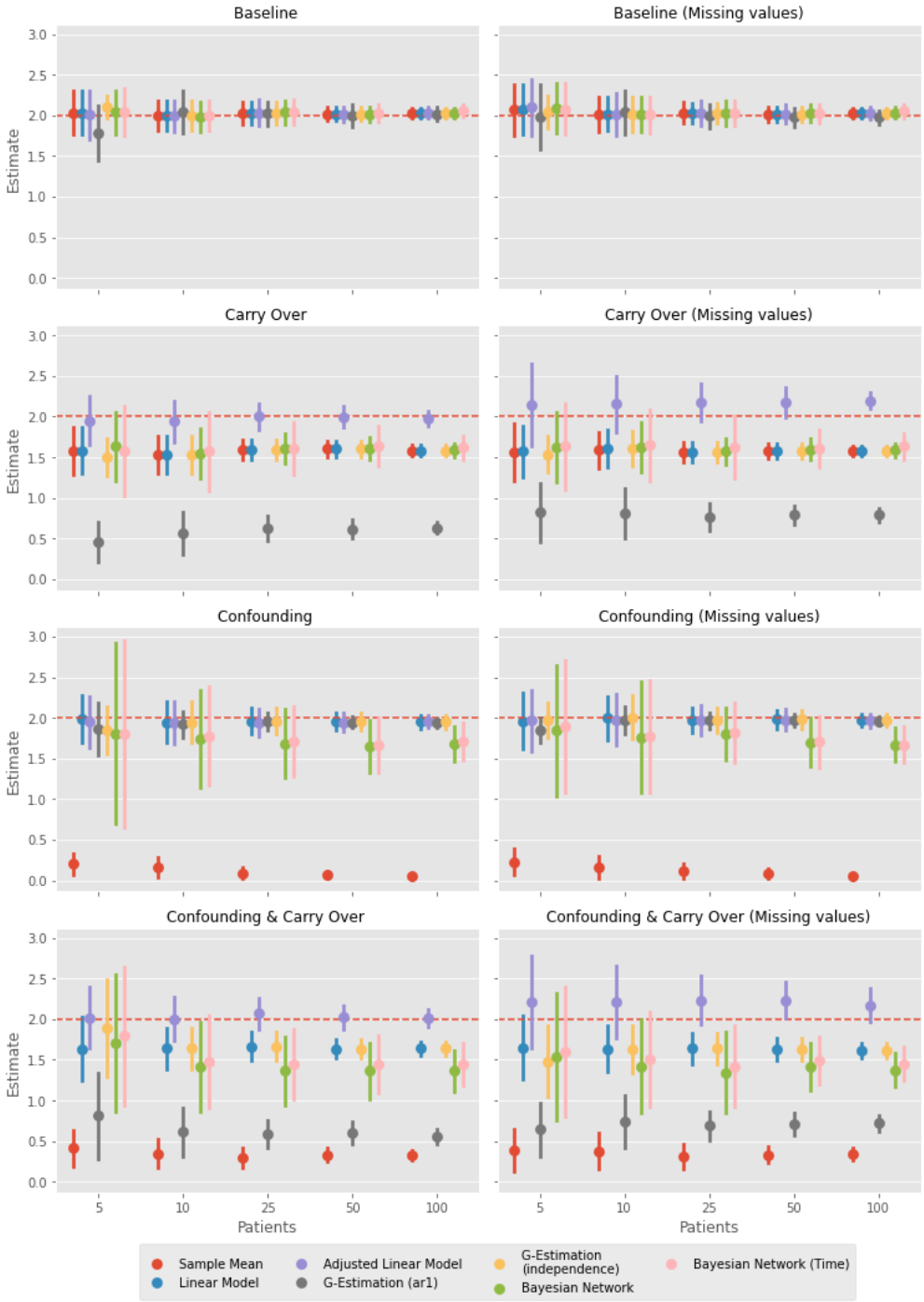


Figure 4: Overview of the estimates of the treatment effect differences (y-axis) with standard error bars, across the four scenarios with and without missing values, for different sample sizes on the x-axis.

### 3.1 Evaluation on Data without Carry-Over and Confounding

In the first scenario without confounding and without carry-over effects, all methods provided unbiased estimates of the true treatment effect difference of 2. As can be expected, the estimates were more efficient i.e. had smaller standard errors for larger sample sizes. For the smallest sample size with 5 patients, the G-Estimation (ar1) model slightly

underestimated the treatment effect. On the data with missing values, the models also provided unbiased treatment effect estimates, with slightly larger standard errors.

### 3.2 Evaluation on Data with Carry-Over

In the second scenario, we added a wash-in and a wash-out effect to our treatment variables. With that, the true simulated treatment effect slightly increased over time up to the full treatment effect. As a result, the sample mean model, linear model, G-estimation (independence) and the Bayesian Network all underestimated the treatment effect. Compared to the unadjusted Bayesian Network, the Bayesian Network with time adjustment did not improve the estimate and also underestimated the treatment effect difference. G-estimation (ar1) assumes an ar1 dependence structure within the data, violated through the exponential decay. Hence, it strongly underestimated the treatment effect in simulation with carry over effects. As only model, the adjusted linear model yielded unbiased estimates of the treatment effect for all sample sizes.

When there was missing data, the adjusted linear model tended to overestimate the treatment effect. The other models performed similarly when there were missing values compared to no missing values but with a slightly increased standard error of the treatment estimates.

### 3.3 Evaluation on Data with Confounding

By adding an effect from Activity to Treatment, we introduced confounding effects through Activity. The sample mean model heavily underestimated the treatment effect for all sample sizes. The linear model, adjusted linear model, G-Estimation (ar1) and G-Estimation (independence) all provided unbiased estimates for all sample sizes. Both Bayesian Networks slightly underestimated the treatment effect difference, which became more apparent for larger sample sizes. As another observation, Bayesian Networks yielded larger standard errors compared to linear models. By increasing the sample size, all standard errors decreased. In this scenario, missing values did not affect the performance of any model.

### 3.4 Evaluation on Data with Carry-over and Confounding

The last investigated scenario contained both carry-over and confounding. We observed that sample mean and G-estimation (ar1) strongly underestimated the treatment effect, both for complete data and data with missing values. The Bayesian Networks provided treatment effect difference estimates of about 1.5, hence underestimating the effect difference, and also yielded larger standard errors compared to the other methods. The Bayesian Network with time adjustment yielded slightly better results than the Bayesian Network without time adjustment, but did not provide a major improvement. The linear model and G-estimation (independence) provided less biased treatment effect estimates but still also underestimated the treatment effect. Finally, the adjusted linear model provided again good results in this scenario. Across all sample sizes with complete data, the model yielded unbiased estimates of the treatment effect of 2. In the data with missing values, the model slightly overestimated the treatment effect.

Overall, the adjusted linear model yielded robust results in all scenarios with complete data. Only in the data with missing values and carry-over effects, the approach tended to overestimate the effect difference.

## 4 Discussion

In this study, as a first contribution, we provided an extensive framework for simulating data for a series of N-of-1 trials by marrying stochastic processes with time-varying treatment effects embedded in a DAG. In our complex simulation models, we made assumptions on the causal structure underlying chronic low back pain and provided a detailed study design that could be translated into an actual conducted series of N-of-1 trials. As a second contribution, we then evaluated and compared different models for estimating the treatment effect under the presence of carry-over effects, confounding effects, and missing values.

As one main finding of our evaluation of the different statistical methods, we conclude that simple statistical models can provide unbiased treatment estimates across different scenarios. Furthermore, we showed that if carry-over effects are present and have not been prevented through the study design of the N-of-1 trial, it is possible to still obtain unbiased treatment effect estimates when the carry-over effects are modeled in the analysis. This adds an interesting novel perspective, in contrast to previous studies which have largely focused on the removal of carry-over through study design and have recommended against the adjustment for carry-over effects in the statistical modeling Araujo et al. [2016]. For this situation, we provide a simple method to incorporate carry-over effects into linear regression models.

Finally, our results showed that G-estimation and Bayesian Networks can provide unbiased and efficient treatment estimates as well, however, do suffer from limitations in some scenarios.

In more detail, simple methods like sample mean comparisons and linear models are easy to apply and evaluate. They were also robust against missing values and applicable for any sample size and delivered good results on the data sets without strong carry-over and without confounding variables. This is in line with the results from previous studies that t-test yield robust and valid results Chen and Chen [2014], Araujo et al. [2016]. On the other hand, sample mean comparisons do not consider carry-over effects and time dependencies and did not yield good results on the scenarios in the presence of confounding variables. Linear models performed better, but do not directly consider carry-over effects. In order to model carry-over effects, an adjusted linear model for wash-in and wash-out effect was introduced. It yielded unbiased estimates for the treatment effects across all data sets, except for some scenarios with missing data. Here, it overestimated the treatment effects but still had less bias than all other methods. All other investigated methods were not able to yield unbiased treatment effect estimates when there were carry-over effects. G-estimation performed very similar to unadjusted linear models, yielding unbiased treatment effect estimates across many scenarios when there were no carry-over effects, and was robust against missing values. However, it proved to be very important how the dependence structure in the GEE model was specified, and the AR(1) yielded largely biased estimates when there was carry-over as the simulated data did not fully fit the AR(1) assumption. Finally, we investigated two implementations of Bayesian Networks which showed robust results with respect to variations in sample size and missing values when there were no carry-over effects, similarly to G-Estimation. Interestingly, the Bayesian Networks that incorporated time lags to model the time dependencies in the data did not outperform the Bayesian Network without time dependencies. In the time dependent model Bayesian Network, we included lags of 1 in the network. However, the exponential decay used in the simulation takes multiple previous states of the treatment into account, which are not reflected in the model. We assume that this misspecification of the time dependency was the cause that the performance of the networks was not improved. For fitting Bayesian Networks, a graph has to be constructed in a first step. This can be computed based on the data, but is not recommended [Hernan and Robins, 2019, Chapter 6.5]. Assuming a pre-specified DAG is preferred for interpretability, similar to all other investigated methods. It should be noted that we obtained parameter estimates from Bayesian Networks in order to compare the results to the other methods in this study, however, the full posterior distribution of the parameters are estimated in Bayesian Networks, allowing for other analyses and interpretations as well.

As one limitation in our simulation study, we only included linear dependencies. In follow-up studies, nonlinear dependencies could be incorporated to provide even more realistic data models. The nonspecific low back pain application that we considered provided a complex N-of-1 trial and necessitated a complex generation of the DAG and simulation. For this study, we generated an outcome variable measured on an ordinal scale. In the analysis, however, we modeled the variable as a truncated Gaussian outcome. While this provides some model misspecification of all models that we investigated, we chose this evaluation to mimic a situation that occurs very often in practical analyses. In follow-up studies, other outcome distributions and other statistical models for the analysis could be investigated. Further, we performed some experiments with missing values and analyzed the data in a complete-case analysis. In practical applications, it is recommended to include some form of imputation, for example, multiple imputation. This would be especially important for the application of time-depending methods and when the data is not missing completely at random.

In follow-up analyses, it would be interesting to compare these methods in a real series of N-of-1 trials on chronic low back pain. Furthermore, additional methods like propensity score matching and inverse probability weighting could be interesting for analyzing aggregated N-of-1 trials, especially when there is missing data and selection bias. We plan to further develop the R package with all implemented methods to handle plausibility checks and include further automated tests, and to provide a computationally more efficient process of estimating $\gamma$ and $\tau$ in the adjusted linear model compared to the currently implemented grid search. Finally, we think that incorporating an adjustment for carry-over effects into G-estimation or Bayesian Networks, and investigating the use of autoregressive moving average models including exogenous covariates (ARIMAX), in addition to methods to control for selection bias, can provide even more powerful and robust tools to estimate causal treatment effects in series of N-of-1 trials.

## Backmatter

### Competing interests

The authors declare that they have no competing interests.

### Author's contributions

Conceptualization, T.G. and S.K.; methodology, T.G. and S.K.; formal analysis and data generation, T.G.; writing-original draft preparation, T.G.; writing-review and editing, S.K., J.S and T.G.; supervision, S.K. and B.A.

All authors read and agreed to the final version of the manuscript.

### Acknowledgements

The authors wish to thank Dr. Tilman Engel for critical feedback on chronic nonspecific low back pain for the generation of the directed acyclic graph.

### Availability of data and materials

The datasets generated and analysed during the current study are available in the `HIAlab/sinot` repository at github.com/HIAlab/sinot. The R code and Python code used for the analysis in this study is available at GitHub:

- R package for the statistical analysis: github.com/HIAlab/cinof1
- Python package for the data simulation: github.com/HIAlab/sinot

### Supplemental material

The following supporting information are attached:

- Supplementary Text 1: Stochastic Time-Series Data Generation Model
- Supplementary Text 2: Details on simulation study on nonspecific low back pain
- Supplementary Text 3: Simulation of Missing Values
- Supplementary Text 4: Full Results of the Simulation Study

## References

J. Nikles and G. Mitchell. *The Essential Guide to N-of-1 Trials in Health*. Springer, Dordrecht, 2015. ISBN 978-94-017-7200-6.

Karina W. Davidson, Michael Silverstein, Ken Cheung, Rocco A. Paluch, and Leonard H. Epstein. Experimental Designs to Optimize Treatments for Individuals: Personalized N-of-1 Trials. *JAMA Pediatrics*, 175(4):404–409, 04 2021. ISSN 2168-6203. doi:10.1001/jamapediatrics.2020.5801. URL `https://doi.org/10.1001/jamapediatrics.2020.5801`.

Bas C. Stunnenberg, Joost Raaphorst, Hans M. Groenewoud, Jeffrey M. Statland, Robert C. Griggs, Willem Woertman, Dick F. Stegeman, Janneke Timmermans, Jaya Trivedi, Emma Matthews, Christiaan G. J. Saris, Bas J. Schouwenberg, Gea Drost, Baziel G. M. van Engelen, and Gert Jan van der Wilt. Effect of mexiletine on muscle stiffness in patients with nondystrophic myotonia evaluated using aggregated n-of-1 trials. *JAMA*, 320(22):2344, 12 2018. doi:10.1001/jama.2018.18020. URL `https://doi.org/10.1001/jama.2018.18020`.

Salima Punja, Christopher H. Schmid, Lisa Hartling, Liana Urichuk, Catherine Jane Nikles, and Sunita Vohra. To meta-analyze or not to meta-analyze? a combined meta-analysis of n-of-1 trial data with rct data on amphetamines and methylphenidate for pediatric adhd. *Journal of Clinical Epidemiology*, 76:76–81, 8 2016. doi:10.1016/j.jclinepi.2016.03.021. URL `https://doi.org/10.1016/j.jclinepi.2016.03.021`.

J. Blackston, Andrew Chapple, James McGree, Suzanne McDonald, and Jane Nikles. Comparison of aggregated n-of-1 trials with parallel and crossover randomized controlled trials using simulation studies. *Healthcare*, 7(4):137, 11 2019. doi:10.3390/healthcare7040137. URL `https://doi.org/10.3390/healthcare7040137`.

Alexander L. Green, Amjad Shad, Rosemary Watson, Dipankar Nandi, John Yianni, and Tipu Z. Aziz. N-of-1 trials for assessing the efficacy of deep brain stimulation in neuropathic pain. *Neuromodulation: Technology at the Neural Interface*, 7(2):76–81, April 2004. doi:10.1111/j.1094-7159.2004.04010.x. URL `https://doi.org/10.1111/j.1094-7159.2004.04010.x`.

Fernando Sierra-Arango, D. M. Castaño, Jennifer D. Forero, Erika D. Pérez-Riveros, Gerardo Ardila Duarte, Maria L. Botero, Andres Cárdenas, and Jose De la Hoz-Valle. A randomized placebo-controlled n-of-1 trial: The effect of proton pump inhibitor in the management of gastroesophageal reflux disease. *Canadian Journal of Gastroenterology and Hepatology*, 2019:1–9, 2019. doi:10.1155/2019/3926051. URL `https://doi.org/10.1155/2019/3926051`.

Schmid CH, Duan N., the DEcIDE Methods Center N-of-1 Guidance Panel. Statistical design and analytic considerations for n-of-1 trials. In Kravitz RL and Duan N and eds and the DEcIDE Methods Center N-of-1 Guidance Panel, editor, *Design and Implementation of N-of-1 Trials: A User's Guide*, chapter 4, pages 33–53. Rockville, MD: Agency for Healthcare Research and Quality, Rockville, MD, 2014. URL https://effectivehealthcare.ahrq.gov/products/n-1-trials/research-2014-1/.

David D. Odineal, Maria T. Marois, Deborah Ward, Christopher H. Schmid, Rima Cabrera, Ida Sim, Youdan Wang, Barth Wilsey, Naihua Duan, Stephen G. Henry, and Richard L. Kravitz. Effect of mobile device-assisted n-of-1 trial participation on analgesic prescribing for chronic pain: Randomized controlled trial. *Journal of General Internal Medicine*, 35(1):102–111, 2019. doi:10.1007/s11606-019-05303-0. URL https://doi.org/10.1007/s11606-019-05303-0.

Charlotte Vrinten, Alexander F Lipka, Erik W van Zwet, Kirsten J M Schimmel, Martina C Cornel, Marja R Kuijpers, Yechiel A Hekster, Stephanie S Weinreich, and Jan J G M Verschuuren. Ephedrine as add-on therapy for patients with myasthenia gravis: protocol for a series of randomised, placebo-controlled n-of-1 trials. *BMJ Open*, 5(7):e007863, 2015. doi:10.1136/bmjopen-2015-007863. URL https://doi.org/10.1136/bmjopen-2015-007863.

Emily Herrett, Elizabeth Williamson, Kieran Brack, Danielle Beaumont, Alexander Perkins, Andrew Thayne, Haleema Shakur-Still, Ian Roberts, Danielle Prowse, Ben Goldacre, Tjeerd van Staa, Thomas M MacDonald, Jane Armitage, Jon Wimborne, Paula Melrose, Jayshireen Singh, Lucy Brooks, Michael Moore, Maurice Hoffman, and Liam Smeeth. Statin treatment and muscle symptoms: series of randomised, placebo controlled n-of-1 trials. *BMJ*, page n135, 2021. doi:10.1136/bmj.n135. URL https://doi.org/10.1136/bmj.n135.

Xinlin Chen and Pingyan Chen. A comparison of four methods for the analysis of n-of-1 trials. *PLoS ONE*, 9(2):e87752, 2 2014. doi:10.1371/journal.pone.0087752. URL https://doi.org/10.1371/journal.pone.0087752.

Joyce P. Samuel, Jon E. Tyson, Charles Green, Cynthia S. Bell, Claudia Pedroza, Don Molony, and Joshua Samuels. Treating hypertension in children with n-of-1 trials. *Pediatrics*, 143(4):e20181818, 2019. doi:10.1542/peds.2018-1818. URL https://doi.org/10.1542/peds.2018-1818.

Tony Zhou, Jennifer L. Dickson, and J. Geoffrey Chase. Autoregressive modeling of drift and random error to characterize a continuous intravascular glucose monitoring sensor. *Journal of Diabetes Science and Technology*, 12 (1):90–104, 7 2017. doi:10.1177/1932296817719089. URL https://doi.org/10.1177/1932296817719089.

Deborah R. Zucker, Robin Ruthazer, and Christopher H. Schmid. Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. *Journal of Clinical Epidemiology*, 63(12):1312–1323, December 2010. doi:10.1016/j.jclinepi.2010.04.020. URL https://doi.org/10.1016/j.jclinepi.2010.04.020.

Artur Araujo, Steven Julious, and Stephen Senn. Understanding variation in sets of n-of-1 trials. *PLOS ONE*, 11 (12):e0167167, December 2016. doi:10.1371/journal.pone.0167167. URL https://doi.org/10.1371/journal.pone.0167167.

Carrol Gamble, Ashma Krishan, Deborah Stocken, Steff Lewis, Edmund Juszczak, Caroline Doré, Paula R. Williamson, Douglas G. Altman, Alan Montgomery, Pilar Lim, Jesse Berlin, Stephen Senn, Simon Day, Yolanda Barbachano, and Elizabeth Loder. Guidelines for the content of statistical analysis plans in clinical trials. *JAMA*, 318(23):2337, 12 2017. doi:10.1001/jama.2017.18556. URL https://doi.org/10.1001/jama.2017.18556.

J. Pearl, M. Glymour, and N.P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, West Sussex, 2016. ISBN 9781119186847.

M.A. Hernan and J.M. Robins. *Causal Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. Taylor & Francis, London, 2019. ISBN 9781420076165.

Bethany Percha, Edward B Baskerville, Matthew Johnson, Joel T Dudley, and Noah Zimmerman. Designing robust n-of-1 studies for precision medicine: Simulation study and design recommendations. *Journal of Medical Internet Research*, 21(4):e12641, 4 2019. doi:10.2196/12641. URL https://doi.org/10.2196/12641.

Alex Burdorf and G Sorock. Positive and negative evidence of risk factors for back disorders. *Scandinavian Journal of Work, Environment & Health*, 23(4):243–256, 1997. doi:10.5271/sjweh.217. URL https://doi.org/10.5271/sjweh.217.

Marco Scutari, Pietro Auconi, Guido Caldarelli, and Lorenzo Franchi. Bayesian networks analysis of malocclusion data. *Scientific Reports*, 7(1), November 2017. doi:10.1038/s41598-017-15293-w. URL https://doi.org/10.1038/s41598-017-15293-w.

Ashley I Naimi, Stephen R Cole, and Edward H Kennedy. An introduction to g methods. *International Journal of Epidemiology*, page dyw323, December 2016. doi:10.1093/ije/dyw323. URL https://doi.org/10.1093/ije/dyw323.

Søren Højsgaard, Ulrich Halekoh, and Jun Yan. The r package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2):1–11, 2005. doi:10.18637/jss.v015.i02. URL `https://www.jstatsoft.org/index.php/jss/article/view/v015i02`.