STRengthening Analytical Thinking for Observational Studies (STRATOS):

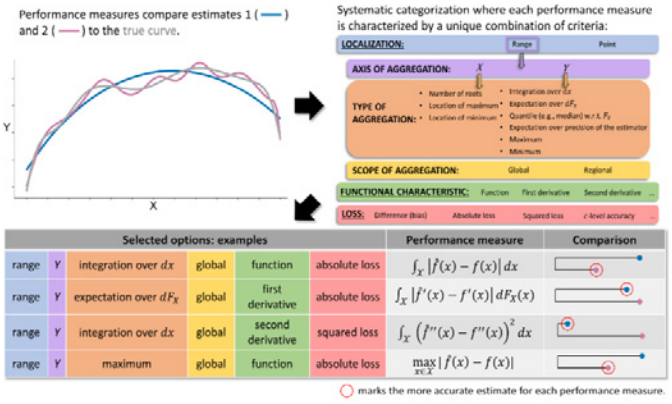
PERFORMANCE MEASURES OF ESTIMATORS FOR NON-LINEAR ASSOCIATIONS BETWEEN A CONTINUOUS VARIABLE AND AN OUTCOME: A CATEGORIZATION AND ILLUSTRATION OF VARIOUS CRITERIA

Theresa Ullmann 1 , Georg Heinze 1 , Michal Abrahamowicz 2 , Aris Perperoglou 3 , Willi Sauerbrei 4 , Matthias Schmid 5 , Daniela Dunkler 1 , for TG2 of the STRATOS initiative

1 Institute of Clinical Biometrics, Center for Medical Data Science, Medical University of Vienna, Vienna, Austria | 2 Department of Epidemiology & Biostatistics, McGill University, Montreal, Canada | 3 Statistics and Data Science Innovation Hub, GSK, Stevenage, UK | 4 Institute of Medical Biometry and Statistics, Medical Center, University of Freiburg, Freiburg, Germany | 5 Department of Medical Biometry, Informatics, and Epidemiology, University Hospital Bonn, Bonn, Germany.

In regression modeling, continuous predictors are often assumed to have a linear association with the outcome. In practice, however, associations may be more complex than that. Several methods were proposed for estimating non-linear associations, like (fractional) polynomials and various spline-based approaches. How should researchers choose among the many available methods? Simulation studies that systematically compare methods under different conditions can help answer this question. However, such studies remain relatively scarce, limiting the evidence available to guide methodological choices [1].

Suppose we decided to perform such a simulation study. The top left panel of Figure 1 shows some hypothetical results: two estimates for the association between a variable and an outcome, obtained using two different methods, to be compared with the true curve . Several performance measures have been proposed for such comparisons [2, 3], but until recently, no systematic overview of such measures existed. In line with the aim of STRATOS Topic Group 2 (TG2) to develop guidance for variable and functional form selection in multivariable analyses, this gap was recently closed by a TG2 article [4, https://doi.org/10.1002/wics.70042].



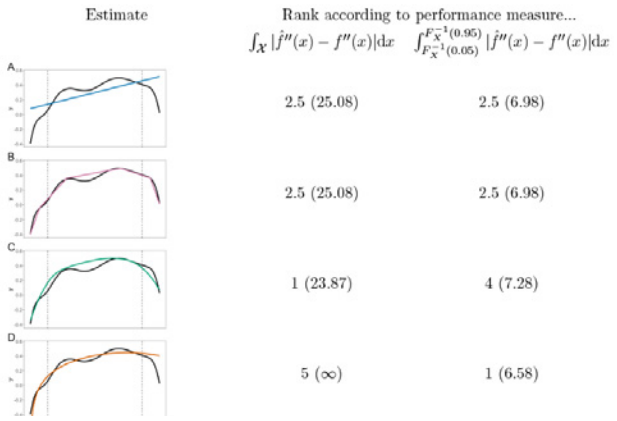
Title: Figure 1. Reprinted from [4] under a CC BY 4.0 license.

The article introduced a categorization of measures which is shown in the top right panel of Figure 1. First, performance measures differ in whether they compare the estimated and true curves over a range of predictor values or only at a single point. If the comparison is

made over a range, measures also differ in how the information is aggregated (for example, by integrating a loss function), and whether it is considered on the whole range of the variable or a subrange. Second, the curves themselves, their first derivatives or their second derivatives can be compared. Third, different types of losses can be chosen to measure the deviation of the estimated curve from the true curve.

Selecting a specific option for each of these aspects gives rise to a specific performance measure. Some examples of resulting measures are shown in the bottom panel of Figure 1. Notably, the ranking of the blue estimate (1) and the pink estimate (2) depends on the chosen performance measure. For example, the first measure prefers the estimate 2 because it is overall rather close to the true curve, while the third measure, which uses the second derivative to penalize "wiggliness," clearly prefers estimate 1.

Another example is shown in Figure 2 where five different estimates are compared. The first performance measure aggregates over the full range of , while the second measure considers only the range between the 5% and 95% percentile of the distribution of which here was Beta(2,2). Next to each estimate, its rank according to the respective performance measure is shown. The values of the measures are shown in brackets. According to the first measure, estimate C is the best one. Estimate D is ranked last because its performance value diverges when evaluated over the full range, as the second derivative diverges to -∞ at a fast rate for . On the other hand, the second measure ranks D on top and C in 4th place, because now the area around zero, where the second derivative of estimate D diverges to , is excluded. More examples can be found in a Shiny app at https://thullmann.shinyapps.io/presplinesshinyapp/.



Title: Figure 2. Reprinted from [4] under a CC BY 4.0 license.

To sum up, different performance measures capture different aspects of an estimated curve. Thus, well-balanced simulation studies that assess estimators should use more than just one measure. However, our categorization includes a very large number of measures, which raises the question which subset should be used in practice. The integrated squared difference can always be included. If excessive wiggliness should be penalized, measures which evaluate the second derivative should be included. Further measures may be considered depending on the context of an evaluation. For example, if the maximum of the curve is of special interest, then measures should evaluate the deviation of the estimated from the true maximum. If the distribution of X is unbounded, restricting the performance evaluation to a subrange is recommended. Our current work focuses on developing practical guidance on selecting suitable performance measures for different settings.

Our article focused on performance measures in the context of descriptive modeling, where the focal interest lies in evaluating the shape of an unadjusted or adjusted association. For explanatory modeling (causal inference), the performance measures considered in our article could be used for various purposes, e.g., to evaluate estimated dose-response curves or the association of a treatment effect with a continuous predictive biomarker. A STRATOS project from TG6 considered performance measures for prediction modeling [5].

References

[1] Sauerbrei, W., Perperoglou, A., Schmid, M., Abrahamowicz, M., Becher, H., Binder, H., Dunkler, D., Harrell Jr, F.E., Royston, P., Heinze, G. and TG2 of the STRATOS initiative, 2020. State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. Diagnostic and Prognostic Research 4, no. 1: 3. https://doi.org/10.1186/s41512-020-00074-3

[2] Govindarajulu, U. S., D. Spiegelman, S. W. Thurston, B. Ganguli, and E. A. Eisen. 2007. "Comparing Smoothing Techniques in Cox Models for Exposure–Response Relationships." Statistics in Medicine 26, no. 20: 3735–3752. https://doi.org/10.1002/sim.2848

[3] Binder, H., W. Sauerbrei, and P. Royston. 2013. "Comparison Between Splines and Fractional Polynomials for Multivariable Model Building With Continuous Covariates: A Simulation

Study With Continuous Response." Statistics in Medicine 32, no. 13: 2262–2277. https://doi.org/10.1002/sim.5639

[4] Ullmann, T., Heinze, G., Abrahamowicz, M., Perperoglou, A., Sauerbrei, W., Schmid, M., Dunkler, D., for TG2 of the STRATOS Initiative, 2025. A systematic categorization of performance measures for estimated non-linear associations between an outcome and continuous predictors. Wiley Interdisciplinary Reviews: Computational Statistics 17, no. 3: e70042. https://doi.org/10.1002/wics.70042

[5] Van Calster, B., Collins, G.S., Vickers, A.J., Wynants, L., Kerr, K.F., Barreñada, L., Varoquaux, G., Singh, K., Moons, K.G., Hernandez-Boussard, T., Timmerman, D., McLernon, D.J., Van Smeden, M., Steyerberg, E.W., Topic Group 6 of the STRATOS initiative, 2025. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. The Lancet Digital Health 7, no. 12: 100916. https://doi.org/10.1016/j.landig.2025.100916

DONATE TO THE IBS TRAVEL AWARDS FUND

The International Biometric Society Travel Awards Fund helps support the allocation of travel grants to a selected group of members from low- and middle-income countries (LMICs), with the goal of allowing them to travel to IBS events during their grant year. Donations are now being accepted online at our Donations page.

You may be interested to know that the Society has instituted a matching funds policy. For every dollar (USD) an individual, company, IBS Region or IBS Network donates, IBS will match it up to $20,000. This gives us the opportunity to fund more travel, bringing professionals from eligible countries together to collaborate, share knowledge, learn new perspectives and build lifelong friendships! Donations are vital to the continued success of this program, as the IBS is typically able to fund less than half of the applicants who apply. Donations also foster our membership growth in LMICs and serve our global audience. Please donate today. Note that each donation received from an individual or region will be acknowledged with a note of appreciation and a receipt. Please contact us directly at ibs@biometricsociety.org with any questions related to tax deductions or related matters.

Advertise with Us!

Do you have vacancies to fill? Are you running a promotion, courses, selling software, or consultancy services of interest to statisticians, biometricians, and mathematicians? Consider advertising on the IBS Website or in the Biometric Bulletin. Click below for additional information and ad rates. Please email the International Business Office (ibs@biometricsociety.org) with any questions. Note that members can also place employment ads online through our new "Employment Opportunities" online community. You must first login to submit your posting!

Advertising Specs & Rates - 2026