Annex A; Clustering Technique

Reasoning of applying a subgroup analysis

Statistical analysis in biomechanical research is generally performed using a single group design [1], which assumed that injury related factors (movement deficiencies) are equal across a population. However, movement strategies may differ significantly across individuals and applying a single group analysis could mask movement deficiencies present [1 2] within a movement [3 4]. Another design is the single subject design, which assumes that every individual has a unique movement strategy and consequently unique movement deficiencies [1]. However, findings are dependent on the studied athlete itself and its current mental and physical condition during the data capture [5]. Consequently, findings are limited because the number of observations might affect findings (due to practice or fatigue effects), the generalization of findings is problematic and the comparison of interventions is difficult - as the ordering of interventions might affect results [5 6]. One alternative design, combining the strengths of the single subject and group design, is the analysis of subgroups. A subgroup analysis accounts for different movement strategies across individuals by classifying individuals into subgroups (clusters) based on their movement strategies. Consequently, a subgroup analysis does not mask movement deficiencies (or at least reduce the risk of masked movement deficiencies), allows the generalization of findings, the comparison of interventions and its benefit has been demonstrated in previous research [7-10] [11-13]

Methodology – Subject Score Generation

Before classifying the captured kinematic and kinetic measures, subject scores were calculated using the idea of Analysis of Characterizing Phases [14]. Analysis of

Characterizing Phases detects phases of variation (pattern characterizing phases) within the kinematic and kinetic measures, allowing the calculation of a subject's score that captures the behaviour of a subject within a phase of variation. Phases of variation were identified using VARIMAX rotated principal components that together described 99% of the variances in the data [13]. Subsequently, subject scores (SS) were calculated as the summed difference between a subjects waveform (Wave) and the average wavefrom ($\overline{Wave}$) for every time point (*i*) within the identified phase (*k*). This was completed for every kinematic and kinetic measure creating a matrix of subjects scores (rows = number of phases; columns = number of subjects; see Equation A.1).

$$SS_{k,n} = \sum_{i=start\ phase}^{end\ phase} Wave_n(i) - \overline{Wave}(i) \qquad \text{A.1.}$$

To maximize the ability to identify movement strategies in further analysis, this matrix was normalized by transforming the similarity score matrix into its correlation matrix (Equation A.2). This step quantifies numerically the relationship between the similarity scores, which cannot be described by distances of the generated similarity scores). The correlation matrix ($\hat{P}$; $\hat{P} \in \mathbb{R}^{322x322}$) was calculated using the Pearson's r-value (*corr*) utilizing the subject scores (*SS*) of the subject *n* (*n* = 1, 2, … number of subject) and *i* (*i* = 1, 2, … number of subject).

$$\left[\hat{P}\right]_{(i,n)} = corr_{(i,n)} = \frac{1}{N-1}\sum_{k=1}^{N}\frac{\left(SS_{i,k} - \mu_i\right) * \left(SS_{n,k} - \mu_n\right)}{\sigma_i * \sigma_n} \qquad \text{A.2}$$

where μ is the average and σ the standard deviation for subject *i* and *n* of their corresponding SS, which were calculated using the identified phases of variance (k = 1,2, … N, where N is the number of identified key phases)

Methodology - Decision Number of Clusters

Gap statistic was used to decide number of clusters (*k*) within the sample [15]. Gap statistics compares the within-cluster dispersion of a data set [$\log(W_k)$] for a number of requested cluster solution (e.g. *k* = 2 to 25) to the average within-cluster dispersion ($\bar{W}_k$) cluster solution *k* computed from B reference data sets (uniform copy of the real data) that hold a null distribution (e.g. no underpinning pattern; Equation A.3 and A.4).

$$gap(k) = \bar{W} - \log(W_k) \quad \text{A.3}$$

where

$$\bar{W} = \frac{1}{B}\sum_{b} \log(W^*_{k,b}) \quad \text{A.4}$$

The within-cluster dispersion is calculated as defined in equation A.5

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \quad \text{A5}$$

where *D* represents the sum of the pairwise distances between all points/subjects in the cluster *r*. [15] suggest that the optimal number of clusters is when the gap at *k* is greater or equal to the k-1 cluster minus its standard deviation (sd_k) for the first time of the within-cluster dispersion of the computed B reference data sets (Equation A.6, A.7 and A.8).

$$sd_k = \sqrt{\frac{1}{B}\sum_{b} \log(W^*_{k,b} - \bar{W})} \quad \text{A.6}$$

$$s_k = sd_k\sqrt{1 + 1/B} \quad \text{A.7}$$

$$gap(k) \geq gap(k+1) - s_{k+1} \quad \text{A.8}$$

The interested reader is referred to the text of[15] or [16] for further information.

Methodology – Optimal number of Clusters

When generating the gap curve for 2 to 25 clusters, the A.8 defined criteria was met first at 3 clusters (Figure 6).

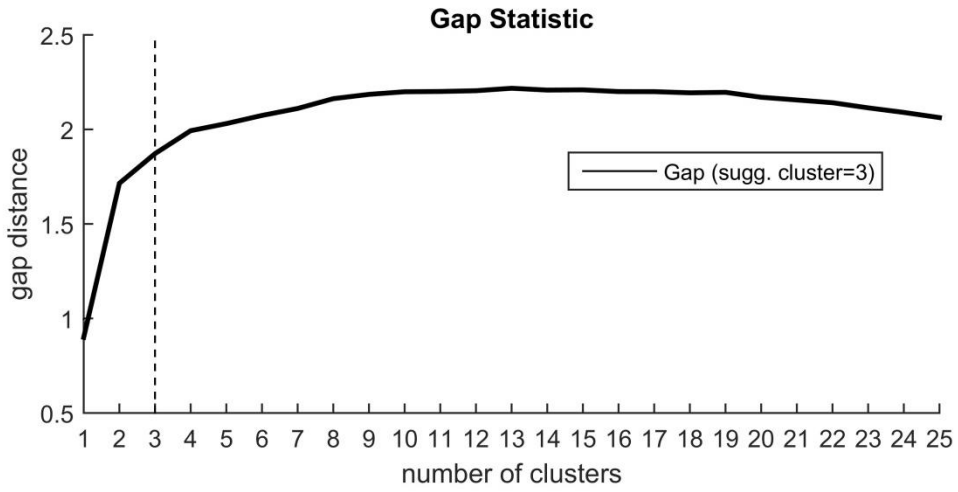

Figure 1: Illustrates findings of the gap statistics for the groin pain sample

Methodology – Data Clustering

After the number of clusters was identified, the correlation matrix was used as input for a hierarchical clustering approach to separate the sample into three subgroups. Hierarchical clustering follows the idea that the distance of the observed measures to each other represents the similarity of individuals, which is illustrated in Figure 2 or further explained in [17] .

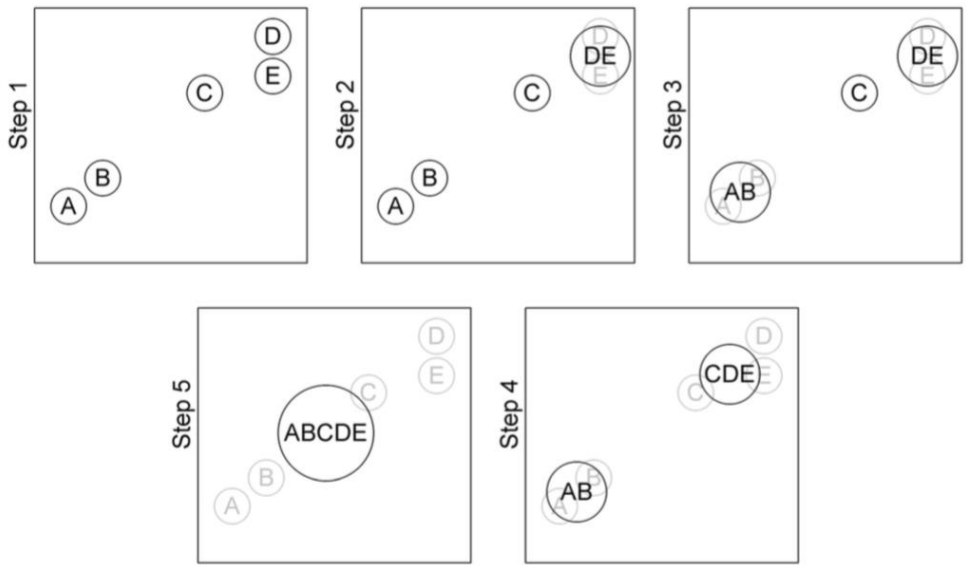

Figure 2: A hierarchal cluster algorithm starts with considering each individual as one group (Step1). Subsequently, it calculates the distance between every individual and searches for the two individuals with the smallest distance to each other (e.g. the two most similar individuals), which are then merged into one group. In the next iteration the two merged individuals are considered as one group that is located at the mid-point between both individuals (Step2). The hierarchical clustering algorithm repeats this process until the requested number (in example 5) of clusters is reached (Step3-5).

The hierarchical algorithm calculated pairwise distances using Euclidean distance, and created a hierarchical cluster tree using the nearest distance [16] The quality of the hierarchical clustering was measured by calculating the cophenetic correlation coefficient between the hierarchical cluster tree and the pairwise distances [16 18]Hierarchical clustering properties were changed if the cophenetic correlation coefficient was less than 0.7 (a low or medium correlation between the hierarchical cluster tree and the pairwise distances). The generated hierarchical cluster tree and the pairwise distances generated a cophenetic correlation coefficient above 0.7.

1. Bates BT. Single-subject methodology: an alternative approach. Med Sci Sports Exerc 1996;**28**(5):631-8
2. Vanezis A, Lees A. A biomechanical analysis of good and poor performers of the vertical jump. Ergonomics 2005;**48**(11-14):1594-603 doi: 10.1080/00140130500101262[published Online First: Epub Date]|.
3. Stergiou N. *Innovative Analyses of Human Movement*. 1st ed. Leeds, UK: Human KInetics, 2004.
4. Stergiou N, Scott MM. Baseline measures are altered in biomechanical studies. J Biomech 2005;**38**(1):175-8 doi: 10.1016/j.jbiomech.2004.03.007[published Online First: Epub Date]|.
5. Backman CL, Harris SR. Case studies, single-subject research, and N of 1 randomized trials: comparisons and contrasts. American journal of physical medicine & rehabilitation / Association of Academic Physiatrists 1999;**78**(2):170-6
6. Morgan DL, Morgan R. *Single-Case Reserach Methods for the Behavioural and Health Sciences*: SAGE publications, 2008.
7. Carriero A, Zavatsky A, Stebbins J, Theologis T, Shefelbine SJ. Determination of gait patterns in children with spastic diplegic cerebral palsy using principal components. Gait Posture 2009;**29**(1):71-5 doi: 10.1016/j.gaitpost.2008.06.011[published Online First: Epub Date]|.
8. O'Byrne JM, Jenkinson A, O'Brien TM. Quantitative analysis and classification of gait patterns in cerebral palsy using a three-dimensional motion analyzer. J Child Neurol 1998;**13**(3):101-8
9. Kienast G, Bachmann D, Steigerwalt AG, Zwick EB, Saraph V. Determination of gait patterns in children with cerebral palsy using cluster

analysis

. Gait and POsture 1999;**10**(1):57
10. Toro B, Nester CJ, Farren PC. Cluster analysis for the extraction of sagittal gait patterns in children with cerebral palsy. Gait Posture 2007;**25**(2):157-65 doi: 10.1016/j.gaitpost.2006.02.004[published Online First: Epub Date]|.
11. O'Malley MJ, Abel MF, Damiano DL, Vaughan CL. Fuzzy clustering of children with cerebral palsy based on temporal-distance gait parameters. IEEE Trans Rehabil Eng 1997;**5**(4):300-9
12. von Tscharner V, Enders H, Maurer C. Subspace identification and classification of healthy human gait. PLoS One 2013;**8**(7):e65063 doi: 10.1371/journal.pone.0065063[published Online First: Epub Date]|.
13. Richter C, O'Connor NE, Marshall B, Moran K. Clustering vertical ground reaction force curves produced during countermovement jumps. J Biomech 2014;**47**(10):2385-90 doi: 10.1016/j.jbiomech.2014.04.032[published Online First: Epub Date]|.
14. Richter C, O'Connor NE, Marshall B, Moran K. Analysis of characterizing phases on waveform: an application to vertical jumps. J Appl Biomech 2014;**30**(2):316-21 doi: 10.1123/jab.2012-0218[published Online First: Epub Date]|.
15. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2001;**63**(2):411-23
16. Sage.Martinez W, Martinez A, Solka J. Exploratory Data Analysis with MATLAB: CRC Press, 2004.
17. Segaran T. Programming collective intelligence: building smart web 2.0 Applications. 1st edition (August 26, 2007) ed: O'Reilly, 2007.
18. Sokhal RR. The comparison of dendrograms by objectve methods. Taxon 1962;**11**(2):33-40