Comparing Multisource Exchangeability Models with other Bayesian Model Approaches in the Context of Basket Trials

Table of contents

1 Internship Goals

✓ Gain an overview of existing basket trial designs and available R packages

✓ What is MEM, what is the theory behind, understand the code in the R package basket extending the package bhmbasket integrating custom models and the MEM

✓ Problem: Performance of MEM Code

✓ Compare methods in bhmbasket with MEM via simulation scenarios (changing the number of cohorts, the number of subjects within cohorts, true effects in baskets, etc.) and evaluating the operating characteristics (True and False Discovery Rates, Means Squared Error, etc.)

2 Achievements during the Internship

1. **Familiarizing with the topic:** I began by familiarizing myself with the subject matter through various sources, including works by Kaizer, Kane, Hobbs, and Stephan's master's thesis. This initial exploration gave me a rough idea of the different Bayesian Hierarchical Model (BHM) methods.
2. **Literature Review:** I conducted a literature review to identify other methods, with a particular focus on available R packages, see Chapter "Overview Methods".
3. **Available Package Comparison:** I compared all the packages I found, using custom-written code. I wrote a script that calibrates the Type I Error Rate (TOER) and ultimately plots all the results. I incorporated parallelization and analysis of simulated unique trial realizations into the code. However, I encountered performance issues due to the package `basket`, which implemented the method MEM. (Conducted performance analyses using `profvis`.)
4. **bhmbasket Package Extension Part I:** I extended the `bhmbasket` package. The extension allowed for the inclusion of custom models where the code was run with `RJAGS`.
5. **Incorporation of Custom Models:** I integrated custom models like Bayesian Model Averaging (BMA) and Proof of Concept (POC) by Jin, Liu, et al. (2020).
6. **Further Understanding:** I revisited Stephan's master's thesis to gain a deeper understanding, especially of the second part where the scenario analyses occur.

7. **Second Literature Search:** I conducted a second literature search, focusing on methods that were not published as packages but had associated papers and potentially existing code with JAGS models.

8. **Learning BUGS:** I familiarized myself with BUGS and tried to implement more complex methods like `BLAST` by Chu and Yuan (2018b), `corBHM` by Jin, Riviere, et al. (2020) and `calBHM` by Chu and Yuan (2018a). However, these variance-driven models required extensive data access for model calculation, which was kind of to overfitting. As a result, we decided to not go further into these methods.

9. **Refocusing on BHM and MEM Comparison:** I refocused on comparing BHM and MEM. I tried to understand the theory behind MEM in detail.

10. **Code Optimization and Analysis:** I revisited and optimized my initial code, focusing only on `basket` and `bhmbasket`. I conducted performance analyses and identified issues with `basket`, attempting to implement Markov Chain Monte Carlo (MCMC) in C++. This did not solve the problem of performance.

11. **bhmbasket Package Extension Part II:** I integrated MEM into `bhmbasket`, by calling the basket function in the bhmbasket, which does not solve the performance issue.

12. **basket Code Understanding and Optimization:** Understanding each step of the code and attempting to optimize the basket code I need for the integration in bhmbasekt. I believe there is still room for improvement, particularly in terms of parallelization.

13. **Optimizing and Integation** of a 6 times faster version of basket in bhmbasket. Details see in the corresponding chapter.

14. **Problem shooting of the doparallel** / do future problem in the bhmbasket package. A problem with the parallalization occured in the main package which had to be fixed.

15. **Set up Scenarios** for the analysis and thought about how to plot them.

16. **First Evaluation** I made the first evaluation and added some new scenarios. Since PoC performed quite good and MEM quite bad, I explored tuning of the prior parameters of those methods.

17. **Searching better Prior Matrix** I am trying different MEM prior matrices, since in the results, it can be seen that the prior tends to give a high probability of pooling all baskets together.

18. **Documentation and Presentation** Write and a documentation and presentation about my work.

19. **Cleaning and Uploading the code**

3 Definition: Basket Trial

- Definition: Investigate treatment effectiveness across diseases with common features, e.g., a genetic mutation
- Each basket: Patients with same medication and disease
- Benefit: Information borrowing enhances understanding of treatment response

  ⇒ Enables decision-making on futility and efficacy of treatment in each basket
- Variety of statistical designs for basket trials available, like: Pooling, stratification, Bayesian techniques

4 Overview Methods

- MEM for basket trials with one primary cohort and supplemental cohorts is presented in: Kaizer, Koopmeiners, and Hobbs (2018)
- Goes more into details of the mathematical background of MEM for basket trials also presenting more precisely the case study VEMURAFENIB (Kane developed the method from Hobbs in R): Hobbs and Landin (2018)
- Based on the paper by Kaizer, basket trials are analysed using MEM, the package basket() is presented with a case study VEMURAFENIB: Kane et al. (2019)
- Bayesian hierarchical modelling for information sharing also implemented in R as Bayesian Cluster Hierarchical Model (BCHM): Chen and Lee (2020)
- Bayesian adaptive design methodology for oncology basket trials with binary endpoints using a Bayesian model averaging framework, also implemented in bmabasket: Psioda et al. (2021)
- Overview over the published designs which are implemented in open source software: Meyer et al. (2021)
- Developed the method from Hobbs/ Kane added the possibility of local echangeability: Liu et al. (2022)
- Presentation of the R package baskexact, which implements the power prior basket trial design from Baumann et al (2024) and a design by Fujikawa et al. (2020): Baumann (2024)
- Comprehensive review of around 20 different basket trial designs: Pohl, Krisam, and Kieser (2021) this can be found in Figure 1.

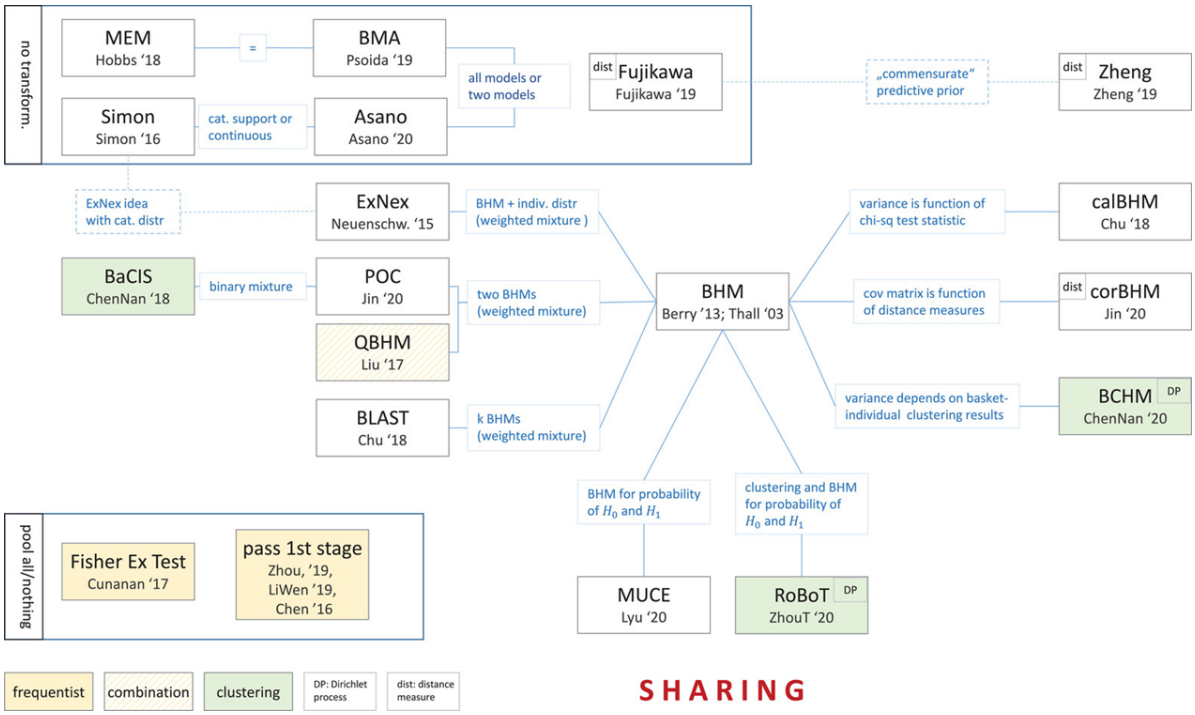


Figure 1: Overview table of different basket trial analysis approaches. Including the name of the technique, author and publication year. In white: Bayesian technique; in yellow: frequentest technique; in green: technique includes clustering. cat. = categorial. Connections between approaches symbolized by arrows. Graphic from Pohl, Krisam, and Kieser (2021).

- **Frequentist Approaches** Use a 'pool all or nothing' method. Data is evaluated at a predefined interim node of the trial, leading to a decision to either pool all baskets into one dataset or treat each basket independently. The decision is based on predefined rules.
- **Mean-driven Models** Extend the basic BHM with additional models to cover a wider range of possible distributions.
- **Non-transformed and Beta-binomial Model** Enable sharing of information directly on the different response rates, using the conjugate property of binomial data in combination with beta distributed response rates, resulting in a beta-binomial model.
- **Variance-driven BHM** The sharing in BHM is guided by the mean and variance. Larger variance allows for less sharing and more diverse values, while smaller variance results in more sharing and similar basket-individual parameters.
- **Hypothesis-driven models** Information is shared among baskets with similar probabilities for the hypothesis, not based on similar response rates.

5 Methods in bhmbasket

The methods already available in bhmbasket are Pooled, Stratified, Berry, Exnex and Exnex Adjusted.

6 Go/ No Go Decision

- **Positive effect**: Certain level of tumor cell diameter shrinking indicates positive patient response (binary endpoint, e.g., objective response according to RECIST v1.1.).
- **Cohort data:** Each cohort has a number of patients n_j and responders p_j. Estimation of response rate p_j for each basket j using posterior distribution.
- **Go decision:** $P(p_j|data > p_{boundary,j}) > \gamma$
- γ evidence level is fix, decision boundaries $p_{boundary,j}$ will be adjusted to control the type one error rate (TOER) under the α level on the following hypotheses for the best negative case: $H_0 : p_j \leq p_0$, $H_1 : p_j > p_1, p_0$ null and p_1 target rate. Normally, $p_0 = p_1$. The best negative case is the Scenario with the best true underlying response rates where we assume the treatment having an overall negative response.

6.1 TOER Calibration

The decision rule $P(p_j|data > p_{boundary,j}) > \gamma$ is equivalent to $p_{boundary,j} < q_{1-\gamma,j}$.

How to fix the evidence level γ and evaluation rule $p_{boundary,j}$?

Choose γ, e.g. as 0.2 and calculate $p_{boundary,j}$ for each cohort j s.t. the type-I-error rate (TOER) for the **best negative case** is under a certain threshold α, which we could choose 0.1.

How to do so?

In the best negative case we want to decide for NoGo so $p_{boundary,j} \geq q_{1-\gamma,j}$.

For each basket j we look at the posterior empirical $1-\gamma$ quantiles q of each run r (=1000). For each j, we save the $q_j(1), ..., q_j(r)$. An example for a given j and fixed r of $q_j(r)$ is the x-value of the red line in figure Figure 2.

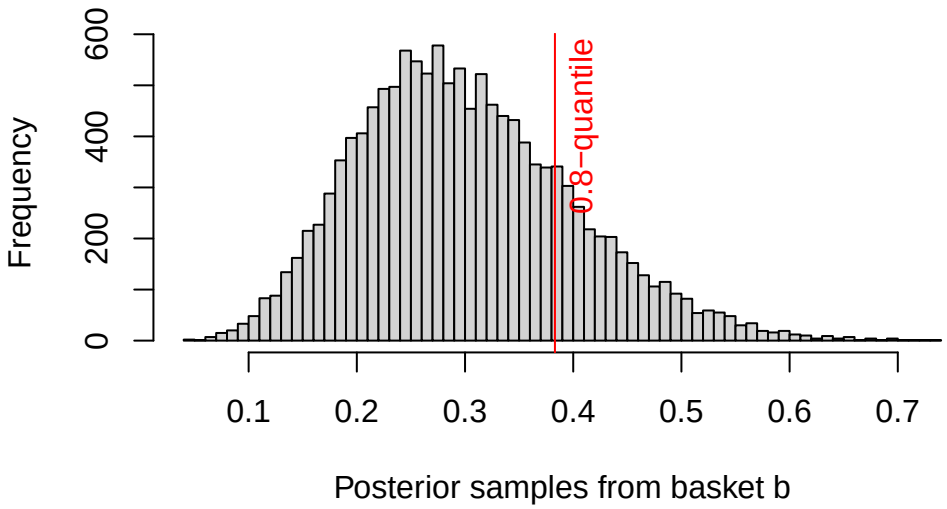


Figure 2: Posterior distribution of reponse rate of a fixed basket b and fixed run r. The red line indicates the 0.8-quantile of the distribution.

In Figure 3 all 1000 $1-\gamma$-quantiles are again plotted in a histogram plot. For $p_{boundary,j} \geq q_{1-\gamma,j}$ being hold for all runs, we choose $p_{boundary,b} > 0.8$. But since we also allow for a type one error (FalseGo) of less then 10% ($= \alpha$), we choose $p_{boundary,j} = 1-\alpha$ - quantile of the 1000 $1-\gamma$ quantiles. In this case for basket j $p_{boundary,b}$ would be 56.8%.

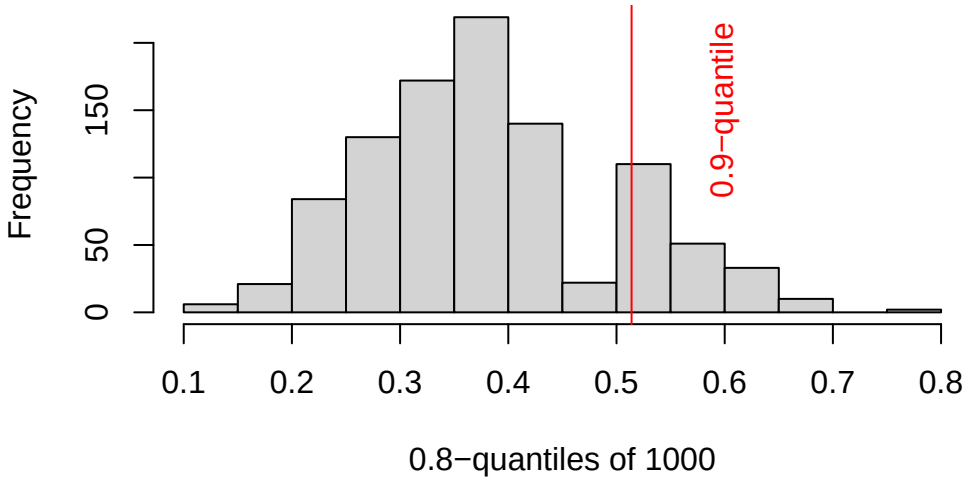


Figure 3: Histrogram of the 1000 0.8-quantiles (so 1-γ-quantile) of posterior distribution of reponse rate of a fixed basket b. The red line indicates the 0.9-quantile (so 1-α-quantile).

6.1.1 Problems with TOER Calibration

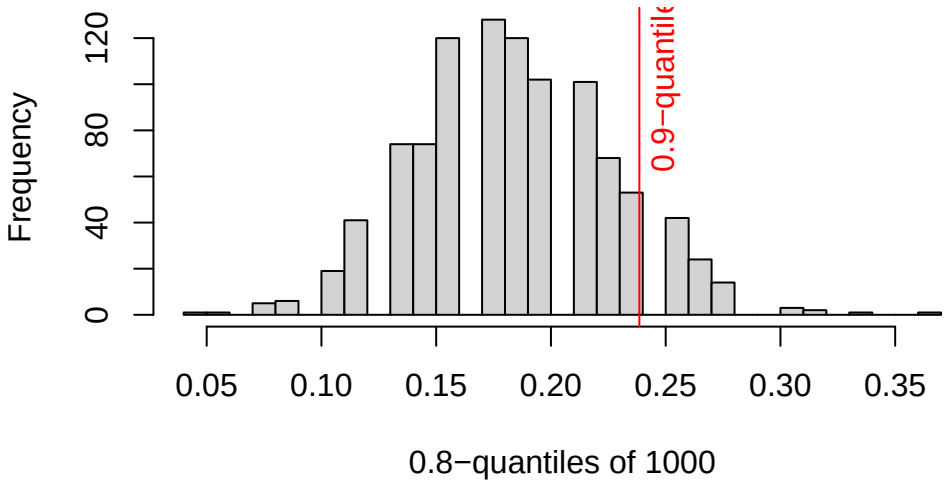


Figure 4: Histrogram of the 1000 0.8-quantiles of posterior distribution of reponse rate of a fixed basket b. The red line indicates the 0.9-quantile (so $1-\alpha$-quantile).

Solution: instead of choosing $p_{boundary,j} = 1-\alpha$ - quantile of the 1000 $1-\gamma$ quantile we add a small positive value as 0.0001. Otherwise even with the TOER calibration in some cases the percentage of go in the best negative case (here this is the all negative case) will exceed the $\alpha = 0.1$ boundary. See Figure 5.

7 Multisource Exchangeable Model (MEM)

MEM was first proposed by Hobbs and Landin (2018), Kaizer, Koopmeiners, and Hobbs (2018) and was then implemented in an R package ("Basket: Basket Trial Analysis" (n.d.)) by Kane et al. (2019). In this chapter I want to get a introduction into the method, therefore I orientate myself a the paper just named.

7.1 Introduction

Multisource Exchangeability Models (MEMs) are Bayesian hierarchical models that capture exchangeability relationships between different subgroups (baskets). In this setting, we as-

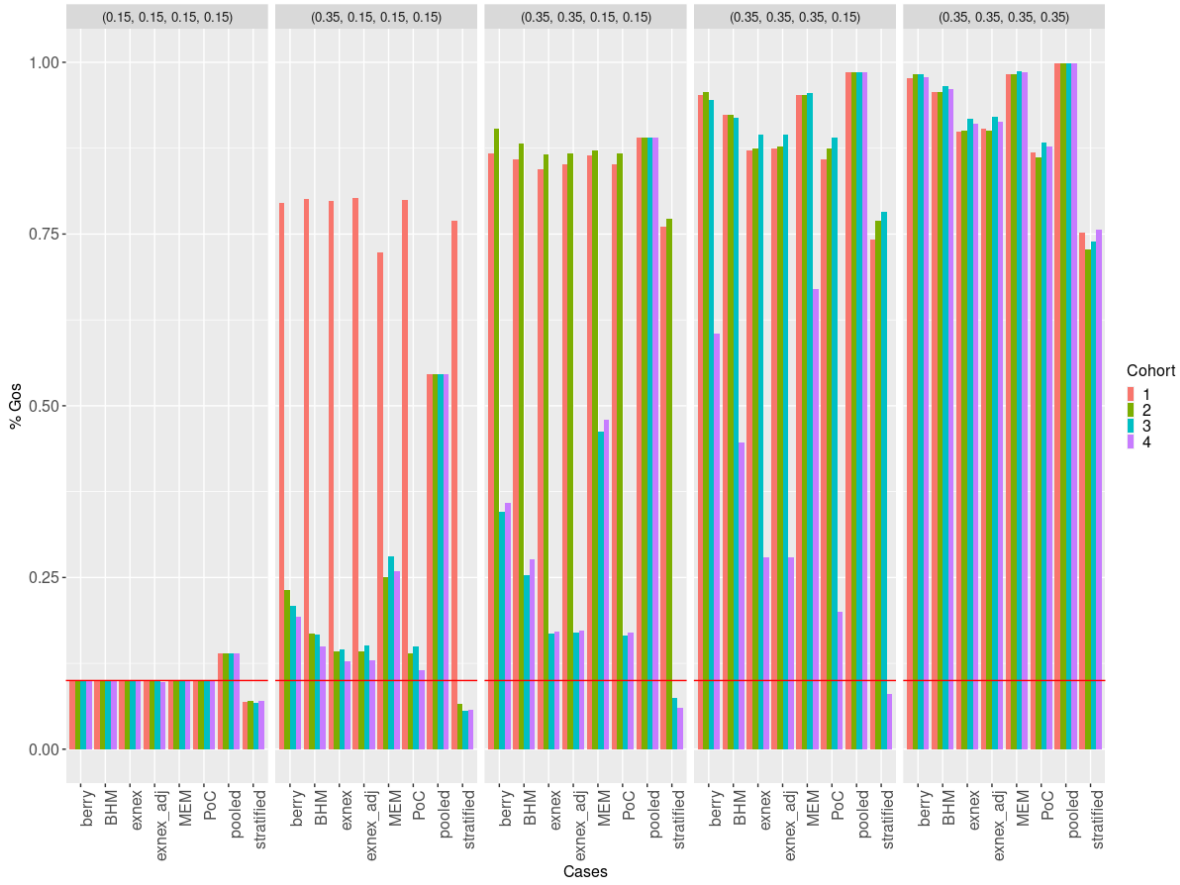


Figure 5: Probability of a Go decision in each cohort when increasing the number of cohorts with active treatment cohort when choosing the decision rule equal to the $1-\alpha$ - quantile. False Go rates = 0.1, k = 1, γ= 0.8, total number of cohorts = 4, p+ = 0.35, p-= 0.15, runs =1000. Showing the problem of % - Go exceeding the 10% level.

sume that some baskets may share similar treatment effects, but this is not known a priori. The exchangeability relationships are modeled using an exchangeability matrix Ω, where each entry $\Omega_{i,j}$ indicates whether baskets i and j are exchangeable (i.e., share the same treatment effect). This always means exchangeable to 100% (can be pooled) or not 0% (we assume to be stratified).

The goal is to derive the posterior distribution of the exchangeability matrix Ω and subsequently compute the posterior distribution of the treatment success probabilities p_j, taking into account the uncertainty in Ω.

7.2 Assumption on the Prior $P(\Omega)$

Let Ω be a symmetric $N \times N$ matrix, where each off-diagonal element $\Omega_{i,j} \in \{0, 1\}$ indicates whether baskets i and j are exchangeable. The diagonal elements are fixed as $\Omega_{i,i} = 1$, since each basket is trivially exchangeable with itself. There are $K = \prod_{j=1}^{N-1} 2^j$ possible configurations for Ω and since this number is increasing exponentially with N, we simplify this problem in assuming that each pair of baskets (i, j) is independently exchangeable with probability $P(\Omega_{i,j} = 1)$, leading to the prior which reflects the prior belief about the exchangeability of baskets:

$$P(\Omega) = \prod_{i<j} P(\Omega_{i,j} = 1)^{I(\Omega_{i,j}=1)} (1 - P(\Omega_{i,j} = 1))^{1-I(\Omega_{i,j}=1)}$$

- **Exchangeability Indicator**: The indicator function $I(\Omega_{i,j} = 1)$ equals 1 when baskets i and j are exchangeable, and 0 otherwise.
- **Prior for Exchangeability**: $P(\Omega_{i,j} = 1)$ represents the prior probability that baskets i and j are exchangeable. This can be based on prior knowledge or assumed to be a fixed value across all pairs.

7.3 Posterior Distribution $P(\Omega|r, n, a, b)$

To compute the posterior distribution, we apply Bayes' theorem:

$$P(\Omega|r, n, a, b) \propto P(r|\Omega, n, a, b)P(\Omega)$$

Where:

- $P(\Omega|r, n, a, b)$ is the posterior distribution of the exchangeability matrix Ω given the observed data r, the sample sizes n, and the prior hyperparameters a and b.

- $P(r|\Omega, n, a, b)$ is the likelihood of the data r (number of responders), given the exchangeability matrix Ω, the sample sizes n, and hyperparameters a and b.
- $P(\Omega)$ is the prior probability of the exchangeability matrix Ω which has to be fixed in advance.

7.3.1 Likelihood $P(r|n, a, b)$

Assuming at the beginning no exchangeability, so using the stratified method. The likelihood $P(r|n, a, b)$ is based on the Beta-Binomial distribution, which models the number of successes (responders) r_j out of n_j trials (patients) for each basket j, given the success probabilities p_j. These probabilities are assumed to follow a Beta distribution with hyperparameters a and b. The likelihood for a single basket j, given its success probability p_j, follows a Binomial distribution:

$$P(r_j|p_j, n_j) = Bin(r_j|p_j, n_j) = \binom{n_j}{r_j} p_j^{r_j}(1-p_j)^{n_j-r_j}$$

And so the likelihood of r, given p, n, is:

$$P(r|p, n) = \prod_{j=1}^{N} \binom{n_j}{r_j} p_j^{r_j}(1-p_j)^{n_j-r_j}$$

Where the prior for p_j, $P(p_j|a, b)$, is given by the Beta distribution:

$$P(p_j|a, b) = Beta(p_j|a, b) = \frac{p_j^{a-1}(1-p_j)^{b-1}}{B(a, b)}$$

To derive the posterior distribution, we integrate out the success probabilities p_j, which follow a Beta distribution with hyperparameters a and b:

$$P(r|n, a, b) = \int \prod_{j=1}^{N} \binom{n_j}{r_j} p_j^{r_j}(1-p_j)^{n_j-r_j} P(p_j|a, b)\, dp_j$$

This integral is the well-known Beta-Binomial integral and has a closed-form solution:

$$P(r_j|n_j, a, b) = \frac{B(r_j + a, n_j - r_j + b)}{B(a, b)}$$

Thus, the likelihood for all baskets, accounting for exchangeability, becomes:

$$P(r|n,a,b) = \prod_{j=1}^{N} \frac{B(r_j + a, n_j - r_j + b)}{B(a,b)}$$

7.3.2 Likelihood $P(r|\Omega, n, a, b)$

However, in the MEM setting, when we incorporate the exchangeability matrix Ω, the likelihood for the treatment success probabilities depends on the pooling of responder data for baskets that are considered exchangeable. For a fixed Ω the Likelihood is given:

$$P(r|\Omega,n,a,b) = \prod_{i=1}^{N} \left[\frac{B(a + \Omega[i,\ ]r,\ b + \Omega[i,\ ](n-r))}{B(a,b)} \prod_{j=1}^{N} \left(\frac{B(a + r_j,\ b + n_j - r_j)}{B(a,b)} \right)^{1-\Omega[i,j]} \right]$$

- **Pooling Baskets (Exchangeable Baskets)**: The first Beta function in the likelihood term handles the case where baskets are exchangeable:

  $$\frac{B(a + \Omega[i,\ ]r,\ b + \Omega[i,\ ](n-r))}{B(a,b)}$$

  Here, $\Omega[i,\ ]r$ refers to the sum of the responder counts r_j for all baskets j that are exchangeable with basket i (i.e., where $\Omega[i,j] = 1$). Similarly, $\Omega[i,\ ](n-r)$ sums the counts of non-responders for baskets that are exchangeable with basket i.

- **Independent Baskets (Non-Exchangeable Baskets)**: The second product handles the case where baskets i and j are **not** exchangeable:

  $$\left(\frac{B(a + r_j,\ b + n_j - r_j)}{B(a,b)} \right)^{1-\Omega[i,j]}$$

  This term is used when baskets are not exchangeable ($\Omega[i,j] = 0$). It treats the likelihood for basket j independently, using its own responder and non-responder counts.

7.3.3 Concluding

The posterior distribution of the exchangeability matrix Ω is proportional to the product of the likelihood and the prior:

$$P(\Omega|r,n,a,b) \propto \left(\prod_{i=1}^{N} \left[\frac{B(a + \Omega[i,\ ]r, b + \Omega[i,\ ](n-r))}{B(a,b)} \prod j = 1^{N} \left(\frac{B(a + r_j, b + n_j - r_j)}{B(a,b)} \right)^{1-\Omega[i,j]} \right] \right) \cdot \prod_{i<j} P(\Omega i, j = 1)^{I(\Omega_{i,j}=1)} (1 - P(\Omega i, j = 1))^{1-I(\Omega_{i,j}=1)}$$

7.4 Posterior Distribution of p_j

Once we have the posterior distribution of Ω, we can compute the posterior distribution of the treatment success probabilities p_j. The posterior distribution of p_j is a mixture of Beta distributions, where the weights are determined by the posterior probabilities of the different exchangeability structures Ω.

The marginal posterior distribution of p_j is given by:

$$P(p_j|r, n, a, b) = \sum_{\Omega} P(p_j|r, \Omega, n, a, b)P(\Omega|r, n, a, b)$$

Where $P(p_j|r, \Omega, n, a, b)$ follows a Beta distribution if the baskets are exchangeable:

$$P(p_j|r, \Omega, n, a, b) \sim \text{Beta}\left(a + \sum_h \Omega_{j,h} r_h, b + \sum_k \Omega_{j,k}(n_k - r_k)\right)$$

This reflects the updated belief about the success probability p_j based on the observed responder data r_j, the exchangeability structure Ω, and the prior parameters a and b.

8 R Package: Implementation of MEM with Example

In this chapter, we're going to break down how the code for a method MEM works. We'll also learn how to calculate the posterior distribution of the exchangeability matrix Ω and how to get posterior samples.

As we learned in the previous chapter, MEMs was first introduced by Hobbs and Kaizer. Later, it was implemented in a R package by Kane. This package is called basket.

We'll use what we learned from the previous chapter to understand how this code works and how to use it. So, let's dive in and start exploring the world of MEMs, the MEM matrix, and posterior samples.

8.1 Example of the Exchangeability Matrix Ω

An example of an Ω can be found in Figure 6.

But the problem is that Ω is unknown and has to be detected.

$$\Omega = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

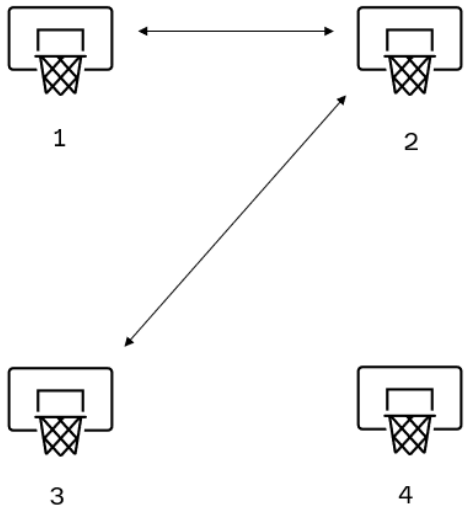


Figure 6

8.2 Configurations of Exchangeability Matrix Ω

- $K = \prod_{j=1}^{N-1} 2^j$ possible configurations for Ω

  N=4 $\Rightarrow$ K=64

$$\begin{bmatrix} - & - & - & - \\ 2 & - & - & - \\ 2 & 2 & - & - \\ 2 & 2 & 2 & - \end{bmatrix}$$

- Assuming independence: We only have $\frac{N(N-1)}{2}$ possible configurations left (triangular numbers)

  N=4 $\Rightarrow$ K=6

$$\begin{bmatrix} - & - & - & - \\ 1 & - & - & - \\ 2 & 3 & - & - \\ 4 & 5 & 6 & - \end{bmatrix}$$

- Since distribution of Ω is unknown and depends on n and r we use a Markov Chain Monte-Carlo algorithm (namely the Metropolis) to sample from the distribution.

8.3 Finding Posterior Distribution of Ω using the Metropolis Algorithm

Call the number of different configurations $d = \frac{N(N-1)}{2} = 6$.

1. Choose an arbitrary initial Ω_0, e.g. identity matrix.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

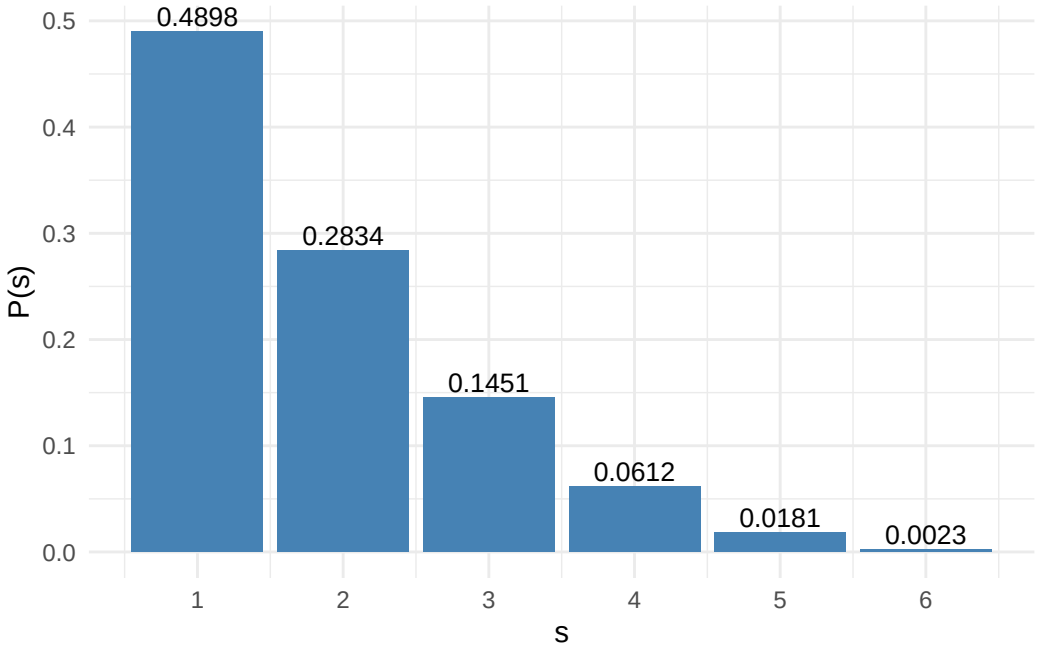


Figure 7: The function P(s)

2. Choose a proposal function to receive a new sample Ω'.

   2.1. Therefore, randomly sample s out of $1, ..., d$ with the probability $P(s) = \frac{(d+1-s)^3}{1^3+...+d^3}$ which can also between in Figure 7. Higher numbers are less probable. Assume we got $s = 2$.

   2.2. Randomly define the vector v of length s by drawing s-times uniformly out of the set $\{1, ..., d\}$ without replacement. Let $v = \{3, 6\}$.

2.3. Define the proposal matrix for the $t+1$ iteration Ω' as following: Let

$$\Omega_t = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

Using the index matrix

$$\begin{bmatrix} NA & 1 & 2 & 3 \\ 1 & NA & 4 & 5 \\ 2 & 4 & NA & 6 \\ 3 & 5 & 6 & NA \end{bmatrix}$$

setting the entries in v to 0 if they are already 1, otherwise to 1. So,

$$\Omega' = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

3. Calculate the acceptance ratio $\rho = exp(log(f(\Omega')) - log(f(\Omega_t))) = \frac{f(\Omega')}{f(\Omega_t)}$ with $f(\Omega)$ be the function that is proportional to the desired probability function $P(\Omega)$, given: $f(\Omega) = \left(\prod_{i=1}^{N} \left[\frac{B(a+\Omega[i,\ ]r,\ b+\Omega[i,\ ](n-r))}{B(a,b)} \prod_{j=1}^{N} \left(\frac{B(a+r_j,\ b+n_j-r_j)}{B(a,b)}\right)^{1-\Omega[i,j]}\right]\right)$ $\cdot \prod_{m=1}^{N(N+1)/2} \left(W[\Delta(m)]^{\Omega[\Delta(m)]} \cdot (1 - W[\Delta(m)])^{(1-\Omega[\Delta(m)])}\right)$. Were Δ is the function that maps index $\{1, ..., \frac{N(N-1)}{2}\}$ in

$$\begin{bmatrix} - & - & - & - \\ 1 & - & - & - \\ 2 & 3 & - & - \\ 4 & 5 & 6 & - \end{bmatrix}$$

to the corresponding entry in the matrix. And W is the predefined prior weight matrix, e.g.

$$\begin{bmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{2} \\ \frac{1}{2} & 1 & \ddots & \vdots \\ \vdots & & \ddots & \frac{1}{2} \\ \frac{1}{2} & \cdots & \frac{1}{2} & 1 \end{bmatrix}$$

.

4. Accept or reject

   4.1. Generate a uniform random number $u \in [0, 1]$.

   4.2. If $u \leq \rho$, then *accept* the candidate by setting $\Omega_{t+1} = \Omega_t$,

   4.3. If $u > \rho$, then *reject* the candidate and set $\Omega_{t+1} = \Omega_t$ instead.

8.3.1 Saving Information from the MCMC

- Save Ω_t for $t = n_{burnin+1}, ..., n_{iteration}$, e.g., $mcmc_{iteration} = 200.000$, $mcmc_{burnin} = 50.000$.
- Transform information. Example: $N = 4$ and for fixed t let

$$\Omega_t = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

- Introduce $2^{N-1\times N}$-encoding matrix Φ containing as rows the possible configuration for exchangeability of basket j with the other baskets. The columns correspond to the following basket order [bj| b1| ...| bj-1| bj+1|...| bN].
  - In our case, if j=3: [b3|b1|b2|b4].
  - This means in the first row we have the case that basket 3 is not exchangeable with the other baskets. In the second row basket 3 is exchangeable with basket 1, etc.
  - So in the case $N = 4$:

$$\Phi = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

- b1 is exchangeable with b3 and b4 $\Rightarrow$ combination in row 7 of Φ (read columns of Φ as [b1| b2|b3|b4])
- b2 is exchangeable with no baskets $\Rightarrow$ combination in row 1 of Φ (read the columns of Φ as [b2| b1|b3|b4])
- b3 is exchangeable with b1 $\Rightarrow$ combination in row 2 of Φ (read the columns of Φ as [b3| b1|b2|b4])
- b3 is exchangeable with b1 $\Rightarrow$ combination in row 2 of Φ (read the columns of Φ as [b4| b1|b2|b3])
- Putting this information in matrix Ψ_t. Ψ_t has one entries in the ith row and jth column if for the jth basket the exchange combination can be found in the ith row of the encoding matrix Φ.

$$\Psi_t = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

- Calculate $\Psi = \frac{\sum_{t=n_{burnin+1}}^{n_{iternation}} \Psi_t}{n_{iteration} - n_{burnin}}$. An example for responses = (4, 2, 7, 1) and size= (8,8,8,8) is:

$$\sum_{t=n_{burnin+1}}^{n_{iternation}} \Psi_t = \begin{bmatrix} 27464 & 9001 & 129899 & 19545 \\ 2 & 42239 & 22253 & 19174 \\ 12218 & 3 & 12398 & 123 \\ 560 & 67307 & 4 & 4797 \\ 401 & 288 & 50851 & 5 \\ 15592 & 47206 & 46 & 12 \\ 6 & 45211 & 70690 & 31 \\ 3 & 7 & 955 & 79 \\ 1 & 34 & 8 & 1344 \\ 247 & 1 & 30 & \end{bmatrix} \Rightarrow \Psi = \begin{bmatrix} 0.18 & 0.06 & 0.87 & 0.13 \\ 0.28 & 0.15 & 0.13 & 0.08 \\ 0.08 & 0 & 0 & 0.45 \\ 0.03 & 0 & 0 & 0.34 \\ 0.10 & 0.31 & 0 & 0 \\ 0.30 & 0.47 & 0 & 0 \\ 0.01 & 0 & 0 & 0 \\ 0.01 & 0 & 0 & 0 \end{bmatrix}$$

8.4 Sampling from the Posterior Distribution of p_j

- The matrix Ψ provides the probability of the different exchangeability possibilities for each basket (column).
- To obtain a sample from the posterior distribution of p_j, for each basket j we processed as follows:
  - An exchange model ω (row) in Φ is randomly selected for the basket j. This selection is multinomially distributed with the probability vector $\Psi[\ , j]$.
  - A random sample from the posterior distribution of p_j is drawn from $Beta(a + \omega r, b + \omega(n - r))$.
  - This process is repeated for each basket run times.

8.5 Adding Ψ as Output

To better observe which baskets the model assumes to be more exchangeable, I added the Ψ_t to the analysis output. Here is an example output from a fixed scenario, first run:

```
$posterior_quantiles
             p_1        p_2        p_3
2.5%  0.03239645 0.02866864 0.02085829
5%    0.03972592 0.03563848 0.02811741
10%   0.04915679 0.04441044 0.03742892
20%   0.06256930 0.05700554 0.04989249
50%   0.09573349 0.08634508 0.07867800
80%   0.14452478 0.12403467 0.11346243
95%   0.22567518 0.17351447 0.15288617
97.5% 0.27116237 0.19913254 0.16933765
Mean  0.10927293 0.09312688 0.08302714
SD    0.06268537 0.04404920 0.03843094

$Psi
       basket 1   basket 2   basket 3
[1,] 0.05289333 0.01126000 0.04004667
[2,] 0.22844667 0.09794667 0.06916000
[3,] 0.08498667 0.12662000 0.24129333
[4,] 0.63367333 0.76417333 0.64950000

$Encoder
       Var1 Var2
[1,] 1    0    0
[2,] 1    1    0
[3,] 1    0    1
[4,] 1    1    1
```

Calculating the mean over the output of all runs, we receive a Ψ-matrix which is indicating the exchangeable probabilities:

```
       basket 1    basket 2   basket 3
[1,] 0.02191867 0.006419333 0.01132267
[2,] 0.10719133 0.052152667 0.04724933
[3,] 0.07965733 0.095156667 0.11778733
[4,] 0.79123267 0.846271333 0.82364067
```

This output can be interpreted as shown in Figure 8 . So what do you think, are the response rates of the different baskets similar?

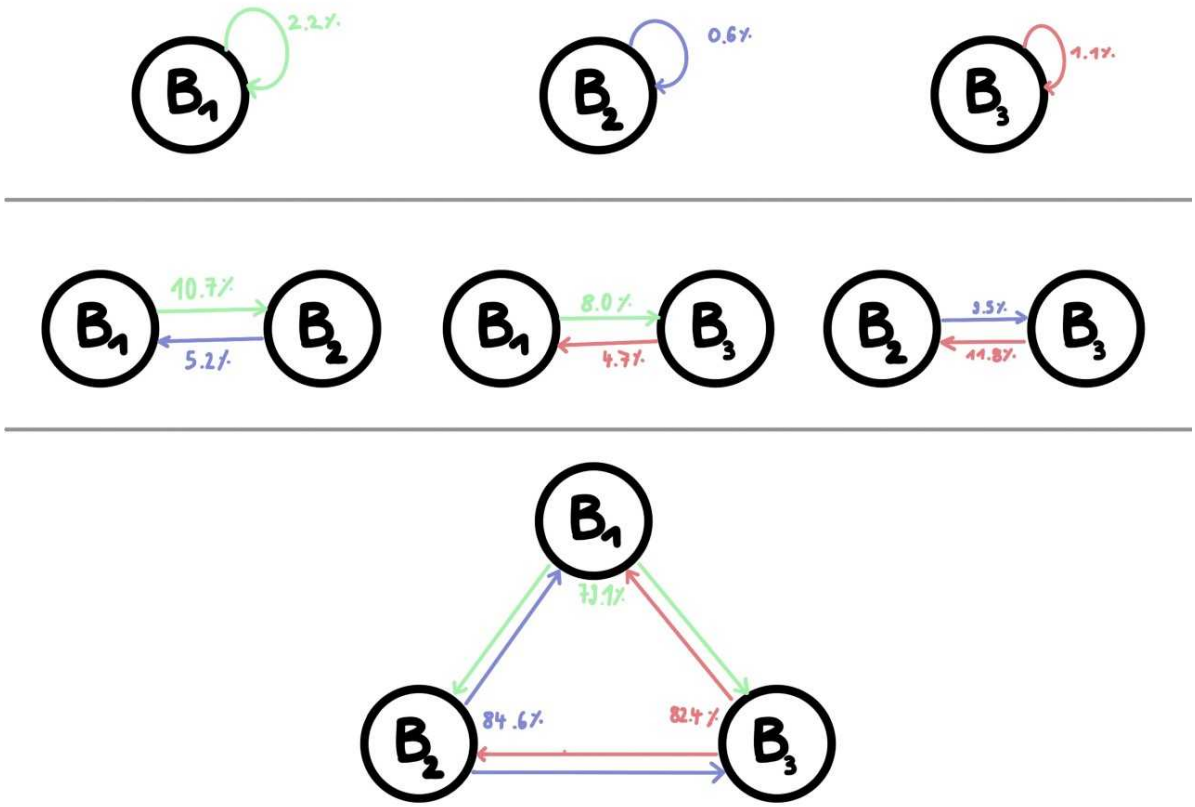


Figure 8: How to read generated matrix Ψ. B1, B2, B3 stand for the different baskets/ cohorts. The percentage on the arrows is the exchange probability.

9 Optimizing the Basket Function

The goal of the internship is it to compare the basket trial methods already included in bhmbasket with MEM. In a first step, I extended the bhmbasket package and made the option method="MEM" available. The first idea was that the package bhmbasket just calls the basket package. The problem was that the performance of basket is bad even with parallelization (workers=32). Therefore I wrote my one code, an optimized version of the basket package where I could reduced the performance by the factor 6. In the following I will explain the changes I made. The biggest changes I've added in the MCMC since these is the part which costs the most performance time.

9.1 Discarding Unnecessary Calculations

The function basket in the package basket is calculating several different characteristics (as the effective size, the MAP, ...). For the integration into bhmbasket package only the output of the posterior samples from the posterior distribution is of interest, which is why I deleted all the other calculations.

9.2 Transformation into the Log-Space

However, I noticed that the transformation into the logarithmic space was not consistently applied. Specifically, the rules of logarithmic calculations, which would typically convert products and exponentiations into sums and products respectively, were not fully utilized. This inconsistency led me to consider simplifying the $log(f)$ to:

$$\begin{aligned} &\sum_{i=1}^{N} log\left(B\left(a+\Omega[i,\ ]r,\ b+\Omega[i,\ ](n-r)\right)\right) - log\left(B(a,b)\right) \\ &+\sum_{j=1}^{N}\left(1-\Omega[i,j]\right)\left(log\left(B\left(a+r_j,\ b+n_j-r_j\right)\right) - log\left(B(a,b)\right)\right) \\ &+\sum_{m=1}^{N(N+1)/2}\left(\Omega[\Delta(m)]log\left(W[\Delta(m)]\right) + \left(1-\Omega[\Delta(m)]\right) log\left(1-W[\Delta(m)]\right)\right) \end{aligned}$$

In the original calculations, the following term was repeatedly computed, which significantly impacted performance:

$$\begin{aligned} \log(f(\Omega)) = \sum_{i=1}^{N}\log\left[\frac{B(a+\Omega[i,\ ]r, b+\Omega[i,\ ](n-r))}{B(a,b)}\prod_{j=1}^{N}\left(\frac{B(a+r_j, b+n_j-r_j)}{B(a,b)}\right)^{1-\Omega[i,j]}\right] \\ +\log\left(\prod_{m=1}^{N(N+1)/2}\left(W[\Delta(m)]^{\Omega[\Delta(m)]}\cdot\left(1-W[\Delta(m)]\right)^{(1-\Omega[\Delta(m)])}\right)\right) \end{aligned}$$

In addition to this, I also carried out minor transformations into the logarithmic space.

9.3 Calculating only the Difference in the First Factor of the Acceptance Ratio

Another improvement was instead of calculating in each step the $log(f(\Omega'))$ and $log(f(\Omega_t))$ only calculating the difference in the chance. For a better understanding we look at the following example:

Let

$$\Omega_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \ \Omega' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

So the only differences in the two matrices are in the second and third row/column.

$$\frac{f(\Omega')}{f(\Omega_t)} = \frac{\left(\prod_{i=1}^{N} \left[\frac{B(a+\Omega'[i,\ ]r,\ b+\Omega'[i,\ ](n-r))}{B(a,b)} \prod_{j=1}^{N} \left(\frac{B(a+r_j,\ b+n_j-r_j)}{B(a,b)} \right)^{1-\Omega'[i,j]} \right] \right)}{\left(\prod_{i=1}^{N} \left[\frac{B(a+\Omega_t[i,\ ]r,\ b+\Omega_t[i,\ ](n-r))}{B(a,b)} \prod_{j=1}^{N} \left(\frac{B(a+r_j,\ b+n_j-r_j)}{B(a,b)} \right)^{1-\Omega_t[i,j]} \right] \right)} \cdot \frac{\prod_{m=1}^{N(N+1)/2} \left(W[\Delta(m)]^{\Omega'[\Delta(m)]} \cdot (1-W[\Delta(m)])^{(1-\Omega'[\Delta(m)])} \right)}{\prod_{m=1}^{N(N+1)/2} \left(W[\Delta(m)]^{\Omega_t[\Delta(m)]} \cdot (1-W[\Delta(m)])^{(1-\Omega_t[\Delta(m)])} \right)}$$

We are leaving the second factor (and simplify it later as far as possible when taking it into the log) and looking in more detail into the first factor which we already transformed in the chapter before by haven taken the log.

$$\sum_{i=1}^{N}\left[log\left(\frac{B(a+\Omega'[i,\ ]r,\ b+\Omega'[i,\ ](n-r))}{B(a,b)}\right)+\sum_{j=1}^{N}(1-\Omega'[i,j])\ log\left(\frac{B(a+r_j,\ b+n_j-r_j)}{B(a,b)}\right)\right]$$

$$-\sum_{i=1}^{N}\left[log\left(\frac{B(a+\Omega_t[i,\ ]r,\ b+\Omega_t[i,\ ](n-r))}{B(a,b)}\right)+\sum_{j=1}^{N}(1-\Omega_t[i,j])\ log\left(\frac{B(a+r_j,\ b+n_j-r_j)}{B(a,b)}\right)\right]$$

$$=\sum_{i=1}^{N}log\left(\frac{B(a+\Omega'[i,\ ]r,\ b+\Omega'[i,\ ](n-r))}{B(a,b)}\right)-log\left(\frac{B(a+\Omega_t[i,\ ]r,\ b+\Omega_t[i,\ ](n-r))}{B(a,b)}\right)$$

$$+\sum_{j=1}^{N}(1-\Omega'[i,j])\ log\left(\frac{B(a+r_j,\ b+n_j-r_j)}{B(a,b)}\right)-(1-\Omega_t[i,j])\ log\left(\frac{B(a+r_j,\ b+n_j-r_j)}{B(a,b)}\right)$$

$$=\sum_{i=1}^{N}\left[log\left(\frac{B(a+\Omega'[i,\ ]r,\ b+\Omega'[i,\ ](n-r))}{B(a+\Omega_t[i,\ ]r,\ b+\Omega_t[i,\ ](n-r))}\right)+\sum_{j=1}^{N}(\Omega_t[i,j]-\Omega'[i,j])\ log\left(\frac{B(a+r_j,\ b+n_j-r_j)}{B(a,b)}\right)\right]$$

$$=\sum_{i=1}^{N}\begin{cases}0 & \text{if } \Omega_t[i,\ ]=\Omega'[i,\ ]\\ log\left(\frac{B(a+\Omega'[i,\ ]r,\ b+\Omega'[i,\ ](n-r))}{B(a+\Omega_t[i,\ ]r,\ b+\Omega_t[i,\ ](n-r))}\right)+\sum_{j=1}^{N}\begin{cases}0 & \text{if } i=j\\ (\Omega_t[i,j]-\Omega'[i,j])\ log\left(\frac{B(a+r_j,\ b+n_j-r_j)}{B(a,b)}\right) & \text{else}\end{cases} & \text{else}\end{cases}$$

$$=\sum_{i=1}^{N}\left[log\left(\frac{B(a+\Omega'[i,\ ]r,\ b+\Omega'[i,\ ](n-r))}{B(a+\Omega_t[i,\ ]r,\ b+\Omega_t[i,\ ](n-r))}\right)+\sum_{j=1}^{N}(\Omega_t[i,j]-\Omega'[i,j])\ log\left(\frac{B(a+r_j,\ b+n_j-r_j)}{B(a,b)}\right)\right]$$

$$=\sum_{i=1}^{N}\begin{cases}0 & \text{if } \Omega_t[i,\ ]=\Omega'[i,\ ]\\ log\left(B(a+\Omega'[i,\ ]r,\ b+\Omega'[i,\ ](n-r))\right)-log(B(a+\Omega_t[i,\ ]r,\ b+\Omega_t[i,\ ](n-r)))+ & \\ \sum_{j=1}^{N}\begin{cases}0 & \text{if } i=j\\ (\Omega_t[i,j]-\Omega'[i,j])\ \left(log(B(a+r_j,\ b+n_j-r_j))-log(B(a,b))\right) & \text{else}\end{cases} & \text{else}\end{cases}$$

So firstly Ω_t and Ω' where compared and only the differences where newly calculated.

9.4 Verification of the Optimized Code

In Figure 9 are example runs where I plotted the empiric posterior density functions for each basket, in blue the output from the basket package, in red my output.

9.5 Code

The code will be available in bitbucket and in my student folder.

10 Compared Methods

In the following the methods we compare with the MEM technique will be explained. We use the following notation:

N : number of cohorts

r_j : number of responses in the jth cohort

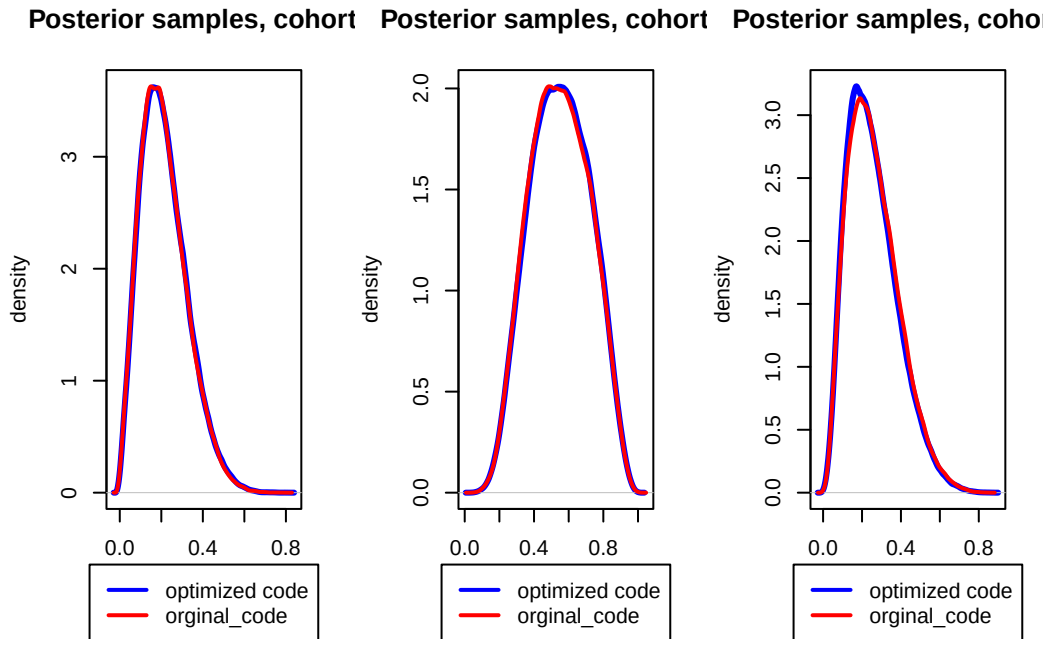
n_j : number of patients in the jth cohort

p_j : underlying response rate

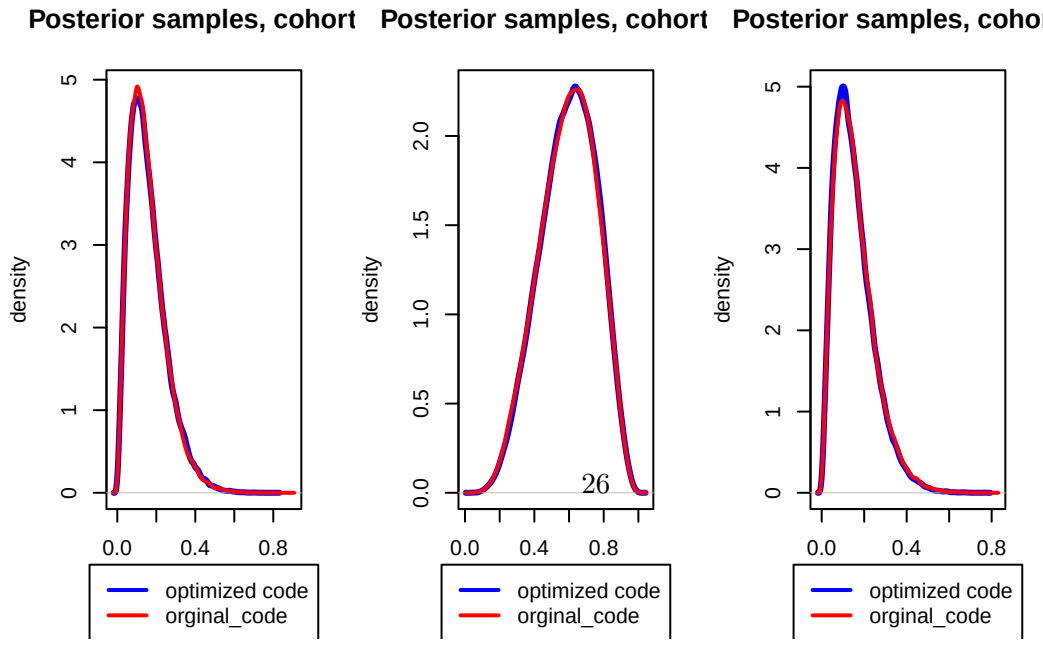
```
[1] 1 5 2

Warning: executing %dopar% sequentially: no parallel backend registered

value: 35
value: -1
value: -1
```




```
[1] 1 5 1
value: 134
value: -1
value: -1
```




```
[1] 2 4 0
value: 1
```

$\hat{p_j}$: assumed underlying response rate (target rate)

$logit(x) = log\,(\frac{x}{1-x})$: the logit function

10.1 Stratified Method

The Stratification method involves analyzing each cohort separately, under the assumption that the response rate of each cohorts are independent and do not affect each others. The number of responses in each cohort follows a binomial distribution. A conjugate beta prior is used for the response rate:

$$r_j \sim Bin(p_j, n_j),\ j = 1, ..., N$$
$$p_j \sim Beta(\hat{p_j}, 1 - \hat{p_j})$$

The posterior distribution of the response rate of each cohort j is than given by:

$$Beta(\hat{p_j} + r_j, 1 - \hat{p_j} + n_j - r_j)$$

However, this method can lead to response rate estimates with a small bias, a large variance and an increase risk of FalseGos due to the smaller number of observations. (Wojciekowski (2018))

10.2 Pooled Method

The Pooling method treats all cohorts as a single group, assuming that they all have the same underlying response rate. The unit-information variance of a beta-distributed random variable is calculated, and the response rate with the largest unit-information variance is selected for the overall variance: $a = arg\ min_j(\hat{p_j} - 0.5)$.

The posterior distribution of the response rate under the pooled analysis of each cohort is then calculated:

$$Beta\left(a + \sum_{j=1}^{J} r_j, 1 - a + \sum_{j=1}^{J} n_j - \sum_{j=1}^{J} r_j\right)$$

This method can result in a smaller variance but potentially larger bias if the assumption of exchangeability among cohorts is violated. It may also lead to overlooking potentially interesting cohorts. (Wojciekowski (2018))

10.3 BHM by Thall

Thall et al. (2003) introduced a Bayesian hierarchical technique for basket trials where the cohorts are exchangeable but the level of "borrowing" is adjusted on the observed data. The response rates p_j are random but adhere to a common underlying distribution. The overall mean of the log-odds of the response rates for all cohorts is represented by the mean μ. This mean is influenced by the information from all cohorts. The variance τ^2 affects the level of borrowing across the cohorts. ($\tau^2 = 0$: all cohorts have the same posterior distribution. On the other hand, when $\tau^2 \longrightarrow \infty$: no borrowing). The more similar the observed number of responses r_j in the cohorts, the smaller the variance of the posterior distribution of τ^2 , the more similar the posterior distribution of and p_j, and the greater the borrowing. The same principles apply to the methods proposed by Berry et al. (2013) and Neuenschwander et al. (2016), and therefore to the ExNex Adjusted method as well. (Wojciekowski (2018))

$$
\begin{aligned}
r_j &\sim Bin(p_j, n_j) \\
logit(p_j) &\sim N(\mu, \tau^2) \\
\mu &\sim N(m, \nu) \\
\tau &\sim HN(s_\tau)
\end{aligned}
$$

equivalent to

$$
\begin{aligned}
r_j &\sim Bin(p_j, n_j) \\
p_j &\sim Logit - normal(m_\mu, \nu_\mu + \tau^2) \\
\tau &\sim HN(s_\tau)
\end{aligned}
$$

I used a slight different prior (truncated t distribution for the precision) for BHM than for berry, which could also cause differences in the results between BHM and the in the following presented method by Berry:

BHM BUGS code:

```
model{
 mu ~ dnorm(mean_mu,precision_mu)
 sigma ~ dt(0,precision_tau,1)T(0,)

 for(j in 1:J){
 r[j] ~ dbin(p[j],n[j])
 p[j] <- exp(theta[j])/(1+exp(theta[j]))
 theta[j] ~ dnorm(mu,1/(sigma)^2)
 }

}
```

10.4 Berry Method

Different tumor sites have distinct historical response rates, or the standard treatments and their effectiveness can vary by tumor site. Berry et al. (2013) suggested a modification to Thall et al. (2003)'s simple model that takes into account. They suggested modeling the difference between the response rate and the assumed positive response rate on the log-odds scale $logit(p_j) - logit(\hat{p_j})$, assuming that the difference on this scale is exchangeable. (Wojciekowski (2018))

$$
\begin{aligned}
r_j \sim Bin(p_j, n_j),\ j = 1, ..., K \\
logit(p_j) - logit(\hat{p_j}) \sim N(\mu, \tau^2) \\
\mu \sim N(m_\mu, \nu_\mu) \\
\tau \sim HN(s_\tau)
\end{aligned}
$$

Berry BUGS Code:

```
model {

        ## Prior
        mu       ~ dnorm(mean_mu, precision_mu)

        tau      ~ dnorm(0, precision_tau)I(0.001, )
        tau_prec <- pow(tau, -2)

        for (j in 1:J) {

          ## Prior
          theta[j] ~ dnorm(mu, tau_prec)

          ## Likelihood
          logit(p[j]) <- theta[j] + logit(p_t[j])
          r[j] ~ dbin(p[j], n[j])

        }

      }
```

10.5 EXNEX Method (Exchangeable/Non-Exchangeable)

Neuenschwander et al. (2016) proposed a different approach for dealing with non-exchangeable cohorts. Problem: The response rates of the cohorts might not be exchangeable, for instance,

when there are outlying cohorts. Their proposed method aims to make the simple Bayesian hierarchical model more robust against such scenarios. They model the log-odds of the response rates with a normal mixture distribution, where one component allows borrowing (exchangeable), while the other one does not (non-exchangeable). w_j represents the weight of the borrowing component of the mixture distribution of $logit(p_j)$. (Wojciekowski (2018))

$$\begin{aligned} r_j|p_j \sim Bin(p_j, n_j),\ j = 1, ..., J \\ \text{With probability } w_j : logit(p_j) \sim N(\mu, \tau^2) \\ \mu \sim N(m_\mu, \nu_\mu) \\ \tau \sim HN(s_\tau) \\ \text{and with probability } 1 - w_j : logit(p_j) \sim N(\mu_j, \tau_j^2) \end{aligned}$$

10.6 EXNEX Adjusted Method

The ExNex Adjusted method is a proposal that combines the methods by Berry et al. (2013), and Neuenschwander et al. (2016). It incorporates both the adjustment $logit(p_j) - logit(\hat{p_j})$ and the mixture prior for $logit(p_j) \sim w_j N(\mu, \tau^2) + (1 - w_j)N(\mu_j, \tau_j^2)$, with the aim of adequately addressing varying prior assumptions about the response rates and different borrowing structures. (Wojciekowski (2018))

$$\begin{aligned} r_j|p_j \sim Bin(p_j, n_j),\ j = 1, ..., J \\ \text{With probability } w_j : logit(p_j) - logit(\hat{p_j}) \sim N(\mu, \tau^2) \\ \mu \sim N(m_\mu, \nu_\mu) \\ \tau \sim HN(s_\tau) \\ \text{and with probability } 1 - w_j : logit(p_j) \sim N(\mu_j, \tau_j^2) \end{aligned}$$

10.7 Custom Methods

I worked on expanding the R-package bhmbasket (available on CRAN), which already included the methods pooled, stratified, berry, exnex, and exnex adjusted. The goal was to enhance this package to support custom models that are executed via RJAGS. RJAGS is an interface for the statistical programming language R, which enables the calculation of Bayesian models via JAGS (Just Another Gibbs Sampler). With this expansion, the bhmbasket package became more adaptable, as various models could now be adjusted and tested. As an example of Custom Methods, I added the model of Thall et al. (2003) discussed above, as well as a model from Jin, Liu, et al. (2020), which I will introduce next.

10.7.1 PoC by Jin

The proposed Bayesian method by Jin, Liu, et al. (2020) aims to detect Proof of Concept (PoC) in early-phase oncology basket trials. The method leverages a hierarchical Bayesian framework to pool information across cancer indications, grouping them into *sensitive* and *insensitive* subgroups. The key idea is to model the response rate p_i of each indication i and classify indications based on their treatment sensitivity. The hypothesis testing problem can be formulated as:

$$H_0 : p_i \leq q_i^0,\ \forall\ i = 1, ..., N\ vs.\ H_1 : p_i \geq q_i^1,\ \exists\ i \in \{1, ..., N\}$$

with q_i^0 being the historical benchmark response rate for standard of care and q_i^1 the target response rate. So we can formulate the model as:

$$r_i \sim Bin(n_i, p_i)$$
$$\theta_i = log\left(\frac{p_i}{1-p_i}\right) \sim N\left(log\left(\frac{q_i^*}{1-q_i^*}\right) + \delta_i, \tau_1^2\right)$$
$$q_i^* = \frac{q_i^0 + q_i^1}{2},\quad \delta_i = d_i\mu_1 + (1-d_i)\mu_0,\quad \tau_1^2 \sim IG(1,1)$$

$$d_i \sim Ber(\pi),\quad \mu_1 \sim TrN(g_1, \tau_2^2, (0,\infty)) < 0,\quad \mu_0 \sim TrN(g_0, \tau_2^2, (-\infty, 0)) > 0$$
$$\pi \sim Unif(0,1), \tau_2^2 \sim IG(1,1)$$

$$g_1 = 1/N \sum_{i=1}^{N}(log(q_i^1/(1-q_i^1)) - log(q_i^*/(1-q_i^*))),\quad g_0 = 1/N \sum_{i=1}^{N}(log(q_i^0/(1-q_i^0)) - log(q_i^*/(1-q_i^*)))$$

- d_i: Sensitivity indicator ($d_i = 1$) if sensitive, ($d_i = 0$) otherwise.
- μ_1 : Mean response rate for the sensitive subgroup ($\mu_1 > 0$).
- μ_0: Mean response rate for the insensitive subgroup ($\mu_0 < 0$).
- q^*: A midpoint between the historical benchmark response rate (q_0) and target response rate (q_1).
- Mean logit response rate, either μ_1 (sensitive subgroup) or μ_0 (insensitive subgroup). These are modeled hierarchically to allow information sharing across indications.

11 Code Implementation

I extended the R-package bhmbasket such that it is possible to call customized methods and the MEM method. For a performance analysis follow the steps described in the package documentation.

11.1 MEM Call

For using the MEM method in the bhmbasket package it is sufficient to specifiy the method in the performAnalyses function:

```
devtools::load_all("~/bhmbasket_extension/bhmbasket")
```

```
i Loading bhmbasket
```

```
library(doFuture)
```

```
scenario_list <- simulateScenarios(
n_subjects_list = list(c(10, 20, 30)),
response_rates_list = list(c(0.2, 0.4, 0.5)),
n_trials = 3)
```

```
performAnalyses(
scenario_list,
evidence_levels = c(0.025, 0.05),
method_names = c("MEM"),
target_rates = rep(0.35, 3))
```

11.2 Custom Method (Code Example)

For a custom method one need to first define the method and save the RJAGS code in a txt-file. In the example of the PoC Method from Jin, Liu, et al. (2020). The txt file contains the following code:

```
model
{
  for (i in 1:J)
  {
    r[i] ~ dbin(p[i],n[i])
    p[i] <- exp(theta[i])/(1+exp(theta[i]))
    theta[i] <- q[i] + delta[i] + e[i]

    e[i] ~ dnorm(0,tausq[i])
    delta_branch[i,1] <- mu1;
    delta_branch[i,2] <- mu2;
    if_branch[i] <- 1+ step(-d[i] + 0.5) # 1 for d=1, 2 for d=0
```

```
    delta[i] <- delta_branch[i,if_branch[i]]

    d[i] ~ dbern(pi)
    tausq[i] ~ dgamma(1,1)
  }
  pi ~ dunif(0,1)
  mu1 ~ dnorm(g1,tausq2) T(0, )
  mu2 ~ dnorm(g0,tausq2) T(, 0)
  tausq2 ~ dgamma(0.01,0.01)
}
```

and in the R code we defined the parameters needed:

```
q0 = 0.15 # historical benchmark response rate
q1 = 0.35 # target response rate
custom_methods <- list(
  PoC = list(
    j_data = list(
      q             = rep(log(((
        q0 + q1
      ) / 2) / (1 - (
        q0 + q1
      ) / 2)), 3),
      p_t           = rep(0.5, 3),
      g1            = log(q1 / (1 - q1)) - log((q1 + q0) / 2 / (1 - (q1 +
                                                                    q0) / 2)),
      g0            = -log((q1 + q0) / 2 / (1 - (q1 + q0) / 2)) + log(q0 /
                                                                        (1 - q0))
    ),
    j_model_file = "~/bhmbasket_extension/bhmbasket/inst/jags_models/PoC.txt",
    j_parameters = c("pi", "d", "p", "delta", "tausq", "mu1", "mu2")
  ),

  BHM = list(
    j_data = list(
      mean_mu       = -1.1734601,
      precision_mu  = (1 / 100),
      precision_tau = 1 / (25 ^ 2)
    ),
    j_model_file = "~/bhmbasket_extension/bhmbasket/inst/jags_models/BHM.txt",
    j_parameters = c("p", "mu", "sigma")
```

```
  )
)
```

and in the performAnalyses function you call the custom functions by:

```
performAnalyses(
      scenario_list = scenarios_list,
      custom_methods = custom_methods,
      method_names = c("custom"),
      target_rates = rep(0.35, 3),
      evidence_levels = c(0.05, 0.1)
    )
```

11.3 Availability of the Extension

I will make the extension available on bitbucket.

12 Scenario Analysis

Finally, I can examine the new implemented methods in the package bhmbasket. Therefore, I constructed scenario analysis. I orientated myself in the way of defining those scenarios, constructing the plots and evaluating the plots at Wojciekowski (2018). These analyses have the goal to find out how the MEM model performs in comparison to the other ones. In these simulations, the operating characteristics, such as true and false discovery rates, and predictive probability of 'go' given preliminary observations were evaluated for several scenarios. These evaluations were conducted for basket trials with different design options, as including the number of cohorts, number of subjects within cohorts, effect sizes across cohorts, staggered approaches. I only used 1000 number of runs due to the long running time of MEM (even if I improved the performance). Therefore small variations of usally constant outcomes are caused by the low number of runs. Since in all analyses, except the ones we adapted especially on it, we used as target rates 0.35 for all cohorts. Therefore the method Exnex and Exnex Adjusted won't differentiate in those cases. (Since we substrate 0.35 for all cohorts, this is not changing the outcome).

k will be the number of go cohorts we need to achieve a overall go.

12.1 Changing Number of Cohorts with Positive Response

12.1.1 True Go

?@fig-TG-ATC illustrates the variations in the true Go rates of various methods due to the rise in the number of active treatment cohorts. These variations are observed in four different scenarios: nugget (1/4), half-half (2/4), inverse nugget (3/4), and all positive (4/4). These scenarios are analyzed with a consistent false Go rate of 0.1, using decision rule parameters = 0.7 (evaluation rule) and k = 1 (amount of Go cohorts for a overall Go).

The methods PoC and ExNex and ExNex Adjusted have the highest true Go rates in any scenario. For the nugget as well as for the half-half scenario MEM had the lowest true Go rate (excluding pooled) around 0.12 worse than the PoC method in the nugget case (compare Table 1).

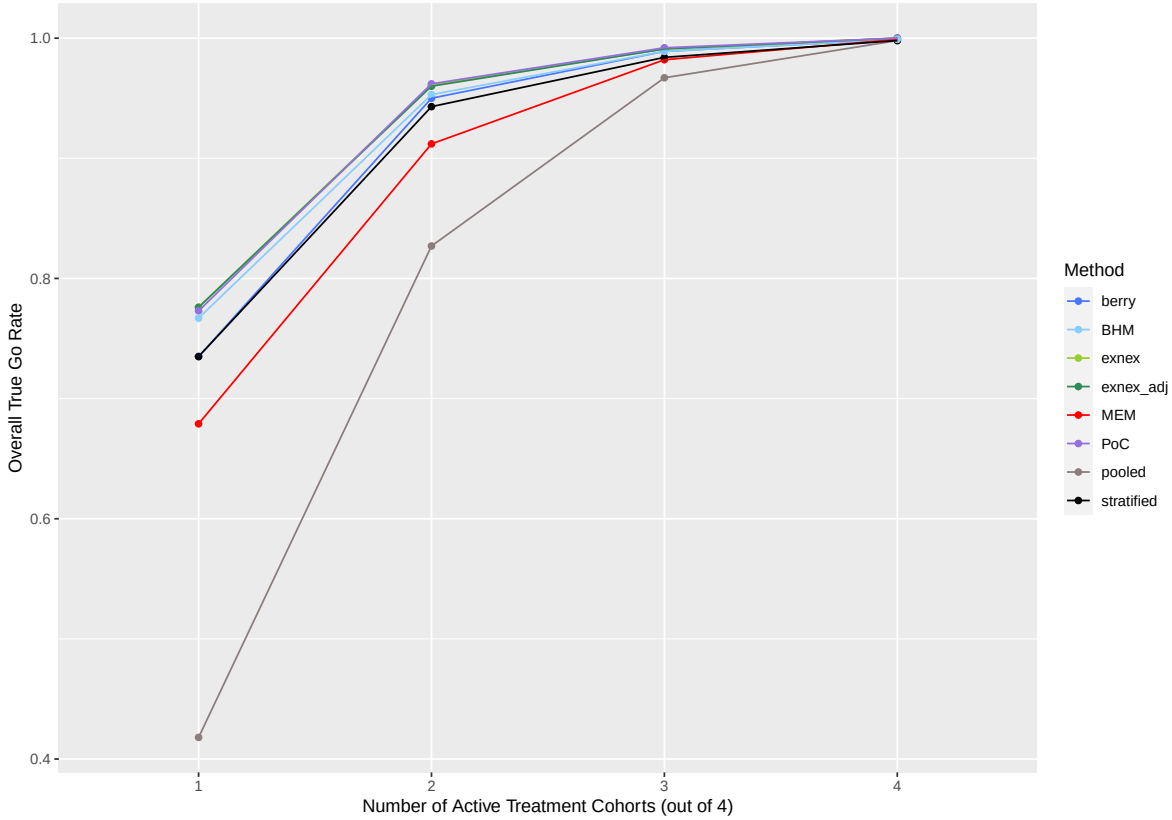


Figure 10: The true go rate when increasing number of cohorts with positive responses, false go rates for the best negative case= $\alpha = 0.1$, k = 1, γ= 0.7, total number of cohorts = 4, p+ = 0.35, p-= 0.15, runs =1000.

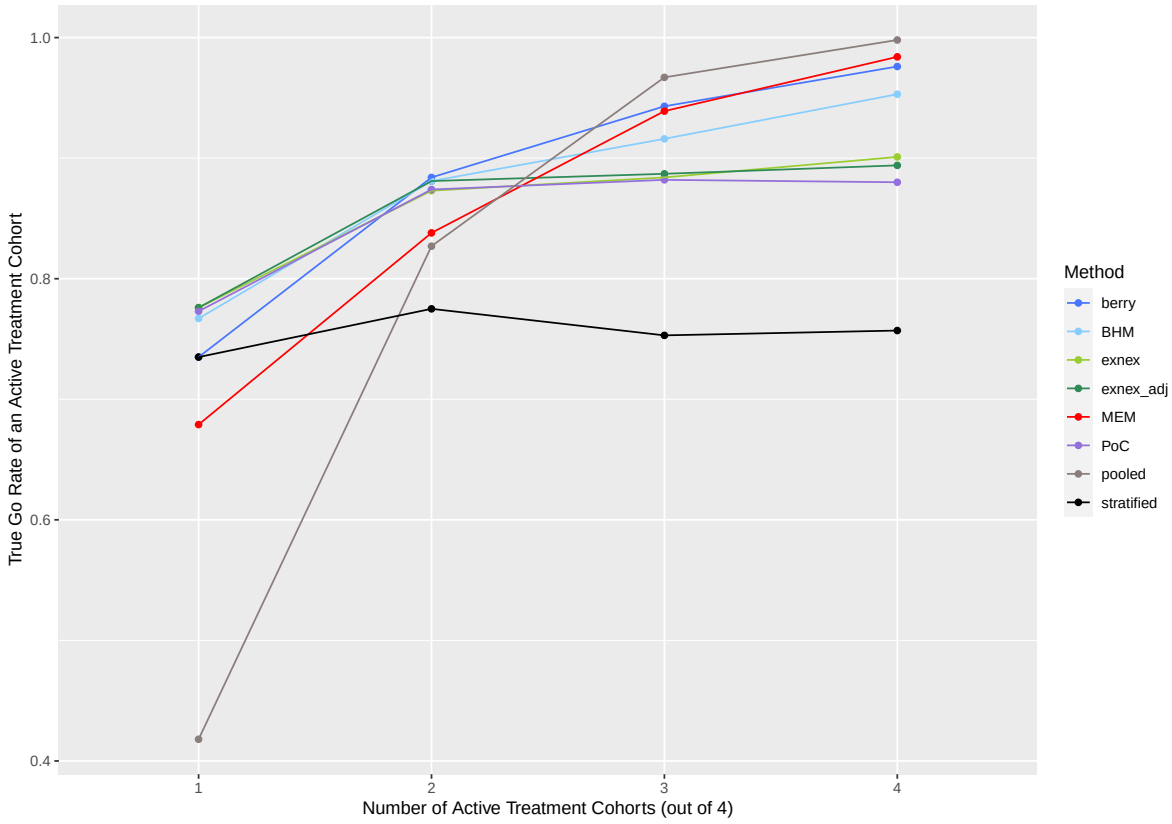


Figure 11: The true go rate when increasing number of cohorts with positive responses, false go rates for the best negative case= $\alpha = 0.1$, k = 1, γ= 0.7, total number of cohorts = 4, p+ = 0.35, p-= 0.15, runs =1000.

12.1.2 MSE

The number of active treatment groups directly impacts the borrowing capacity of hierarchical analysis techniques across these groups. Consequently, it's anticipated that the Mean Squared Error (MSE) for both the hierarchical analysis techniques and the Pooling method will decline as the count of active treatment cohorts rises.

Figure 12 illustrates the MSEs of the posterior response rates' mean of an active treatment cohort in the nugget (1), half-half (2), inverse nugget (3), and all positive cases (4).

The diagram indicates that the MSEs for the Stratification method remain constant, regardless of the number of active treatment cohorts. In the nugget scenario, the Pooling method's MSE is the highest, while the PoC method's MSE is the lowest. The method MEM to be compared

Table 1: The true go rate when increasing number of cohorts with positive responses, false go rates for the best negative case= $\alpha = 0.1$, k = 1, γ= 0.7, total number of cohorts = 4, p+ = 0.35, p-= 0.15, runs =1000.

	Method	0	1	2	3	4
1:	BHM	0	0.767	0.953	0.989	0.999
2:	MEM	0	0.679	0.912	0.982	0.999
3:	PoC	0	0.773	0.962	0.992	1.000
4:	berry	0	0.735	0.950	0.989	0.999
5:	exnex	0	0.776	0.961	0.991	1.000
6:	exnex_adj	0	0.776	0.960	0.991	1.000
7:	pooled	0	0.418	0.827	0.967	0.998
8:	stratified	0	0.735	0.943	0.984	0.998

has after the Pooling the highst MSE in the Nugget Case (which is 0.008 higher, than the MSE of the PoC Method).

In the half-half scenario, all methods except for Stratification show a decrease in MSEs. The method by Berry et al. and PoC have the lowest MSE, and the MSEs of the ExNex and ExNex Adjusted methods are lower than the MSE of the Stratification method. Still pooling has the highest MSE followed by MEM, but both MSE decreased dramatically.

In the inverse nugget scenario, all methods show a further decrease in MSEs. The Stratification method and the PoC have the highest and second highest MSEs, respectively. In contrast to the half-half scenario the Pooling performs the best followed by MEM.

In the all positive scenario, all methods except for Stratification and PoC show a further decrease in MSEs, maintaining the same order as in the inverse nugget scenario.

The figure indicates that the more active treatment cohorts there are, the lower the MSEs of all methods except for Stratification and PoC. The Pooling method's MSE decreases the fastest, while the method MEM has the second fastest rate of decrease from the nugget to the half-half Case.

The fundamental concept of the PoC is to categorize the cohorts into two distinct groups: those that are sensitive to the medication and those that are not. Consequently, a greater number of active treatment cohorts are also classified as sensitive. However, this classification does not affect the Mean Squared Error because these cohorts were sampled from the same underlying response rate.

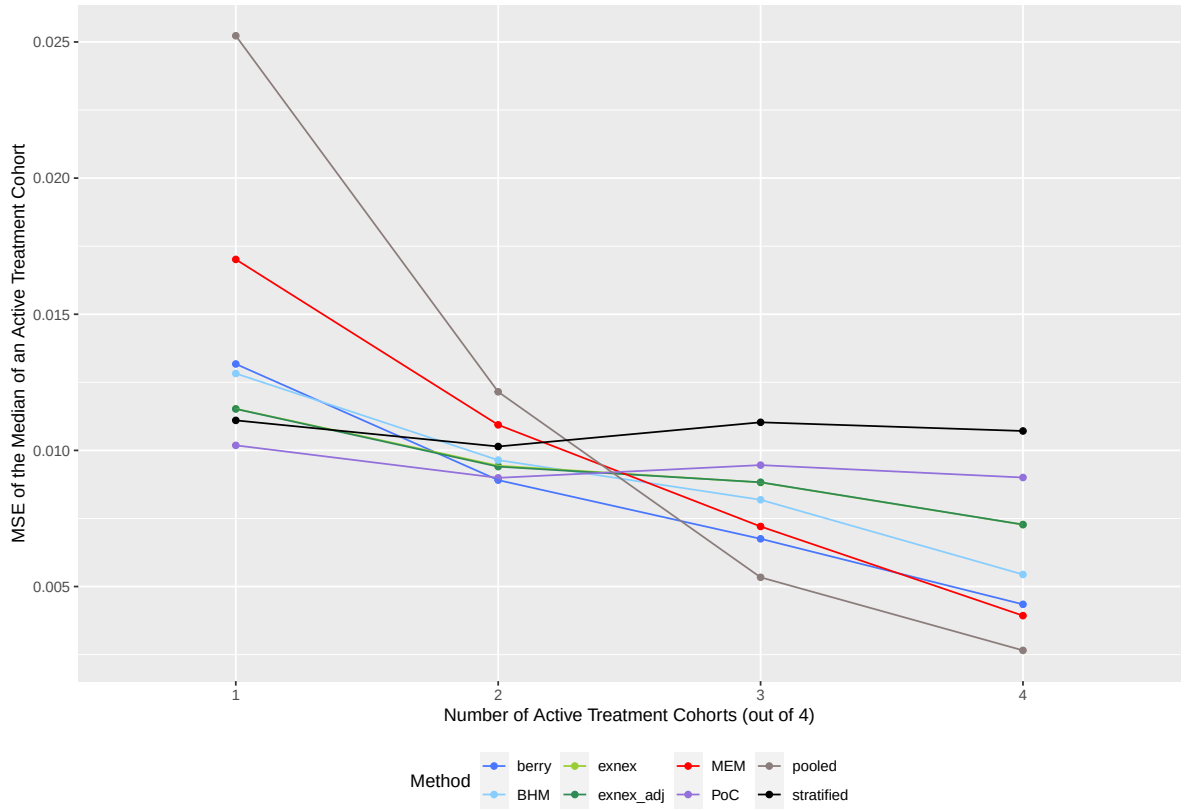


Figure 12: MSE of the posterior distributions' means of the response rates of an active treatment cohort when increasing number of cohorts with positive responses, false go rates for the best negative case= $\alpha = 0.1$, k = 1, γ= 0.7, total number of cohorts = 4, $p_+ = 0.35$, $p\_$= 0.15, runs =1000.

In Figure 13, we observe the Mean Squared Errors (MSE) of the methods increasing the number of active treatment cohorts. MEM has low MSEs in the all-negative ("No-go") and all-positive scenarios, indicating good accuracy there. However, MSE is significantly higher in the outlier groups of the nugget and inverse nugget cases, as well as across all groups in the half-half case. The PoC and Berry methods generally perform well with lower MSE in the nugget, half-half, and inverse nugget cases, showing more stable results in these varied scenarios.

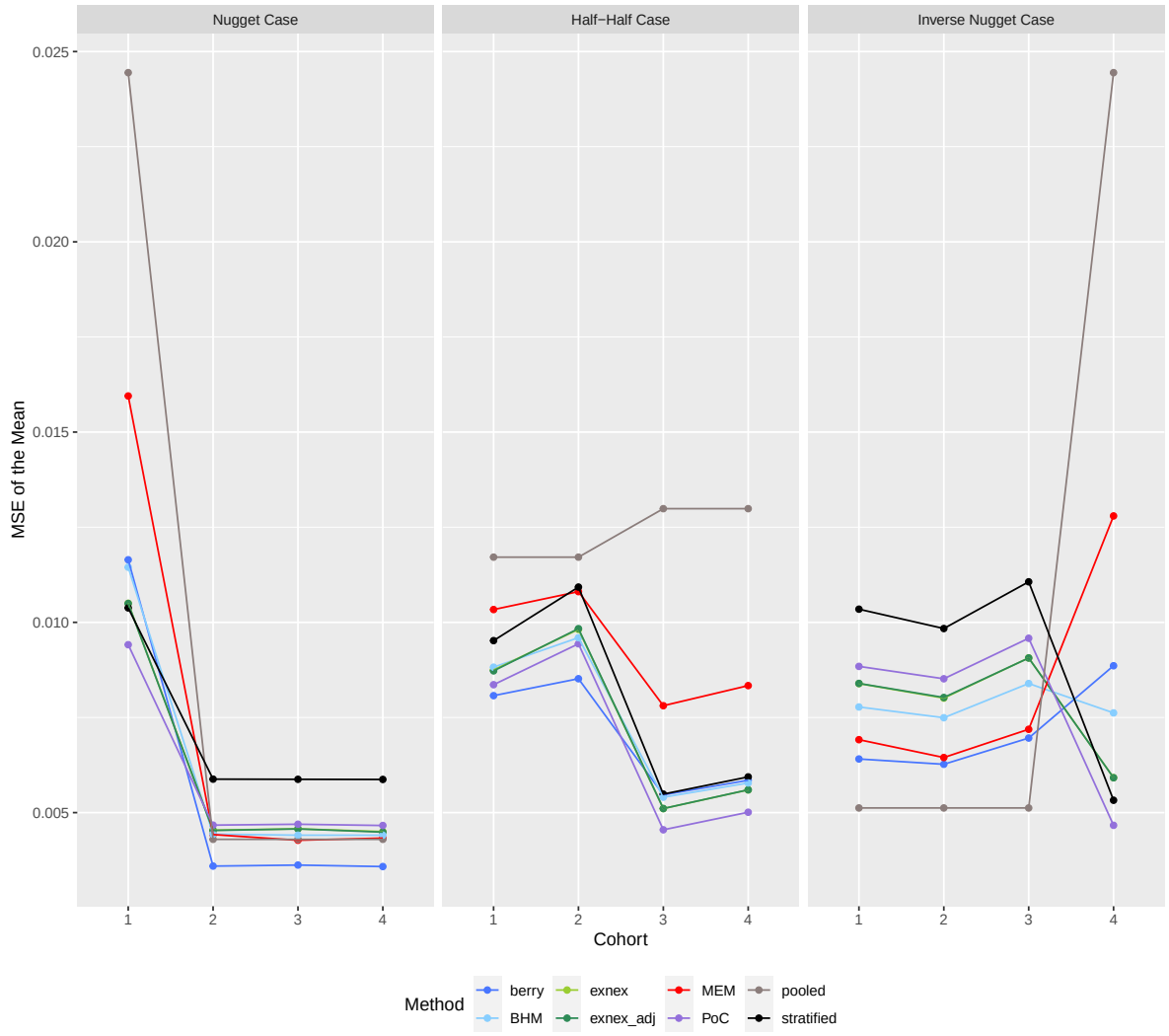


Figure 13: MSE of the posterior distributions' mean of the response rate when increasing number of cohorts with positive responses, false go rates for the best negative case= $\alpha = 0.1$, k = 1, γ= 0.7, total number of cohorts = 4, $p_+ = 0.35$, $p\_$= 0.15, runs =1000..

12.1.3 Mean

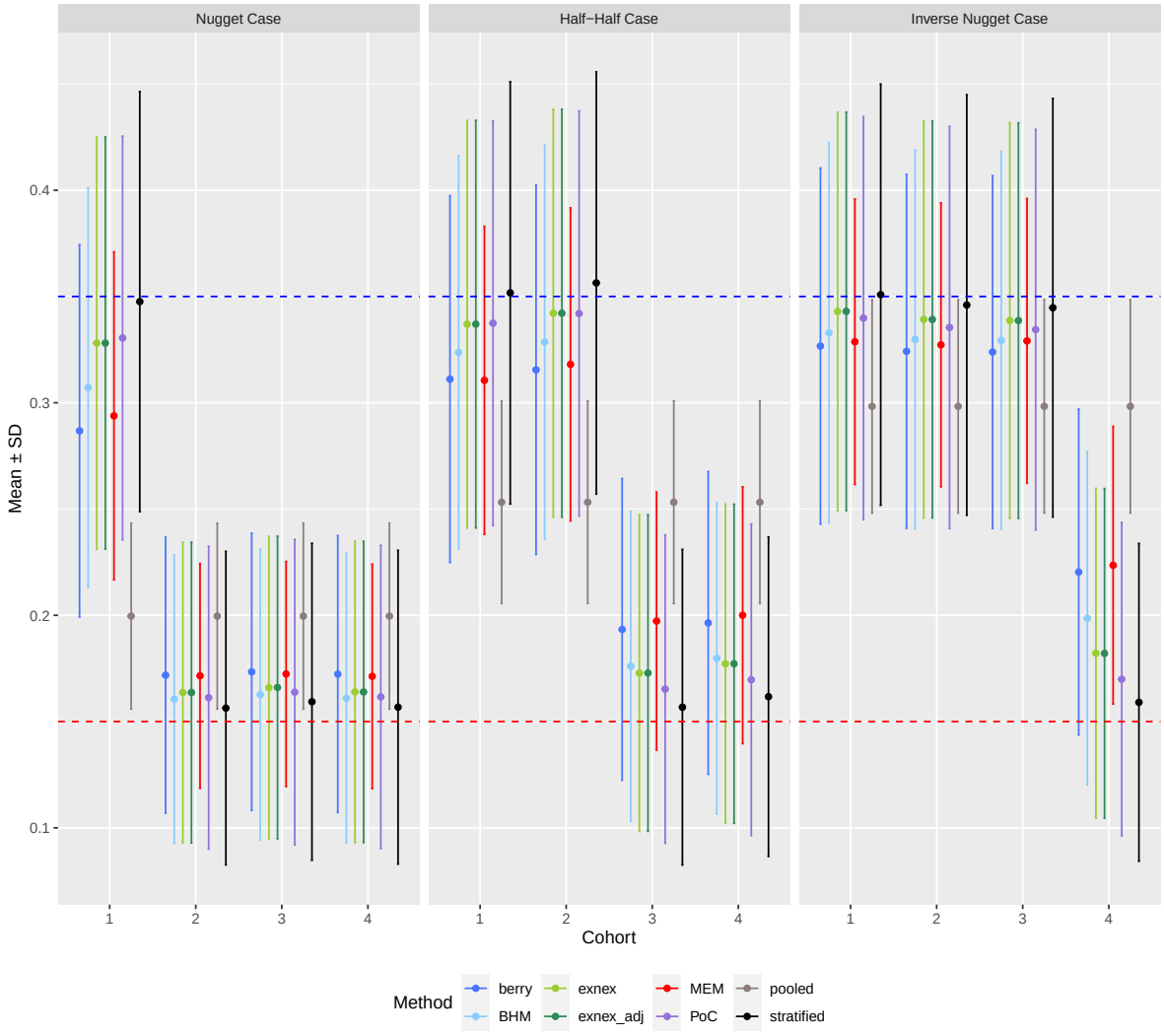


Figure 14: Mean and standard Deviation of the posterior distribution of the response rate of an active treatment cohort when increasing number of cohorts with positive responses, false go rates for the best negative case= $\alpha = 0.1$, k = 1, γ= 0.7, total number of cohorts = 4, $p_+ = 0.35$, $p\_$= 0.15, runs =1000.

12.1.4 Median

The Method MEM shows a relatively narrow credible interval compared to other methods, as seen in Figure 15 and in the more clear Figure 41 where only the median and the 95% - credible

intervals of one active treatment cohort are visualized. Its credible interval is almost as small as that of the Pooled method, which has the smallest intervals across all five scenarios.

Three scenarios stand out: the nugget, half-half, and inverse nugget cases.

In the nugget and half-half cases, all methods' credible intervals cover the true response rate, except for the Pooled method. MEM and Berry have very small median values for the active treatment cohort, but these values are influenced by the other three non-active cohorts. (Here, we exclude the Pooled method from consideration.) The Stratified method's median is close to the true response rate in both cases, but it has a wide credible interval. The ExNex and ExNex_adj methods have medians closer to the true response rate than Stratified, with slightly narrower credible intervals.

In the inverse nugget case, we observe a similar effect, but here the negative control cohort's posterior distribution median is influenced by the three active treatment cohorts.

Overall, MEM has much narrower credible interval compared to other methods. However, there is still a to high borrowing effect between the active and non-active treatment cohorts. This effect could potentially be reduced by refining the MEM method with an informative prior matrix. Which I tried in the section "Changing the prior matrix for MEM matrix".

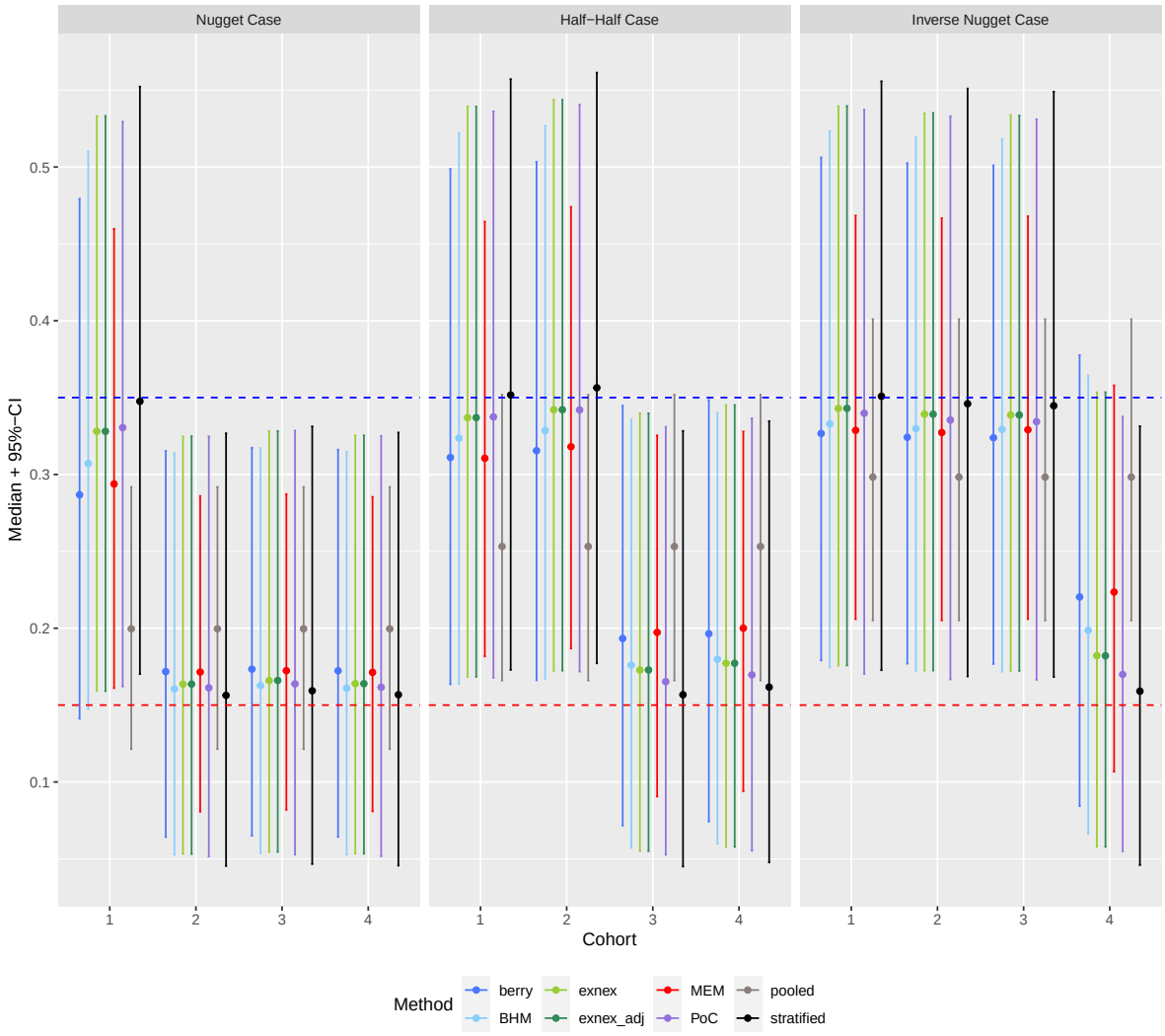


Figure 15: Median with 95% credible interval of the posterior distribution of the response rate when increasing number of cohorts with positive responses, false go rates for the best negative case= $\alpha = 0.1$, k = 1, γ= 0.7, total number of cohorts = 4, $p_+ = 0.35$, $p\_$= 0.15, runs =1000.

Since there is no big different in examining the mean with its standard deviation or the median with a 95% credible intervall, I will only use the median in the further analyses.

12.1.5 %Go

As illustrated in Figure 16, MEM's probability of recommending a "Go" decision (Go probability) is relatively high in cases where the true scenario is "No-go," especially in heterogeneous cases. In the nugget case, MEM's Go probability for the "Go" cohort is also lower than in other methods, suggesting excessive influence or "sharing" across groups. In these heterogeneous cases, PoC and ExNex (or its adjusted form, ExNex_adjusted) show good balance in true Go and true No-go percentages.

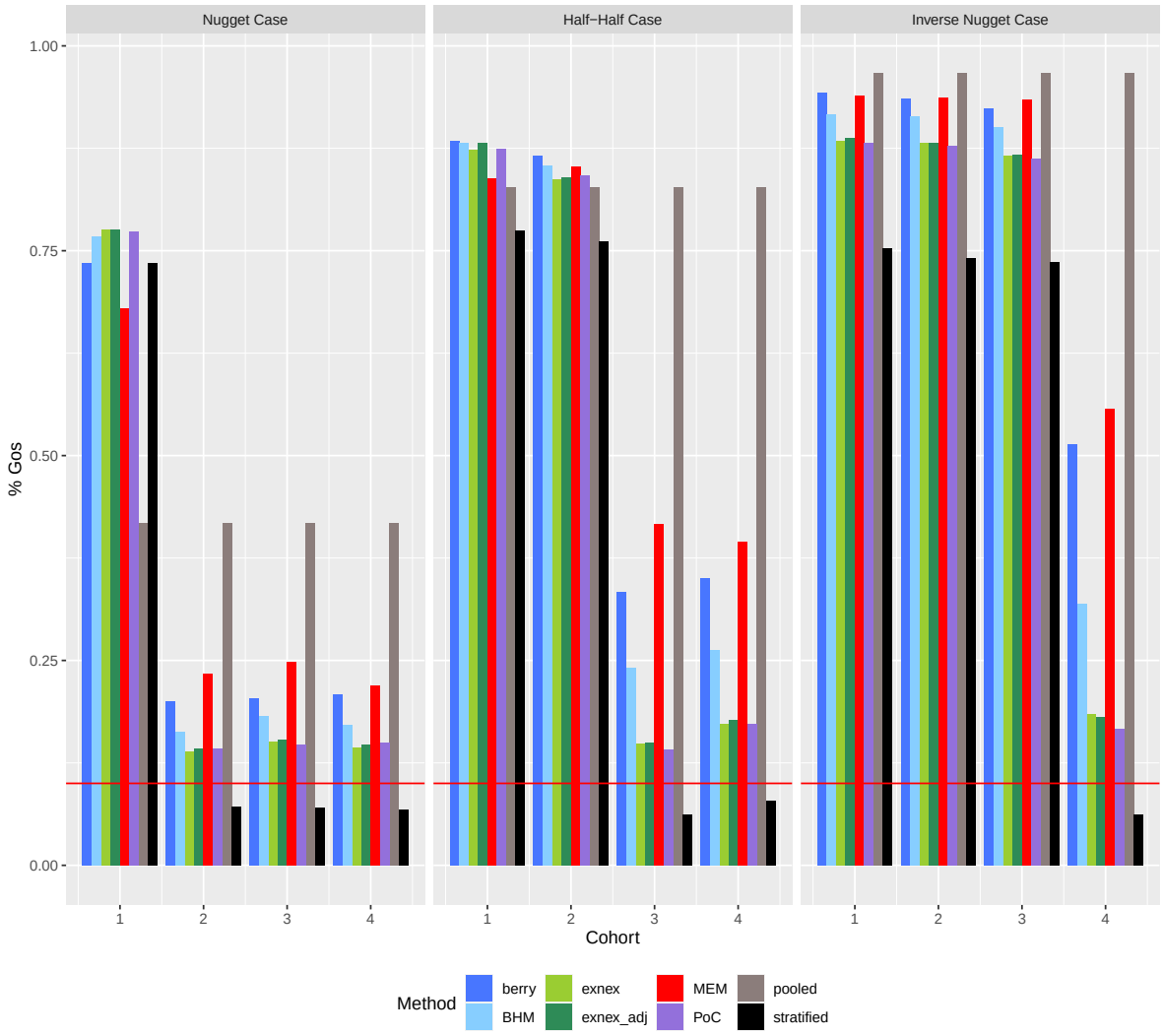


Figure 16: Probability of a Go decision when increasing number of cohorts with positive responses, false go rates for the best negative case= $\alpha = 0.1$, k = 1, γ= 0.7, total number of cohorts = 4, $p_+ = 0.35$, p_-= 0.15, runs =1000.

12.1.6 P-Boundary

Figure 17 examines the probability boundaries when varying the evaluation rate γ. For a gamma value of 0.2, it is quite noticeable that methods which borrow a significant amount of information between cohorts, such as pooled and MEM, tend to have the smallest p-boundary. On the other hand, methods with less information borrowing have a higher p-boundary, with stratified (which does not borrow any information) having the highest p-boundary.

This characteristic holds true for all gamma values, however, the difference between the p-boundaries decreases with higher gamma values. This makes sense, as for methods like pooled and MEM, we have a smaller variance and non-active treatment cohorts also tend to have less frequent higher values. This results in a lower decision boundary as it was calibrated.

The differences between the decision boundaries become smaller with higher gamma values, as the probability of having multiple outliers is simply low.

For all other analyses we chose a gamma of 0.7 and an alpha of 0.1 where the decision threshold for Go or No-go lies between approximately 0.15 and 0.175.

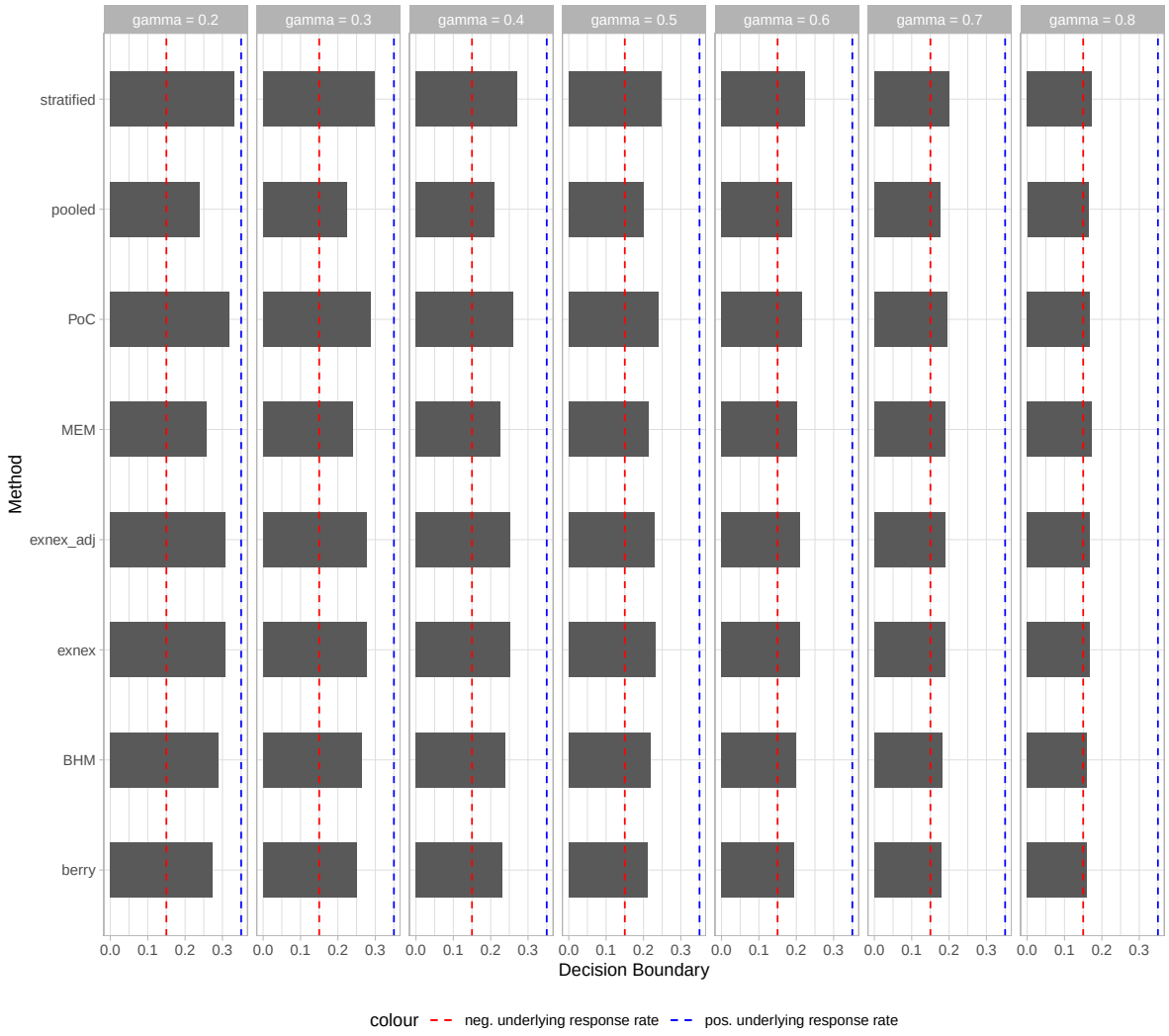


Figure 17: Threshold for go or nogo decisions, fixed for each method seperatly, for a fixed evaluation rate $\gamma = 0.7$ adjusting such that the falsego rate for the best negative case (here the all negative case) being less $\alpha = 0.1$. k = 1, total number of cohorts = 4, p_+ = 0.35, p_-= 0.15, runs =1000.

Figure 18 shows the overall go probabilities. We assumed the same definition as in Wojciekowski (2018). A overall go is a overall true go, if at least for k go cohorts they are also predicted as go, otherwise it is a overall false go. (Example: For k=1 and 3 cohorts, with the true decisions: go, no go and no go but predicted as: no go, go and no go, respectively, this would be a overall false go.) So there is the possibility of a predicted overall go being an overall true go, an overall false go or an overall false no go.

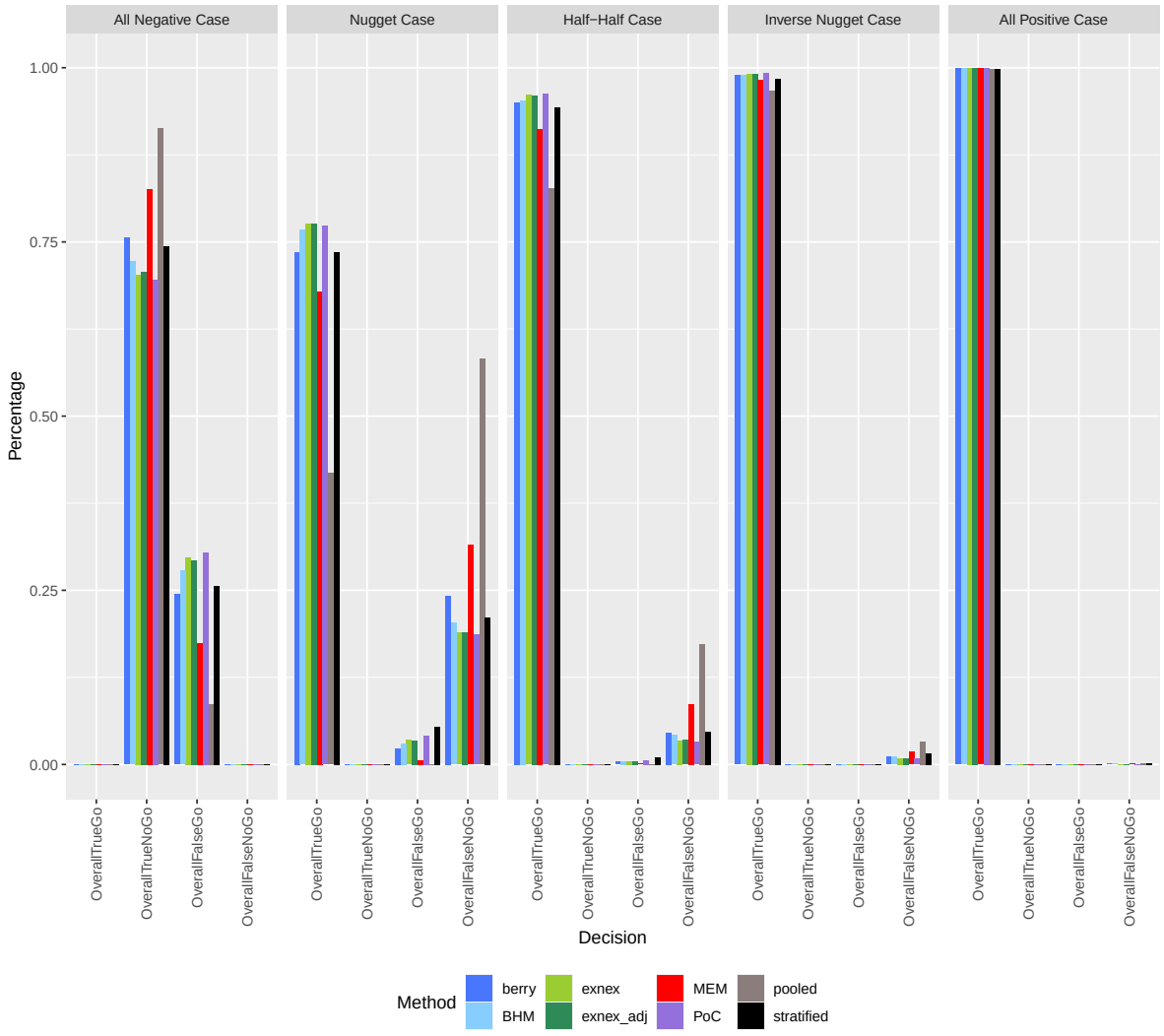


Figure 18: Percentage of overall true go, false go, true no go, false no go when increasing number of cohorts with positive responses, false go rates for the best negative case= $\alpha = 0.1$, k = 1, γ= 0.7, total number of cohorts = 4, $p_+ = 0.35$, $p\_$= 0.15, runs =1000.

12.2 Changing the True Underlying Positive Response Rates

Figure 19 illustrates how each method's true Go rate changes with different positive response rates, keeping the negative response rate fixed at 0.15 and maintaining a false Go rate of 0.1. The decision rule is parameterized by $\gamma = 0.7$ and $k = 1$. As expected, the overall true Go rate increases as the underlying positive response rate rises. However, MEM consistently shows the

lowest true Go rates at each level of the positive response rate (except for the Pooled method), although MEM's performance does approach the other methods' rates as the positive response rate increases. The PoC method achieves the best true Go rates across all positive response levels (staying almost constant), indicating its strong performance for Go decision accuracy.

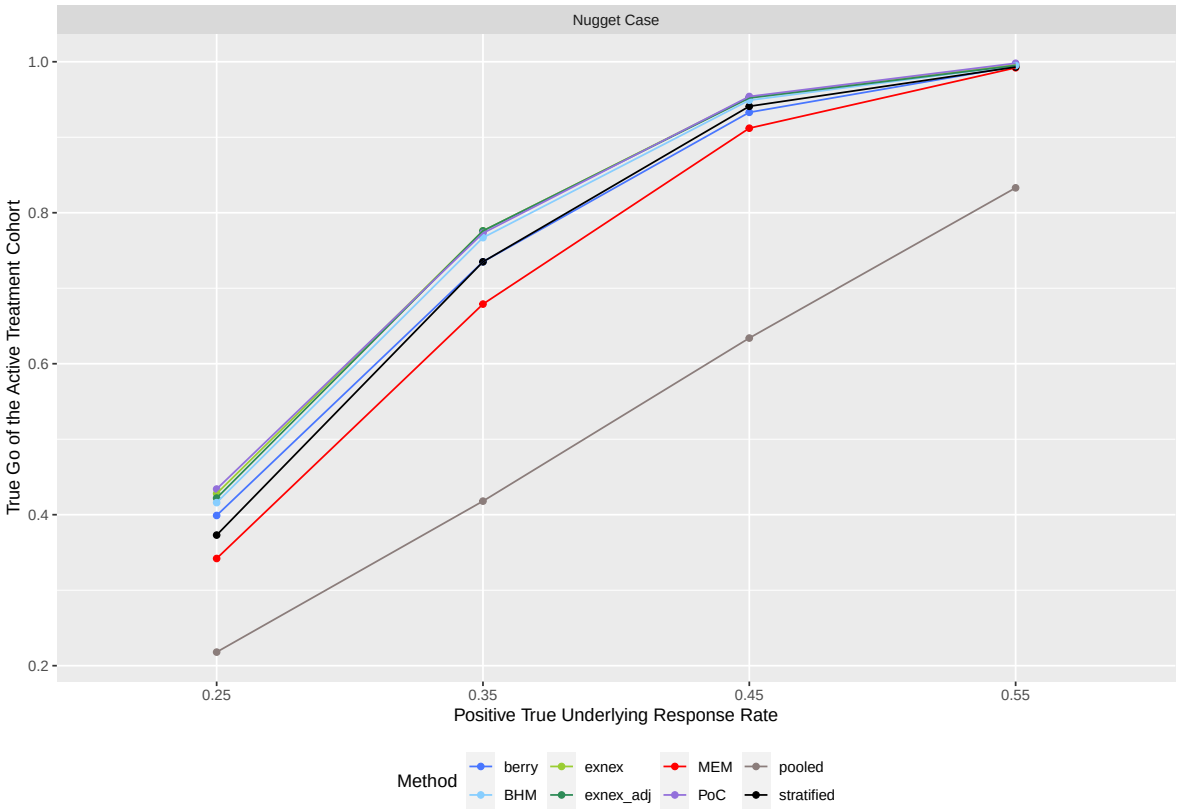


Figure 19: True overall go rate when increasing number of cohorts with positive responses, false go rates for the best negative case= $\alpha = 0.1$, k = 1, γ= 0.7, total number of cohorts = 4, $p_+ = 0.25/0.35/0.45/0.55$, $p\_$= 0.15, target rate = 0.35, runs =1000.

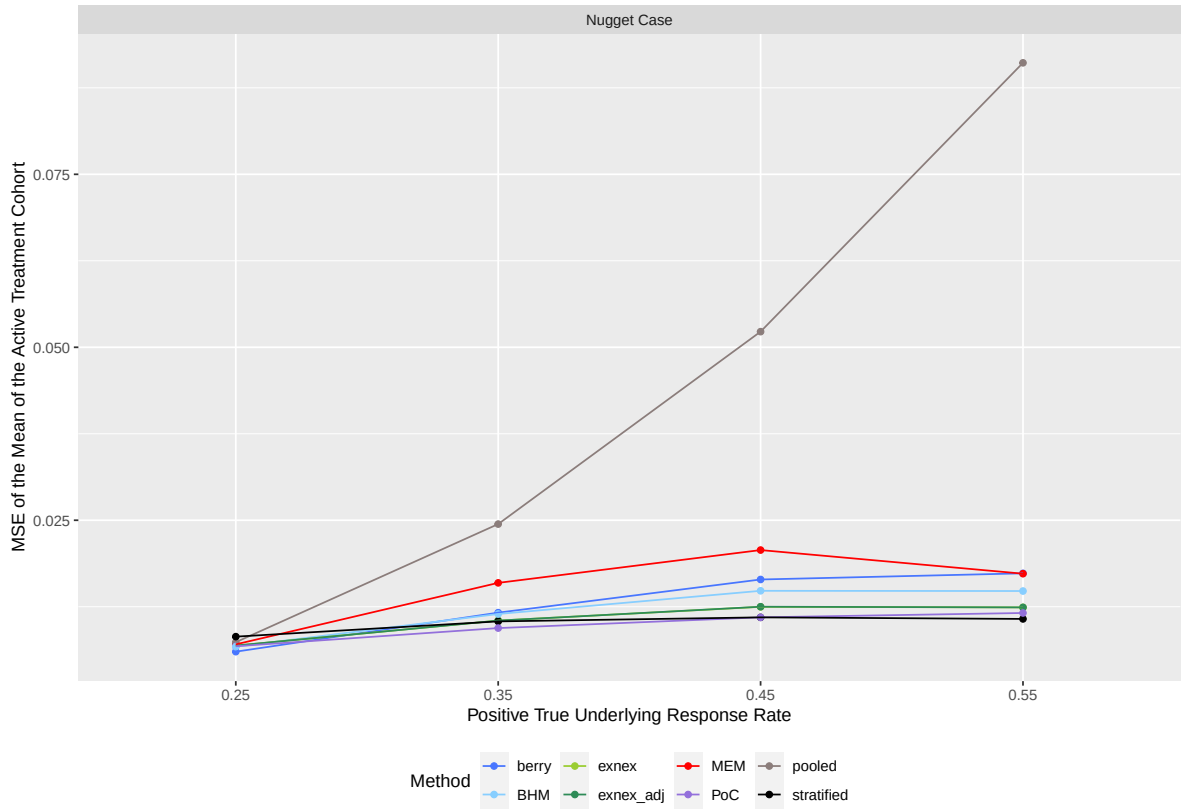


Figure 20: True overall go rate when increasing number of cohorts with positive responses, false go rates for the best negative case= $\alpha = 0.1$, k = 1, γ= 0.7, total number of cohorts = 4, $p_+ = 0.25/0.35/0.45/0.55$, $p\_$= 0.15, target rate = 0.35, runs =1000.

12.3 Adjusting Assumed Positive Response Rates

In this section, we adjusted the target rate in the scenario to match the true underlying response rate (adapted case). In the unadapted case, we maintained the target rate at the 0.35 level for every cohort, regardless of whether it was an active or non-active treatment cohort. Therefore, we examined the cases nugget, half-half, and inverse nugget. Consequently, the four cohorts of the nugget case are analyzed with the true underlying response rates and target rates of 0.35, 0.15, 0.15, and 0.15, the cohorts of half-half with 0.35, 0.35, 0.15, and 0.15, and the cohorts of inverse nugget with 0.15, 0.35, 0.35, and 0.35.

The cohorts of the No-go case are analyzed with the respective target rates of the Go cases nugget, half-half, and inverse nugget to adjust the boundary level. It is expected that the actual Go rates of the method by Berry et al. (2013) and the method ExNex Adjusted, as well as for the MEM method, will improve, as they adjust for the target rate.

The actual Go rates of the other methods are not expected to change, because their prior values are, although dependent on the target rates, derived to result in vague prior distributions. The changes induced by the correct target rates in these cases are depicted in Figure 21. For almost all methods there are almost constant. For MEM it is even decreasing in the nugget and half-half case which is odd.

Figure 22 displays the methods' MSE of the mean in the nugget, half-half, and inverse nugget cases for equal and correct target rates across cohorts. The MSE decreases when the target rate equals the true underlying response rate in the nugget and half-half cases, but decreases the most for the method by Berry et al. (2013) and the method ExNex Adjusted. As a result, the MSE of the method by Berry et al. (2013) with the correct target rates is the lowest MSE in the nugget and half-half cases, and, along with the MSE of the method Pooling, also the lowest in the inverse nugget case. The MSE of MEM is the highest regardless of the adaptation in the nugget case, regardless of whether it's pooled. The adaptation does not have a significant influence on the MSE in all cases.

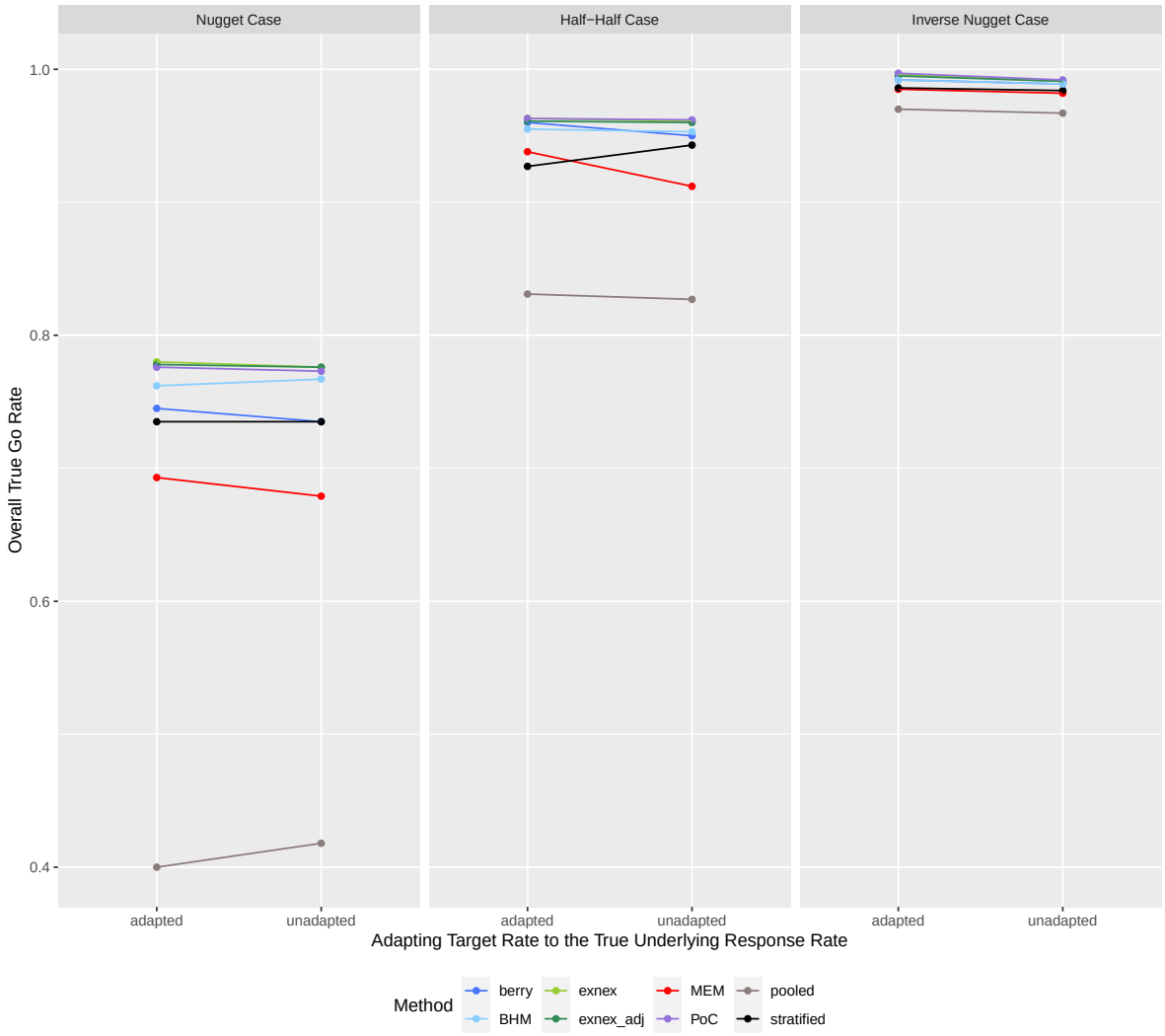


Figure 21: Overall truego rate when changing the assumed positive response rate (target rate), false go rates for the best negative case= $\alpha = 0.1$, k = 1, γ= 0.7, total number of cohorts = 4, $p_+ = 0.25/0.35/0.45/0.55$, $p\_$= 0.15, target rate = 0.35 in the undapted case, target_rate = underlying response rate in the adapted case, runs =1000.

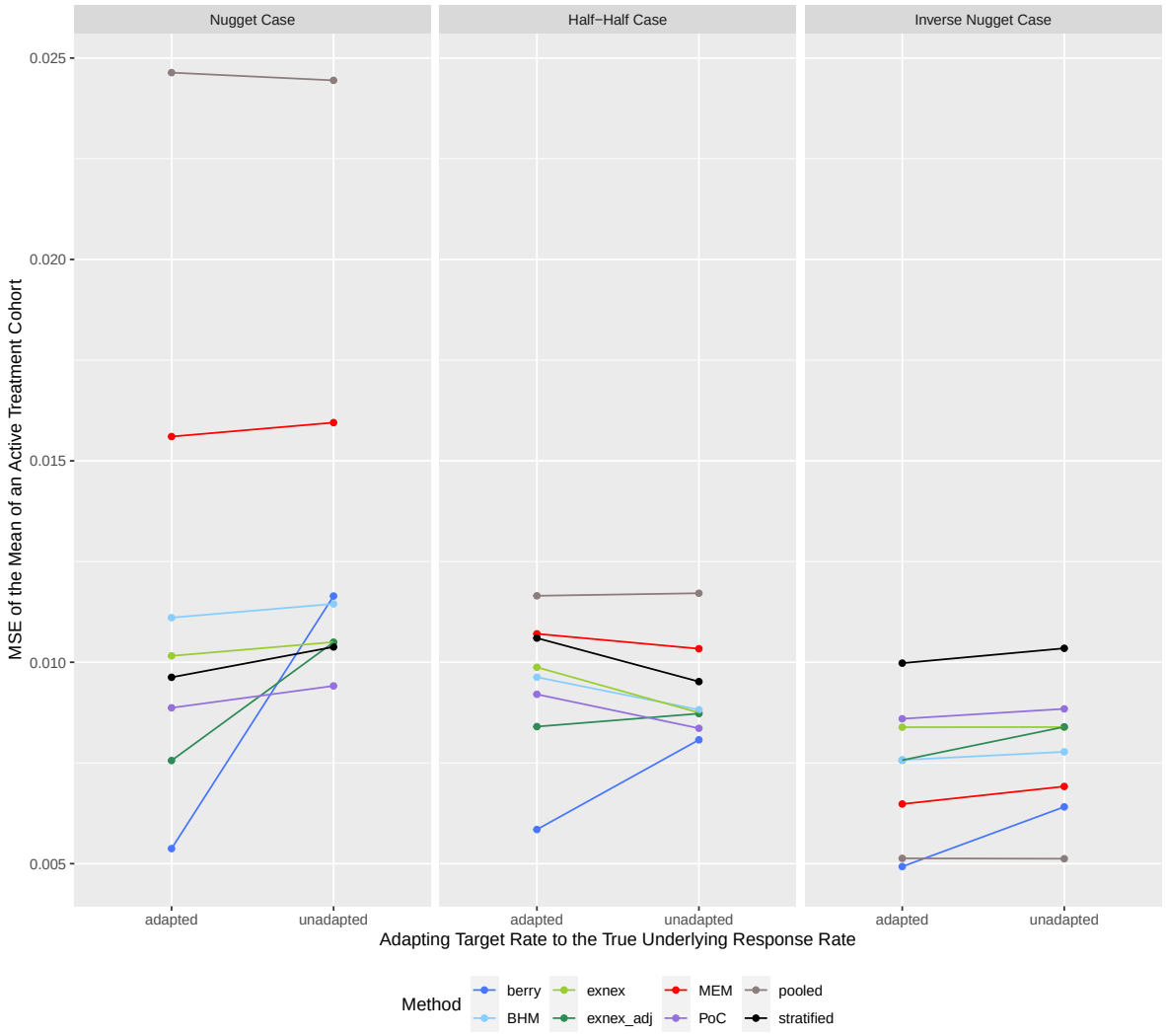


Figure 22: MSE of the posterior distributions' mean of the response rate of an active treatment cohort when changing the assumed positive response rate (target rate), falsego rate for the best negative case= 0.1, k = 1, γ= 0.7, Total number of cohorts = 4, p_+ = 0.35, $p\_$= 0.15, target rate = 0.35 in the undapted case, target rate = underlying response rate in the adapted case, runs =1000.

12.4 Number of Cohorts

Figure 23 shows the false No-Go rates with a fixed negative response rate of 0.15 and a positive rate of 0.35, as the number of cohorts increases. The analysis uses a constant false Go rate of 0.1, with decision rule parameters $\gamma = 0.7$ and $k = 1$, in a nugget case where there is one

active treatment cohort. As expected, the overall false No-Go rate rises as the number of negative-response cohorts increases. For MEM, this rate is particularly high across number of cohorts, only slightly better than the Pooled method. Notably, the gap between MEM and other methods widens as the cohort number grows. The PoC method consistently achieves the best results, with almost constant false No-Go rates across all cohort number.

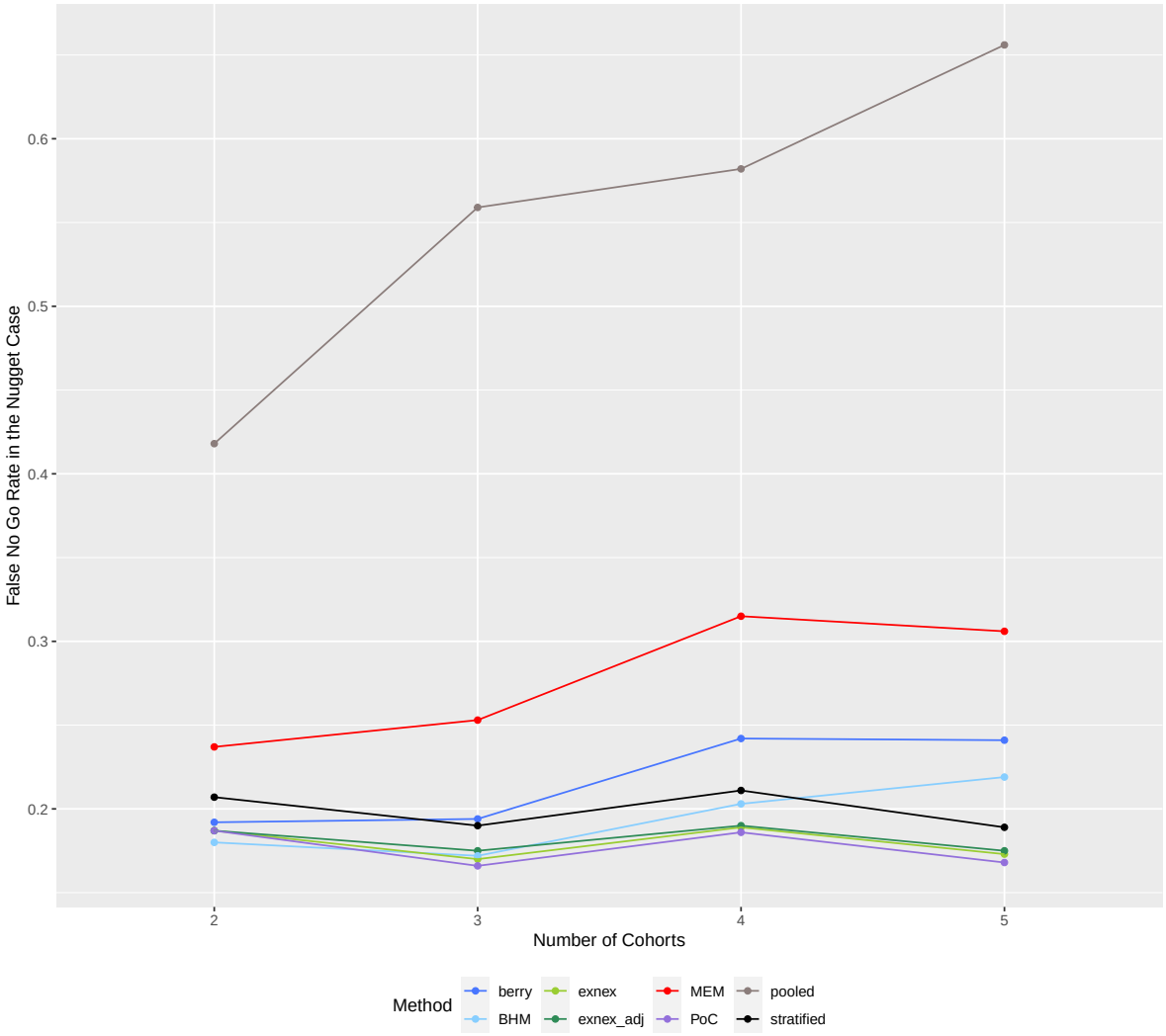


Figure 23: Overall false nogo rate when increasing the number of cohorts. falsego rates for the best negative case= 0.1, k = 1, γ= 0.7, total number of cohorts = 4, p_+ = 0.35, $p\_$= 0.15, target rate = 0.35, cohort size = 20 patients per cohort, runs =1000.

Figure 24 illustrates the true Go rates with the same fixed response rates (0.15 for negative and 0.35 for positive) while increasing the number of cohorts. The scenario again maintains

a constant false Go rate of 0.1 and decision rule parameters $\gamma = 0.7$ and $k = 1$, in the nugget case with a single active treatment cohort. As expected, the overall true Go rate increases as the number of cohorts with a negative response grows. For MEM, the true Go rate remains low compared to other methods across cohort sizes, only performing better than the Pooled method. Furthermore, the difference between MEM and the better-performing methods increases as the cohort size grows. PoC and Exnex maintains the highest true Go rates across all cohort sizes, showing its consistent performance.

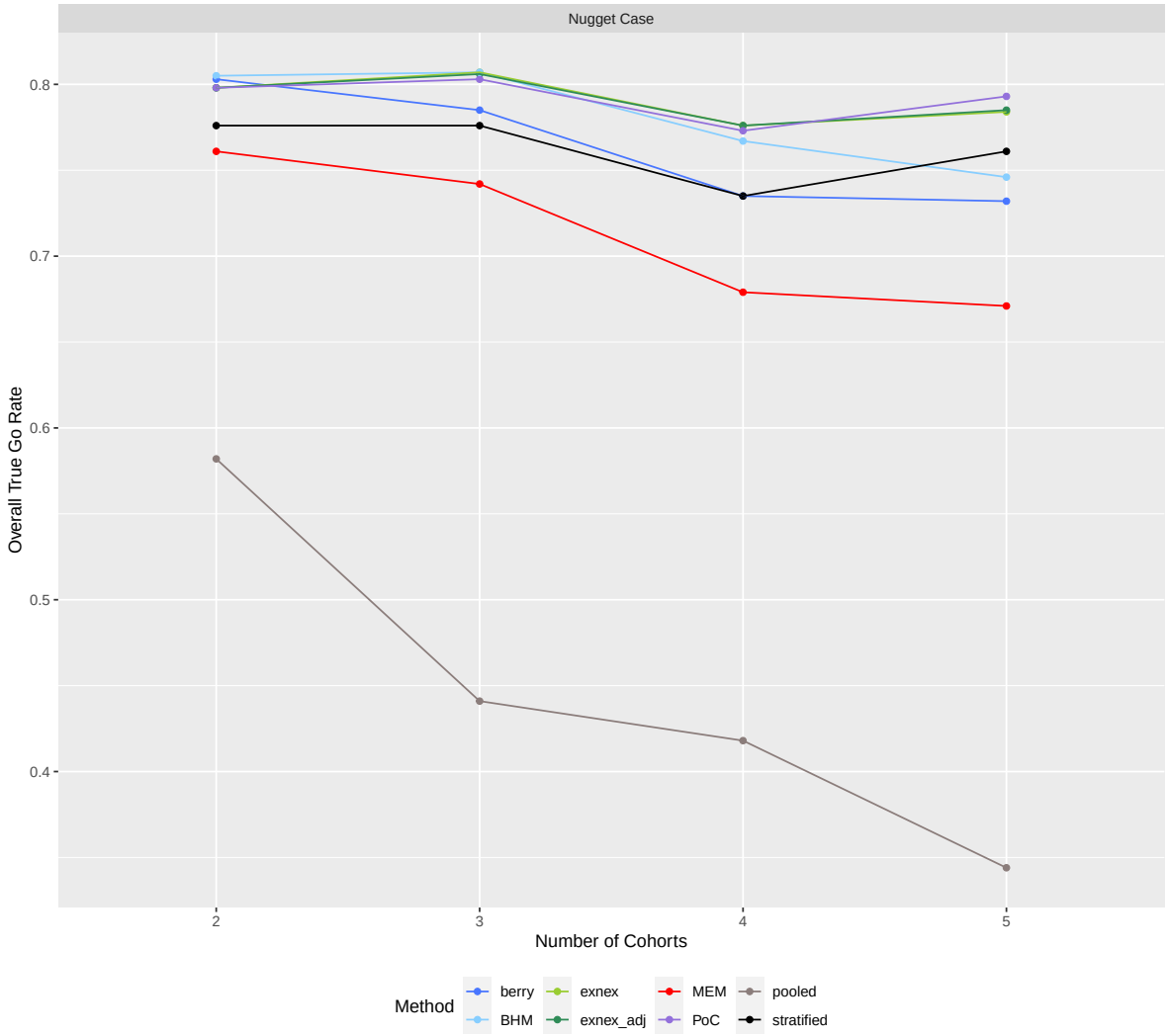


Figure 24: Overall true go rate when increasing the number of cohorts. falsego rates for the best negative case= 0.1, k = 1, γ= 0.7, total number of cohorts = 4, p_+ = 0.35, $p\_$= 0.15, target rate = 0.35, cohort size = 20 patients per cohort, runs =1000.

Figure 25 also provides insights into the Mean Squared Error (MSE) trends in this scenario with a fixed negative response rate of 0.15 and a positive rate of 0.35 as cohort numbers increase. MEM displays relatively high MSEs across cohort sizes, with performance deteriorating as the number of cohorts grows. PoC, on the other hand, maintains the lowest MSE values across all cohort sizes, indicating its stability and accuracy in handling larger numbers of cohorts.

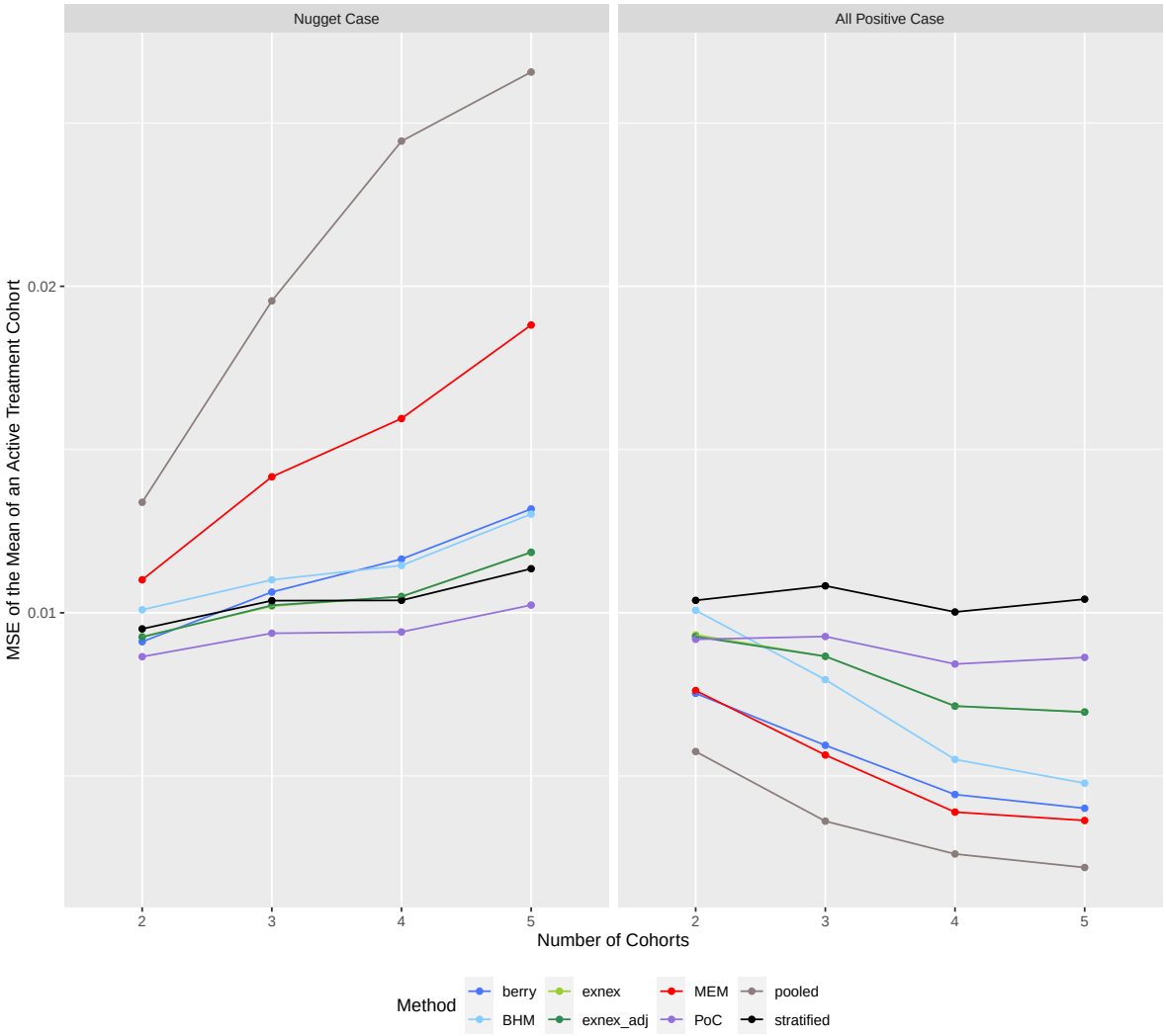


Figure 25: MSE of the posterior distributions' mean of the response rate of an active treatment cohort when increasing the number of cohorts. False go rates for the best negative case= 0.1, k = 1, γ= 0.7, total number of cohorts = 4, p_+ = 0.35, p_-= 0.15, target rate = 0.35, cohort size = 20 patients per cohort, runs =1000.

12.5 Number of Patients in Cohorts

The impact of the number of patients in the cohorts in the nugget and all positive case is explored. It is hypothesized that a larger number of patients in the cohorts would reduce the variance of the response rates' log-odds of the analysis methods, leading to a better distinction between the Go and No-go cohorts. Figure 26 illustrates the relationship between the true Go rates and five, ten, 20, and 30 patients in four cohorts in the nugget and all positive cases, with a fixed false Go rate of 0.1 and decision rule parameters = 0.7 and k = 1.

As anticipated, the true Go rates improve for all methods in both the nugget and the all positive case. In the nugget case, the rate of increase of the methods ExNex, ExNex Adjusted, Stratification, and the method by Berry et al. is approximately the same, and larger than the rate of increase of the method Pooling. In the all positive case, the rate of increase in true Go rates of the methods are roughly the same. In the nugget scenario, the MEM method exhibits the most significant increase in the TrueGo rate (by 0.2) when the number of patients increases from 5 to 10. Interestingly, even though the number of patients per cohort does not increase in increments of ten, the TrueGo rate continues to rise by 0.1 per increment. In the all positive case MEM behaves like Berry or BHM. Everytime PoC has the best TrueGo rate.

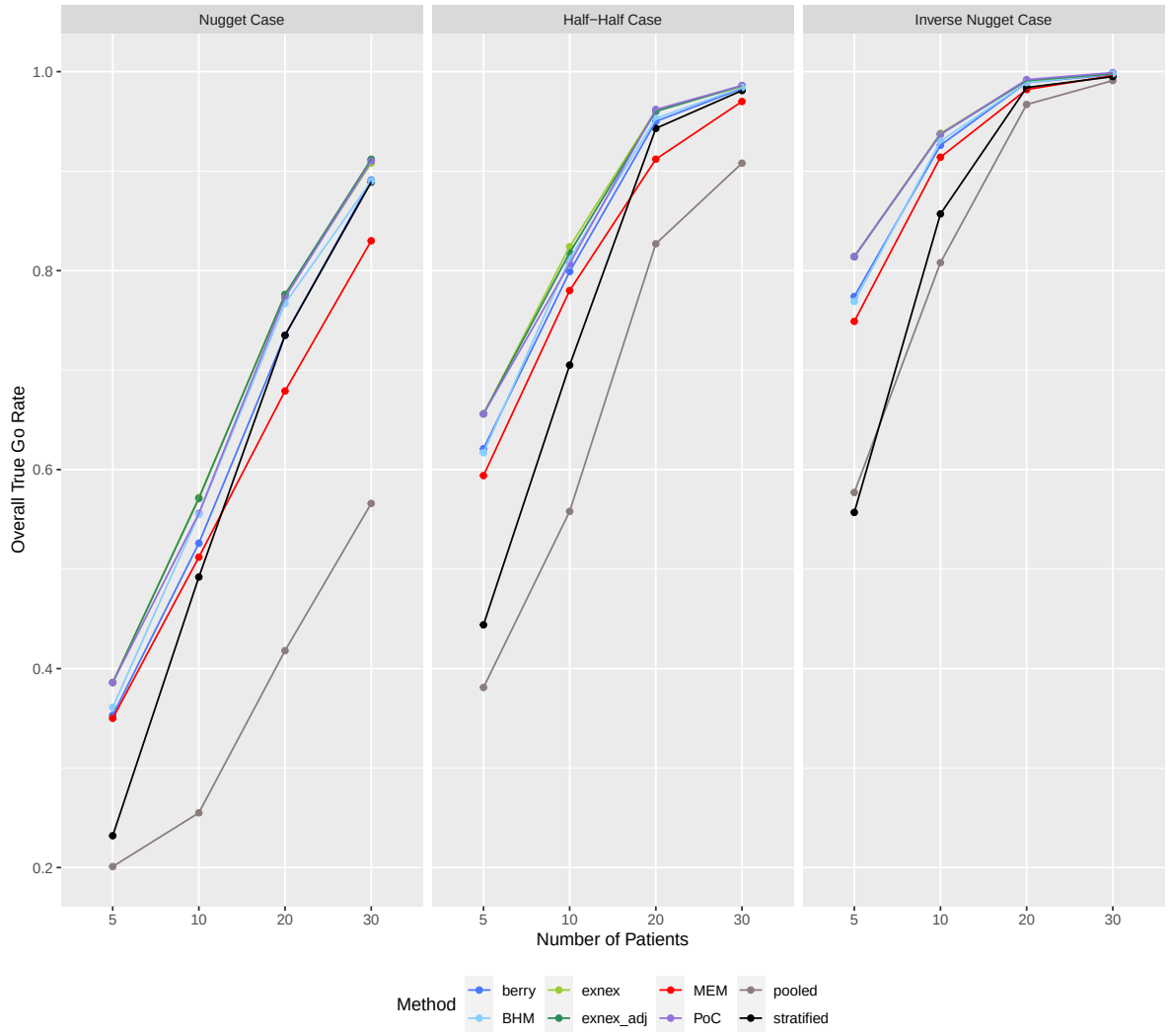


Figure 26: Overall true go rate when increasing the number of patients in the cohorts. False go rates for the best negative case= 0.1, k = 1, γ= 0.7, total number of cohorts = 4, p_+ = 0.35, $p\_$= 0.15, target rate = 0.35, cohort size = 5/ 10/ 20/ 30 patients per cohort, runs =1000.

Assuming that a larger number of patients per cohort would reduce the variance of the cohorts' posterior response rates, it is expected that increasing numbers of patients in the cohorts would positively impact the methods' MSEs. Figure 27 displays the methods' MSEs for an active treatment cohort with five, ten, 20, and 30 patients in the nugget and all positive cases. It demonstrates that the MSEs of the methods decrease in both the nugget and the all positive case with increasing numbers of patients in the cohorts.

In the nugget case: The rate of decrease is for every method almost the same except for MEM and pooled which both have a stronger decrease from five to ten patients than in further increase of patient numbers. Between 10 and 30 the TrueGo rates of the pooled methods stay almost constant and the decrease in MEMs MSE continues but slower as before. Exept for patient cohort of 5 (where berry has the smallest MSE) the PoC has the smallest MSE.

In the all positive case, the smaller the MSE of a method, the smaller the rate of decrease through additional patients per cohorts in the all positive case. So by increasing the number of patients the MSE of the methods approach. The lowest and highest MSE for all patients have the pooled and stratified version, respectively.

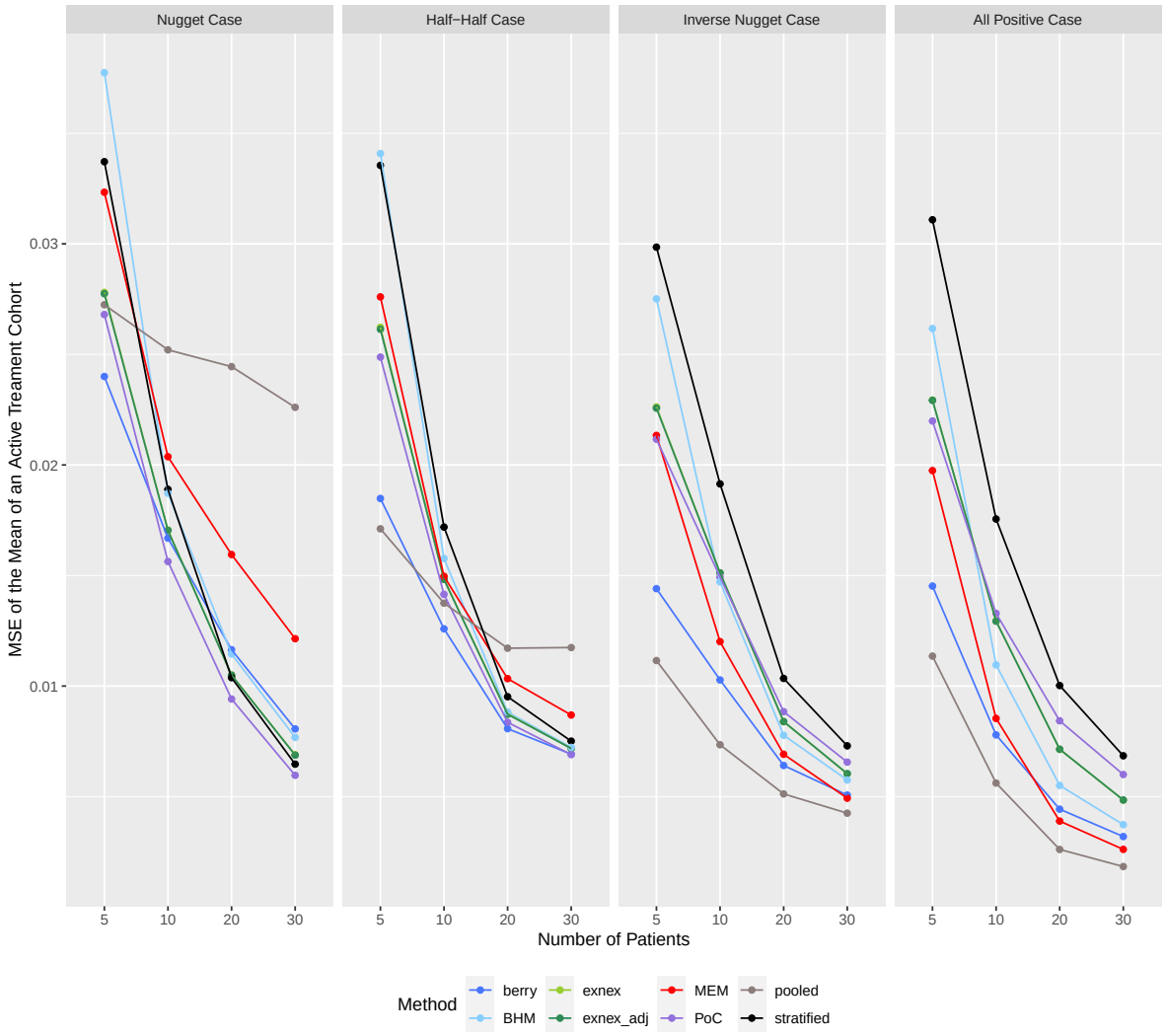


Figure 27: MSE of the posterior distributions' mean of the response rate of an active treatment cohort when increasing the number of patients in the cohorts. False go rates for the best negative case= 0.1, k = 1, γ= 0.7, total number of cohorts = 4, $p_+ = 0.35$, $p\_ = 0.15$, target rate = 0.35, cohort size = 5/ 10/ 20/ 30 patients per cohort, runs =1000.

12.6 Probability Threshold γ

The gamma level is set to 0.7 in all the other scenarios, here we want to determine the influence of changing this parameter. (For more details please check the chapter Go/No Go Decision) The influence of the value for is uncertain, as a higher value for could be offset by a lower value

for p-Boundary, which is chosen to achieve a specific false Go rate of 0.1 in the best negative case. Figure 29 illustrates the effect of different values of on the true Go rates of the analysis methods for the nugget, half-half, inverse nugget and all positive case, with a constant false Go rate of 0.1 and overall decision parameter k = 1.

The figure also shows that all the methods are not really affected by changes of γ .

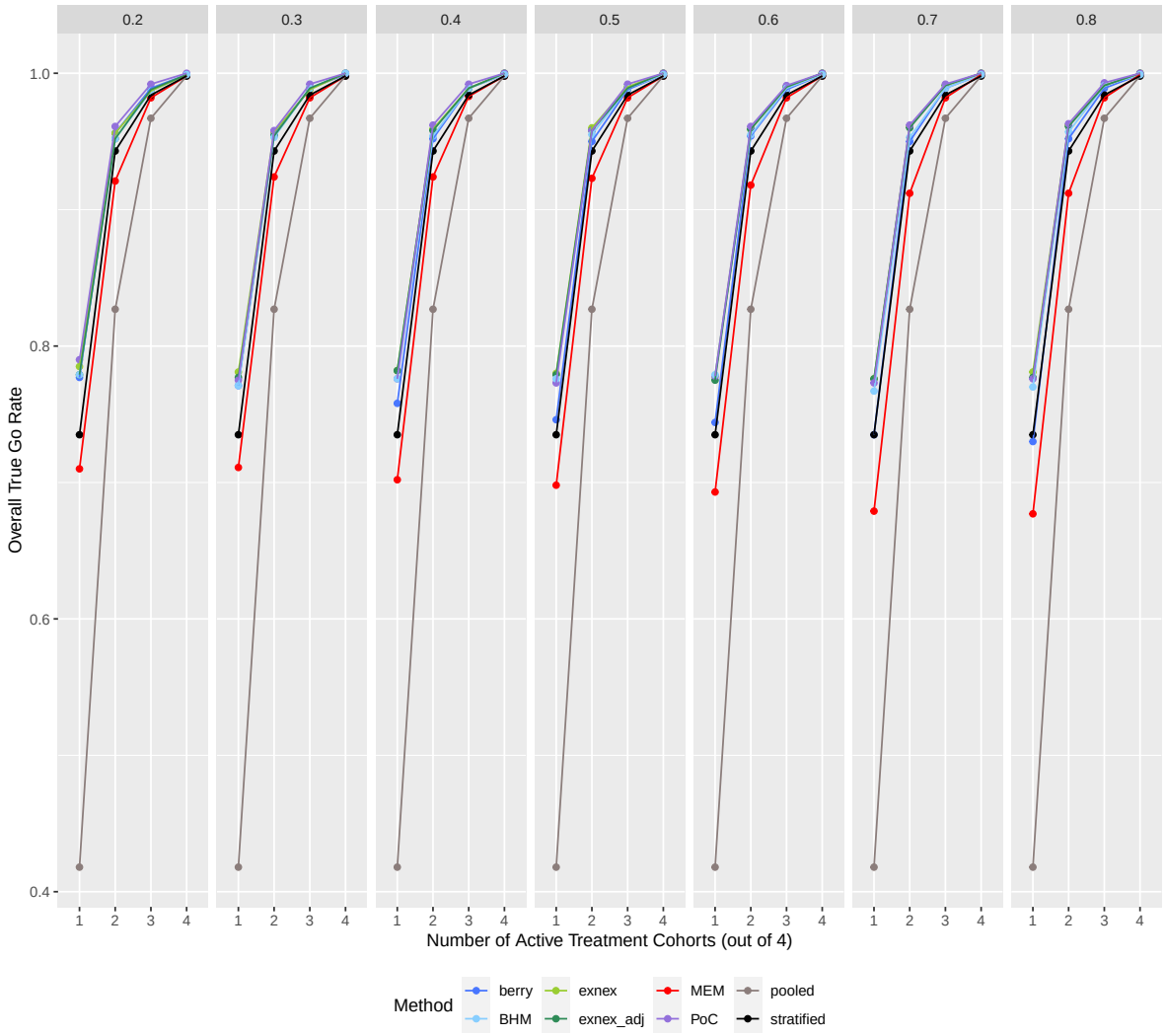


Figure 28: Overall TrueGo Rate when increasing the evidence level γ and the number of active treatment cohorts. Adapting in each case the false go rates for the best negative case= 0.1, k = 1, γ= 0.7, total number of cohorts = 4, p_+ = 0.35, p_-= 0.15, target rate = 0.35, cohort size = 20 patients per cohort, runs =1000.

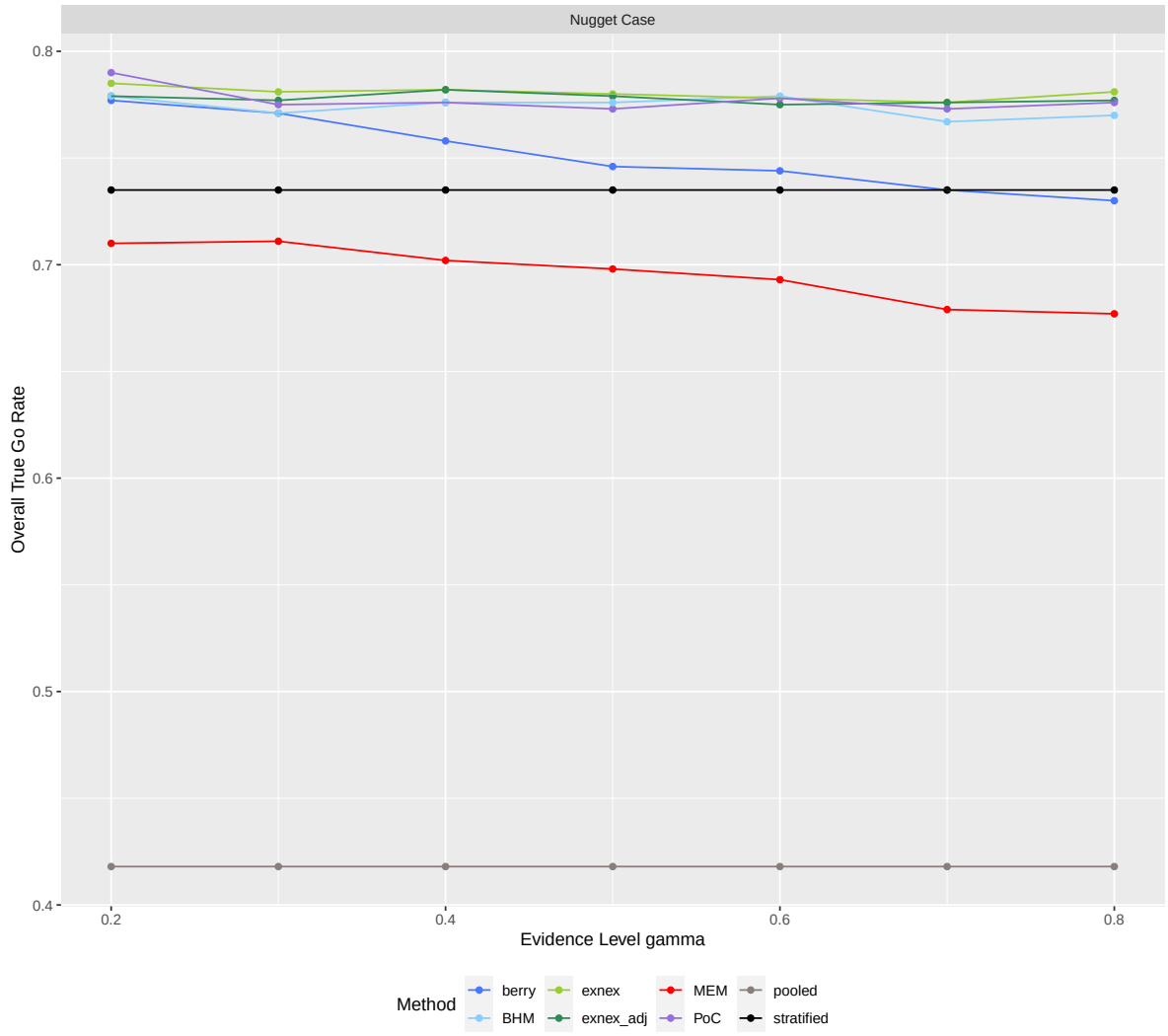


Figure 29: Overall TrueGo Rate when increasing the evidence level γ. Adapting in each case the false go rates for the best negative case= 0.1, k = 1, total number of cohorts = 4, $p_+ = 0.35$, $p\_= 0.15$, target rate = 0.35, cohort size = 20 patients per cohort, runs =1000.

As it is is interesting to see I also plotted the p-boundary level for the corresponding γ level (where the TOER was always adjusted on α-level of 0.1.) One can observe in Figure 30 that when increasing γ the p-boundary level is decreasing almost linearly. Which makes sense since for a higher γ level the that the TOER of 0.1 is still achieved the p-boundary has to be smaller.

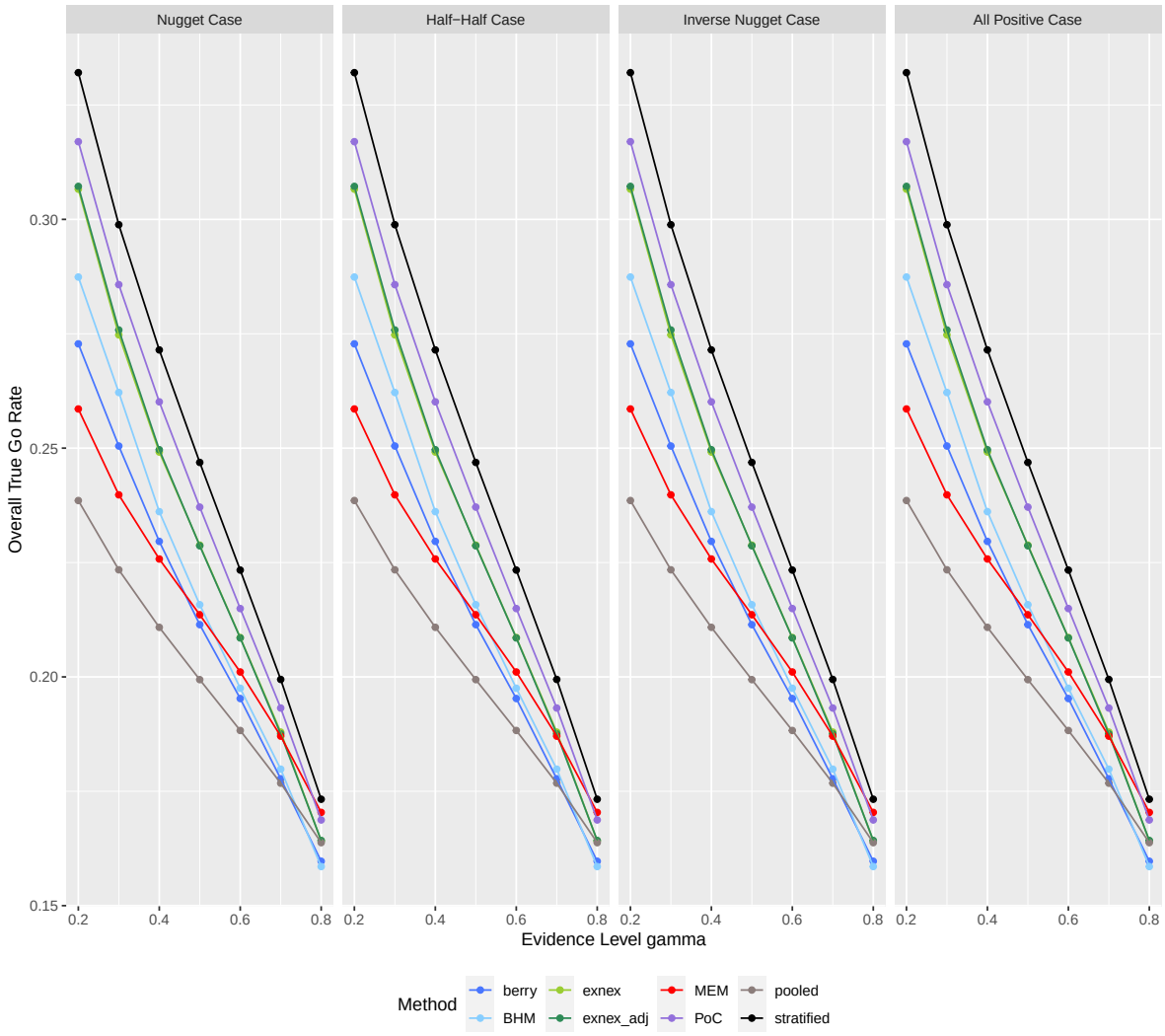


Figure 30: Decision boundary when increasing the evidence level γ. Adapting in each case the false go rates for the best negative case= 0.1, k = 1, total number of cohorts = 4, $p_+ = 0.35$, $p\_= 0.15$, target rate = 0.35, cohort size = 20 patients per cohort, runs =1000.

12.7 Study by Hyman et al.

In the study by Hyman et al. (2015), patients with advanced, non-curable cancers were examined, all of whom had specific genetic mutations that were potentially responsive to targeted therapies. The cancers included melanoma, lung cancer, colorectal cancer, uterine cancer, and other solid tumors. The goal of the study was to assess the effectiveness of specific

treatment in patients with these specific genetic alterations.

For the analysis of this real-world dataset, we used six cohorts with different response rates and set an adjusted alpha level for an all-negative scenario, where each cohort had a response rate of 0.15. The threshold for a "Go" decision was set with $k = 1$.

As expected, the Pooled method performed poorly, as it was unable to account for differences between cohorts. Interestingly, the method that best estimated the true median response rate (indicated by the dashed line) and had the smallest credible interval (CI) was clearly MEM. In the previously tested synthetic scenarios, MEM had not always performed as well, likely due to its complexity and computational demands in calculating an exchangeability matrix. However, in this real-world dataset, MEM demonstrated a clear advantage, probably because it better modeled the cohort-specific effects.

In simpler scenarios, other, less complex methods can yield adequate results, as they avoid the potential overfitting seen with MEM. However, MEM's strong performance in this real-world example highlights its value in complex, nuanced applications. The other methods (excluding Pooled) showed somewhat similar median estimates to MEM but with notably larger credible intervals, especially in the Stratified method. Figure 31 shows that MEM outperforms in terms of both accuracy and precision in this real-world context — which, ultimately, is the most important criterion for evaluating a method's practical relevance.

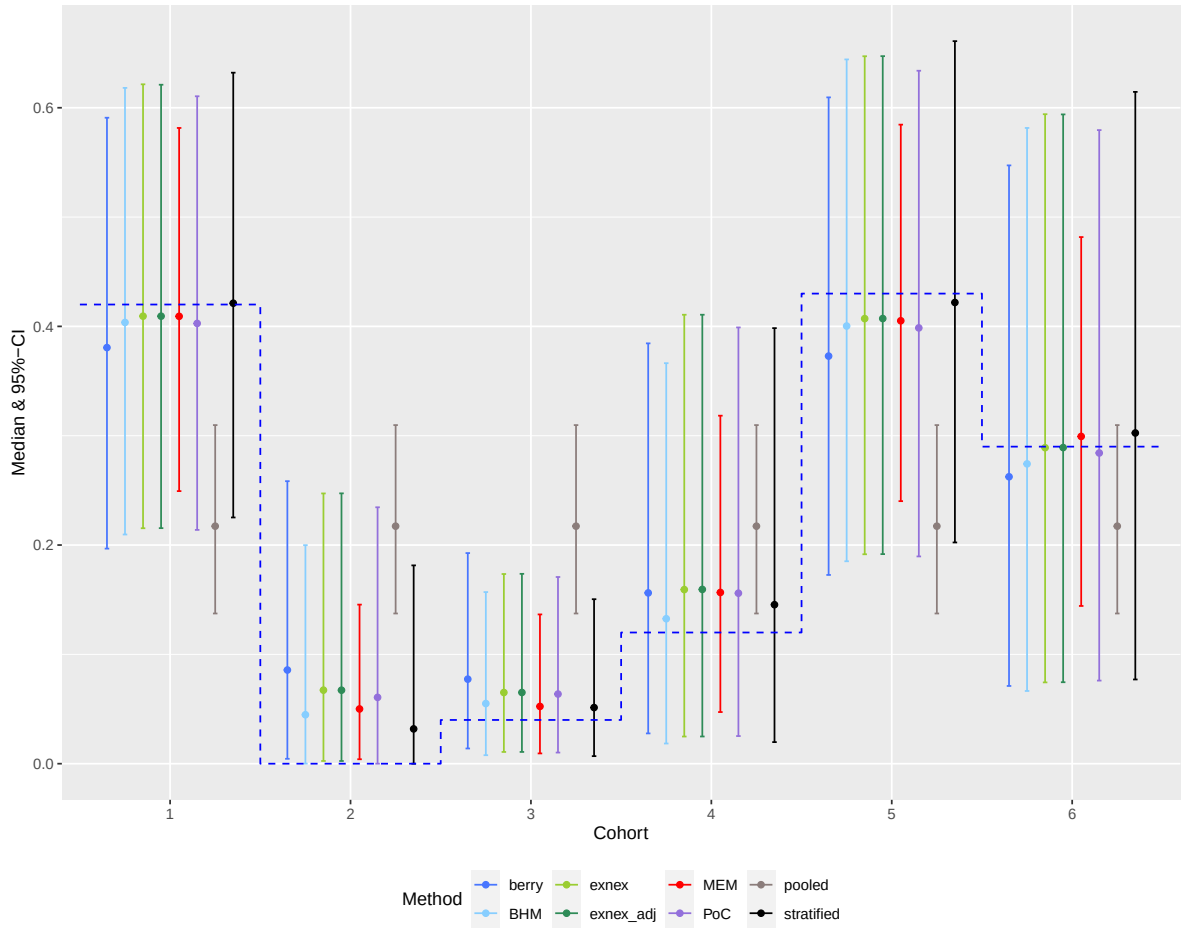


Figure 31: Median and 95% credible interval of the posterior distribution of the response rates for the study cohorts by Hyman et al. (2015) . Adapting in each case the false go rates for the best negative case= 0.1, k = 1, γ= 0.7, total number of cohorts = 6, p_+ = 0.35, $p\_$= 0.15, target rate = 0.35, cohort size = (19, 10, 26, 8, 14, 7) patients per cohort, runs =1000.

In **?@fig-MSE-heyman** one observes the MSE of the mean of each cohort. The MSE for the forth cohort of MEM is very high compared to the other MSEs. But when checking the code this seams not being incorrect. The MEM method has for this cohort the highest bias (compared to other methods, except pooled), and a small variance. But since the bias goes squared into the MSE it has such a high MSE. Pooled has a small MSE since it variance is so small.

`Estimates for the MEM method:`

```
> var_estimates
           [,1]
p_1 0.009347481
p_2 0.000800732
p_3 0.001468459
p_4 0.016689154
p_5 0.013288959
p_6 0.022762979

> bias_estimates
            [,1]
p_1 -0.010764937
p_2  0.050067487
p_3  0.012389935
p_4  0.036625664
p_5 -0.024852438
p_6  0.009368789

> mse_estimates
           [,1]
p_1 0.009463365
p_2 0.003307485
p_3 0.001621969
p_4 0.018030593
p_5 0.013906602
p_6 0.022850753
```

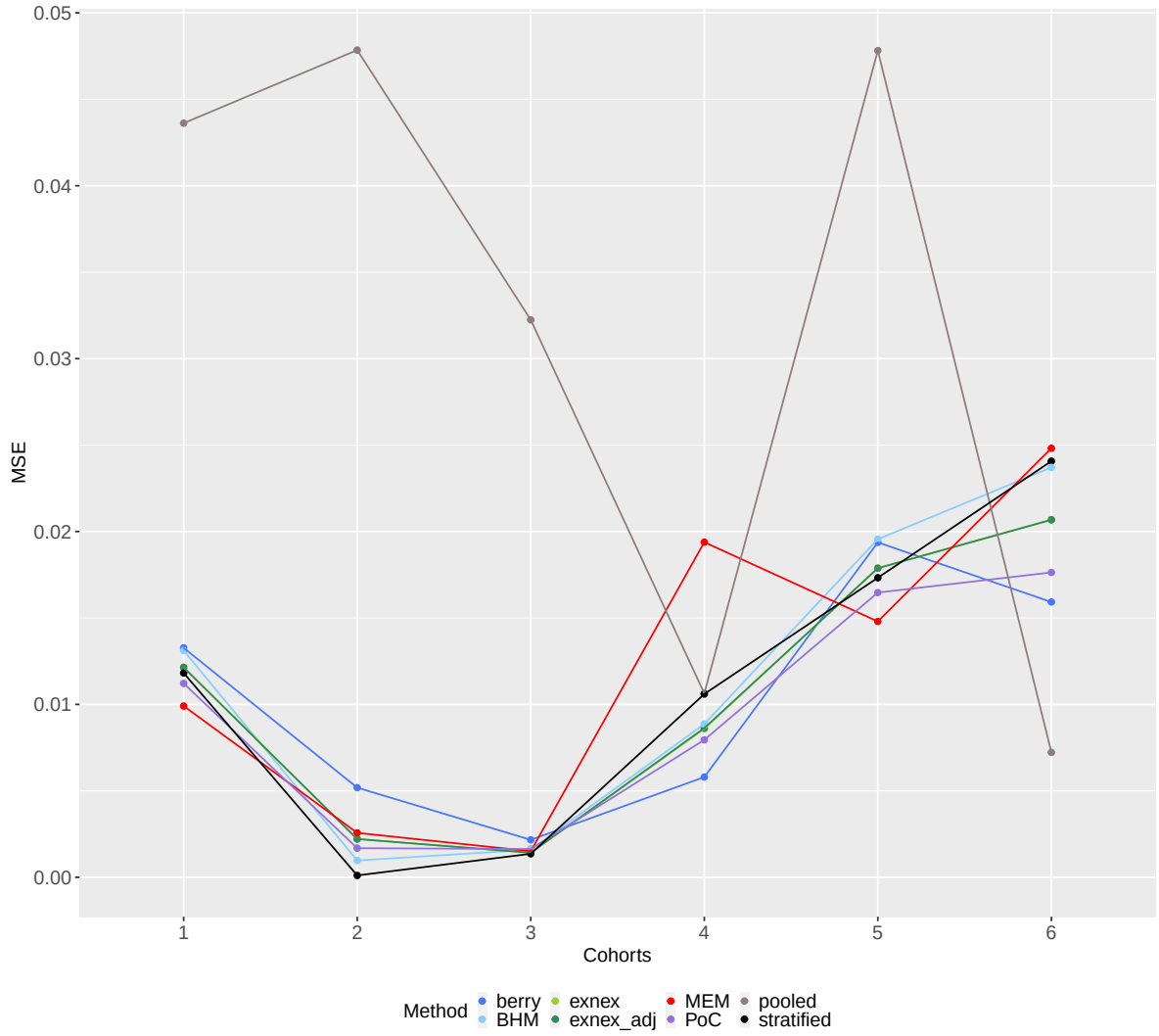


Figure 32: MSE of the posterior distributions' means of the response rates of an active treatment cohort for the study cohort by Hyman et al. (2015) . Adapting in each case the false go rates for the best negative case= 0.1, k = 1, γ= 0.7, total number of cohorts = 6, $p_+ = 0.35$, $p\_= 0.15$, target rate = 0.35, cohort size = 20 patients per cohort, runs =1000.

12.8 Study Talimogene Laherparepvec

https://clinicaltrials.gov/study/NCT02509507?term=Basket%20trial&limit=100&aggFilters=results:with&ran
measures

DCR per modified irRC-RECIST was defined as percentage of participants that had a BOR in 1 of the following: CR, PR or SD. (8. Chapter outcomes)

Cohorts: Hormone Receptor Positive Breast Cancer (HRBC), Triple Negative Breast Cancer (TNBC), Cutaneous Squamous Cell Carcinoma (CSCC), Basal Cell Carcinoma (BCC), Colorectal Adenocarcinoma (CRC)

Table 2: Study Talimogene Laherparepvec

tumor type	HRBC	TNBC	CSCC	BCC	CRC
responses	2	4	2	3	3
size	10	18	10	5	10

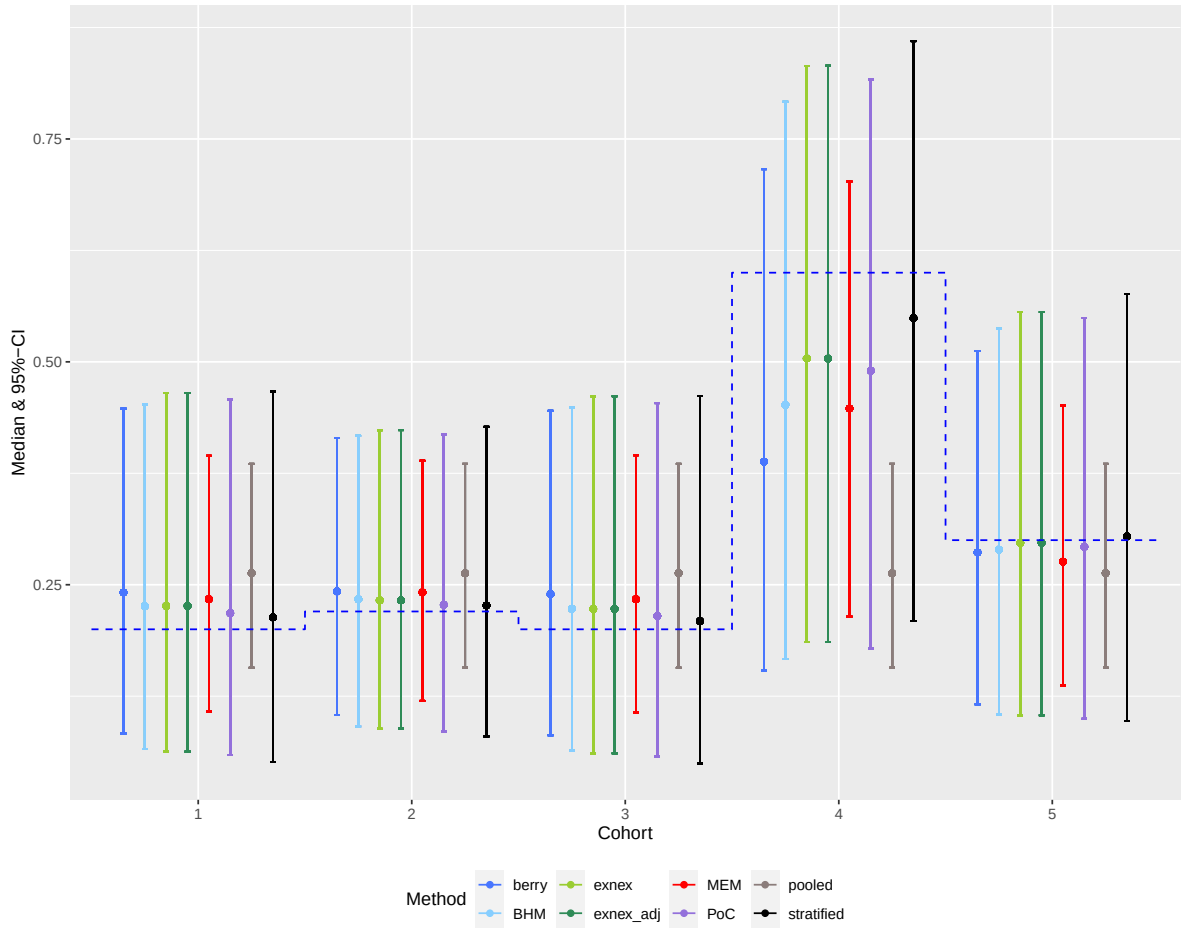


Figure 33: MSE of the posterior distributions' mean of the response rate for the study cohorts by Talimogene Laherparepvec. Adapting in each case the false go rates for the best negative case= 0.1, k = 1, γ= 0.7, total number of cohorts = 6, p_+ = 0.35, p_-= 0.15, target rate = 0.35, cohort size = (10, 18, 10, 5, 10) patients per cohort, runs =1000.

12.9 With Intermediate Cases

Generally, the MEM method has a small confidence interval, but its median is not as close to the true underlying response rate as it is in the case of PoC, Berry, and Exnex. Despite MEM having a high MSE, it performs quite well in the Hyman scenario. As a result, I conducted an intermediate case study where a larger variety of true underlying response rates were present in the basket trial. See Figure 34. Even though MEM performed very well for the Study of Hyman et al. (2015) here it tends to pool to much like in the Scenario Analyses.

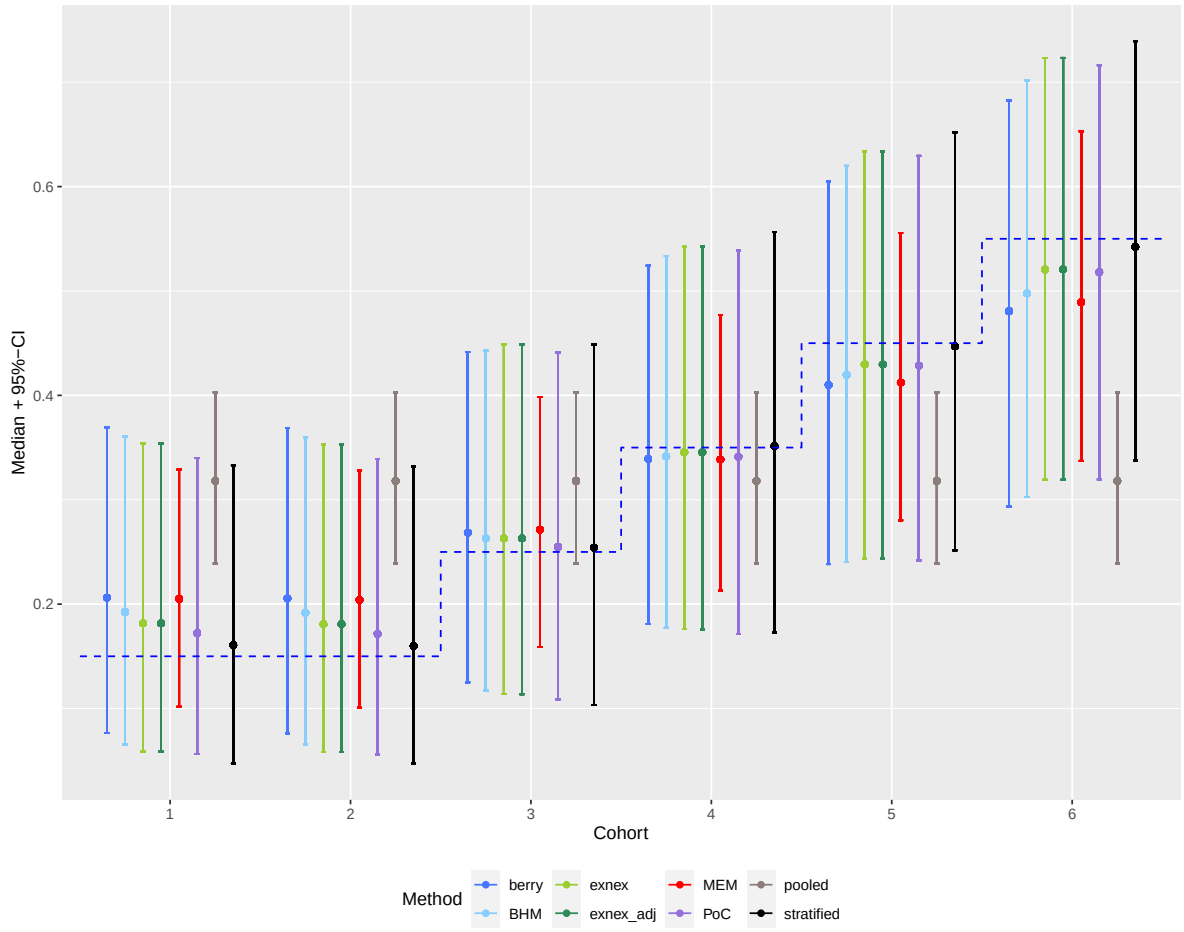


Figure 34: Median and 95% credible interval of the posterior distribution of the response rates for the study cohorts by Hyman et al. (2015). Adapting in each case the false go rates for the best negative case= 0.1, k = 1, γ= 0.7, Total number of cohorts = 6, p_+ = 0.35, p_-= 0.15, target rate = 0.35, cohort size = 20 patients per cohort, runs =1000.

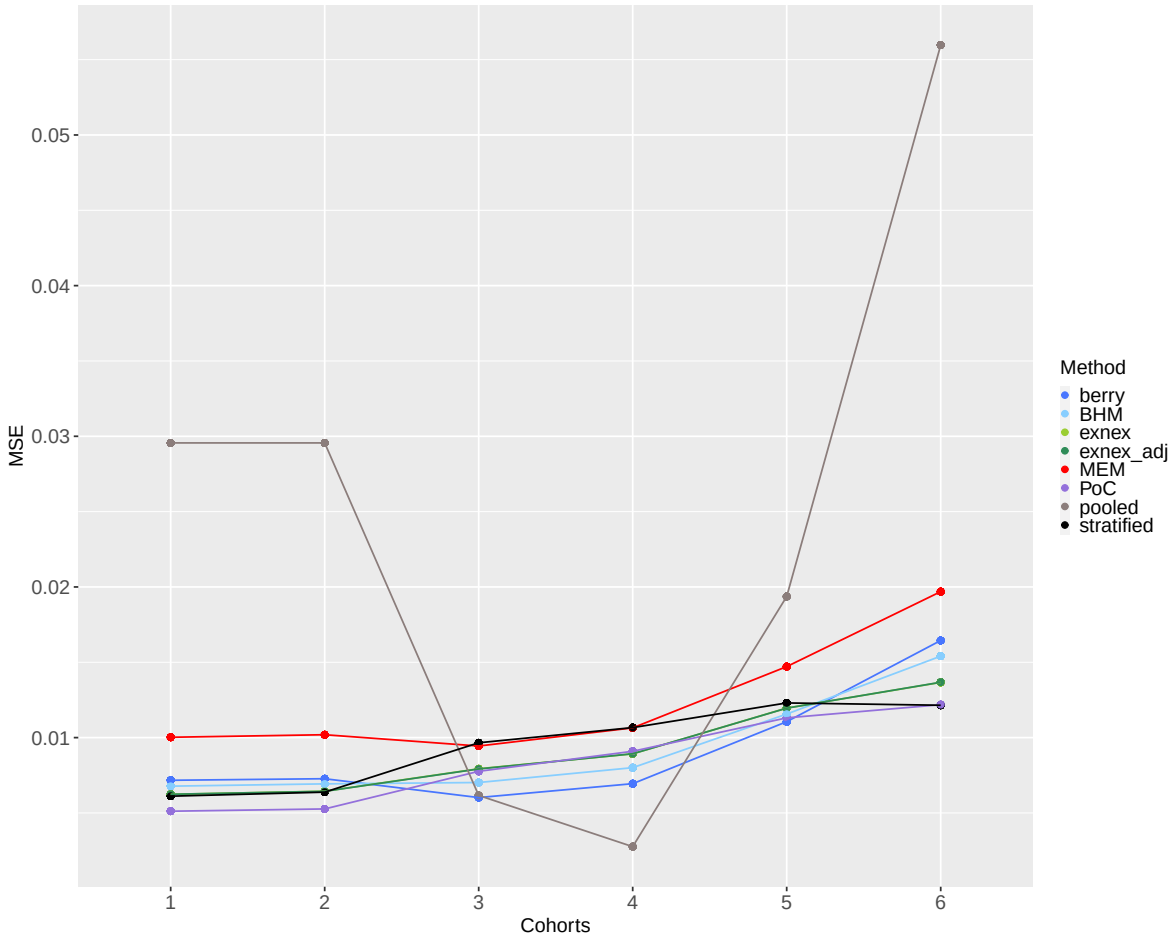


Figure 35: Median and 95% credible interval of the posterior distribution of the response rates for the study cohorts by Hyman et al. (2015). Adapting in each case the false go rates for the best negative case= 0.1, k = 1, γ= 0.7, Total number of cohorts = 6, p_+ = 0.35, $p\_$= 0.15, target rate = 0.35, cohort size = 20 patients per cohort, runs =1000.

12.10 Varying α

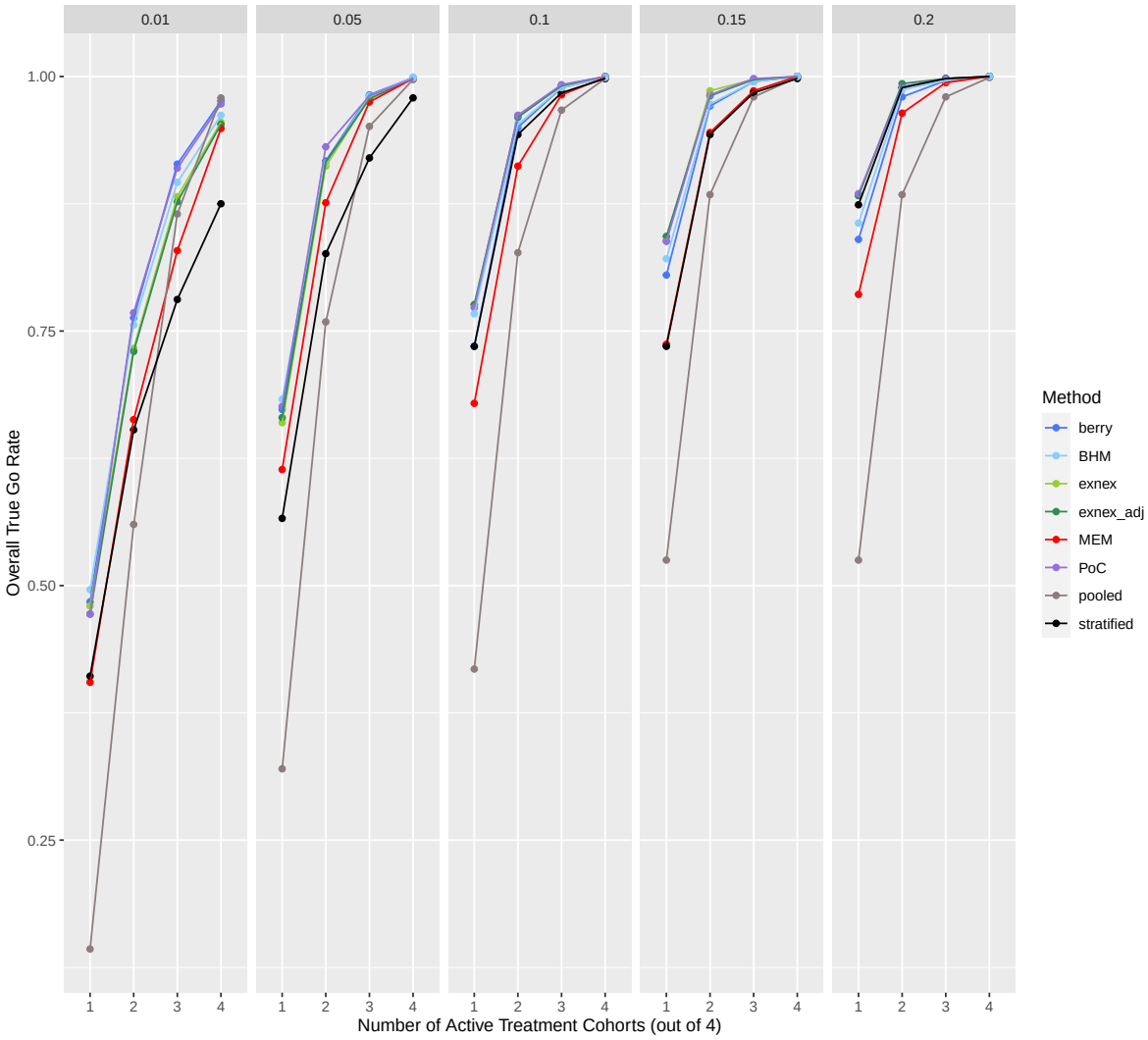


Figure 36: Overall true go rate when increasing the α level and the number of active treatment cohorts. Adapting in each case the false go rates for the best negative case for α = 0.01, 0.05, 0.1, 0.15, 0.2, k = 1, γ= 0.7, total number of cohorts = 4, p_+ = 0.35, $p\_$= 0.15, target rate = 0.35, cohort size = 20 patients per cohort, runs =1000.

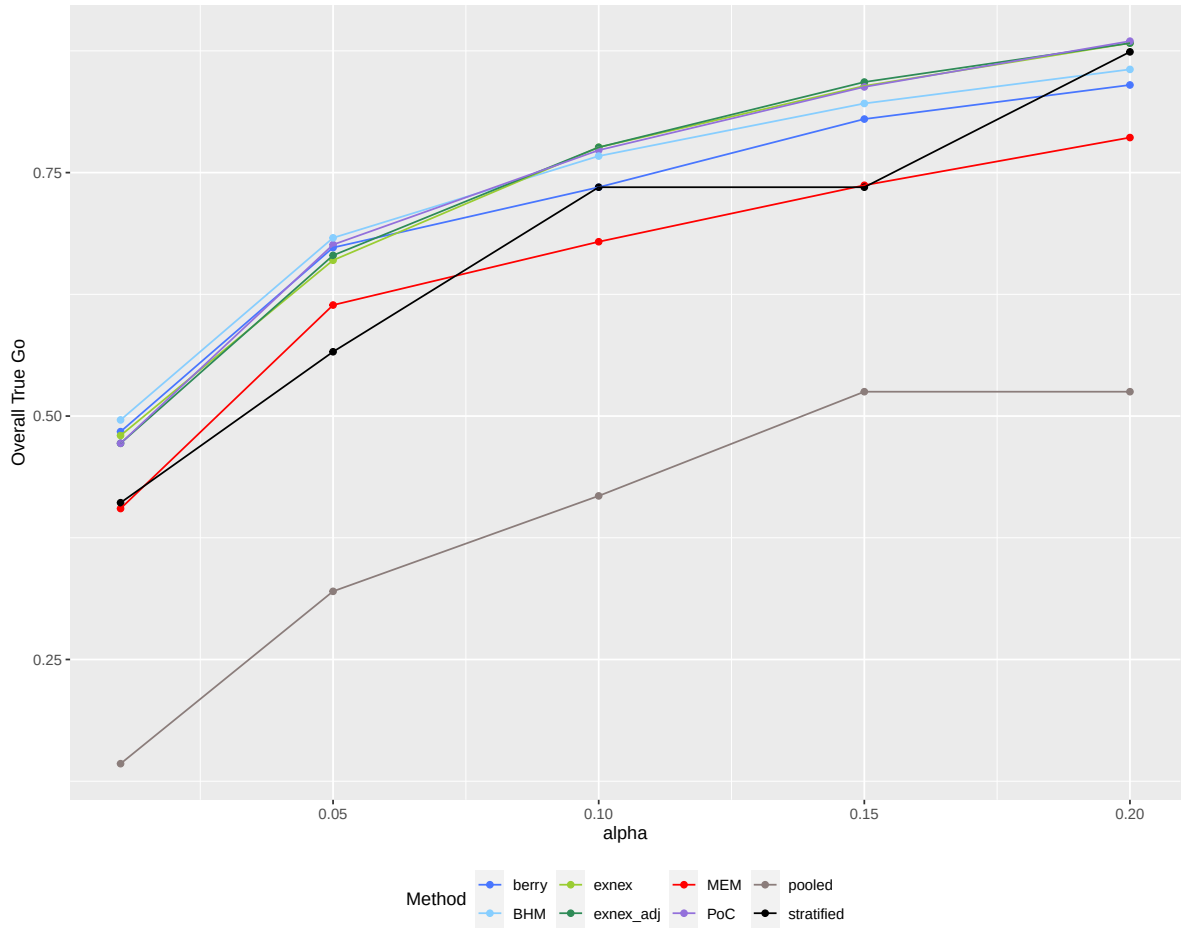


Figure 37: Overall true go rate when increasing the α level. Adapting in each case the false go rates for the best negative case $\alpha = 0.01, 0.05, 0.1, 0.15, 0.2$, k = 1, γ= 0.7, total number of cohorts = 4, $p_+ = 0.35$, $p\_$= 0.15, target rate = 0.35, cohort size = 20 patients per cohort, runs =1000.

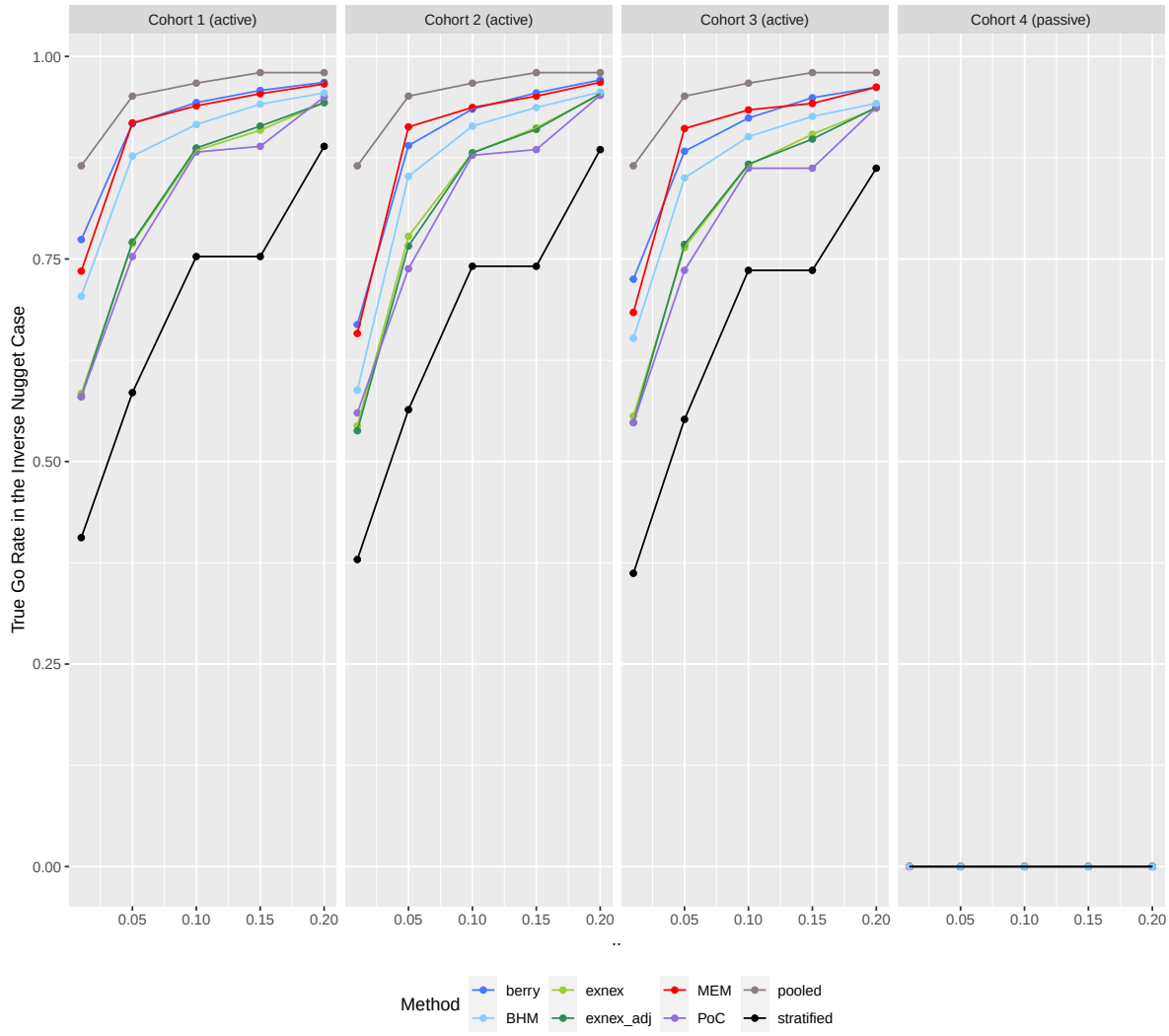


Figure 38: True go rate in the inverse nugget case when increasing the α level. Adapting in each case the false go rates for the best negative case $\alpha = 0.01, 0.05, 0.1, 0.15, 0.2$, k = 1, γ= 0.7, Total number of cohorts = 4, $p_+ = 0.35$, $p\_$= 0.15, target rate = 0.35, cohort size = 20 patients per cohort, runs =1000.

12.11 Changing the prior matrix for MEM matrix

The idea is to adapt the prior matrix W for the MEM method. In the analyses this was assumed to be a non-informative prior matrix with ones on the diagonal entries and 0.5 else. So the probability of two matrices being exchangeable is 0.5 a priori. Since the MEM method tends to share to much, I wanted to adapt the prior matrix for each run. In each run we have the $r_1, ..., r_N$ and the $n_1, ..., n_N$, so my idea of a new prior matrix was the distance matrix:

$P(\Omega_{i,j} = 1) = 1 - |\frac{r_i}{n_i} - \frac{r_j}{n_j}|$ for $i,\ j \in 1, ..., N$. So if they are near each other the probability of exchangeability is high. (So for each run a different prior matrix). The following results were received.

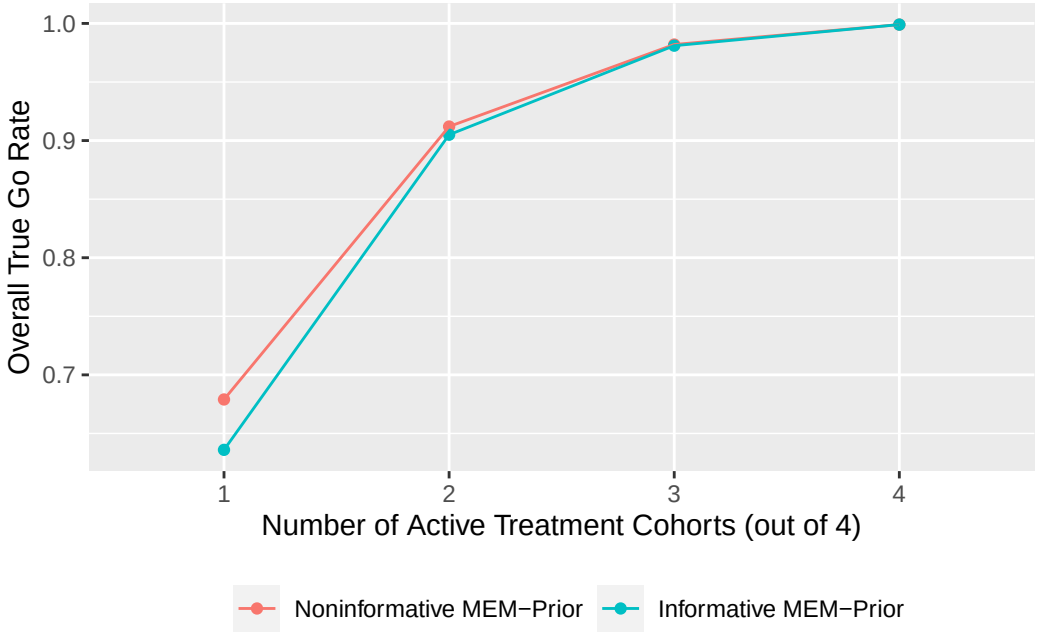


Figure 39: The True Go Rate when Increasing number of cohorts with positive responses, False Go Rates for the Best Negative Case= 0.1, k = 1, γ= 0.7, Total number of cohorts = 4, p+ = 0.35, p-= 0.15, runs =1000.

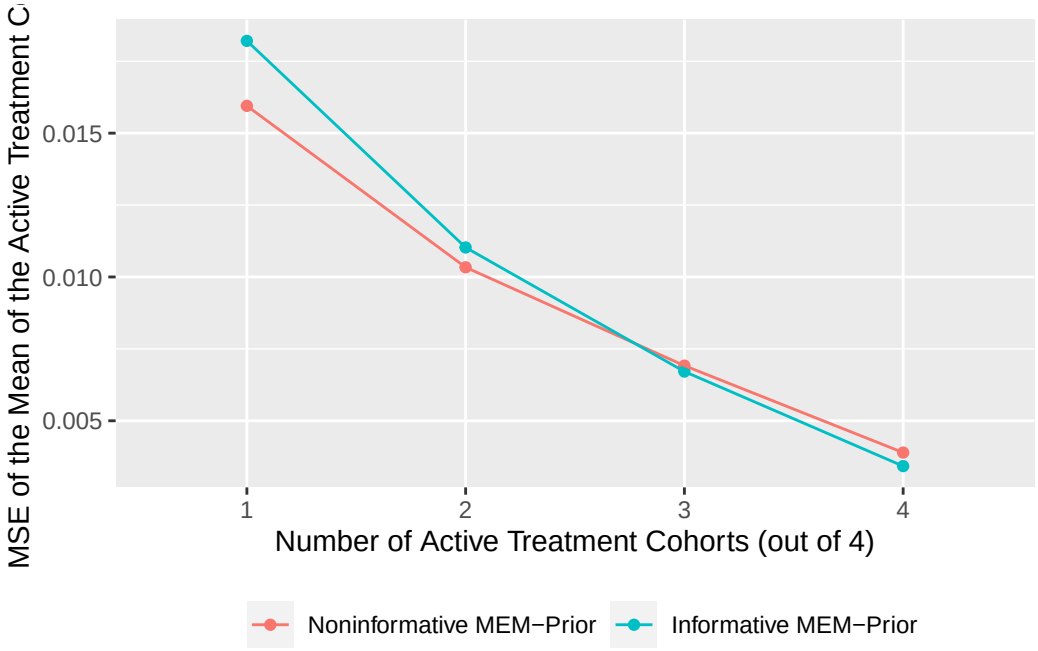


Figure 40: MSE of the Mean of the Active Treatment Cohort when Increasing number of cohorts with Active Treatment Cohort. False Go Rates = 0.1, k = 1, γ= 0.7, Total number of cohorts = 4, p+ = 0.35, p-= 0.15, runs =1000.

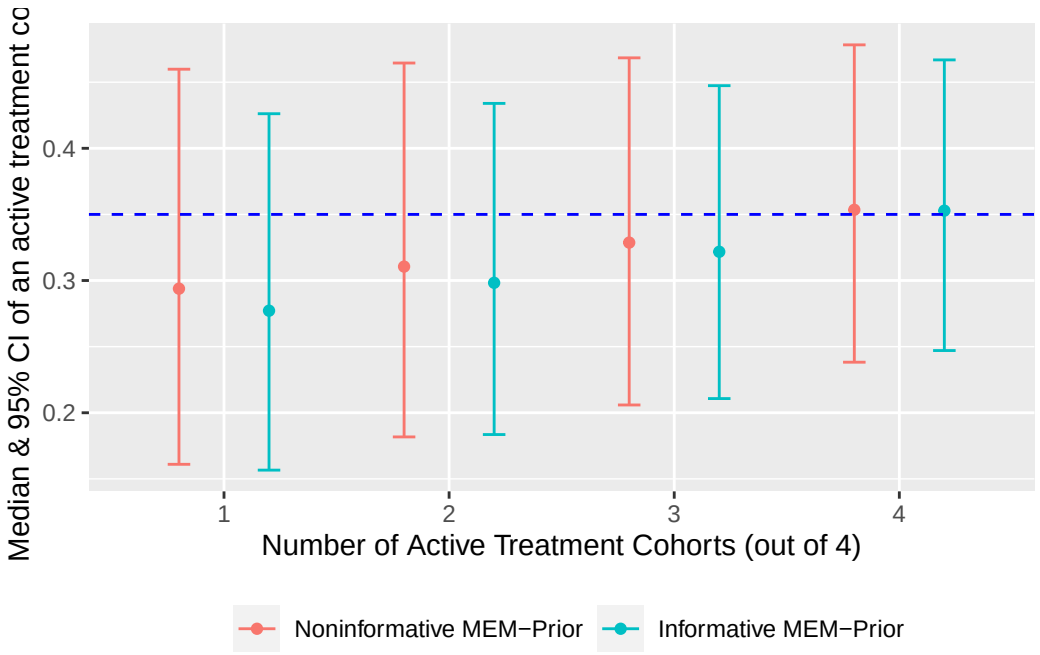
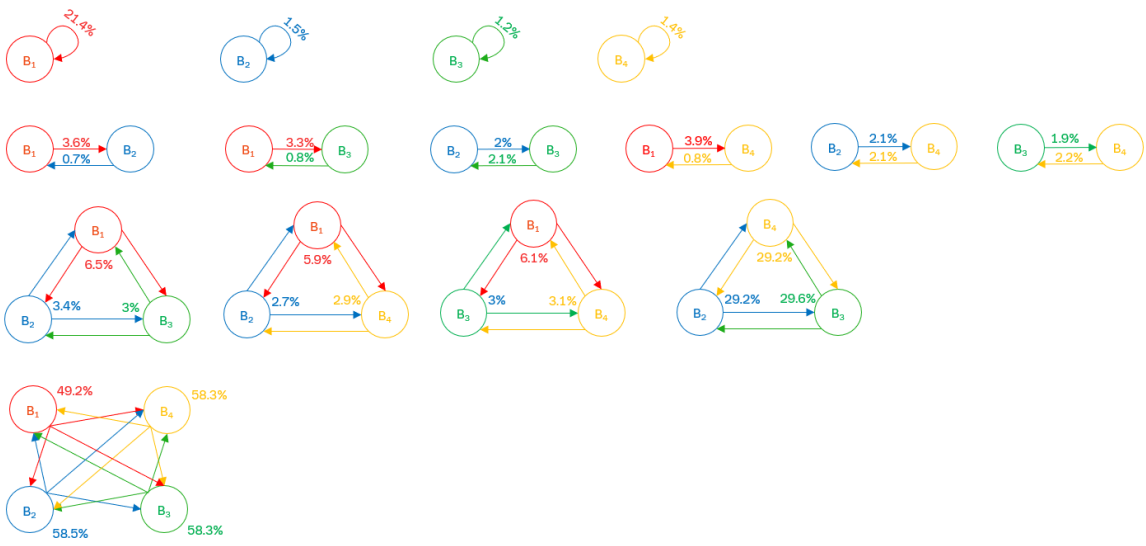


Figure 41: Median with 95% Credible Interval of an Active Treatment Cohort when Increasing number of cohorts with Active Treatment Cohort. False Go Rates = 0.1, k = 1, γ= 0.7, Total number of cohorts = 4, p+ = 0.35, p-= 0.15, runs =1000.

The results are even worse due to the prior, since it favors also sharing between groups which shouldn't share information. E.g., $r_i/n_i = 0.15$ and $r_j/n_j = 0.35$ the $P(\Omega_{i,j} = 1) = 0.8$ so it is much to high. This one can also see in the posterior Ψ matrix. An example for a nugget cohort case $n = 20, p_+ = 0.35, p_- = 0.15, \gamma = 0.7, \alpha = 0.1, run = 1000$, B_1 is the active treatment cohort:



So the posterior distribution for Ψ gives the most probability to the case of pooling all matrices together, even if we are in the nugget case.

To address this problem, I choose another prior matrix choice namely: Calculate all $|\frac{r_i}{n_i} - \frac{r_j}{n_j}|$ and for each couple of baskets where the distance is smaller half of the maximal distance $P(\Omega_{i,j} = 1) = 0.99$ otherwise $P(\Omega_{i,j} = 1) = 0.01$. This is ultimately a strong choice, but it does not advocate for pooling all the baskets together.

This is the Posterior Distribution Matrix Ψ for the Nugget Case. (4 Cohorts, 20 Patients, p+=0.35, p-=0.15). The probability of pooling all baskets together is much less.

```
       basket 1   basket 2   basket 3   basket 4
[1,] 0.51264283 0.05581363 0.04482052 0.04809569
[2,] 0.11636355 0.06961467 0.06866617 0.05965173
[3,] 0.10589840 0.08672247 0.08190285 0.09699464
[4,] 0.05991089 0.07640394 0.07091741 0.06156488
[5,] 0.08894806 0.08651669 0.09154663 0.10958167
[6,] 0.04914617 0.07308253 0.07999419 0.07122561
[7,] 0.06027052 0.53870702 0.54883062 0.54014367
[8,] 0.00681958 0.01313906 0.01332161 0.01274211
```

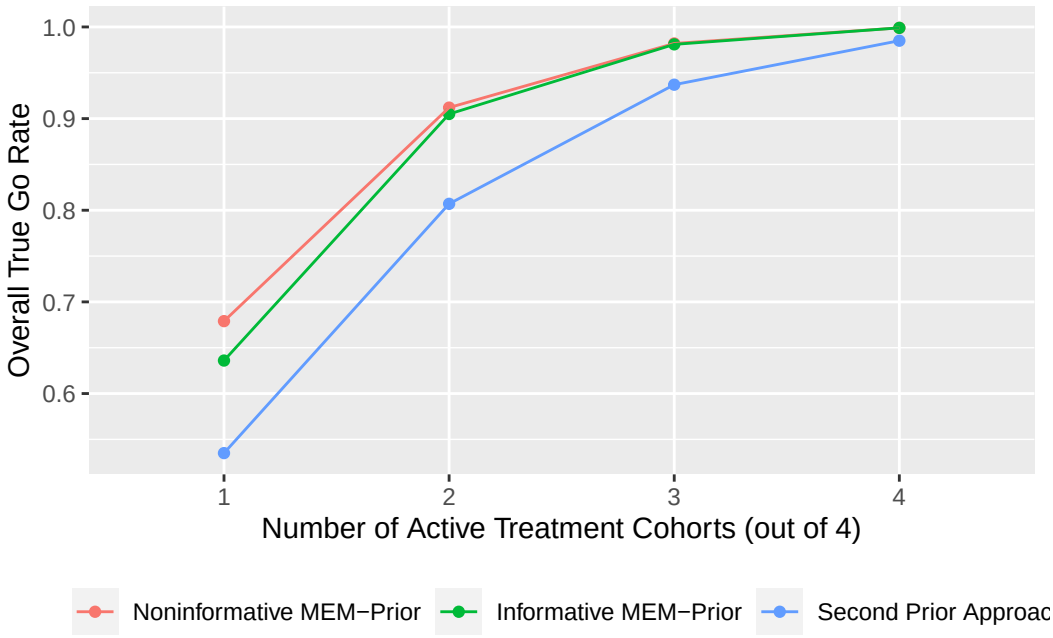


Figure 42: The True Go Rate when Increasing number of cohorts with positive responses, False Go Rates for the Best Negative Case= 0.1, k = 1, γ= 0.7, Total number of cohorts = 4, p+ = 0.35, p-= 0.15, runs =1000.

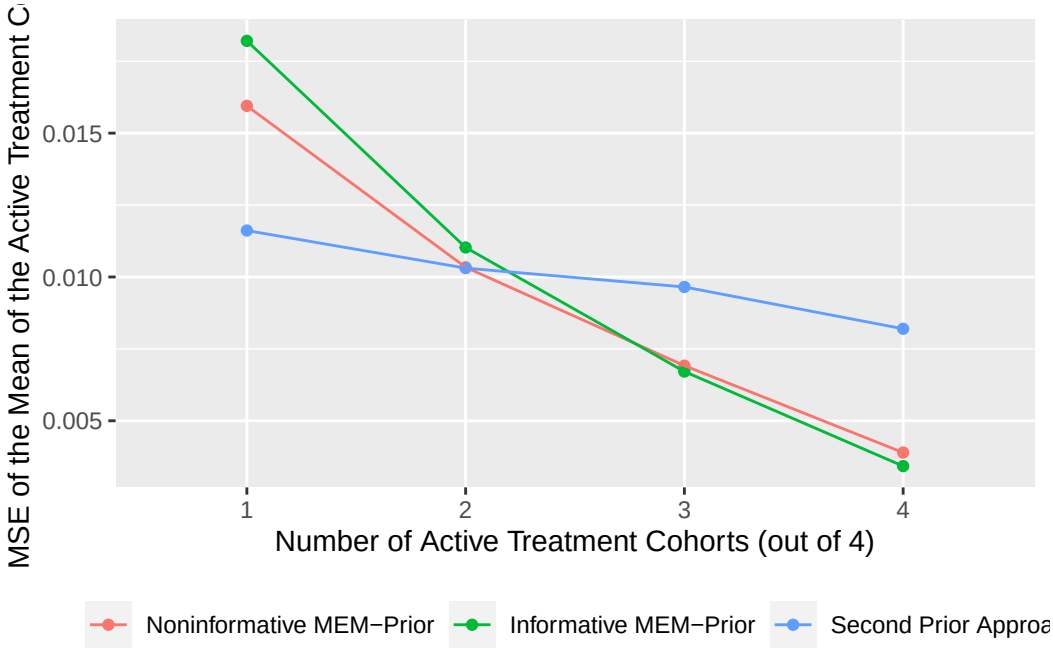


Figure 43: MSE of the Mean of the Active Treatment Cohort when Increasing number of cohorts with Active Treatment Cohort. False Go Rates = 0.1, k = 1, γ= 0.7, Total number of cohorts = 4, p+ = 0.35, p-= 0.15, runs =1000.

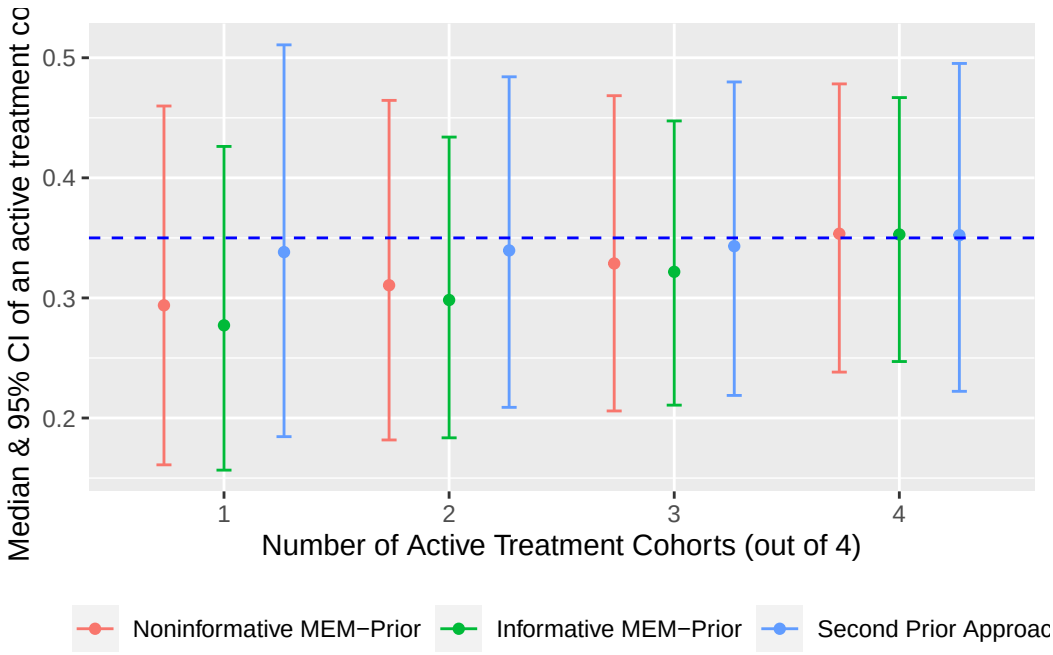


Figure 44: Median with 95% Credible Interval of an Active Treatment Cohort when Increasing number of cohorts with Active Treatment Cohort. False Go Rates = 0.1, k = 1, γ= 0.7, Total number of cohorts = 4, p+ = 0.35, p-= 0.15, runs =1000.

Interpreting the Second Prior Approach on could say, that this performs better for the nugget case (MSE (Figure 43) and Median is better, overall Go not (Figure 42)). Also the results in Figure 44 are impressive. But the other approaches have better MSE for more active treatment cohorts (inverse nugget, all positive). So MEM has potential, maybe choosing the values 0.9 and 0.1 instead of 0.99 and 0.01 as prior values would be a good choose, since in the last prior choice tends to share too less information.

References

"Basket: Basket Trial Analysis." n.d. Accessed November 5, 2024. https://cran.r-project.org/web/packages/basket/index.html.

Baumann, Lukas. 2024. "Baskexact: An r Package for Analytical Calculation of Basket Trial Operating Characteristics." *SoftwareX* 27.

Berry, Scott M, Kristine R Broglio, Susan Groshen, and Donald A Berry. 2013. "Bayesian hierarchical modeling of patient subpopulations: Efficient designs of Phase II oncology clinical trials." *Clinical Trials* 10 (5): 720–34. https://doi.org/10.1177/1740774513497539.

Chen, Nan, and J Jack Lee. 2020. "Bayesian Cluster Hierarchical Model for Subgroup Borrowing in the Design and Analysis of Basket Trials with Binary Endpoints." *Statistical Methods in Medical Research* 29 (9): 2717–32.

Chu, Yiyi, and Ying Yuan. 2018a. "A Bayesian Basket Trial Design Using a Calibrated Bayesian Hierarchical Model." *Clinical Trials* 15 (2): 149–58.

———. 2018b. "BLAST: Bayesian Latent Subgroup Design for Basket Trials Accounting for Patient Heterogeneity." *Journal of the Royal Statistical Society Series C: Applied Statistics* 67 (3): 723–40.

Hobbs, Brian P, and Rick Landin. 2018. "Bayesian Basket Trial Design with Exchangeability Monitoring." *Statistics in Medicine* 37 (25): 3557–72.

Hyman, David M., Igor Puzanov, Vivek Subbiah, Jason E. Faris, Ian Chau, Jean-Yves Blay, Jürgen Wolf, et al. 2015. "Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations." *New England Journal of Medicine* 373 (8): 726–36. https://doi.org/10.1056/NEJMoa1502309.

Jin, Jin, Qianying Liu, Wei Zheng, Zhenming Shun, Tun Tun Lin, Lei Gao, and Yingwen Dong. 2020. "A Bayesian Method for the Detection of Proof of Concept in Early Phase Oncology Studies with a Basket Design." *Statistics in Biosciences* 12: 167–79.

Jin, Jin, Marie-Karelle Riviere, Xiaodong Luo, and Yingwen Dong. 2020. "Bayesian Methods for the Analysis of Early-Phase Oncology Basket Trials with Information Borrowing Across Cancer Types." *Statistics in Medicine* 39 (25): 3459–75.

Kaizer, Alexander M, Joseph S Koopmeiners, and Brian P Hobbs. 2018. "Bayesian Hierarchical Modeling Based on Multisource Exchangeability." *Biostatistics* 19 (2): 169–84.

Kane, Michael J, Nan Chen, Alexander M Kaizer, Xun Jiang, H Amy Xia, and Brian P Hobbs. 2019. "Analyzing Basket Trials Under Multisource Exchangeability Assumptions." *arXiv Preprint arXiv:1908.00618*.

Liu, Yilin, Michael Kane, Denise Esserman, Ondrej Blaha, Daniel Zelterman, and Wei Wei. 2022. "Bayesian Local Exchangeability Design for Phase II Basket Trials." *Statistics in Medicine* 41 (22): 4367–84.

Meyer, Elias Laurin, Peter Mesenbrink, Tobias Mielke, Tom Parke, Daniel Evans, Franz König, and EU-PEARL (EU Patient-cEntric clinicAl tRial pLatforms) Consortium. 2021. "Systematic Review of Available Software for Multi-Arm Multi-Stage and Platform Clinical Trial Design." *Trials* 22: 1–14.

Neuenschwander, Beat, Simon Wandel, Satrajit Roychoudhury, and Stuart Bailey. 2016. "Robust Exchangeability Designs for Early Phase Clinical Trials with Multiple Strata." *Pharmaceutical Statistics* 15 (2): 123–34.

Pohl, Moritz, Johannes Krisam, and Meinhard Kieser. 2021. "Categories, Components, and Techniques in a Modular Construction of Basket Trials for Application and Further Research." *Biometrical Journal* 63 (6): 1159–84.

Psioda, Matthew A, Jiawei Xu, QI Jiang, Chunlei Ke, Zhao Yang, and Joseph G Ibrahim. 2021. "Bayesian Adaptive Basket Trial Design Using Model Averaging." *Biostatistics* 22 (1): 19–34.

Thall, Peter F, J Kyle Wathen, B Nebiyou Bekele, Richard E Champlin, Laurence H Baker, and Robert S Benjamin. 2003. "Hierarchical Bayesian Approaches to Phase II Trials in

Diseases with Multiple Subtypes." *Statistics in Medicine* 22 (5): 763–80.
Wojciekowski, Stephan. 2018. "Bayesian Basket Trials in Oncology - Methods and Go / No-Go Decision Making." Master's Thesis, Berlin, Germany: Humboldt-Universität zu Berlin.