Research Master's programme Methodology and Statistics for the Behavioural, Biomedical and Social Sciences
Utrecht University, the Netherlands

MSc Thesis Emilia Susanna Löscher (8470014)
TITLE: Miscalibration due to heterogeneity in received treatment in prognostic models: a simulation study
May 2023

Supervisors:
Dr. Kim Luijken (University Medical Center Utrecht)
Dr. Maarten van Smeden (University Medical Center Utrecht)
Dr. Ben Van Calster (KU Leuven and Leiden University Medical Center)

Second grader:
Prof. dr. Irene Klugkist (Utrecht University)

Preferred journal of publication: Statistics in Medicine
Word count: 5112

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Miscalibration due to heterogeneity in received treatment in prognostic models: a simulation study

Emilia S. Löscher[1,2]

[1]Department of Social Sciences, Utrecht University, Utrecht, The Netherlands
[2]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

Correspondence
Emilia S. Löscher, Department of Social Sciences, Utrecht University.
Email: e.s.loscher@students.uu.nl

Abstract

Heterogeneity in received treatment can occur in a setting in which a prognostic model was derived versus where a prognostic model is validated or applied, e.g., due to regional differences or treatment policies. We investigated to what extent predictive performance of prognostic models is influenced by heterogeneity in treatment effect size and treatment proportion across derivation and validation. A large and a small sample size simulation study were used and the prediction modeling approaches Ignore Treatment and Treatment-Naïve were considered. For the Ignore Treatment approach, the treatment variable is not used at model derivation. For the Treatment-Naïve approach, the model is derived solely on the untreated individuals. Measures of predictive performance considered were calibration (calibration-in-the-large & calibration slope), discrimination (c-statistic), and overall performance (scaled Brier score). The results from both studies indicated that heterogeneity in received treatment led to miscalibrated predicted risks. The Ignore treatment approach led to substantial underestimation and overestimation of risks on average for increase and decrease in treatment effect size or treatment proportion, respectively. Using the Treatment-Naïve approach, the predicted risks are on average consistently overestimating the risk of the event, also for homogeneity in received treatment. We advise discussing whether heterogeneity in treatment effect size and/or treatment proportion is expected during the data preparation phase of a new prognostic model. Potential miscalibration of risks in case of treatment heterogeneity should be considered when planning to implement the prognostic model in a real-word setting.

KEYWORDS:
Prognostic models, Treatment heterogeneity, Predictive performance, Calibration

1 | INTRODUCTION

Prognostic models are used to provide risk predictions to inform the treatment decision-making process in a medical setting. At external validation and clinical application, prognostic models often have worse predictive performance compared to performance at model derivation [1]. This worsened performance can be a consequence of differences in patient characteristics and/or prevalence rates [2], or sub-optimal modeling strategies at derivation [3]. In this paper we focus on an additional potential reason for compromised external model performance: heterogeneity in received treatment.

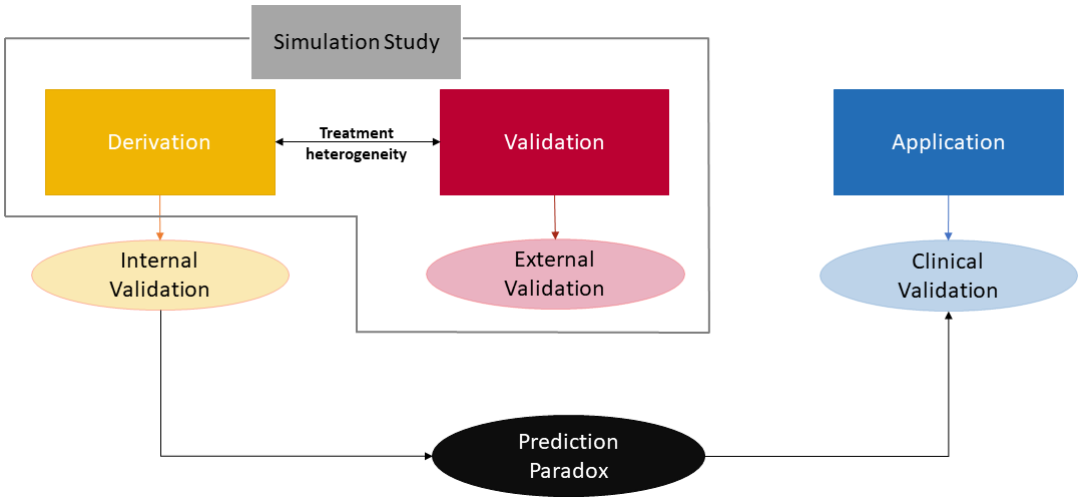


Figure 1 Visualization of the stages of a prognostic model and the scope of the presented study.

Heterogeneity in received treatment refers to any differences in characteristics concerning received treatment between the derivation and (external) validation as well as differences between derivation and implementation population. In this study, by heterogeneity in received treatment, we mean to refer to any differences in treatment between the derivation and (external) validation population (Figure 1). The treatment characteristics considered are treatment effect size and treatment proportion.

When a prognostic model is applied in a real-world setting, human behavior can lead to invalidation of predicted risk. For example, predicted risks can become invalid when treatment decisions are based on predicted risks originating from a prognostic model. Predicted risks are frequently interpreted as the risk of an outcome for an untreated person[4]. If an individual receives treatment for the modeled outcome, this can reduce their risk of that outcome and the initial predicted risk can become invalid[5,6]. This mechanism is referred to as the "Prediction Paradox"[7]. There are many aspects that potentially influence whether and to what extent predicted risks become invalid. To investigate whether treatment assignment is an explanation of the Prediction Paradox, a first step is to quantify the impact of heterogeneity in received treatment across settings on the external performance of a prognostic model when treatment is ignored in the modeling process.

Potential consequences of using approaches where treatment is ignored at prognostic model derivation have been described in literature[4,8]. For an Ignore Treatment approach, for which treatment is not incorporated at model derivation, Pajouheshnia and colleagues[4] state that an increased treatment use at validation will likely affect predictive performance. For the Treatment-Naïve approach where only untreated individuals are used to derive the prognostic model, problems occur when the predicted risks are falsely interpreted in such a way as if the prediction target was the risk for untreated as well as treated individuals, and not only for the untreated population[9]. In this context, Groenwold and colleagues[8] state that when the using the Treatment-Naïve approach and not accounting for treatment at validation, this will lead to predicted risk which seem to be too high. To the best of our knowledge there have not been any studies published studying the scope of influence of varying extents of treatment heterogeneity on predictive performance for both approaches in a systematic manner. Alternative approaches where treatment itself is explicitly modeled[8,9] were not considered. We also excluded a counterfactual hypothetical approach as no validation technique for the counterfactual hypothetical approach has been developed, yet[10].

The aim of this paper was to systematically investigate how the predictive performance of a prognostic model at external validation is affected by heterogeneity in received treatment across the derivation and validation population. This was done by performing two simulation studies in which the predictive performance at external validation for different extents of heterogeneity in treatment effect size and treatment proportion were evaluated. In Section 2, the details for both simulation studies are provided, comprising the aims, data generating mechanisms, estimands, methods, performance measures, and error handling. The results of the simulation studies are presented in Section 3. Section 4 contains a discussion of results, linking to existing literature, and implications for future work.

2 | METHODS

This study comprised a large sample size simulation study (Study 1) as well as a small sample size study (Study 2). Details for both studies are reported according to the ADEMP approach [11]. The code to reproduce the presented analyses and visualizations is available at https://github.com/emsulo/Master-Thesis. Ethical approval was granted by the Ethical Review Board of the Faculty of Social and Behavioural Sciences of Utrecht University and filed under number 22-1815.

2.1 | Aims

The focus of the simulation studies was on quantifying the impact of heterogeneity in treatment effect size and proportion of patients treated across settings of derivation and validation on predictive performance. Study 1 was used to identify possible impact of heterogeneity using a large sample size ($n_{1,d} = n_{1,v} = 100,000$) at derivation and validation for a simplified setting using only one continuous predictor. Study 2 investigated the impact of heterogeneity on predictive performance using a smaller sample size at derivation ($n_{2,d} = 3,500$) and external validation ($n_{2,v} = 3,527$) and eight correlated continuous predictors. In both studies, we used the two modeling approaches Ignore Treatment and the Treatment-Naïve.

2.2 | Data Generating Mechanisms

In both studies, we generated M standard-normally distributed predictors (P) and a binary treatment variable (T), as well as a binary outcome (O) for both the derivation and the validation data. The derivation and validation data only differ regarding received treatment.

The probability of an individual i being treated ($T_i = 1$) was based on predictors $P_1, \dots, P_M$ and given by

$$p_i(T_i = 1 | P_{1i}, \dots, P_{Mi}) = \frac{1}{1 + e^{-\eta_i}}, \qquad (1)$$
$$\eta_i = a + \sum_{m=1}^{M} P_{mi} \quad \text{for } i = 1, \dots, n,$$

where n denotes the sample size and the intercept a was optimized such that the overall treatment proportion corresponded to the treatment proportion of the respective simulation study and scenario (Appendix A). Using a coefficient of log-odds of 1 for all predictors simplified the optimization of a and we implicitly assumed that all predictors are equally important for treatment assignment.

The outcome variable was generated based on the treatment and the predictor(s). The probability of individual i experiencing an event ($O_i = 1$) was given by

$$p_i(O_i = 1 | T_i, P_{1i}, \dots, P_{Mi}) = \frac{1}{1 + e^{-\theta_i}}, \qquad (2)$$
$$\theta_i = \beta_0 + \beta_1 \cdot T_i + \beta_2 \cdot \sum_{m=1}^{M} P_{mi} \quad \text{for } i = 1, \dots, n,$$

where n denotes the sample size and β_1 corresponds to the treatment effect size depending on the scenario. β_0 and β_2 were optimized such that data was generated for which the expected c-statistic in all derivation sets was 0.80 and the proportion of observed outcomes corresponded to the targeted outcome proportion for the respective simulation study (Appendix B). The factor β_2 for all predictors in Equation 2 reflects the assumption made that the relation between each of the predictors with the outcome is constrained to be the same.

Study 1 (Large Sample Simulation)

For Study 1, there was one predictor simulated ($M = 1$) in each of the derivation and validation data sets. The targeted outcome proportion was 0.5.

Homogeneity and heterogeneity in treatment effect size and treatment proportion were considered. The values used for treatment effect size in terms of log-odds were -0.2, -0.5, and -0.8, corresponding to a small, medium, and large effect size, respectively [12]. Hence, differences in log-odds of 0.3 and 0.6 were studied. The log-odds are negative as treatment has a risk-reducing

effect. In the following, the absolute values of the log-odds are used. That way a numerically larger effect size corresponds to a larger effect strength. The treatment proportions in a population were either 0.1, 0.5, or 0.9 resulting in treatment proportion in- and decreases of 0.4 and 0.8 across derivation and validation being studied.

A full factorial design for the settings for the derivation and validation data set provided us with $N_1 = 81$ different simulation scenarios (Table A1). For each scenario data of sample of size $n_{1,d} = n_{1,v} = 100,000$ was generated for derivation and validation, respectively. This way we expected around 50,000 events for each derivation set. Using the formula presented by Riley and colleagues in Figure 2[13] this sample size corresponds to a mean absolute prediction error smaller 0.001. Each scenario was simulated once.

Study 2 (Small Sample Simulation)

Eight standard-normally distributed, correlated predictors ($M = 8$; $cor(P_i, P_j) = 0.2, i \neq j$) were simulated in each of the derivation and validation data sets in Study 2. The targeted outcome proportion was 0.2 to reflect commonly occurring class imbalance.

For Study 2, we chose settings for which the maximal absolute difference was 0.3 for treatment effect size and 0.4 for treatment proportion across the derivation and validation data set. Heterogeneity in treatment effect size was limited to an in- or decrease of 0.3 in terms of log-odds. The values for treatment effect size in terms of log-odds were -0.2, -0.5, and -0.8, corresponding to a small, medium, and large effect size, respectively[12], and are henceforth noted as absolute values. For heterogeneity in treatment proportion a difference of 0.05 was added. Thereby, the focus was on an in- or decrease in treatment proportion by 0.05 or 0.4. The treatment proportion in a population was either 0.1, 0.15, 0.5, 0.55, 0.9, or 0.95.

The resulting total number of scenarios was $N_2 = 91$ (Table A2). For each scenario a sample of size $n_{2,d} = 3,500$ and $n_{2,v} = 3,527$ was generated for derivation and validation, respectively. The sample size for derivation was calculated such that the required number of events for the development of a clinical prediction model for a desired shrinkage level of 90% presented by Riley and colleagues[13] was reached in the untreated group used for derivation for the Treatment-Naïve approach (Appendix C.1). The required sample size for external validation was calculated based on Riley and colleagues[14] (Appendix C.2). For each scenario 1,000 iterations were performed.

2.3 | Estimands: Prediction Targets

The prediction target of the the Ignore Treatment approach is the risk of the binary adverse event given the M predictors for individuals in a population with a treatment policy similar to that in the derivation setting. The prediction target of the Treatment-Naïve approach is the risk of the binary adverse event given the M predictors for individuals in a population that remains untreated.

2.4 | Methods: Prediction Approaches

The Ignore Treatment and the Treatment-Naïve approach were used to derive models for risk predictions for the binary adverse event using M continuous predictors based on each of the simulated derivation data sets with varying settings for the effect size and proportion treated. The models derived based on the derivation data sets were validated on the derivation data sets (apparent performance) and the validation data sets (external performance). The focus was on the external performance with underlying heterogeneity in treatment effect size or treatment proportion across the derivation and validation data sets.

Ignore Treatment approach

For the Ignore Treatment approach, the variable treatment is not incorporated at any point when predicting the risk of the adverse event[10]. The fitted model is a logistic regression model which has only one predictor (P_1) for Study 1 and eight predictors ($P_1, \dots, P_8$) for Study 2.

Treatment-Naïve approach

For the Treatment-Naïve approach, the variable treatment is used to filter the data set. Only individuals who were not treated are used to derive the prognostic model[9]. The fitted model is a logistic regression model with only one predictor (P_1) for Study 1 and eight predictors ($P_1, \dots, P_8$) for Study 2.

2.5 | Performance Measures

Predictive performance was assessed using measures for calibration, discrimination, and overall performance. Calibration was measured with the calibration-in-the-large coefficient and the calibration slope[15]. The calibration-in-the-large coefficient is obtained by comparing the average predicted risks with the overall event rate and has a target value of 0[2]. The calibration slope has a target value of 1 and is measuring the spread of the predicted risks relative to the observed proportions[2]. As a measure for the model discrimination, the c-statistic was used. A value of 1 corresponds to perfect discrimination of the model between individuals with and without an event[2,15]. For a c-statistic of 0.5 the model performs just as well as random guessing[16]. The Brier score was used to measure overall performance where a smaller Brier score corresponds to better overall performance[15,17]. All of these performance measures were calculated using the `val.prob` function from the `rms` package[18]. For performance comparisons we used a scaled version of the Brier score[19]. The results for Study 2 were pooled over the 1,000 iterations for each scenario by calculating the median and plotting box plots for each of the performance measures. Furthermore, we collected descriptives on the generated data sets including the outcome proportion, the treatment proportion, and the outcome proportion for the untreated population per data set.

For all analyses, we used R version 4.2.1[20], including packages `dplyr`[21], `rms`[18], and `ggplot2`[22] for visualizations. Logistic regression models were fit with the `glm` function from the `stats`[20] package.

2.6 | Error Handling

In both studies, the `tryCatch.W.E` function from the `simsalapar`[23] package was used to record potential warnings and errors occurring in any of the scenarios. At model derivation, warnings regarding non-convergence of the logistic model were captured and saved for each iteration for both modeling approaches. The same was done for warnings concerning the fitted probabilities being numerically 0 or 1. If both warnings occurred, only the last (regarding fitted probabilities) was captured by the `tryCatch.W.E` function and saved.

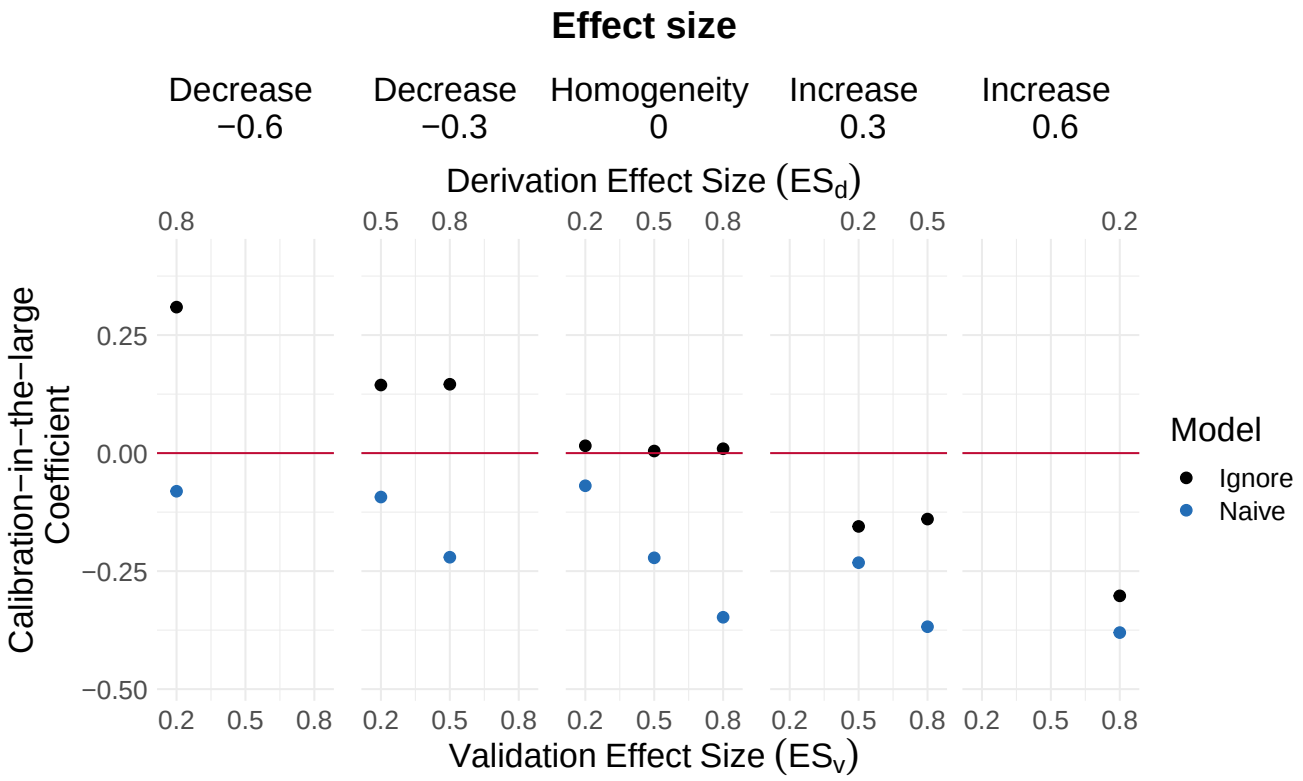
Furthermore, when risk predictions were obtained for the Treatment-Naïve approach using the `predict` function from the `stats`[20] package, potential warnings regarding potentially misleading predicted risks due to rank-deficiency of fit were tracked. To allow for continuation of the simulation study run in case errors occurred when performance measures are calculated for Treatment-Naïve approach, the `tryCatch.W.E` function was used to record an error occurring for an apparently singular matrix and the simulation continued with the following iteration. For Study 1, data would have been newly generated if an error occurred as only one iteration was performed per scenario. If an error had occurred in Study 2, the simulation iteration would have been discarded. No warnings or errors occurred and neither the new generation of data for Study 1 nor the omitting of iterations for Study 2 were implemented.

3 | RESULTS

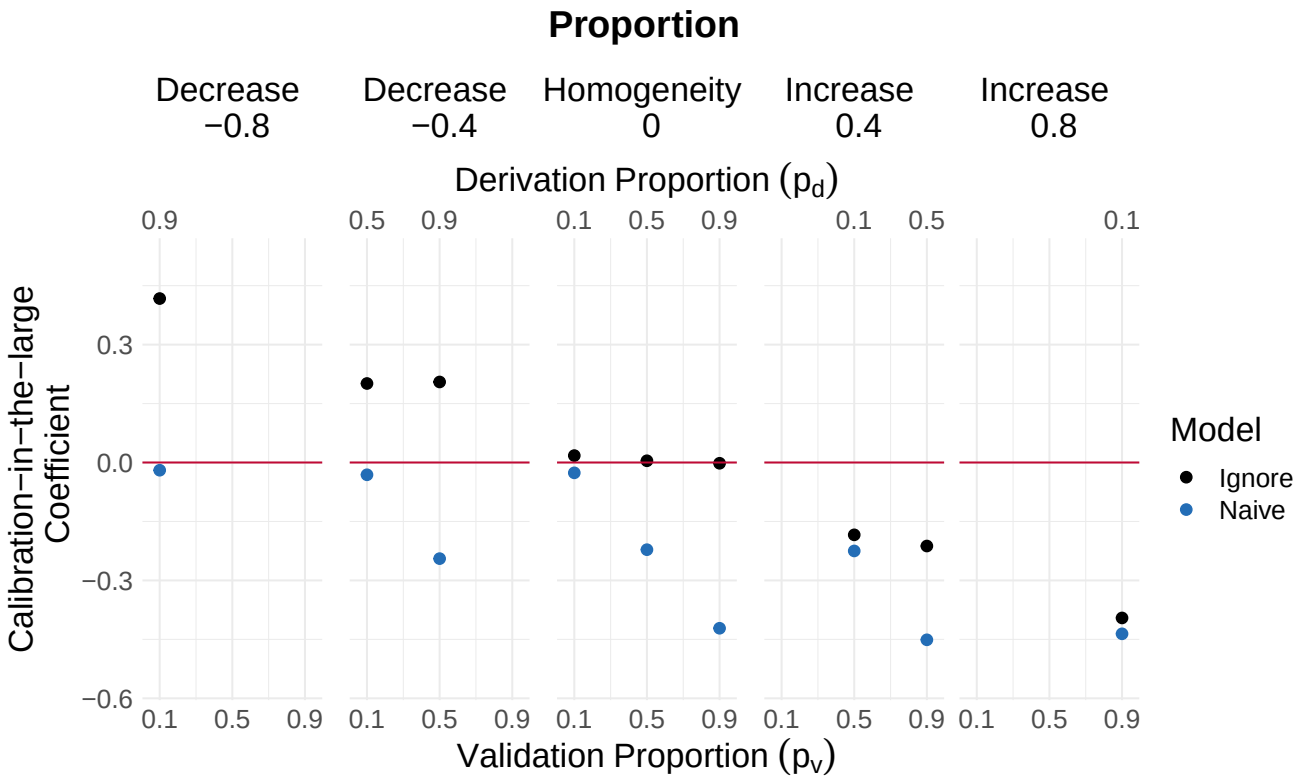
We assessed the simulated derivation data regarding outcome proportion and c-statistic, and found that the averaged observed characteristics corresponded to the targeted c-statistic of 0.8 and to the targeted outcome proportion 0.5 and 0.2, for Study 1 and Study 2, respectively (Table B5). Results for heterogeneity in either treatment effect size or treatment proportion at a time are presented here. Visualizations of the results regarding predictive performance for simultaneous heterogeneity in treatment effect size and treatment proportion can be found in the associated Shiny App (https://emilialoescher.shinyapps.io/6_visualization_app/) and the Supplementary Materials. The latter also include Tables for those results and are available at https://github.com/emsulo/Master-Thesis.

3.1 | Study 1 (Large Sample Simulation)

For the Treatment-Naïve approach, there was miscalibration across all scenarios. The risks were consistently overestimated on average (calibration-in-the-large coefficient < 0) and predicted risks were too extreme compared to observed proportions (calibration slope < 1). The extent of overestimation depended solely on the treatment characteristics at validation and was most pronounced for scenarios with a proportion of 0.9 receiving treatment at validation and scenarios where the effect of the treatment was 0.8 at validation (Figure 2). The performance regarding discrimination for the Treatment-Naïve approach was the same as for the Ignore Treatment approach described below with little variation across the scenarios (Figure D3). For heterogeneity in

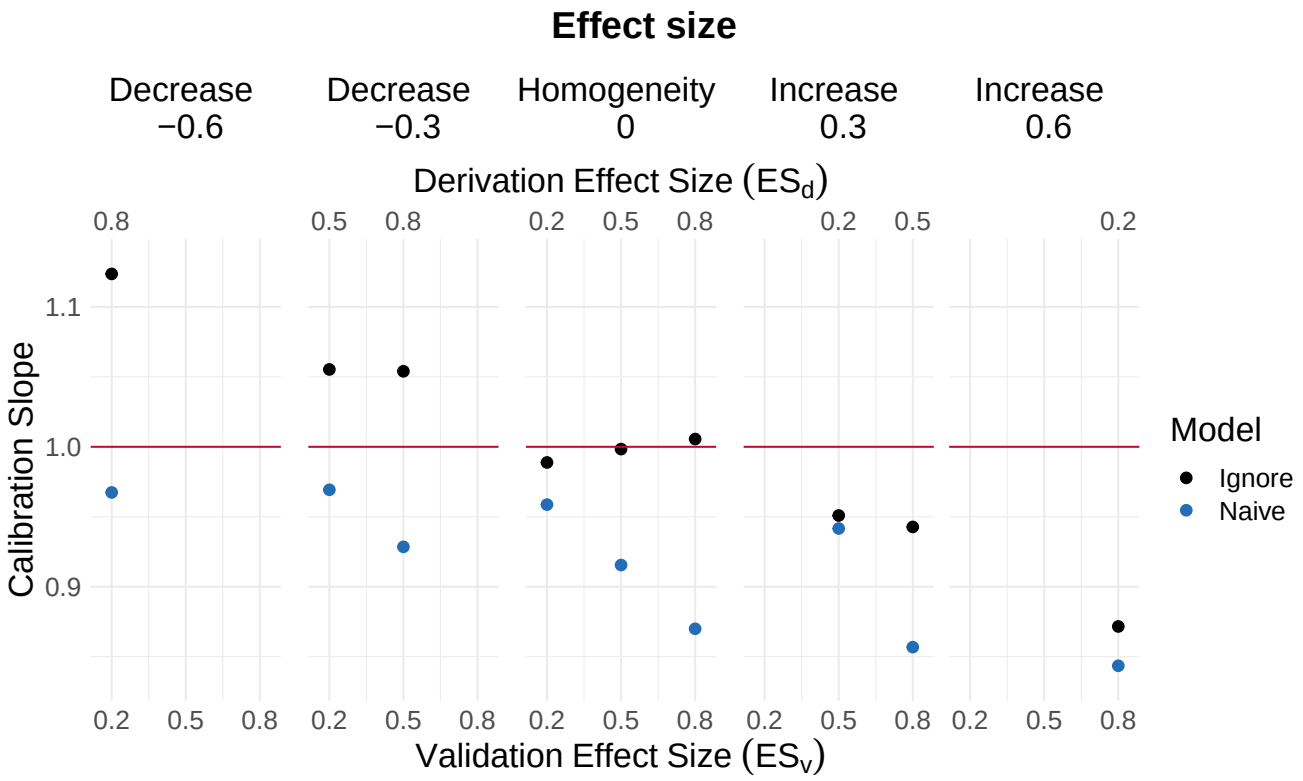


(a)

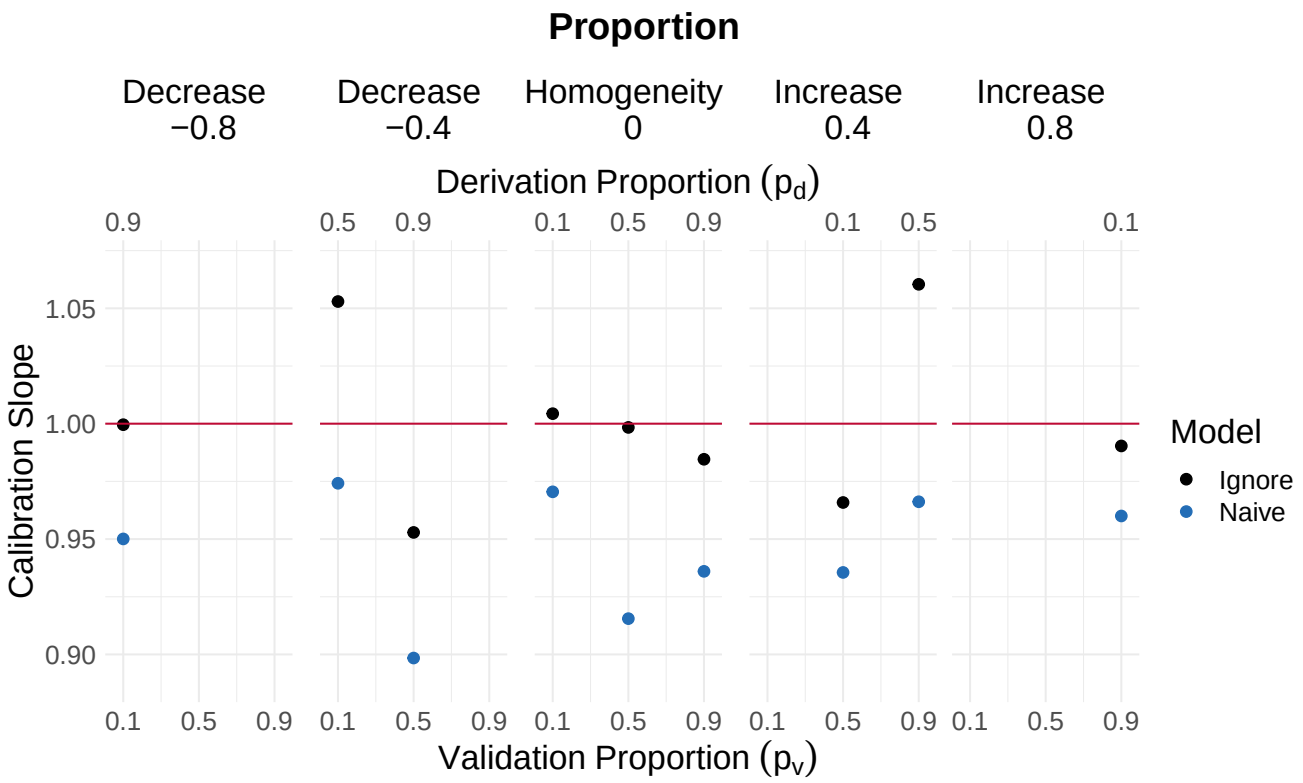


(b)

Figure 2 Study 1 - Calibration-in-the-large coefficient for (a) varying or equal treatment effect and constant treatment proportion of 0.5 and (b) varying or equal treatment proportion and constant medium treatment effect (0.5).



(a)



(b)

Figure 3 Study 1 - Calibration slope for (a) varying or equal treatment effect and constant treatment proportion of 0.5 and (b) varying or equal treatment proportion and constant medium treatment effect (0.5).

treatment proportion for both approaches, there was no clear influence of the extent of heterogeneity on discrimination detected. For the Treatment-Naïve approach, overall performance was lower when a decrease in effect size was present, and higher for an increase in the effect size across derivation and validation. For the treatment proportions, the overall performance depended on the extent of heterogeneity as well as the precise treatment proportion at derivation and validation (Figure D4).

3.1.1 | Homogeneity in received treatment

In case of homogeneity in received treatment, the predicted risks were well-calibrated for the Ignore Treatment approach. The calibration-in-the-large coefficient was close to 0 (-0.01 - 0.02) and the calibration slope close to 1 (0.98 - 1.01) across scenarios (Tables E6 & E7). Regarding discrimination, the performance was similar to the target c-statistic of 0.80 at derivation (0.80 - 0.81; Tables E6 & E7). The scaled Brier score lay between 0.27 and 0.30 for the Ignore treatment approach (Tables E6 & E7).

3.1.2 | Increased proportion treated or treatment effect size at validation

For an increase in either treatment property from derivation to validation, the predicted risks were miscalibrated. Using the Ignore Treatment approach resulted in overestimation of risks of the event (calibration-in-the-large coefficient < 0). The extent of overestimation was larger the larger the increase in the respective treatment property (Figure 2). The greatest overestimation for an increase in treatment effect size was observed for an increase from small to large effect size (increase by 0.6 in terms of log-odds) and a constant treatment proportion of 0.9 with a calibration-in-the-large coefficient of -0.54 (Table E6). For an increase in treatment proportion by 0.8 and a constant treatment effect size of 0.8 overestimation was most extreme with a calibration-in-the-large coefficient of -0.65 (Table E7).

With respect to the calibration slope, the obtained predicted risks were too extreme compared to the observed proportions for an increase in treatment effect size (calibration slope ranging from 0.87 to 0.98; Figure 3 (a) & Table E6). For scenarios with an increase in treatment proportion, the predicted risks were too extreme compared to observed proportions for scenarios with a treatment proportion of 0.5 in the validation data set with a calibration slope between 0.91 and 0.97 (3 (b) & Table E7) and too moderate for scenarios with a treatment proportion of 0.5 at derivation with a calibration slope between 1.01 and 1.10 (Figure 3 (b) & Table E7). Too moderate predicted risks refer to predicted risks being to close to the overall event fraction.

An increase in effect size by 0.6 (in terms of log-odds) led to slightly poorer performance regarding discrimination for a constant treatment proportion of 0.1 and 0.5 with a c-statistic between 0.78 and 0.79 (Figure D3 (a) & Table E6). The scaled Brier scores for an increase in treatment effect lay between 0.22 and 0.28 and were lower compared to scenarios with treatment homogeneity (Figure D4 (a) & Table E10). This corresponds to better overall performance. For an increase in treatment proportion, overall performance was not dependent on the degree of heterogeneity. Instead, overall performance was worse for scenarios with a derivation proportion of 0.5 and better whenever the validation proportion was 0.5 (Figure D4 (a)).

3.1.3 | Decreased proportion treated or treatment effect size at validation

A decrease in either treatment effect size or treatment proportion resulted in underestimation of risks (calibration-in-the-large coefficient > 0). Underestimation was more extreme the larger the decrease (Figures 2). Additionally, the extent of underestimation on average depended on the underlying constant treatment characteristic with more extreme underestimation for a treatment proportion of 0.9 and a large treatment effect size and vice versa (Figures D1 & D2). Regarding the calibration slope for the Ignore Treatment approach, the obtained predicted risks were too close to the overall event fraction (calibration slope > 1) for a decrease in treatment effect.

A decrease in treatment effect size led to slightly better discrimination performance measured by the c-statistic (0.80 - 0.82) compared to scenarios with homogeneity in received treatment (Table E6). For scenarios with a constant treatment proportion of 0.1 or 0.5, a decrease in treatment effect led to worse overall performance compared to homogeneity in treatment. This was most pronounced for a constant treatment proportion of 0.5 (Figure D4 (a)).

3.2 | Study 2 (Small Sample Simulation)

The results for Study 2 are presented in terms of the median of the performance measures across the $1,000$ iterations as well as the variation across scenarios. It has to be kept in mind that the maximum extents of heterogeneity considered in Study 2 were differences of 0.4 for treatment proportion and differences of 0.3 for treatment effect size.

The results in Study 2 were in accordance with the results in Study 1, except that in Study 2 the predictive performance for the Treatment-Naïve approach regarding discrimination and overall performance was consistently, slightly worse in median compared to results obtained for the Ignore Treatment approach (Figures D7 & D8).

3.2.1 | Homogeneity in received treatment

As in Study 1, the predicted risks were well-calibrated in case of homogeneity in received treatment for the Ignore Treatment approach with the calibration-in-the-large coefficient being close to 0 and the calibration slope being close to 1 (Figures 4 & 5). Under homogeneity in treatment, the c-statistic was similar to the target c-statistic of 0.80 at derivation across all scenarios (0.78 - 0.80) and the scaled Brier score lay between 0.20 and 0.21 for the Ignore Treatment approach (Tables E6 & E7).

3.2.2 | Increased proportion treated or treatment effect size at validation

For an increase in treatment proportion by 0.05, the predicted risks were well-calibrated (Figures 4 & 5). The calibration-in-the-large coefficient ranged from -0.06 to 0.05 and the calibration slope lay between 0.97 and 1.00 (Tables E10 & E11). For a medium increase in treatment proportion (i.e. difference of 0.4) or an increase in treatment effect size by 0.3 in terms of log-odds, the risks were on average overestimated (Figure 4). Overestimation was more pronounced for scenarios where the treatment proportion remained 0.9 and for cases with a large treatment effect size across derivation and validation (Figure D5). For an increase in treatment effect size overestimation was most profound for scenarios with a constant treatment proportion of 0.9 with a calibration-in-the-large coefficient of -0.30 (Table E10). The largest extent of overestimation was observed for an increase in treatment proportion from 0.1 to 0.5 with a constant large treatment effect size (0.8) with a median calibration-in-the-large coefficient of -0.39 (Table E11).

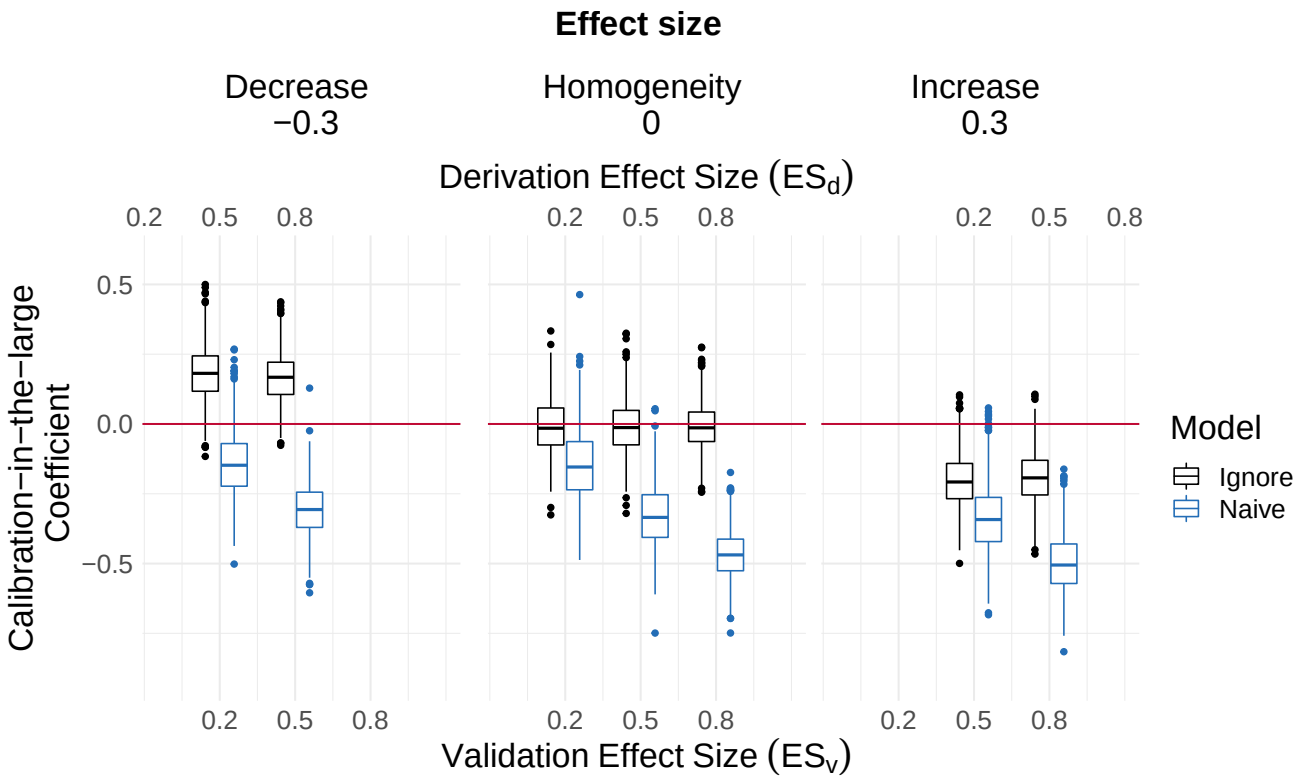
Increase in treatment proportion from derivation to validation resulted in too extreme predicted risks compared to observed proportions (calibration slope < 1). When there was an increase in treatment proportion, predicted risks were too extreme if the treatment proportion was 0.5 in the validation data set and too moderate compared to observed proportions if the treatment proportion was 0.5 in derivation population (Figure 5). These patterns were more distinct the larger the underlying treatment effect size was (Figure D6). The extent of predicted risks being too extreme or too moderate compared to observed proportions was small as the calibration slope only ranged from 0.95 to 1.04 across all scenarios with an increase in either treatment characteristic.

The c-statistic was reduced for scenarios with strong treatment effect (0.8), a derivation treatment proportion of 0.5 and increase in treatment proportion (Figure D7). An increase in the effect size led to slightly poorer performance regarding discrimination for a constant treatment proportion of 0.5 (Figure D7). The scaled Brier score was larger for scenarios with an increase in treatment effect size or treatment proportion corresponding to worse overall performance (Figure D8). In Study 1, results showed a smaller scaled Brier score for an increase in either treatment property.

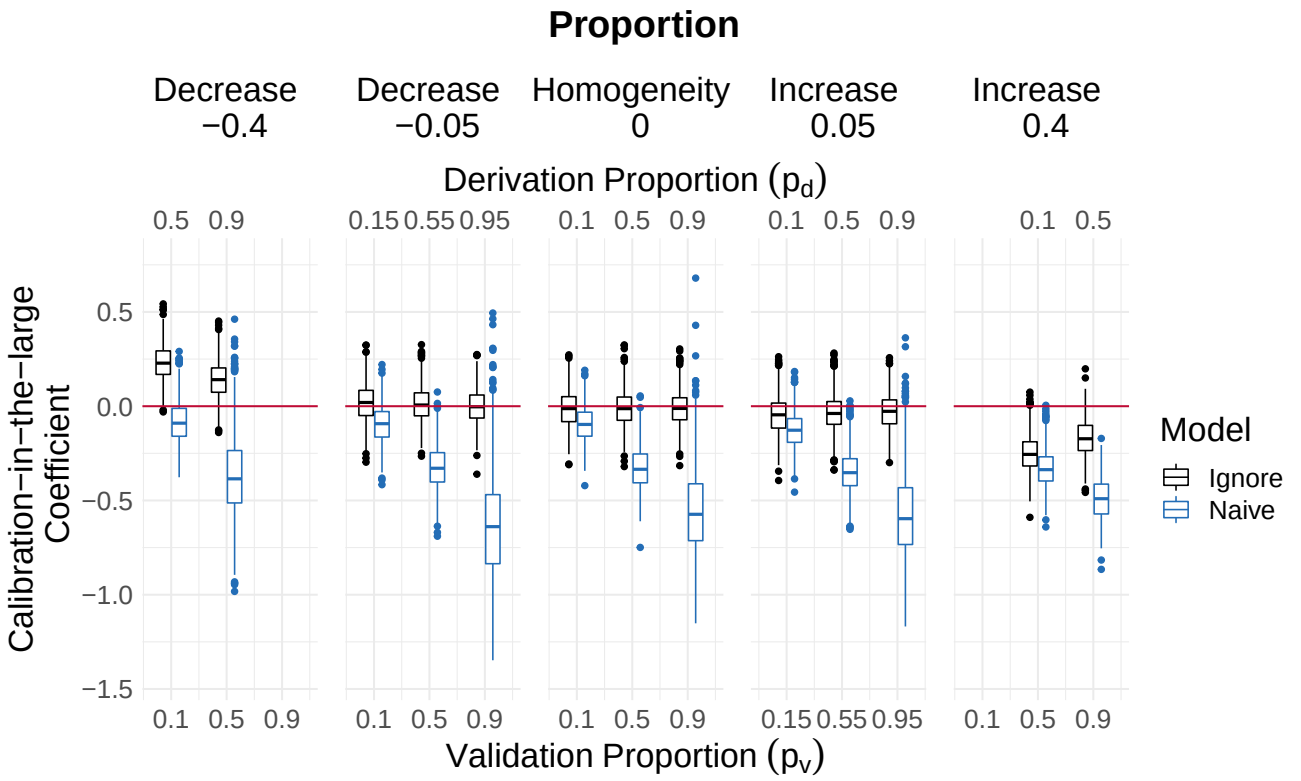
3.2.3 | Decreased proportion treated or treatment effect size at validation

As for increase in treatment characteristics, there only was misestimation of risks for a decrease in treatment proportion of 0.4. For a decrease by 0.05 the predicted risks were well-calibrated (Figures 4 & 5). For the decrease in treatment proportion of 0.4 as well as for a decrease in treatment effect size, the risks of the event were on average underestimated (Figure 4). For scenarios where the treatment proportion was 0.9 and for cases with large treatment effect sizes across derivation and validation underestimation was more distinct (Figure D5). Decrease in treatment proportion from derivation to validation led to too moderate predicted risks compared to observed proportions (Figure 5). When there was a decrease in treatment proportion, the same patterns as described for an increase in treatment proportion were present. The calibration slope lay between 0.93 and 1.03 across all scenarios with a decrease in either treatment characteristic (Tables E10 & E11).

Better discrimination performance was observed for scenarios with a decrease in treatment effect size (Figure D7). When there was a decrease in treatment proportion, discrimination tended to be worse for scenarios with stronger treatment effect and a derivation treatment proportion of 0.5 (Figure D7). For a decrease in treatment effect size or treatment proportion, the overall performance was better compared to scenarios with homogeneity in received treatment (Figure D8). As pointed out above, this presents a difference to the results in Study 1.

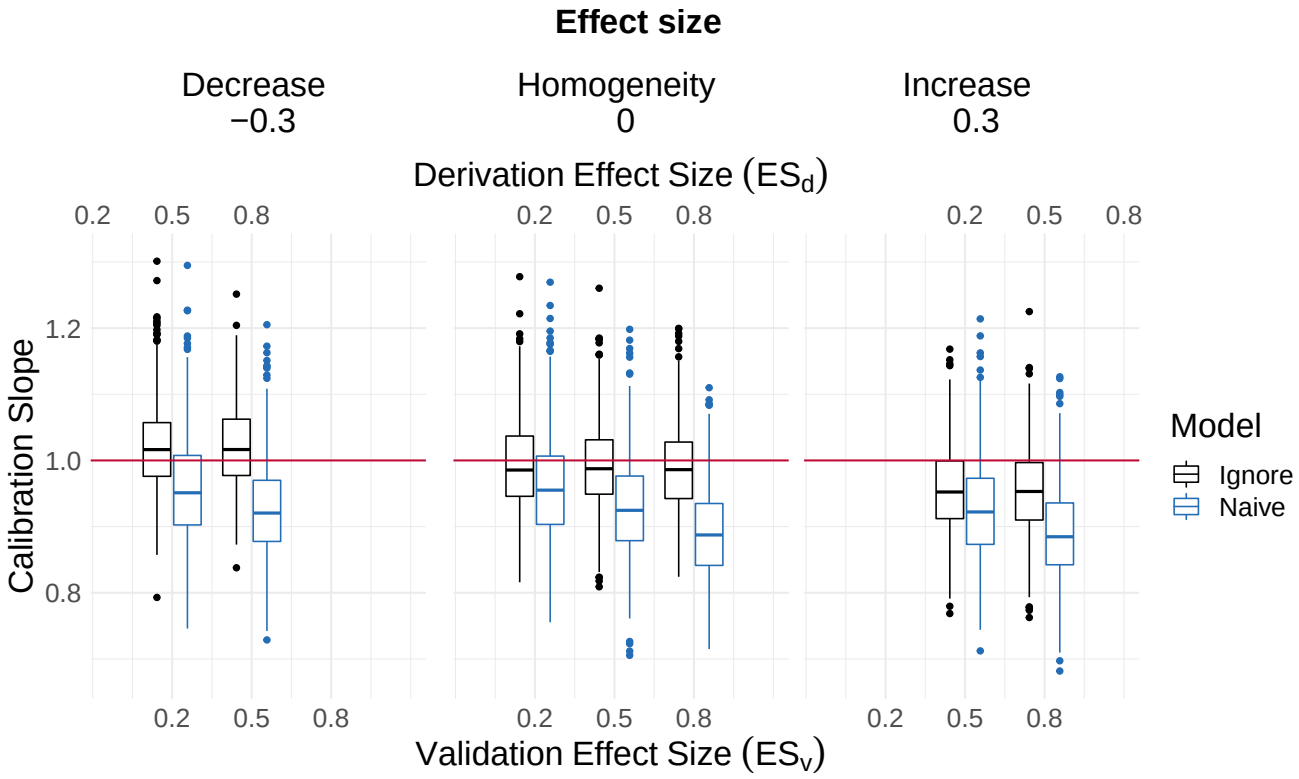


(a)

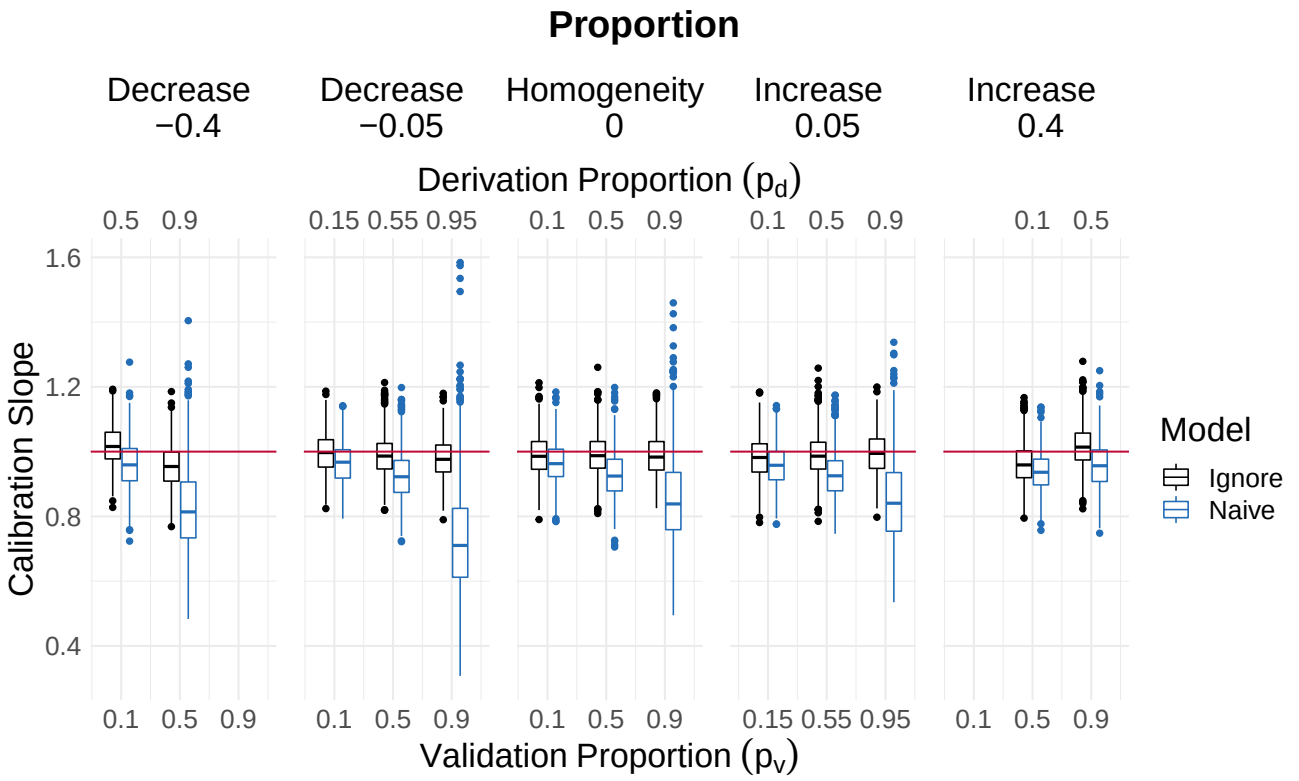


(b)

Figure 4 Study 2 - Calibration-in-the-large coefficient for (a) varying or equal treatment effect and constant treatment proportion of 0.5 and (b) varying or equal treatment proportion and constant medium treatment effect (0.5).



(a)



(b)

Figure 5 Study 2 - Calibration slope for (a) varying or equal treatment effect and constant treatment proportion of 0.5 and (b) varying or equal treatment proportion and constant medium treatment effect (0.5).

The presented results are in agreement with theoretical considerations. For the Ignore Treatment approach, the observed effects can be explained by the fact that at derivation the reduced risk of a patient treated with an effective treatment is not modeled accordingly. By assuming similar treatment policies, we assume to have the same risk reducing effect in the validation population. However, if less patients are treated at validation, this is not the case and on average the risk is underestimated. The same explanation holds if the treatment is less effective at validation. Vice versa, the risks are overestimated in case a larger proportion receives treatment or a more effective treatment is provided at validation as the risk reduction is greater than modeled at derivation. For the Treatment-Naïve approach, the model is derived solely on untreated patients. If patients in the validation data set were treated with an effective, thus, risk reducing treatment, the risk of each of those treated patients is overestimated leading to an overestimation on average. The more people receive treatment (proportion) and the more risk-reducing the treatment is (effect size), the larger the extent of that overestimation.

4 | DISCUSSION

The simulation studies illustrated that the predictive performance of a prognostic model at external validation was affected when the proportion that received treatment or the treatment effect size in a validation setting differed from the treatment received in the derivation setting. The effects of heterogeneity in received treatment on predictive performance were different for the Treatment-Naïve and the Ignore Treatment modeling approach. The Treatment-Naïve approach (i.e. using only untreated individuals for model derivation) provided miscalibrated predicted risks even for homogeneity in received treatment. The predicted risks obtained overestimated the risks of an event and were too extreme compared to the observed proportions. The Ignore Treatment approach (i.e. the treatment variable was not incorporated at any point of modeling), performed well regarding overall performance and discrimination and provided well-calibrated risks for treatment homogeneity and small differences in treatment proportions of 0.05 across derivation and validation. In scenarios with heterogeneity in treatment effect size or greater heterogeneity in treatment proportion, miscalibration was observed. The direction of misestimation of predicted risks depended on whether there was an increase or decrease in either treatment property. These results showed that heterogeneity in received treatment should be considered as one of the underlying reasons of predicted risks becoming invalid in the context of the Prediction Paradox.

Our findings underscored the theoretical considerations presented by Pajouheshnia and colleagues[4] and Groenwold and colleagues[8]. We did find that the predictive performance in case of an increase in treatment for the Ignore Treatment approach was affected[4] and that predicted risk using the Treatment-Naïve approach seem to overestimate the risk of the event of interest[8]. Beyond the theoretical considerations, we were able to provide a systematic overview over the extents of effects across different performance measures and emphasize that also underestimation of risks can occur for the Ignore Treatment approach. The suggested requirement of similar treatment policies[10] resulted in well-calibrated models for the Ignore Treatment approach. We also showed that small differences in treatment proportion do barely reduce performance regarding calibration. However, in accordance with previous studies[4,9], similar treatment policies did not guarantee well-calibrated risks in case of the Treatment-Naïve approach.

We point out that the implementation of the scaled Brier score we used followed the definition by Steyerberg and colleagues[19] using the predicted risks for scaling. However, in other implementations the observed risks are used for scaling, e.g., in the `psfmi`[24] package. A limitation of the presented study is that the small sample size study is only representative of a specific setting for realistic prognostic models (with 8 correlated predictors with equal effect size and an outcome proportion of 0.2). Additionally, it is likely that the ratio of effects of the predictor on the outcome and of the treatment on the outcome matters for strength of effects described. This has to be kept in mind when drawing conclusions from the aforementioned results. Another limitation is that only a setting where all predictors which were used for data generation were used at model derivation. As commonly not all predictors of the outcome are available for model derivation (e.g., due to costly measurements)[8], this constitutes a limitation of this study.

Future research could explore the impact and its extent of heterogeneity in received treatment in a greater variety of settings. For example, scenarios in which a time component is added and treatment is started after baseline could be considered. It could also be of interest to perform simulation studies which closely mirror a specific clinical setting. Other aspects of received treatment (e.g., heterogeneity in treatment effect size) could also be included in future studies.

Conclusion

We investigated to what extent predictive performance of prognostic models is influenced by heterogeneity in treatment effect size and treatment proportion across derivation and validation using a large and a small sample size simulation study. Miscalibration and variance in predictive performance was found for heterogeneity in treatment effect size and treatment proportion between the derivation and (external) validation population. The impact of heterogeneity in received treatment on predictive performance differed for the two considered modeling approaches, Ignore Treatment and Treatment-Naïve. Implications from the presented results are that discussing the possible presence of treatment heterogeneity between the derivation and later application target populations is advised to be part of the data preparation phase when developing a new prognostic model. Furthermore, stakeholders should be aware of the over- and underestimation of predicted risks occurring depending on the extent and direction of treatment heterogeneity and the used modeling approach.

References

1. Steyerberg EW. *Clinical prediction models*. Springer New York . 2009.

2. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative . Calibration: the Achilles heel of predictive analytics. *BMC medicine* 2019; 17(1).

3. Pajouheshnia R, Damen JAAG, Groenwold RHH, Moons KGM, Peelen LM. Literate Programming. *Diagnostic and Prognostic Research* 2017; 1(15).

4. Pajouheshnia R, Peelen LM, Moons KGM, Reitsma JB, Groenwold RHH. Accounting for treatment use when validating a prognostic model: a simulation study. *BMC medical research methodology* 2017; 17(1).

5. Sperrin M, Jenkins D, Martin GP, Peek N. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success.. *Journal of the American Medical Informatics Association : JAMIA* 2019; 26(12): 1675–1676. doi: https://doi.org/10.1093/jamia/ocz197

6. Cheong-See F, Allotey J, Marlin N, et al. Prediction models in obstetrics: understanding the treatment paradox and potential solutions to the threat it poses. *BJOG: An International Journal of Obstetrics & Gynaecology* 2016; 123(7): 1060-1064. doi: https://doi.org/10.1111/1471-0528.13859

7. Peek N, Sperrin M, Mamas M, van Staa T, Buchan I. Rapid Response: Hari Seldon, QRISK3, and the Prediction Paradox. 2017. *BMJ*, Available online at: https://www.bmj.com/content/357/bmj.j2099/rr-0, last accessed on 04.05.2023.

8. Groenwold RH, Moons KG, Pajouheshnia R, et al. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *Journal of Clinical Epidemiology* 2016; 78: 90-100. doi: https://doi.org/10.1016/j.jclinepi.2016.03.017

9. Sperrin M, Martin G, Staa T, Peek N, Buchan I. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. *Statistics in medicine* 2017; 37. doi: 10.1002/sim.7913

10. van Geloven N, Swanson S, Ramspek C, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *European Journal of Epidemiology* 2020; 35: 619-630. doi: https://doi.org/10.1007/s10654-020-00636-1

11. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019; 38(11): 2074-2102.

12. Pencina MJ, D'Agostino RB, Pencina KM, Janssens ACJW, Greenland P. Interpreting Incremental Value of Markers Added to Risk Prediction Models. *American Journal of Epidemiology* 2012; 176(6): 473–481. doi: https://doi.org/10.1093/aje/kws207

13. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020; 368. doi: 10.1136/bmj.m441

14. Riley RD, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Statistics in Medicine* 2021; 40(19): 4230-4251. doi: https://doi.org/10.1002/sim.9025

15. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures.. *Epidemiology (Cambridge, Mass.)* 2010; 21(1): 128–138.

16. Agresti A. *An Introduction to Categorical Data Analysis*. John Wiley Sons, Hoboken, New Jersey, USA. second ed. 2007.

17. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 1950; 78(1): 1 – 3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

18. Harrell Jr FE. *rms: Regression Modeling Strategies*. 2022. R package version 6.3-0.

19. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures.. *Epidemiology* 2010; 21(1): 128-138. doi: https://doi.org/10.1097/EDE.0b013e3181c30fb2

20. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2022.

21. Wickham H, François R, Henry L, Müller K. *dplyr: A Grammar of Data Manipulation*. 2022. R package version 1.0.10.

22. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York . 2016.

23. Hofert M, Mächler M. Parallel and Other Simulations in R Made Easy: An End-to-End Study. *Journal of Statistical Software* 2016; 69(4): 1–44. doi: 10.18637/jss.v069.i04

24. Heymans M. *psfmi: Prediction Model Selection and Performance Evaluation in Multiple Imputed Datasets*. 2020. R package version 0.5.0.

25. Ensor J, Martin EC, Riley RD. *pmsampsize: Calculates the Minimum Sample Size Required for Developing a Multivariable Prediction Model*. 2022. R package version 1.1.2.

26. Snell KI, Archer L, Ensor J, et al. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *Journal of Clinical Epidemiology* 2021; 135: 79-89. doi: https://doi.org/10.1016/j.jclinepi.2021.02.011

APPENDIX

A SCENARIOS

A.1 Study 1

Table A1 Treatment effect size and treatment proportion at derivation and validation for Study 1; each treatment effect size row is combined with each treatment proportion row where the effect sizes are given as absolute values of the log-odds in parentheses; resulting in $N_1 = 81$ scenarios

	Derivation		Validation
Treatment effect size			
Homogeneity	small (0.20)	→	small (0.20)
	medium (0.50)	→	medium (0.50)
	large (0.80)	→	large (0.80)
Medium increase	small (0.20)	→	medium (0.50)
(0.3)	medium (0.50)	→	large (0.80)
Large increase (0.6)	small (0.20)	→	large (0.80)
Medium decrease	large (0.80)	→	medium (0.50)
(-0.3)	medium (0.50)	→	small (0.20)
Large decrease (-0.6)	large (0.80)	→	small (0.20)
Treatment proportion			
Homogeneity	0.10	→	0.10
	0.50	→	0.50
	0.90	→	0.90
Medium increase	0.10	→	0.50
(0.4)	0.50	→	0.90
Large increase (0.8)	0.10	→	0.90
Medium decrease	0.50	→	0.10
(-0.4)	0.90	→	0.50
Large decrease (-0.8)	0.90	→	0.10
Targets			
C-statistic	0.80		
Outcome proportion	0.50		

A.2 Study 2

Table A2 Treatment effect size and treatment proportion at derivation and validation for Study 2; each treatment effect size row is combined with each treatment proportion row where the effect sizes are given as log-odds in parentheses; resulting in $N_2 = 91$ scenarios

	Derivation		Validation
Treatment effect size			
Homogeneity	small (0.20)	→	small (0.20)
	medium (0.50)	→	medium (0.50)
	large (0.80)	→	large (0.80)
Medium increase	small (0.20)	→	medium (0.50)
(0.3)	medium (0.50)	→	large (0.80)
Medium decrease	large (0.80)	→	medium (0.50)
(-0.3)	medium (0.50)	→	small (0.20)
Treatment proportion			
Homogeneity	0.10	→	0.10
	0.50	→	0.50
	0.90	→	0.90
Small increase	0.10	→	0.15
(0.05)	0.50	→	0.55
	0.90	→	0.95
Medium increase	0.10	→	0.50
(0.4)	0.50	→	0.90
Small decrease	0.15	→	0.10
(-0.05)	0.55	→	0.50
	0.95	→	0.90
Medium decrease	0.50	→	0.10
(-0.4)	0.90	→	0.50
Targets			
C-statistic	0.80		
Outcome proportion	0.20		

B OPTIMIZATION

B.1 Procedure

A large sample size of 100, 000 was used for all optimization procedures.

The aim was to obtain a set of values for the intercept *a* in Equation 1 leading to the respective treatment proportions for each of the simulation scenarios. The intercept *a* was optimized 1, 000 times for each treatment proportion using the `optimize` function from the `stats`[20] package. Twice, treatment variables for 1, 000 individuals each were obtained using the optimized parameters. One set used the means over the obtained optimized parameter across all iterations. The other set used the one optimized parameter across all iteration which led to a treatment proportion closest to the targeted treatment proportion. The minimum difference between the targeted treatment proportions and the mean treatment proportion obtained for both sets was calculated. For each treatment proportion, the optimized intercepts used to obtain the minimum difference were used as the final optimized intercepts (Tables B3 & B4).

For the second optimization, the intercept (β_0) and the predictor effect size (β_2) in Equation 2 were optimized such that the targeted c-statistic of 0.80 and a outcome proportion of 0.20 at derivation were reached. Using the `optimize` function from the `stats` [20] package for the optimization of β_0 and β_2 was not possible as the algorithm did not converge. Instead both parameters were optimized following a grid approach. The effect size was constrained to be the same for all predictors. Three rounds of grid searches were performed. In the first round a broader grid for both parameters with larger steps was executed five times with different seeds. For the two following rounds, the grids were adapted manually. This included setting the limits of the grid closer to the optimal results obtained in the previous grid and decreasing the step-size. This was done for each scenario separately. After the last grid search, the optimized parameters for β_0 and β_2 were saved for each of the simulation scenarios (Tables B3 & B4). The success of the optimization was evaluated on the created data set for the simulation study (Table B5). The details on used grids are included in the simulation study code available at https://github.com/emsulo/Master-Thesis.

B.2 Optimized coefficients

Study 1

Table B3 Optimized coefficients for Study 1 with targeted c-statistic = 0.80 and outcome proportion of 0.50, for the proportion treated at derivation (p_d) and the effect size as log-odds at derivation (ES_d) including the optimized a, β_0, and β_2 for each derivation setting. Treatment parameters for validation sets (p_v and ES_v) take on the same values as p_d and ES_d.

Scenario	p_d	a	$ES_d(\beta_1)$	β_0	β_2
1	0.10	-2.55	-0.20	0.04	1.45
2	0.50	-0.00	-0.20	0.12	1.45
3	0.90	2.58	-0.20	0.32	1.52
4	0.10	-2.55	-0.50	0.07	1.47
5	0.50	-0.00	-0.50	0.24	1.54
6	0.90	2.58	-0.50	0.49	1.45
7	0.10	-2.55	-0.80	0.06	1.49
8	0.50	-0.00	-0.80	0.42	1.61
9	0.90	2.58	-0.80	0.70	1.48

Study 2

Table B4 Optimized coefficients for Study 2 with targeted c-statistic = 0.80 and outcome proportion of 0.20, for the proportion treated at derivation (p_d) and the effect size as log-odds at derivation (ES_d) including the optimized a, β_0, and β_2 for each derivation setting. Treatment parameters for validation sets (p_v and ES_v) take on the same values as p_d and ES_d according to Table A2.

Scenario	p_d	a	ES_d (β_1)	β_0	β_2
1	0.10	-2.56	-0.20	-1.83	0.31
2	0.15	-2.08	-0.20	-1.83	0.31
3	0.50	0.00	-0.20	-1.66	0.31
4	0.55	0.23	-0.20	-1.65	0.31
5	0.90	2.55	-0.20	-1.52	0.30
6	0.95	3.36	-0.20	-1.56	0.30
7	0.10	-2.56	-0.50	-1.84	0.32
8	0.15	-2.08	-0.50	-1.72	0.31
9	0.50	0.00	-0.50	-1.52	0.32
10	0.55	0.23	-0.50	-1.48	0.31
11	0.90	2.55	-0.50	-1.30	0.30
12	0.95	3.36	-0.50	-1.22	0.30
13	0.10	-2.56	-0.80	-1.84	0.32
14	0.15	-2.08	-0.80	-1.65	0.31
15	0.50	0.00	-0.80	-1.08	0.31
16	0.55	0.23	-0.80	-1.21	0.32
17	0.90	2.55	-0.80	-1.04	0.31
18	0.95	3.36	-0.80	-0.97	0.30

B.3 Descriptives of target characteristics after optimization

Table B5 Descriptives for the targeted characteristics of the derivation data sets for Study 1 and Study 2.

		Target	Mean	Median	Min	Max	SD
Study 1	outcome proportion	0.50	0.5039	0.5031	0.4936	0.5264	0.0082
	c-statistic	0.80	0.8021	0.8022	0.7933	0.8160	0.0045
Study 2	outcome proportion	0.20	0.2088	0.2075	0.1927	0.2464	0.0134
	c-statistic	0.80	0.8017	0.8015	0.7916	0.8139	0.0047

C REQUIRED SAMPLE SIZE CALCULATIONS

C.1 Derivation

We used the `pmsampsize` function from the `pmsampsize` package[25] to calculate the required minimal sample size for developing a prediction model according to Riley and colleagues[13]. The applied criteria were the following:

- binary outcome (`type = "b"`)
- c-statistic of 0.80 (`cstatistic = 0.80`)
- outcome rate of 0.20 (`prevalence = 0.20`)
- 8 parameters (`parameters = 8`)
- seed of 1705 (`seed = 1705`)
- desired shrinkage level of 0.9 (`shrinkage = 0.9`)

As the minimum required sample size for model derivation should also be reached if only the untreated 10% of the population are used for derivation under the Treatment-Naïve approach (treatment proportion = 0.9), the thereby obtained minimal required sample size of 350 was multiplied by 10. This resulted in a final derivation population sample size of $3,500$ for the small sample size study.

C.2 External Validation

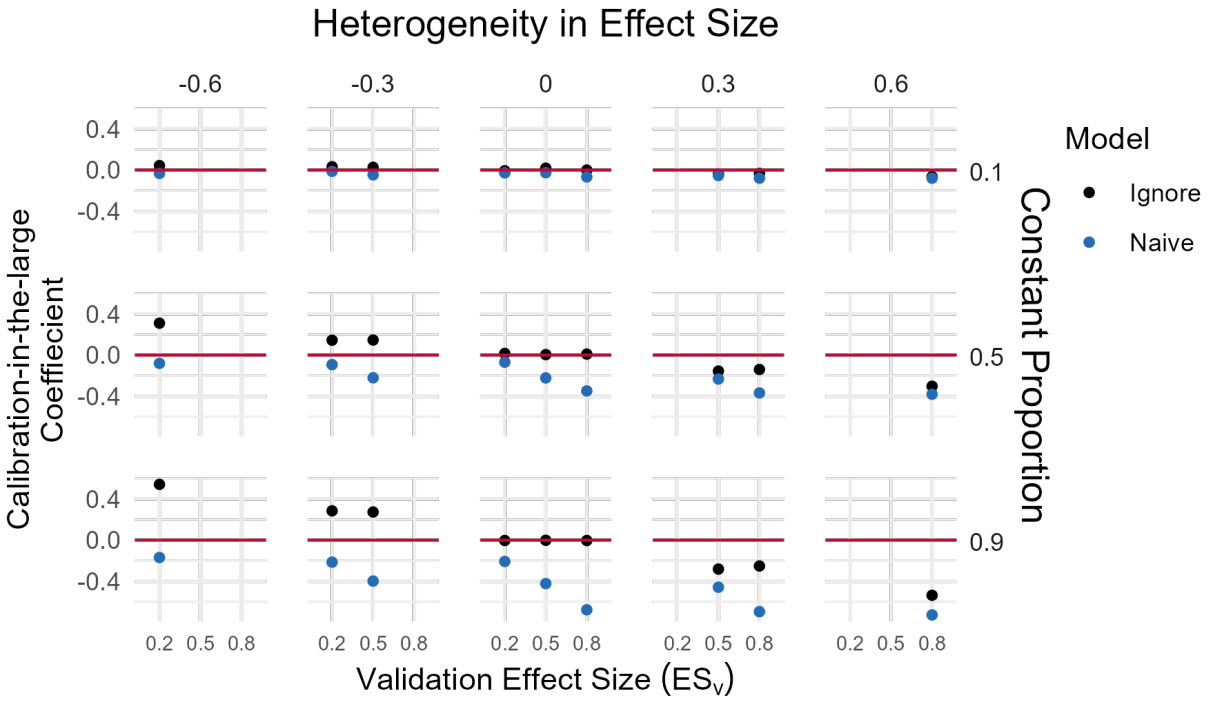
The procedure described by Riley and colleagues[14] to calculate the minimal required sample size for the eternal validation data set was used. We only executed the procedure for a targeted calibration slope and while this tends to give the highest required sample size, we are aware that there is no guarantee that the obtained required sample size regarding targets for calibration-in-the-large, discrimination, or clinical utility could not be larger[14].

We looked at the distributions of the linear predictors in form of histograms, and found the assumption of normality needed for the calculations to be justified in spite of some variation across simulation scenarios. Figure 1 in the paper by Riley and colleagues[14] was used to eye-ball the underlying distribution of the linear predictor and we conclude that the linear predictors follow a $\mathcal{N}(-1.39, 1)$ distribution. It is assumed that the model is well calibrated ("weak" calibration[26]) and the targeted standard error for the calibration slope was chosen as 0.05. Following the steps described in Box 1[14] and using $10,000$ hypothetical participants for Part (C) as well as Equation 7, we obtained a required minimal sample size of $3,527$ for the small sample size study.

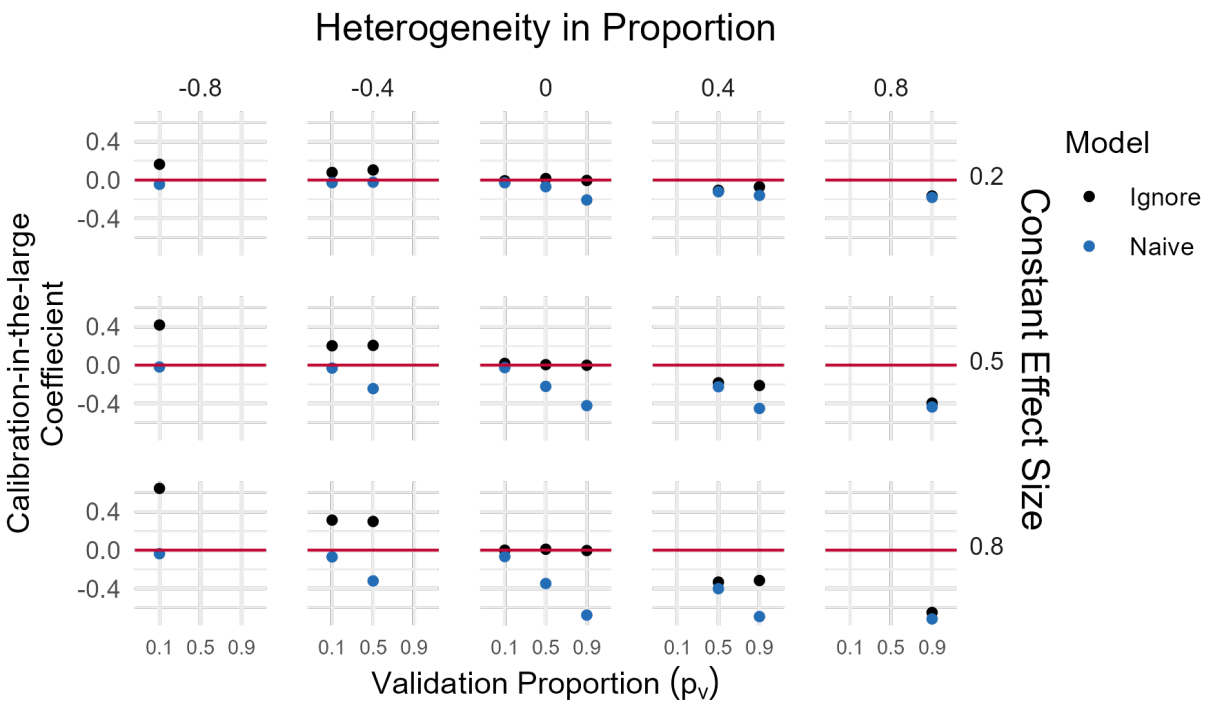
D RESULT VISUALIZATION

Below, the plots for the results for the calibration-in-the-large coefficient, calibration slope, c-statistic, and the scaled Brier score are provided for all considered underlying constant treatment characteristics. Visualizations of results for cases of heterogeneity in treatment effect size and proportion simultaneously in form of loop-plots can be found in the Shiny App (https://emilialoescher.shinyapps.io/6_visualization_app/). Those figures as well as the corresponding result tables are also included in the Supplementary Materials accessible at https://github.com/emsulo/Master-Thesis.

D.1 Study 1

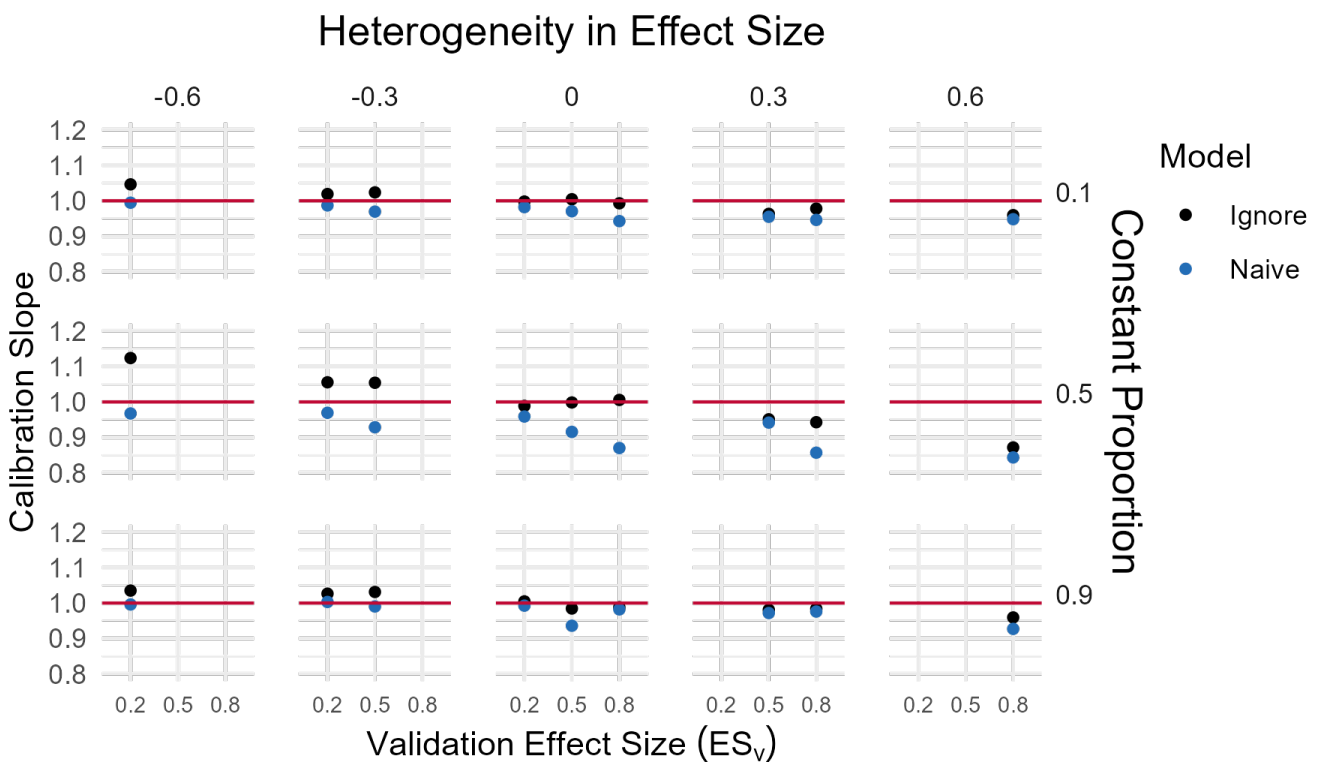


(a)

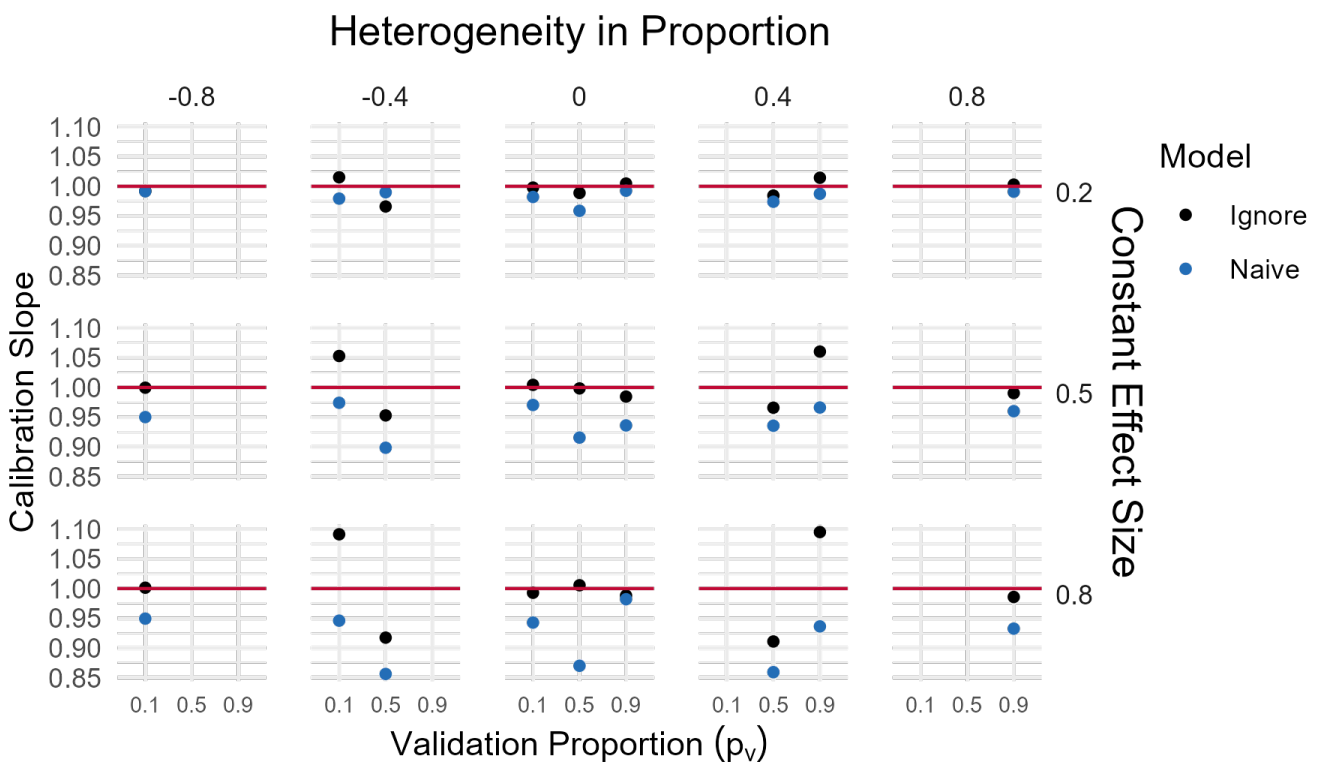


(b)

Figure D1 Calibration-in-the-large coefficient for the different cases for (a) varying or equal treatment effect and constant treatment proportion and (b) varying or equal treatment proportion and constant treatment effect.

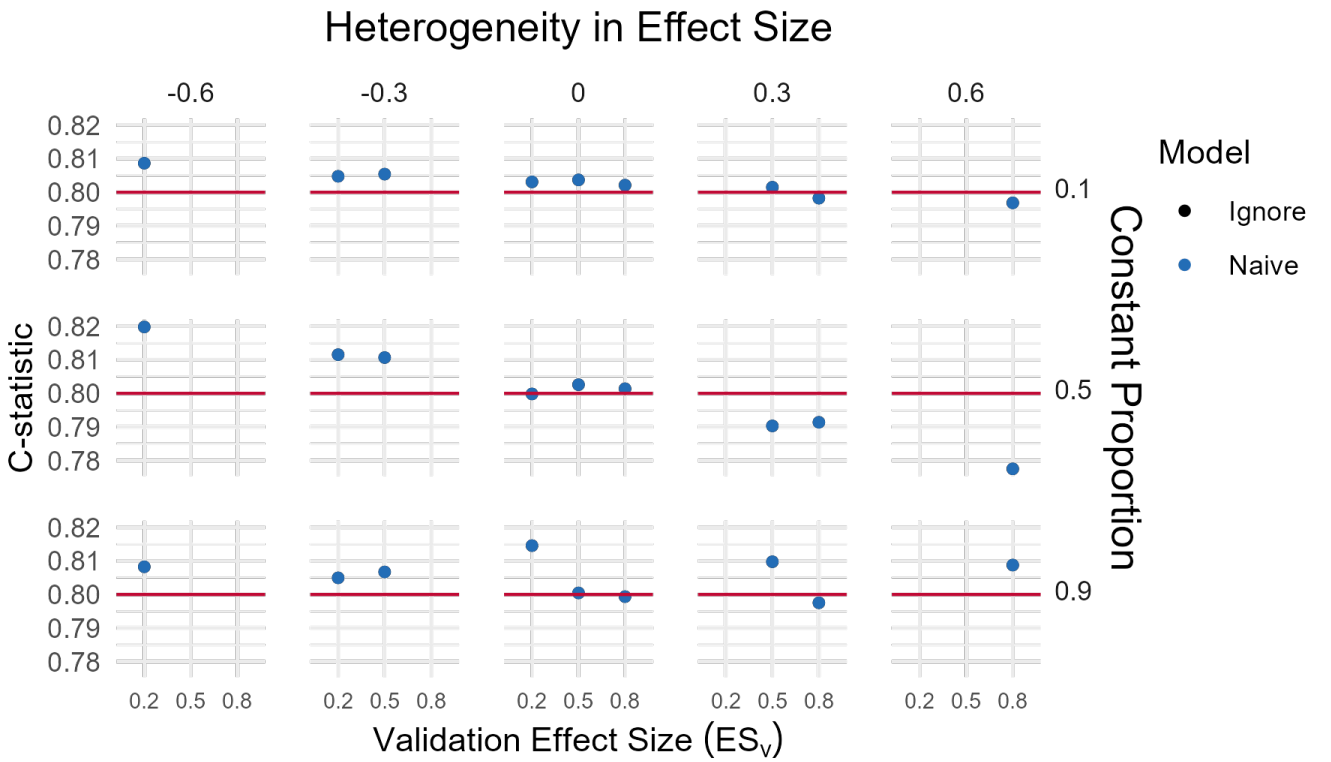


(a)

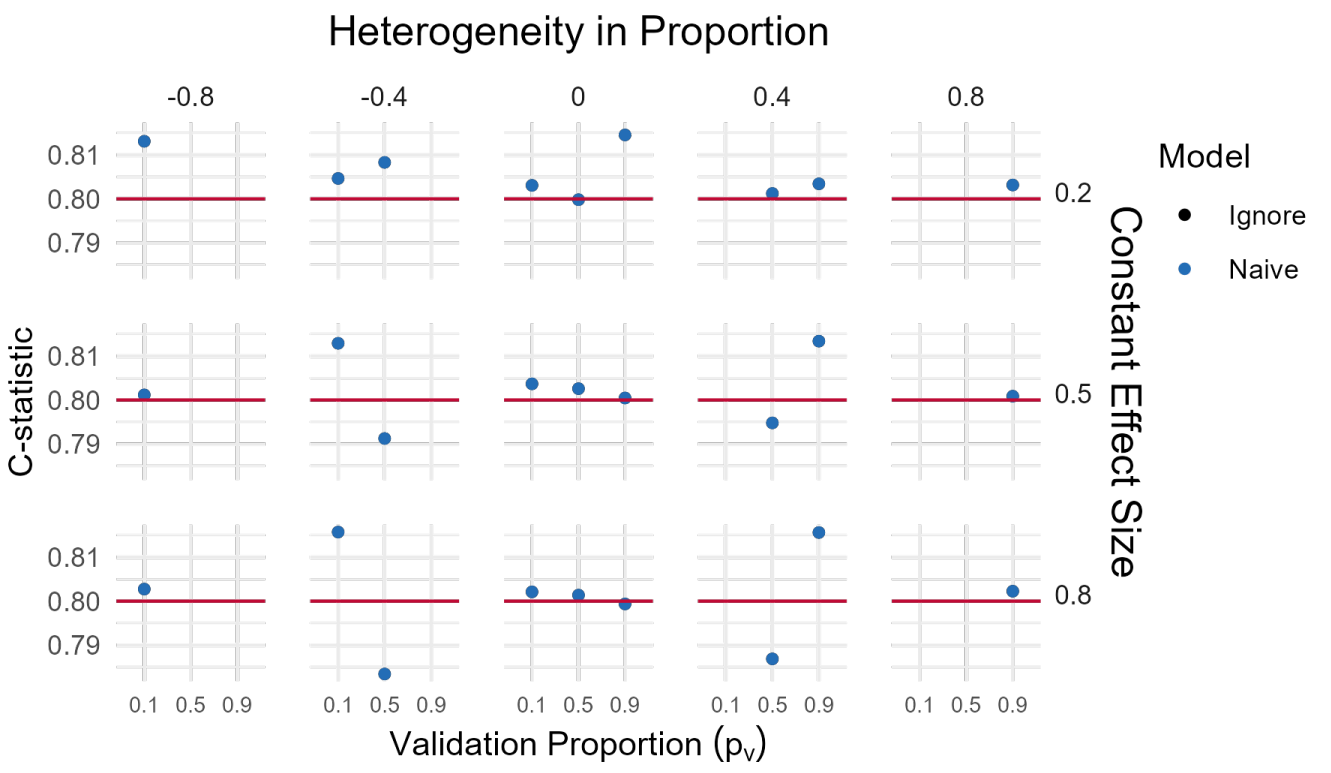


(b)

Figure D2 Calibration slope for the different cases for (a) varying or equal treatment effect and constant treatment proportion and (b) varying or equal treatment proportion and constant treatment effect.

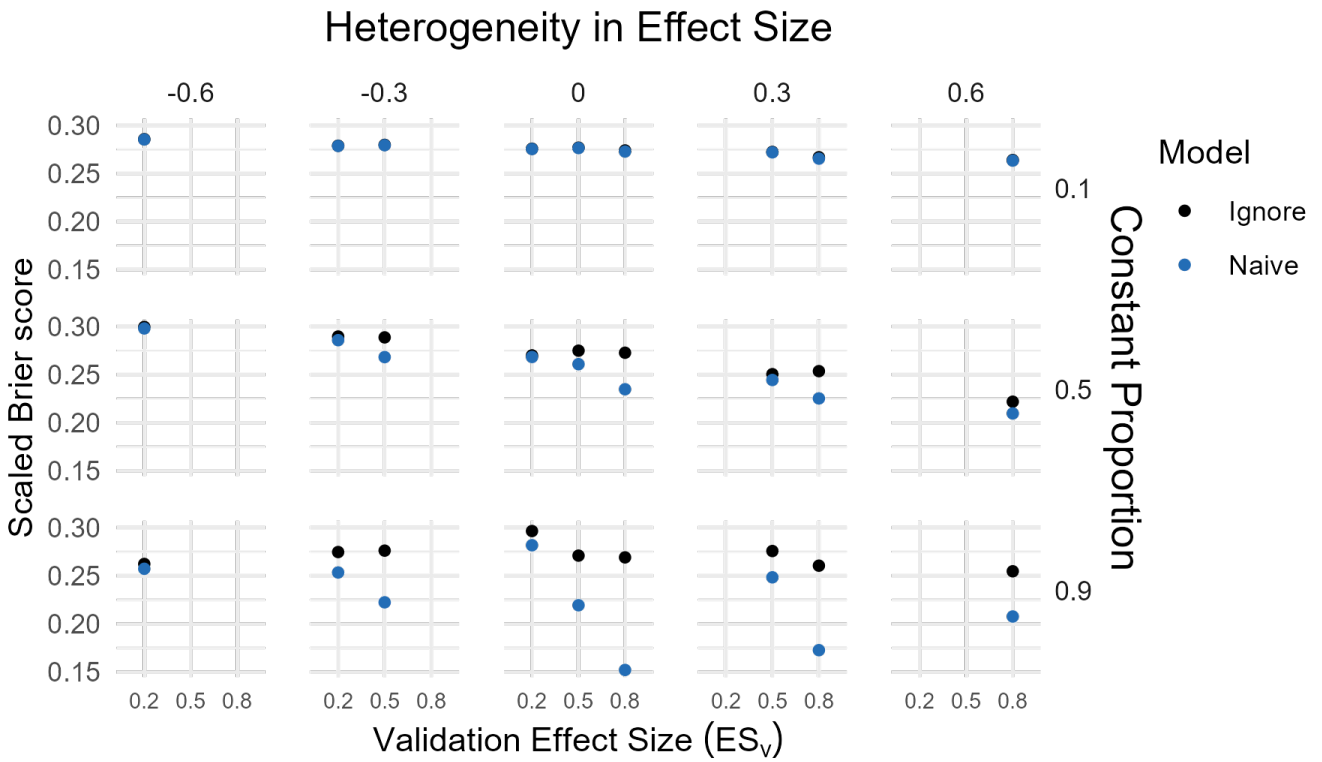


(a)

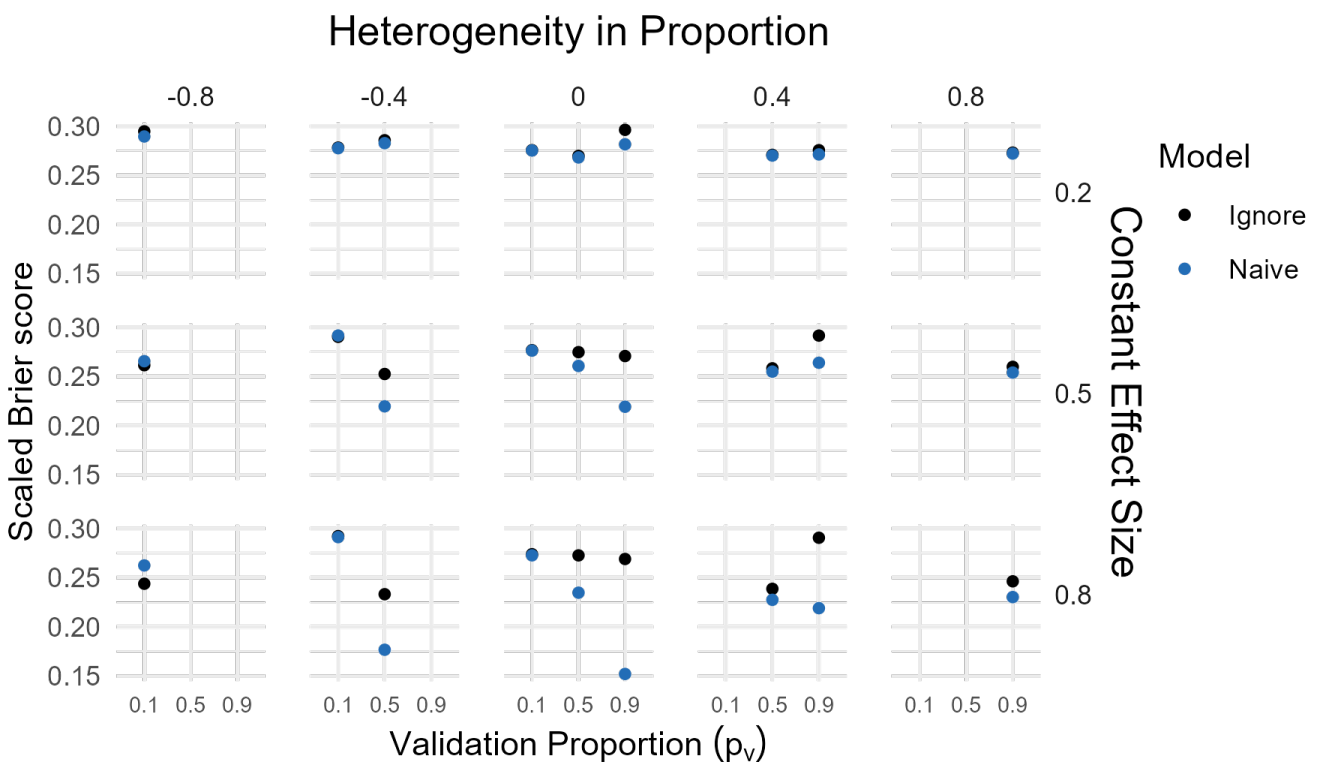


(b)

Figure D3 C-statistic for the different cases for (a) varying or equal treatment effect and constant treatment proportion and (b) varying or equal treatment proportion and constant treatment effect.



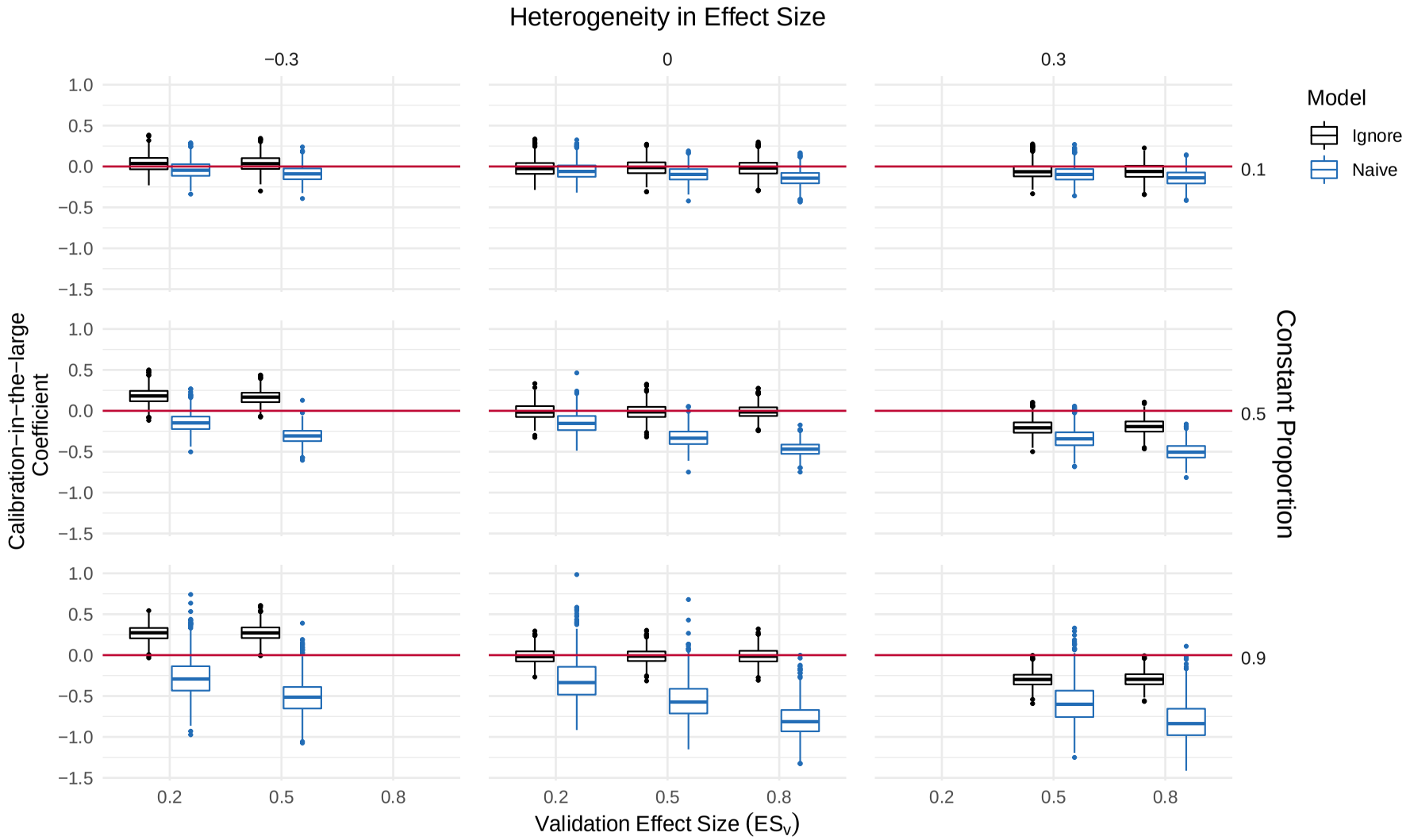
(a)



(b)

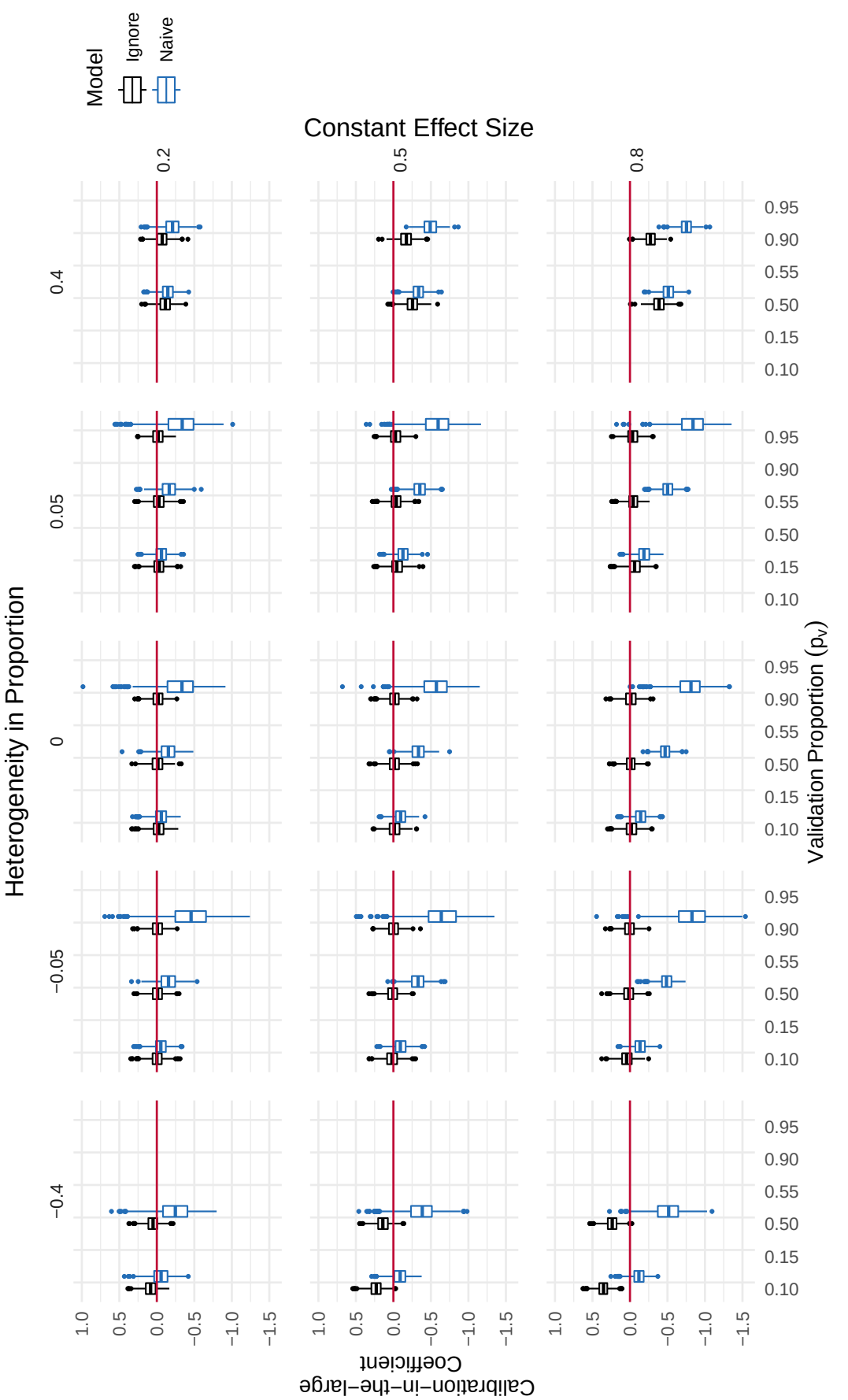
Figure D4 Calibration slope for the different cases for (a) varying or equal treatment effect and constant treatment proportion and (b) varying or equal treatment proportion and constant treatment effect.

D.2 Study 2



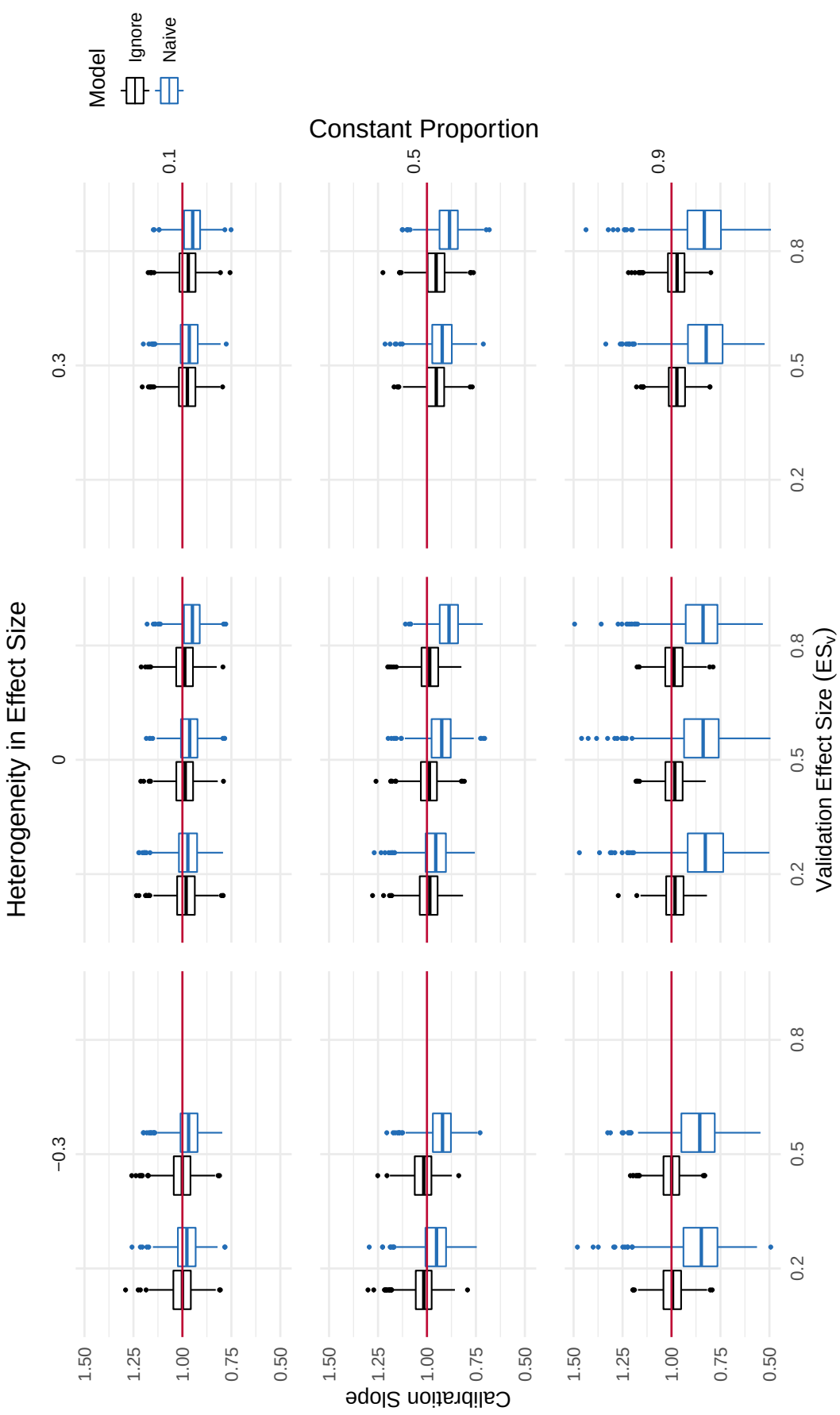
(a)

Figure D5 Calibration-in-the-large coefficient for the different cases for (a) varying or equal treatment effect and constant treatment proportion.



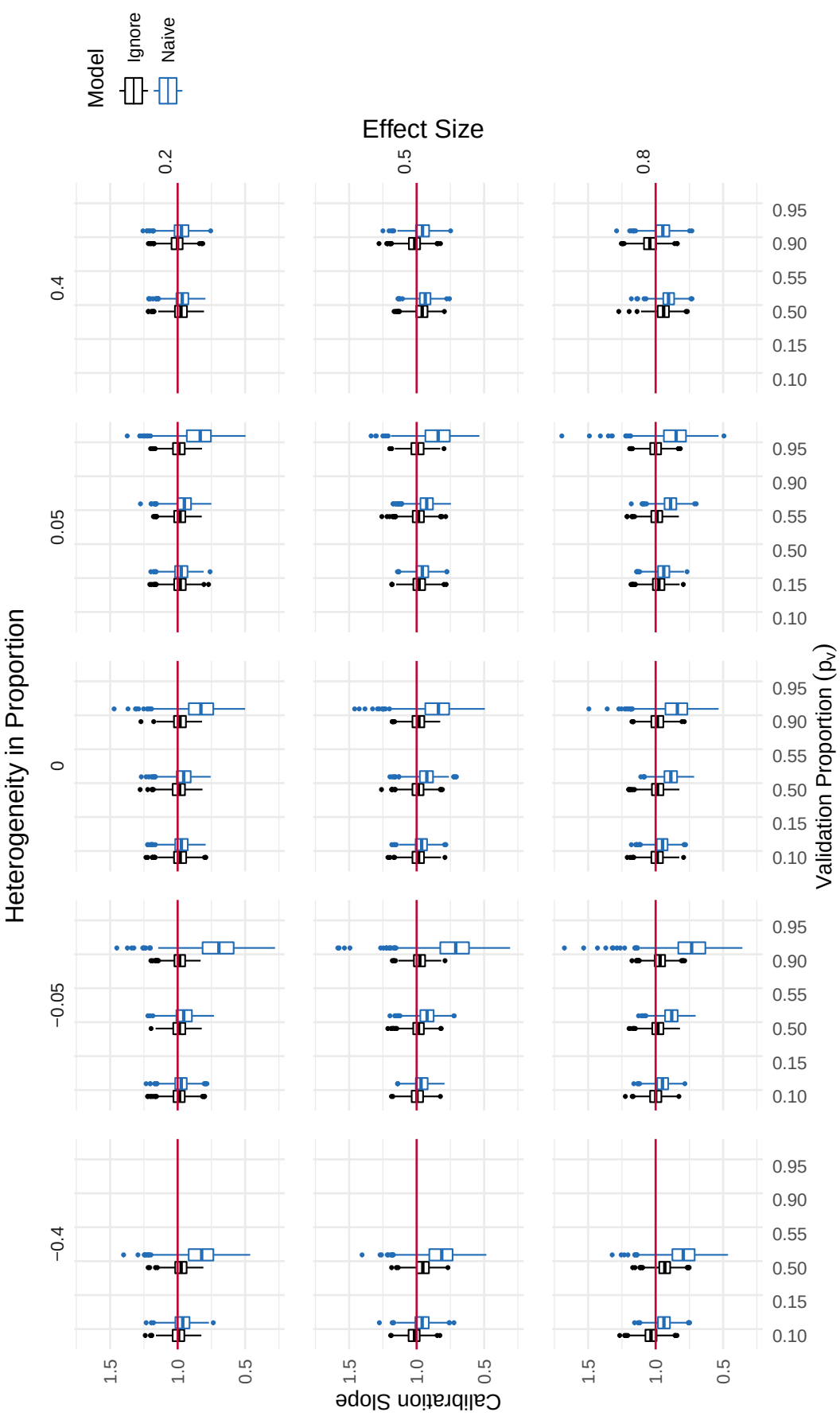
(b)

Figure D5 Calibration-in-the-large coefficient for the different cases for (b) varying or equal treatment proportion and constant treatment effect.



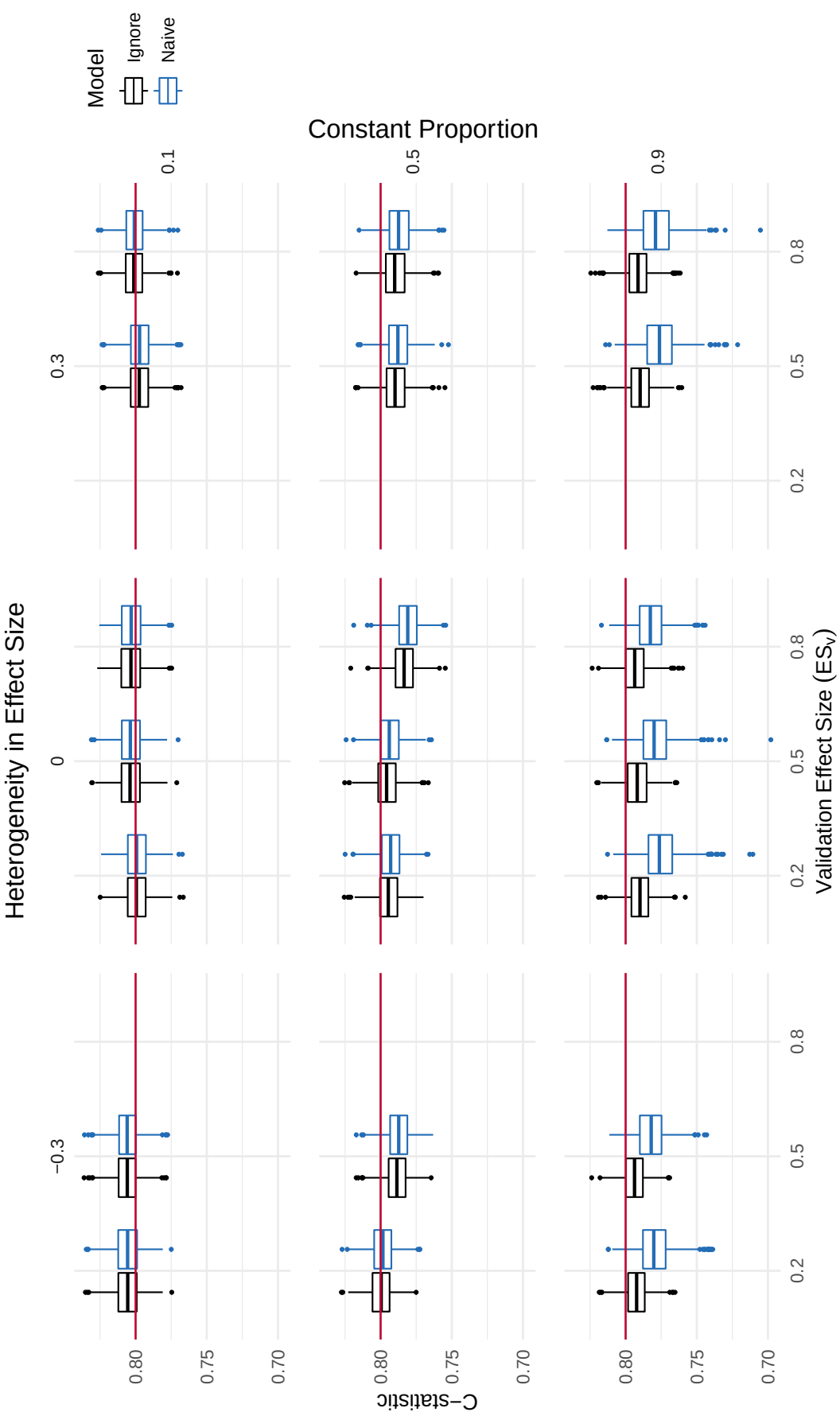
(a)

Figure D6 Calibration slope for the different cases for (a) varying or equal treatment effect and constant treatment proportion.



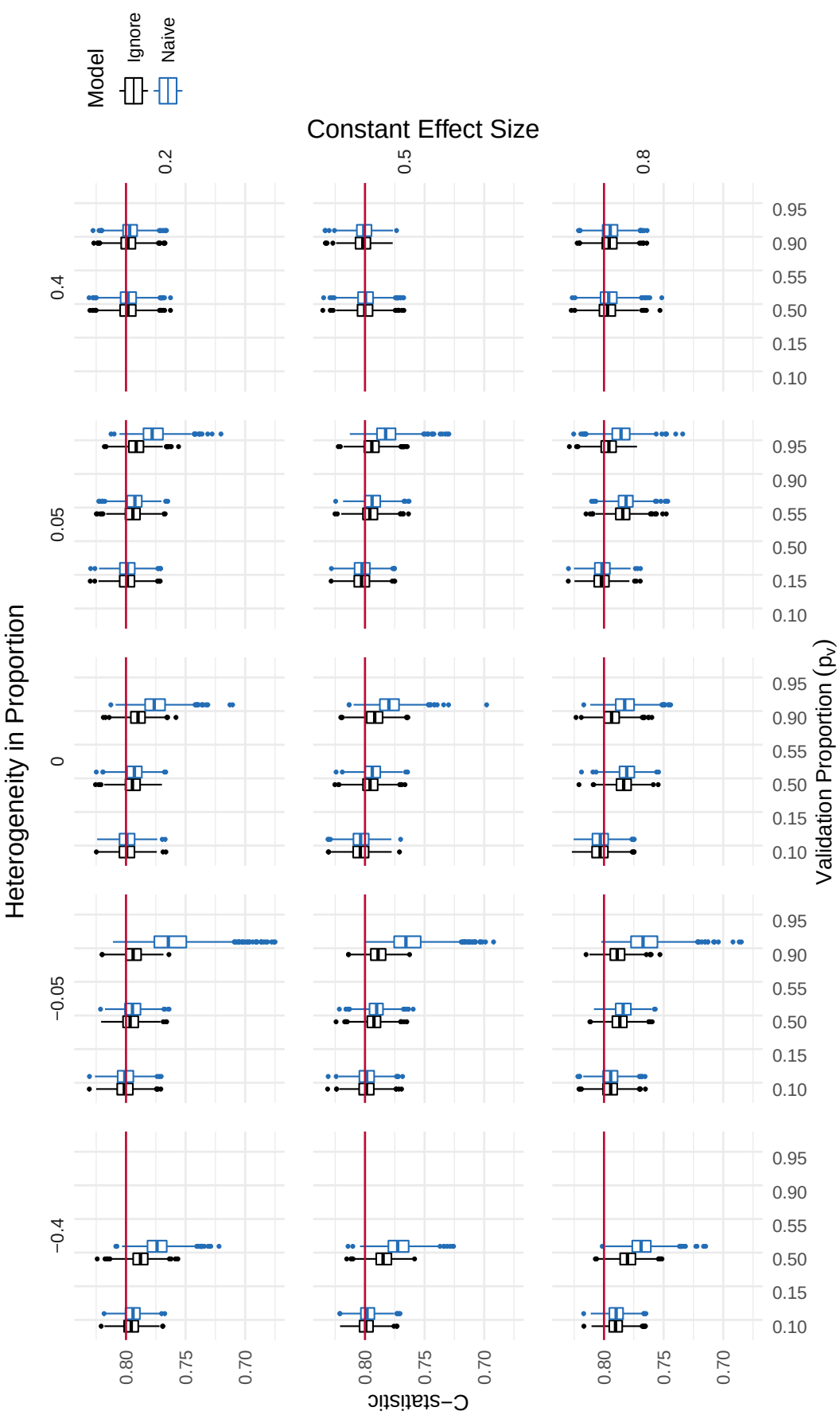
(b)

Figure D6 Calibration slope for the different cases for (b) varying or equal treatment proportion and constant treatment effect.



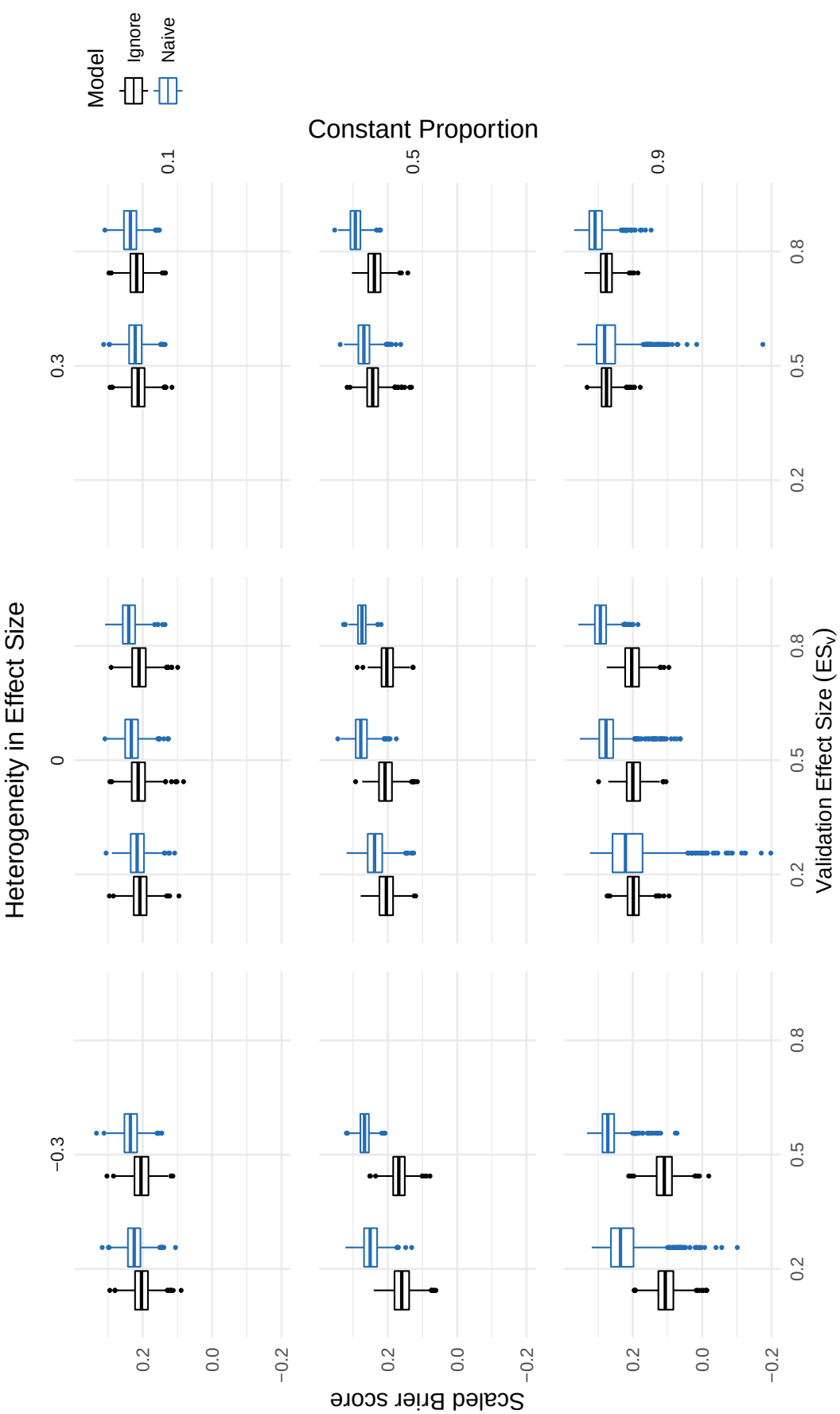
(a)

Figure D7 C-statistic for the different cases for (a) varying or equal treatment effect and constant treatment proportion.



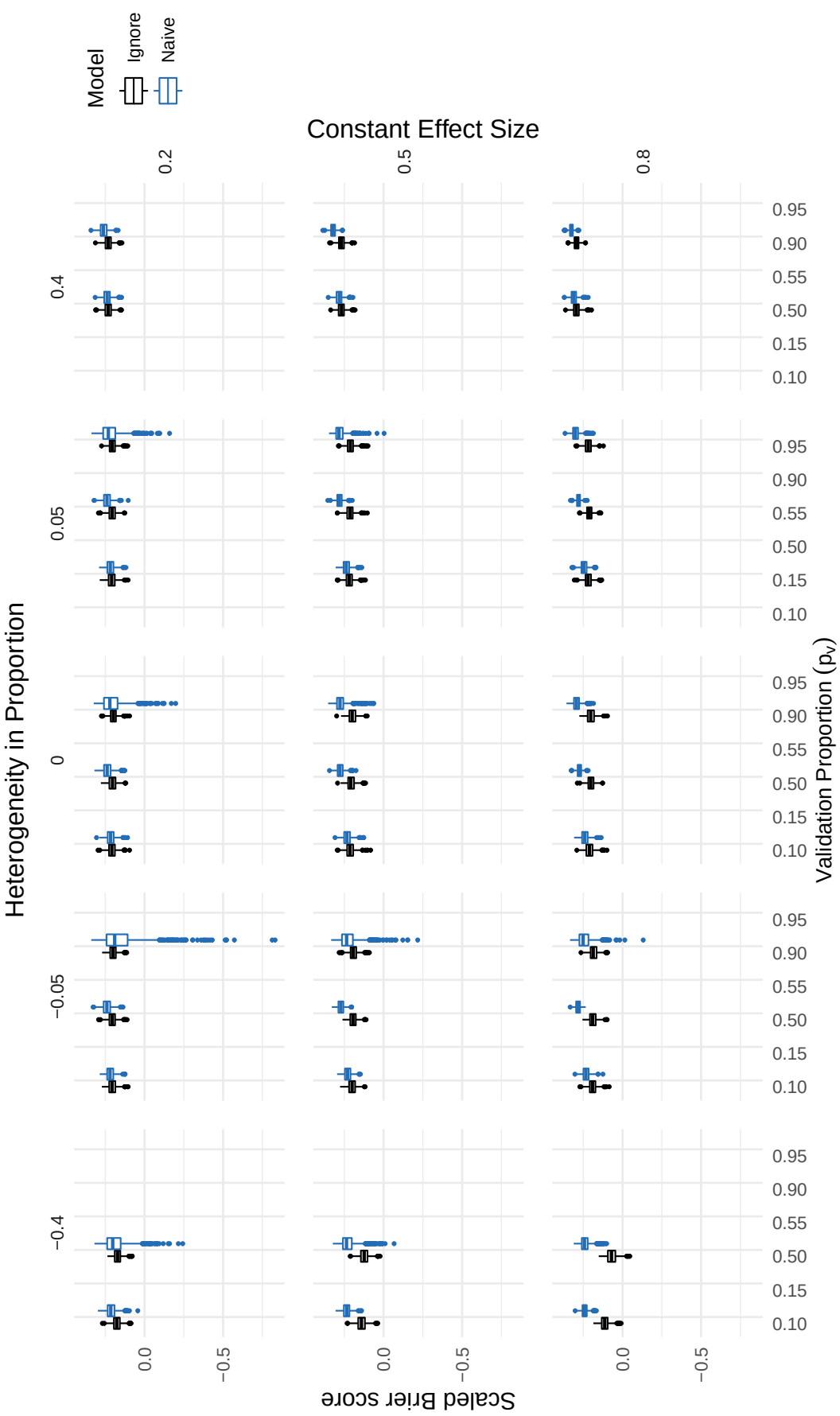
(b)

Figure D7 C-statistic for the different cases for (b) varying or equal treatment proportion and constant treatment effect.



(a)

Figure D8 Scaled Brier score for the different cases for (a) varying or equal treatment effect and constant treatment proportion.



(b)

Figure D8 Scaled Brier score for the different cases for (b) varying or equal treatment proportion and constant treatment effect.

E RESULT TABLES

E.1 Study 1 - Ignore Treatment approach

Table E6 Study 1 results for the Ignore Treatment approach for the performance measures Calibration-in-the-large coefficient, Calibration slope, c-statistic, and scaled Brier score with heterogeneity in treatment effect size.

Effect size	$p_d = p_v$	ES_d	ES_v	Calibr. coeff.[1]	Calibr. slope[2]	c-statistic	Brier
Homogeneity	0.10	0.20	0.20	-0.0076	0.9979	0.8031	0.2757
	0.10	0.50	0.50	0.0176	1.0043	0.8037	0.2768
	0.10	0.80	0.80	-0.0004	0.9929	0.8021	0.2739
	0.50	0.20	0.20	0.0155	0.9888	0.7998	0.2699
	0.50	0.50	0.50	0.0044	0.9984	0.8026	0.2748
	0.50	0.80	0.80	0.0093	1.0056	0.8014	0.2727
	0.90	0.20	0.20	-0.0042	1.0047	0.8146	0.2964
	0.90	0.50	0.50	-0.0019	0.9846	0.8005	0.2709
	0.90	0.80	0.80	-0.0040	0.9883	0.7994	0.2690
Medium increase	0.10	0.20	0.50	-0.0383	0.9631	0.8015	0.2724
(0.3)	0.10	0.50	0.80	-0.0334	0.9780	0.7982	0.2668
	0.50	0.20	0.50	-0.1553	0.9509	0.7903	0.2504
	0.50	0.50	0.80	-0.1397	0.9427	0.7914	0.2536
	0.90	0.20	0.50	-0.2811	0.9810	0.8098	0.2756
	0.90	0.50	0.80	-0.2512	0.9838	0.7975	0.2604
Large increase	0.10	0.20	0.80	-0.0645	0.9595	0.7968	0.2639
(0.6)	0.50	0.20	0.80	-0.3024	0.8716	0.7775	0.2218
	0.90	0.20	0.80	-0.5356	0.9593	0.8088	0.2546
Medium decrease	0.10	0.50	0.20	0.0316	1.0190	0.8048	0.2788
(-0.3)	0.10	0.80	0.50	0.0263	1.0236	0.8054	0.2797
	0.50	0.50	0.20	0.1442	1.0553	0.8115	0.2895
	0.50	0.80	0.50	0.1458	1.0540	0.8107	0.2886
	0.90	0.50	0.20	0.2838	1.0264	0.8050	0.2746
	0.90	0.80	0.50	0.2722	1.0312	0.8067	0.2760
Large decrease	0.10	0.80	0.20	0.0426	1.0463	0.8087	0.2856
(-0.6)	0.50	0.80	0.20	0.3092	1.1236	0.8198	0.2995
	0.90	0.80	0.20	0.5405	1.0351	0.8083	0.2622

[1] Calibration-in-the-large coefficient
[2] Calibration slope

Table E7 Study 1 results for the Ignore approach for the performance measures Calibration-in-the-large coefficient, Calibration slope, c-statistic, and scaled Brier score with heterogeneity in treatment proportion.

Proportion	$ES_d = ES_v$	p_d	p_v	Calibr. coeff.[3]	Calibr. slope[4]	c-statistic	Brier
Homogeneity	0.20	0.10	0.10	-0.0076	0.9979	0.8031	0.2757
	0.20	0.50	0.50	0.0155	0.9888	0.7998	0.2699
	0.20	0.90	0.90	-0.0042	1.0047	0.8146	0.2964
	0.50	0.10	0.10	0.0176	1.0043	0.8037	0.2768
	0.50	0.50	0.50	0.0044	0.9984	0.8026	0.2748
	0.50	0.90	0.90	-0.0019	0.9846	0.8005	0.2709
	0.80	0.10	0.10	-0.0004	0.9929	0.8021	0.2739
	0.80	0.50	0.50	0.0093	1.0056	0.8014	0.2727
	0.80	0.90	0.90	-0.0040	0.9883	0.7994	0.2690
Medium increase	0.20	0.10	0.50	-0.1067	0.9844	0.8012	0.2709
(0.4)	0.20	0.50	0.90	-0.0703	1.0144	0.8034	0.2757
	0.50	0.10	0.50	-0.1841	0.9658	0.7948	0.2584
	0.50	0.50	0.90	-0.2125	1.0604	0.8134	0.2917
	0.80	0.10	0.50	-0.3325	0.9109	0.7869	0.2387
	0.80	0.50	0.90	-0.3161	1.0951	0.8156	0.2906
Large increase	0.20	0.10	0.90	-0.1658	1.0029	0.8032	0.2732
(0.8)	0.50	0.10	0.90	-0.3955	0.9904	0.8009	0.2599
	0.80	0.10	0.90	-0.6484	0.9859	0.8023	0.2463
Medium decrease	0.20	0.50	0.10	0.0789	1.0152	0.8047	0.2783
(-0.4)	0.20	0.90	0.50	0.1049	0.9660	0.8083	0.2858
	0.50	0.50	0.10	0.2010	1.0529	0.8129	0.2904
	0.50	0.90	0.50	0.2049	0.9529	0.7912	0.2526
	0.80	0.50	0.10	0.3125	1.0914	0.8157	0.2923
	0.80	0.90	0.50	0.2985	0.9174	0.7834	0.2332
Large decrease	0.20	0.90	0.10	0.1640	0.9921	0.8131	0.2948
(-0.8)	0.50	0.90	0.10	0.4172	0.9996	0.8012	0.2615
	0.80	0.90	0.10	0.6428	1.0014	0.8027	0.2438

[3]Calibration-in-the-large coefficient
[4]Calibration slope

E.2 Study 1 - Treatment-Naïve approach

Table E8 Study 1 results for the Treatment-Naïve approach for the performance measures Calibration-in-the-large coefficient, Calibration slope, c-statistic, and scaled Brier score for heterogeneity in treatment effect size.

Effect size	$p_d = p_v$	ES_d	ES_v	Calibr. coeff.[5]	Calibr. slope[6]	c-statistic	Brier
Homogeneity	0.10	0.20	0.20	-0.0277	0.9821	0.8031	0.2754
	0.10	0.50	0.50	-0.0263	0.9705	0.8037	0.2764
	0.10	0.80	0.80	-0.0671	0.9427	0.8021	0.2727
	0.50	0.20	0.20	-0.0692	0.9587	0.7998	0.2683
	0.50	0.50	0.50	-0.2218	0.9155	0.8026	0.2608
	0.50	0.80	0.80	-0.3477	0.8700	0.8014	0.2347
	0.90	0.20	0.20	-0.2070	0.9926	0.8146	0.2816
	0.90	0.50	0.50	-0.4217	0.9360	0.8005	0.2193
	0.90	0.80	0.80	-0.6760	0.9821	0.7994	0.1521
Medium increase	0.10	0.20	0.50	-0.0536	0.9548	0.8015	0.2719
(0.3)	0.10	0.50	0.80	-0.0788	0.9460	0.7982	0.2654
	0.50	0.20	0.50	-0.2321	0.9416	0.7903	0.2444
	0.50	0.50	0.80	-0.3678	0.8568	0.7914	0.2252
	0.90	0.20	0.50	-0.4567	0.9718	0.8098	0.2483
	0.90	0.50	0.80	-0.6939	0.9760	0.7975	0.1727
Large increase	0.10	0.20	0.80	-0.0792	0.9481	0.7968	0.2634
(0.6)	0.50	0.20	0.80	-0.3802	0.8435	0.7775	0.2096
	0.90	0.20	0.80	-0.7257	0.9271	0.8088	0.2076
Medium decrease	0.10	0.50	0.20	-0.0132	0.9869	0.8048	0.2786
(-0.3)	0.10	0.80	0.50	-0.0466	0.9694	0.8054	0.2794
	0.50	0.50	0.20	-0.0930	0.9693	0.8115	0.2857
	0.50	0.80	0.50	-0.2206	0.9285	0.8107	0.2681
	0.90	0.50	0.20	-0.2141	1.0030	0.8050	0.2533
	0.90	0.80	0.50	-0.3974	0.9902	0.8067	0.2223
Large decrease	0.10	0.80	0.20	-0.0330	0.9946	0.8087	0.2853
(-0.6)	0.50	0.80	0.20	-0.0808	0.9674	0.8198	0.2979
	0.90	0.80	0.20	-0.1687	0.9959	0.8083	0.2572

[5] Calibration-in-the-large coefficient
[6] Calibration slope

Table E9 Study 1 results for the Treatment-Naïve approach for the performance measures Calibration-in-the-large coefficient, Calibration slope, c-statistic, and scaled Brier score for heterogeneity in treatment proportion.

Proportion	$ES_d = ES_v$	p_d	p_v	Calibr. coeff.[7]	Calibr. slope[8]	c-statistic	Brier
Homogeneity	0.20	0.10	0.10	-0.0277	0.9821	0.8031	0.2754
	0.20	0.50	0.50	-0.0692	0.9587	0.7998	0.2683
	0.20	0.90	0.90	-0.2070	0.9926	0.8146	0.2816
	0.50	0.10	0.10	-0.0263	0.9705	0.8037	0.2764
	0.50	0.50	0.50	-0.2218	0.9155	0.8026	0.2608
	0.50	0.90	0.90	-0.4217	0.9360	0.8005	0.2193
	0.80	0.10	0.10	-0.0671	0.9427	0.8021	0.2727
	0.80	0.50	0.50	-0.3477	0.8700	0.8014	0.2347
	0.80	0.90	0.90	-0.6760	0.9821	0.7994	0.1521
Medium increase	0.20	0.10	0.50	-0.1212	0.9740	0.8012	0.2702
(0.4)	0.20	0.50	0.90	-0.1599	0.9874	0.8034	0.2714
	0.50	0.10	0.50	-0.2250	0.9355	0.7948	0.2550
	0.50	0.50	0.90	-0.4511	0.9662	0.8134	0.2640
	0.80	0.10	0.50	-0.4013	0.8592	0.7869	0.2275
	0.80	0.50	0.90	-0.6916	0.9363	0.8156	0.2190
Large increase	0.20	0.10	0.90	-0.1804	0.9910	0.8032	0.2723
(0.8)	0.50	0.10	0.90	-0.4359	0.9600	0.8009	0.2541
	0.80	0.10	0.90	-0.7160	0.9325	0.8023	0.2303
Medium decrease	0.20	0.50	0.10	-0.0274	0.9793	0.8047	0.2777
(-0.4)	0.20	0.90	0.50	-0.0220	0.9900	0.8083	0.2828
	0.50	0.50	0.10	-0.0312	0.9742	0.8129	0.2917
	0.50	0.90	0.50	-0.2446	0.8984	0.7912	0.2197
	0.80	0.50	0.10	-0.0705	0.9460	0.8157	0.2913
	0.80	0.90	0.50	-0.3206	0.8562	0.7834	0.1767
Large decrease	0.20	0.90	0.10	-0.0447	0.9916	0.8131	0.2896
(-0.8)	0.50	0.90	0.10	-0.0196	0.9501	0.8012	0.2656
	0.80	0.90	0.10	-0.0368	0.9494	0.8027	0.2624

[7]Calibration-in-the-large coefficient
[8]Calibration slope

E.3 Study 2 - Ignore Treatment approach

Table E10 Study 2 results for the Ignore Treatment approach comprising the median, mean, and variance for the performance measures Calibration-in-the-large coefficient, Calibration slope, c-statistic, and scaled Brier score for heterogeneity in treatment effect size.

	Setting			Calibr. coeff.[9]			Calibr. slope[10]			c-statistic			Brier		
Effect size	$p_d = p_v$	ES_d	ES_v	median	mean	var	median	mean	var	median	mean	var	median	mean	var
Homo-	0.10	0.20	0.20	-0.0252	-0.0211	1.01e-02	0.9809	0.9836	4.51e-03	0.7995	0.7991	8.21e-05	0.2083	0.2069	8.09e-04
geneity	0.10	0.50	0.50	-0.0129	-0.0137	8.75e-03	0.9856	0.9885	4.00e-03	0.8039	0.8036	9.06e-05	0.2131	0.2120	8.15e-04
	0.10	0.80	0.80	-0.0180	-0.0166	9.75e-03	0.9879	0.9887	4.16e-03	0.8033	0.8033	9.12e-05	0.2107	0.2106	7.97e-04
	0.50	0.20	0.20	-0.0152	-0.0088	9.42e-03	0.9855	0.9911	4.26e-03	0.7946	0.7946	8.39e-05	0.2040	0.2036	7.85e-04
	0.50	0.50	0.50	-0.0129	-0.0127	8.62e-03	0.9875	0.9912	3.78e-03	0.7958	0.7954	8.29e-05	0.2079	0.2058	7.32e-04
	0.50	0.80	0.80	-0.0137	-0.0109	6.91e-03	0.9861	0.9873	3.87e-03	0.7835	0.7834	7.82e-05	0.2026	0.2013	5.50e-04
	0.90	0.20	0.20	-0.0182	-0.0138	8.15e-03	0.9822	0.9844	4.07e-03	0.7899	0.7898	8.04e-05	0.1985	0.1980	7.13e-04
	0.90	0.50	0.50	-0.0119	-0.0105	8.11e-03	0.9832	0.9877	3.93e-03	0.7919	0.7918	8.99e-05	0.1999	0.1987	7.58e-04
	0.90	0.80	0.80	-0.0133	-0.0103	9.18e-03	0.9871	0.9881	4.13e-03	0.7936	0.7931	8.73e-05	0.2030	0.2017	7.49e-04
Medium	0.10	0.20	0.50	-0.0639	-0.0574	8.95e-03	0.9746	0.9761	4.01e-03	0.7974	0.7973	8.81e-05	0.2132	0.2128	7.63e-04
increase	0.10	0.50	0.80	-0.0606	-0.0604	9.69e-03	0.9700	0.9741	3.82e-03	0.8013	0.8011	7.71e-05	0.2174	0.2169	8.19e-04
(0.3)	0.50	0.20	0.50	-0.2078	-0.2039	9.01e-03	0.9523	0.9557	3.89e-03	0.7899	0.7896	9.52e-05	0.2434	0.2425	6.29e-04
	0.50	0.50	0.80	-0.1931	-0.1913	8.17e-03	0.9531	0.9542	3.91e-03	0.7901	0.7896	9.56e-05	0.2385	0.2381	6.25e-04
	0.90	0.20	0.50	-0.2976	-0.2969	7.97e-03	0.9726	0.9750	3.90e-03	0.7898	0.7899	8.97e-05	0.2757	0.2753	4.68e-04
	0.90	0.50	0.80	-0.2958	-0.2932	8.34e-03	0.9715	0.9764	4.12e-03	0.7913	0.7912	9.75e-05	0.2765	0.2759	5.75e-04
Medium	0.10	0.50	0.20	0.0370	0.0388	9.99e-03	0.9998	1.0018	4.21e-03	0.8055	0.8056	8.51e-05	0.2036	0.2029	8.78e-04
decrease	0.10	0.80	0.50	0.0343	0.0372	9.93e-03	1.0004	1.0044	4.22e-03	0.8058	0.8058	8.22e-05	0.2050	0.2035	8.35e-04
(-0.3)	0.50	0.50	0.20	0.1814	0.1810	8.88e-03	1.0164	1.0186	4.00e-03	0.7996	0.7996	7.54e-05	0.1597	0.1592	9.29e-04
	0.50	0.80	0.50	0.1674	0.1649	7.60e-03	1.0165	1.0203	3.72e-03	0.7886	0.7885	7.29e-05	0.1681	0.1680	6.71e-04
	0.90	0.50	0.20	0.2736	0.2707	8.42e-03	0.9948	0.9964	4.08e-03	0.7923	0.7922	7.68e-05	0.1063	0.1049	1.09e-03
	0.90	0.80	0.50	0.2716	0.2750	9.00e-03	0.9992	1.0016	3.89e-03	0.7937	0.7936	7.62e-05	0.1092	0.1081	1.13e-03

[9]Calibration-in-the-large coefficient
[10]Calibration slope

Table E11 Study 2 results for the Ignore Treatment approach comprising the median, mean, and variance for the performance measures Calibration-in-the-large coefficient, Calibration slope, c-statistic, and scaled Brier score for heterogeneity in treatment proportion.

	Setting			Calibr. coeff.[11]			Calibr. slope[12]			c-statistic			Brier		
Proportion	$ES_d = ES_v$	p_d	p_v	median	mean	var	median	mean	var	median	mean	var	median	mean	var
Homo-	0.20	0.10	0.10	-0.0252	-0.0211	1.01e-02	0.9809	0.9836	4.51e-03	0.7995	0.7991	8.21e-05	0.2083	0.2069	8.09e-04
geneity	0.20	0.50	0.50	-0.0152	-0.0088	9.42e-03	0.9855	0.9911	4.26e-03	0.7946	0.7946	8.39e-05	0.2040	0.2036	7.85e-04
	0.20	0.90	0.90	-0.0182	-0.0138	8.15e-03	0.9822	0.9844	4.07e-03	0.7899	0.7898	8.04e-05	0.1985	0.1980	7.13e-04
	0.50	0.10	0.10	-0.0129	-0.0137	8.75e-03	0.9856	0.9885	4.00e-03	0.8039	0.8036	9.06e-05	0.2131	0.2120	8.15e-04
	0.50	0.50	0.50	-0.0129	-0.0127	8.62e-03	0.9875	0.9912	3.78e-03	0.7958	0.7954	8.29e-05	0.2079	0.2058	7.32e-04
	0.50	0.90	0.90	-0.0119	-0.0105	8.11e-03	0.9832	0.9877	3.93e-03	0.7919	0.7918	8.99e-05	0.1999	0.1987	7.58e-04
	0.80	0.10	0.10	-0.0180	-0.0166	9.75e-03	0.9879	0.9887	4.16e-03	0.8033	0.8033	9.12e-05	0.2107	0.2106	7.97e-04
	0.80	0.50	0.50	-0.0137	-0.0109	6.91e-03	0.9861	0.9873	3.87e-03	0.7835	0.7834	7.82e-05	0.2026	0.2013	5.50e-04
	0.80	0.90	0.90	-0.0133	-0.0103	9.18e-03	0.9871	0.9881	4.13e-03	0.7936	0.7931	8.73e-05	0.2030	0.2017	7.49e-04
Small	0.20	0.10	0.15	-0.0334	-0.0270	8.85e-03	0.9811	0.9844	3.95e-03	0.7989	0.7990	8.09e-05	0.2094	0.2091	7.79e-04
increase	0.20	0.50	0.55	-0.0294	-0.0246	9.35e-03	0.9820	0.9857	3.73e-03	0.7942	0.7944	8.04e-05	0.2068	0.2062	7.40e-04
(0.05)	0.20	0.90	0.95	-0.0192	-0.0152	8.08e-03	0.9899	0.9930	4.04e-03	0.7914	0.7913	8.29e-05	0.2057	0.2040	6.59e-04
	0.50	0.10	0.15	-0.0457	-0.0461	9.19e-03	0.9819	0.9808	4.03e-03	0.8029	0.8027	8.18e-05	0.2183	0.2176	7.83e-04
	0.50	0.50	0.55	-0.0385	-0.0354	9.06e-03	0.9860	0.9888	4.32e-03	0.7959	0.7954	8.27e-05	0.2126	0.2124	7.11e-04
	0.50	0.90	0.95	-0.0272	-0.0286	8.96e-03	0.9946	0.9950	4.13e-03	0.7943	0.7943	8.51e-05	0.2113	0.2104	7.58e-04
	0.80	0.10	0.15	-0.0616	-0.0625	9.93e-03	0.9760	0.9782	3.79e-03	0.8022	0.8018	8.52e-05	0.2195	0.2181	8.01e-04
	0.80	0.50	0.55	-0.0429	-0.0413	6.98e-03	0.9893	0.9920	3.62e-03	0.7842	0.7843	7.66e-05	0.2123	0.2120	5.10e-04
	0.80	0.90	0.95	-0.0343	-0.0349	8.44e-03	1.0005	1.0003	3.80e-03	0.7959	0.7962	8.21e-05	0.2183	0.2177	7.02e-04

[11]Calibration-in-the-large coefficient
[12]Calibration slope

Medium	0.20	0.10	0.50	-0.1138	-0.1095	9.35e-03	0.9753	0.9761	4.36e-03	0.7984	0.7981	1.00e-04	0.2311	0.2305	7.06e-04
increase	0.20	0.50	0.90	-0.0718	-0.0669	8.23e-03	1.0006	1.0045	4.12e-03	0.7983	0.7980	8.23e-05	0.2308	0.2306	6.59e-04
(0.4)	0.50	0.10	0.50	-0.2556	-0.2521	9.35e-03	0.9590	0.9626	3.83e-03	0.8002	0.7998	9.37e-05	0.2679	0.2680	6.34e-04
	0.50	0.50	0.90	-0.1727	-0.1688	8.88e-03	1.0140	1.0176	4.13e-03	0.8018	0.8018	8.58e-05	0.2687	0.2683	5.63e-04
	0.80	0.10	0.50	-0.3876	-0.3880	9.05e-03	0.9408	0.9436	3.78e-03	0.7970	0.7969	1.06e-04	0.2945	0.2937	5.61e-04
	0.80	0.50	0.90	-0.2725	-0.2719	6.82e-03	1.0431	1.0439	4.18e-03	0.7956	0.7954	8.34e-05	0.2924	0.2926	3.60e-04
Small	0.20	0.15	0.10	-0.0015	-0.0012	9.10e-03	0.9858	0.9897	3.88e-03	0.8014	0.8007	8.47e-05	0.2067	0.2052	8.11e-04
decrease	0.20	0.55	0.50	-0.0104	-0.0061	8.41e-03	0.9873	0.9893	3.98e-03	0.7965	0.7962	8.44e-05	0.2053	0.2047	7.28e-04
(-0.05)	0.20	0.95	0.90	-0.0080	-0.0066	9.28e-03	0.9852	0.9857	3.89e-03	0.7940	0.7937	8.46e-05	0.2002	0.1999	7.02e-04
	0.50	0.15	0.10	0.0197	0.0181	8.70e-03	0.9971	0.9955	3.81e-03	0.7988	0.7984	8.52e-05	0.2005	0.2004	7.64e-04
	0.50	0.55	0.50	0.0064	0.0097	8.86e-03	0.9863	0.9881	3.96e-03	0.7925	0.7924	7.38e-05	0.1946	0.1941	6.95e-04
	0.50	0.95	0.90	-0.0031	-0.0006	8.08e-03	0.9763	0.9799	3.64e-03	0.7890	0.7889	7.44e-05	0.1924	0.1918	7.41e-04
	0.80	0.15	0.10	0.0454	0.0427	8.55e-03	0.9997	1.0015	3.89e-03	0.7944	0.7949	7.96e-05	0.1916	0.1904	7.68e-04
	0.80	0.55	0.50	0.0156	0.0181	7.83e-03	0.9829	0.9855	3.75e-03	0.7867	0.7868	8.09e-05	0.1904	0.1886	7.06e-04
	0.80	0.95	0.90	0.0031	0.0059	7.73e-03	0.9669	0.9698	3.66e-03	0.7889	0.7886	8.03e-05	0.1868	0.1842	7.56e-04
Medium	0.20	0.50	0.10	0.0829	0.0832	9.22e-03	0.9927	0.9946	3.94e-03	0.7956	0.7956	7.87e-05	0.1770	0.1760	8.33e-04
decrease	0.20	0.90	0.50	0.0502	0.0533	8.31e-03	0.9742	0.9768	3.70e-03	0.7878	0.7878	8.10e-05	0.1719	0.1708	7.47e-04
(-0.4)	0.50	0.50	0.10	0.2284	0.2321	8.42e-03	1.0159	1.0184	3.69e-03	0.7994	0.7989	7.23e-05	0.1408	0.1393	9.40e-04
	0.50	0.90	0.50	0.1408	0.1399	8.21e-03	0.9541	0.9546	3.70e-03	0.7848	0.7845	8.32e-05	0.1234	0.1232	9.75e-04
	0.80	0.50	0.10	0.3524	0.3559	7.08e-03	1.0343	1.0340	4.02e-03	0.7903	0.7901	6.55e-05	0.1134	0.1130	8.57e-04
	0.80	0.90	0.50	0.2367	0.2364	7.88e-03	0.9321	0.9329	3.61e-03	0.7803	0.7801	8.37e-05	0.0706	0.0686	1.24e-03

E.4 Study 2 - Treatment-Naïve approach

Table E12 Study 2 results for the Treatment-Naïve approach comprising the median, mean, and variance for the performance measures Calibration-in-the-large coefficient, Calibration slope, c-statistic, and scaled Brier score for heterogeneity in treatment effect size.

	Setting			Calibr. coeff.[13]			Calibr. slope[14]			c-statistic			Brier		
Effect size	$p_d = p_v$	ES_d	ES_v	median	mean	var	median	mean	var	median	mean	var	median	mean	var
Homo-	0.10	0.20	0.20	-0.0252	-0.0211	1.01e-02	0.9809	0.9836	4.51e-03	0.7995	0.7991	8.21e-05	0.2083	0.2069	8.09e-04
geneity	0.10	0.50	0.50	-0.0129	-0.0137	8.75e-03	0.9856	0.9885	4.00e-03	0.8039	0.8036	9.06e-05	0.2131	0.2120	8.15e-04
	0.10	0.80	0.80	-0.0180	-0.0166	9.75e-03	0.9879	0.9887	4.16e-03	0.8033	0.8033	9.12e-05	0.2107	0.2106	7.97e-04
	0.50	0.20	0.20	-0.0152	-0.0088	9.42e-03	0.9855	0.9911	4.26e-03	0.7946	0.7946	8.39e-05	0.2040	0.2036	7.85e-04
	0.50	0.50	0.50	-0.0129	-0.0127	8.62e-03	0.9875	0.9912	3.78e-03	0.7958	0.7954	8.29e-05	0.2079	0.2058	7.32e-04
	0.50	0.80	0.80	-0.0137	-0.0109	6.91e-03	0.9861	0.9873	3.87e-03	0.7835	0.7834	7.82e-05	0.2026	0.2013	5.50e-04
	0.90	0.20	0.20	-0.0182	-0.0138	8.15e-03	0.9822	0.9844	4.07e-03	0.7899	0.7898	8.04e-05	0.1985	0.1980	7.13e-04
	0.90	0.50	0.50	-0.0119	-0.0105	8.11e-03	0.9832	0.9877	3.93e-03	0.7919	0.7918	8.99e-05	0.1999	0.1987	7.58e-04
	0.90	0.80	0.80	-0.0133	-0.0103	9.18e-03	0.9871	0.9881	4.13e-03	0.7936	0.7931	8.73e-05	0.2030	0.2017	7.49e-04
Medium	0.10	0.20	0.50	-0.0639	-0.0574	8.95e-03	0.9746	0.9761	4.01e-03	0.7974	0.7973	8.81e-05	0.2132	0.2128	7.63e-04
increase	0.10	0.50	0.80	-0.0606	-0.0604	9.69e-03	0.9700	0.9741	3.82e-03	0.8013	0.8011	7.71e-05	0.2174	0.2169	8.19e-04
(0.3)	0.50	0.20	0.50	-0.2078	-0.2039	9.01e-03	0.9523	0.9557	3.89e-03	0.7899	0.7896	9.52e-05	0.2434	0.2425	6.29e-04
	0.50	0.50	0.80	-0.1931	-0.1913	8.17e-03	0.9531	0.9542	3.91e-03	0.7901	0.7896	9.56e-05	0.2385	0.2381	6.25e-04
	0.90	0.20	0.50	-0.2976	-0.2969	7.97e-03	0.9726	0.9750	3.90e-03	0.7898	0.7899	8.97e-05	0.2757	0.2753	4.68e-04
	0.90	0.50	0.80	-0.2958	-0.2932	8.34e-03	0.9715	0.9764	4.12e-03	0.7913	0.7912	9.75e-05	0.2765	0.2759	5.75e-04
Medium	0.10	0.50	0.20	0.0370	0.0388	9.99e-03	0.9998	1.0018	4.21e-03	0.8055	0.8056	8.51e-05	0.2036	0.2029	8.78e-04
decrease	0.10	0.80	0.50	0.0343	0.0372	9.93e-03	1.0004	1.0044	4.22e-03	0.8058	0.8058	8.22e-05	0.2050	0.2035	8.35e-04
(-0.3)	0.50	0.50	0.20	0.1814	0.1810	8.88e-03	1.0164	1.0186	4.00e-03	0.7996	0.7996	7.54e-05	0.1597	0.1592	9.29e-04
	0.50	0.80	0.50	0.1674	0.1649	7.60e-03	1.0165	1.0203	3.72e-03	0.7886	0.7885	7.29e-05	0.1681	0.1680	6.71e-04
	0.90	0.50	0.20	0.2736	0.2707	8.42e-03	0.9948	0.9964	4.08e-03	0.7923	0.7922	7.68e-05	0.1063	0.1049	1.09e-03
	0.90	0.80	0.50	0.2716	0.2750	9.00e-03	0.9992	1.0016	3.89e-03	0.7937	0.7936	7.62e-05	0.1092	0.1081	1.13e-03

Table E13 Study 2 results for the Treatment-Naïve approach comprising the median, mean, and variance for the performance measures Calibration-in-the-large coefficient, Calibration slope, c-statistic, and scaled Brier score for heterogeneity in treatment proportion.

Setting				Calibr. coeff.[15]			Calibr. slope[16]			c-statistic			Brier		
Proportion	$ES_d = ES_v$	p_d	p_v	median	mean	var	median	mean	var	median	mean	var	median	mean	var
Homo-	0.20	0.10	0.10	-0.0252	-0.0211	1.01e-02	0.9809	0.9836	4.51e-03	0.7995	0.7991	8.21e-05	0.2083	0.2069	8.09e-04
geneity	0.20	0.50	0.50	-0.0152	-0.0088	9.42e-03	0.9855	0.9911	4.26e-03	0.7946	0.7946	8.39e-05	0.2040	0.2036	7.85e-04
	0.20	0.90	0.90	-0.0182	-0.0138	8.15e-03	0.9822	0.9844	4.07e-03	0.7899	0.7898	8.04e-05	0.1985	0.1980	7.13e-04
	0.50	0.10	0.10	-0.0129	-0.0137	8.75e-03	0.9856	0.9885	4.00e-03	0.8039	0.8036	9.06e-05	0.2131	0.2120	8.15e-04
	0.50	0.50	0.50	-0.0129	-0.0127	8.62e-03	0.9875	0.9912	3.78e-03	0.7958	0.7954	8.29e-05	0.2079	0.2058	7.32e-04
	0.50	0.90	0.90	-0.0119	-0.0105	8.11e-03	0.9832	0.9877	3.93e-03	0.7919	0.7918	8.99e-05	0.1999	0.1987	7.58e-04
	0.80	0.10	0.10	-0.0180	-0.0166	9.75e-03	0.9879	0.9887	4.16e-03	0.8033	0.8033	9.12e-05	0.2107	0.2106	7.97e-04
	0.80	0.50	0.50	-0.0137	-0.0109	6.91e-03	0.9861	0.9873	3.87e-03	0.7835	0.7834	7.82e-05	0.2026	0.2013	5.50e-04
	0.80	0.90	0.90	-0.0133	-0.0103	9.18e-03	0.9871	0.9881	4.13e-03	0.7936	0.7931	8.73e-05	0.2030	0.2017	7.49e-04
Small	0.20	0.10	0.15	-0.0334	-0.0270	8.85e-03	0.9811	0.9844	3.95e-03	0.7989	0.7990	8.09e-05	0.2094	0.2091	7.79e-04
increase	0.20	0.50	0.55	-0.0294	-0.0246	9.35e-03	0.9820	0.9857	3.73e-03	0.7942	0.7944	8.04e-05	0.2068	0.2062	7.40e-04
(0.05)	0.20	0.90	0.95	-0.0192	-0.0152	8.08e-03	0.9899	0.9930	4.04e-03	0.7914	0.7913	8.29e-05	0.2057	0.2040	6.59e-04
	0.50	0.10	0.15	-0.0457	-0.0461	9.19e-03	0.9819	0.9808	4.03e-03	0.8029	0.8027	8.18e-05	0.2183	0.2176	7.83e-04
	0.50	0.50	0.55	-0.0385	-0.0354	9.06e-03	0.9860	0.9888	4.32e-03	0.7959	0.7954	8.27e-05	0.2126	0.2124	7.11e-04
	0.50	0.90	0.95	-0.0272	-0.0286	8.96e-03	0.9946	0.9950	4.13e-03	0.7943	0.7943	8.51e-05	0.2113	0.2104	7.58e-04
	0.80	0.10	0.15	-0.0616	-0.0625	9.93e-03	0.9760	0.9782	3.79e-03	0.8022	0.8018	8.52e-05	0.2195	0.2181	8.01e-04
	0.80	0.50	0.55	-0.0429	-0.0413	6.98e-03	0.9893	0.9920	3.62e-03	0.7842	0.7843	7.66e-05	0.2123	0.2120	5.10e-04
	0.80	0.90	0.95	-0.0343	-0.0349	8.44e-03	1.0005	1.0003	3.80e-03	0.7959	0.7962	8.21e-05	0.2183	0.2177	7.02e-04

[15]Calibration-in-the-large coefficient
[16]Calibration slope

Medium	0.20	0.10	0.50	-0.1138	-0.1095	9.35e-03	0.9753	0.9761	4.36e-03	0.7984	0.7981	1.00e-04	0.2311	0.2305	7.06e-04
increase	0.20	0.50	0.90	-0.0718	-0.0669	8.23e-03	1.0006	1.0045	4.12e-03	0.7983	0.7980	8.23e-05	0.2308	0.2306	6.59e-04
(0.4)	0.50	0.10	0.50	-0.2556	-0.2521	9.35e-03	0.9590	0.9626	3.83e-03	0.8002	0.7998	9.37e-05	0.2679	0.2680	6.34e-04
	0.50	0.50	0.90	-0.1727	-0.1688	8.88e-03	1.0140	1.0176	4.13e-03	0.8018	0.8018	8.58e-05	0.2687	0.2683	5.63e-04
	0.80	0.10	0.50	-0.3876	-0.3880	9.05e-03	0.9408	0.9436	3.78e-03	0.7970	0.7969	1.06e-04	0.2945	0.2937	5.61e-04
	0.80	0.50	0.90	-0.2725	-0.2719	6.82e-03	1.0431	1.0439	4.18e-03	0.7956	0.7954	8.34e-05	0.2924	0.2926	3.60e-04
Small	0.20	0.15	0.10	-0.0015	-0.0012	9.10e-03	0.9858	0.9897	3.88e-03	0.8014	0.8007	8.47e-05	0.2067	0.2052	8.11e-04
decrease	0.20	0.55	0.50	-0.0104	-0.0061	8.41e-03	0.9873	0.9893	3.98e-03	0.7965	0.7962	8.44e-05	0.2053	0.2047	7.28e-04
(-0.05)	0.20	0.95	0.90	-0.0080	-0.0066	9.28e-03	0.9852	0.9857	3.89e-03	0.7940	0.7937	8.46e-05	0.2002	0.1999	7.02e-04
	0.50	0.15	0.10	0.0197	0.0181	8.70e-03	0.9971	0.9955	3.81e-03	0.7988	0.7984	8.52e-05	0.2005	0.2004	7.64e-04
	0.50	0.55	0.50	0.0064	0.0097	8.86e-03	0.9863	0.9881	3.96e-03	0.7925	0.7924	7.38e-05	0.1946	0.1941	6.95e-04
	0.50	0.95	0.90	-0.0031	-0.0006	8.08e-03	0.9763	0.9799	3.64e-03	0.7890	0.7889	7.44e-05	0.1924	0.1918	7.41e-04
	0.80	0.15	0.10	0.0454	0.0427	8.55e-03	0.9997	1.0015	3.89e-03	0.7944	0.7949	7.96e-05	0.1916	0.1904	7.68e-04
	0.80	0.55	0.50	0.0156	0.0181	7.83e-03	0.9829	0.9855	3.75e-03	0.7867	0.7868	8.09e-05	0.1904	0.1886	7.06e-04
	0.80	0.95	0.90	0.0031	0.0059	7.73e-03	0.9669	0.9698	3.66e-03	0.7889	0.7886	8.03e-05	0.1868	0.1842	7.56e-04
Medium	0.20	0.50	0.10	0.0829	0.0832	9.22e-03	0.9927	0.9946	3.94e-03	0.7956	0.7956	7.87e-05	0.1770	0.1760	8.33e-04
decrease	0.20	0.90	0.50	0.0502	0.0533	8.31e-03	0.9742	0.9768	3.70e-03	0.7878	0.7878	8.10e-05	0.1719	0.1708	7.47e-04
(-0.4)	0.50	0.50	0.10	0.2284	0.2321	8.42e-03	1.0159	1.0184	3.69e-03	0.7994	0.7989	7.23e-05	0.1408	0.1393	9.40e-04
	0.50	0.90	0.50	0.1408	0.1399	8.21e-03	0.9541	0.9546	3.70e-03	0.7848	0.7845	8.32e-05	0.1234	0.1232	9.75e-04
	0.80	0.50	0.10	0.3524	0.3559	7.08e-03	1.0343	1.0340	4.02e-03	0.7903	0.7901	6.55e-05	0.1134	0.1130	8.57e-04
	0.80	0.90	0.50	0.2367	0.2364	7.88e-03	0.9321	0.9329	3.61e-03	0.7803	0.7801	8.37e-05	0.0706	0.0686	1.24e-03