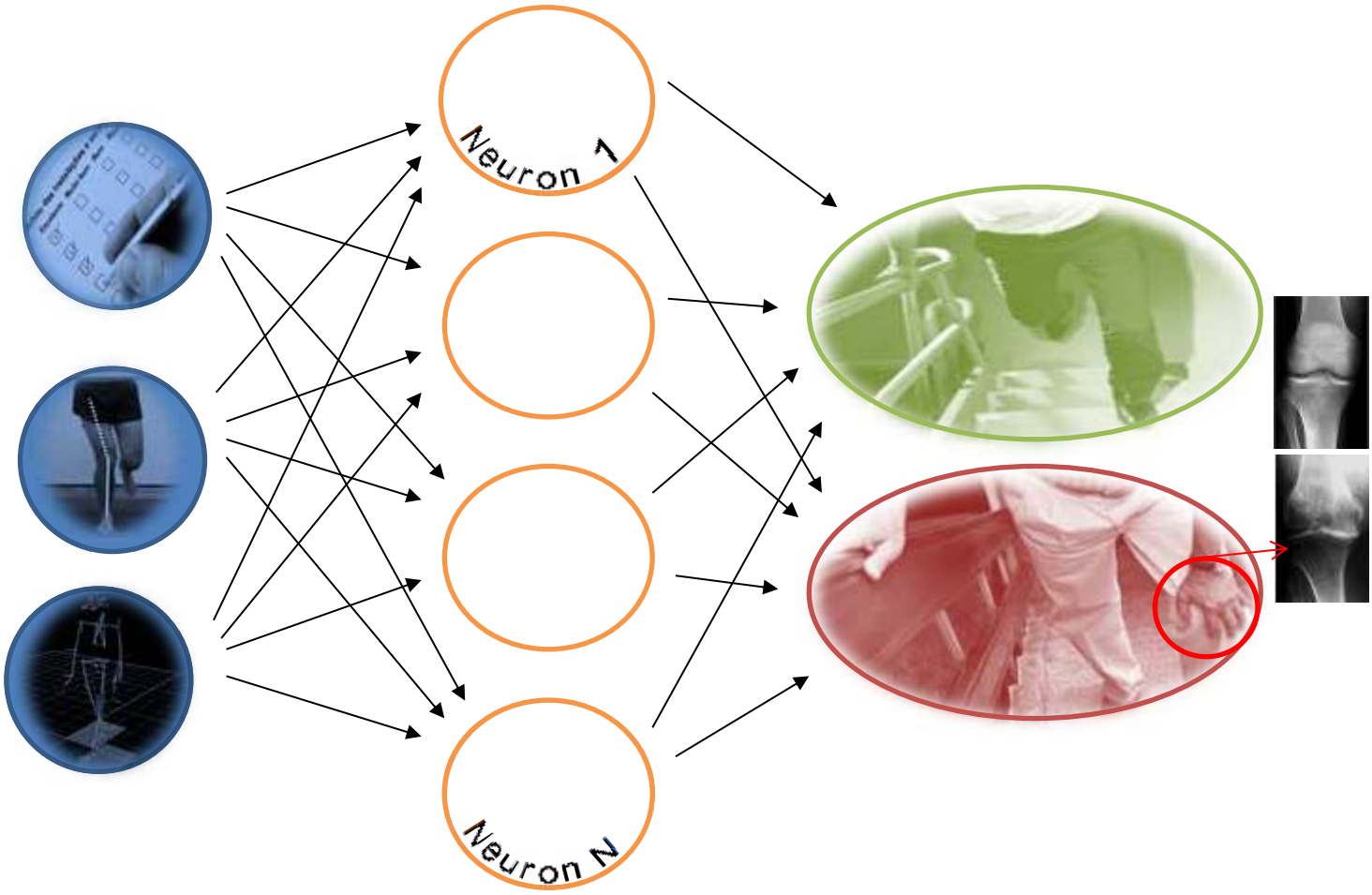
**Human Movement Sciences**
**Sport, Exercise & Health (research)**

Research Internship Research Master 2015-2016
(course code B_RIRM)

# Classifying physical function in patients with knee osteoarthritis and healthy controls based on kinematics of gait and single-limb mini squat using a supervised machine learning approach




VU University Amsterdam
Faculty of Human Movement Sciences
Qualification: MSc in Human Movement Sciences
Research internship Research Master

Author: N.M. van Mastrigt

Supervisor(s): prof. dr. ir. J. Harlaar
prof. dr. A. Daffertshofer

*October 2016*

**Summary**

*Background:*
Knee osteoarthritis (KOA) significantly affects physical function. Conventional physical function measures do not include movement quality information and do not correlate well. We proposed a machine learning approach including kinematics of gait and single-limb mini squat (SLMS) for classifying physical function.

*Aims:*
The aims of this study were: to highlight advantages and limitations of 1) conventional univariate and multivariate statistical methods, and 2) a new machine learning approach using a multidimensional data set of patients with KOA and healthy controls, and 3) to evaluate classification performance of binary classifications discriminating patients and controls based on input feature sets containing conventional physical function measures with or without kinematic parameters and time series.

*Methods*:
In 40 patients and 25 controls kinematics were recorded while walking or performing the SLMS test. Conventional physical function measures, spatiotemporal characteristics and discrete kinematic parameters were compared between patients and controls using multiple independent t-tests. Kinematics time series were analysed using principal component analysis (PCA). An artificial neural network supervised machine learning approach was used for classifying subjects as patient or healthy controls based on all previous input features. Different feature sets were statistically compared in terms of classification accuracy, sensitivity and specificity.

*Results*:
Univariate statistics did not take into account covariance between variables so redundant information was measured. This decreased the chance of finding significant differences due to Bonferroni corrections. PCA builds on covariance and allowed large data reductions but interpretation appeared to be less intuitive. Machine learning was able to handle multimodal data sets and resulted in high classification accuracies, sensitivities and specificities.

*Conclusions*:
Machine learning is a useful approach in handling multimodal data sets but it should be used in combination with educated guesses. In future research, this approach could be useful for classifying physical function in patients with KOA in multiple classes by unsupervised machine learning.

**Key words**

*Knee osteoarthritis; Physical function; Supervised machine learning*

**Table of Contents**

# 1 Introduction

*Problem statement*
Three-dimensional (3D) movement analysis has become an established method in studying biomechanical responses to pathology in several functional tasks. Mostly, gait is analysed. It provides detailed information about joint angles, forces and moments over time during a movement. Quantifying movement offers the opportunity to compare groups with and without a certain pathology or before and after a therapy which may assist in gaining insight in the cause, progression and treatment of the pathology. In the clinic, the use of 3D movement analysis could assist in diagnosing, monitoring disease progression, evaluating treatment effects and treatment planning.
Despite its principal benefits, 3D movement analysis is not widely used in the clinic yet[1], besides gait analysis in children with cerebral palsy[2]. This may have several causes. In the first place, a large amount of data is produced[3,4]: marker trajectories, and angles, forces, moments and powers of multiple joints are calculated over time. In the second place, the data set is high-dimensional[4]: time series of multiple dimensions (for example, angles and EMG) as well as discrete anthropometric variables are included. In the third place, different movement time series are highly interdependent[4,5], in general in a non-linear manner[4]. In addition, the total variability is high due to inter- and intra-subject and inter- and intra-trial variability and variability in marker placement, which weakens statistical power[4]. These problems are significant barriers with respect to interpretability, which makes clinical decision making assisted by the results of a 3D movement analysis difficult[6].

*Standard statistical approaches*
The question rises how to reduce the data in a meaningful manner that allows statistical comparison between groups or conditions[1,4], for example between groups with and without a certain pathology or between pre- and post-treatment conditions. Two common approaches are to select discrete parameters from kinematic time series and to do multiple t-tests, or to reduce the amount of time series by means of a principal component analysis (PCA).
The first approach is a univariate statistical approach that is often used in studying gait. Common practice in the field of analyzing gait data is the definition and selection of parameters such as peak values and ranges at certain percentages of the gait cycle from angle, force, moment or power waveforms[4,7–10]. Patients' parameters can be compared to average "normal" parameters or literature values to draw conclusions that are clinically relevant, such as detection of abnormality[11].
Although this data reduction is often based on an educated guess, the choice of which parameters to select is very subjective[1,12,13]. Moreover, there can be significant inter-subject variation in the observed gait waveforms[14] and selection of predefined parameters from an atypical waveform may require user intervention[1,12]. It may be impossible to define certain parameters, for example in pathological data[4,15,16]. And, temporal information as represented by the complete waveform is lost[1,4] by this way of reducing data.
Whereas univariate statistical methods cannot account for covariance between variables, multivariate statistical methods can. Principal component analysis is a multivariate statistical method that is also suitable for analyzing time series. It re-expresses the data by projecting them onto less, linearly uncorrelated principal component (PCs) axes that lie in the orthogonal directions of maximal variance (of the projected subspace). It can be used to reduce the amount of time series needed to describe a movement[17]. This approach may be considered more objective compared to simple univariate statistics, since the data reduction is based on features that are extracted by the analysis technique instead of the user, and because data from the entire gait cycle are considered[1]. This comes at the cost of a less intuitive interpretation of the results. However, if results are interpreted, they could give valuable insight in human coordination patterns using only a hand full of variables[17,18].

*Machine learning*
Along with the progress in 3D movement analysis acquisition techniques, data analysis techniques have improved. In addition to conventional summary statistics as mean and standard deviation, more elaborate measures such as Lyapunov exponents and other stability-related measures can be calculated. Kaptein et al. [13] hint at a potential caveat in movement analysis: they warn for a bias when selecting measures a priori without thorough theoretical considerations. They recommend a more unbiased method to determine a measures relevance: machine learning combined with a priori knowledge (i.e. the aforementioned educated guess). This minimizes user intervention by combining many candidate measures and classifying according to the ones that discriminate[13]. This novel multivariate approach has the unique property of being able to handle multimodal data since artificial neural network machine learning algorithms are based on non-linear statistics[19]. The conventional uni- and multivariate approaches are both linear.

*Classifying physical function in patients with knee osteoarthritis*
The machine learning approach was used to classify physical function in patients with knee osteoarthritis and healthy controls based on a multidimensional data set. Knee osteoarthritis (KOA) is a degenerative joint disease that leads to pain, stiffness, decreased range of motion (ROM), progressive articular cartilage breakdown and changes in underlying subchondral bone and synovium of the knee[20,21]. OA is the

most prevalent joint disease worldwide[20,22,23] and the knee is one of the most affected sites[24]. The aetiology of KOA is not yet fully understood[25] but risk factors include aging, BMI, female sex, decreased quadriceps strength, malalignment and genetics[26]. As a consequence of the unknown aetiology, treatment mainly focuses on symptom relief[21,27].
From patient perspective, physical function is an important treatment outcome measure[28] and optimizing it is a major aim in KOA treatment[29]. It is defined as an individual's ability to perform activities of daily living[29–32]. Currently, two physical function outcome measures exist: patient-reported outcome measures (PROMs) and performance-based (PB) tests. In PB tests, *capacity* or maximal performance is measured[32,33]. PROMs, by contrast, are questionnaires that measure perceived level of functioning during daily activities[30–32], so *perception of performance* is measured[33]. In KOA, a common PROM is the Knee injury and Osteoarthritis Outcome Score (KOOS) questionnaire[34]. A PB test that is suitable for KOA patients is the single-limb mini squat (SLMS) test[35], which measures the maximum amount of knee bendings in 30 seconds. The two types of measurement are considered complementary[33,36,37], which is illustrated by low-to-moderate correlations between PROMs and PB tests[28,36,38,39]. Consequently, to evaluate physical function, both tools have to be used[36,40]. However, no information on movement quality is used to evaluate physical function. Such information can be gathered by visual analysis of movement quality by clinicians[41] and by 3D movement analysis of gait and SLMS.
I used an artificial neural network machine learning approach for classifying physical function in patients with knee osteoarthritis. This approach generally allows for the combination of basic personal characteristics, conventional physical function measures, clinicians' movement quality ratings and kinematic parameters and time series (i.e. multimodal data). As a first step towards an automatic classification system to assist in screening, diagnosing, and planning and evaluating treatment[12], a binary classification discriminating patients and healthy controls was tested. Previous research resulted in maximal classification accuracies from 72-100%[1,12,22,42,43]. Only two articles included gait kinematics[1,43] and only angles of the affected knee were analysed. I proposed that including whole-body kinematics may enhance classification accuracy, since OA in one joint has strong influence on kinetics and kinematics in other joints[44]. Since a physical function classification tool should include multiple movements resembling daily activities, SLMS kinematics were also included in the data set.

*Aims*
The aims of this study were: 1) to highlight advantages and limitations of existing univariate and multivariate statistical methods using a multidimensional data set containing basic personal characteristics, conventional physical function measures, clinicians' movement quality ratings and gait and SLMS kinematic parameters and time series of patients with KOA and healthy controls, 2) to test a new multivariate supervised machine learning approach on this data set and highlight advantages and limitations of this approach, and 3) to evaluate classification accuracy, sensitivity and specificity of binary classifications discriminating patients and healthy controls based on input feature sets containing conventional physical function measures with or without kinematic parameters and time series.

## 2 Methods & Procedures

### *2.1 Participants*

Participants were 40 patients with knee osteoarthritis and 25 age- and gender-matched healthy controls. Patients were recruited from two orthopedic departments in Stockholm, Sweden. They were included if they (1) had physician diagnosed primary KOA and (2) were scheduled for unilateral total knee replacement surgery within one month after data collection. Healthy controls without any known musculoskeletal disease were recruited through acquaintances. This control group was matched to the KOA group by age strata across five age groups (40-49, 50-59, 60-69, 70-79, 80-89 years of age).
Additional inclusion criteria for both patients and controls were (1) to be able to walk 10 meters repeatedly without the use of a walking aid, and (2) to be able to understand verbal and written information in Swedish. Participants were excluded if they had had total joint replacement of the hip or knee in the last 12 months or other major orthopedic surgery in the lower extremities, if they had rheumatoid arthritis, diabetes mellitus, a neurologic disease and/or another condition affecting walking ability.

### *2.2 Data collection of subject characteristics and gait and SLMS kinematics*

All data were collected in the gait laboratory of the Karolinska Institutet by a physical therapist assisted by a technician. After signing informed consent and collection of basic personal information such as age, gender, height, weight and BMI, physical function was measured via existing measures: two patient-reported outcome measure questionnaires and one performance-based test. PROMs were the Knee injury and Osteoarthritis Outcome Score (KOOS) questionnaire[34], measuring functional status and quality of life[45] (QoL), and the EQ5D-3L[46], measuring the health-related QOL[46]. The performance-based test was the single-limb mini squat test[35]. It aims to measure lower extremity function[40] resembling conditions of daily life, such as stair ascent/descent[47], by counting the maximal amount of knee bendings on a single leg as possible in 30 seconds. Subjects performed a single-limb mini squat test on both legs. They were free to choose which leg to start with. Subjects were instructed to start in stance, with the long axis of their stance foot aligned with a straight line and their toes placed on a perpendicular line. Fingertip support for balance was provided. The aim was to bend their knee as many times as possible in 30 seconds, without bending forward from the hip and until the line along the toes could no longer be seen (at about 30º of knee flexion)[35]. Prior to or after the SLMS tests, subjects were asked to repetitively walk 6 meters at their own preferred speed.
For each SLMS and gait trial, three-dimensional movement analysis data were recorded by the Vicon Nexus Plug-In-Gait model[48]; resulting in 3D time series of marker positions, center-of-mass (CoM) and joint angles of the upper and lower body. Each SLMS test was recorded with a frontal video camera capturing the lower body and trunk to be able to collect clinicians' movement quality ratings.

### *2.3 Data collection of clinicians' SLMS movement quality ratings*

A clinician questionnaire including SLMS videos of 55 subjects was designed. Fifteen raters were recruited from two orthopedic departments of hospitals in Stockholm (Karolinska University Hospital and Ortho Center). Both medical doctors (N = 4) and physical therapists (N = 11) working with patients with knee osteoarthritis were included.
One video per subject was selected for inclusion in a clinician questionnaire. For patients, this was the video of the affected leg. For controls, the video of a random leg was selected. For time reasons, only the last ten seconds of squat performance were included. To ensure consistent video exposure duration to the raters, the squats of patients who could not perform 10 seconds of squatting were looped. The 55 videos were presented in random order. Raters were blind to the health status of the subject in the video. Raters were instructed to watch the full 10 seconds of squatting and in the subsequent 20 seconds rate overall movement quality on a 4-point ordinal scale, with a score of 1 representing 'poor' and a score of 4 representing 'good' overall movement quality. Raters were not given guidelines on which to base their ratings, as in [41,49]. Rather, they were asked to rate overall movement quality and provide comments on which movement characteristics they based their rating. Raters were provided with 5 example videos to get familiar with the protocol. Raters were instructed not to discuss their ratings with anyone and rewinding was not allowed.
Prior to data collection, a test rating session was organized. Feedback on the lay-out and instructions in the questionnaire, time to rate and comment, the amount of rating scales to fill in and the visibility of video loops was incorporated in the final protocol as described above.

### *2.4 Data pre-processing*

Before data pre-processing, for each control a random leg was selected to analyse as if it were the affected leg in patients. All data were pre-processed in the Vicon Nexus Plug-In-Gait system and in Matlab version R2013a. Output of the Vicon system consists of marker position data, from which center of mass position data and joint angles are calculated. Marker position data were filtered using Woltring's generalized cross-validation quintic spline with a predicted mean square error of 15mm to remove noise[48]. Gaps were interpolated using a cubic spline routine[48]. Sampling rate of the system was 100 Hz. Further pre-processing was conducted in Matlab as described below for the gait and SLMS kinematics separately. Clinicians' ratings were transformed (1 representing good and 4 representing poor movement quality) to make interpretation more comparable with KL scale (increasing KL score indicating more severe KOA).

### 2.4.1 Gait

For each person 5-16 gait trials were performed and recorded. The recorded part of each gait trial consisted of two consecutive gait cycles (GC) (leg 1 and leg 2). Heel strike was manually defined using the Vicon interface. Per subject, one gait trial was selected. Since in many trials, marker trajectories were noisy or missing, not all angle curves could be calculated. Only trials in which all angle curves were available during one gait cycle of the affected leg in patients and of the randomly selected leg in controls, were included. Since marker trajectories of the shoulder, elbow, wrist, spine, neck, and head angles were noisy in most of the gait trials, these angles were excluded from analysis (likewise for the CoM data). If none of the trials met the criteria, the subject was excluded from the statistical analyses on gait kinematic data. If multiple trials met the criteria, the earliest trial was selected.
The affected gait cycle was time normalized into 100 intervals (%GC), resulting in trunk and pelvis angles, and hip, knee and ankle angles of the affected side as well as of the contralateral (CL) side (Appendix Table A1). Gait spatiotemporal characteristics as listed in Table A2 (Appendix) were directly calculated using Vicon software. Discrete kinematic parameters listed in Table A2 (Appendix) were calculated manually from the time normalized gait cycle angle curves.

### 2.4.2 SLMS

Each subject performed one SLMS test per leg, resulting in two SLMS test trials per subject. For patients, the trial of the affected leg was selected; for controls, the randomly chosen leg trial. Only trials in which all angle curves and all three CoM position coordinates (Appendix Table A1) were available were included. As in the gait trials, shoulder, elbow, wrist, spine, neck and head angles were noisy in most of the trials and were excluded from analyses. For patients who could not perform the SLMS with the affected leg, no kinematic data were available so these subjects are also not included in the analyses. CoM position data were transformed from lab coordinates (left and right) into body coordinates (medial and lateral relative to the toe marker of the performing leg) to be able to compare CoM position between subjects performing the SLMS test with right and left leg.
For the selection of discrete kinematic parameters, whole test time series were used. For the purpose of visual comparison of squat cycle (SC) time series patients and controls, cycle starts and ends were defined by the time frames of knee extension peaks. Cycles were time normalized into 100 intervals (%SC). A mean time normalized squat cycle was calculated for each subject.
The spatiotemporal SLMS characteristics and discrete kinematic parameters listed in Table A2 (Appendix) were calculated in Matlab.

## 2.5 Statistical methods

Statistical analyses were conducted with IBM SPSS Statistics version 20.0. For the machine learning approach, the Matlab multivariate statistical toolbox UPMOVE[50] was used.

### 2.5.1 Univariate statistics

Based on previous research, an educated guess was taken about kinematic or spatiotemporal features that discriminate between healthy subjects and patients with KOA. Hypotheses for spatiotemporal characteristics and kinematic parameters are listed in Table A2 (Appendix).
Multiple independent t-tests were performed for the subject characteristics, PROM and PB test scores, clinician's rating and gait and SLMS spatiotemporal characteristics and discrete parameters derived from gait and SLMS kinematics. For data measured on nominal or ordinal scale and data that violated the assumption of normality, the non-parametric Mann-Whitney U-test was performed.
To correct for the increased chance of type I error (finding a false significant difference) as a result of doing multiple univariate tests[51], a Bonferroni correction was applied by dividing the single-test critical p-value of 0.05 by the amount of tests performed (N = 44, see Appendix C1). The adjusted critical p-value for each test to reject the null hypothesis was 0.0011 (rounded off).
Correlations between all subject characteristics (except for gender and KL score), PROM and PB test scores, clinicians' ratings and gait and SLMS spatiotemporal characteristics and kinematic parameters were calculated to gain insight in the relationships between variables. If data violated the assumption of normality or were measured on a nominal or ordinal scale, Spearman's rho was calculated instead of Pearson's r. Percentage of variable pairs significantly correlating as well as the average correlation were calculated.

### 2.5.2 Multivariate statistics: principal component analysis

Principal component analysis[17] (PCA) was conducted on the gait and SLMS time series.

#### 2.5.2.1 PCA input

$\underline{PCA_{Gait}}$ For every subject, a matrix of time series was constructed. This matrix consisted of 24 columns (see Appendix Table A1), each representing one joint angle. The 100 rows represented each time point on the gait cycle. Per column, the mean was subtracted (i.e. columns were detrended). All subjects' data were concatenated, resulting in a detrended angle PCA input matrix of $N_{subjects} \times 100$ rows by 24 columns.
$\underline{PCA_{SLMS}}$ For each subject, a matrix of time series was constructed. This matrix consisted of 25 columns (see Appendix Table A1), each representing one joint angle or CoM position. The rows represented each time point during the SLMS test. For controls, this matrix had 25 columns by 3000 rows. Since some

patients could not squat for 30 seconds, their number of rows is smaller. Per column, the mean was subtracted (i.e. columns were detrended). All subjects' matrices were concatenated, resulting in a detrended angle PCA input matrix of $N_{subjects}$ x $N_{framesoftest}$ rows by 25 columns.

*2.5.2.2 PCA output*

Output of the PCA consists of eigenvalues, indicating the amount of variance in the original data the corresponding PC accounts for, eigenvectors, describing the orientation of the PC axes in the old coordinate system (consisting of the initial variables), and time series projections on the PC axes.

Eigenvalues PC eigenvalue spectrum was plotted and the reduced number of PCs was determined using a 90% trace criterion as in [17,52] based on the "Cumulative percentage of total variation"-rule of Joliffe (2002)[53].

Eigenvector coefficients The contribution of each original variable to each of the PCs was inspected by plotting the eigenvector coefficients[17] of each PC. To evaluate which original variables contributed significantly to each PC, a broken stick test was applied[54]. Phase relations between variables within each PC can be evaluated by interpreting the signs of the eigenvector coefficients. Opposite signs indicate an anti-phasic relation (180º phase difference) between the original variables, equal signs indicate an in-phase relation between the original variables (0º phase difference).

Time projections on PCs In combination with the eigenvector coefficients, we evaluated the time projections on each corresponding PC. For $PCA_{Gait}$, gait cycle time projections on each of the relevant PCs were averaged over all subjects, patients and controls. For $PCA_{SLMS}$, for each subject, the time projection of each squat cycle in the SLMS test was selected from the SLMS test time projection on each of the relevant PCs. Each squat cycle time projection was time normalized into 100 intervals (%SC) and for each subject, an average squat cycle time projection on the relevant PCs was calculated. The average squat cycle time projections were averaged over all subjects, patients and controls.

Frequency Power spectral density was estimated of the concatenated time normalized cycle time projections using Welch's method[55] with a window size of 100 and 30% overlap. Frequencies of the time projections were identified by visual inspection of the peaks in the power spectral density – frequency plots.

Phase In order to check for pairs of PCs that are not independent, relative phases were calculated. Two variables having a 90º phase difference are not independent from each other, as is the case with sine and cosine time series (that per definition have a 90º phase difference). If you know the sine, the time series of the cosine is completely known. However, in state space sine and cosine describe a circle[17], which can only be described in two dimensions. Identifying pairs of PCs with a 90º phase difference indicates that the amount of dimensions to describe the original data in could be even more reduced by one PC per PC pair.

For each PC, continuous phase was determined using the Hilbert transform as described in [18]. Relative phase was calculated between all PCs with the same frequency. Average and standard deviation of phase and relative phase were calculated using directional (circular) statistics for all subjects, only patients and only controls. To test for significant differences in relative phase between patients and controls, the circular Kuiper two-sample test[56] was conducted.

Differences between patients and controls The variance of the time projection ($\sigma_\xi^2$) of each subject's gait/squat cycle on each PC was calculated and used as a measure of strength of that PC in that subject. To identify significant differences in the contribution of patients and controls to each PC, for each PC an independent t-test was conducted between $\sigma_\xi^2$ values of patients and controls. For data that violated the assumption of normality, the non-parametric Mann-Whitney U-test was performed. To correct for the increased chance of type I error (finding a false significant difference) as a result of doing multiple univariate tests, a Bonferroni correction was applied[57] by dividing the single-test critical p-value of 0.05 by the amount of tests performed (N = 24 (gait), N = 25 (SLMS), see Appendix Table A1). The adjusted critical p-value for each test to reject the null hypothesis was therefore 0.0021 (rounded off) for gait data and 0.0020 for SLMS data.

*2.6 Multivariate statistics: supervised machine learning*

Supervised machine learning was used for classifying physical function in patients with KOA and healthy controls. Due to the small data set[58], a binary classification was made using KL score as a physical function class label. A subject belongs either to the class of patients or to the class of healthy controls. Patients have a KL score of 1 or higher (per definition) and healthy controls do not have a KL score (coded as a 0). During learning, these class labels provide feedback (i.e. supervision) to the machine learner.

*2.6.1 Algorithm*

Learning vector quantization (LVQ)[59] is a distance-based competitive-layer artificial neural network algorithm. It consists of a competitive layer of neurons and a second output layer that has as many neurons as pre-defined classes. Each competitive-layer neuron is defined by a codebook vector of weights for all input features. The data set is divided in a training set of subjects for training the network and a test set for evaluating machine learner performance. Each subject is represented by an input vector containing the input features of that subject. During training subjects also have a target vector containing their class label. For each input vector, Euclidean distance to all neurons is calculated; the closest neuron wins. During training, codebook vectors of neurons are updated: if the prediction is correct according to the target vector, the winning neuron moves towards the input vector. If it was not correct, it moves

away and its weights are updated accordingly. During testing, no adaptations are made anymore. Instead, machine learner classification accuracy is evaluated by calculating the amount of correctly classified subjects as a percentage of the total amount of subjects in the test set. Sensitivity was calculated as the amount of correctly classified patients as a percentage of the total amount of patients in the test set. Specificity was calculated as the amount of correctly classified controls as a percentage of the total amount of controls in the test set.

### *2.6.2 Settings*

Our neural network consisted of a hidden competitive layer of 20 neurons and 2 output neurons (Figure 1). As recommended by Kohonen, the LVQ1 algorithm is used with learning rate of 0.05[60]. For each run, a random half of the subjects was selected to serve as the training set. Since the data set consisted of 24 controls and 31 patients, this resulted in 12 controls and 16 patients. The other half of the cases served as the test set. In each training phase, the input vectors of the training set were presented 500 times[60]. Each run resulted in one accuracy, sensitivity and specificity value. To be able to evaluate classification accuracy, sensitivity and specificity statistically, 50 runs per input feature set were executed.

### *2.6.3 Input features*

Input features are summarized in Appendix B. Subject characteristics, spatiotemporal characteristics and discrete kinematic parameters as studied in the univariate statistical method were used as input features. Additional input features were the cycle time projections (ξ) per PC and variance of those cycle time projections as derived from the $PCA_{Gait}$ and $PCA_{SLMS}$. For the reason of computational speed, only time projections and variance of those time projections on the reduced number of PCs as defined with the 90% trace criterion were used as input features. If variance of time projections on higher PCs had turned out to be significantly different between patients and controls, the time projections and variance of those time projections were used as input feature, too. Eight different sets of input features were defined, combining the subsets "Subject characteristics", "Gait" and "SLMS" and in- and excluding the KL score. The largest set included 45 features.


**Figure 1. Artificial neural network consisting of an input layer of N neurons, a competitive hidden layer of X neurons and an output layer predicting two classes (Y).** Each arrow represents a neuronal connection with weight $W_{Feat\_n,Neuron\_x}$ (input-hidden connection) or with weight $W_{Neuron\_x,Class\_y}$ (hidden-output connection).

### *2.6.4 Validation*

As a first test of machine learner validity, only the KL score was used as input feature. Because class labels are based on these KL scores, this should result in a classification accuracy of 100%. The same was realized by adding the KL score to the full feature set.

### *2.6.5 Influence of training set proportion*

To evaluate if increasing the proportion of the training set would result in higher classification accuracies, for the maximal feature set (all features without the KL score) two additional analyses were conducted. In the first, training set size was minimized, resulting in a training set of 1 control and 1 patient. In the second, training set size was maximized, resulting in a training set of 23 controls and 30 patients. Minimal and maximal training set proportion percentages were calculated for controls (1/24*100 for minimal and 23/24*100 for maximal) and applied to the patients to calculate the amount of patients in the minimal and maximal training set. Values for patients were rounded off.

### *2.6.6 Statistical evaluation of classification accuracies, sensitivities and specificities*

To test for significant differences in classification accuracy between the different sets of input features, a Friedman's ANOVA was conducted followed by post-hoc Wilcoxon signed-rank tests if Friedman's ANOVA indicated a significant influence of input feature set on classification accuracy. In the post-hoc tests, a Bonferroni correction was applied since multiple univariate dependent t-tests are applied. Statistical tests were non-parametric and dependent as recommended by [61].

### *2.6.7 Evaluation of discriminative features*

To evaluate the discriminative capacity of each feature, for each run of the maximal feature set (without KL score) weights of the connections between the input neurons, hidden competitive layer neurons and output neurons of the trained network (see Figure 1) were used to calculate the relative contribution (%) of each input feature data point to the class prediction of the machine learner using the method of Garson[62] described by Goh[63]. Average relative contribution for each input feature data point and standard deviation were calculated over all runs and both output neurons.

The maximal feature set contains 45 features. Most features exist of only one data point, but the principal components (100 data points per PC) exist of more than one data point, resulting in a total of 1863 data points. The contributions of each data point are summed to extract one relative contribution value per PC. Since PC contributions are the sum of multiple data points, comparing these values with the

contribution of single data points is unfair. Therefore, relative contributions of each feature were compared visually with the relative contribution if each data point of the feature set would have contributed equally to classification (100/1863 = 0.054% (rounded off) per data point, so 0.054% * $N_{datapoints}$).

## 3 Results

Only data of subjects with both gait and SLMS kinematics were analysed except in the PCA, resulting in a subset of 31 patients and 24 controls. Due to problems with marker visibility and noise, gait kinematics were not available or excluded for 1 patient and 1 control. For 8 patients and 1 control, no SLMS kinematics were used in analyses, either because the patient could not perform the test or because of problems with marker visibility and noise or with saving the recordings.

### *3.1 Univariate statistics*

#### *3.1.1 Subject characteristics, PROM and PB test scores, clinicians' ratings*

Subject characteristics, PROM and PB test scores and clinicians' ratings are summarized in Table B1 (Appendix) and visualized in Figure 2. No significant differences between patients and controls were found in gender distribution, age and height. All patients had moderate to severe KOA as indicated by KL scores of grade 3 and 4. Patients had significantly higher weight and BMI compared to controls. All PROM scores were significantly lower (i.e. worse) for patients. In addition, patients performed significantly less knee bendings in the SLMS test and their movement quality received significantly worse clinicians' ratings compared to controls.

#### *3.1.2 Gait spatiotemporal characteristics and kinematics*

Regarding gait spatiotemporal characteristics, patients had significantly lower speed and stride length compared to controls (Appendix Table B2, Figure 2). Regarding the gait kinematics, no significant differences were found in trunk and pelvis parameters. Peak hip adduction during stance was significantly lower in patients. Patients had significantly lower peak knee extension and knee flexion-extension range of motion compared to controls. No differences were found in ankle kinematic parameters.

#### *3.1.3 SLMS spatiotemporal characteristics and kinematics*

For the SLMS spatiotemporal characteristics, patients had a significantly lower limb symmetry index compared to controls (Appendix Table B2, Figure 2). For the SLMS kinematics, no difference was found in trunk tilt parameters. Knee flexion-extension range of motion was significantly lower in patients.

#### *3.1.4 Correlations*

Of the 861 pairs of discrete variables, 368 correlations were significant ($p < 0.05$) (42.7%). Average absolute significant Spearman's rho was 0.46, with a minimum value of 0.17 and maximum value of 1.

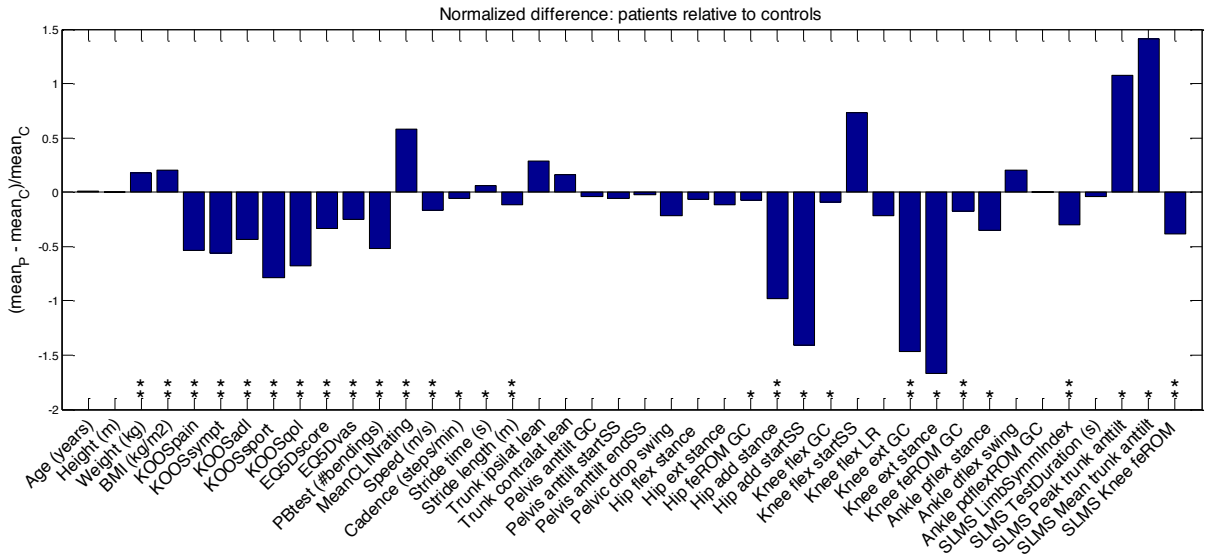



**Figure 2. Normalized difference in means between patients and controls of subject characteristics, PROM and PB test scores, clinicians' rating, spatiotemporal gait characteristics, discrete gait variables, spatiotemporal SLMS characteristics and discrete SLMS characteristics of patients and controls.** Bars indicate difference between patients and controls, divided by the mean of all controls. Single stars indicate a significant difference between patients and controls at the level of $p<0.05$; double stars indicate a significant difference after Bonferroni correction.

### 3.2 Multivariate statistics: principal component analysis (PCA)

#### 3.2.1 Eigenvalues and amount of PCs

For both PCAs the first five principal components were needed to explain >90% of the variance in data (Appendix Table D1 & E1).

#### 3.2.2 Eigenvectors and time projections of $PCA_{Gait}$

Significant eigenvector coefficients for $PCA_{Gait}$ are indicated by a star in Figure D2 (Appendix). Average time projections of patients and controls on each PC are depicted in Figure 3. Frequencies are depicted in Figure D1 (Appendix) and phase relations are summarized in Table D3 (Appendix).

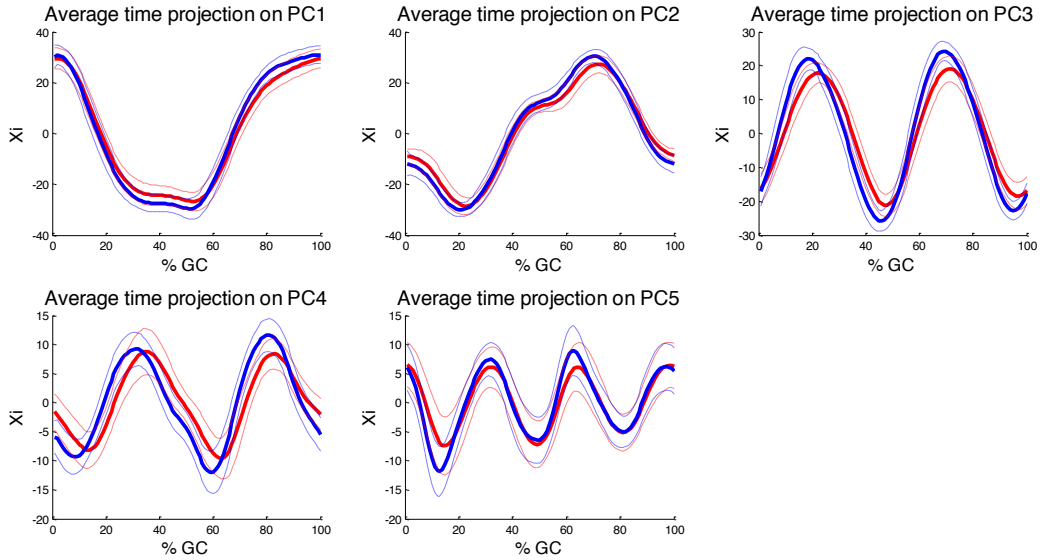



**Figure 3. Average time projections of patients (red) and controls (blue) on PC1-5 in $PCA_{Gait}$.** Bold lines indicate average time projections, dotted lines indicate standard deviations. Note the difference in y-axis scaling between subplots.

Significant contributors to $PC1_{gait}$ were the hip flexion angle and contralateral hip flexion angle. Since the eigenvector coefficient of the affected hip had a positive sign and the coefficient of the contralateral hip a negative, the hips moved in anti-phase with each other (Appendix Figure D2, upper left panel). For $PC2_{gait}$, the knee flexion angles were significant eigenvector coefficients. Again, the knees moved in anti-phase (Appendix Figure D2, upper right panel) and in phase with the hip flexion pattern described by $PC1_{gait}$. Both $PC1_{gait}$ and $PC2_{gait}$ had a frequency of once per gait cycle (stride event). However, an average phase difference of approximately 90º existed between $PC1_{gait}$ and $PC2_{gait}$ (Appendix Table D3), indicating that PC1 and PC2 are not independent but together describe a pattern of 180º phase difference between the hips of both sides and between the knees of both sides, but with a 90º phase difference between hips and knee of the same body side. The hips flex a quarter of a gait cycle later than the knees (Appendix Figure D2, upper panels and Figure 3).

For $PC3_{gait}$, again knee flexion angles were the significant contributors. However, here the knees flexed in phase with each other (Appendix Figure D2, middle left panel). So, on top of the anti-phase knee flexion pattern described by $PC1_{gait}$, an in-phase knee flexion pattern existed that had a frequency of twice per gait cycle, so once per step (Appendix Figure D1 and Figure 3). It thus represents a step event. On top of the already existing knee flexion pattern of $PC2_{gait}$, during each step both knees exhibited flexion simultaneously.

$PC4_{gait}$ had the same frequency as $PC3_{gait}$ (Appendix Figure D1), but a phase difference of approximately -90º (Appendix Table D3). Significant eigenvector coefficients of $PC4_{gait}$ were ankle dorsiflexion angles, as well as the contralateral foot progression angle and hip flexion angle. The hips flexed and ankles dorsiflexed in an in-phase pattern that is in phase with the knee flexion pattern described by $PC3_{gait}$ but that is delayed with a quarter gait cycle (-90º phase shift). The feet toed in (foot progression angle) in anti-phase with that pattern (Appendix Figure D2, middle right panel). So, during each step, a quarter gait cycle later than the knee flexion in $PC3_{gait}$, both ankles dorsiflex and toe out simultaneously. On top of the anti-phasic stride hip flexion pattern of $PC1_{gait}$, an in-phase hip flexion pattern existed that occurs twice per stride and a quarter gait cycle later than the in-phase step knee flexion pattern of $PC3_{gait}$.

The frequency of $PC5_{gait}$ was three times per gait cycle (Appendix Figure D1). Significant contributors to $PC5_{gait}$ were – again - ankle dorsiflexion angles, but also ankle and hip internal rotation angles (Appendix Figure D2). However, the coordination patterns described $PC5_{gait}$ were all anti-phasic regarding body sides, with ankle internal rotation occurring in anti-phase with the other two joint movements.

$PCA_{Gait}$ revealed significant differences between $\sigma_\xi^2$ of PCs 1-3 (Appendix Table D2). Patients had significantly less variance in their time projections than controls on those PCs. So, patients exhibited the stride anti-phasic hip flexion pattern and the anti-phasic stride knee flexion pattern combined with the in-phase

step knee flexion pattern less than controls. The total explained variance by the PCs with a significant different $\sigma_\xi^2$ was 83.89% (Appendix Table D1).

### *3.2.3 Eigenvectors and time projections of PCA$_{SLMS}$*

All relevant PCs in PCA$_{SLMS}$ had a basic frequency of once per squat cycle (Appendix Figure E1). Average time projections are depicted in Figure 4. Although visual inspection of Figure 4 indicates differences in time projection trajectories of PC2$_{SLMS}$ – PC19$_{SLMS}$, this could not be quantified by the variance of the time projections: PCA$_{SLMS}$ only revealed a significant difference between $\sigma_\xi^2$ values of PC2$_{SLMS}$ and PC19$_{SLMS}$ (Appendix Table E2).

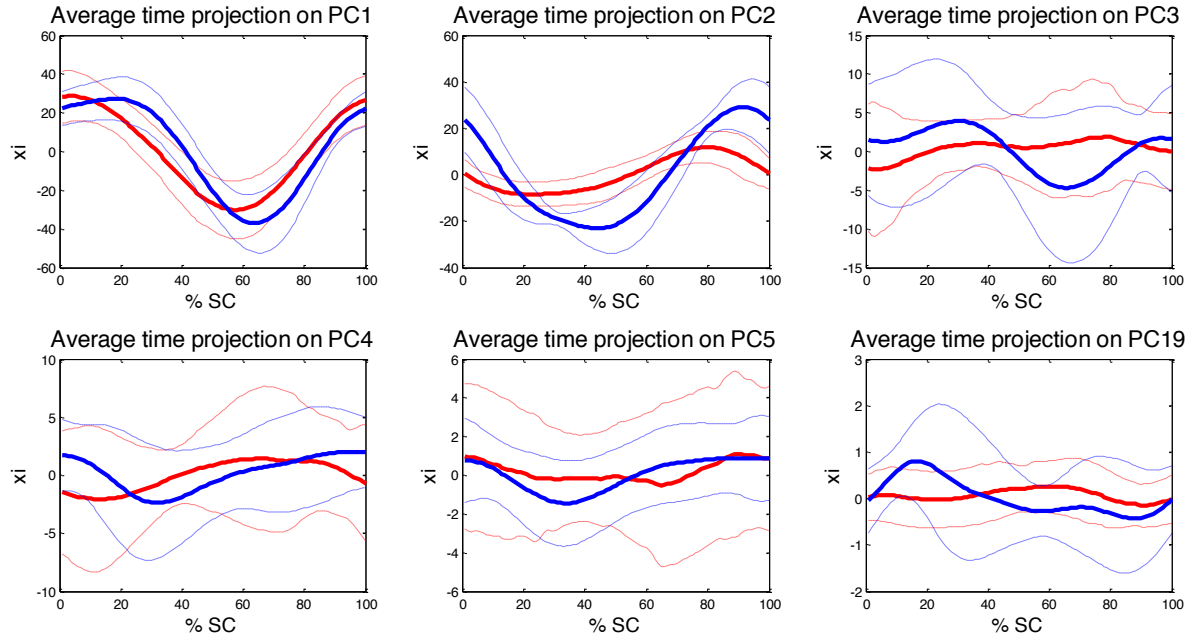



**Figure 4. Average time projections of patients (red) and controls (blue) on PC1-5 and 19 in PCA$_{SLMS}$.** Bold lines indicate average time projections, dotted lines indicate standard deviations. Note the difference in y-axis scaling between sub-plots.

PC1$_{SLMS}$ and PC2$_{SLMS}$ had an average phase difference of 90º, but standard deviation was high. Significant differences in average phase differences between PCs were found between patients and controls for the combinations PC1$_{SLMS}$ – PC3$_{SLMS}$, PC2$_{SLMS}$- PC4$_{SLMS}$ and PC2$_{SLMS}$ – PC19$_{SLMS}$ (Appendix Table E3).
Eigenvector coefficients are depicted in Figure E2 (Appendix). The only significant contributor to PC1$_{SLMS}$ was the upward CoM coordinate. For both patients and controls, the PC1$_{SLMS}$ time projection decreases after a slight rise and subsequently increases again, reflecting the drop and rise in vertical CoM position during a squat.
The upward CoM coordinate also significantly contributed to PC2$_{SLMS}$. Other significant contributors of this PC were the knee flexion angle and ankle dorsiflexion angle of the squat leg. The time projection of PC2$_{SLMS}$ has a phase difference of approximately 90º with PC1 in both patients and controls (Appendix Table E3) and describes a pattern in which CoM downward movement starts earlier compared to PC1 (and earlier in patients compared to controls, based on visual inspection). So on top of PC1, a phase shifted downward CoM movement combined with in phase knee extension and ankle plantar-flexion exists. This pattern is significantly more pronounced in controls compared to patients (Appendix Table E2).
The coordination pattern described by PC2$_{SLMS}$ is somewhat counterintuitive. However, the combined knee extension and ankle plantar-flexion during the descending phase of the squat come on top of the pattern described by PC1$_{SLMS}$. Although the only significant eigenvector coefficient was the upward CoM coordinate, in Figure E2 (Appendix) it can be seen that in PC1$_{SLMS}$, the upward CoM coordinate is in anti-phase with hip and knee flexion and ankle dorsiflexion. So the PC1$_{SLMS}$ pattern describes combined hip and knee flexion and ankle dorsiflexion during the descending squat phase.
In PC1$_{SLMS}$, the backward CoM coordinate was a non-significant eigenvector coefficient that was in anti-phase with the upward CoM coordinate. So in PC1$_{SLMS}$, a higher upward CoM position comes with slightly less backward lean. The backward CoM coordinate was the only significant contributor to PC3$_{SLMS}$ and it was in phase with the upward CoM coordinate. On top of the coupled squat rise – forward lean pattern, a squat rise – backward lean pattern is present. In controls, the time projection followed a trajectory similar to the time projection on PC1$_{SLMS}$, but patients seemed to keep their backward lean more constant over the squat cycle. This difference is visible by a significant difference in relative phase between PC1$_{SLMS}$ and PC3$_{SLMS}$ between patients and controls (Appendix Table E3).
The significant eigenvector coefficients of PC4$_{SLMS}$, the contralateral hip and knee flexion angle, moved in phase with each other but in anti-phase with the previous PC contributors. Their time trajectories

seemed to be phase shifted relative to each other (Figure 4), with an earlier decrease in contralateral hip and knee extension in patients compared to controls. This indicates a pattern of increasing contralateral knee and hip flexion during the downward squat phase and decreasing it during the upward squat phase.
Significant contributors to $PC5_{SLMS}$ was the mediolateral CoM coordinate, which was in phase with the contributors of PC1-3. During the downward phase of each squat, patients and controls move their CoM from medial to lateral.
Since $PC19_{SLMS}$ appeared to have significantly lower variance of the time projections of patients compared to controls (Appendix Table E2), it is interesting to evaluate which variables significantly contribute to this PC. Significant contributors to $PC19_{SLMS}$ were the trunk anterior tilt and internal rotation angle, pelvic anterior tilt angle, hip and knee internal rotation and ankle dorsiflexion angle. Ankle dorsiflexion, trunk and pelvis anterior tilt occur in phase with each other, but in anti-phase with trunk internal rotation. Trunk and hip and knee of the squat leg internally rotate in phase with each other. This indicates that this pattern of ankle dorsiflexion and anterior tilt with simultaneous trunk, hip and knee external rotation is less pronounced in patients. However, the total explained variance by this PC was only 0.11% (Appendix Table E1) so it indicates only slight differences between patients and controls.

### *3.3 Multivariate statistics: supervised machine learning*

#### *3.3.1 Validation*

As expected, the classification accuracy of the input feature set consisting of the KL score was 100% (median), with an interquartile range of 0. However, when adding all other features, classification accuracy significantly decreased to a median of 81.5 [18.5] (median [IQR]) (Appendix Table F1, Figure 5).

#### *3.3.2 Classification accuracies, sensitivity and specificity values*

Classification accuracies of the different input feature sets are summarized in Table F1 (Appendix, left column) and Figure 5 (left panel). For classifications without KL score, maximal classification accuracy was 96.3% [3.7] (median [IQR]) when using the "Subject characteristics" subset either with or without the "Gait" subset. Minimal classification accuracies were achieved by the subset "SLMS" either with (74.1% [11.1) or without (74.1% [18.5]) the "Subject characteristics" subset.
Friedman's ANOVA indicated a statistically significant difference in classification accuracy between the sets of input features, $\chi^2(8) = 266.613$, $p < 0.01$. Post-hoc tests are summarized in Table F1 (Appendix, right column) and revealed a significantly higher classification accuracy of the "Subject characteristics" and "Subject characteristics + Gait" sets compared to all other sets. Additionally, the "Gait" set had a significantly higher classification accuracy compared to the "SLMS" set and the "Subject characteristics + SLMS" set.

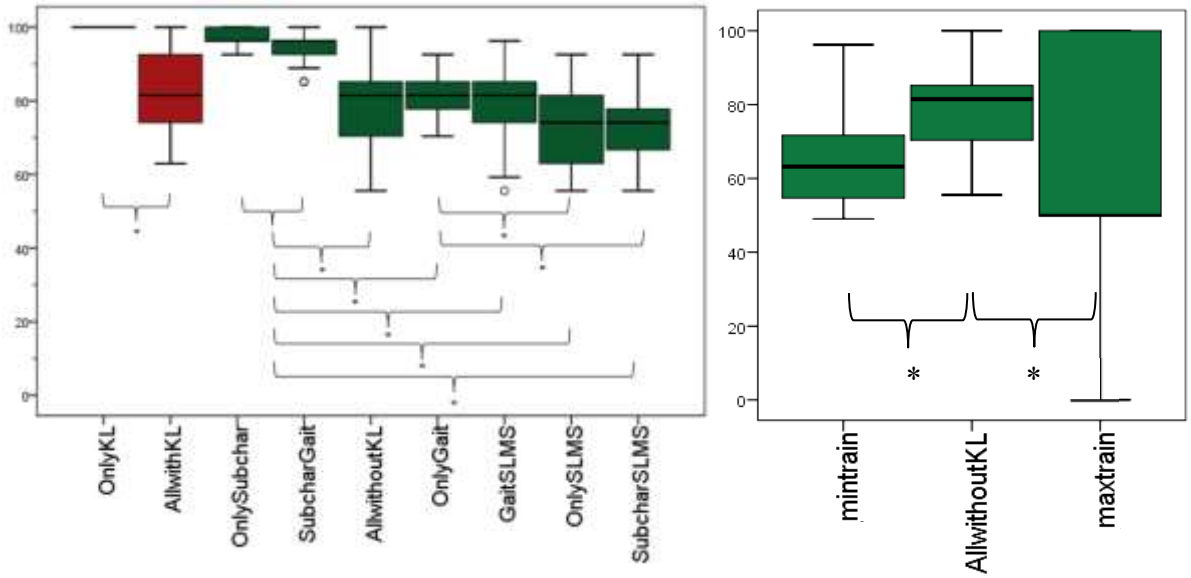



**Figure 5. Classification accuracies for each set of input variables (left panel) and for the complete feature set without KL-score with minimal, standard and maximal training set (right panel).** Tests including the KL score are colored red, tests without KL score are colored green. Boxes indicate median, first and third quartile. Whiskers indicate minima and maxima, unless outliers are present (dots). Stars indicate significant difference after Bonferroni correction.

Sensitivities (Appendix Table F3, Figure 6 (left panel)) ranged from 100% [0.0] (median [IQR]) for the "Subject characteristics" set to 79.6% [8.3] for the "Subject characteristics + SLMS" set. There was a statistically significant difference in sensitivity between the sets of input features, $\chi^2(7) = 193.634$, $p < 0.01$. Post-hoc tests revealed a significantly higher sensitivity of both the "Subject characteristics" and "Subject characteristics + Gait" sets compared to all other sets. The "Gait" set had a significantly higher sensitivity compared to the "Gait + SLMS", "SLMS" and "Subject characteristics + SLMS" sets.

Specificities (Appendix Table F4, Figure 6 (right panel)) ranged from 100.0% [0.0] (median [IQR]) for the "Subject characteristics + Gait" set to 92.6 [7.4] for the "Gait" set. A statistically significant difference in specificity existed between the sets of input features, $\chi 2(7) = 77.271$, $p < 0.01$. Post-hoc tests revealed a significantly higher specificity of the "Subject characteristics" and "Subject characteristics + Gait" set compared to all other sets.

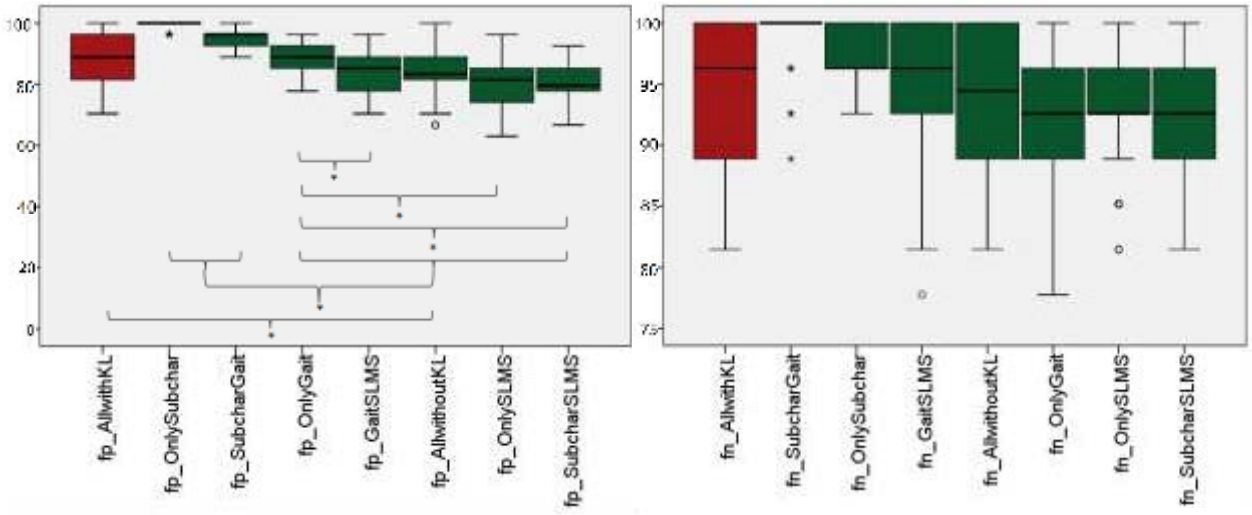



**Figure 6. Sensitivity (left panel) and specificity (right panel) for 7 different input feature sets.** Sensitivity is defined as the percentage of cases correctly classified as patients. Specificity is defined as the percentage of cases correctly classified as controls. Tests including the KL score are coloured red, tests without KL score are colored green. Boxes indicate median, first and third quartile. Whiskers indicate minima and maxima, unless outliers are present (dots). Stars indicate significant difference after Bonferroni correction. Note the difference in scale of the vertical axes and the different order of input feature sets on the horizontal axes.

### *3.3.3 Influence of increasing training set proportion*

A statistically significant difference in classification accuracy of the "All features without KL score" set was found between running the set with different training set proportions, $\chi^2(2) = 18.121$, $p < 0.01$. Classification accuracies are summarized in Table F2 (Appendix) and Figure 5 (right panel). Post-hoc tests revealed significantly lower classification accuracies of both the minimal (63.2% [17.5], median [IQR]) and maximal (50.0% [50.0]) training set sizes compared to the standard training set size (81.5 [14.8]). No significant difference was found between the classification accuracies of the minimal and maximal training sets.

### *3.3.4 Evaluation of discriminative features*

Relative contributions of each feature to classification are listed in Table G1 (Appendix) and visualized in figure 7. Gender, age, PB test score and all PROM scores except the EQ5D score had a relatively higher contribution compared to the situation in which all data points contribute equally to classification. This is also true for cadence, peak hip flexion and hip flexion ROM during gait, peak knee flexion over the gait cycle and during LR, knee flexion-extension ROM during gait, ankle plantar-dorsiflexion ROM over the gait cycle and pelvic anterior tilt during the gait cycle and at the start and end of single-stance. For SLMS kinematics, knee flexion-extension ROM and all anterior trunk tilt parameters contributed more than average to the classification. The variances of xi on the first five PCs of gait and on the first six PCs of SLMS contributed more than average per data point to class prediction. The relative contributions of PC1-3 of gait and PC1-2 of SLMS were higher compared to the situation with equal data point contributions.

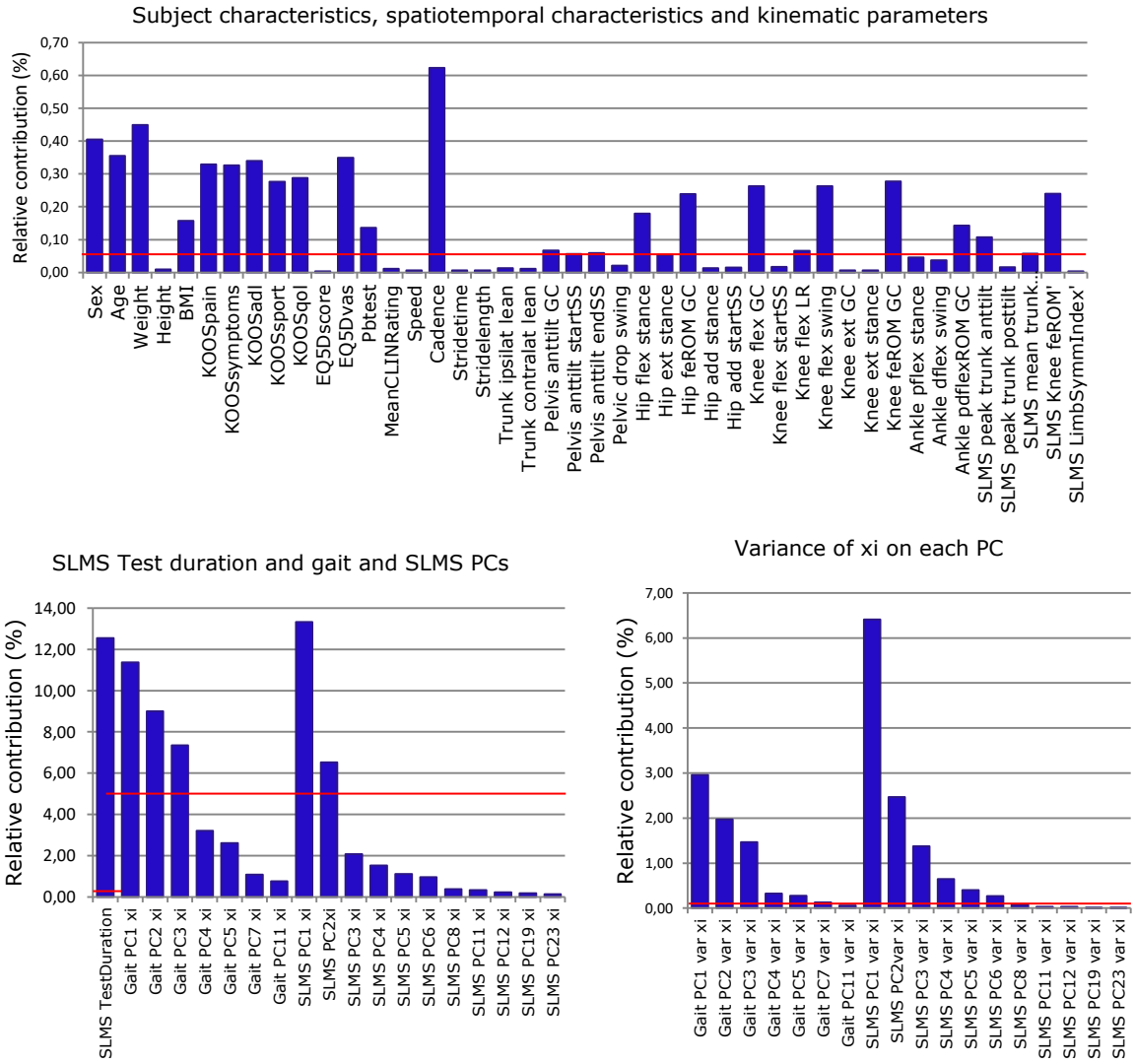



**Figure 7. Relative contribution of input features to class prediction.** Upper panel: subject characteristics, spatiotemporal characteristics and kinematic parameters for gait and SLMS. Lower left panel: SLMS test duration and gait and SLMS PCs. Lower right panel: variance of xi on each PC of gait and SLMS. Red line indicates the contribution if each data point had contributed equally (calculated as $N_{datapoints}$*100/1863 = 0.054%*$N_{datapoints}$). Note the different scales of the vertical axes. No 95% confidence intervals are depicted due to their small size.

# 4 Discussion

The aims of this study were to highlight advantages and limitations of existing univariate and multivariate statistical methods and a new multivariate supervised machine learning approach on a multidimensional data set of patients with KOA and healthy controls, and to evaluate classification accuracy, sensitivity and specificity of binary classifications discriminating patients and healthy controls based on input feature sets containing conventional physical function measures with or without kinematic parameters and time series.

## *4.1 Univariate statistics*

The conventional univariate statistical approach revealed significant differences between patients and controls in weight, BMI and all of the PROM scores, with patients having significantly lower physical function scores than controls. Patients had a significantly worse clinicians' rating and performed significantly less knee bendings in the SLMS test. After Bonferroni correction in only 5 of the 29 proposed gait and SLMS spatiotemporal characteristics and kinematic parameters significant differences were found. Patients had a significantly lower gait speed and stride length, lower peak hip adduction during stance, a lower peak knee extension over the gait cycle and a lower knee flexion-extension ROM both during gait and SLMS. In addition, patients had a significantly lower limb symmetry index.

I tested 44 variables based on an educated guess. It can be tempting to include even more variables in the hope of finding significant differences, even without educated guess. However, the sky is not the limit. As indicated by the high percentage of significantly correlated variables (42.7%) and the average absolute significant correlation of 0.46 with a range from 0.17 to 1, the covariance between variables is high. This means that some of the information is redundant. The higher the amount of variables to test, the lower the critical p-value level will be due to the Bonferroni correction[51] to correct for the chance of finding a significant difference by chance (Type I error) and thereby decreases the chance of finding a significant difference that does exist. For example, in the present study both cadence and stride time were (erroneously) included in the univariate statistical approach. These variables are inversely related (correlation = -1) and thus, redundant information is measured. Doing two multiple t-tests will now result in a critical p-value level of 0.05/2, which decreases the chance of finding a significant difference relative to the critical p-value level of 0.05 if only one t-test had been conducted. This illustrates an important practical and theoretical disadvantage of the univariate statistical approach.

## *4.2 Multivariate statistics: PCA*

Both principal component analyses resulted in a data reduction of 24 or 25 time series to 5 principal components explaining 90% of the variance in the data. This again points out the high covariance between variables (here: time series). $PCA_{Gait}$ revealed two pairs of PCs, the first pair describing stride events and the first pair describing step events. PC1 and 2 describe a pattern of anti-phasic hip flexion pattern preceded by an anti-phasic knee flexion pattern a quarter gait cycle earlier during each stride. PC3 and PC4 describe a pattern of in-phase knee flexion followed by in-phase hip flexion and ankle dorsiflexion patterns a quarter gait cycle later during each step. The last PC captures ankle dorsiflexion and hip and ankle internal rotations at a frequency of 3 times per gait cycle. Patients were found to exhibit the stride pattern to a lesser extent than controls, as well as the in-phase step knee flexion pattern. This is in accordance with the finding of a significantly lower knee flexion-extension ROM during the gait cycle in patients in the univariate statistical approach. However, differences in hip flexion and ankle dorsiflexion were only significant without Bonferroni correction in that approach. Previous PCA of gait of patients with KOA and healthy controls only included kinematic (and kinetic) time series of the knee, so results are not very comparable. However, from the knee angle curves, the knee flexion-extension ROM proved to be the most discriminative PC[1], which is supported by our finding.

$PCA_{SLMS}$ revealed five PCs with a frequency of once per squat cycle. Eigenvector coefficients were dominated by the CoM position time series. For the upward CoM coordinate, this can be explained by the fact that the standard deviation of the original time series was approximately 2-3x higher compared to all other time series.

The pair of PC1 and PC2 together described downward-upward CoM movement achieved by a basic pattern of in-phase ipsilateral hip and knee flexion and ankle dorsiflexion (PC1) with on top of it a pattern of in-phase ipsilateral knee extension and ankle plantarflexion (PC2). PC1 and 3 revealed a pattern of in-phase upward-backward CoM movement (PC3) on top of the upward–forward CoM movement (PC1). Contralateral in-phase hip and knee flexion during the downward squat phase was captured by PC4 and PC5 captured lateral CoM movement during the downward squat phase.

Only a significant difference in variance of the time projections on $PC2_{SLMS}$ and $PC19_{SLMS}$ were found between patients and controls. Patients exhibited the pattern of downward CoM position with in-phase ipsilateral hip and knee flexion combined with ankle dorsiflexion ($PC2_{SLMS}$) to a lesser extent than controls. This is in accordance with the finding of a significantly lower knee flexion-extension ROM during the SLMS test in patients in the univariate statistical approach. $PC19_{SLMS}$ was also found to have significantly lower variance in time projections of patients compared to controls. Patients exhibited a pattern of ankle dorsiflexion and anterior tilt with simultaneous trunk, hip and knee external rotation to a lesser extent. This difference is only subtle since $PC19_{SLMS}$ explains only 0.11% of the variance in the original data, but still, discriminative capacity can be high[52].

Although only two PCs were found to significantly differ between patients and controls in $PCA_{SLMS}$, visual inspection of the time projections (Figure 5) indicated differences between patients and controls in timing of squat coordination. This could not be quantified by the variance of time projections, except in $PC2_{SLMS}$ and $PC19_{SLMS}$. An appropriate method would be to calculate the average relative phase between patient and control time projections. Differences may be attributable to the slower squat speed in patients as indicated by the significantly lower amount of squats in 30 seconds (Table 3, Figure 1) and to the fact that patients moved less smooth compared to controls (reported by raters, results presented elsewhere). The latter might be related to the subtle differences in squat pattern described by $PC19_{SLMS}$.
A major advantage of PCA over the univariate statistical approach is that it accounts for the covariance between variables, as illustrated by the high eigenvalues of the first five PCs (> 90%). This makes it possible to analyse large numbers of time series and to reveal complicated coordination patterns by interpreting eigenvalues, eigenvectors and time projections of a small number of PCs. The approach is considered less biased compared to the previous, since only the selection of input time series is subjective. The extraction of the reduced amount of meaningful variables (here: PCs) is fully objective. Together, this makes PCA more suitable for exploratory studies[13]. In the present study, the use of PCA made it possible to include not only affected body side angle curves, as is common practice in the univariate statistical approach, but also unaffected body side angles.
A disadvantage of PCA is that interpretation is time-consuming and less straightforward compared to the univariate statistical approach. In addition, both PCA and the univariate statistical methods are based on linear statistics so only linear relations can be revealed. Lastly, if you want to compare differences between two groups after PCA, a measure has to be selected to quantify the difference and again, multiple univariate statistical tests have to be conducted. So, still some subjectivity is present[13].

### *4.3 Multivariate statistics: supervised machine learning*

The validation with a feature set of only KL score resulted in the expected classification accuracy of 100%. This was expected since subjects were either classified as patient or healthy control, with a patient defined as someone with a KL score (coded as a number between 1-4) and a healthy control as someone without a KL score (coded as a zero). However, adding all other feature subsets significantly decreased classification accuracy to 81.5% [18.5] (median [IQR]). The feature set including all features did not significantly differ from the feature set including all features but no KL score regarding accuracy and specificity. However, the feature set including all features had a significant higher sensitivity.
The lower accuracy after adding features to the KL score indicates that the machine learner is influenced by non-relevant information. Ideally, this is not the case so further improvement of the machine learner algorithm is needed and classification accuracies, sensitivities and specificities reported in this study must be taken with caution. A method to improve the performance of the machine learner is to increase the weights of relevant input features manually. This is an important advantage of supervised learning. However, the relevance of features for classification is often not known. Still, in case of a validation trial, weights of the reference (here: KL score) can be increased. Network connections will adapt and network performance can be evaluated without the reference input feature.

Maximal classification accuracies, sensitivities and specificities were achieved by the "Subject characteristics" and "Subject characteristics + Gait" input feature sets, that had significantly higher accuracies and sensitivities than all other input feature sets. Adding the "SLMS" subset significantly decreased classification accuracies and sensitivities, except for accuracy after adding the "SLMS" subset to the "Gait" subset. The "Gait" subset had significantly higher accuracies and sensitivities compared to the "Subject characteristics + SLMS" and "SLMS" input feature sets.
The maximal classification accuracy of 96.3% [3.7] is high compared to previous literature. Classification accuracies based on principal component analyses of gait knee angles, forces and moments alone[1] or combined with gait spatiotemporal characteristics[52] were 92% and 94%, respectively, but instead of machine learning, linear discriminant analysis was used. In the study by Kirkwood et al. [43] the same approach with gait knee angles alone led to a classification accuracy of 71.8%. Machine learning has only been used to classify patients and healthy controls based on knee forces and moments. Classification accuracies were 82.6% using a Bayes algorithm[22], and 91% using a k-nearest neighbour algorithm[12,42]. Sensitivity and specificity were only reported in the study by Kotti et al. [22]: a sensitivity of 0.77 and a specificity of 0.79 were achieved. Compared to these values, our maximal sensitivity 100.0 [0.0] (median [IQR]) and specificity (100.0 [0.0]) are high, as well as the sensitivity (100.0 [0.0] and 96.3 [4.6] for the "Subject characteristics" and "Subject characteristics + Gait" feature sets respectively) and specificity (96.3 [3.7] and 100.0 [0.0] for the "Subject characteristics" and "Subject characteristics + Gait" feature sets respectively) of our feature sets with highest classification accuracies (96.3% [3.7]).

Since highest classification accuracies, sensitivities and specificities were highest for the "Subject characteristics" input feature set, there is no need for adding gait or SLMS kinematic input features to conventional physical function measures. Adding gait kinematics did not decrease classification performance, but adding SLMS kinematic input features did significantly decrease classification performance. This might have to do with the fact that, although only two PCs of SLMS significantly differed after Bonferrroni correction between patients and controls, input to the machine learner consisted of all significant PCs without Bonferroni correction (Appendix Table E2). Figure 7 indicates that only the first two PC time se-

ries of SLMS have a relative contribution higher than in case of equal data point contribution. This may indicate that the "SLMS" feature set contains much non-relevant information.

Variances of the time series of the principal components of both gait and SLMS turned out to be important discriminative features. When comparing their relative contribution (consisting of 1 data point) with the PC time series input features (consisting of 100 data points), the variance contributions appear to be in the same order of magnitude as the time series contributions but they are achieved by only 1% of the data points. In addition, only the first three (gait) and two (SLMS) PC time series had a relative contribution of more than 5.37% (rounded off, threshold of 100 data points * 0.054%) whereas variance contributions were almost all higher than the threshold. Therefore, in future research it would be informative to evaluate classification performance after removing the PC time series or tuning down their connection weights as described above.
When comparing the information that contributes to classification with the univariate statistical tests, it appears that the machine learner's predictions rely to a large extent to pelvis, hip, knee and ankle angles in the sagittal plane, whereas in the univariate statistical tests only significant differences were found in the hip frontal plane angle and knee sagittal angle. A similar pattern is seen in the SLMS kinematic parameters: the machine learner relies not only on the knee flexion-extension ROM but also on anterior trunk tilt, whereas no significant differences were found in trunk parameters in the univariate statistical approach. Interestingly, the machine learner's relative contributions of cadence is high but the contributions of speed and stride length are not (which appeared to significantly differ between patients and healthy controls). In addition, clinicians' SLMS movement quality rating appeared to be negligible compared to all other features.
These findings emphasize the difference between finding significant differences in variables and their contribution to discrimination in an artificial neural network. Garson's method[62] is a relatively simple way to get insight in importance of a feature for classification which allows for step-wise removal of non-relevant features in an attempt to improve classification performance. In future research, the current algorithm could be improved in this manner.

Since the data set used in the present was only small, not many training cases were present, which was expected to lower the classification accuracy. A method to enlarge the amount of training cases presented to the machine learner during training is to increase the proportion of cases used as the training set. However, the data set appeared too small to show the expected trend. Classification accuracies of the minimal and maximal training set sizes were both significantly lower compared to the standard training set size and classification accuracy of the maximal training set was not significantly higher compared to that of the minimal training set. However, we expect that the expected trend would have been visible in a larger data set.

Despite the small data set, the multivariate machine learning approach resulted in high classification accuracies, sensitivities and specificities. An important advantage of this method is that multimodal data can be handled due to the non-linear nature of the artificial neural network algorithm: both discrete subject characteristics and time series can be included and both continuous measures as well as nominal or ordinal measures can be included. Secondly, interpretation of the results is relatively simple. A third advantage is that in the future, a large set of input features can be defined automatically, followed by the automatic classification according to those features that discriminate[13]. Such a shotgun approach makes the process of feature selection for classification more objective although the definition of a threshold for selecting discriminating features still requires user intervention.
On the other hand, this brings the risk that the machine learner becomes a black box. We support the recommendation of Kaptein et al. [13] to use this machine learning approach in combination with educated guesses, making theory-based decisions. Two other important disadvantages are that large databases to allow for sufficient subjects-to-classes ratio[57] and measures-to-subjects ratio[13], and computers with high computational power are required.

The present study was designed as a first step in classifying physical function in patients with KOA and healthy controls using a machine learning approach. A binary classification was tested distinguishing between patients (assumed to have an impaired physical function) and healthy controls (assumed to have unimpaired physical function) based on conventional physical function measures and movement quality information. The next step would be to classify not according to KL-score (i.e. being patient or not) but in an undefined amount of groups with different physical function levels. This could be achieved by adopting an unsupervised learning approach. That doesn't require a reference classification (i.e. supervision) to train the machine learner with. This could assist in gaining new insights in the concept of physical function and its contributors. The final step would be to develop a machine learning algorithm that assists in screening, diagnosis and monitoring of patients with KOA and healthy controls.

## 5 Conclusions

Our first aim was to highlight advantages and disadvantages of two conventional univariate and multivariate statistical approaches as well as a new, multivariate machine learning approach in evaluating differences in physical function between patients with KOA and healthy controls. The conventional univariate approach of doing multiple independent t-tests between groups turned out to be an easy method with easy interpretation. An important disadvantage is that due to the univariate nature of the tests, covariance between variables is not taken into account. So, information could be redundant and the more t-tests are conducted, the lower the statistical power. Principal component analysis does take into account covariance between variables. Complex information can be gained by interpreting a low amount of outcome variables, but the interpretation of the results is time-consuming. The machine learning approach is able to handle multimodal data in one analysis due to the non-linear nature of the artificial neural network algorithm. Still, interpretation of the results is relatively simple. The approach offers the opportunity of more objective feature selection and extraction but I recommend to use the machine learning approach in combination with a priori knowledge. Major disadvantages are the need for good computers and large databases.
Our second aim was to evaluate the classification performance of the new machine learning approach in classifying physical function in patients with KOA and healthy controls. Maximal classification accuracies (96.3%), sensitivities (100%) and specificities (96.3%) were achieved by an input feature set without gait and SLMS kinematics. Adding gait kinematic input features did not deteriorate performance significantly, but adding SLMS did. This indicates that adding information on movement quality does not assist in classifying physical function in patients with KOA and healthy controls. However, the data presented here should be taken with caution since the validation of the algorithm was not sufficient yet. In future research, the algorithm should be improved so that classification only occurs based on discriminative features. In addition, this study was only a first step in classifying physical function in patients with KOA. In the future, the present study should be repeated with a larger database and an unsupervised learning approach to classify into more than 2 classes and without the need of a reference classification. Such an unsupervised learning approach allows for the design of a classification that classifies physical function in a not predefined, meaningful amount of classes representing different levels of physical function. This might give new insights in the concept of physical function and its contributors. On the long term, this could lead to a machine learning algorithm that helps in screening, diagnosis and monitoring of patients with KOA and healthy controls.

## 6 Acknowledgements

I want to thank prof. dr. ir. Jaap Harlaar and prof.dr. Andreas Daffertshofer for being my supervisors, dr. Eva Weidenhielm-Broström, Josefine Naili Eriksson (PhD) and Mikael Reimeringer for collecting the data and Sanna Aufwerber (PhD) and dr. Maura Daly Iversen for the help with organizing movement quality rating sessions. Of course I also want to thank the clinicians of the Karolinska Institutet and OrthoCenter Stockholm for their participation in the rating sessions.

## 7 References


1. Deluzio, K. J. & Astephen, J. L. Biomechanical features of gait waveform data associated with knee osteoarthritis. An application of principal component analysis. *Gait Posture* (2007).
2. Frank M. Chang, MD, Jason T. Rhodes, MD, MS, Katherine M. Davies, BA, and James J. Carollo, PhD, P. Gait analysis influencs care of children with CP. *http://lermagazine.com/article/gait-analysis-influences-care-of-children-with-cp* (2011).
3. Deluzio, K. J., Wyss, U. P., Costigan, P. A., Sorbie, C. & Zee, B. Gait assessment in unicompartmental knee arthroplasty patients: Principal component modelling of gait waveforms and clinical status. *Hum. Mov. Sci.* (1999).
4. Chau, T. A review of analytical techniques for gait data. Part 1: Fuzzy, statistical and fractal methods. *Gait and Posture* **13**, 49-66 (2001).
5. Schwartz, M. H. & Rozumalski, A. The gait deviation index: A new comprehensive index of gait pathology. *Gait Posture* **28,** 351–357 (2008).
6. Simon, S. R. Quantification of human motion: Gait analysis - Benefits and limitations to its application to clinical problems. *J. Biomech.* **37,** 1869–1880 (2004).
7. Chau, T. A review of analytical techniques for gait data. Part 2: neural network and wavelet methods. *Gait Posture* **13,** 102–120 (2001).
8. Stauffer, R. N., Chao, E. Y. & Györy, A. N. Biomechanical gait analysis of the diseased knee joint. *Clin. Orthop. Relat. Res.* 246–55 (1977).
9. Andriacchi, T. P., Galante, J. O. & Fermier, R. W. The influence of total knee-replacement design on walking and stair-climbing. *J. Bone Joint Surg. Am.* **64,** 1328–35 (1982).
10. Schnitzer, T. J., Popovich, J. M., Andersson, G. B. & Andriacchi, T. P. Effect of piroxicam on gait in patients with osteoarthritis of the knee. *Arthritis Rheum.* **36,** 1207–13 (1993).
11. Carriero, A., Zavatsky, A., Stebbins, J., Theologis, T. & Shefelbine, S. J. Determination of gait patterns in children with spastic diplegic cerebral palsy using principal components. *Gait Posture* **29,** 71–75 (2008).
12. Mezghani, N., Husse, S., Turcot K. & de Guise, J.A. Automatic Classification of Asymptomatic and Osteoarthritis Knee Gait Patterns Using Kinematic Data Features and the Nearest Neighbor Classifier. *IEEE Trans. Biomed. Eng.* **55,** 1230–1232 (2008).
13. Kaptein, R. G., Wezenberg D., IJmker, T., Houdijk, H., Beek, P.J., Lamoth, C.J. & Daffertshofer A.. Shotgun approaches to gait analysis: insights & limitations. *J. Neuroeng. Rehabil.* **11,** 120 (2014).
14. Winter, D.A. Systems approach to the biomechanical assessment of pathological gait. *Klavora Mot. Learn. b-352.* (1980).
15. Chao, E. Y., Laughman, R. K., Schneider, E. & Stauffer, R. N. Normative data of knee joint motion and ground reaction forces in adult level walking. *J. Biomech.* **16,** 219–33 (1983).
16. Whittle, M. W. & Jefferson, R. J. Functional biomechanical assessment of the Oxford Meniscal Knee. *J. Arthroplasty* **4,** 231–43 (1989).
17. Daffertshofer, A., Lamoth, C. J. C., Meijer, O. G. & Beek, P. J. PCA in studying coordination and variability: A tutorial. *Clin. Biomech.* **19,** 415–428 (2004).
18. Lamoth, C. J. C., Daffertshofer, a., Huys, R. & Beek, P. J. Steady and transient coordination structures of walking and running. *Hum. Mov. Sci.* **28,** 371–386 (2009).
19. Lafuente, R., Belda, J. M., Sánchez-Lacuesta, J., Soler, C. & Prat, J. Design and test of neural networks and statistical classifiers in computer-aided movement analysis: a case study on gait analysis. *Clin. Biomech. (Bristol, Avon)* **13,** 216–229 (1998).
20. Dieppe, P. A. & Lohmander, L. S. Pathogenesis and management of pain in osteoarthritis. *Lancet* **365,** 965–73 (2005).
21. Andriacchi, T. P., Favre, J., Erhart-Hledik, J. C. & Chu, C. R. A Systems View of Risk Factors for Knee Osteoarthritis Reveals Insights into the Pathogenesis of the Disease. *Ann. Biomed. Eng.* **43,** 376–387 (2014).
22. Kotti, M., Duffell, L. D., Faisal, A. A. & McGregor, A. H. The Complexity of Human Walking: A Knee Osteoarthritis Study. *PLoS One* **9,** e107325 (2014).
23. Bijlsma, J. W., Berenbaum, F. & Lafeber, F. P. Osteoarthritis: an update with relevance for clinical practice. *Lancet* **377,** 2115–2126 (2011).
24. Fibel, K. H. State-of-the-Art management of knee osteoarthritis. *World J. Clin. Cases* **3**(2), 89-101 (2015).
25. Adatia, A., Rainsford, K. D. & Kean, W. F. Osteoarthritis of the knee and hip. Part I: Aetiology and pathogenesis as a basis for pharmacotherapy. *Journal of Pharmacy and Pharmacology* **64**, 617-615 (2012).
26. Altman, R. D. Classification of disease: Osteoarthritis. *Semin. Arthritis Rheum.* **20,** 40–47 (1991).
27. Broström, E. W., Esbjörnsson, A. C., Von Heideken, J. & Iversen, M. D. Gait deviations in individuals with inflammatory joint diseases and osteoarthritis and the usage of three-dimensional gait analysis. *Best Pract. Res. Clin. Rheumatol.* **26,** 409–422 (2012).
28. Maly, M. R., Costigan, P. A. & Olney, S. J. Determinants of Self-Report Outcome Measures in People With Knee Osteoarthritis. *Arch. Phys. Med. Rehabil.* **87,** 96–104 (2006).
29. Alghadir, A., Anwer, S. & Brismée, J.-M. The reliability and minimal detectable change of Timed Up and Go test in individuals with grade 1 – 3 knee osteoarthritis. *BMC Musculoskelet. Disord.* **16,** 174 (2015).
30. Terwee, C. B., Mokkink, L.B., Steultjens, M.P.M. & Dekker, J.. Self-reported physical functioning was more influenced by pain than performance-based physical functioning in knee-osteoarthritis patients. *J. Clin. Epidemiol.* **45**, 890-892 (2006).
31. Terwee, C. B., Coopmans, C., Peter, W., Roorda, L.D., Poolman R.W., Scholtes, V.A.B., Harlaar J. & de

Vet, H.C.W. Development and Validation of the Computer-Administered Animated Activity Questionnaire to Measure Physical Functioning of Patients With Hip or Knee Osteoarthritis. *Phys. Ther.* **94,** 251–261 (2014).
32. Dobson, F., Hinman, R.S., Hall. M., Terwee, C.B., Roos, E.M. & Bennell, K.L.. Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: A systematic review. *Osteoarthr. Cartil.* **20**, 1548-1562 (2012).
33. Peter, W. F., Loos M, de Vet HC, Boers M, Harlaar J, Roorda LD, Poolman RW, Scholtes VA, Boogaard J, Buitelaar H, Steultjens M, Roos EM, Guillemin F, Rat AC, Benedetti MG, Escobar A, Østerås N & Terwee CB. Development and Preliminary Testing of a Computerized Animated Activity Questionnaire in Patients With Hip and Knee Osteoarthritis. *Arthritis Care Res. (Hoboken).* **67,** 32–39 (2015).
34. Roos, E. M., Roos, H. P., Lohmander, L. S., Ekdahl, C. & Beynnon, B. D. Knee Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure. *J. Orthop. Sports Phys. Ther.* **28,** 88–96 (1998).
35. Bremander, a. B., Dahl, L. L. & Roos, E. M. Validity and reliability of functional performance tests in meniscectomized patients with or without knee osteoarthritis. *Scand. J. Med. Sci. Sport.* **17,** 120-127 (2007).
36. Stratford, P. W., Kennedy, D., Pagura, S. M. C. & Gollish, J. D. The relationship between self-report and performance-related measures: questioning the content validity of timed tests. *Arthritis Rheum.* **49,** 535–40 (2003).
37. F. Dobson, Hinman, R.S., Roos, E.M., Abbott, J.H., Stratford, P., Davis, A.M., Buchbinder, R., Snyder-Mackler, L., Henrotin, Y., Thumboo, J., Hansen, P. & Bennell, K.L. OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. *Osteoarthr. Cartil.* **21,** 1042–1052 (2013).
38. Stevens-Lapsley, J. E., Schenkman, M. L. & Dayton, M. R. Comparison of self-reported knee injury and osteoarthritis outcome score to performance measures in patients after total knee arthroplasty. *PM R* **3,** 541–9; 549 (2011).
39. Gandhi, R., Tsvetkov, D., Davey, J. R., Syed, K. A. & Mahomed, N. N. Relationship between self-reported and performance-based tests in a hip and knee joint replacement population. *Clin. Rheumatol.* **28,** 253–7 (2009).
40. Roos, E. M., Bremander, A. B., Englund, M. & Lohmander, L. S. Change in self-reported outcomes and objective physical function over 7 years in middle-aged subjects with or at high risk of knee osteoarthritis. **67**, 505-510 (2008).
41. Chmielewski, T. L. Hodges, J.M., Horodyski, M., Bishop, M.D., Conrad, B.P. & Tillman, S.M. Investigation of clinician agreement in evaluating movement quality during unilateral lower extremity functional tasks: a comparison of 2 rating methods. *J. Orthop. Sports Phys. Ther.* **37,** 122–9 (2007).
42. Mezghani, N., Turcot K. & de Guise, J.A. Asymptomatic and knee osteoarthritis automatic gait pattern analysis using a wavelet representation of kinetic data and the nearest neighbor classifier. *Age* **63,** 66–67 (2007).
43. Kirkwood, R. N., Resende, R.A., Magalhaes, C.M.B., Gomes, H.A., Mingoti, S.A. & Sampaio, R.F. Application of principal component analysis on gait kinematics in elderly women with knee osteoarthritis. *Rev Bras Fisioter* **15,** 52–8 (2011).
44. Schmitt, D., Vap, A., Queen, R. M. & Krzyzewski, M. W. Effect of end-stage hip, knee, and ankle osteoarthritis on walking mechanics. *Gait Posture* (2015).
45. de Groot, I. B., Favejee, M. M., Reijman, M., Verhaar, J. A. N. & Terwee, C. B. The Dutch version of the Knee Injury and Osteoarthritis Outcome Score: a validation study. 6, 16 *Health Qual. Life Outcomes* (2008).
46. EuroQol. A new facility for the measurement of health-related quality of life. *Health Policy* **16,** 199–208 (1990).
47. Ageberg, E. Bennell, K.L., Hunt, M.A., Simic, M., Roos, E.M & Creaby, M.W. Validity and inter-rater reliability of medio-lateral knee motion observed during a single-limb mini squat. *BMC Musculoskelet. Disord.* **11,** 265 (2010).
48. Vicon. Plug-In Gait. (2002).
49. Weeks, B. K., Carty, C. P. & Horan, S. A. Kinematic predictors of single-leg squat performance: a comparison of experienced physiotherapists and student physiotherapists. *BMC Musculoskelet. Disord.* **13,** 207 (2012).
50. Kaptein R, D. A. UPMOVE: a Matlab Toolbox for the Analysis and Classification of Human Gait. *[http://www.upmove.org]*
51. Field A. *Discovering statistics using SPSS*. (SAGE, 2009).
52. Astephen, J. L. & Deluzio, K. J. A multivariate gait data analysis technique: application to knee osteoarthritis. Proceedings of the Institution of Mechanical Engineers. Part H, *Journal of engineering in medicine* (2004).
53. Joliffe, I. in *Principal Component Analysis (2nd edition)* 11–114 (2002).
54. Peres-Neto, P. R., Jackson, D. A. & Somers, K. . Giving meaningful interpretation to ordination axes: assessing loading signficance in principal component analysis. *Ecology* **84,** 2347–2363 (2003).
55. Welch, P. The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.* **AU-15,** 70–73 (1967).
56. Arsham H. Kuiper's P-value as a measuring tool and decision procedure for the goodness-of-fit test. *J. Appl. Stat.* **15,** 131–137 (1988).
57. Vincent, W. J. *Statistics in kinesiology*. (Human Kinetics, 2005).
58. Sahiner, B., Chan, H.-P. & Hadjiiski, L. Classifier performance prediction for computer-aided diagnosis

using a limited dataset. *Med. Phys.* **35,** 1559–1570 (2008).
59. Kohonen, T. in *The Handbook of Brain Theory and Neural Networks* 631–635 (2003).
60. WEKA classification algorithms. http://wekaclassalgos.sourceforge.net/. Last accessed 26-10-2016.
61. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **7,** 1–30 (2006).
62. Garson, G. D. Interpreting neural-network connection weights. *AI Expert* **6,** 47–51 (1991).
63. Goh, A. T. C. Back-propagation neural networks for modeling complex systems. *Artif. Intell. Eng.* **9,** 143–151 (1995).
64. Astephen, J. L., Deluzio, K. J., Caldwell, G. E. & Dunbar, M. J. Biomechanical changes at the hip, knee, and ankle joints during gait are associated with knee osteoarthritis severity. *J. Orthop. Res.* **26,** 332–341 (2008).
65. Kaufman, K. R., Hughes, C., Morrey, B. F., Morrey, M. & An, K. N. Gait characteristics of patients with knee osteoarthritis. *J. Biomech.* **34,** 907–915 (2001).
66. Weidow, J., Tranberg, R., Saari, T. & Kärrholm, J. Hip and knee joint rotations differ between patients with medial and lateral knee osteoarthritis: gait analysis of 30 patients and 15 controls. *J. Orthop. Res.* **24,** 1890–9 (2006).
67. Gök, H., Ergin, S. & Yavuzer, G. Kinetic and kinematic characteristics of gait in patients with medial knee arthrosis. *Acta Orthop. Scand.* **73,** 647–52 (2002).
68. Hunt, M. A., Wrigley, T. V., Hinman, R. S. & Bennell, K. L. Individuals with severe knee osteoarthritis (OA) exhibit altered proximal walking mechanics compared with individuals with less severe OA and those without knee pain. *Arthritis Care Res. (Hoboken).* **62,** 1426–1432 (2010).
69. Mills, K., Hunt, M. A. & Ferber, R. Biomechanical deviations during level walking associated with knee osteoarthritis: A systematic review and meta-analysis. *Arthritis Care Res.* (2013).
70. Huang, S.-C. *et al.* Effects of severity of degeneration on gait patterns in patients with medial knee osteoarthritis. *Med. Eng. Phys.* **30,** 997–1003 (2008).
71. Baliunas, A. J. *et al.* Increased knee joint loads during walking are present in subjects with knee osteoarthritis. *Osteoarthritis Cartilage* **10,** 573–9 (2002).
72. Al-Zahrani KS, B. A. A study of the gait characteristics of patients with chronic osteoarthritis of the knee. *Disabil Rehabil.* Mar 20;24(5):275-80. (2002).
73. Naili, J. E., Gutierrez-Farewik, E. M., Reimeringer, M., Esbjornsson, A.-C. & Brostrom, E. W. 3D evaluation of the single limb mini squat test in patients with knee osteoarthritis . *Gait Posture* **Conference,** S6–S7 (2013).
74. Bejek, Z., Paróczai, R., Illyés, A. & Kiss, R. M. The influence of walking speed on gait parameters in healthy people and in patients with osteoarthritis. *Knee Surg. Sports Traumatol. Arthrosc.* **14,** 612–22 (2006).

## Appendix A. 3D movement analysis data

| Table A1. Angle and CoM time series used in kinematic analyses | | | |
|---|---|---|---|
| Segment/joint | Time series | Gait | SLMS |
| Trunk | Posterior tilt | X | X |
| | Contralateral tilt | X | X |
| | Internal rotation | X | X |
| Pelvis | Anterior tilt | X | X |
| | Ipsilateral upward list | X | X |
| | Internal rotation | X | X |
| Hip & CL hip | Flexion | X | X |
| | Adduction | X | X |
| | Internal rotation | X | X |
| Knee & CL knee | Flexion | X | X |
| | Varus | X | X |
| | Internal rotation | X | X |
| Ankle & CL ankle | Dorsiflexion | X | X |
| | Internal rotation | X | X |
| | Foot progression angle: toe in | X | |
| CoM position relative to toe marker of squat leg | Mediolateral CoM position | | X |
| | Backward-forward CoM position | | X |
| | Upward-downward CoM position | | X |
| Number of time series | | 24 | 25 |

| Table A2. Educated guess for gait and SLMS spatiotemporal characteristics and kinematic parameters discriminating between patients and controls | | | |
|---|---|---|---|
| Gait spatiotemporal characteristics | | Gait kinematic parameters | |
| Speed (m/s) | $<^{1,27,64–67}$ | Peak trunk ipsilateral lean during GC (º) | $>^{68}$ |
| Cadence (steps/min) | $<^{67,69}$ | Peak trunk contralateral lean during GC (º) | $>^{68}$ |
| Stride time (s) | $>^{64,67}$ | Peak pelvic anterior tilt during GC, at start and end SS (º) | $>^{70}$ |
| Stride length (m) | $<^{64,66,67,71}$ | Peak pelvic drop during swing (º) | $<^{70}$ |
| | | Peak hip flexion during stance (º) | $>^{44}$ |
| SLMS spatiotemporal characteristics | | Peak hip extension during stance (º) | $<^{44,72}$ |
| Decreased number of repetitions in the affected leg (PB test score) | $<^{73}$ | Hip flexion-extension ROM during GC (º) | $<^{64}$ |
| Limb symmetry index (= repetitions affected leg / healthy leg) | $<^{73}$ | Peak hip adduction during stance (º) | $<^{68,70}$ |
| Test duration (s) | < | Hip adduction at the start of SS (º) | $<^{70}$ |
| | | Peak knee flexion during GC (º) | $<^{64,65,72}$ |
| SLMS kinematic parameters | | Knee flexion at the start of SS (º) | $<^{70}$ |
| SLMS Peak & mean anterior trunk tilt (º) | $>^{73}$ | Peak knee flexion during LR (º) | $<^{70,72}$ |
| SLMS Knee flexion-extension ROM (º) | $<^{73}$ | Peak knee extension during GC (º) | $<^{74}$ |
| | | Peak knee extension during stance (º) | $<^{44}$ |
| | | Knee flexion-extension ROM during GC (º) | $<^{65,74}$ |
| | | Peak ankle plantarflexion during stance (º) | $<^{44}$ |
| | | Peak ankle dorsiflexion during swing (º) | $<^{72}$ |
| | | Ankle plantar-dorsiflexion ROM over GC (º) | $<^{64}$ |
| > Patients' values are increased relative to controls' values. < Patients' values are reduced relative to controls' values. All parameters refer to the affected side. GC = gait cycle; ROM = range of motion; SS = single stance; LR = loading response | | | |

## Appendix B. Machine learning input features

**Table B1. Input feature subsets used in the machine learning approach**

| Subject characteristics | Gait | SLMS |
|---|---|---|
| *Basic personal information* | *Spatiotemporal characteristics* (Table 2) | *Spatiotemporal characteristics* (Table 2) |
| Sex | Speed | LimbSymmIndex |
| Age | Cadence | TestDuration |
| Weight | Stride time | |
| Height | Stridel ength | |
| BMI | | |
| | *Discrete kinematic parameters* (Table 2) | *Discrete kinematic parameters* (Table 2) |
| *KL (only included if stated explicitly)* | Peak trunk ipsilateral and contralateral lean during GC (2) | Peak trunk anterior tilt and mean trunk anterior tilt (2) |
| | Peak pelvic anterior tilt during GC, at start and end SS, peak pelvic drop during swing (4) | Knee flexion-extension ROM (1) |
| *PROMs* | Peak hip flexion and extension during stance, hip flexion-extension ROM during GC, peak hip adduction during stance and hip adduction at start of SS (5) | |
| KOOS 5scales | Peak knee flexion during GC, at start SS, during LR, peak knee extension during GC and stance, knee flexion-extension ROM during GC (6) | |
| EQ5D score+VAS | Peak ankle plantarflexion during stance, peak ankle dorsiflexion during swing, ankle plantar-dorsiflexion ROM over GC (3) | |
| | | |
| *PB test score* | *Kinematics from $PCA_{Gait}$* | *Kinematics from $PCA_{SLMS}$* |
| *ClinRating* | Time projections (ξ) on selected PCs | Time projections (ξ) on selected PCs |
| | Variances of the time projections of the selected PCs ($\sigma_\xi^2$) | Variances of the time projections of the selected PCs ($\sigma_\xi^2$) |

## Appendix C. Results of univariate statistical approach

| Table C1. Subject characteristics, PROM scores, PB test score and clinicians' ratings | | | | |
|---|---|---|---|---|
| | **Patients (N=31)** | **Controls (N=24)** | | |
| | **Mean (SE)** | **Mean (SE)** | **p** | **r** |
| **Basic personal information** | | | | |
| Gender (number of males/females) | 20x F; 11x M | 16x F; 8x M | 0.87 | -0.02 |
| Age (years) | 66.0 (1.3) | 65.2 (1.9) | 0.70 | 0.05 |
| Height (m) [a] | 1.7 [0.2] | 1.7 [0.1] | 0.48 | -0.10 |
| **Weight** (kg) | 85.4 (2.2) | 72.7 (2.5) | **p < 0.05*** | 0.46 |
| **BMI** (kg/m$^2$) | 29.9 (0.7) | 24.9 (0.6) | **p < 0.05*** | 0.56 |
| **KL score** (scale 1-4, 0 indicates no KL score) | 2x 3a; 5x 3b 9x 4a; 4x 4b | 24x 0 | **p < 0.05*** | -0.90 |
| | | | | |
| **PROM scores** | | | | |
| **KOOSpain** (scale 0-100) [a] | 44.4 [19.4] | 100.0 [5.6] | **p < 0.05*** | -0.86 |
| **KOOSsympt** (scale 0-100) [a] | 35.7 [38.4] | 96.4 [10.7] | **p < 0.05*** | -0.84 |
| **KOOSadl** (scale 0-100) [a] | 55.1 [19.5] | 100.0 [6.6] | **p < 0.05*** | -0.86 |
| **KOOSsport** (scale 0-100) [a] | 10.0 [28.8] | 100.0 [22.5] | **p < 0.05*** | -0.84 |
| **KOOSqol** (scale 0-100) [a] | 31.3 [17.2] | 100.0 [18.8] | **p < 0.05*** | -0.86 |
| **EQ5Dscore** (scale -0.594 – 1) [a] | 0.7 [0.1] | 1.0 [0.2] | **p < 0.05*** | -0.85 |
| **EQ5Dvas**[b] (scale 0-100) [a] | 71.5 [32.5] | 89.5 [17.3] | **p < 0.05*** | -0.54 |
| | | | | |
| **PBtest** (amount of knee bendings in 30s) | 13.9 (1.4) | 28.7 (2.1) | **p < 0.05*** | 0.64 |
| **MeanCLINrating**[c] **(transformed)** (scale 1-4) [a] | 3.0 [1.0] | 2.0 [0.0] | **p < 0.05*** | -0.67 |

[a] The non-parametric counterpart of the independent t-test (Mann-Whitney U test) is used. Instead of mean (SE), median [IQR] is reported.
[b] indicates four missing values for patients, resulting in $N_{patients}$ = 27 and $N_{controls}$ = 24.
[c] indicates one missing value for controls, resulting in $N_{patients}$ = 31 and $N_{controls}$ = 23.
***** indicates significance after Bonferroni correction: p < (0.05/amount of tests), with total amount of tests = 44.
All scales are presented from minimal (affected) to optimal (healthy) score.

| Table C2. Gait and SLMS spatiotemporal characteristics and kinematic parameters | | | | |
|---|---|---|---|---|
| | **Patients (N=31)** | **Controls (N=24)** | | |
| | **Mean (SE)** | **Mean (SE)** | **p** | **r** |
| **Gait spatiotemporal characteristics** | | | | |
| **Speed** (m/s) | 1.10 (0.03) | 1.32 (0.04) | **p < 0.05*** | 0.52 |
| Cadence (steps/min) | 111.04 (1.30) | 117.74 (2.12) | p < 0.05 | 0.36 |
| Stride time (s) [a] | 1.07 [0.09] | 1.01 [0.11] | p < 0.05 | -0.39 |
| **Stride length** (m) | 1.19 (0.03) | 1.34 (0.02) | **p < 0.05*** | 0.49 |
| | | | | |
| **Gait kinematics (affected side)** | | | | |
| Peak trunk ipsilateral lean (º) | 2.72 (0.40) | 2.11 (0.44) | 0.31 | 0.14 |
| Peak trunk contralateral lean (º) | 1.93 (0.25) | 1.66 (0.33) | 0.50 | 0.09 |
| Peak pelvic anterior tilt during GC (º) | 12.07 (0.70) | 12.54 (0.85) | 0.67 | 0.06 |
| Pelvic anterior tilt at start SS (º) | 10.05 (0.69) | 10.61 (0.90) | 0.62 | 0.07 |
| Pelvic anterior tilt at end SS (º)[a] | 10.96 [7.40] | 11.44 [5.27] | 0.91 | -0.02 |
| Peak pelvic drop during swing (º) | 3.17 (0.35) | 4.06 (0.41) | 0.10 | 0.22 |
| Peak hip flexion during stance (º) | 30.90 (1.19) | 33.15 (1.02) | 0.16 | 0.19 |
| Peak hip extension during stance (º) | 9.96 (0.95) | 11.28 (1.35) | 0.41 | 0.11 |
| Hip flexion-extension ROM during GC (º) | 41.74 (0.73) | 45.04 (1.04) | p < 0.05 | 0.34 |
| **Peak hip adduction during stance** (º) | 0.10 (1.16) | 4.41 (0.73) | **p < 0.05*** | 0.37 |
| Hip adduction at the start of SS (º) | -1.52 (1.15) | 3.76 (0.80) | p < 0.05 | 0.44 |
| Peak knee flexion during GC (º) [a] | 49.20 [9.16] | 54.02 [7.79] | p < 0.05 | -0.36 |
| Knee flexion at the start of SS (º) | 4.64 (1.22) | 2.68 (0.84) | 0.19 | 0.18 |
| Peak knee flexion during LR (º) | 12.45 (1.49) | 15.92 (0.97) | 0.23 | -0.16 |
| **Peak knee extension during GC** (º) | -1.65 (1.19) | 3.51 (0.86) | **p < 0.05*** | 0.44 |
| Peak knee extension during stance (º) | -1.91 (1.20) | 2.86 (0.95) | p < 0.05 | 0.40 |
| **Knee flexion-extension ROM during GC** (º) | 47.88 (1.17) | 57.97 (1.05) | **p < 0.05*** | 0.65 |
| Peak ankle plantarflexion during stance (º) | 6.34 (0.85) | 9.77 (0.73) | p < 0.05 | 0.38 |
| Peak ankle dorsiflexion during swing (º) | 6.23 (0.56) | 5.19 (0.34) | 0.12 | 0.23 |
| Ankle plantar-dorsiflexion ROM over GC (º) | 24.80 (0.99) | 24.85 (1.02) | 0.97 | 0.01 |
| | | | | |
| **SLMS spatiotemporal characteristics** | | | | |
| **LimbSymmIndex** (affected / healthy leg repetition ratio[73]) [a] | 0.7 [0.5] | 1.0 [0.1] | **p < 0.05*** | -0.62 |
| TestDuration (s) [a] | 30.0 [0.0] | 30.0 [0.0] | 0.07 | -0.24 |
| | | | | |
| **SLMS kinematics** | | | | |
| SLMS Peak anterior trunk tilt (º) | 19.61 (1.91) | 9.46 (1.72) | p < 0.05 | 0.38 |
| SLMS Mean anterior trunk tilt (º) | 10.72 (1.46) | 4.44 (1.48) | p < 0.05 | 0.38 |
| **SLMS Knee flexion-extension ROM** (º) | 33.20 (1.52) | 54.01 (1.78) | **p < 0.05*** | 0.77 |

[a] The non-parametric counterpart of the independent t-test (Mann-Whitney U test) is used. Instead of mean (SE), median [IQR] is reported.
[b] indicates four missing values for patients, resulting in $N_{patients}$ = 27 and $N_{controls}$ = 24.
[c] indicates one missing value for controls, resulting in $N_{patients}$ = 31 and $N_{controls}$ = 23.
***** indicates significance after Bonferroni correction: p < (0.05/amount of tests), with amount of tests = 44.

## Appendix D. $PCA_{Gait}$

**Table D1. Percentage of the variance explained by the eigenvalues of the first five principal components for gait and additional eigenvalues of significantly different PCs**

| Eigenvalue λ | $PCA_{Gait}$ (%) |
|---|---|
| $\lambda_1$ | 38.82* |
| $\lambda_2$ | 26.60* |
| $\lambda_3$ | 18.47* |
| $\lambda_4$ | 3.91 |
| $\lambda_5$ | 3.08 |
| **$\Sigma\lambda_{1-5}$** | **90.88** |
| | |
| $\Sigma\lambda_{allsignPCs}$ | 83.89 |

* indicates significant difference in variance of cycle time projection on each PC between patients and controls

**Table D2. Significant differences in variance of gait cycle time projection on each $PC_{gait}$ between patients and controls**

| | $\sigma_\xi^2$ | | | |
|---|---|---|---|---|
| | **Patients** | **Controls** | | |
| PC | mean (SE) | mean (SE) | p | r |
| **1** | 472.9 (17.01) | 566.7 (24.03) | **p < 0.05*** | 0.39 |
| **2** | 321.9 (12.36) | 391.8 (14.24) | **p < 0.05*** | 0.42 |
| **3** | 203.8 (7.83) | 303.9 (12.75) | **p < 0.05*** | 0.67 |
| 4 | 44.6 (3.26) | 61.9 (4.80) | p < 0.05 | 0.37 |
| 5[a] | 33.9 [21.2] | 44.3 [14.1] | 0.052 | 0.42 |
| 7 | 14.6 (1.05) | 20.8 (2.25) | p < 0.05 | 0.40 |
| 11[a] | 6.9 [6.8] | 5.3 [4.2] | p < 0.05 | 0.19 |

[a] The non-parametric counterpart of the independent t-test (Mann-Whitney U test) is used. Instead of mean (SE), median [IQR] is reported.
***** indicates significance after Bonferroni correction: p < (0.05/amount of tests), with amount of tests = 24.

**Table D3. Average relative phase (°) between PCs of gait with similar frequencies.**

| Gait | | | |
|---|---|---|---|
| | f (/GC) | Mean | SD |
| PC1-PC2 | 1 | 92.6 | 8.9 |
| PC3-PC4 | 2 | -91.4 | 11.7 |

Relative phases are calculated as the phase of the PC with the highest number minus the phase of the PC with the lowest number.
* indicates a significant difference in average relative phase between patients and controls (p < 0.01).

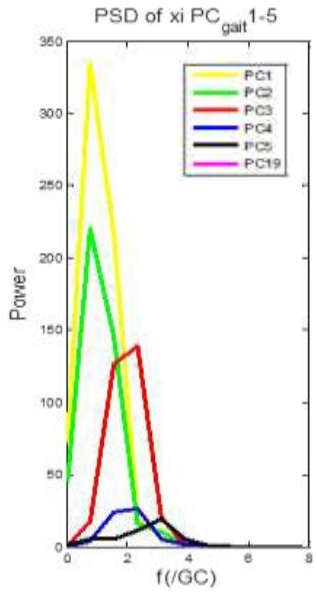



**Figure D1. Power spectral densities for time projections on PC1-5 in $PCA_{Gait}$.**

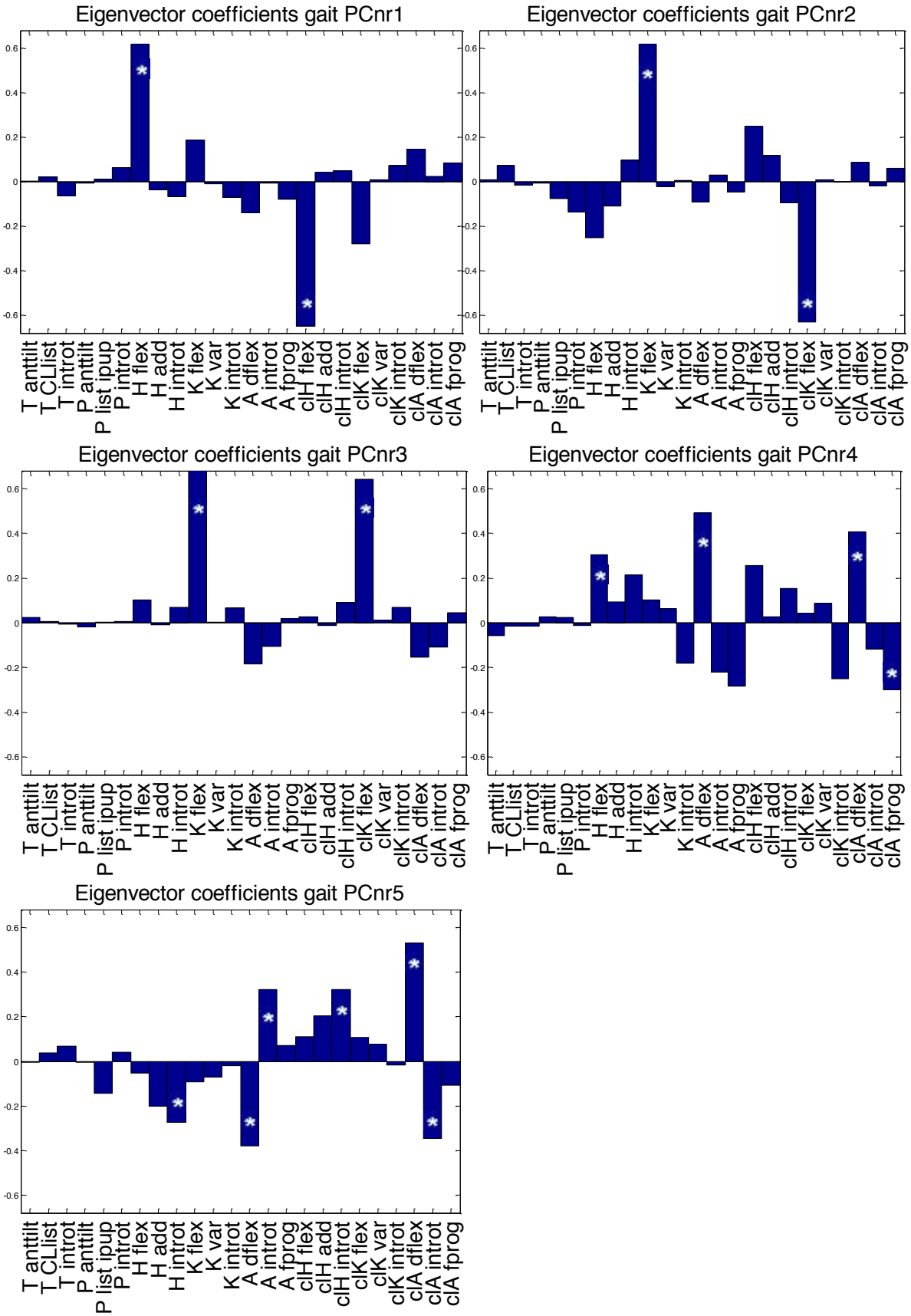



**Figure D2. PCA$_{Gait}$: Eigenvector coefficients of principal component 1-5**. Eigenvector coefficients quantify the contribution of each original time series to that principal component. Stars indicate significance according to the broken stick test. Angle abbreviations on the horizontal axis: T = trunk, P = pelvis, H = hip, K = knee, A = ankle, cl = contralateral, anttilt = anterior tilt, ip up = ipsilateral up, CLlist = contralateral list, introt = internal rotation, flex = flexion, var = varus, dflex = dorsiflexion, fprog = foot progression.

## Appendix E. $PCA_{SLMS}$

**Table E1. Percentage of the variance explained by the eigenvalues of the first five principal components for SLMS and additional eigenvalues of significantly different PCs**

| Eigenvalue λ | $PCA_{SLMS}$ (%) |
|---|---|
| $\lambda_1$ | 54.97 |
| $\lambda_2$ | 21.99* |
| $\lambda_3$ | 8.51 |
| $\lambda_4$ | 3.71 |
| $\lambda_5$ | 3.48 |
| **$\Sigma\lambda_{1-5}$** | **92.66** |
| | |
| $\lambda_{19}$ | 0.11* |
| $\Sigma\lambda_{allsignPCs}$ | 22.10 |

* indicates significant difference in variance of cycle time projection on each PC between patients and controls

**Table E2. Significant differences in variance of SLMS test time projection on each $PC_{SLMS}$ between patients and controls**

| | $\sigma_\xi^2$ | | | |
|---|---|---|---|---|
| | **Patients** | **Controls** | | |
| PC | Median (IQR)[a] | Median (IQR)[a] | p | r |
| 1 | 452.1 [567.8] | 659.4 [772.0] | 0.204 | 0.15 |
| **2** | 72.3 [99.7] | 530.8 [380.7] | **p < 0.05*** | 0.33 |
| 3 | 66.0 [60.9] | 60.1 [60.2] | 0.872 | 0.05 |
| 4 | 24.2 [33.7] | 36.6 [31.7] | 0.139 | 0.16 |
| 5 | 43.4 [29.7] | 33.5 [29.2] | 0.06 | 0.18 |
| 6 | 26.3 [27.6] | 12.1 [13.7] | p < 0.05 | 0.22 |
| 8 | 4.8 [7.6] | 8.4 [8.1] | p < 0.05 | 0.19 |
| 11 | 3.9 [3.7] | 5.6 [3.5] | p < 0.05 | 0.19 |
| 12 | 3.3 [3.3] | 5.1 [3.4] | p < 0.05 | 0.21 |
| **19** | 0.9 [0.5] | 1.3 [1.4] | **p < 0.05*** | 0.24 |
| 23 | 0.3 [0.5] | 0.5 [0.5] | p < 0.05 | 0.03 |

[a] For SLMS data, all tests indicating a significant difference in $\sigma_\xi^2$ were non-parametric.
* indicates significance after Bonferroni correction: p < (0.05/amount of tests), with amount of tests = 25.

**Table E3. Average relative phase between PCs of SLMS with similar frequencies.**

| **SLMS** | | | | | | |
|---|---|---|---|---|---|---|
| | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC19** |
| **PC1** | | 92.4 (28.3) | -13.2 (71.9)[a] | 44.6 (64.3) | 25.2 (70.2) | 17.0 (77.4) |
| **PC2** | | | -16.7 (73.9) | 11.0 (59.8)[b] | 1.7 (77.6) | -20.7 (69.4)[c] |
| **PC3** | | | | 21.3 (73.7) | -8.9 (79.2) | -6.8 (78.8) |
| **PC4** | | | | | 23.5 (76.7) | -28.6 (75.0) |
| **PC5** | | | | | | -28.6 (75.0) |

Relative phases are calculated as the phase of the PC with the highest number minus the phase of the PC with the lowest number. Reported values represent mean (SD).
[a] A significant difference existed in average relative phase between patients (16.8 (63.3)) and controls (-51.5 (52.8)), p < 0.01.
[b] A significant difference existed in average relative phase between patients (29.5 (54.5)) and controls(-12.8 (60.9)), p < 0.01.
[c] A significant difference existed in average relative phase between patients (17.7 (70.6)) and controls (-69.9 (50.1)), p < 0.01.

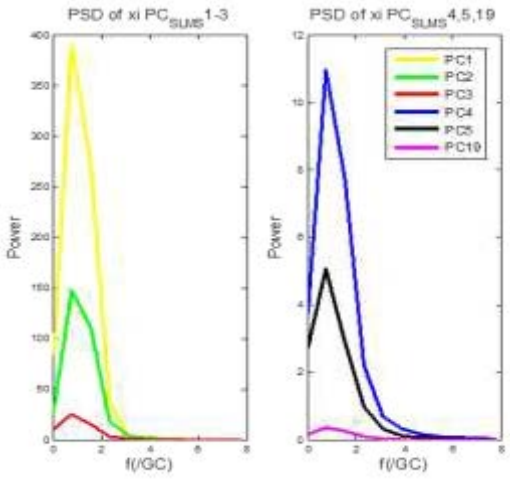



**Figure E1. Power spectral densities for time projections on PC1-3 (left panel) and PC4,5 and 19 in $PCA_{SLMS}$ (right panel).**

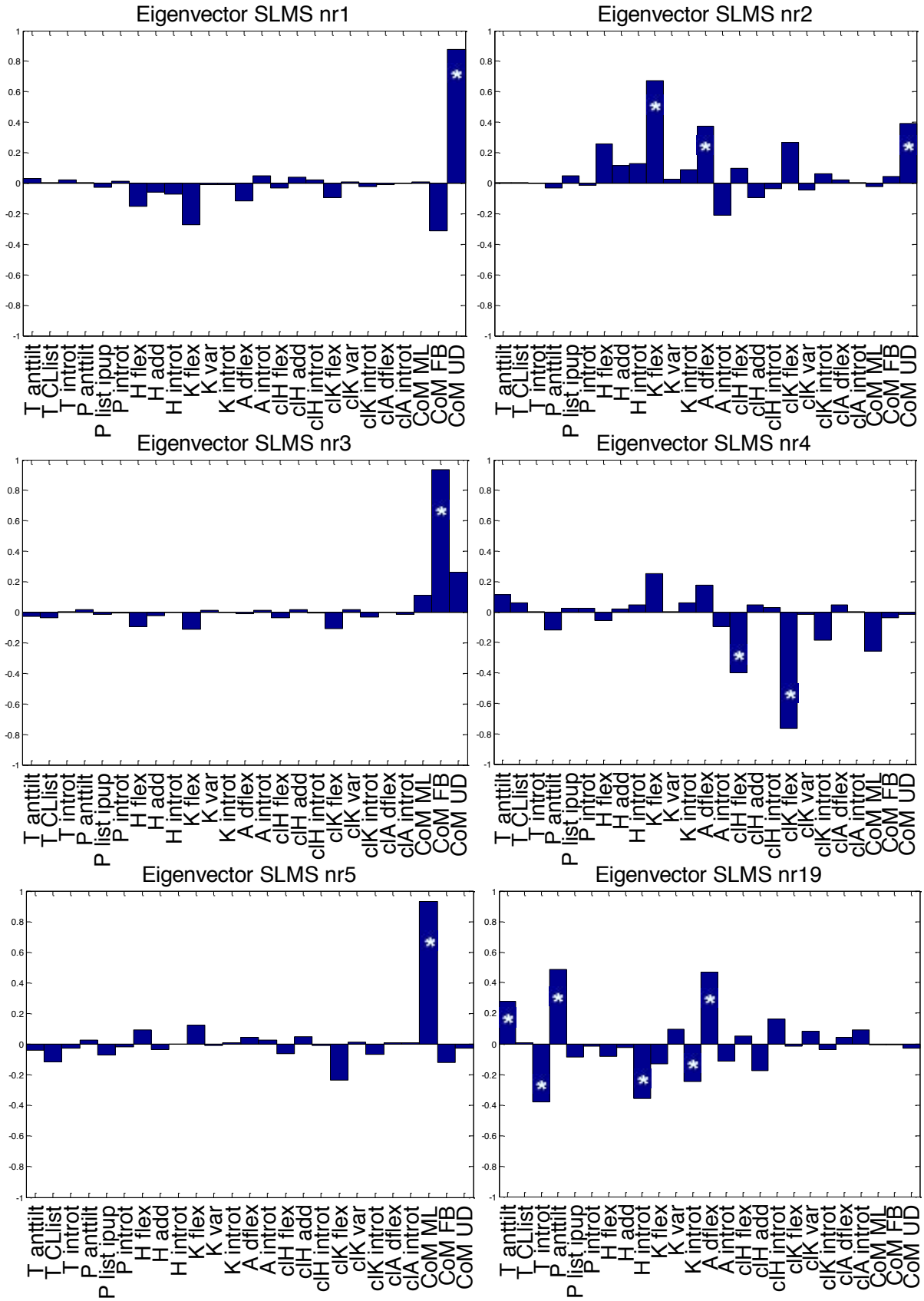



**Figure E2. PCA$_{SLMS}$: Eigenvector coefficients of principal component 1-5 & 19.** Eigenvector coefficients quantify the contribution of each original time series to that principal component. Stars indicate significance according to the broken stick test. Angle abbreviations on the horizontal axis: T = trunk, P = pelvis, H = hip, K = knee, A = ankle, CoM = center of mass, cl = contralateral, anttilt = anterior tilt, ip up = ipsilateral up, CList = contralateral list, introt = internal rotation, flex = flexion, var = varus, dflex = dorsiflexion, fprog = foot progression, ML = mediolateral, FB = forward-backward, UD = upwards-downwards.

## Appendix F. Classification accuracies, specificities and sensitivities

**Table F1. Classification accuracies (%) of 8 different input feature sets and results of post-hoc tests after Friedman's ANOVA**

| Input feature set | Median [IQR][a] | | Median [IQR][a] | p | r |
|---|---|---|---|---|---|
| With KL score | | | | | |
| **Only KL** | 100.0 [0.0] | **All with KL** | 81.5 [18.5] | **p<0.05*** | |
| All with KL | 81.5 [18.5] | | | | |
| Without KL score | | | | | |
| **All without KL** | 81.5 [14.8] | **Only subchar** | 96.3 [3.7] | **p<0.05*** | -0.61 |
| | | Only gait | 81.5 [7.4] | 0.125 | -0.15 |
| | | Only SLMS | 74.1 [18.5] | 0.054 | -0.19 |
| | | **Subchar+gait** | 96.3 [3.7] | **p<0.05*** | -0.58 |
| | | Subchar+SLMS | 74.1 [11.1] | p<0.05 | -0.26 |
| | | Gait+SLMS | 81.5 [12.0] | 0.935 | -0.01 |
| **Only subchar** | 96.3 [3.7] | **Only gait** | 81.5 [7.4] | **p<0.05*** | -0.62 |
| | | **Only SLMS** | 74.1 [18.5] | **p<0.05*** | -0.60 |
| | | Subchar+gait | 96.3 [3.7] | p<0.05 | -0.29 |
| | | **Subchar+SLMS** | 74.1 [11.1] | **p<0.05*** | -0.62 |
| | | **Gait+SLMS** | 81.5 [12.0] | **p<0.05*** | -0.60 |
| **Only gait** | 81.5 [7.4] | **Only SLMS** | 74.1 [18.5] | **p<0.05*** | -0.40 |
| | | **Subchar+gait** | 96.3 [3.7] | **p<0.05*** | -0.60 |
| | | **Subchar+SLMS** | 74.1 [11.1] | **p<0.05*** | -0.43 |
| | | Gait+SLMS | 81.5 [12.0] | 0.074 | -0.18 |
| **Only SLMS** | 74.1 [18.5] | **Subchar+gait** | 96.3 [3.7] | **p<0.05*** | -0.61 |
| | | Subchar+SLMS | 74.1 [11.1] | 0.743 | -0.03 |
| | | Gait+SLMS | 81.5 [12.0] | p<0.05 | -0.25 |
| **Subchar+gait** | 96.3 [3.7] | **Subchar+SLMS** | 74.1 [11.1] | **p<0.05*** | -0.61 |
| | | **Gait+SLMS** | 81.5 [12.0] | **p<0.05*** | -0.60 |
| Subchar+SLMS | 74.1 [11.1] | Gait+SLMS | 81.5 [12.0] | p<0.05 | -0.25 |
| Gait + SLMS | 81.5 [12.0] | | | | |

[a] Since the Friedman ANOVA is a non-parametric test, medians and interquartile range are reported.
***** indicates significance after Bonferroni correction: p < (0.05/amount of tests), with amount of tests = 22.

**Table F2. Classification accuracies (%) of the full input feature set (without KL score) with minimal, standard and maximal training set sizes and results of post-hoc tests after Friedman's ANOVA**

| Training set | Training set proportion | Median [IQR][a] | Training set for compari-son | Median [IQR][a] | p | r |
|---|---|---|---|---|---|---|
| **Standard** | 50% | 81.5 [14.8] | **Minimal** | 63.2 [17.5] | **p < 0.05*** | -0.51 |
| | | | **Maximal** | 50.0 [50.0] | **p < 0.05*** | -0.30 |
| Minimal | 4.2% | 63.2 [17.5] | Maximal | 50.0 [50.0] | 0.809 | -0.02 |
| Maximal | 95.8% | 50.0 [50.0] | | | | |

[a] Since the Friedman ANOVA is a non-parametric test, medians and interquartile range are reported.
***** indicates significance after Bonferroni correction: p < (0.05/amount of tests), with amount of tests = 3.

| **Table F3. Percentage of cases incorrectly classified as patients (sensitivities) for 7 different input feature sets and results of post-hoc tests after Friedman's ANOVA** | | | | | |
|---|---|---|---|---|---|
| Input feature set | Median [IQR][a] | | Median [IQR][a] | p | r |
| With KL score | | | | | |
| All with KL | 88.9 [14.8] | All without KL | 83.3 [8.3] | .083 | -0.17 |
| Without KL score | | | | | |
| **All without KL** | 83.3 [8.3] | **Only subchar** | 100.0 [0.0] | **p<0.05*** | -0.59 |
| | | Only gait | 88.9 [7.4] | p<0.05 | -0.29 |
| | | Only SLMS | 81.5 [11.1] | p<0.05 | -0.21 |
| | | **Subchar+gait** | 96.3 [4.6] | **p<0.05*** | -0.52 |
| | | Subchar+SLMS | 79.6 [8.3] | p<0.05 | -0.25 |
| | | Gait+SLMS | 85.2 [11.1] | 0.644 | -0.05 |
| **Only subchar** | 100.0 [0.0] | **Only gait** | 88.9 [7.4] | **p<0.05*** | -0.61 |
| | | **Only SLMS** | 81.5 [11.1] | **p<0.05*** | -0.61 |
| | | **Subchar+gait** | 96.3 [4.6] | **p<0.05*** | -0.51 |
| | | **Subchar+SLMS** | 79.6 [8.3] | **p<0.05*** | -0.62 |
| | | **Gait+SLMS** | 85.2 [11.1] | **p<0.05*** | -0.61 |
| **Only gait** | 88.9 [7.4] | **Only SLMS** | 81.5 [11.1] | **p<0.05*** | -0.49 |
| | | **Subchar+gait** | 96.3 [4.6] | **p<0.05*** | -0.54 |
| | | **Subchar+SLMS** | 79.6 [8.3] | **p<0.05*** | -0.53 |
| | | **Gait+SLMS** | 85.2 [11.1] | **p<0.05*** | -0.37 |
| **Only SLMS** | 81.5 [11.1] | **Subchar+gait** | 96.3 [4.6] | **p<0.05*** | -0.60 |
| | | Subchar+SLMS | 79.6 [8.3] | 0.847 | -0.02 |
| | | Gait+SLMS | 85.2 [11.1] | p<0.05 | -0.22 |
| **Subchar+gait** | 96.3 [4.6] | **Subchar+SLMS** | 79.6 [8.3] | **p<0.05*** | -0.61 |
| | | **Gait+SLMS** | 85.2 [11.1] | **p<0.05*** | -0.59 |
| Subchar+SLMS | 79.6 [8.3] | Gait+SLMS | 85.2 [11.1] | p<0.05 | -0.23 |
| Gait + SLMS | 85.2 [11.1] | | | | |

[a] Since the Friedman ANOVA is a non-parametric test, medians and interquartile range are reported.
***** indicates significance after Bonferroni correction: p < (0.05/amount of tests), with amount of tests = 22.

| **Table F4. Percentage of cases incorrectly classified as controls (specificities) for 7 different input feature sets and results of post-hoc tests after Friedman's ANOVA** | | | | | |
|---|---|---|---|---|---|
| Input feature set | Median [IQR][a] | | Median [IQR][a] | p | r |
| With KL score | | | | | |
| All with KL | 96.3 [11.1] | All without KL | 94.4 [11.1] | 0.61 | -0.05 |
| Without KL score | | | | | |
| **All without KL** | 94.4 [11.1] | **Only subchar** | 96.3 [3.7] | **p < 0.05*** | -0.35 |
| | | Only gait | 92.6 [7.4] | 0.19 | -0.13 |
| | | Only SLMS | 92.6 [4.6] | 0.72 | -0.04 |
| | | **Subchar+gait** | 100.0 [0.0] | **p < 0.05*** | -0.45 |
| | | Subchar+SLMS | 92.6 [7.4] | 0.49 | -0.07 |
| | | Gait+SLMS | 96.3 [8.3] | 0.48 | -0.07 |
| **Only subchar** | 96.3 [3.7] | **Only gait** | 92.6 [7.4] | **p < 0.05*** | -0.43 |
| | | **Only SLMS** | 92.6 [4.6] | **p < 0.05*** | -0.41 |
| | | Subchar+gait | 100.0 [0.0] | p < 0.05 | -0.29 |
| | | **Subchar+SLMS** | 92.6 [7.4] | **p < 0.05*** | -0.45 |
| | | **Gait+SLMS** | 96.3 [8.3] | **p < 0.05** | -0.29 |
| **Only gait** | 92.6 [7.4] | Only SLMS | 92.6 [4.6] | 0.57 | -0.06 |
| | | **Subchar+gait** | 100.0 [0.0] | **p < 0.05*** | -0.55 |
| | | Subchar+SLMS | 92.6 [7.4] | 0.55 | -0.06 |
| | | Gait+SLMS | 96.3 [8.3] | 0.10 | -0.17 |
| **Only SLMS** | 92.6 [4.6] | **Subchar+gait** | 100.0 [0.0] | **p < 0.05*** | -0.50 |
| | | Subchar+SLMS | 92.6 [7.4] | 1.00 | 0.00 |
| | | Gait+SLMS | 96.3 [8.3] | 0.20 | -0.13 |
| **Subchar+gait** | 100.0 [0.0] | **Subchar+SLMS** | 92.6 [7.4] | **p < 0.05*** | -0.52 |
| | | **Gait+SLMS** | 96.3 [8.3] | **p < 0.05*** | -0.45 |
| Subchar+SLMS | 92.6 [7.4] | Gait+SLMS | 96.3 [8.3] | 0.20 | -0.13 |
| Gait + SLMS | 96.3 [8.3] | | | | |

[a] Since the Friedman ANOVA is a non-parametric test, medians and interquartile range are reported.
***** indicates significance after Bonferroni correction: p < (0.05/amount of tests), with amount of tests = 22.

## Appendix G. Relative contribution of each input feature to classification

**Table G1. Relative contribution (%) of each input feature to classification using the maximal feature set (without KL score)**

| | Mean (%) | SD | Larger[a] |
|---|---|---|---|
| **Basic personal information** | | | |
| **Gender** | 0.41 | 0.02 | **>** |
| **Age** | 0.36 | 0.02 | **>** |
| Height | 0.01 | 0.00 | |
| **Weight** | 0.45 | 0.02 | **>** |
| **BMI** | 0.16 | 0.01 | **>** |
| | | | |
| **PROM scores** | | | |
| **KOOSpain** | 0.33 | 0.02 | **>** |
| **KOOSsympt** | 0.33 | 0.02 | **>** |
| **KOOSadl** | 0.34 | 0.02 | **>** |
| **KOOSsport** | 0.28 | 0.02 | **>** |
| **KOOSqol** | 0.29 | 0.02 | **>** |
| EQ5Dscore | 0.00 | 0.00 | |
| **EQ5Dvas** | 0.35 | 0.02 | **>** |
| | | | |
| **PBtest** | 0.14 | 0.02 | **>** |
| MeanCLINrating | 0.01 | 0.00 | |
| | | | |
| **Gait spatiotemporal characteristics** | | | |
| Speed | 0.01 | 0.00 | |
| **Cadence** | 0.62 | 0.03 | **>** |
| Stride time | 0.01 | 0.00 | |
| Stride length | 0.01 | 0.00 | |
| | | | |
| **Gait kinematics (affected side)** | | | |
| Peak trunk ipsilateral lean | 0.01 | 0.00 | |
| Peak trunk contralateral lean | 0.01 | 0.00 | |
| **Peak pelvic anterior tilt during GC** | 0.07 | 0.01 | **>** |
| **Pelvic anterior tilt at start SS** | 0.06 | 0.01 | **>** |
| **Pelvic anterior tilt at end SS** | 0.06 | 0.01 | **>** |
| Peak pelvic drop during swing | 0.02 | 0.00 | |
| **Peak hip flexion during stance** | 0.18 | 0.01 | **>** |
| Peak hip extension during stance | 0.05 | 0.01 | |
| **Hip flexion-extension ROM during GC** | 0.24 | 0.01 | **>** |
| Peak hip adduction during stance | 0.01 | 0.01 | |
| Hip adduction at the start of SS | 0.01 | 0.01 | |
| **Peak knee flexion during GC** | 0.26 | 0.02 | **>** |
| Knee flexion at the start of SS | 0.02 | 0.01 | |
| **Peak knee flexion during LR** | 0.07 | 0.01 | **>** |
| Peak knee extension during GC | 0.01 | 0.00 | |
| Peak knee extension during stance | 0.01 | 0.00 | |
| **Knee flexion-extension ROM during GC** | 0.28 | 0.02 | **>** |
| Peak ankle plantarflexion during stance | 0.05 | 0.01 | |
| Peak ankle dorsiflexion during swing | 0.04 | 0.01 | |
| **Ankle plantar-dorsiflexion ROM over GC** | 0.14 | 0.01 | **>** |
| | | | |
| **SLMS spatiotemporal characteristics** | | | |
| **LimbSymmIndex** | 0.00 | 0.00 | |
| TestDuration | 12.55 | 2.30 | **>** |
| | | | |
| **SLMS kinematics** | | | |
| **SLMS Peak anterior trunk tilt** | 0.11 | 0.02 | **>** |
| **SLMS Mean anterior trunk tilt** | 0.06 | 0.01 | **>** |
| **SLMS Knee flexion-extension ROM** | 0.24 | 0.01 | **>** |
| | | | |
| **$PCA_{Gait}$** | | | |
| **Gait PC1 xi** | 11.37 | 0.24 | **>** |
| **Gait PC2 xi** | 9.01 | 0.11 | **>** |

| | | | |
|---|---|---|---|
| **Gait PC3 xi** | 7.35 | 0.10 | **>** |
| Gait PC4 xi | 3.21 | 0.03 | |
| Gait PC5 xi | 2.61 | 0.03 | |
| Gait PC7 xi | 1.08 | 0.02 | |
| Gait PC11 xi | 0.76 | 0.01 | |
| **Gait PC1 var xi** | 2.95 | 0.90 | **>** |
| **Gait PC2 var xi** | 1.97 | 0.73 | **>** |
| **Gait PC3 var xi** | 1.46 | 0.69 | **>** |
| **Gait PC4 var xi** | 0.32 | 0.37 | **>** |
| **Gait PC5 var xi** | 0.28 | 0.50 | **>** |
| **Gait PC7 var xi** | 0.13 | 0.36 | **>** |
| **Gait PC11 var xi** | 0.06 | 0.30 | **>** |
| | | | |
| **$PCA_{SLMS}$** | | | |
| **SLMS PC1 xi** | 13.33 | 2.36 | **>** |
| **SLMS PC2xi** | 6.53 | 1.66 | **>** |
| SLMS PC3 xi | 2.08 | 1.13 | |
| SLMS PC4 xi | 1.52 | 0.93 | |
| SLMS PC5 xi | 1.12 | 0.70 | |
| SLMS PC6 xi | 0.96 | 0.46 | |
| SLMS PC8 xi | 0.38 | 0.18 | |
| SLMS PC11 xi | 0.33 | 0.18 | |
| SLMS PC12 xi | 0.24 | 0.11 | |
| SLMS PC19 xi | 0.18 | 0.08 | |
| SLMS PC23 xi | 0.13 | 0.06 | |
| **SLMS PC1 var xi** | 6.41 | 1.22 | **>** |
| **SLMS PC2var xi** | 2.47 | 0.42 | **>** |
| **SLMS PC3 var xi** | 1.37 | 0.20 | **>** |
| **SLMS PC4 var xi** | 0.65 | 0.13 | **>** |
| **SLMS PC5 var xi** | 0.40 | 0.08 | **>** |
| **SLMS PC6 var xi** | 0.27 | 0.06 | **>** |
| **SLMS PC8 var xi** | 0.07 | 0.01 | **>** |
| SLMS PC11 var xi | 0.03 | 0.00 | |
| SLMS PC12 var xi | 0.03 | 0.00 | |
| SLMS PC19 var xi | 0.01 | 0.00 | |
| SLMS PC23 var xi | 0.01 | 0.00 | |
| | | | |
| **Sum** | 100 | | 100 |
| | | | |

[a] Larger compared to the situation in which each data point equally contributes to classification (calculated as $N_{datapoints}$*100/1863 = 0.054%*$N_{datapoints}$).