MRS SUSANNE JAUHIAINEN (Orcid ID : 0000-0001-8553-8018)

Article type : Original Article

A hierarchical cluster analysis to determine whether injured runners exhibit similar kinematic gait patterns.

Running title: Hierarchical clustering of injured runners

Susanne Jauhiainen[1,*], Andrew J. Pohl[2], Sami Äyrämö[1], Jukka-Pekka Kauppi[1], Reed Ferber[2,3,4].

[1] Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland;[2] Faculty of Kinesiology, University of Calgary, Calgary, Alberta, Canada;[3] Faculty of Nursing, University of Calgary, Calgary, Alberta, Canada;[4] Running Injury Clinic, Calgary, Alberta, Canada

Corresponding Author:

M.Sc. Susanne Jauhiainen,
Faculty of Information Technology, University of Jyväskylä, P.O. Box 35, FI-40014, University of Jyväskylä, Finland; Tel. +358408053652; susanne.m.jauhiainen@jyu.fi; Orcid ID: 0000-0001-8553-8018

Acknowledgements

Susanne Jauhiainen was funded by the Jenny and Antti Wihuri Foundation (grant 00180121). Reed Ferber and Andrew Pohl were funded by the Natural Sciences and Engineering Research Council of Canada (NSERC grant 1030390).

Disclosure of interest

The authors report no conflicts of interest

A hierarchical cluster analysis to determine whether injured runners exhibit similar kinematic gait patterns.

Abstract

Previous studies have suggested that runners can be subgrouped based on homogeneous gait patterns, however, no previous study has assessed the presence of such subgroups in a population of individuals across a wide variety of injuries. Therefore, the purpose of this study was to assess whether distinct subgroups with homogeneous running patterns can be identified among a large group of injured and healthy runners and whether identified subgroups are associated with specific injury location. Three-dimensional kinematic data from 291 injured and healthy runners, representing both sexes and a wide range of ages (10-66 years) was clustered using hierarchical cluster analysis. Cluster analysis revealed five distinct subgroups from the data. Kinematic differences between the subgroups were compared using one-way analysis of variance (ANOVA). Against our hypothesis, runners with the same injury types did not cluster together, but the distribution of different injuries within subgroups was similar across the entire sample. These results suggest that homogeneous gait patterns exist independent of injury location and that it is important to consider these underlying patterns when planning injury prevention or rehabilitation strategies.

Keywords: Running, Injury, Kinematics, Unsupervised machine learning

Introduction

Running is a popular sport for developing and maintaining cardiovascular fitness[1,2], despite a relatively high injury rate. Estimates of the annual prevalence of lower extremity running related injuries (RRI) vary between 19.4% to 79.3%[3], while a widely accepted estimate is that 50% of runners experience an RRI annually[1]. Naturally, with such a high injury rate, the etiology of RRIs has received much attention within the gait analysis community to determine injury etiology for specific RRIs.

Multiple studies have suggested that there are similar kinematic gait patterns for runners with similar injury locations. For example, studies involving common knee RRIs, such as patellofemoral pain (PFP)[4–7] and iliotibial band syndrome (ITBS)[8,9] have reported that runners exhibit increased peak hip adduction and peak knee internal rotation. As well, similarities in ankle-related RRIs have been reported in individuals with Achilles tendinopathy (AT)[10] and those with posterior tibial tendon dysfunction[7] suggesting these injured runners exhibit increased time to peak pronation[10]. However, a recent study by Bramah et al.[11] investigated 72 injured runners, with PFP, ITBS, AT, and medial tibial stress syndrome (MTSS), and 36 healthy runners and reported that all injured runners exhibited greater contralateral pelvic drop and forward trunk lean as well as more knee extension and ankle dorsiflexion at initial contact irrespective of injury location. Therefore, further research is necessary to determine if common kinematic gait patterns exist for specific injuries and/or anatomical location of injuries.

One method in which to gain some understanding as to the etiology of RRIs may be to examine whether unsupervised machine learning techniques can identify homogenous subgroups within a large population of injured runners and healthy individuals. Unsupervised machine learning works by discovering underlying patterns and associations in datasets without any a priori information aside from a set of input variables[12]. These statistical methods have demonstrated success in uncovering underlying structure within datasets in previous sport science research[13–17]. For example, race patterns in elite swimmers[13], different golf swing patterns[14], and different gait patters in injured and healthy runners[15–17], have all been successfully identified using unsupervised machine learning methods. Another study used unsupervised machine learning to create a movement profile for healthy controls walking and assessed the deviation of patients with gait problems from this profile[18]. However, to our knowledge very few studies have utilized this approach for a large population of injured and healthy runners. Recognition of distinct gait

patterns can be utilized in injury rehabilitation protocols and in future biomechanical investigations seeking to better understand injury etiology.

Thus, the main purpose of this exploratory study was to investigate whether a large group of injured and healthy runners can be clustered into subgroups of homogeneous gait patterns based on 3D kinematic data. We hypothesized that runners with injuries at similar locations would exhibit similarities in gait patterns and consequently be identified within the same homogenous subgroups. Similarly, we assumed the gait pattern of healthy runners would be distinct from those identified within injured homogenous subgroups. A secondary purpose was to analyze differences in 3D kinematics between the formed subgroups to better understand differences in running kinematics between the subgroups.

Materials and methods

Participants

A sample of 291injured and healthy runners were queried from an existing database of running kinematics[19,20]. Participants were recruited through standard advertisements (e.g., posters, facebook posts) approved by the Ethics Board. In this study, runners were identified for inclusion provided they: (i) had 3D gait analysis performed by the same experienced clinician using the same motion capture system, (ii) were known to be injury free for the 6 months preceding data collection (25 runners) or were suffering some form of lower body injury at the time of data collection (266 runners) and (iii) contained complete kinematic data with no missing values. A standard approach to define RRI based on Yamato et al. was used "*Running-related musculoskeletal pain in the lower limbs that causes a restriction on or stoppage of running (distance, speed, duration, or training) for at least 7 days or 3 consecutive scheduled training sessions, or that requires the runner to consult a physician or other health professional."*[23]. The runners (146 females, age 39.51±11.21 years) were defined either as competitive (n=57) or recreational (n=234)[22] based on age, sex, and most recent race performance (10 km, half-marathon, or marathon) and the World Masters Association Age Grading Performance Tables[1]. Injuries were diagnosed and injury history was collated by a licensed health professional (e.g., physiotherapist, medical doctor, athletic therapist). Injuries were grouped by location with: 72

[1] http://www.usatf.org/Resources-for---/Masters/LDR/Age-Grading.aspx. Accessed May 21, 2019.

knee, 58 ankle/foot, 51 hip/pelvis, 42 thigh, 39 lower leg (shin) identified. Across the 266 injured runners, a wide variety of RRI were reported. However, the main injuries included patellofemoral pain (n=44), iliotibial band syndrome (n=29), achilles tendinopathy (n=15), plantar fasciitis (n=14), and medical tibial stress syndrome (n=12). The occurrence and distribution of these main injuries were consistent with previously reported epidemiological investigations[24]. Other injuries included specific muscle strains (e.g. gastrocnemius, hamstring, hip flexor), tendinopathies (e.g. tibialis posterior tendinopathy, patellar tendinopathy), as well as generalized joint pain. None of the injured participants described any pain during the treadmill running procedure.
Twenty-five individuals were confirmed as injury free for at least six months prior to data collection. Participant characteristics are summarized by injury location in Table 1. Data collection was approved by the University of Calgary's Conjoint Health Research Ethics Board (CHREB: REB15- 0557). Before data collection, all participants provided a written informed consent to participate.
[Table 1 somewhere here]

Data collection

Three-dimensional (3D) marker trajectory data were captured via an 8-camera VICON motion capture system (MX3+, Vicon Motion Systems Oxford, UK) at 200 Hz while participants ran on a treadmill (Bertec Corporation, Columbus, OH). Spherical retro-reflective markers were placed on anatomical landmarks and rigid plates with clusters of 3-4 markers were placed on each of seven lower body segments as per Pohl et al.[25]. Anatomical markers and segmental clusters were placed by a single examiner with over twenty years' experience in clinical gait analysis and physical therapy[26,27]. This marker-set consisted of seven rigid segments and has been reported to produce reliable kinematic waveforms[25]. To allow for unobstructed movement during running, anatomical markers were removed following a one second static trial where subjects stood upon a template with their feet positioned straight ahead and 0.3m apart with arms crossed over their chest.

Following a warmup period of 2-5 minutes, kinematic data were collected for approximately 60 seconds while participants ran at a constant self-selected speed between 1.84 and 3.37 m/s. In order to standardize the footwear condition, each participant wore the same shoes (Pegasus, Nike, Beaverton, USA). It should be noted that while the use of standardized shoes has

advantages, it might also alter running patterns as it is possible that some participants may not be accustomed to the shoes.

Data processing

Key gait events, foot strike and toe off were identified using a Principal Component Analysis (PCA) approach as described elsewhere[28,29]. Joint angles within each movement plane were extracted using 3D GAIT custom software (Running Injury Clinic Inc., Calgary, Alberta, Canada), and time normalized to 100 data points per gait cycle: 35 data points for stance and 65 data points for the swing phase.

Each subject's running pattern was described by the median for each of 62 kinematic (e.g., peak knee flexion and adduction angles, heel strike angle) and functional variables (e.g., step width, vertical oscillation, stride rate and length) extracted from each gait cycle. A minimum of ten gait cycles were included but generally approximately 30–40 consecutive running strides were collected for processing and analysis. All variables were extracted from frontal and sagittal plane motion given the limited reliability of transverse plane angles during motion capture analysis[27,30,31]. A full description of these variables is provided in Table A1 found in the appendices.

A matrix (291 subjects x 62 variables) was created with each column normalized to have a mean of zero and standard deviation of one. A PCA was performed on the data to reduce multi-collinearity between biomechanical variables. A subset of principal components (PCs) were chosen so that 80% of the total variance in the dataset was explained by the selected PCs[32].

Cluster analysis

A hierarchical cluster analysis (HCA) was used to identify subgroups with homogeneous gait patterns. A hierarchical cluster tree, a dendrogram, was formed with the *linkage*-function in *Statistics and machine learning toolbox 11.0* of MATLAB. The function was used with the Ward's linkage method and Euclidean distance. The subgroups were formed in an agglomerative manner, i.e. starting with each observation as their own subgroup and at every step pairing the two closest subgroups together until only one group remains. The final number of subgroups was chosen based on a stopping rule (a large percentage decrease in the coefficient followed by a plateau)[33,34]. The number of subgroups was also confirmed by visual inspection of the

dendrogram[16,17].

Interpretation and comparison of subgroups

After forming the subgroups, a univariate analysis of variance (ANOVA) was used to determine which PCs separated each subgroup from the others[17]. The PCs were then interpreted by calculating the loadings of variables to determine which variables comprised the PCs[35]. Demographic information (height, weight, age, and running speed) of the subgroups were also compared using ANOVA. Normality of variables was tested via a Shaphiro-Wilk test and equal variances with Levene's test and in the cases where assumptions were not met, non-parametric Kruskal Wallis tests were used instead. When significant differences occurred, post-hoc tests were performed using Tukey's test. The proportion of injuries and males/females in subgroups were compared using Chi-squared test. For all tests, a significance limit of α=0.05 was chosen and adjusted with Bonferroni's correction and Cohen's effect size d was calculated where appropriate[36].

The injury distribution of the formed subgroups was assessed with the adjusted Rand index[37]. The Rand index objectively measures the similarity between two different clusterings of the same data. If $X = \{x_1, x_2, ..., x_n\}$ is the set of observations and $P = \{P_1, P_2, ..., P_{K_1}\}$ and $P' = \{P'_1, P'_2, ..., P'_{K_2}\}$ are two different partitions of X, where n is the number of observations in data and K_1 and K_2 are the number of subgroups in partitions P and P' respectively, the Rand Index is calculated by using all possible pairs of observations in X. Defining s as the number of pairs that are clustered to the same subgroup in both P and P', and d as the number of pairs that are not clustered to the same subgroup in either P or P', finally the Rand Index can be written as $R = \frac{s+d}{\binom{n}{2}}$, where the denominator is the total number of pairs. Simply put, the index measures the proportion of similar pairings, over all possible pairs of observations. The index receives a value between 1 and 0, with 1 indicating the clusterings are exactly the same while 0 indicates that clustering do not agree on any parts. The adjusted version works similarly but is corrected for chance. The index was calculated between the clustering labels and the injury class labels with a custom Matlab script[2]. All data processing and analysis were performed on MATLAB R2016b (MathWorks Inc).

[2] https://se.mathworks.com/matlabcentral/fileexchange/49908-adjusted-rand-index. Accessed May 21, 2019.

Results

The first 16 PCs, explaining 80.98% of the total variance, were chosen as input for the HCA method. The dendrogram for the clustering results is outlined in Figure 1. Between four and five subgroups, there was a large decrease in the agglomeration coefficients (21.0% based on min–max normalized linkage distances), followed by a plateau between five and six (6.0%). Therefore, the number of subgroups was set to five and the result was also confirmed by visual inspection of the dendrogram.

[Figure 1 somewhere here]

Despite five distinct subgroups being identified (average linkage distance of 39.32 between subgroups), the population of injured and healthy runners was randomly dispersed amongst the subgroups and this was confirmed with the very low Rand index score of r=0.012 when the cluster partition and the original injury classification were compared. The proportion of injured and healthy runners was not different between the subgroups (X^2=0.53, p=0.99) and similarly there was no evidence to suggest a difference in any of the injury types/locations between the subgroups (X^2=20.20, p=0.251).

The demographics of the subgroups are described in Table 2. The proportion of males and females was different between the subgroups (X^2=53.85, p<0.01) with the proportion of males in S1 (73.7%) being higher (X^2=35.00, p<0.01, d=0.80) compared to other subgroups and similarly, the proportion of females in S5 (63.3%) was higher (X^2=31.03, p<0.01, d= 0.81). There was evidence to suggest differences in weight (X^2=61.80, p<0.01), height (X^2=22.30, p<0.01), and running speed (X^2=56.18, p<0.01), but not in age (F=2.47, p=0.18).

[Table 2 somewhere here]

There were differences between the subgroups in the five first PCs. The amount of variance explained by the individual PCs was 13.44, 12.34, 10.01, 6.53, and 6.06 percent for the first five PCs respectively. PC1 was primarily loaded on frontal plane hip variables and stride rate, vertical oscillation, and swing time. PC2 consisted of frontal plane knee variables, stride length, vertical oscillation, and swing time. PC3 was comprised of ankle and foot frontal plane variables.

PC4 consisted of variables describing ankle eversion, excursion and peak eversion velocity. PC5 consisted of heel strike angle and knee adduction excursion. Clear differences in running biomechanics between the subgroups were found.

Subgroup 1

S1 was separated from the other subgroups by PC2 ($p<0.001$, $F=135.13$, $d=2.08$) Compared to the other four subgroups, S1 had the largest peak knee adduction (-3.16±4.08 deg), the least knee abduction (-8.60±4.32 deg), and exhibited greater knee flexion (-49.01±4.03 deg). S1 also exhibited the second largest stride length (1.97±0.17 m), vertical oscillation (89±14.32 m), and swing time (0.45±0.03 s) compared to the other subgroups. Also, 73.68% of runners in S1 were males.

Subgroup 2

Subgroup S2 was separated from the other subgroups by PC1 ($p<0.001$, $F=109.69$, $d=2.05$) and PC2 ($p<0.001$, $F=33.72$, $d=1.16$). S2 exhibited the smallest knee flexion peak (-44±5.47 deg), second smallest hip adduction (8.87±3.95 deg) and knee abduction (-11.79±4.00 deg). The S2 subgroup also exhibited the highest stride rate (87.12±4.42 strikes/min) as well as the lowest swing time (0.40±0.03 s), stride length (1.66±0.16 m), and vertical oscillation (72.04±10.10 m) compared to the other subgroups.

Subgroup 3

S3 was separated from the other subgroups by PC1 ($p<0.001$, $F=92.27$, $d=2.12$). The S3 subgroup exhibited the second highest hip adduction peak (12.07±4.00 deg), hip adduction excursion (10.13±2.79 mm), and hip abduction velocity peak (160.76±44.96 deg). They also had the lowest stride rate (77.12±3.25 strikes/min), highest swing time (0.47±0.03 s), and the most vertical oscillation (104±13.10 m) compared to the other four subgroups.

Subgroup 4

S4 was separated from the others by PC3 ($p<0.001$, $F=25.44$, $d=1.20$), PC4 ($p<0.001$, $F=32.53$, $d=1.30$), and PC5 ($p<0.001$, $F=20.14$, $d=1.04$). Compared to the other four subgroups, S4 had the largest heel strike angle (17.10±4.90 deg) and largest foot progression angle (-15±4.96 deg) along with the second largest offset (42.27±8.40 % of gait cycle) and onset (13.96±2.67 % of gait cycle) rearfoot eversion.

Subgroup 5

Subgroup S5 was separated by PC1 ($p<0.001$, $F= 30.65$, $d=1.10$), PC2 ($p<0.001$, $F=36.05$, $d=1.17$), and PC3 ($p<0.001$, $F=103.80$, $d=1.75$). S5 exhibited the highest offset (54.95±17.13 % of gait cycle) and onset (15.66±3.86 % of gait cycle) rearfoot eversion, longest time to peak pronation (0.29±0.13 % of gait cycle) as well as the smallest foot progression angle (-8.45±4.54 deg) compared to the other four subgroups. S5 also demonstrated the largest hip adduction velocity peak (173.87±52.67 deg /s), hip adduction excursion (10.61±3.16 mm), and hip adduction peak (13.03±4.28 deg). Also, 78.95% of the runners in S5 were females.

[Figure 2 somewhere here]

Discussion

The primary purpose of our study was to investigate whether distinct subgroups with homogeneous running gait patterns could be identified from a large group of injured and healthy runners using an unsupervised hierarchical cluster analysis. Five subgroups were identified, however, contrary to our initial hypothesis, individuals with similar injuries (or no injury) did not cluster together. Instead, different types of injuries, and healthy control subjects, were evenly distributed across the five subgroups.

These results support previous research that has shown that there are similarities in kinematics between individuals with different injuries[11] and refutes the premise that injury location is related to similarities in gait kinematic patterns. Moreover, the gait pattern of healthy runners was not distinct of that of the injured and suggests that there is not a single ’protective gait pattern’ reducing the likelihood of developing RRI. However, future prospective studies are necessary to support or refute this premise. Regardless, our results also show that in a large group of runners with different injuries, representing both sexes and a wide distribution of ages, exhibit biomechanical running patterns that can be subgrouped into five distinct patterns.

Specific gait patterns can be observed within each subgroup. Specifically, S1 consisted of mostly male runners whose knee collapsed and flexed the most during running and ran at the fastest pace. Runners in S2 exhibited overall limb stiffness, observed as the least amount of peak knee flexion as well as second least amount of hip adduction and knee abduction. Runners in S3 had the second largest hip adduction peak angle, hip adduction excursion and hip abduction peak

velocity. S4 consisted of runners that exhibited a pronounced heelstrike and a large foot progression angle during running. Runners in S5 had the highest hip adduction peak and hip adduction excursion as well as the smallest foot progression angle, as well as the most rearfoot eversion and time to peak pronation. S5 also had a high ratio of females and they ran at the slowest pace.

The results of the current study suggest that it is possible that the traditional method of creating a "cluster" of subjects based on a pre-defined injury does not consider that variance of gait biomechanical patterns exists independent of the injury location/category. Thus, we propose that in order to discover these inter-relationships between movement patterns and injuries better, it is necessary to segment, or sub-type, according to gait patterns as an initial step in developing rehabilitation protocols and with respect to future biomechanical investigations seeking to better understand injury etiology. We also suggest that future prospective studies should employ PCA and HCA approaches for large cohorts of injured and pain-free runners in order to determine whether biomechanical sub-types, or unique homogeneous clusters, are potentially related to higher rates of injury.

Previous studies have suggested that atypical biomechanical patterns can lead to injuries by causing excessive repetitive tissue loading during running[10,38]. In addition, associations between certain injuries and kinematic gait patterns have been detected in multiple studies[4,6,10]. In support of this premise, a study by Braham et al.[11] reported that runners with different injuries all exhibited similar patterns among each other. However, this study[11] only involved 72 injured runners and 36 healthy controls and a simple logistic regression model to determine which kinematic parameters could best separate the two groups. In contrast, the results of the current study used a much larger cohort and employed an unsupervised machine learning approach to reveal that certain running patterns cannot be conclusively linked to injury location and that homogeneous kinematic subgroups exist regardless of injury location.

Our study benefits from a large cohort of injured and healthy runners along with robust data collection procedures. Moreover, all data were collected by a single examiner with over 20 years' experience. This is an especially important point when using unsupervised machine learning methods to identify subgroups, as these methods might pick up patterns originating from subtly different marker placements resulting from different examiners[39,40]. Specifically, Osis et al.[26] reported that a novice examiner, with 6-years of experience and trained by the same expert examiner used in the current study, made improvements in their consistency over the course of

one-year of training. However, systematic differences were apparent in data collected during the end of the year. Thus, future research involving a large cohort should take into consideration the number of people collecting the data and/or use appropriate feedback methods[40,41] to minimize marker placement error.

Limitations in the current study are acknowledged. First, given our data source was created by amalgamating data collected for specific purposes, running speed was not uniformly controlled within this study. Deviations from preferred or self-selected speed have been shown to result in deviations from typical gait patterns in walking[42] and future research should consider this factor. Second, the variables were averaged over the gait cycles, while variability in movement patterns has been associated with injuries in previous studies[43–45]. Future studies could benefit from considering the variability across gait cycles. Second, the current study was retrospective in nature and future research should prospectively follow runners to determine whether similar subgroups exist prior to injury development. In addition, each injury group included several types of injuries, that might have different effects on gait. Lastly, the data for the present study were collected in a laboratory setting whilst running on a treadmill. Previous studies[46,47] have suggested that a laboratory-based setting limits our ability to study the multifactorial nature of RRIs. Therefore, future studies should utilise inertial measurement units (IMUs) to quantify running gait patterns in real-world environments and determine whether these homogenous subgroups exist.

Perspective

This study showed that among a large population of runners with different injuries, representing both sexes and a wide distribution of ages, distinct subgroups exist with homogeneous running gait patterns. Interestingly, these patterns were not related to injury location, but different type of injuries were randomly distributed throughout the subgroups, together with healthy individuals. These results suggest that the location of injury is not related to specific gait kinematic patterns and this should be considered when planning future research studies or when developing rehabilitation and injury prevention strategies. Therefore, we recommend that when performing a clinical examination of an injured runner, individual presentation plays a larger role than attempting to determine whether they are exhibiting a gait pattern previously associated with a specific injury. Finally, based on the results of this study, prediction of injuries, based on whether or not an individual exhibits specific kinematic gait patterns, is not supported.

References

1. Fields KB, Sykes JC, Walker KM, Jackson JC. Prevention of running injuries. *Curr Sports Med Rep*. 2010;9(3):176-182.

2. Van Middelkoop M, Kolkman J, Van Ochten J, Bierma-Zeinstra SMA, Koes BW. Risk factors for lower extremity injuries among male marathon runners. *Scand J Med Sci Sports*. 2008;18(6):691-697.

3. Van Gent RN, Siem D, van Middelkoop M, Van Os AG, Bierma-Zeinstra SMA, Koes BW. Incidence and determinants of lower extremity running injuries in long distance runners: a systematic review. *Br J Sports Med*. 2007;41(8):469-480.

4. Luz BC, dos Santos AF, de Souza MC, de Oliveira Sato T, Nawoczenski DA, Serrão FV. Relationship between rearfoot, tibia and femur kinematics in runners with and without patellofemoral pain. *Gait Posture*. 2018;61:416-422.

5. Watari R, Kobsar D, Phinyomark A, Osis S, Ferber R. Determination of patellofemoral pain sub-groups and development of a method for predicting treatment outcome using running gait kinematics. *Clin Biomech*. 2016;38:13-21.

6. Noehren B, Hamill J, Davis I. Prospective evidence for a hip etiology in patellofemoral pain. *Med Sci Sports Exerc*. 2013;45(6):1120-1124.

7. Rabbito M, Pohl MB, Humble N, Ferber R. Biomechanical and clinical factors related to stage I posterior tibial tendon dysfunction. *J Orthop Sport Phys Ther*. 2011;41(10):776-784.

8. Ferber R, Noehren B, Hamill J, Davis I. Competitive female runners with a history of iliotibial band syndrome demonstrate atypical hip and knee kinematics. *J Orthop Sport Phys Ther*. 2010;40(2):52-58.

9. Miller RH, Lowry JL, Meardon SA, Gillette JC. Lower extremity mechanics of iliotibial band syndrome during an exhaustive run. *Gait Posture*. 2007;26(3):407-413.

10. Ogbonmwan I, Kumar BD, Paton B. New lower-limb gait biomechanical characteristics in individuals with Achilles tendinopathy: A systematic review update. *Gait Posture*.

2018;62:146-156.

11. Bramah C, Preece SJ, Gill N, Herrington L. Is there a pathological gait associated with common soft tissue running injuries? *Am J Sports Med*. 2018;46(12):3023-3031.

12. Friedman J, Hastie T, Tibshirani R. *The Elements of Statistical Learning*. Vol 1. Springer series in statistics New York; 2001.

13. Chen I, Homma H, Jin C, Yan H. Identification of elite swimmers' race patterns using cluster analysis. *Int J Sports Sci Coach*. 2007;2(3):293-303.

14. Ball KA, Best RJ. Different centre of pressure patterns within the golf stroke I: Cluster analysis. *J Sports Sci*. 2007;25(7):757-770.

15. Kulmala J-P, Äyrämö S, Avela J. Knee extensor and flexor dominant gait patterns increase the knee frontal plane moment during walking. *J Orthop Res*. 2013;31(7):1013-1019.

16. Watari R, Osis ST, Phinyomark A, Ferber R. Runners with patellofemoral pain demonstrate sub-groups of pelvic acceleration profiles using hierarchical cluster analysis: an exploratory cross-sectional study. *BMC Musculoskelet Disord*. 2018;19(1):120.

17. Phinyomark A, Osis S, Hettinga BA, Ferber R. Kinematic gait patterns in healthy runners: A hierarchical cluster analysis. *J Biomech*. 2015;48(14):3897-3904.

18. Barton GJ, Hawken MB, Scott MA, Schwartz MH. Movement Deviation Profile: A measure of distance from normality using a self-organizing neural network. *Hum Mov Sci*. 2012;31(2):284-294.

19. Phinyomark A, Petri G, Ibáñez-Marcelo E, Osis ST, Ferber R. Analysis of big data in gait biomechanics: Current trends and future directions. *J Med Biol Eng*. 2018;38(2):244-260.

20. Ferber R, Osis ST, Hicks JL, Delp SL. Gait biomechanics in the era of data science. *J Biomech*. 2016;49(16):3759-3761.

21. Phinyomark A, Hettinga BA, Osis ST, Ferber R. Gender and age-related differences in bilateral lower extremity mechanics during treadmill running. *PLoS One*.

2014;9(8):e105246.

22. Clermont CA, Osis ST, Phinyomark A, Ferber R. Kinematic gait patterns in competitive and recreational runners. *J Appl Biomech*. 2017;33(4):268-276.

23. Yamato TP, Saragiotto BT, Lopes AD. A consensus definition of running-related injury in recreational runners: a modified Delphi approach. *J Orthop Sport Phys Ther*. 2015;45(5):375-380.

24. Taunton JE, Ryan MB, Clement DB, McKenzie DC, Lloyd-Smith DR, Zumbo BD. A retrospective case-control analysis of 2002 running injuries. *Br J Sports Med*. 2002;36(2):95-101.

25. Pohl MB, Lloyd C, Ferber R. Can the reliability of three-dimensional running kinematics be improved using functional joint methodology? *Gait Posture*. 2010;32(4):559-563.

26. Osis ST, Hettinga BA, Macdonald SL, Ferber R. A novel method to evaluate error in anatomical marker placement using a modified generalized Procrustes analysis. *Comput Methods Biomech Biomed Engin*. 2015;18(10):1108-1116.

27. Osis ST, Hettinga BA, Macdonald S, Ferber R. Effects of simulated marker placement deviations on running kinematics and evaluation of a morphometric-based placement feedback method. *PLoS One*. 2016;11(1):e0147111.

28. Osis ST, Hettinga BA, Leitch J, Ferber R. Predicting timing of foot strike during running, independent of striking technique, using principal component analysis of joint angles. *J Biomech*. 2014;47(11):2786-2789.

29. Osis ST, Hettinga BA, Ferber R. Predicting ground contact events for a continuum of gait types: an application of targeted machine learning using principal component analysis. *Gait Posture*. 2016;46:86-90.

30. Reinschmidt C, Van Den Bogert AJ, Nigg BM, Lundberg A, Murphy N. Effect of skin movement on the analysis of skeletal knee joint motion during running. *J Biomech*. 1997;30(7):729-732.

31. Kadaba MP, Ramakrishnan HK, Wootten ME, Gainey J, Gorton G, Cochran GVB. Repeatability of kinematic, kinetic, and electromyographic data in normal adult gait. *J Orthop Res*. 1989;7(6):849-860.

32. Jolliffe I. *Principal Component Analysis*. Springer; 1986.

33. Hair JF, Anderson Jr RE, Tatham RL, Black WC. Multivariate data analysis 7th Ed.(Global Edition). 2009.

34. Kinsella S, Moran K. Gait pattern categorization of stroke participants with equinus deformity of the foot. *Gait Posture*. 2008;27(1):144-151.

35. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat*. 2010;2(4):433-459.

36. Cohen J. Statistical power analysis for the behavioural sciences. 1988.

37. Rand WM. Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc*. 1971;66(336):846-850.

38. Mousavi SH, Hijmans JM, Rajabi R, Diercks R, Zwerver J, van der Worp H. Kinematic risk factors for lower limb tendinopathy in distance runners: A systematic review and meta-analysis. *Gait Posture*. 2019.

39. Gorton III GE, Hebert DA, Gannotti ME. Assessment of the kinematic variability among 12 motion analysis laboratories. *Gait Posture*. 2009;29(3):398-402.

40. Osis ST, Kobsar D, Leigh RJ, Macaulay CAJ, Ferber R. An expert system feedback tool improves the reliability of clinical gait kinematics for older adults with lower limb osteoarthritis. *Gait Posture*. 2017;58:261-267.

41. Macaulay CAJ, Osis ST, Clermont C, Ferber R. The use of real-time feedback to improve kinematic marker placement consistency among novice examiners. *Gait Posture*. 2017;58:440-445.

42. Chung M-J, Wang M-JJ. The change of gait parameters during walking at different

percentage of preferred walking speed for healthy adults aged 20--60 years. *Gait Posture*. 2010;31(1):131-135.

43. Brown C, Bowser B, Simpson KJ. Movement variability during single leg jump landings in individuals with and without chronic ankle instability. *Clin Biomech*. 2012;27(1):52-63.

44. Hamill J, Palmer C, Van Emmerik REA. Coordinative variability and overuse injury. *Sport Med Arthrosc Rehabil Ther Technol*. 2012;4(1):45.

45. Stergiou N, Decker LM. Human movement variability, nonlinear dynamics, and pathology: is there a connection? *Hum Mov Sci*. 2011;30(5):869-888.

46. Ahamed NU, Kobsar D, Benson LC, Clermont CA, Osis ST, Ferber R. Subject-specific and group-based running pattern classification using a single wearable sensor. *J Biomech*. 2019;84:227-233.

47. Benson LC, Clermont CA, Bošnjak E, Ferber R. The use of wearable devices for walking and running gait analysis outside of the lab: A systematic review. *Gait Posture*. 2018;63:124-138.

Tables

Table 1: Characteristics of participants in each injury group.

	Knee	Ankle/foot	Hip/Pelvis	Thigh	Lower leg	Healthy	Other injuries
Male-female	38-34	24-34	20-31	24-18	21-18	14-11	4-0
Age (years)	37.06±12.95	41.10±11.74	40.00±10.42	39.83±8.98	39.46±9.67	40.52±11.94	45.25±8.85
Height (cm)	173.53±9.31	170.35±9.70	171.07±13.34	172.95±7.78	171.54±9.58	172.87±9.14	178.08±10.23
Weight (kg)	69.80±12.27	71.78±17.12	70.53±13.32	71.00±13.23	71.24±11.46	70.81±10.68	78.15±11.22
Running speed (m/s)	2.49±0.26	2.46±0.31	2.49±0.31	2.55±0.24	2.52±0.28	2.59±0.25	2.67±0.22

*Table 2: Characteristics of the five identified subgroups. Significant differences identified: * p<0.05, **p<0.01. Mean ± standard deviation for continuous variables, count (portion in that cluster) for the injury locations.*

Subgroup	S1	S2	S3	S4	S5
Size	95	60	32	28	76
Male/Female	70-25**	22-38	21-11	15-13	16-60**
Age (years)	40.01±10.29	42.48±13.11	37.03±10.93	40.72±9.15	37.14±11.06
Height (cm)	176.45±7.74**	166.17±12.36**	177.32±8.14**	171.84±7.54	169.29±8.70**
Weight (kg)	73.13±12.17**	65.29±11.58**	74.67±12.37**	74.75±13.45	69.50±15.09
Speed (m/s)	2.65±0.25**	2.41±0.26**	2.63±0.23**	2.50±0.27	2.37±0.25**
Healthy	11 (11.6%)	6 (10.0%)	0 (0.0%)	3 (10.7%)	5 (6.6%)
Knee	25 (26.3%)	10 (16.7%)	14 (43.8%)	4 (14.3%)	17 (22.4%)
Ankle/foot	16 (16.8%)	18 (30.0%)	3 (9.4%)	7 (25.0%)	16 (21.1%)
Hip/pelvis	12 (12.6%)	10 (16.7%)	8 (25.0%)	7 (25.0%)	13 (17.1%)
Thigh	18 (19.0%)	7 (11.7%)	4 (12.5%)	4 (14.3%)	10 (13.1%)
Lower leg	12 (12.6%)	7 (11.7%)	3 (9.4%)	3 (10.7%)	14 (18.4%)

Figure legends

Figure 1: Dendrogram of the hierarchical cluster analysis. Linkage distance on the y-axis and individual runners on the x-axis. The five identified subgroups are identified by color. For clarity, not all runners are plotted on the dendrogram and the x-axis labels are omitted.

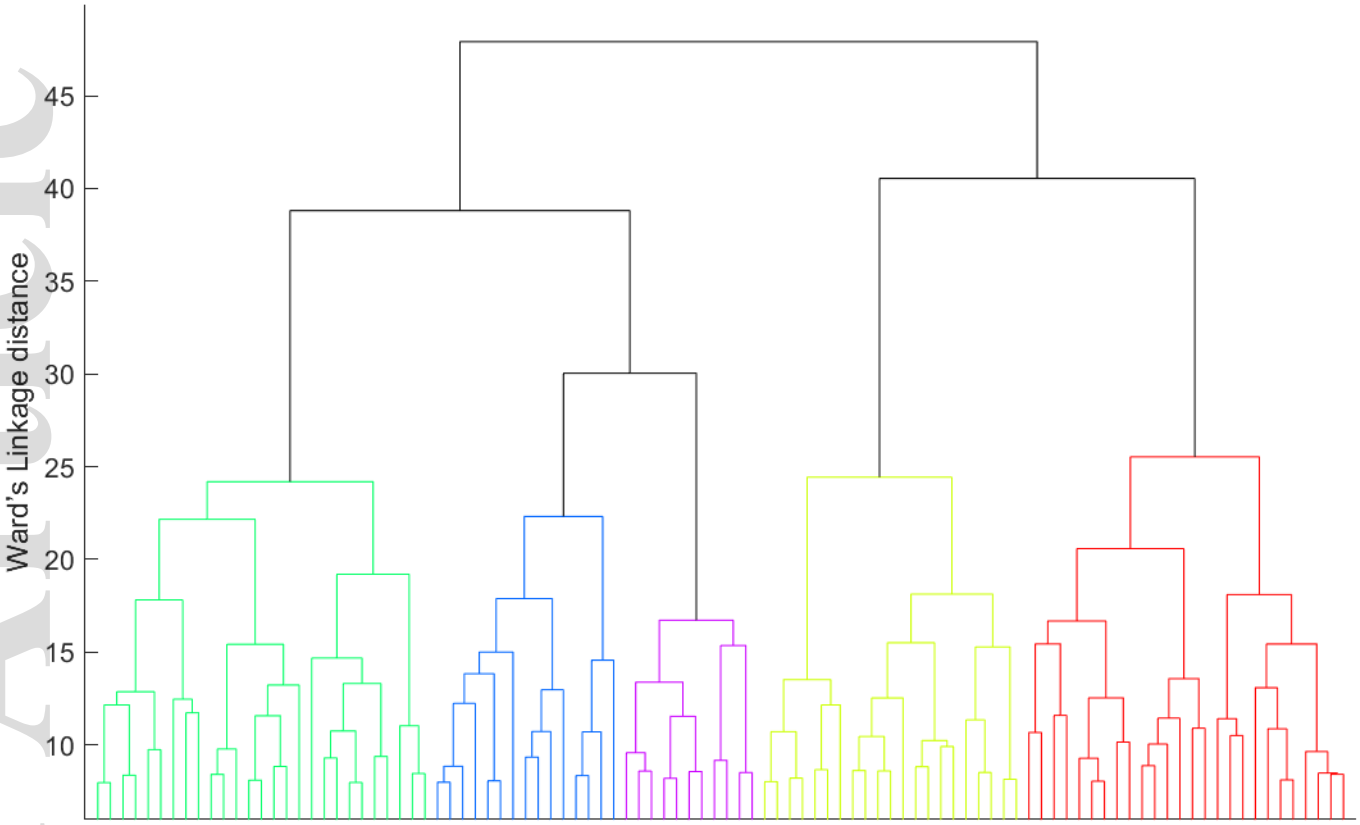
Figure 2: Boxplots highlighting the kinematics for subgroup 1 (left) to subgroup 5 (right). The vertical dashed line corresponds to the average of the study population and values have been normalized to have a mean of zero and standard deviation 1.

Appendices

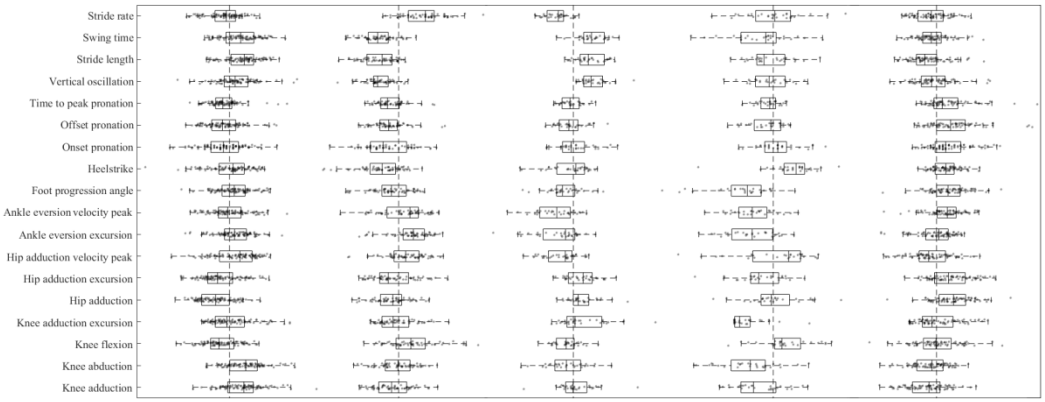
Table A1: Kinematic and functional variables used to describe each subject's running pattern:

	Joint	Variable	Description
Functional	Left/Right Side	Step Width	Side-to-side distance between left and right footsteps (m)
		Stride Rate	Number of foot strikes per minute (strikes per min)
		Stride Length	For-aft distance between left and right footfalls (m)
		Swing Time	Length of time (s) subject spent during the swing phase of gait.
		Stance Time	Length of time (s) subject spent during the stance phase of gait.
		Maximum Heel Whip	Difference between foot external rotation (deg) from toe-off to the point of maximal external rotation during swing phase.
		Vertical Oscilliation	Vertical oscillation of the center of mass (m) during complete gait cycle.
Kinematic	Left/Right Foot	Progression Angle	Angle of foot relative to direction of movement (deg) during stance phase of gait cycle.
		Heel Strike angle	Sagittal plane angle of foot at heel strike (deg).
	Left/Right Ankle	Peak Dorsiflexion	Maximum dorsiflexion angle (deg) experienced during complete gait cycle.
		Peak Eversion	Maximum eversion angle (deg) experienced during complete gait cycle.
		Time to peak pronation	The amount of time (% gait cycle) to reach peak pronation
		Eversion excursion	Difference between eversion angle at toe-off to peak eversion angle (deg).
		Peak eversion velocity	Maximum eversion angle (deg/s) subject experienced during complete gait cycle.
		Onset of Pronation	Point at which the foot reaches a pronated position (% of gait cycle)
		Offset of pronation	Point at which the foot leaves a pronated position (% of gait cycle)
	Left/Right Knee	Peak flexion Angle	Maximum knee flexion angle (deg) experienced during complete gait cycle.
		Peak adduction angle	Maximum knee adduction angle (deg) experienced during complete gait cycle.
		Adduction excursion	Distance (mm) of knee adduction excursion.
		Peak adduction velocity	Maximum knee adduction velocity (deg/s) experienced during complete gait cycle.
		Peak Abduction angle	Maximum knee abduction angle (deg) experienced during complete gait cycle.
		Abduction excursion	Difference between minimum and maximum knee abduction during stance (deg).
		Peak abduction velocity	Maximum knee abduction angle (deg) experienced during complete gait cycle.
	Left/Right Hip	Peak extension angle	Maximum hip extension angle (deg) experienced during complete gait cycle.
		Peak adduction angle	Maximum hip adduction angle (deg) experienced during complete gait cycle.
		Adduction excursion	Distance (mm) of hip adduction during gait cycle.

		Peak abduction velocity	Maximum hip adduction velocity (deg/s) experienced during complete gait cycle.
		Peak adduction velocity	Maximum hip adduction velocity (deg/s) experienced during complete gait cycle.
	Left/Right Pelvis	Peak Pelvic Drop	Maximum frontal plane angle of pelvis segment relative to horizontal (deg) experienced during complete gait cycle.
		Pelvic Drop Excursion	Difference between minimum and maximum pelvic drop during stance phase (deg).
		Peak Pelvic Drop Velocity	Maximum pelvic drop angle velocity (deg/s) experienced during stance.



sms_13624_f1.tif



sms_13624_f2.tif