Pitfalls and Potentials in Simulation Studies

Online Supplement

Samuel Pawel[*†], **Lucas Kook**[*], **Kelly Reeve**

Epidemiology, Biostatistics and Prevention Institute

Center for Reproducible Science

University of Zurich, Hirschengraben 84, CH-8001 Zurich

{samuel.pawel, lucasheinrich.kook, kelly.reeve}@uzh.ch

Here, we describe the outcomes of the pre-registered simulations. Overall, the performance of AINET was virtually identical to elastic net regression. The adaptive penalization weights of AINET do not seem to make a difference for the data generating mechanism considered in our simulations. Moreover, since the data were generated under a process equivalent to a logistic regression model, it is no surprise that for reasonably large sample sizes, logistic regression also performed the best. The only exception are conditions with small sample size and low number of events per variable. Here, AINET and elastic net led to more stable and better calibrated predictions than logistic regression. The random forest was outperformed by AINET in most simulation conditions, with exception of very small sample size and prevalence, as well as when a high correlation between covariates was present. Finally, the performance of the adaptive elastic net was generally worse compared to AINET and elastic net. In the following, we summarize the results for each estimand.

1 Brier score (primary estimand)

Figure 1 shows the differences in mean Brier score between AINET and the other methods stratified by simulation conditions. We see that there is hardly any difference between AINET and the elastic net (EN) across all simulation conditions meaning that predictive performance of both methods seems to be very similar in the investigated scenarios.

The random forest (RF) shows better predictive performance than AINET in conditions with very low sample size ($n = 100$) and prevalence (prev $= 0.01$). For increasing sample size and prevalence, the performance of AINET seems to become more similar or improve over RF when the correlation of the covariates is not too large ($\rho \leq 0.6$) especially for low events per variable (EPV ≤ 1). For highly correlated covariates ($\rho = 0.95$), the performance of AINET is similar or worse across most simulation conditions.

Logistic regression (GLM) showed better predictive performance compared to AINET in most simulation conditions. An exception are the conditions with small sample size ($n = 100$), medium to large prevalence (prev ≥ 0.05) and low events per variable (EPV ≤ 1), where AINET performed better than GLM.

[*]Contributed equally.

[†]Corresponding author: samuel.pawel@uzh.ch

The adaptive elastic net (AEN) method performed worse than AINET in almost all simulation conditions. Only in conditions with very large sample size ($n = 5000$), very small prevalence (prev $= 0.01$), and high events per variable (EPV $= 20$), AEN showed predictive performance on par with AINET.

2 Scaled Brier score (secondary estimand)

Figure 2 shows the differences in scaled Brier score between AINET and the other methods stratified by simulation conditions. The scaled Brier score is useful to compare the actual values of Brier scores across conditions with different prevalence, but not so much to compare Brier scores of different methods within a simulation condition with fixed prevalence.

We see that for most conditions the plots look like a flipped version of the original Brier scores from Figure 1. Therefore, conclusions are mostly the same. For very small sample sizes coupled with low prevalence and low events per variable (the topleft plots), the scaled Brier score indicates superiority of AINET over RF and GLM, which is opposite the conclusion based on the raw Brier score. We advise to interpret these conditions cautiously since the prevalence prediction which is used for scaling is based on the much larger test data set.

3 Log-score (secondary estimand)

Figure 3 shows the differences in log-score between AINET and the other methods stratified by simulation conditions. We see that in certain conditions, the error bars of certain methods are much larger. This is due to the log-score's sensitivity to extreme predictions, which often happen under the RF (and sometimes under the GLM). Despite the larger variability of the log-score, conclusion regarding the comparison between AINET and the other methods are largely the same as under the Brier score.

4 Area under the curve (secondary estimand)

Figure 4 shows the differences in area under the curve (AUC) between AINET and the other methods stratified by simulation conditions. As with the other estimands, AINET shows virtually identical performance as EN regression across all simulation conditions. AINET seems to outperform RF across most simulation conditions, with the exception of a conditions with low sample size ($n = 100$), medium prevalence (prev $= 0.05$), and low events per variable (EPV ≤ 1). GLM, typically outperforms AINET conditions with small to medium sample size ($n \leq 500$), and also in conditions with larger sample size when the events per variable is normal to high (EPV ≥ 10) and the prevalence is small (prev $= 0.01$). Finally, the AEN is worse with respect to AUC than AINET across all simulation conditions.

5 Calibration slope (secondary estimand)

Figure 5 shows boxplots of calibration slopes stratified by simulation condition and method. For each condition the percentage of simulations where no estimate could be obtained is indicated. This usually happened because of extreme (close to zero or one) predictions, or non-convergence of the method itself.

We caution against interpretation of the random forest (RF) calibration slopes because this method often resulted in predicted probabilities of zero or one, so that a calibration slope could not be fitted.

We see that logistic regression (GLM) shows on average optimal calibration slopes in most simulation condition. In cases where it is off one, its calibration slopes are usually too small indicating overoptimistic predictions. In general, worse calibration slopes are obtained for lower event per variable (EPV).

The penalized methods (AINET, EN, AEN) show a more stable behavior, and on average larger calibration slopes than GLM. This is likely confounded by the simulation conditions in which no GLM calibration slope can be estimated, but estimation of the penalized methods' calibration slope is still possible. Among the penalized method's AINET and EN shows relatively similar calibration slopes whereas the AEN shows worse calibration slopes that are more off the value of one.

6 Calibration in the large (secondary estimand)

Figure 6 shows boxplots of calibration in the large estimates stratified by simulation condition and method. For each condition also the percentage of simulations where no estimate could be obtained is indicated. This usually happened because of extreme (close to zero or one) predictions.

We see that the number of simulations with non-estimable calibration is substantially larger when the sample size is small, whereas it decreases for larger sample sizes. An exception is the RF where the number of non-estimable calibrations stays high across most conditions.

While all methods seem to be marginally well calibrated, the penalized methods (AINET, EN, and AEN) show lower numbers of simulations with non-estimable calibration compared to GLM, especially for low to medium sample sizes and low events per variables.

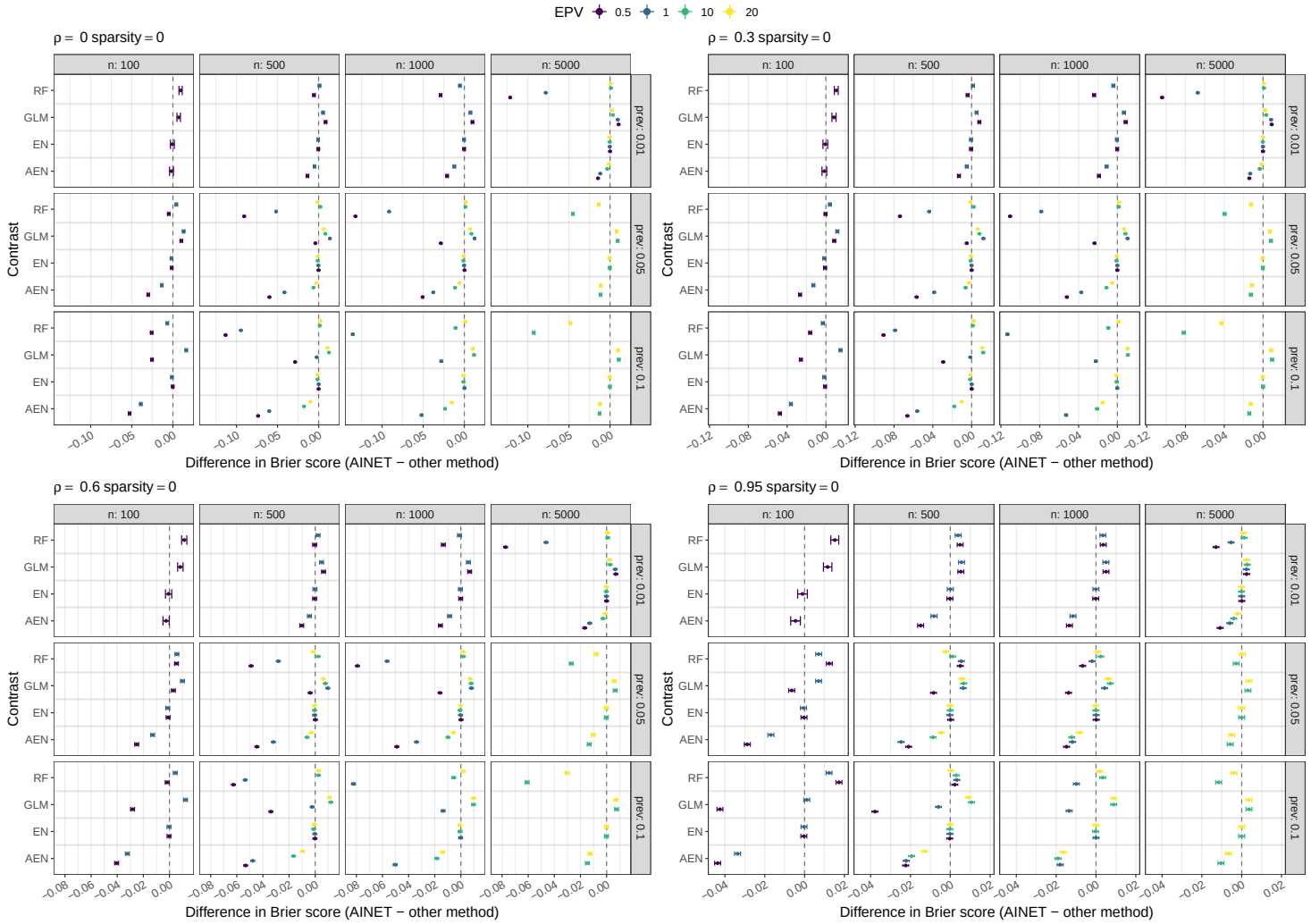

Figure 1: Tie-fighter plot for the difference in Brier score between any method on the y-axis and AINET. The 95% confidence intervals are adjusted per simulation condition using the single-step method. Lower values indicate better performance of AINET.

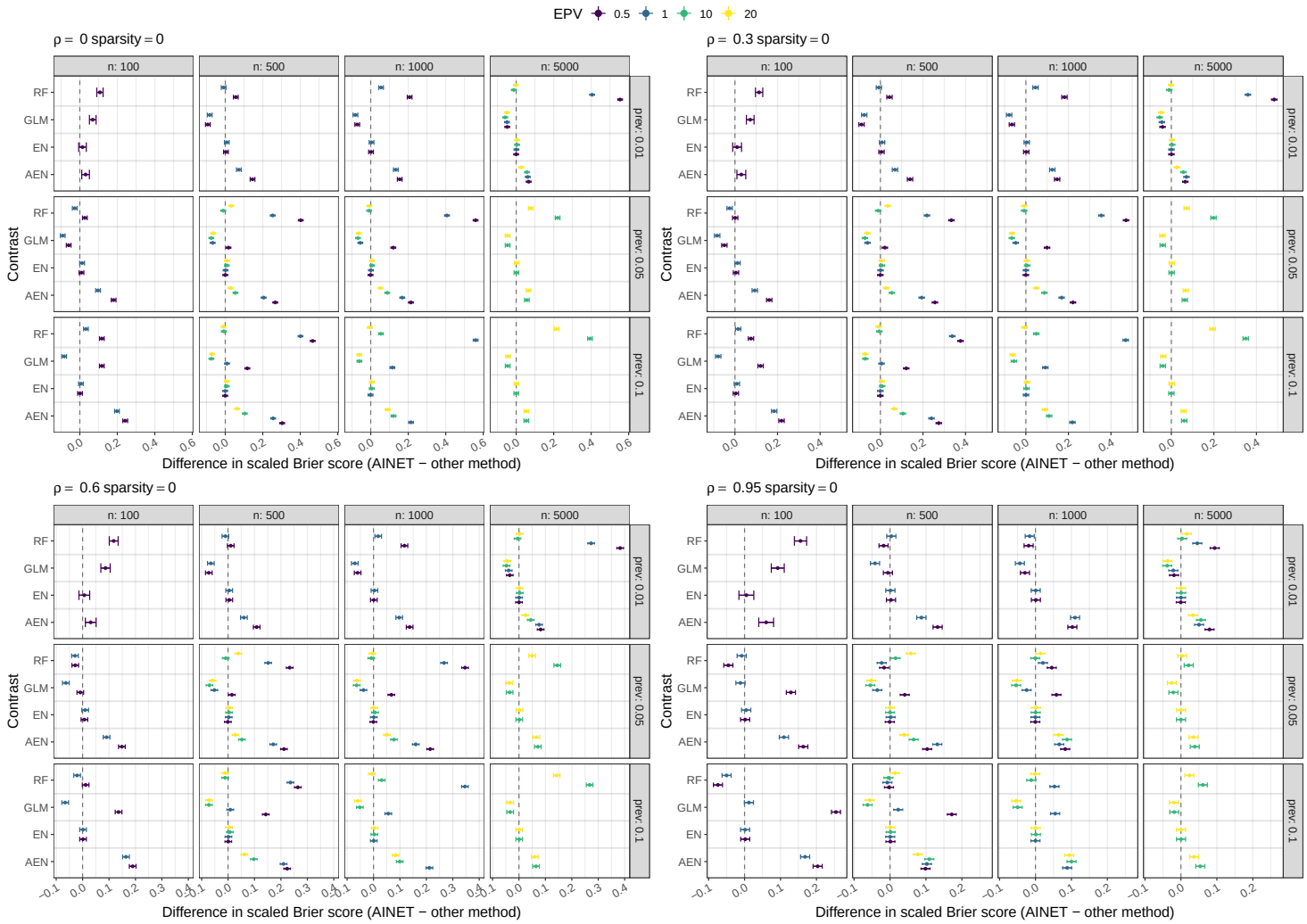

Figure 2: Tie-fighter plot for the difference in scaled Brier score between any method on the y-axis and AINET. The 95% confidence intervals are adjusted per simulation condition using the single-step method. Larger values indicate better performance of AINET.

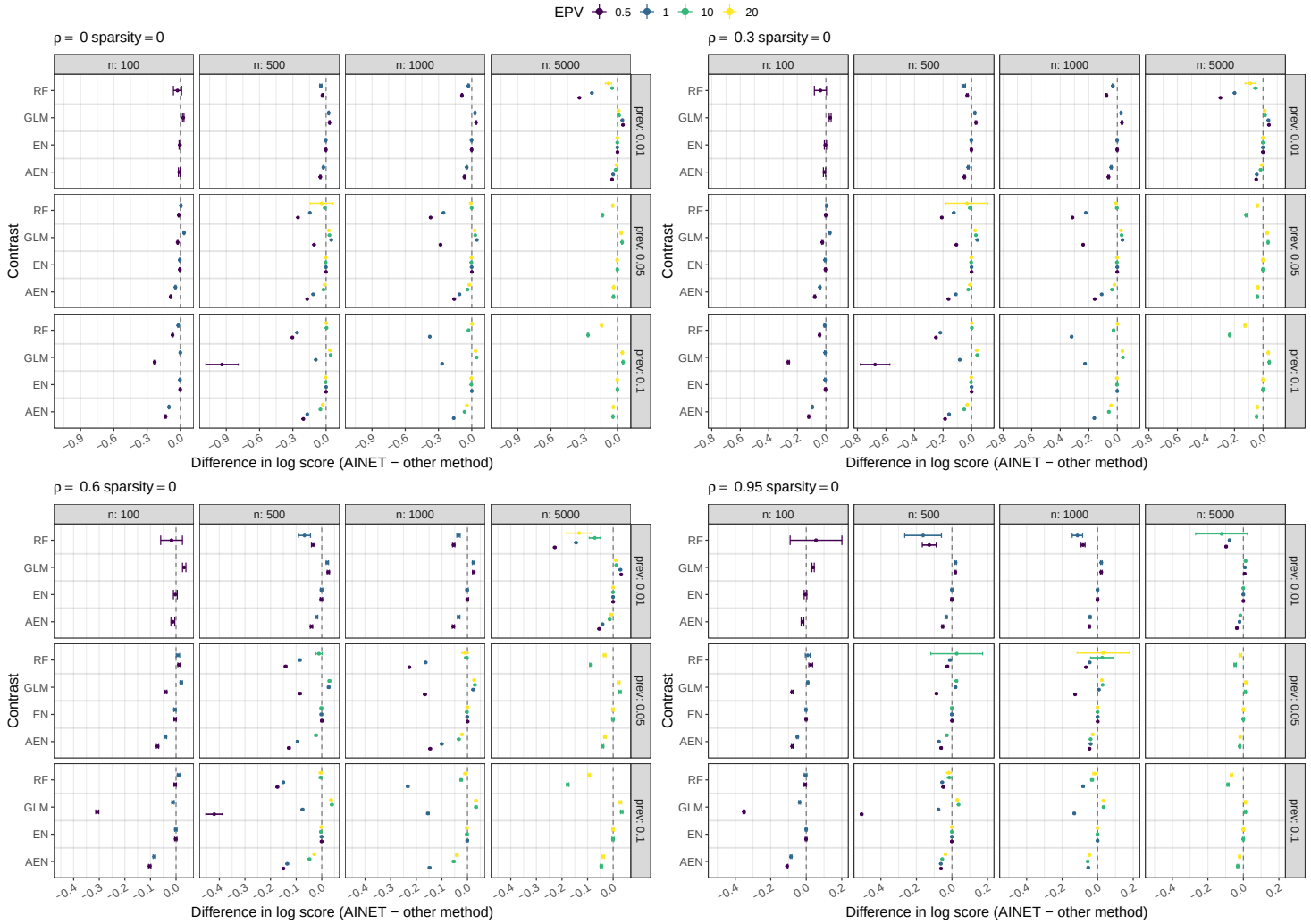

Figure 3: Tie-fighter plot for the difference in log-score between any method on the y-axis and AINET. The 95% confidence intervals are adjusted per simulation condition using the single-step method. Lower values indicate better performance of AINET.

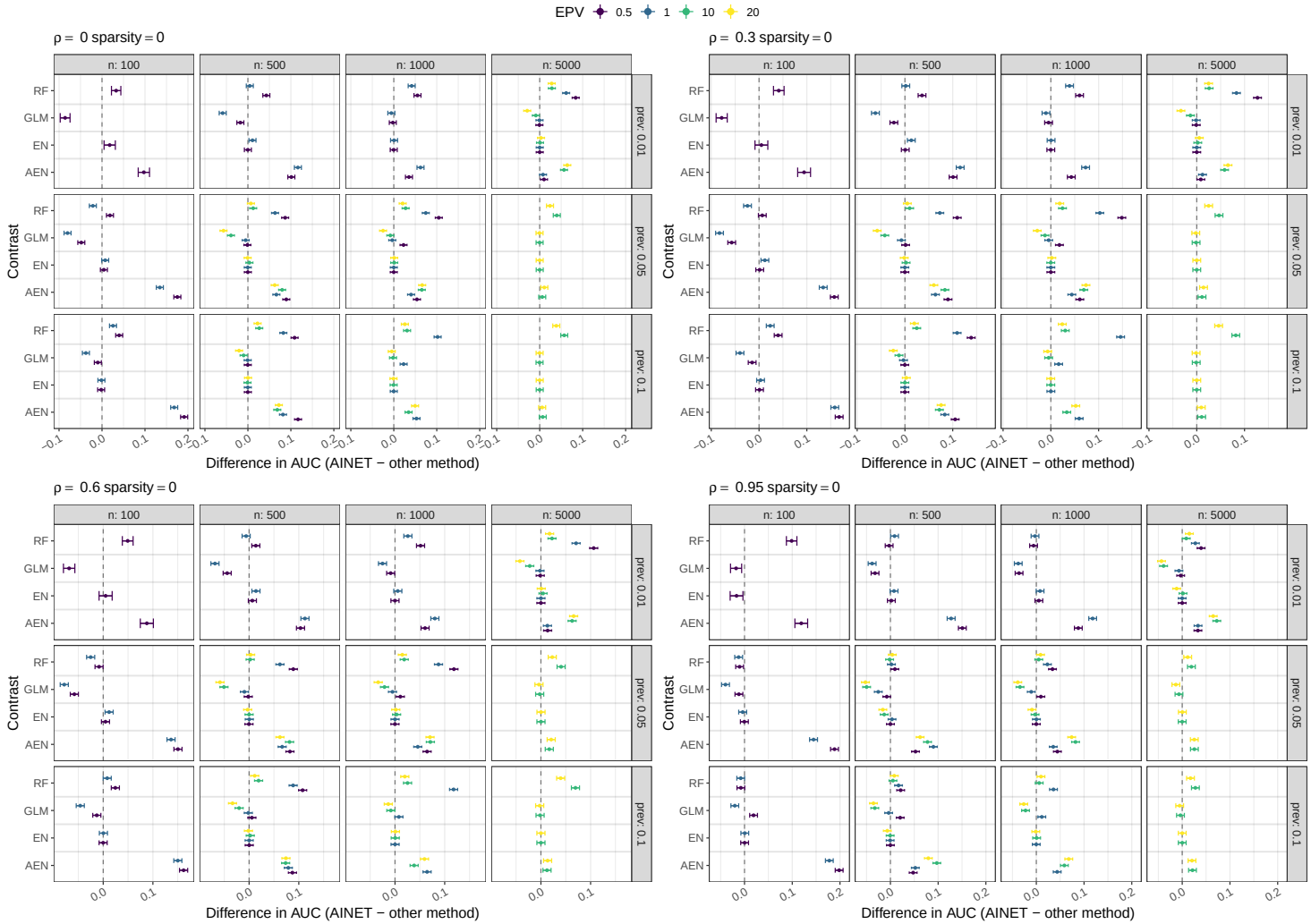

Figure 4: Tie-fighter plot for the difference in AUC between any method on the y-axis and AINET. The 95% confidence intervals are adjusted per simulation condition using the single-step method. Higher values indicate better performance of AINET.

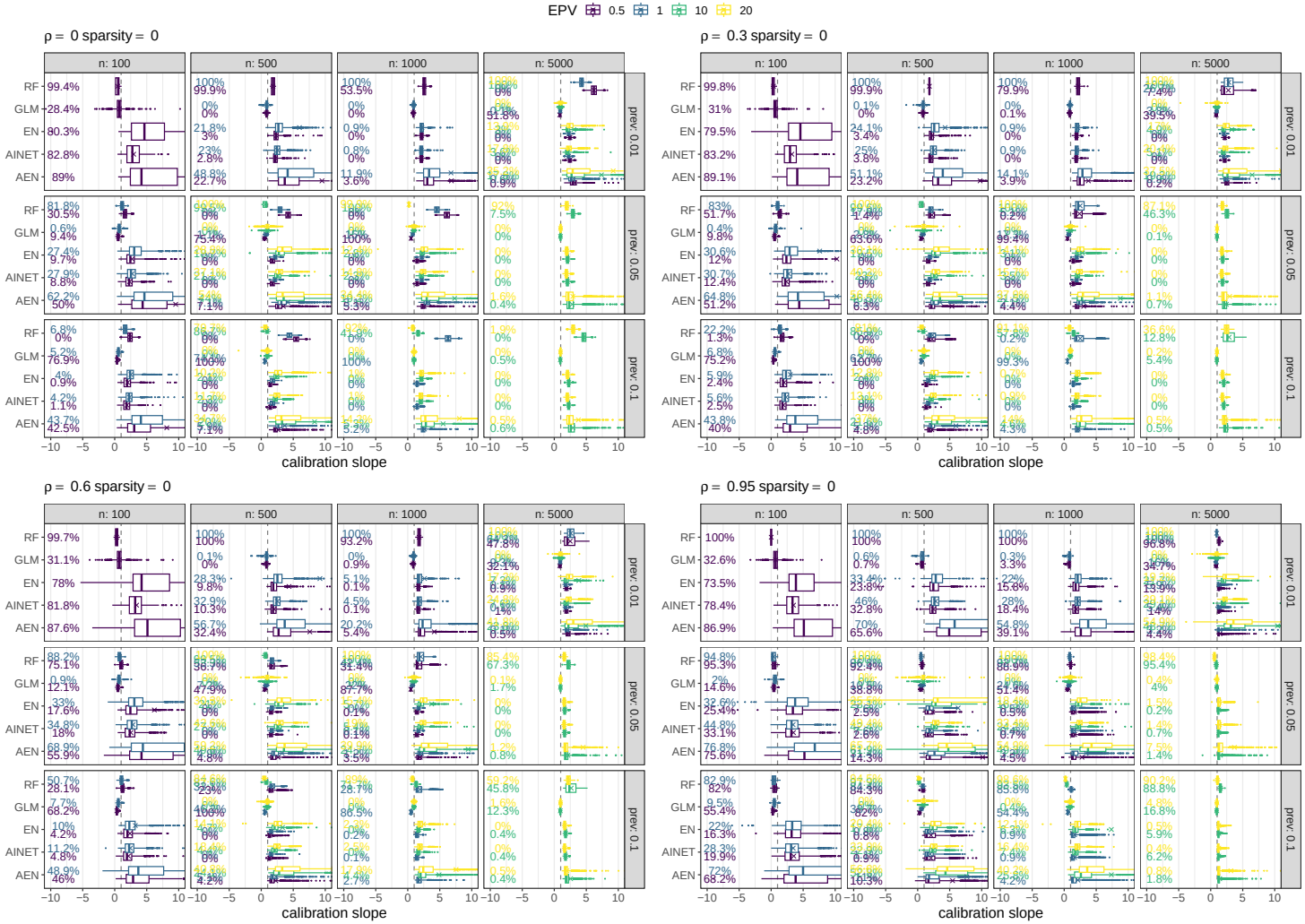

Figure 5: Boxplots of calibration slopes stratified by method and simulation conditions. Mean calibration slope is indicated by a cross. A value of one indicates optimal calibration. Percentage of simulations where calibration slope could not be estimated (due to extreme predictions or complete separation) are also indicated.

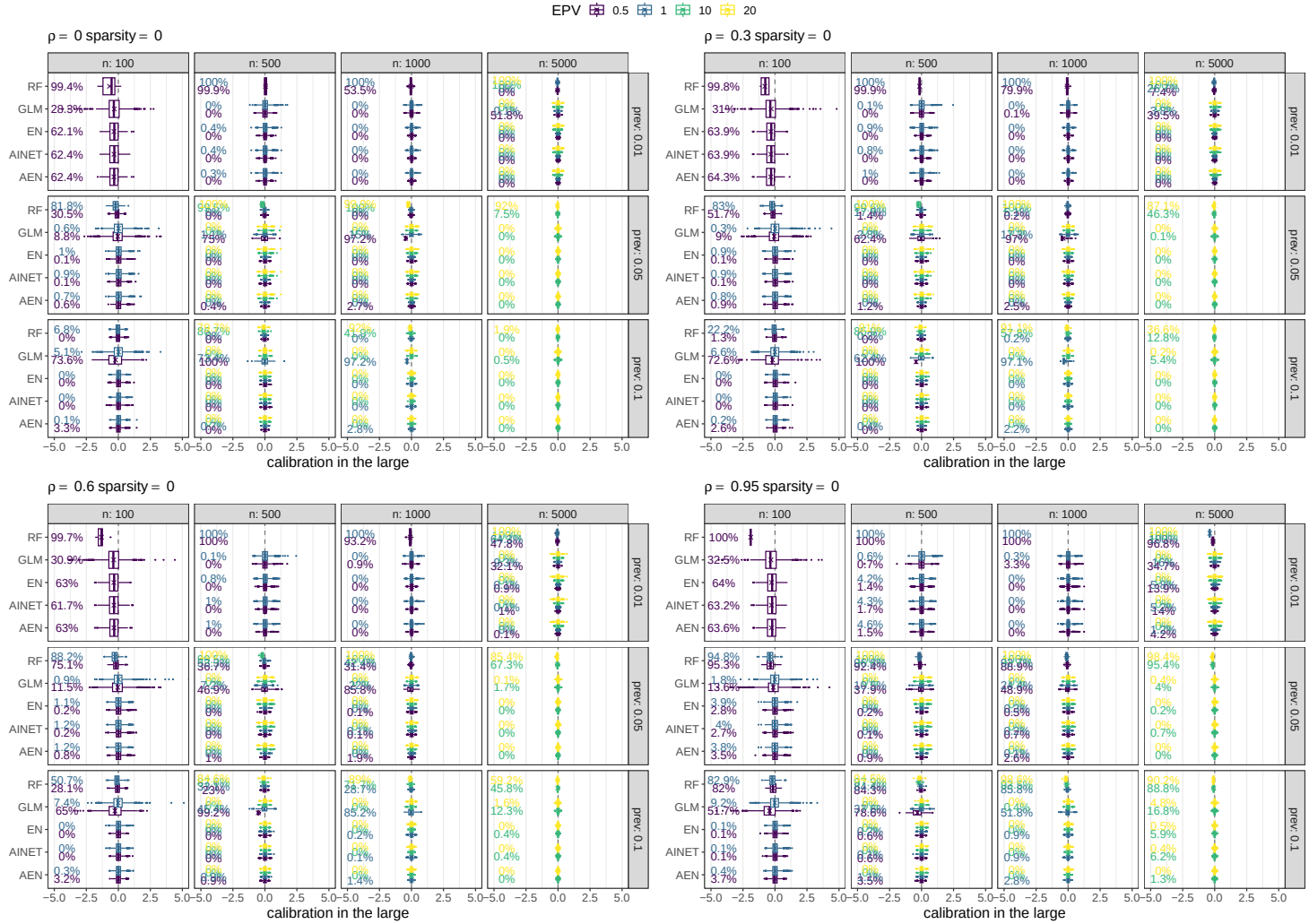

Figure 6: Boxplots of calibration in the large stratified by method and simulation conditions. Mean calibration in the large is indicated by a cross. A value of zero indicates optimal calibration in the large. Percentage of simulations where calibration in the large could not be estimated (due to extreme predictions or complete separation) are also indicated.