The Journal of Open Source Software

# dsBinVal: Conducting distributed ROC analysis using DataSHIELD

**Daniel Schalk** [1,3,4], **Verena Sophia Hoffmann**[2,3], **Bernd Bischl**[1,4], **and Ulrich Mansmann**[2,3]

**1** Department of Statistics, LMU Munich, Munich, Germany **2** Institute for Medical Information Processing, Biometry and Epidemiology, LMU Munich, Munich, Germany **3** DIFUTURE (DataIntegration for Future Medicine, www.difuture.de), LMU Munich, Munich, Germany **4** Munich Center for Machine Learning, Munich, Germany

**DOI:** 10.21105/joss.04545

**Software**

- Review
- Repository
- Archive

**Editor:** Charlotte Soneson

**Reviewers:**

- @patRyserWelch8
- @brunomontezano
- @AnthonyOfSeattle

**Submitted:** 13 June 2022
**Published:** 21 February 2023



## Summary

Our `R` (R Core Team, 2021) package `dsBinVal` implements the methodology explained by Schalk et al. (2022). It extends the ROC-GLM (Pepe, 2000) to distributed data by using techniques of differential privacy (Dwork et al., 2006) and the idea of sharing highly aggregated values only. The package also exports functionality to calculate distributed calibration curves and assess the calibration. Using the package allows us to evaluate a prognostic model based on a binary outcome using the DataSHIELD (Gaye et al., 2014) framework. Therefore, the main functionality makes it able to 1) compute the receiver operating characteristic (ROC) curve using the ROC-GLM from which 2) the area under the curve (AUC) and confidence intervals (CI) are derived to conduct hypothesis testing according to DeLong et al. (1988). Furthermore, 3) the calibration can be assessed distributively via calibration curves and the Brier score. Visualizing the approximated ROC curve, the AUC with confidence intervals, and the calibration curves using `ggplot2` is also supported. Examples can be found in the README file of the repository.

## Statement of need

Privacy protection of patient data plays a major role for a variety of tasks in medical research. Uncontrolled release of health information may cause personal disadvantages for individuals, and the individual patient needs to be protected against personal details becoming visible to people not authorized to know them.

In statistics or machine learning, one of these tasks is to gain insights by building statistical or prognostic models. Prognoses on the development of severe health conditions and covariates encoding critical health information, such as genetic susceptibility, need to be handled with care. Furthermore, using confidential data comes with administrative burdens and mostly requires a consent around data usage. Additionally, the data can be distributed over multiple sites (e.g. hospitals) which makes their access even more challenging. Modern approaches in distributed analysis allow work on distributed confidential data by providing frameworks that allow retrieval of information without sharing of sensitive information. Since no sensitive information is shared through the use of privacy-preserving and distributed algorithms, their use helps to meet administrative, ethical, and legal requirements in medical research as users do not have access to personal data.

One of these frameworks for privacy protected analysis is DataSHIELD (Gaye et al., 2014). It allows the analysis of data in a non-disclosive setting. The framework already provides techniques for descriptive statistics, basic summary statistics, and basic statistical modeling.

Within a multiple sclerosis use case to enhance patient medication in the DIFUTURE consortium of the German Medical Informatics Initiative (Prasser et al., 2018), a prognostic model was developed on individual patient data. One goal of the multiple sclerosis use case is to validate that prognostic model using ROC and calibration analysis on patient data distributed across five hospitals using DataSHIELD.

In this package we close the gap between distributed model building and the validation of binary outcomes also on the distributed data. Therefore, our package seamlessly integrates into the DataSHIELD framework, which does not yet provide distributed ROC analysis and calibration assessment.

## Functionality

The integration of the `dsBinVal` package into the DataSHIELD framework extends its functionality and allows users to assess the discrimination and calibration of a binary classification model without harming the privacy of individuals. Based on privacy-preserving distributed algorithms (Schalk et al., 2022), the assessment of the discrimination is done by the `dsROCGLM()` function that calculates a ROC curve based on the ROC-GLM as well as an AUC with CI. The calibration is estimated distributively using the functions `dsBrierScore()` and `dsCalibrationCurve()`. Additional helper functions, `dsConfusion()` or `dsL2Sens()`, can be used to calculate several measures, e.g. sensitivity, specificity, accuracy, or the F1 score, from the confusion matrix or the L2-sensitivity. Note that measures from the confusion matrix may be disclosive for specific thresholds and are therefore checked and protected by DataSHIELDs privacy mechanisms. During the call to `dsROCGLM()`, parts of the data set are communicated twice, first, to calculate the ROC-GLM based on prediction scores, and second, to calculate the CI of the AUC. In both steps, the information is protected by differential privacy to prevent individuals from re-identification. The amount of noise generated for differential privacy is carefully chosen based on a simulation study that takes the variation of the predicted values into account. We refer to the README file of the repository for a demonstration and usage of the functionality.

**Technical details:** To ensure the functioning of our package on DataSHIELD, it is constantly unit tested on an active DataSHIELD test instance. The reference, username, and password are available at the OPAL documentation in the "Types" section. Parts of the tests also cover checks against privacy breaches by attempting to call functions with data sets that do not pass the safety mechanisms of DataSHIELD. Hence, individual functions attempt to prevent accidental disclosures when data is not sufficient to ensure privacy.

**State of the field:** To the best of our knowledge, there is no distributed ROC-GLM implementation available in `R`. Current state-of-the-art techniques require sharing of sensitive information from the sites and using existing implementation such as `pROC` (Robin et al., 2011) for the ROC curve or standard software for the GLM to calculate the ROC-GLM (as stated by Pepe (2000)).

## Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A and Federal Ministry for Research and Technology (BMFT) under Grant No. 01ZZ1804C (DIFUTURE, MII). The authors of this work take full responsibilities for its content.

## References

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach.

*Biometrics*, 837–845. https://doi.org/10.2307/2531595

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference*, 265–284. https://doi.org/10.1007/11681878_14

Gaye, A., Marcon, Y., Isaeva, J., LaFlamme, P., Turner, A., Jones, E. M., Minion, J., Boyd, A. W., Newby, C. J., Nuotio, M.-L., & others. (2014). DataSHIELD: Taking the analysis to the data, not the data to the analysis. *International Journal of Epidemiology*, *43*(6), 1929–1944. https://doi.org/10.1093/ije/dyu188

Pepe, M. S. (2000). An interpretation for the ROC curve and inference using GLM procedures. *Biometrics*, *56*(2), 352–359. https://doi.org/10.1111/j.0006-341x.2000.00352.x

Prasser, F., Kohlbacher, O., Mansmann, U., Bauer, B., & Kuhn, K. A. (2018). Data integration for future medicine (DIFUTURE). *Methods of Information in Medicine*, *57*(S01), e57–e65. https://doi.org/10.3414/ME17-02-0022

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77. https://doi.org/10.1186/1471-2105-12-77

Schalk, D., Hoffmann, V. S., Bischl, B., & Mansmann, U. (2022). *Distributed non-disclosive validation of predictive models by a modified ROC-GLM*. arXiv. https://doi.org/10.48550/ARXIV.2203.10828