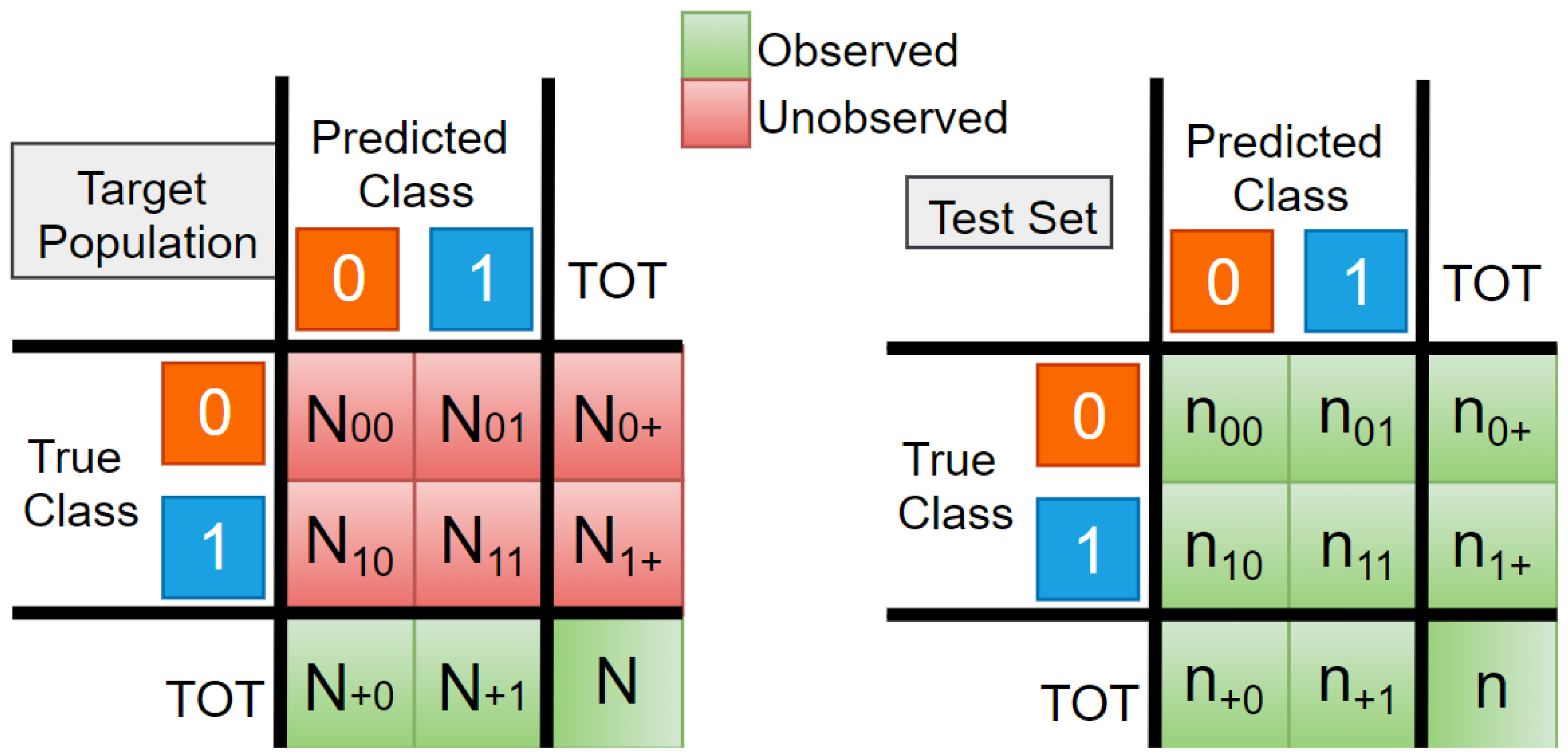
Comparing Correction Methods to reduce Misclassification Bias

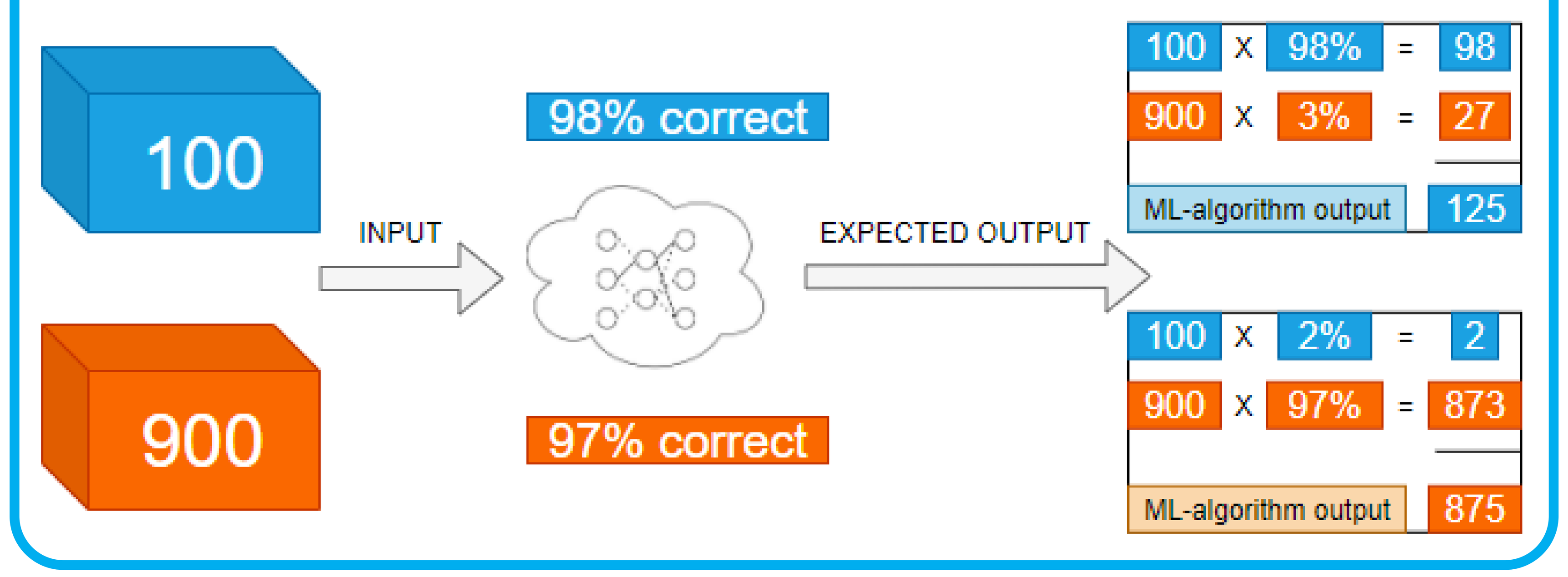
Kevin Kloos*, Quinten Meertens**, Sander Scholtus and Julian Karch
Leiden University, Statistics Netherlands
*k.kloos.2@umail.leidenuniv.nl, **q.a.meertens@uva.nl

1. Problem statement

- **Misclassification bias** is the statistical bias that occurs when aggregating the predictions of a machine learning algorithm.
- The **base rate** refers to the relative occurrence of the class of interest in the *target population*.
- Our aim is to reduce the misclassification bias of the base rate, while taking the variance into account.
- We compare five estimators in terms of their misclassification bias, variance and (R)MSE.



2. Statistically biased



3. Model specification

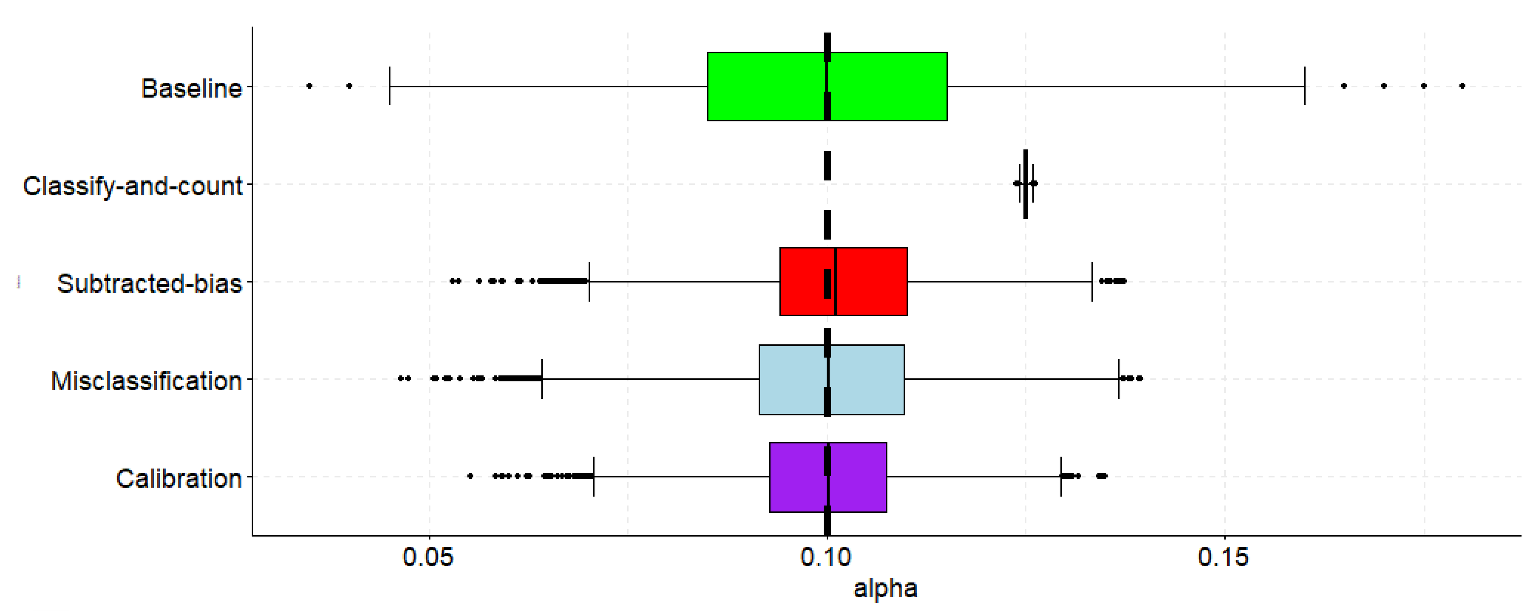
- Sample a test set from the target population.
- Both the observed labels as the true labels of the test set are observed.
- Only the predicted labels in the target population are observed, the true labels are unobserved.
- Estimate the classification probabilities of the target population with a test set, where the size of the test set is much smaller than the size of the target population.

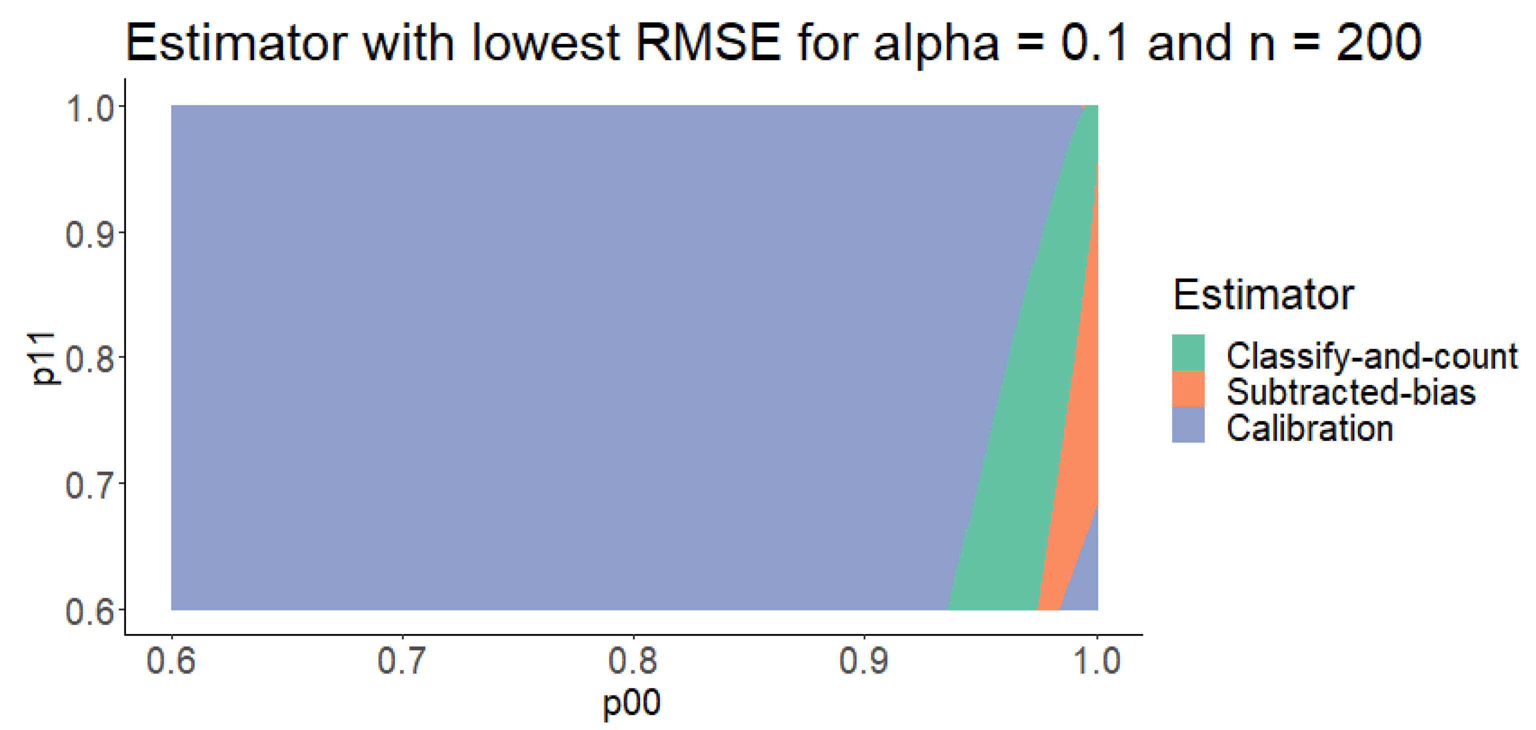
4. Defining the estimators

Estimator	Makes use of test set	Makes use of target population	Estimate classification probabilities	Estimate calibration probabilities	(Asymptotic) Unbiasedness
Baseline	✓	✗	✗	✗	✓
Classify-and-count	✗	✓	✗	✗	✗
Subtracting-bias	✓	✓	✓	✗	✗
Misclassification	✓	✓	✓	✗	✓
Calibration	✓	✓	✗	✓	✓

- **Baseline**: count true values of Class 1 in the test set.
- **Classify-and-count**: count predicted values of Class 1 in target population.
- **Subtracted-bias**: Estimate bias with plug-in estimators and correct. ([3] for more details)
- **Misclassification**: Use estimated classification probabilities ([1] for more details).
- **Calibration**: Use estimated calibration probabilities ([2] for more details).

5. Empirical Evaluation and Conclusion





Take home messages:

- **Classify-and-count** is unbiased only for specific combinations of p_{00}, p_{11} and α (see [3] for more information) and has therefore a low RMSE under these specific circumstances.
- **Calibration** estimator is the most stable in terms of RMSE for all combinations of p_{00}, p_{11} and α.
- **Baseline** estimator performs the best under poor classifiers and small test sets.

$\alpha = 0.10$, $p_{11} = 0.98$, $p_{00} = 0.97$, $n = 200$, $N = 3 \times 10^6$.

Method	*Symbol*	Bias $\times 10^{-2}$	Variance $\times 10^{-5}$	RMSE $\times 10^{-2}$
Baseline	$\hat{\alpha}_a$	0.000	45.000	1.265
Classify-and-count	$\hat{\alpha}^*$	-2.500	0.000	2.500
Subtracted-bias	$\hat{\alpha}_b$	-0.125	13.984	1.189
Misclassification	$\hat{\alpha}_p$	0.000	15.652	1.251
Calibration	$\hat{\alpha}_c$	0.000	11.952	1.093

6. References

[1] J. P. Buonaccorsi, *Measurement Error: Models, Methods, and Applications*, Chapman & Hall/CRC, Boca Raton, FL, 2010.

[2] J. Kuha and C. J. Skinner, *Categorical data analysis and misclassification*, in Survey Measurement and Process Quality, L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, eds., Wiley, Mar. 1997, pp. 633–670.

[3] S. Scholtus and A. van Delden, *On the accuracy of estimators based on a binary classifier*, (2020). Discussion Paper, Statistics Netherlands, The Hague.

Dashboard

Interactive dashboard to explore the effect of the parameters can be found on this GitHub page: `https://github.com/kevinkloos/Misclassification-Bias`