The potential of 'conceptually oriented' pre-processing of covariates to improve prognostic models. A case study in low back pain patients

Anne Molgaard Nielsen[1], Adrian Binding[2], Casey Ahlbrandt-Rains[3], Martin Boeker[3], Stefan Feuerriegel[2] ,Werner Vach[4,5]

[1] *Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark*
[2] *Department of Management, Technology, and Economics, ETH Zurich, Weinbergstr. 56/58, 8092 Zurich, Switzerland*
[3] *Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Stefan-Meier-Str. 26, D-79104 Freiburg i. Br., Germany*
[4] *Department of Orthopaedics and Traumatology, University Hospital Basel, Spitalstr. 21, CH-4031 Basel, Switzerland*
[5] *Nordic Institute of Chiropractic and Clinical Biomechanics, Campusvej 55, DK-5230 Odense M, Denmark*

Background: A conceptually oriented pre-processing of a large number of potential prognostic factors may improve the development of a prognostic model and hence may play an important role in this process. However, it is unclear, whether this assumption holds and which way of pre-processing is optimal.

Aim: This study investigated whether various forms of conceptually oriented pre-processing or the preselection of established factors was superior to using all factors as input.

Methods: We made use of an existing project which developed two conceptually oriented subgroupings of low back-pain patients without taking the outcome into account. Based on the prediction of six outcome variables by seven statistical methods, this type of pre-processing was compared with domain specific principal component scores, medical experts' preselection of established factors as well as with using all 112 available baseline factors.

Results: Subgrouping of patients was associated with low prognostic capacity. Applying a Lasso-based variable selection to all factors or to domain-specific principal component scores performed best. The preselection of established factors showed a good compromise between model complexity and prognostic capacity.

Conclusion: The prognostic capacity is hard to improve by means of a conceptually oriented pre-processing when compared to purely statistical approaches. However, a careful selection of already established factors combined in a simple linear model should be considered as one option when constructing a new prognostic rule based on a large number of potential prognostic factors.

Keywords
Model construction, Prognostic models, Domain knowledge