Randomized Model Selection as a Model Building Strategy in Stepwise Logistic Regression

Marcus Vollmer
Department of Mathematics and Computer Science
University of Greifswald, Germany

ERNST MORITZ ARNDT
UNIVERSITÄT GREIFSWALD

Wissen
lockt.
Seit 1456

Briefly

Selecting predictors?
patients with septic shock or severe sepsis

Logistic Regression » **Discrete Optimization**
improved by genetic algorithm
stable, better models

Introduction

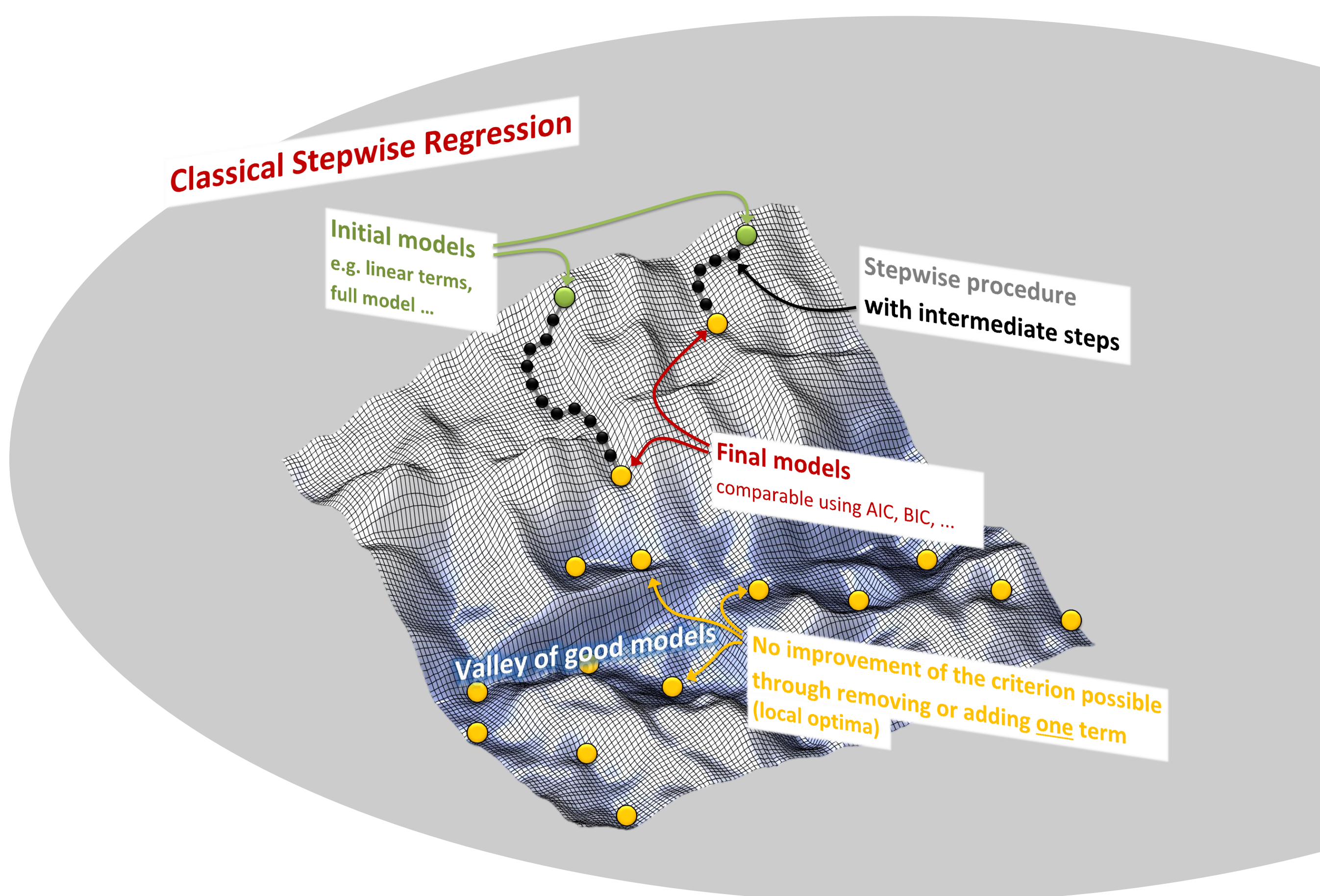
Logistic regression is an approved method in the applied research for predicting the outcome of a binary variable Y using predictor variables X_i (see Harrell[1] and Steyerberg[5]). The coefficients β of the logistic model, $P(Y{=}1|X_i{=}x_i) = 1/\left(1{+}\exp^{-(\beta_0+X_i^T\beta)}\right)$, are estimated by maximum-likelihood estimation. The discrete optimization problem of selecting relevant predictors (a subset of all possible regressors and interactions) is the subject of this research.

How to select a good set of relevant variables, including interactions?

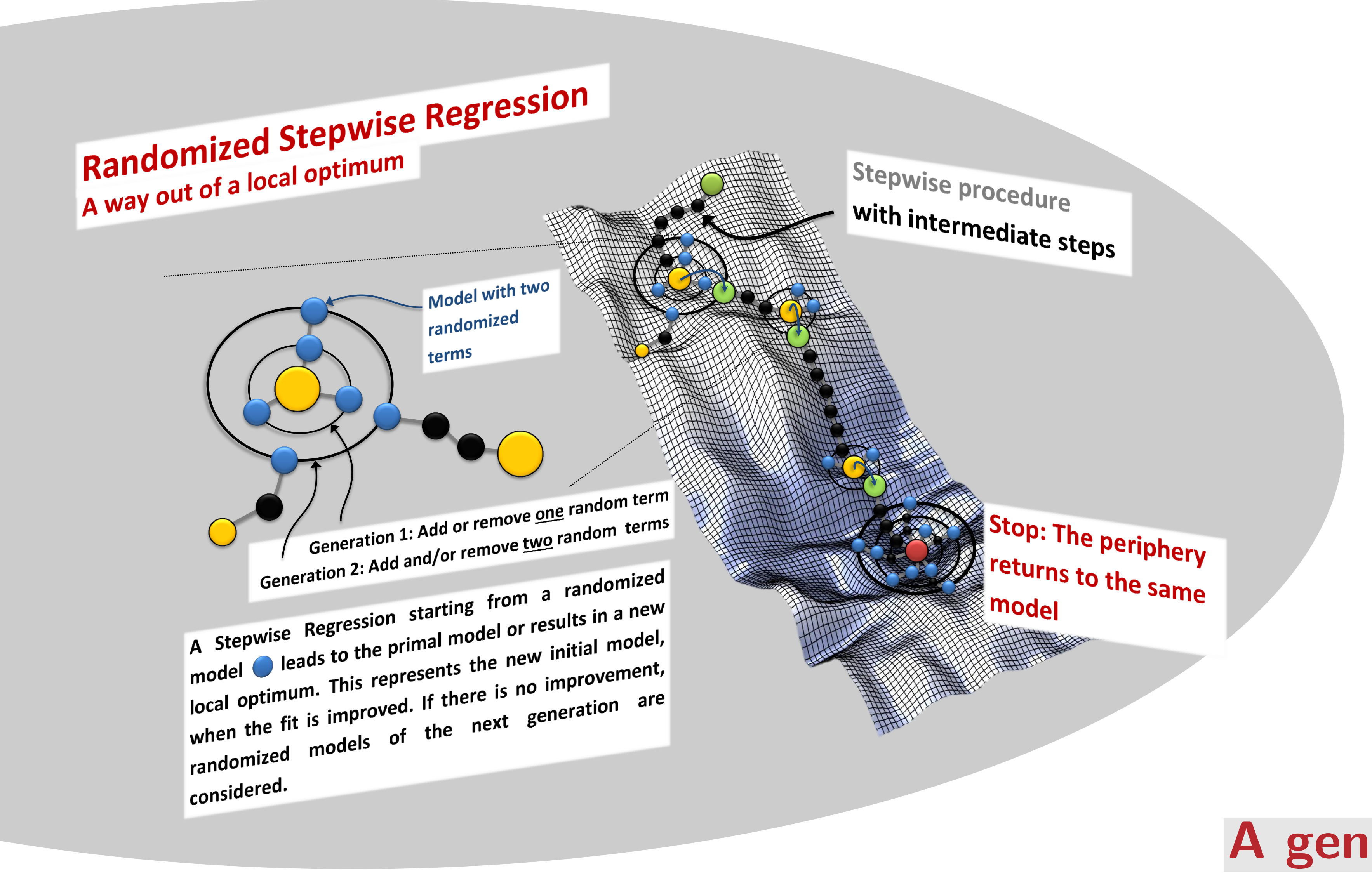
Often the multivariate modeling was performed using the statistically significant univariate variables (e.g. Nguyen[2]). But what about confounding and interaction terms? **Stepwise logistic regression with backward-elimination and forward-selection are commonly used** (in medical publications see Sebat[3] or Shapiro[4]).

Aim: To improve an initial model by adding or removing one predictor (1st Generation).
Comparison: The models of the 1st generation are compared using a criterion for model selection, such as the Bayesian information criterion, with k predictor variables, sample size n: $BIC = -2\log L(\hat{\theta}) + k\log n$
Steps: The initial model will be modified towards the 1st-generation-model with the lowest BIC value iteratively.
Termination: The "mutation" ends, if no improvement is possible.

The classical stepwise regression ends only in a local optimum and depends strongly on the initial model (as illustrated right). It has various disadvantages[5], like instability of the selection and biased estimation of coefficients.


Classical Stepwise Regression
Initial models
e.g. linear terms, full model ...
Stepwise procedure
with intermediate steps
Final models
comparable using AIC, BIC, ...

No improvement of the criterion possible through removing or adding one term (local optima)


Randomized Stepwise Regression
A way out of a local optimum
Model with two randomized terms
Generation 1: Add or remove one random term
Generation 2: Add and/or remove two random terms
A Stepwise Regression starting from a randomized model ● leads to the primal model or results in a new local optimum. This represents the new initial model, when the fit is improved. If there is no improvement, randomized models of the next generation are considered.
Stepwise procedure
with intermediate steps
Stop: The periphery returns to the same model

Randomized Model Selection - heuristic technique & modified stepwise procedure

Models of the j-th generation are neighbor models of the initial model, by adding and removing exactly j regressors. Let l be the maximum number of considered generations. The fixed numbers $g_1, \ldots, g_l$ denotes, how many neighbor models are considered at each generation.

- Initial model randomly chosen
- Classical stepwise procedure tries to optimize the logistic model until a local optimum is reached
- Only g_1 modified models of the first generation are considered
- From each of these g_1 models the classical stepwise procedure tries to optimize
- Either the procedure leads to the previous optimum or it ends in some new local optimum
- The best model, concerning the BIC value, will be selected
- If none of the g_1 final models provides an improvement, g_2 randomly chosen models of the second generation will be considered
- The procedure ends, unless the l-th generation improves the logistic model

The number of generations l, its sizes $g_1, \ldots, g_l$ and the number of involved variables determine the computing time.
In contrast to the classical deterministic procedure, not every possible neighbor model will be considered. Therefore steps to further generations are allowed and provides a way out of a local optimum.

A genetic algorithm was used for optimization.

Application to sepsis data and validation of the new procedure

In a interdisciplinary collaboration with the team of Sepsisdialog (www.sepsisdialog.de) of the Universitätsmedizin Greifswald, managed by Dr. Matthias Gründling, we were able to rate the performance of the new strategy and investigate the intensive care unit survival of 746 patients with septic shock or severe sepsis between 2006 and midyear 2012. In 2008, a continuous training program of diagnostic measures and treatment was implemented.

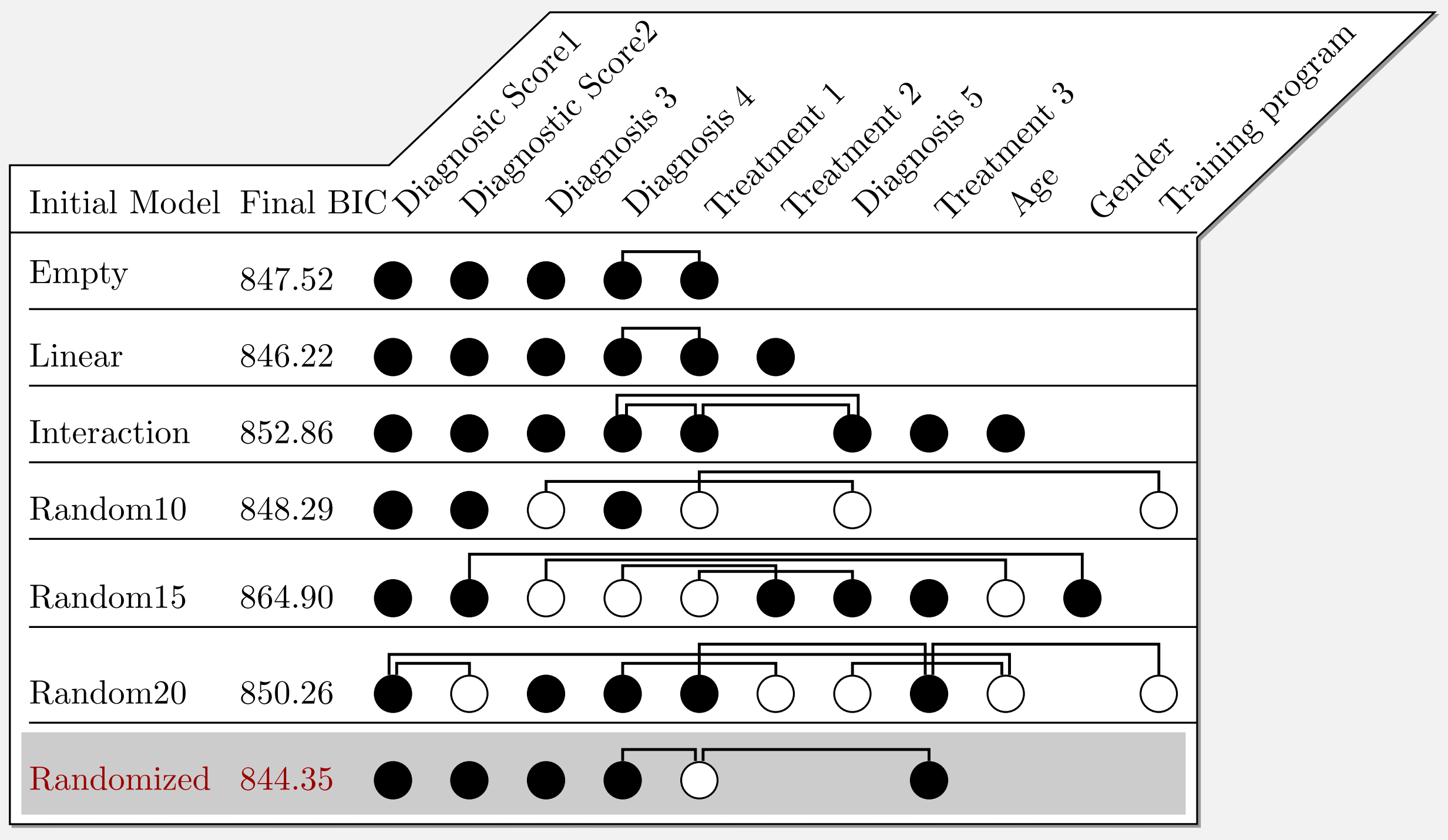
Response Variable: 90-day-survival
Possible Predictors:

- Demographic characteristics *(age, gender)*
- Septic diagnosis *(6 possible predictors)*
- Therapy *(6 possible predictors)*
- Training program group

Model: Logistic regression with bivariate interactions

Classical stepwise logistic regression with different initial models were compared with the randomized approach. The randomized model selection starts with a randomly chosen model with 10 terms. The method proceeds till the fourth generation ($[g_1, \ldots, g_4] = [5, 10, 15, 20]$).

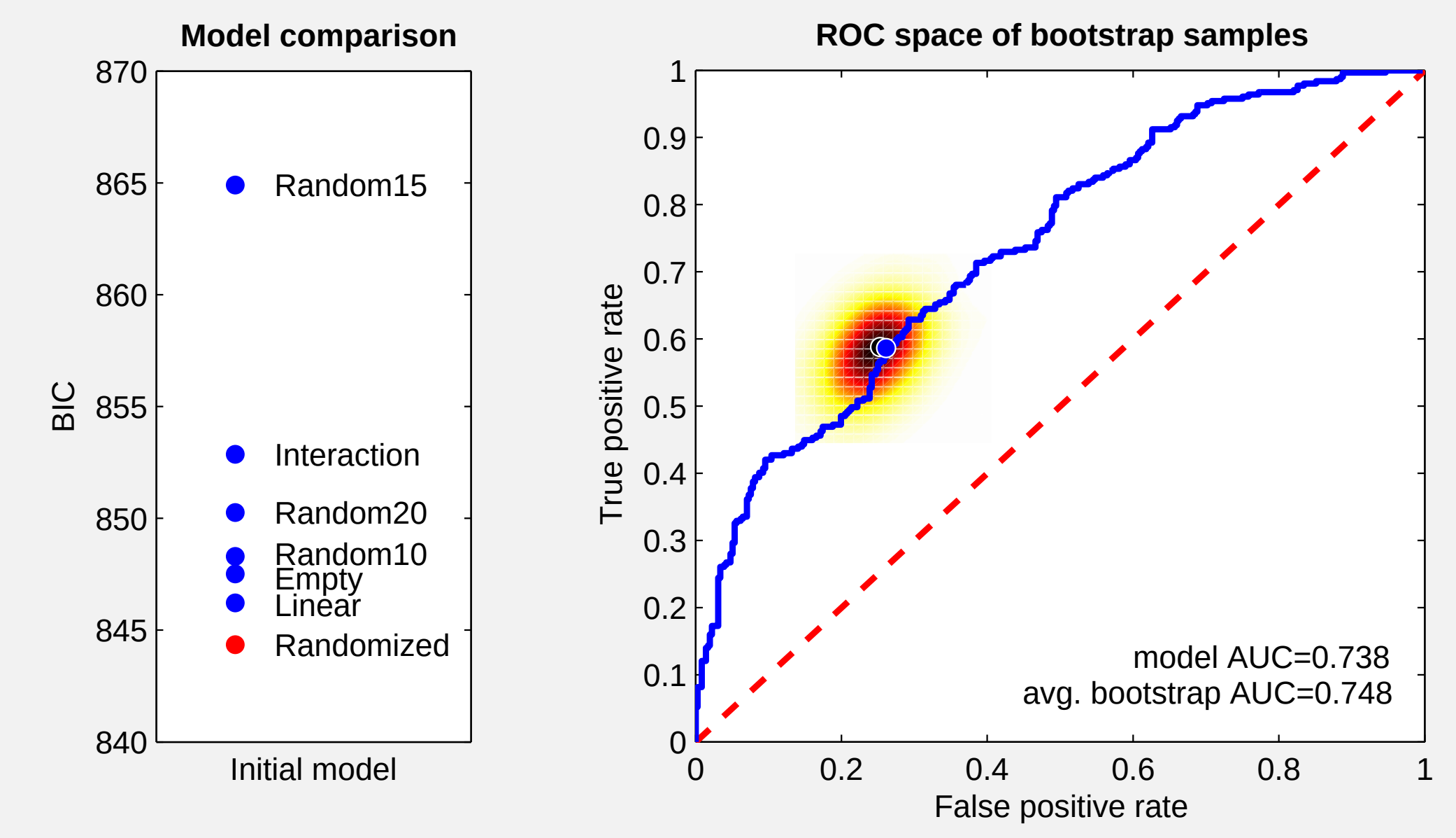
Initial models for the classical stepwise model selection were the empty model, the model with all linear terms, half of all possible interaction terms (Interaction) and three random models including 10, 15 and 20 terms (linear and interaction terms mixed).



Initial Model	Final BIC	Diagnosic Score1	Diagnostic Score2	Diagnosis 3	Diagnosis 4	Treatment 1	Treatment 2	Diagnosis 5	Treatment 3	Age	Gender	Training program
Empty	847.52	●	●	●	●	●						
Linear	846.22	●	●	●	●	●	●					
Interaction	852.86	●	●	●	●	●		●	●	●		
Random10	848.29	●	●	○	●	○		○				○
Random15	864.90	●	●	○	○	○	●	●	●	○	●	
Random20	850.26	●	○	●	●	●	○	○	●	○		○
Randomized	844.35	●	●	●	●	○			●			

Figure 1 : Comparison of the local optima. Classical stepwise regression with six different initial models and randomized model selection. Filled circles: linear terms. Lines between circles: interactions. Unfilled circles: appearance only in interactions.

The best model (lowest BIC) might be the one from the randomized model selection with five linear and two interaction terms. **Bootstrapping (see Steyerberg[5], Chapter 5.3) was used to evaluate the stability of the regression coefficients** and of the AUC value (see figure 2b).



(a) BIC values of the different models.

(b) ROC curve and bootstrapping inference of the best model.

Figure 2 : Validation of the final model.

Conclusions

✓ **Randomized algorithm was more successful**
✓ **Construction ensures to produce better or equivalent models**

✓ **Model selection seems to be more stable**
✓ **Random initial models can be used for enhancement**

✓ **Success of the training program of Sepsisdialog was verified**

[1] Harrell, F. (2001). *Regression Modeling Strategies*. Springer, corrected edition.
[2] Nguyen, H. B. (2004). Early lactate clearance is associated with improved outcome in severe sepsis and septic shock. *Critical Care Medicine*, 32:1637–1642.
[3] Sebat, F. (2007). Effect of a rapid response system for patients in shock on time to treatment and mortality during 5 years. *Critical Care Medicine*, 35:2568–2575.
[4] Shapiro, N. I. (2006). Implementation and outcomes of the Multiple Urgent Sepsis Therapies (MUST) protocol. *Critical Care Medicine*, 34:1025–1032.
[5] Steyerberg, E. W. (2008). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (Statistics for Biology and Health)*. Springer, 1st edition.

Conference on Structural Inference in Statistics
Potsdam, 17.-19. Sept. 2013

Marcus Vollmer
marcus.vollmer@uni-greifswald.de