Supplementary Material to "A computationally fast variable importance test for random forests for high-dimensional data"

Advances in Data Analysis and Classification

Silke Janitza Ender Celik Anne-Laure Boulesteix

Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, D-81377 Munich, Germany.

1 Real data sets

Prostate Cancer Data (Singh et al.; 2002): From 1995 to 1997 samples of prostate tumors and adjacent non-tumor prostate tissue were collected from patients undergoing radical prostatectomy at the Brigham and Women's Hospital. High-quality expression profiles were obtained from 50 non-tumor prostate samples and 52 tumor specimens. The oligonucleotide microarrays contained probes for approximately 12600 genes. We obtained this data set from the website `http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html`.

Breast Cancer Data (van't Veer et al.; 2002): We considered the data set that was previously analyzed by Díaz-Uriarte and De Andres (2006) and made publicly available at the website `http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html`. In this data set there were 33 patients that developed distant metastases within 5 years and 44 that remained disease-free for over 5 years. Missing data was imputed by using 5-nearest neighbor imputation. Further details on transformations of the original data are given in the supplement to the paper of Díaz-Uriarte and De Andres (2006).

Leukemia Data (Golub et al.; 1999): The Leukemia Data consists of 47 patients with acute lymphoblastic leukemia (ALL) and 25 patients with acute myeloid leukemia (AML). The considered data set comprises both, training samples and test samples from Golub et al. (1999) and was retrieved from the Bioconductor package golubEsets. The samples were assayed using Affymetrix Hgu6800 chips and data on the expression of 7129 genes are available.

Colon Cancer Data (Alon et al.; 1999): In this data set, expression levels of 40 tumor and 22 normal colon tissues for 6500 human genes are measured. The considered data set contains the expression of the 2000 genes with highest minimal intensity across the 62 tissues measured using the Affymetrix technology. We obtained this data set from the website `http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html`.

Embryonal Tumor Data (Pomeroy et al.; 2002): This data set includes 60 patients with embryonal tumors of the central nervous system from whom biopsies were obtained before receiving treatment. The data was used to differentiate between patients who are alive after treatment ($n = 21$) and those who succumbed to their disease ($n = 39$) (data set C in the paper by Pomeroy et al.; 2002). RNA was extracted from frozen specimens and was analysed with oligonucleotide microarrays containing 7129 probes from 6817 genes. We obtained the data from the website `http://datam.i2r.a-star.edu.sg/datasets/krbd/NervousSystem/NervousSystem.html`.

2 Further results of simulation studies

2.1 Studies with complete predictor space

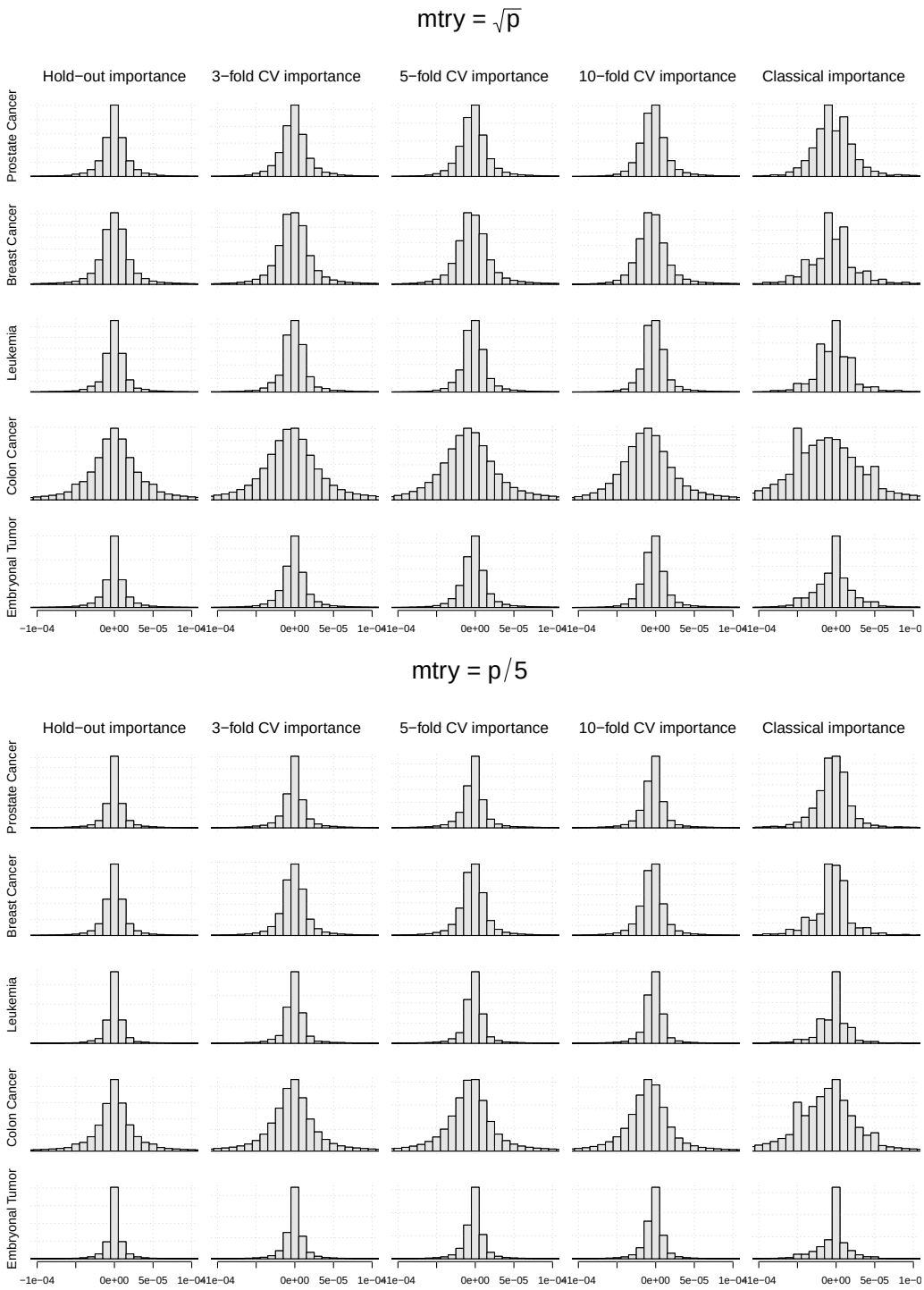

Figure 1: Variable importance null distribution when using the hold-out permutation variable importance measure, the cross-validated importance measure with $k = 3$, $k = 5$, and $k = 10$ and the classical permutation variable importance measure and setting `mtry` to $\sqrt{100}$ (upper) and $\frac{p}{5}$ (lower).

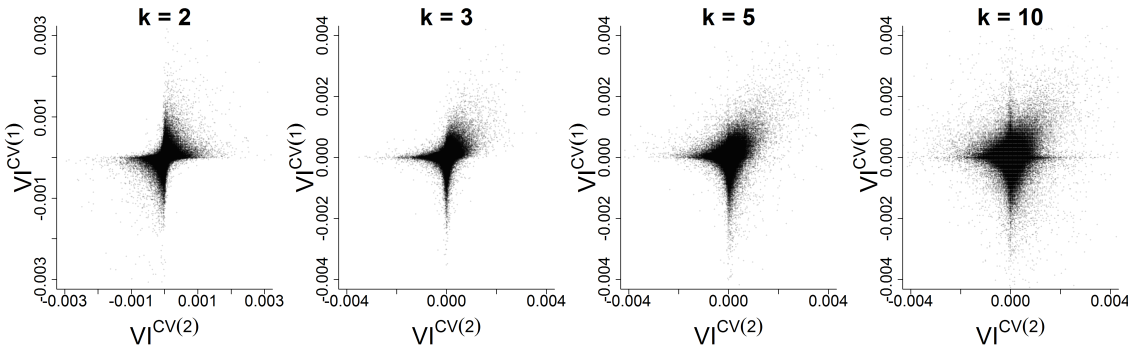

Figure 2: Fold-specific variable importance for the first fold plotted against fold-specific variable importance for the second fold for all variables $X_1, \ldots, X_{2000}$ of Study I (no relevant variables; `mtry` $= \sqrt{p}$) for the Colon Cancer data with $k = 2$, $k = 3$, $k = 5$ and $k = 10$. Five-hundred repetitions of Study I were performed, yielding a total of 2000×500 points shown in each plot.

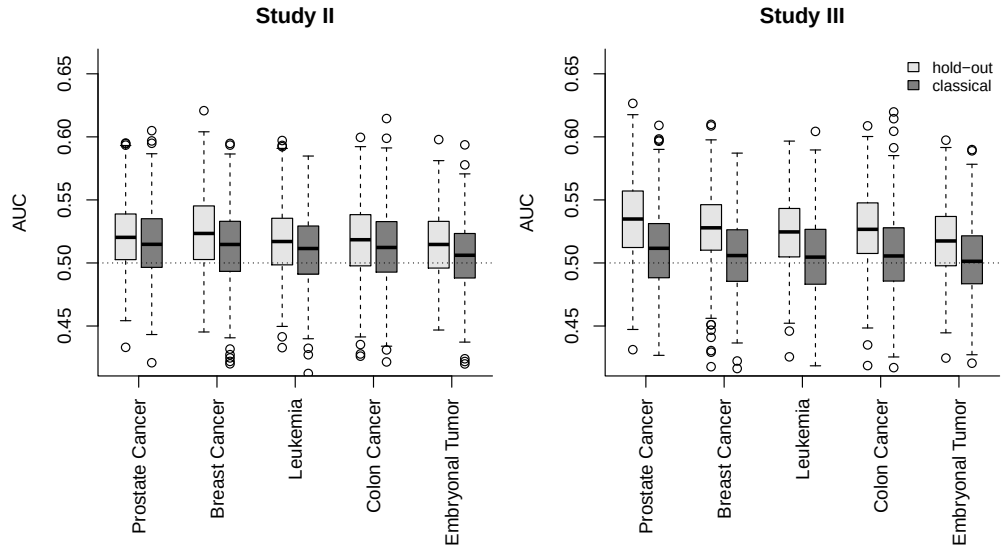

Figure 3: Discriminative ability of the novel hold-out permutation variable importance measure and the classical permutation variable importance measure. Discriminative ability is measured by the area under the curve for Studies II and III (`mtry` always set to $\frac{p}{5}$). Values of 0.5 indicate no discriminative ability (horizontal dotted line).

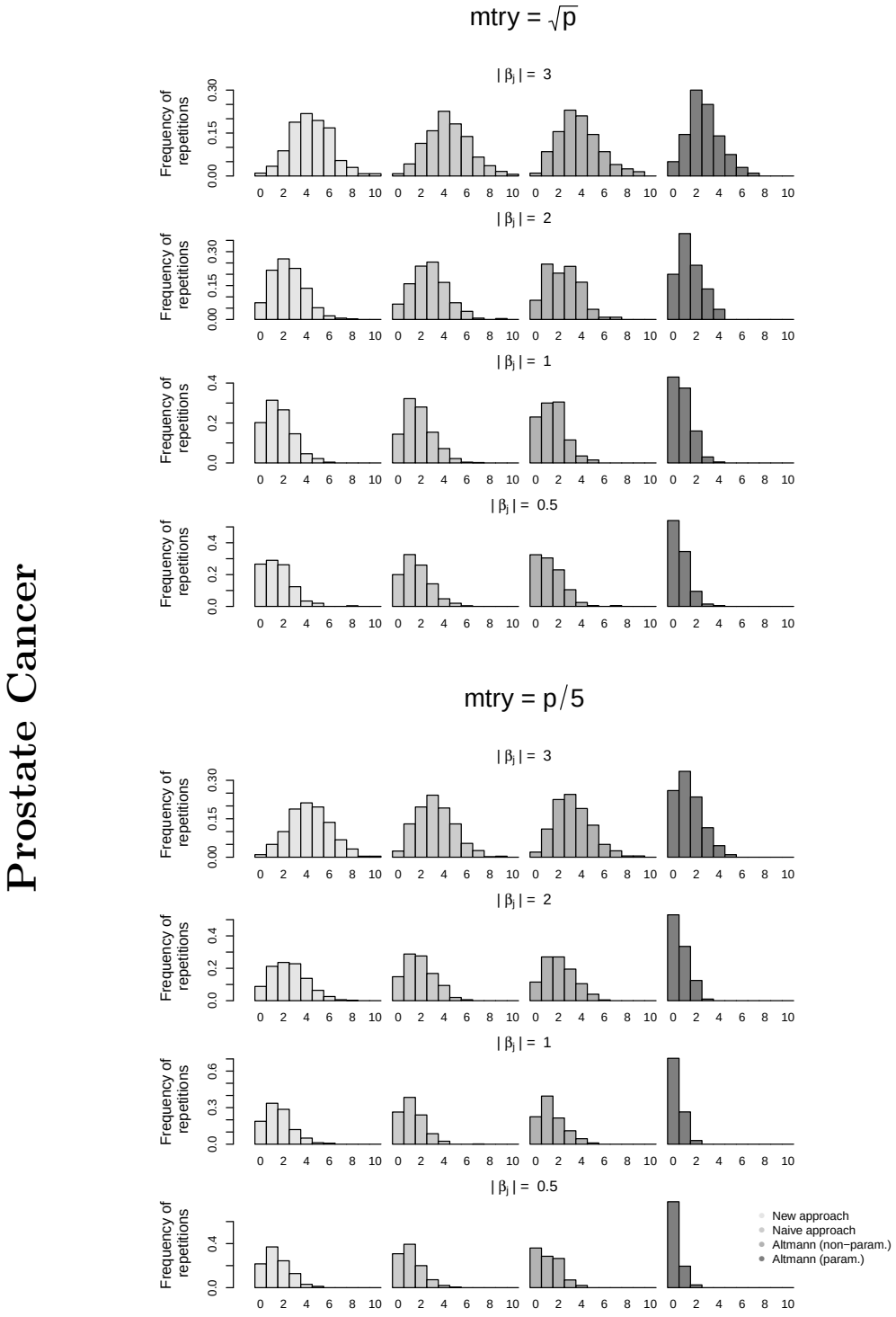

Figure 4: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using our new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with `mtry` set to $\sqrt{p}$ (upper) and $\frac{p}{5}$ (lower).

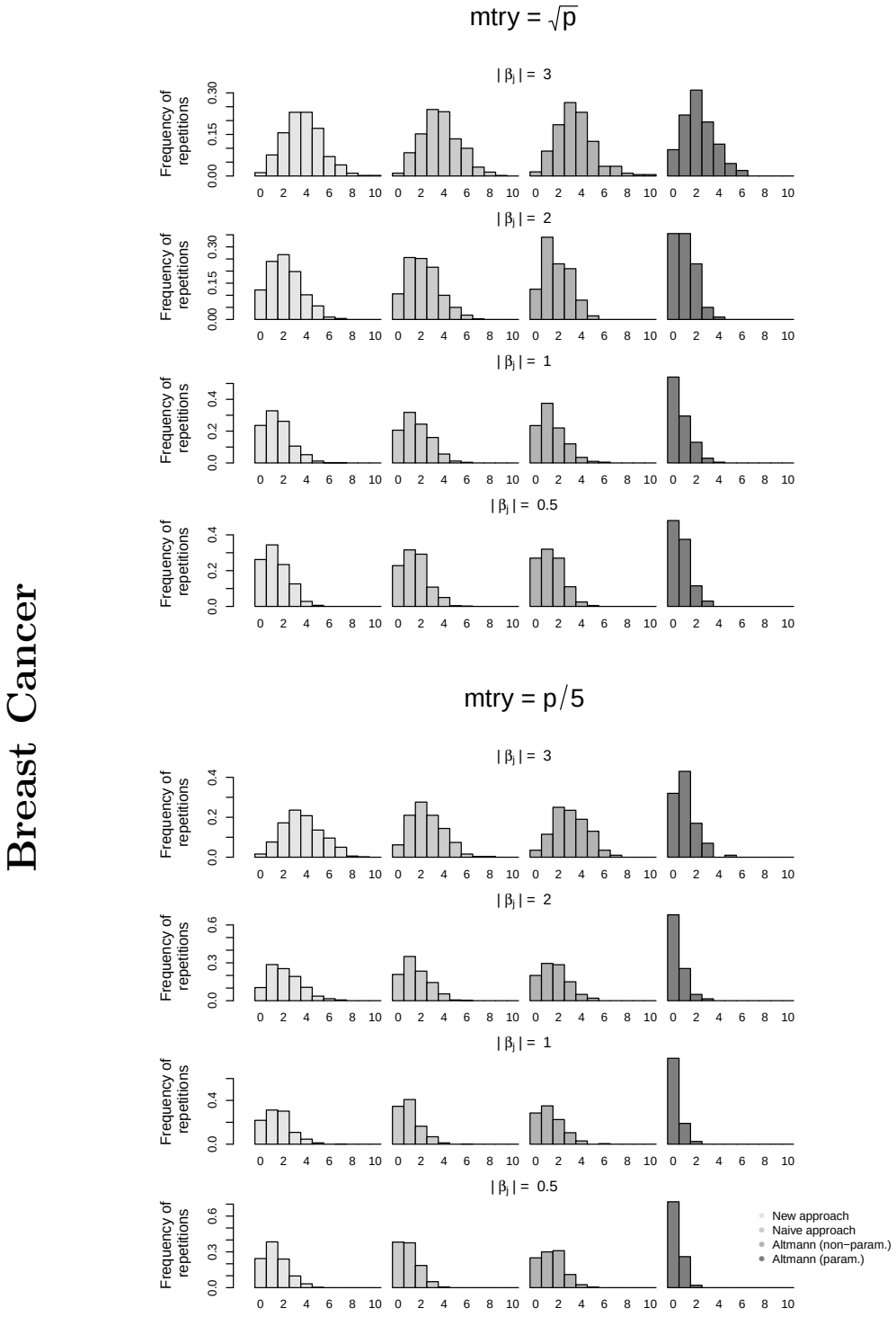

Figure 5: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using our new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with `mtry` set to $\sqrt{p}$ (upper) and $\frac{p}{5}$ (lower).

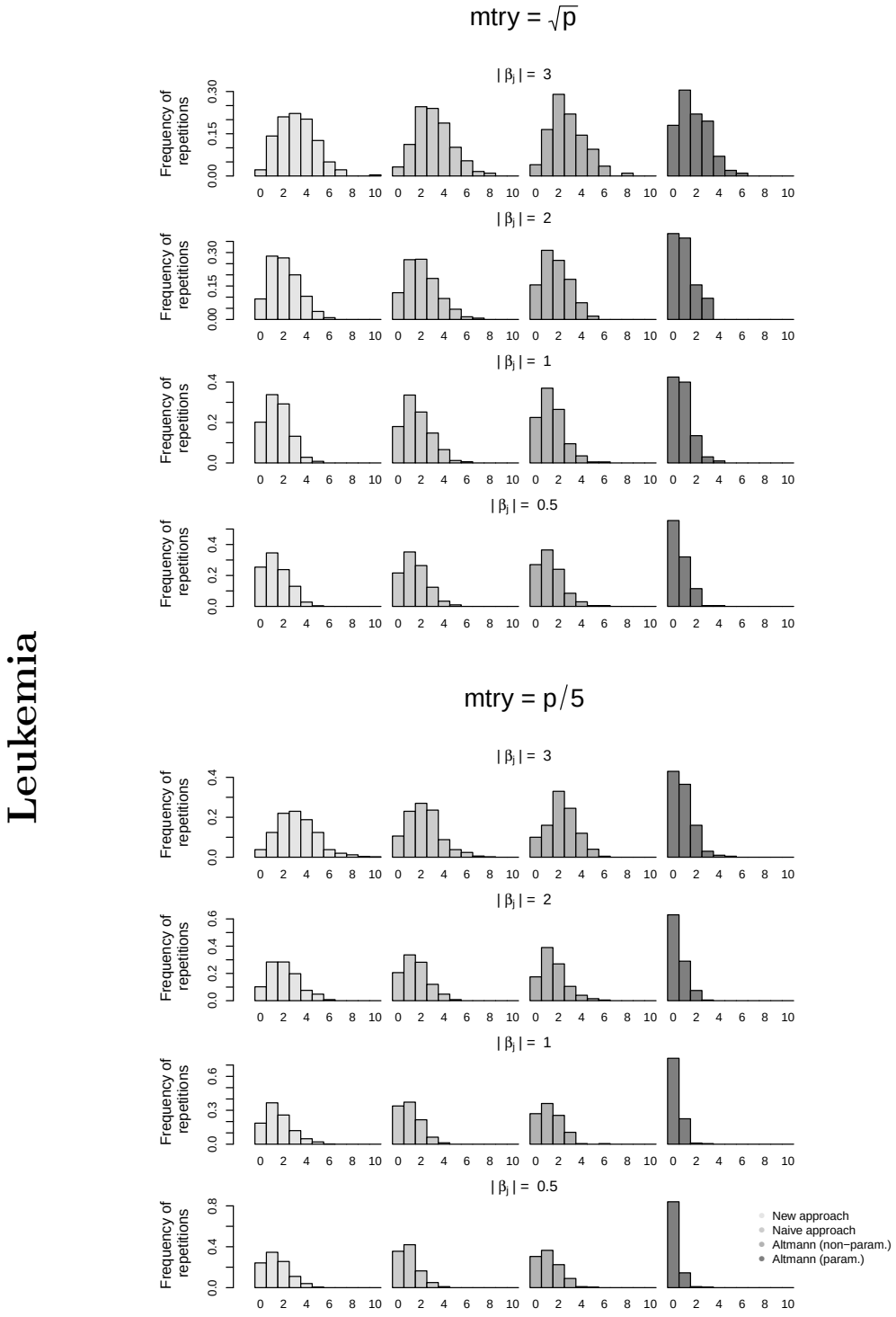

Figure 6: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using our new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with `mtry` set to $\sqrt{p}$ (upper) and $\frac{p}{5}$ (lower).

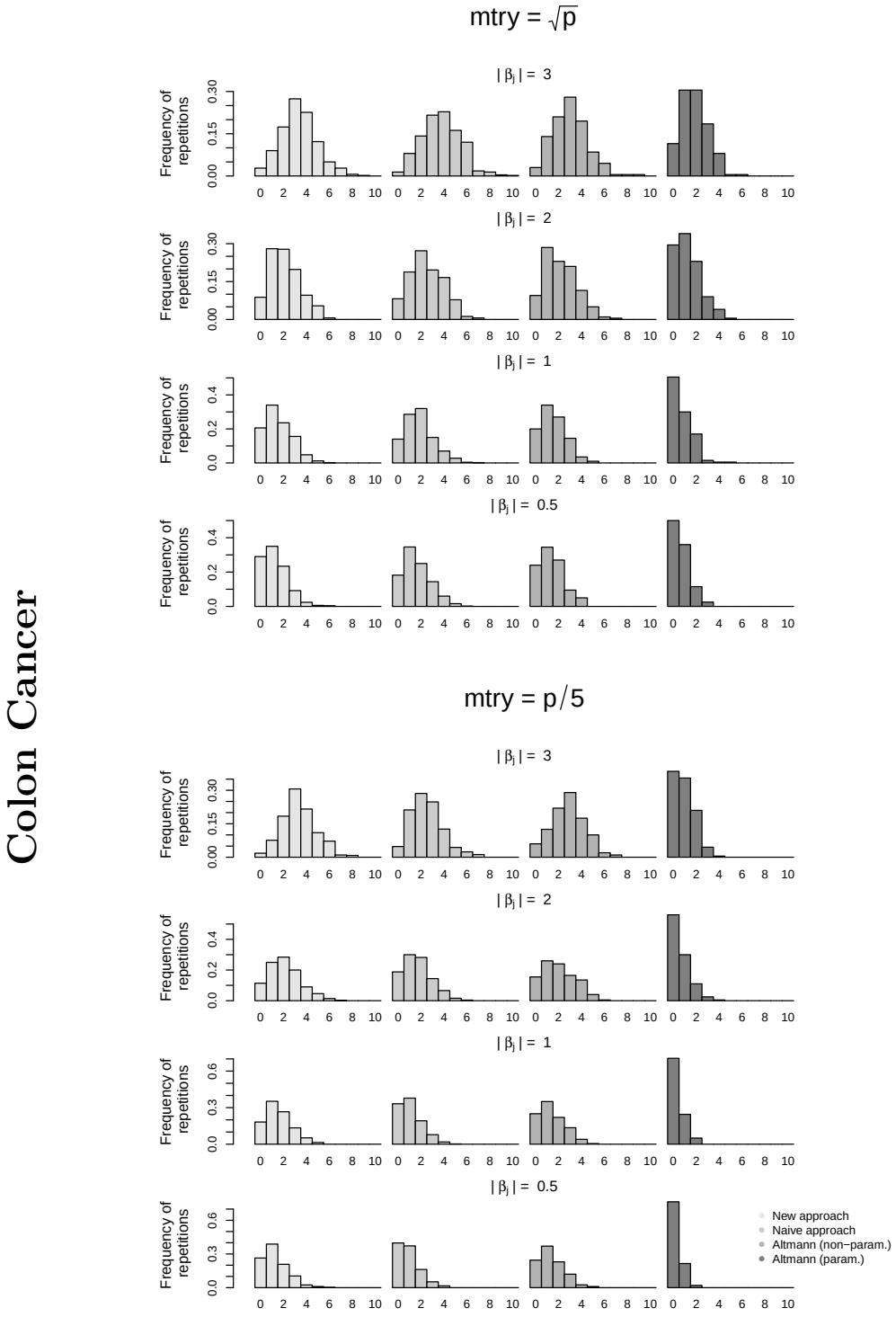

Figure 7: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using our new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with `mtry` set to $\sqrt{p}$ (upper) and $\frac{p}{5}$ (lower).

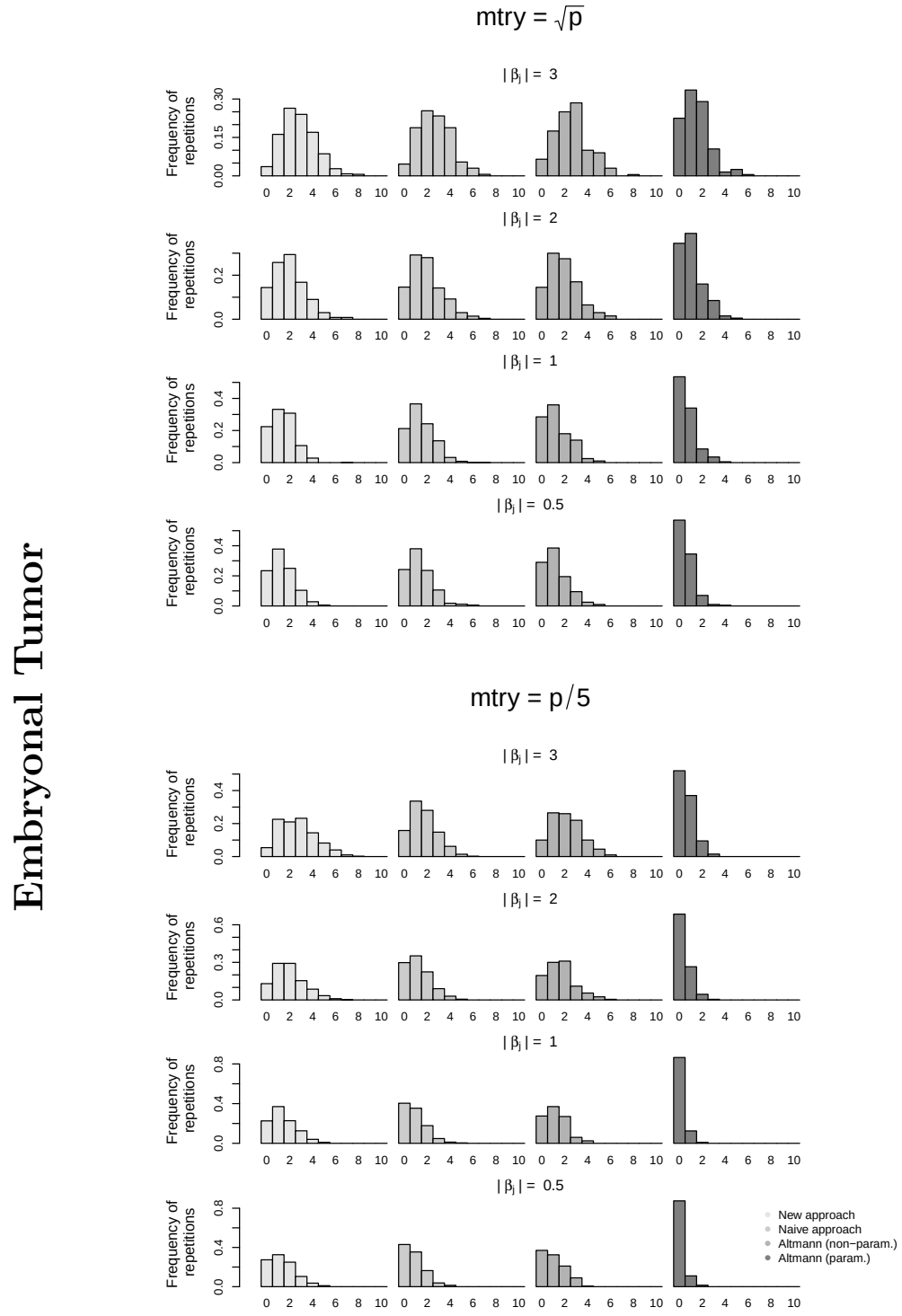

Figure 8: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using our new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with `mtry` set to $\sqrt{p}$ (upper) and $\frac{p}{5}$ (lower).

2.2 Studies with reduced predictor space ($p = 100$)

2.2.1 Study I

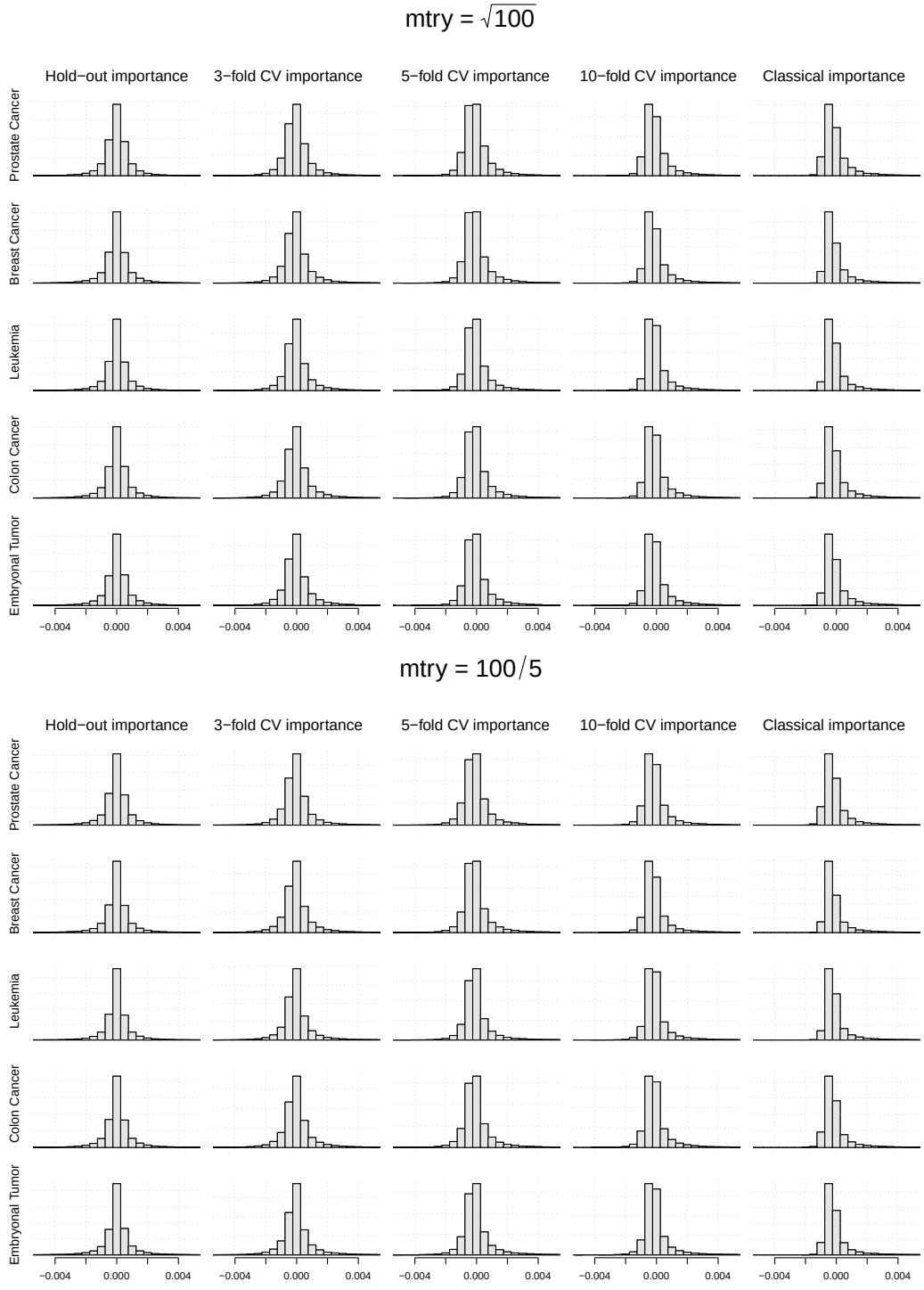

Figure 9: Variable importance null distribution when using the hold-out permutation variable importance measure, the cross-validated importance measure with $k = 3$, $k = 5$, and $k = 10$ and the classical permutation variable importance measure and setting `mtry` to $\sqrt{100}$ (upper) and $\frac{100}{5}$ (lower).

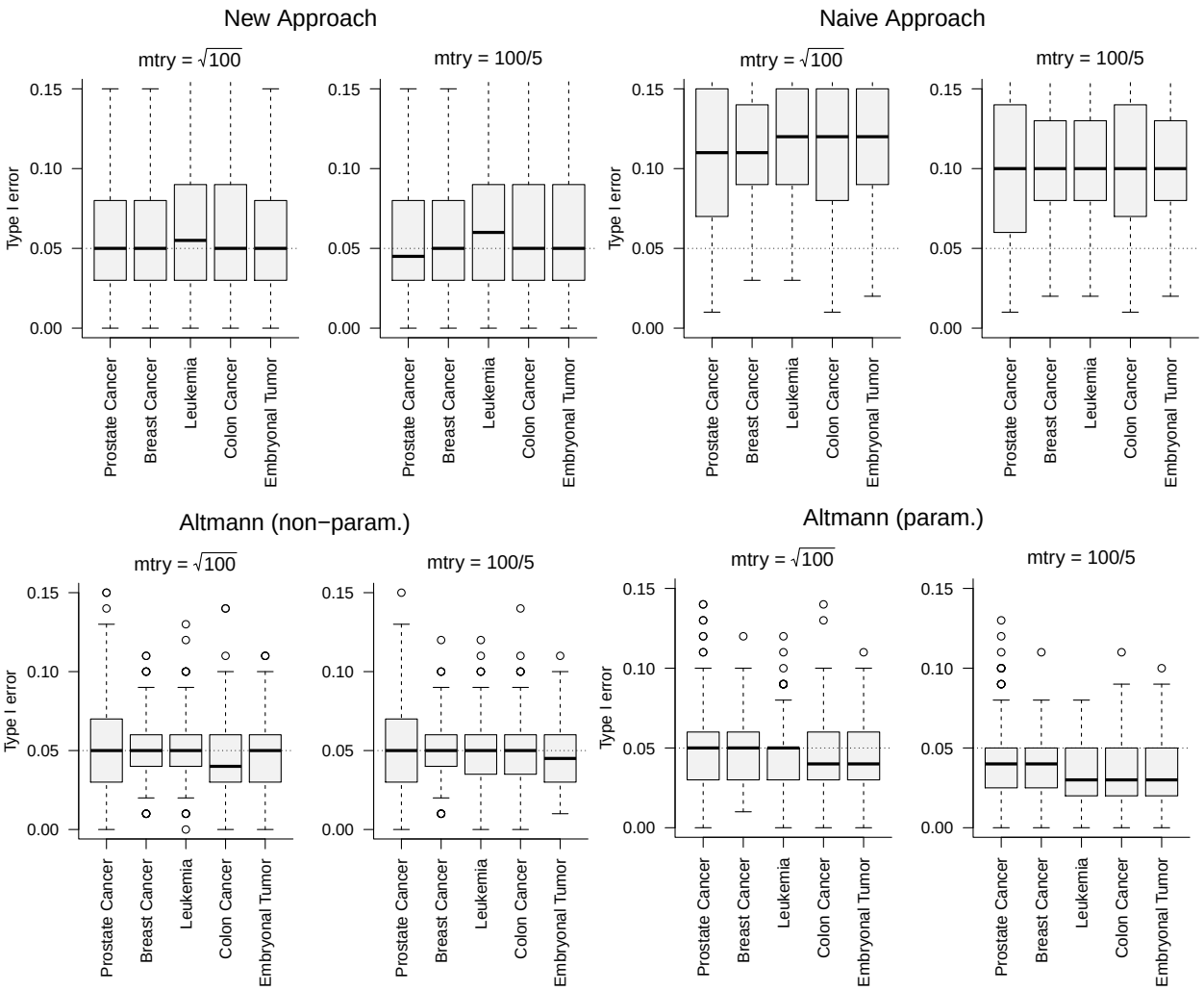

Figure 10: Type I error in Study I for the new testing approach (which uses the hold-out permutation variable importance measure), the naive testing approach (which uses the classical permutation variable importance measure) and the approach of Altmann et al. (2010) (non-parametric and parametric). Hypothesis tests were performed at significance level $\alpha = 0.05$ (dotted horizontal line).

2.2.2 Study II

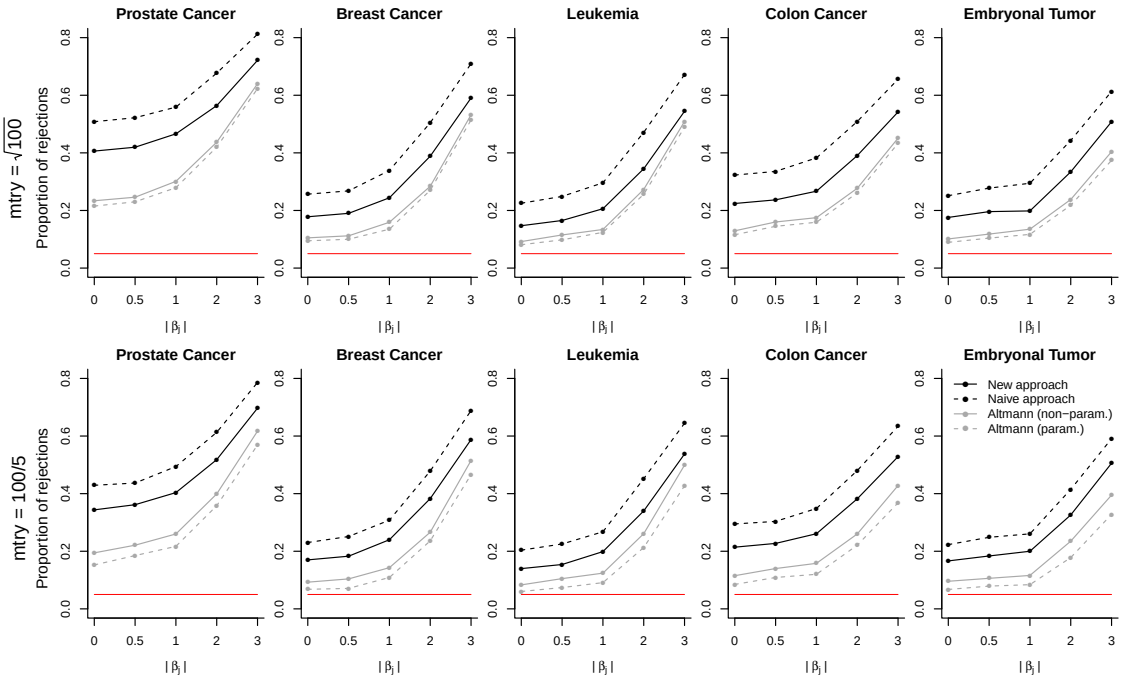

Figure 11: Proportion of rejected null hypothesis among predictor variables with specified absolute effect size. The mean proportions over 500 (200 for the approach of Altmann et al. (2010), resp.) repetitions of Study II are shown when using our novel approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with `mtry` set to $\sqrt{100}$ (upper panel) and $\frac{100}{5}$ (lower panel). The red horizontal line represents the 5% significance level.

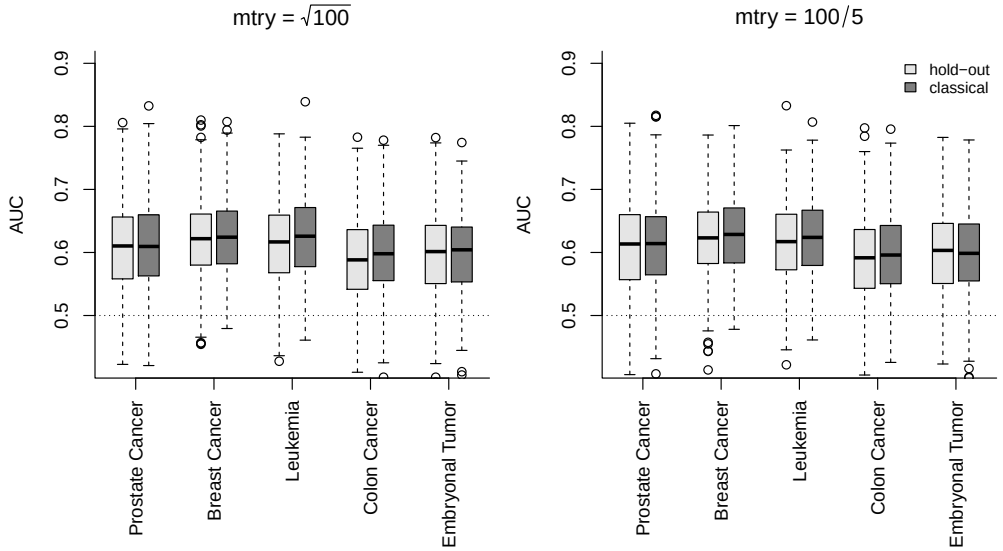

Figure 12: Discriminative ability of the novel hold-out permutation variable importance measure and the classical permutation variable importance measure. Discriminative ability is measured by the area under the curve. Results are shown for `mtry` set to $\sqrt{100}$ (left) and $\frac{100}{5}$ (right). Values of 0.5 indicate no discriminative ability (horizontal dotted line).

2.2.3 Study III

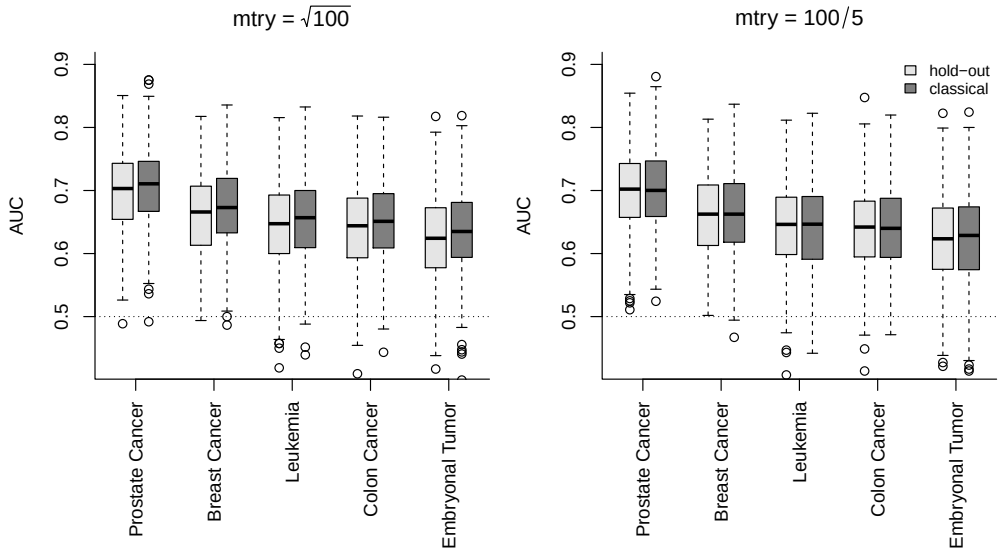

Figure 13: Discriminative ability of the novel hold-out permutation variable importance measure and the classical permutation variable importance measure. Discriminative ability is measured by the area under the curve. Results are shown for `mtry` set to $\sqrt{100}$ (left) and $\frac{100}{5}$ (right). Values of 0.5 indicate no discriminative ability (horizontal dotted line).

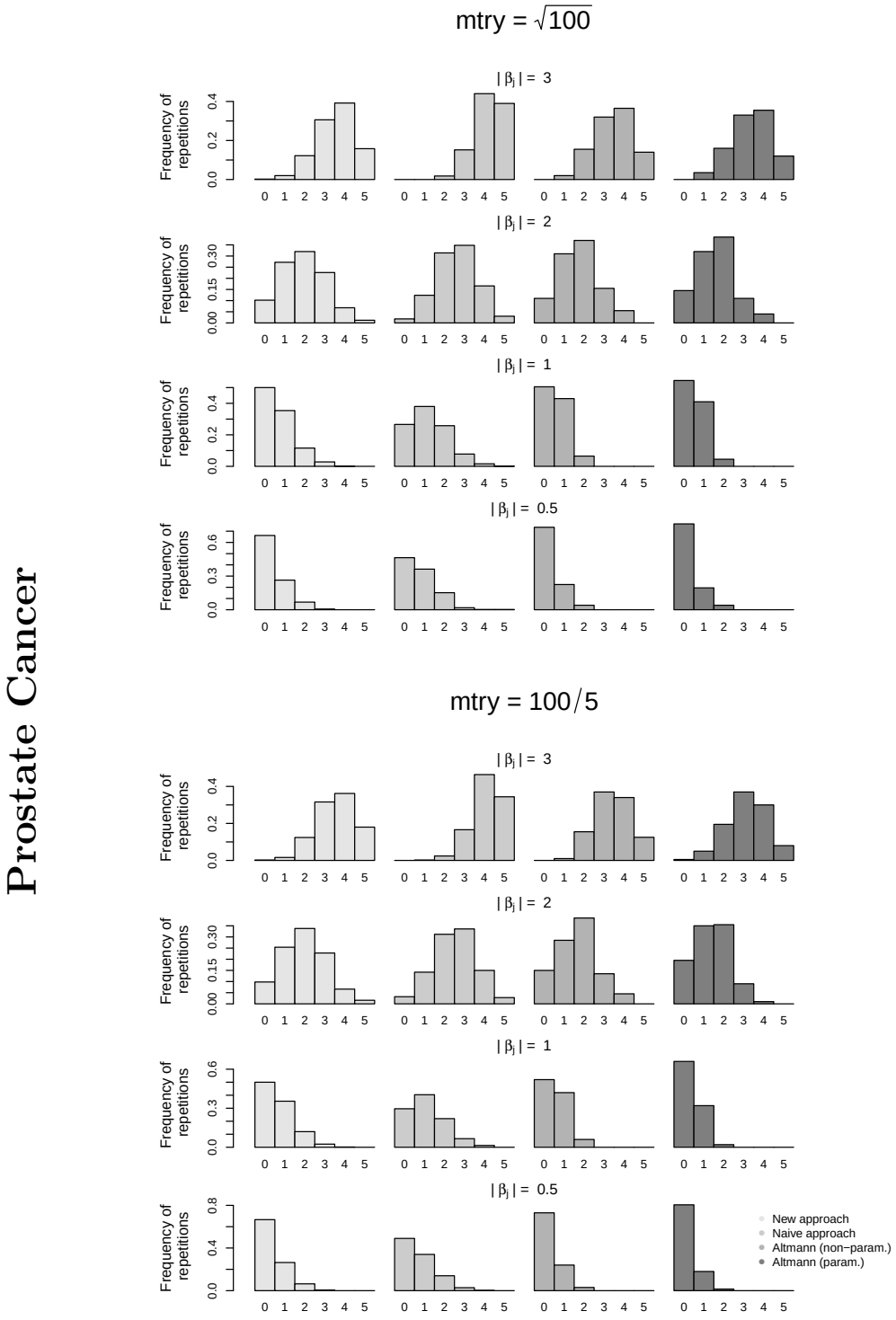

Figure 14: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using our new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with `mtry` set to $\sqrt{100}$ (upper) and $\frac{100}{5}$ (lower).

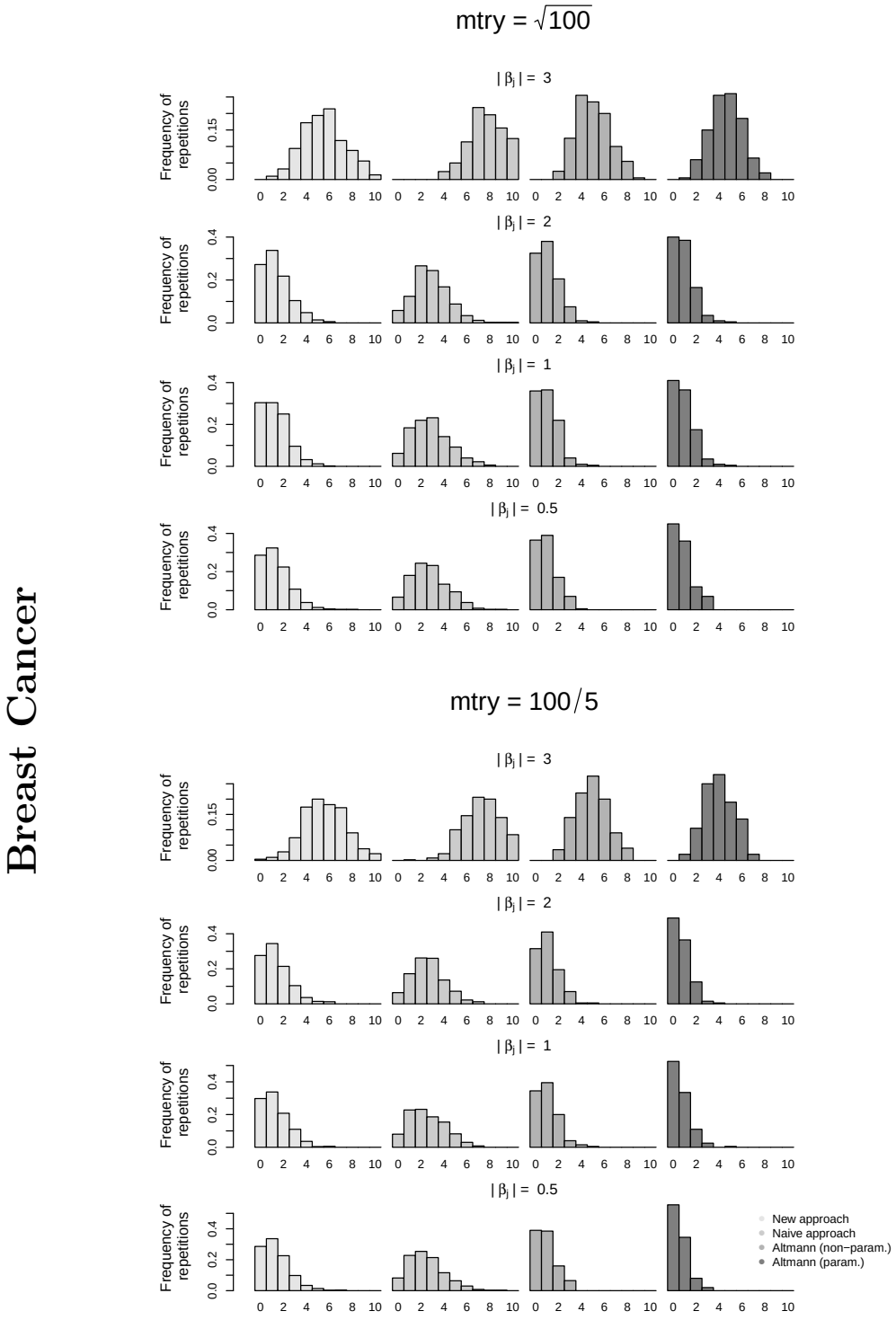

Figure 15: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using our new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with `mtry` set to $\sqrt{100}$ (upper) and $\frac{100}{5}$ (lower).

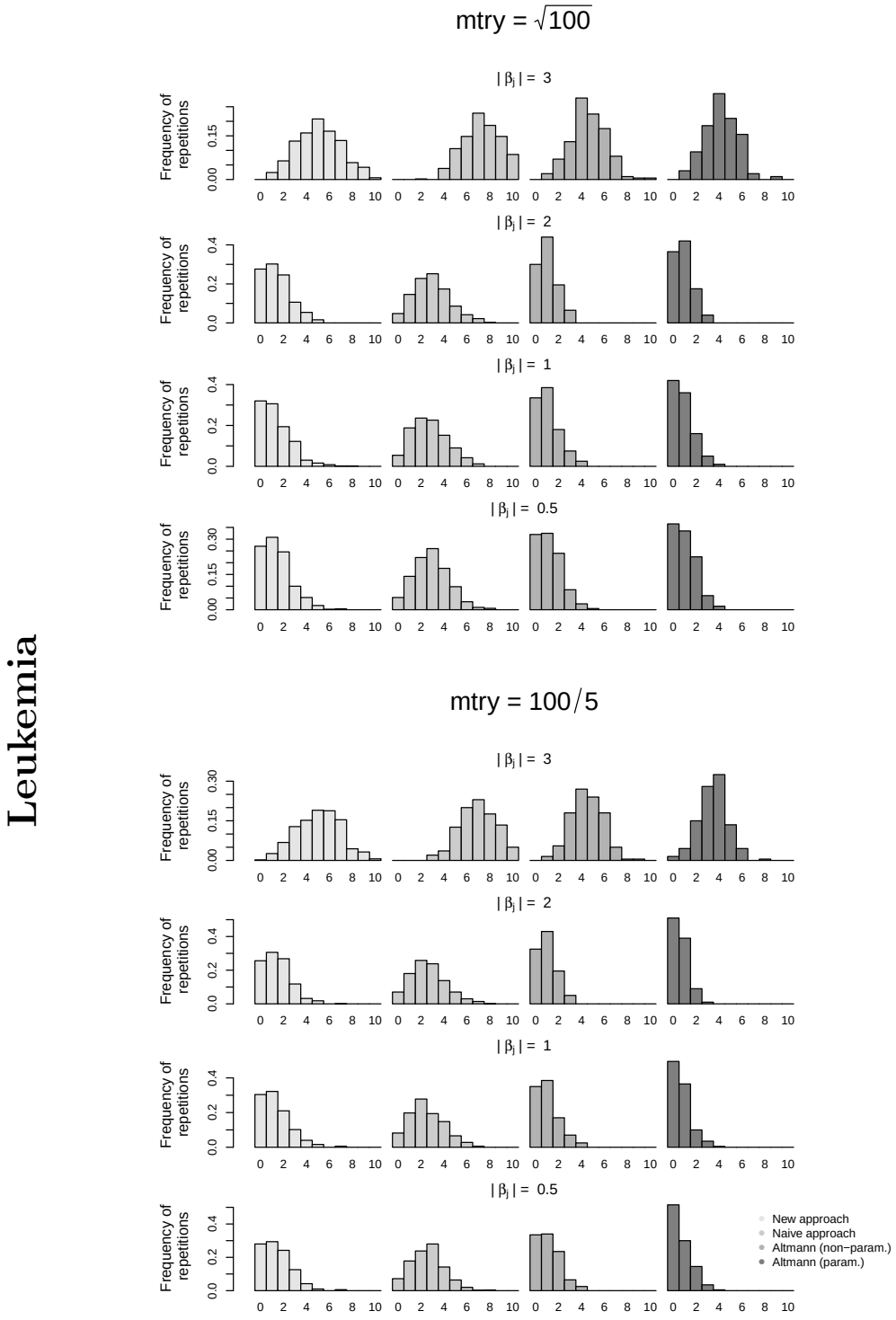

Figure 16: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using our new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with `mtry` set to $\sqrt{100}$ (upper) and $\frac{100}{5}$ (lower).

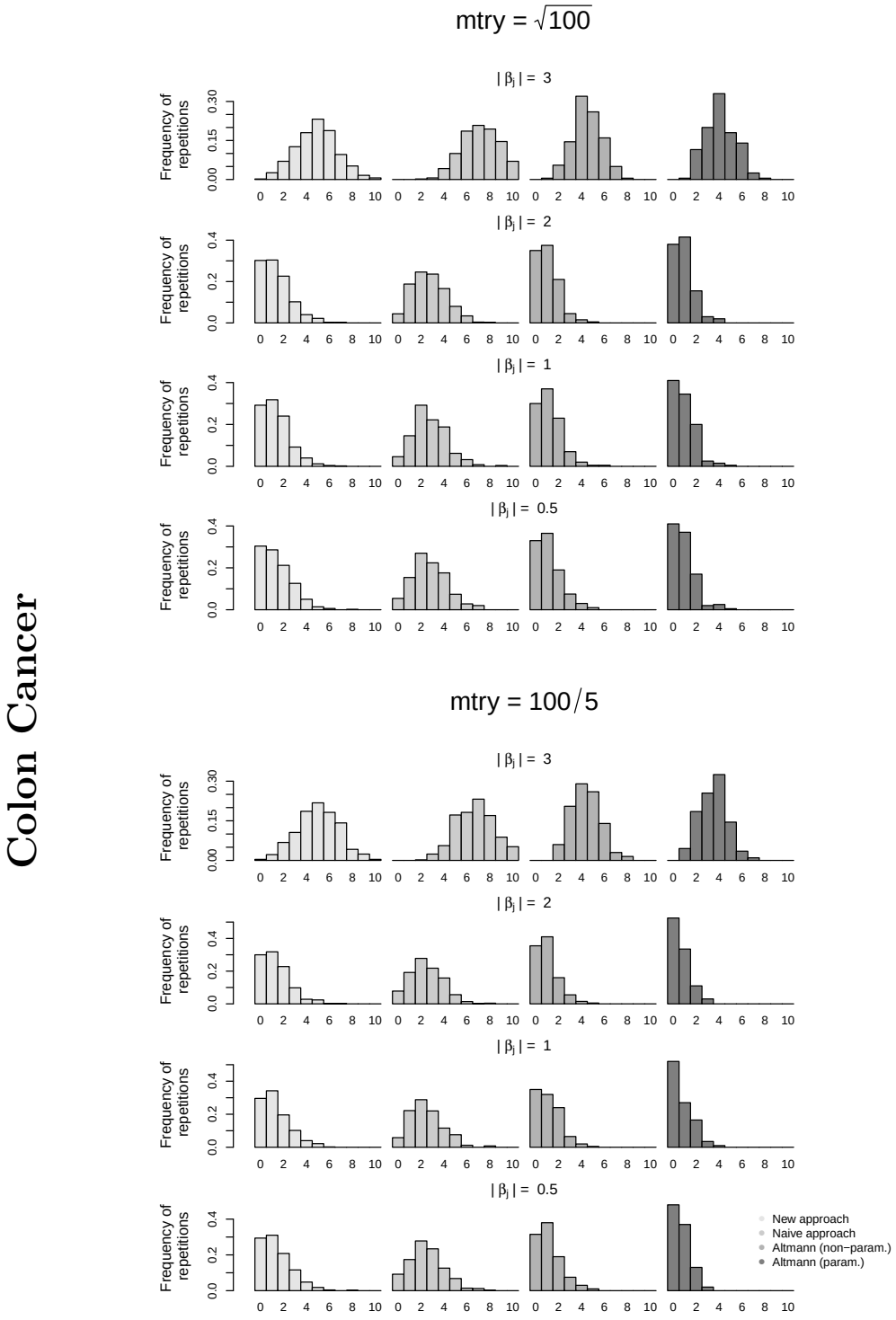

Figure 17: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using our new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with `mtry` set to $\sqrt{100}$ (upper) and $\frac{100}{5}$ (lower).

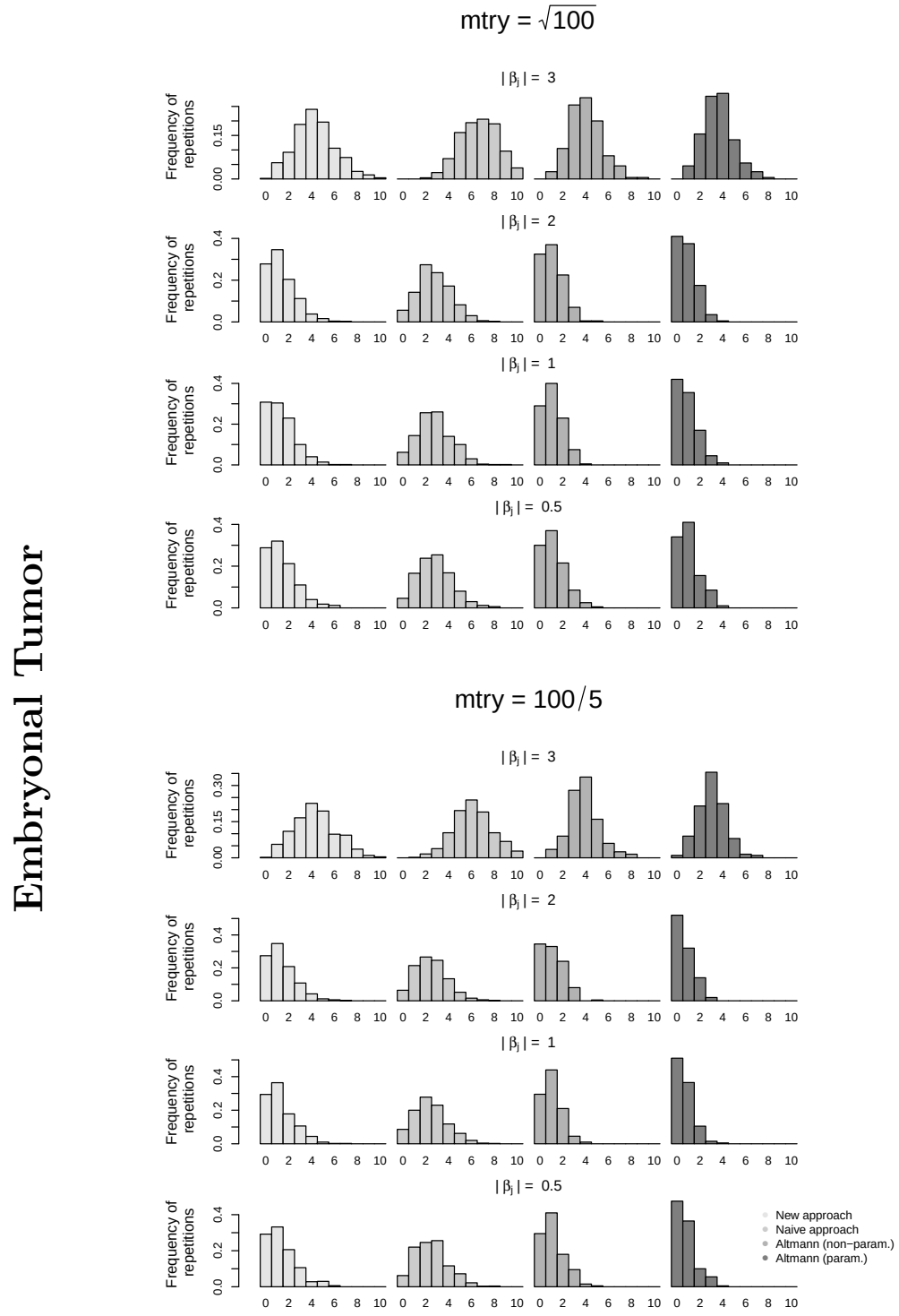

Figure 18: Relative frequency of repetitions of Study III in which the specified number of variables with effect was selected (i.e., variables with p-value below $\alpha = 0.05$). Distributions are shown for variables with specified absolute effect size and when using our new approach, the naive approach and the approach of Altmann et al. (2010) (non-parametric and parametric), with `mtry` set to $\sqrt{100}$ (upper) and $\frac{100}{5}$ (lower).

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Sciences* **96**(12): 6745–6750.

Altmann, A., Toloşi, L., Sander, O. and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure, *Bioinformatics* **26**(10): 1340–1347.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest, *BMC Bioinformatics* **7**(1): 3.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**(5439): 531–537.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., Goumnerova, L. C., Black, P. M., Lau, C. et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature* **415**(6870): 436–442.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P. et al. (2002). Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell* **1**(2): 203–209.

van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer, *Nature* **415**(6871): 530–536.