

# Research question

1. To understand if clusters of homogenous individiauls with low back pain LBP can be derived from baseline demographic, pain, psychological, and physical characteristics (work package [WP] 1).
2. To verify if the mechanism of effect of the GLAD program follows the hypothesized pathway of change (see above) (WP2).
3. In the circumstance where the hypothesized variables did not mediate the outcome, the secondary aim is to propose alternative pathway of change effects (WP3).

# Methods

## Work package 1

### Exploratory

To determine the optimal number of clusters of homogenous individiauls with low back pain LBP derived from baseline demographic, pain, psychological, and physical characteristics. The primary purpose of this analysis is to inform subsequent work packages of the potential for sub-group analysis. For example, if there exist different clinical sub-groups (e.g. high pain-low fear vs low pain-high fear), than it is predicted that different causal mechanisms may mediate the effect of the GLAD program on disability. In this instance, the effect of *clusters* could either be added into the analysis as a moderating variable (e.g. in work package two), or separate analyses could be performed on each cluster, sample size permitting (e.g. one cluster one network in work package 3).

### Variables

1. Fear
2. Illness perception
3. Self-efficacy
4. Physical performance
5. Disability
6. Pain characteristics
7. Demographic characteristics
8. Clinical location

All statistical analyses will be performed in R software. Various unsupervised machine learning methods will be used, such as 1) latent class analysis (1), 2) hierarchical clustering, and 3) principal components plus hierarchical clustering. Cluster method number 3 is especially useful when the number of variables is large, and dimension reduction is first used to create groups of variables that will be subsequently used to create groups of individuals. The optimal number of clusters will be determined using a statistical threshold combined with expert clinical opinion (1). The statistical determination of the optimal number of clusters will be done using the elbow method – which looks at the total within-cluster sum of squares (WSS) as a function of the number of clusters. One should choose a number of clusters so that adding another cluster doesn’t improve much better the total WSS. In addition, an expert judgement of the meaningfulness of the resultant clusters will be made based on several factors: 1) if the clusters are “clinically” relevant, 2) if the clusters are “clinically” distinct, and 3) if the clusters have a sufficiently large number of individuals.

## Work package 2

### Hypotheses

The variables - fear, illness perception, self-efficacy and physical capacity will mediate the relationship between the independent variables (clinician pain beliefs or beliefs about capabilities) and the dependent variables of (disability or pain).

### Independent variables ()

1. (baseline to 4m)
2. (baseline to 4m)

### Mediator variables ():

1. (baseline to 6m)
2. (baseline to 6m)
3. (baseline to 6m)
4. performance (baseline to 6m)

### Outcome variables ()

1. (baseline to 6m)
2. (baseline to 6m)

### Causal mediation analyses

Causal mediation analyses will be performed using the “mediation” package (2). We will construct two models for each independent, mediator, and outcome variable combinations. The mediator model will be fitted as follows:

and the outcome model will be fitted as follows:

For each mediation analysis, we will input the mediator and outcome models into the *mediate* function of the statistical package to compute the average total effect, average causal mediation effect, and average direct effect. The average causal mediation effect is the effect of anticipatory cues on cutting performance, given the influence, anticipatory cues have on stiffness which, in turn, affects cutting performance. The average direct effect is the effect of anticipatory cues on performance when stiffness remained invariant between the two conditions. The average total effect is the sum of the average causal mediation and direct effects. Significance will be determined using an alpha of 0.05.

### Structural equation modelling (SEM)

The limitation of the causal mediation analysis framework above is that it cannot simultaneously test the mediating effect of multiple variables of multiple outcomes. SEM will be used to test a hypothesised causal chain whereby participating in the GLAD program leads to changes in potential mediating variables causing change scores in the study outcomes.

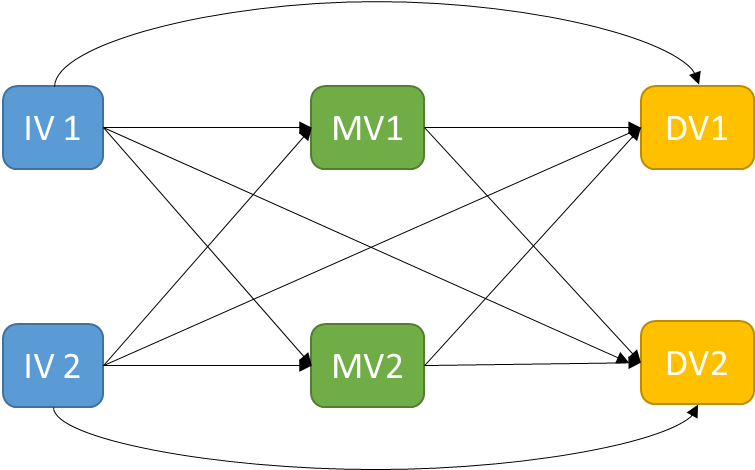


Figure 1. Basic structural model with only two independent, mediating, and dependent variables. The model will be expanded to use the number of mediating and dependent variables present in the study.

SEM will be performed using the *lavaan* package in R. Model fit will be assessed using 4 goodness of fit indices: p-value of chi-square, root mean square error of approximation (RMSEA) with 90% confidence interval, Adjusted Goodness of Fit Index (AGFI), and Bentler Comparative Fit Index (BCFI). An acceptable model fit is expected to reach p>0.05, RMSEA<0.05, AGFI≥0.90 and BCFI≥0.95.

## Work package 3

### Exploratory

If the hypothesized structural model in WP 2 does not provide an adequate model fit, a data-driven statistical approach to learn the multi-variate structural relationships will be undertaken.

### Variables included in the Bayesian Network

1. (baseline to 4m)
2. (baseline to 4m)
3. (baseline to 6m)
4. (baseline to 6m)
5. (baseline to 6m)
6. (baseline to 6m)
7. (baseline to 6m)
8. (baseline to 6m)
9. (6m to 12m)
10. (6m to 12m)
11. (6m to 12m)
12. (6m to 12m)
13. (6m to 12m)
14. (6m to 12m)

### Bayesian network analysis.

All analyses will be performed in R software(3) using the “bnlearn” package (4). BN is a graphical modelling technique (5) used increasingly in the health sciences to understand causal relationships. BN can handle some missing data (6), which makes them practical in longitudinal studies where data sets are often incomplete. BN quantifies the relationships among a set of variables X = {*X1*, …, *XN*}, where *N* is the number of different variables, using a directed acyclic graph (DAG). Each variable is associated with a node and directed arcs represent conditional dependencies between pairs of nodes. Building a BN model using a data-driven approach involves two stages: 1) structural learning - identifying which arcs are present in the DAG; and 2) parameter learning - estimating the parameters that regulate the strength and the sign of the corresponding relationships.

BN can easily include prior knowledge, sourced from the literature and experts, during the model building process. In the BN framework, prior knowledge can be included in the model as blacklist and whitelist arcs. Blacklist arcs are those which contravene known biological/physical mechanisms. We will only blacklist arcs that point backwards in time in the present study.

We will make use of model averaging to reduce the potential of including spurious relationships in the BN, using bootstrap resampling (*B = 200*) and performing structure learning on each of the resulting sample using Structural Expectation-Maximization (EM) (6). Structural EM is a technique which can build BN models in the presence of missing data (6). It does so by building an initial empty BN model using the original complete data, using it to impute missing data, rebuilding the BN model using the imputed complete data, and repeating this sequence until convergence. We will compute an “average” consensus DAG by selecting those arcs that have a frequency of > 50% in the bootstrapped samples, to create a sparse and interpretable network (3).

### Conditional probability queries.

The derived averaged BN model can be considered an “expert system”, which means that we can elicit a sample of realisations of the modelled variables under specific conditions. For each conditional probability query, we will 104 realisations of the variables of interest to obtain precise probability estimates. We will use a technique known as belief updating, which estimates the posterior probability of an event happening based on the available evidence on the values of certain variables. We will also adopt a specific method of belief updating known as logic sampling (5).

# References

1. Kent P, Jensen RK, Kongsted A. A comparison of three clustering methods for finding subgroups in MRI, SMS or clinical data: SPSS TwoStep Cluster analysis, Latent Gold and SNOB. BMC Med Res Methodol. 2014;14:113.

2. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation: R Package for Causal Mediation Analysis. J Stat Softw. 2014;59(5):38.

3. Scutari M, Nagarajan R. Identifying significant edges in graphical models of molecular networks. Artif Intell Med. 2013;57(3):207-17.

4. Scutari M. Learning Bayesian Networks with the bnlearn R Package. J Stat Softw. 2010;35(3):22.

5. Nagarajan R, Scutari M, Lèbre S. Bayesian networks in R with applications in systems biology. Verlag, New York: Springer; 2013.

6. Friedman N. Learning Belief Networks in the Presence of Missing Values and Hidden Variables. Proceedings of the Fourteenth International Conference on Machine Learning. 657145: Morgan Kaufmann Publishers Inc.; 1997. p. 125-33.