Survey Response Data Challange

Corresponding person:	Werner Vach <wv@imbi.uni-freiburg.de>
Source:	Alice Kongsted (a.kongsted@nikkb.dk) and Lise Hestbaek (l.hestbaek@nikkb.dk), Nordic Institute of Chiropractic and Clinical Biomechanics, Odense, Denmark
License:	ODbL
Dataset title:	Baseline assessment and outcome measures of low back pain patients
Dataset name:	data_challenge2
Abstract:	928 low back pain patients filled out self-reported questionnaires plus clinicians recorded additional information from their history and carried out standardised clinical examinations, which resulted in 112 variables at baseline; in addition, the patients filled out outcome-related questionnaires at 2 weeks, 3 months and 12 months, which resulted in three outcome variables at each of these three time points. (The full submitted dataset includes more variables but for the challenge these were removed.)
Subject matter background:	The data stem from a longitudinal observational study of adult patients who were consulting chiropractors in Denmark due to their low back pain (LBP). Participants completed a baseline questionnaire while attending the clinic. The variables covered pain history and work-related questions, and in addition validated questionnaires covering activity limitation, fear avoidance, depression etc. were used. Follow-up questionnaires were sent by mail 2 weeks, 3 months, and 12 months after the baseline consultation. They included measures of global perceived improvement and the RMDQ (Roland Morris Disability Questionnaire). In general, low back pain is a poorly understood condition with considerable negative consequences for global health. A small proportion of patients severely affected by LBP account for most health- and disability-related costs, whereas other people with LBP report no care seeking, activity limitation, or sick leave. Consequently, LBP is a highly diverse condition and there is a need for a better understanding of the mechanisms underlying this heterogeneity. Our understanding of the heterogeneity in LBP and its consequences would be assisted by detailed knowledge about the prognosis of LBP and about the factors associated with the transition from trivial to burdensome LBP. We are interested in a clinically oriented grouping of the patients based on their baseline characteristics in order to obtain groups with similar characteristics, for whom it is reasonable to assume that they have a similar prognosis and may act similarly to interventions. It would be very useful if these patient groups can be characterized conceptually or with reference to a few key variables, such that this grouping can be used later in clinical practice without collecting data on all baseline variables. These baseline variables are the 112 variables from bsex0

	(11th variable) to Start_risk (122nd variable).
Data structure:	object x variables data matrix
Data objects and variables:	The objects are patients who have approached a chiropractor with low back pain. All baseline variables can be used for the first research question.
Data values:	An overview of all variables with explanations in English (including information on admissible values and of the meaning of the values) can be found in variables_data_challenge2.xlsx. All variables have been numerically coded, and missings are presented by NA. Variables 2-10 represent the longitudinal outcomes; they should not be used for the actual clustering but could be used to inform or validate the clustering. After that, you find the baseline variables to be used for clustering. Most variables are based on existing and published instruments. Detailed information on many variables can be found in the references, in particular in Nielsen et al. (2016) about the baseline variables.
Preprocessing:	Preprocessing has been applied to several variables. Information on this can be found in the column 'Response option rescaling (if performed)' in the file data_challenge2.xlsx. For a few variables this column also mentions a possibility of some additional preprocessing.
Researchers using this data set in publications should cite the following paper(s):	When analysing the baseline variables, the following paper should be cited: Nielsen AM, Vach W, Kent P, Hestbaek L, Kongsted A (2016): Using existing questionnaires in latent class analysis: should we use summary scores or single items as input? A methodological study using a cohort of patients with low back pain. Clinical Epidemiology 8, 73-89. DOI: 10.2147/CLEP.S103330
Other relevant papers:	Eirikstoft H, Kongsted A. Patient characteristics in low back pain subgroups based on an existing classification system: a descriptive cohort study in chiropractic practice. Man Ther. 2014;19(1): 65–71. Kongsted A, Vach W, Axo M, Bech RN, Hestbaek L (2014): Expectation of Recovery From Low Back Pain. Spine 39 (1), 91-90. DOI: 10.1097/BRS.0000000000000059 Nielsen AM, Kent P, Hestbaek L, Vach W, Kongsted A (2017). Identifying subgroups of patients using latent class analysis: should we use a single-stage or a two-stage approach? A methodological study using a cohort of patients with low back pain. BMC Musculoskeletal Disorders. The baseline variables have been analysed using a conceptually oriented latent class analysis (Nielsen et al. 2016).
Justification for clustering:	As pointed out in the description of the research questions, the aim is to find a (semi-)automatic classification of the patients based on the baseline variables in order to find clinically applicable and useful groups.
Is an external variable (not part of the	No

data used for clustering) to be used to evaluate the clustering result?	
Analysis has pragmatic aim:	Variables 2-10 could be used to check the prognostic value of a classification based on the baseline variables.
Number of clusters:	To ensure clinical acceptance, it is desirable to have between 3 and about 12 clusters/groups.
Nature of clusters:	It is natural in this setting that some patients are on the borderline between different groups.
All objects clustered or not:	A small group of patients classified as 'unclassifiable' may be acceptable.
Cluster overlap:	If the groups reflect different conceptual characteristics, it may be also natural to allow overlap between groups.
Cluster sizes:	Clusters can vary in size. A large number of small clusters (<2%) would limit the clinical acceptability.
Are there requirements on what should be the unifying/common ground for objects to belong to the same cluster?	There are no further requirements than a sufficient degree of similarity, which allows a conceptual labeling.
Are there requirements on the form of within-cluster heterogeneity?	No
Are there requirements on what should be the discriminating ground for objects to belong to different clusters?	No
Are there requirements on the between-cluster heterogeneity, that is, the structure of between-cluster differences?	No
Should clusters be similar with regard to within-cluster features such as variance or within-cluster structure?	No
Should the clustering be stable (with regard to the influence of outliers, the subset of variables under study, the choice of a dissimilarity measure, other)?	No
Is quality of inferences about population characteristics an issue?	No