



# dsROCGLM: Conducting distributed ROC analysis using DataSHIELD

Daniel Schalk<sup>1,3</sup>, Verena Sophia Hoffmann<sup>2,3</sup>, Bernd Bischl<sup>1</sup>, and Ulrich Mansmann<sup>2,3</sup>

<sup>1</sup> Department of Statistics, LMU Munich, Munich, Germany <sup>2</sup> Institute for Medical Information Processing, Biometry and Epidemiology, LMU Munich, Munich, Germany <sup>3</sup> DIFUTURE (DataIntegration for Future Medicine, [www.difuture.de](http://www.difuture.de)), LMU Munich, Munich, Germany

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: 

Submitted: 29 March 2022

Published: unpublished

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

## Summary

Our R ([R Core Team, 2021](#)) package dsROCGLM 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. 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 ROC curve using the ROC-GLM from which 2) the AUC is derived. Furthermore, 3) confidence intervals after DeLong et al. (1988) are estimated to conduct hypothesis testing of the estimated AUC. Visualizing the approximated ROC curve, the AUC, and the confidence intervals is also supported based on [ggplot2](#). 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 imply personal disadvantages for individuals. The individual patient needs to be protected that personal features become 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. Prognosis on the development of severe health conditions or covariates coding critical health information like genetic susceptibility need to be handled with care. Furthermore, using confidential data comes with administrative burdens and mostly requires a consent about using the data. 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 to work on distributed confidential data by providing frameworks that allow retrieval of information without sharing sensitive information. These techniques alleviate many of the administrative, ethical and legal requirements on medical research.

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, among others, techniques for descriptive statistics, basic summary statistics, or 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. This model is to be validated using ROC analysis on patient data distributed across five hospitals using DataSHIELD. Distributed ROC analysis is currently not available in DataSHIELD.

41 In this package we close the gap between distributed model building and conducting ROC  
42 analysis also on the distributed data. Therefore, our package seamlessly integrates into the  
43 DataSHIELD framework.

44 **Technical details:** To ensure the functioning of our package on DataSHIELD, it is constantly  
45 unit tested on an active DataSHIELD [test instance](#). The reference, username, and password  
46 are available at the [OPAL documentation](#) in the “Types” section.

47 **Related software:** We also implemented the Brier score and calibration curves to assess  
48 the model calibration within the DataSHIELD framework. These functions are available in  
49 the [dsCalibration](#) package. To upload models to the DataSHIELD servers and calculate  
50 predictions can be done using our [dsPredictBase](#) package.

## 51 Acknowledgements

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

## 56 References

- 57 DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under  
58 two or more correlated receiver operating characteristic curves: A nonparametric approach.  
59 *Biometrics*, 837–845. <https://doi.org/10.2307/2531595>
- 60 Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in  
61 private data analysis. *Theory of Cryptography Conference*, 265–284. [https://doi.org/10.1007/11681878\\_14](https://doi.org/10.1007/11681878_14)
- 62 Gaye, A., Marcon, Y., Isaeva, J., LaFlamme, P., Turner, A., Jones, E. M., Minion, J., Boyd,  
63 A. W., Newby, C. J., Nuotio, M.-L., & others. (2014). DataSHIELD: Taking the analysis  
64 to the data, not the data to the analysis. *International Journal of Epidemiology*, 43(6),  
65 1929–1944. <https://doi.org/10.1093/ije/dyu188>
- 66 Pepe, M. S. (2000). An interpretation for the ROC curve and inference using GLM procedures.  
67 *Biometrics*, 56(2), 352–359. <https://doi.org/10.1111/j.0006-341x.2000.00352.x>
- 68 Prasser, F., Kohlbacher, O., Mansmann, U., Bauer, B., & Kuhn, K. A. (2018). Data  
69 integration for future medicine (DIFUTURE). *Methods Inf Med*, 57(S01), e57–e65. <https://doi.org/10.3414/ME17-02-0022>
- 70 R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation  
71 for Statistical Computing. <https://www.R-project.org/>
- 72 Schalk, D., Hoffmann, V. S., Bischl, B., & Mansmann, U. (2022). *Distributed non-disclosive*  
73 *validation of predictive models by a modified ROC-GLM*. arXiv. <https://doi.org/10.48550/ARXIV.2203.10828>