Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Injury attenuates speed-dependent stride length scaling in runners: evidence for a protective locomotor control adaptation from a multi-method analysis

Kofi Nyantakyi Appiah

nyantakyi.appiah@gmail.com

Wesley College of Education

Nathanael Adu

Mampong Technical College of Education

Neelam Sharma

Lovely Professional University

Lawrence Rudolph Essuman

St. Monica's College of Education

Research Article

Keywords: Running-related injury, Locomotor control, Motor adaptation, Stride length scaling, Bayesian hierarchical modelling, Machine learning

Posted Date: June 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9528625/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background: How running-related injury alters locomotor control remains poorly understood. Within the motor adaptation to pain framework, the relationship between running speed and stride length can be conceptualised as a quantifiable control law. We hypothesised that injury reorganises this mapping, producing a shallower speed-stride length slope that limits mechanical loading at higher velocities.

Methods: We conducted a secondary analysis of an open biomechanics dataset (N = 1,798 runners) containing treadmill running kinematics. Discrete gait parameters (stride length, step width, stride rate, swing time) were extracted from 1,556 running trials. Linear mixed models and Bayesian hierarchical inference estimated the speed × injury interaction. Bootstrap resampling (1,000 iterations) and a random forest classifier with SHAP values provided complementary robustness checks and feature importance.

Results: The linear mixed model revealed a significant speed × injury interaction (β = −0.0416, 95% CI: −0.072 to −0.010; p = 0.009), indicating that injured runners increase stride length 0.042 m less per 1 m/s increase in speed compared to healthy runners. Bootstrap resampling showed 90.2% of interaction estimates were negative (95% CI: −0.0706 to 0.0147). The Bayesian model gave a 99.5% posterior probability that the true interaction is negative. Random forest identified speed and stride length as the top predictors of injury status.

Conclusions: Running-related injury attenuates the normal speed-dependent increase in stride length, consistent with a protective motor adaptation that reduces mechanical loading at higher speeds. These findings provide a biomechanical rationale for gait retraining interventions targeting stride length and cadence in injured runners.

Trial registration: Not applicable

Background

The ability to walk and run is fundamental to human movement, and understanding the control, organisation, and adaptation of locomotion is central to kinesiology, rehabilitation, and sport science [1]. Quantitative three-dimensional gait analysis using motion capture systems is the gold standard for investigating the biomechanical aetiology of musculoskeletal injuries, determining treatment options, and evaluating rehabilitation progression [2, 3]. Yet, despite decades of technological advances in camera systems, marker sets, and computational methods, the fundamental approach to gait analysis has remained largely unchanged, with individual laboratories operating in relative isolation [4]. While several large-scale gait databases exist (e.g., OU-ISIR Gait Database, OU-MVLP dataset), they are primarily focused on biometrics or walking in healthy adults. The present dataset is unique in its focus on lower-limb running injuries and provides a valuable resource for hypothesis-driven secondary analyses. Nonetheless, the historical fragmentation has resulted in a critical lack of open-access benchmark datasets for running-related injury, limiting generalizability [5, 6]. Consequently, the vast majority of biomechanical studies are underpowered, focus on a small number of discrete variables, and lack the transparency required for meaningful replication [7].

To address this critical gap, [4] recently published an open biomechanics dataset of unprecedented scale: kinematic data from 1,798 healthy and injured, young and older adults during treadmill walking and/or running at a range of self-selected speeds. The dataset includes raw three-dimensional marker trajectories, detailed participant demographics, injury history, and calculated spatiotemporal parameters. This resource offers a

unique opportunity to move beyond small-sample descriptive studies and test sophisticated, theory-driven hypotheses about the control and organisation of human movement under pathological conditions. The present study addresses these issues by using multilevel models and sensitivity analyses across the speed spectrum.

A leading theoretical framework for understanding how injury alters movement is the motor adaptation to pain model [8]. This theory posits that, rather than a simple reflex inhibition or excitation, pain induces a complex reorganisation of motor commands to redistribute activity within and between muscles, thereby protecting the injured tissue while maintaining task performance [8]. This framework has been supported in various patient populations, including individuals with knee osteoarthritis [9], patellofemoral pain [10], and anterior cruciate ligament injury [11]. In each case, patients show changes in movement and coordination consistent with strategies that unload the painful or unstable structures. However, most of these studies are limited to small, homogeneous samples (often fewer than 50 participants per group), which prevents a rigorous, quantitative evaluation of how injury alters the fundamental relationship between gait parameters, such as the scaling of stride length with running speed. The present study directly addresses this gap by using the large, heterogeneous dataset of [4] and by employing open, reproducible analytical methods.

The relationship between running speed and stride length is a basic but informative feature of human locomotion. In control terms, this speed-stride length relationship can be viewed as a low-dimensional control law that links task demand (speed) to an effector variable (stride length). This makes it an attractive, quantifiable target for studying how injury reorganises locomotor control. In healthy runners, stride length increases approximately linearly with speed (e.g., [12], reported an $r^2 = 0.80$, $p < 0.0001$ for this relationship), reflecting increased propulsive impulse and longer flight phases.. This relationship is not fixed; it can be modified by fatigue, training, and, importantly, injury. The motor adaptation to pain framework would predict that injured runners adopt a shallower slope of stride length against speed, meaning they increase stride length less than healthy runners when moving faster. Such an adaptation would reduce vertical impact forces and joint moments at higher speeds, thereby protecting the injured tissue [8, 12, 13]. Step width, a parameter more closely associated with frontal plane balance, might not show the same adaptation, because it is less directly linked to vertical loading [14, 15] and has been more strongly associated with lateral stability and asymmetry [14]. Here, we explicitly treat the speed-stride length relationship as a control law, and ask whether injury alters the slope of this mapping rather than simply shifting mean stride length.

The present study leverages the large, heterogeneous [4] dataset to test the specific hypothesis derived from motor adaptation theory: injury attenuates the normal, speed-dependent increase in stride length during running. We hypothesise that injured runners will exhibit a shallower slope of stride length against speed compared to healthy runners, whereas step width will show no such interaction. This dissociation would provide evidence for a targeted, functionally relevant motor adaptation rather than a generalised cautious gait pattern.

To test this hypothesis with the highest methodological rigour, we employ a combination of advanced statistical techniques that directly address the nested, continuous nature of gait data. First, we use linear mixed models (LMM) with random intercepts to account for repeated measures within subjects, which is the modern gold standard for analysing biomechanical data with multiple trials per participant [16]. Second, we perform bootstrap resampling [17] at the participant level to obtain robust confidence intervals for the interaction term, circumventing distributional assumptions. Third, we fit a Bayesian hierarchical model using weakly informative priors to compute the posterior probability that the interaction is negative, providing a direct probabilistic

statement about the hypothesis [18, 19]. Fourth, we train a random forest classifier (Liaw & Wiener, 2002) to predict injury status from gait features, and we use SHAP (SHapley Additive ExPlanations) values [20] to interpret which features most strongly discriminate injured from healthy runners. To our knowledge, this combination of frequentist, bootstrap, Bayesian, and machine learning techniques has not been widely applied in clinical gait analysis. This multi-method approach helps ensure that any conclusion is not an artefact of a single statistical framework.

Methods

Dataset and participants

This study is a secondary analysis of the Running Injury Clinic kinematic dataset [4], which is openly available under a CC-BY 4.0 licence. The dataset contains three-dimensional marker trajectory data from 1,798 participants (mean age 34.9 ± 12.4 years; 53% female) who walked or ran on a motorised treadmill. Participants were either healthy (n = 396) or had a lower-extremity running-related injury (n = 1,402) diagnosed by a licensed medical professional. In the original dataset, 'injury' is defined pragmatically as running-related lower limb pain or dysfunction that affects training (for example, continuing to train in pain, reduced training volume/intensity, or missing at least two consecutive workouts; Ferber et al., 2024), and encompasses common diagnoses such as patellofemoral pain, Achilles tendinopathy, iliotibial band syndrome, and plantar fasciopathy. No standardised clinical examination was imposed beyond the original clinic's routine diagnostic process; the dataset reflects naturalistic clinical data. Inclusion criteria for the present secondary analysis were: (i) running trial present, (ii) non-missing speed and stride length values. No other exclusions were applied. All procedures were approved by the University of Calgary's Conjoint Health Research Ethics Board (Ethics IDs: E-21705, E-22194, E-24339), and all participants gave written informed consent.

For the present study, we included only running trials (self-selected speeds ranging from 1.1 to 4.9 m/s). Walking trials were excluded because the primary hypothesis concerned speed-dependent adaptation during running, where impact forces are higher and injury-related adaptations are more likely to manifest [12]. Treating speed as self-selected rather than experimentally fixed reflects an ecologically valid aspect of locomotor control, and our modelling approach leverages this variability by analysing how stride length scales with the continuously varying speed. The following exclusion criteria were applied: (i) walking trials (identified by the 'mode' field in the JSON data), (ii) trials with missing speed values, (iii) trials with missing stride length values. No other exclusions were made. After these exclusions, the final analytic sample consisted of 1,556 trials from 1,272 unique participants (848 injured, 424 healthy).

Data extraction and pre-processing

Data extraction and pre-processing

The original dataset is stored in JSON format, with each file containing raw marker trajectories, participant metadata, and pre-computed discrete gait variables. For each participant, the discrete running variables are located in the *dv_r* field, which includes stride length, step width, stride rate, swing time, stance time, peak joint angles, and other spatiotemporal parameters. We extracted these variables using the *RcppSimdJson* package [21] in R version 4.5.2 [22], which provides fast parsing of large JSON files. A limitation of using pre-computed

discrete variables is the loss of temporal waveform information (e.g., joint angle time series), which may obscure subtle changes in coordination. Additionally, discrete variable extraction can introduce investigator bias if not standardised [23]. We mitigated this by using the dataset's provided variables without modification and by averaging left/right sides.

After extraction, we merged the discrete gait data with the participant metadata (age, gender, body mass index [BMI], injury status, and running speed) using the unique filename identifier. Because the original metadata contained duplicate rows for some (sub_id, filename) combinations, we resolved duplicates by retaining the first occurrence after verifying that duplicate rows were identical in all other fields. In the rare case of discrepancies (e.g., different injury codes), the row with more complete injury data was prioritised. This approach is unlikely to bias results because duplicates arose from administrative errors, not systematic variation. For each trial, values from the left and right sides were averaged to obtain a single value per variable, thereby reducing dimensionality and avoiding side-specific artefacts [24]. The final wide-format dataset contained one row per trial with columns for participant ID, speed, injury status, age, gender, BMI, and the seven gait variables of interest: stride length, step width, stride rate, swing time, stance time, peak knee flexion angle, and peak hip extension angle.

Statistical analysis

All statistical analyses were performed in R version 4.5.2 [22]. The analysis pipeline was designed to test the hypothesis that injury attenuates the speed-dependent increase in stride length, using multiple complementary methods: (1) linear mixed model (LMM) with random intercept; (2) non-parametric bootstrap resampling; (3) Bayesian hierarchical model; (4) random forest classifier with SHAP values.

Linear-mixed model

To account for repeated measures (multiple trials per participant) and the nested structure of the data, we fit a linear mixed model (LMM) with a random intercept for each participant using the *lme4* package [25]. The model was specified as:

$$STRIDE_{LENGTHij} = \beta 0 + \beta 1 \cdot \ speedij + \beta 2 \cdot \ injuredi + \beta 3 \cdot \ (speedij \times \ injuredi) + ui + \epsilon ij$$

Where $ui \sim N(0,\sigma u2)$ is the participant-specific random intercept, and $\epsilon ij \sim N(0,\sigma 2)$ is the residual error. A random slope for speed was attempted but proved unidentifiable because 96% of participants contributed only one trial, leading to a singular fit (estimated random slope variance near zero). This is a common issue in mixed models with limited repeated measures [25]. Therefore we retained only the random intercept. The fixed effects of primary interest were the main effect of speed (β_1) and the speed × injury interaction (β_3). The interaction coefficient directly quantifies whether the change in stride length per unit increase in speed differs between injured and healthy runners. P-values were obtained using Satterthwaite's approximation for degrees of freedom [26].

Bootstrap resampling

To obtain a robust confidence interval for the interaction term without relying on parametric assumptions, we performed non-parametric bootstrap resampling [17]. We resampled participants with replacement (1,000

iterations), refitting a simple linear model (stride length ~ speed × injury) on each bootstrap sample. The 95% percentile confidence interval was calculated from the distribution of the bootstrap interaction coefficients. Percentile bootstrap confidence intervals for interaction terms can be conservative and may have inflated Type I error rates in small samples [27]. In the present large sample (1,556 trials), this concern is mitigated, but the interval should be interpreted alongside the proportion of negative estimates. We also computed the proportion of bootstrap estimates that were negative, which provides an empirical measure of evidence against the null hypothesis of no effect.

Bayesian hierarchical model

As a Bayesian complement to the frequentist LMM, we fitted a hierarchical model using the *rstanarm* package [28]. The model had the same fixed and random effects structure as the LMM, but with weakly informative priors: Normal(0, 2) for the fixed effects (including the interaction) and Normal(1.5, 1) for the intercept. The prior for the residual standard deviation was Exponential(1). These priors were chosen to be weakly informative: Normal(0, 2) allows plausible effect sizes up to ± 4 standard deviations; Normal(1.5, 1) reflects the typical stride length of ~ 1.5 m; Exponential(1) is a standard prior for positive scale parameters. A sensitivity analysis using more diffuse priors (Normal(0, 5)) yielded a posterior probability of interaction < 0 = 0.993, consistent with the main result (see OSF repository). Four Markov chains were run for 2,000 iterations each, with the first 1,000 discarded as warm-up. Convergence was assessed by the potential scale reduction factor (Rhat < 1.05) and effective sample size [29]. The primary Bayesian quantity of interest was the posterior probability that the speed × injury interaction coefficient was negative, which directly answers the question: “How likely is the hypothesised protective adaptation, given the data and priors?”

Random forest and explainable machine learning

To identify which gait features most strongly discriminate injured from healthy runners, we trained a random forest classifier [30] using the *randomForest* package. The predictor variables were: stride length, step width, stride rate, swing time, speed, age, and BMI. The outcome was binary injury status (healthy vs. injured). The forest comprised 500 trees, with the number of predictors considered at each split set to the square root of the total number of predictors (default). Variable importance was assessed by the mean decrease in classification accuracy when a given predictor was permuted [31]. A limitation of this permutation-based importance measure is that it can be affected by correlated predictors and measurement scale [32]. However, the random forest was run on standardised continuous variables, and the overall ranking was consistent with SHAP values.

To interpret the random forest’s predictions, we computed SHAP (SHapley Additive ExPlanations) values using the *DALEX* package [33]. SHAP values provide a unified measure of feature importance that is both locally accurate and globally consistent [34]. SHAP values were calculated for all observations using the predict_parts function with 25 background samples (B = 25). While some guidelines recommend 100–1000 samples, we used 25 for computational efficiency (1,556 observations × 7 features). A stability check on a subset with B = 100 yielded highly correlated SHAP values (r > 0.95), indicating that the approximation was reliable. The resulting summary plot shows the direction and magnitude of each feature’s contribution to the predicted injury status.

Software and reproducibility

All analyses were performed in R version 4.5.2 [22] with the following packages: *data.table* [35], *lme4* [25], *lmerTest* [26], rstanarm [28], *randomForest* [31], *DALEX* [33], and *ggplot2* [36]. Specific package versions are documented in the OSF repository (session-info.txt). A random seed (123) was set prior to all analyses involving random number generation (bootstrap resampling, Bayesian MCMC, random forest). The complete data extraction and analysis scripts, along with the cleaned data and all output files (model summaries, bootstrap distributions, Bayesian posterior draws, figures, and tables), are publicly available on the Open Science Framework (OSF; https://osf.io/2ej9k). The original dataset is also openly accessible [4]. This transparency allows full replication and extension by other researchers.

Results

Participant and trial characteristics

After data cleaning and exclusion of walking trials, the final analytic sample comprised 1,556 running trials from 1,272 unique participants. Walking trials were identified by the 'mode' field in the extracted data ('walk' vs. 'run'). Of these, 848 participants (66.7%) were classified as injured and 424 (33.3%) as healthy. The mean running speed was 2.89 m/s (SD = 0.58, range 1.1–4.9 m/s), and mean stride length was 2.01 m (SD = 0.21). Injured and healthy groups did not differ significantly in age (mean 36.2 vs. 35.7 years, $p = 0.34$) or body mass index (BMI; mean 23.8 vs. 23.5 kg/m², $p = 0.41$), but injured participants had a slightly lower mean running speed (2.84 vs. 2.97 m/s, $p = 0.02$). Demographic and injury characteristics are fully documented in the original dataset description [4]. In line with the original dataset description, the injured group predominantly comprised common running-related overuse conditions of the lower limb (for example, patellofemoral pain, Achilles tendinopathy, iliotibial band syndrome), rather than acute traumatic injuries [4].

Linear-mixed model: speed-stride length interaction

To test the hypothesis that injury modifies the relationship between running speed and stride length, we fitted a linear mixed model with a random intercept for participant. The model explained a substantial proportion of variance (conditional $R^2 = 0.78$). Table 1 presents the fixed effects (intercept, speed, injured status, and their interaction) from this model.

There was a strong main effect of speed ($\beta = 0.1776$, SE = 0.0135, $p < 0.001$), confirming that all participants increased stride length as running speed increased. The main effect of injury status was not significant ($\beta = -0.0337$, SE = 0.0271, $p = 0.213$), indicating that injured runners did not have systematically shorter or longer strides across all speeds. Critically, the speed × injury interaction was statistically significant ($\beta = -0.0416$, SE = 0.0159, $p = 0.009$). This interaction coefficient quantifies the difference in the slope of stride length against speed between injured and healthy runners. For healthy runners, each 1 m/s increase in speed was associated with a 0.178 m increase in stride length; for injured runners, the increase was only 0.136 m per 1 m/s (0.178–0.042).

Table 1
Fixed-effects from linear-mixed model of stride length (random intercept for participant)

Term	Estimate (m)	Standard Error	df	t value	p value
Intercept	1.7226	0.0224	1027.6	76.76	< 0.001
Speed (m/s)	0.1776	0.0135	627.6	13.20	< 0.001
Injured (yes vs. no)	−0.0337	0.0271	1015.5	−1.25	0.213
Speed × Injured	−0.0416	0.0159	635.8	−2.62	**0.009**

Note. The interaction coefficient indicates that injured runners increase stride length 0.042 m less per 1 m/s increase in speed compared to healthy runners.

Bootstrap resampling

To assess the robustness of the interaction without parametric assumptions, we performed non-parametric bootstrap resampling (1,000 iterations) at the participant level. The distribution of the bootstrap interaction coefficients is shown in the inset histogram of Fig. 1. The 95% percentile confidence interval for the speed × injury interaction was [−0.0706, 0.0147]. Notably, 90.2% of the bootstrap estimates were negative, indicating that the preponderance of resampled datasets supported a negative interaction (i.e., a shallower slope in injured runners). This bootstrap evidence, while not reaching the conventional 95% threshold for a two-sided interval, provides moderate support for the hypothesised effect [17]. However, the proportion of bootstrap estimates below zero should be interpreted with caution, as it can be conservative and does not directly correspond to a p-value [27].

Figure 1 Predicted stride length versus running speed from the linear mixed model (solid lines, 95% confidence bands). Inset: bootstrap distribution of the speed × injury interaction coefficient (vertical dashed line = zero; 90.2% of estimates < 0).

Bayesian hierarchical model

The Bayesian hierarchical model, fit with weakly informative priors, converged well (all Rhat = 1.00, effective sample sizes > 1300 for fixed effects). Visual inspection of trace plots showed good mixing, and posterior predictive checks indicated adequate model fit (see OSF repository for diagnostic plots). Table 2 summarises the bootstrap and Bayesian results for the speed × injury interaction. The posterior mean of the interaction coefficient was − 0.0415 (SD = 0.0159), nearly identical to the frequentist estimate. Most importantly, the posterior probability that the true interaction coefficient is negative was 0.995 (99.5%). In Bayesian terms, this means that given the data and the priors, there is a 99.5% probability that injury attenuates the speed-dependent increase in stride length [37]. The 95% credible interval for the interaction was [−0.072, − 0.011], which excludes zero.

Table 2
Bootstrap and Bayesian results for the speed × injury interaction.

Method	Estimate / Statistic	Value
Bootstrap (95% CI)	Confidence interval	[–0.0706, 0.0147]
Bootstrap	Proportion negative	90.2%
Bayesian	Posterior mean (SD)	–0.0415 (0.0159)
Bayesian	95% credible interval	[–0.072, – 0.011]
Bayesian	Posterior probability < 0	99.5%

Note. Bayesian estimates are based on 4 chains, 2,000 iterations each (1,000 warm-up). Priors: Normal(0, 2) for fixed effects; Normal(1.5, 1) for intercept.

Random forest variable importance and SHAP values

To identify which gait features best discriminated injured from healthy runners, we trained a random forest classifier. The model achieved an out-of-bag accuracy of 68.2%, with sensitivity = 0.72, specificity = 0.65, and F1-score = 0.69 (confusion matrix available in the OSF repository). These metrics indicate moderate predictive performance, with slightly better detection of injured runners. Table 3 presents variable importance measured by mean decrease in accuracy when a predictor was permuted. Speed was the most important predictor (mean decrease in accuracy = 35.9), followed by stride length (29.2) and age (26.1). Step width, stride rate, swing time, and BMI showed lower importance (10.1–23.7). These results indicate that the combination of speed and stride length carries the strongest signal for distinguishing injured from healthy runners, consistent with the interaction effect observed in the LMM.

Table 3
Random forest variable importance (mean decrease in accuracy) for predicting injury status.

Variable	Mean Decrease in Accuracy
Speed	35.9
Stride length	29.2
Age	26.1
Stride rate	23.7
BMI	18.6
Swing time	13.1
Step width	10.1

Note. Higher values indicate greater importance for classification. The model used 500 trees; out-of-bag accuracy = 68.2%.

SHAP values provided a more detailed, instance-level explanation of the random forest's predictions [38]. Figure 2 displays the SHAP summary plot. Higher speed and longer stride length strongly push the prediction toward "healthy", whereas lower speed and shorter stride length push the prediction toward "injured". Age also contributed, with older age associated with a higher probability of injury. Step width and swing time had smaller,

more variable contributions. While longer stride lengths are generally associated with higher running speed, excessive stride length (overstriding) can increase injury risk in some contexts [39]. However, in the present dataset, longer stride length at a given speed was protective, possibly because our participants increased stride length in a controlled, efficient manner without overstriding. This pattern reinforces the primary finding: injured runners tend to have a reduced stride length for a given speed, and the model uses this relationship to discriminate groups.

Sensitivity analyses

Sensitivity analyses were conducted to assess the robustness of the primary finding. First, excluding trials with extreme speeds (< 2.0 m/s or > 4.0 m/s) reduced the sample to 222 trials (182 participants); the interaction was not statistically significant (β = − 0.0695, SE = 0.0500, 95% CI: − 0.168 to 0.029, p = 0.166), although the coefficient remained negative and larger in magnitude. The wide confidence interval reflects the substantially reduced sample size (n = 222) and does not rule out a meaningful negative effect. Second, adding body mass index (BMI) as a covariate to the main model (n = 1,556 trials) did not attenuate the interaction; the coefficient was − 0.0440 (SE = 0.0158, p = 0.005). These results suggest that the protective adaptation is robust to adjustment for BMI but may be more variable at extreme speeds or underpowered in the reduced sample. The convergence of the main LMM, bootstrap, Bayesian, and ML analyses supports the presence of a genuine, though not universally consistent, effect.

Fourth, we reran the main linear mixed model including gender as a fixed effect. The speed × injury interaction remained significant (β = − 0.0397, SE = 0.0155, p = 0.010), and gender itself was a significant predictor (β = 0.1816, p < 0.001). The magnitude and significance of the interaction were virtually unchanged from the main model, indicating that the primary finding is not confounded by sex (see OSF repository for full output).

Table 4 summarises the sensitivity analyses.

Table 4
Sensitivity analyses of the speed × injury interaction for stride length.

Analysis	N (trials)	Interaction coefficient (β)	Standard error	p value
Main model (random intercept)	1,556	−0.0416	0.0159	0.009
+ Adjustment for BMI	1,556	−0.0440	0.0158	0.005
Excluding extreme speeds (< 2.0 or > 4.0 m/s)	222	−0.0695	0.0500	0.166
+ Adjustment for gender	1,556	−0.0397	0.0155	0.010

Note. The main model is repeated from Table 1 for comparison. BMI = body mass index. Extreme speed exclusion reduced the sample size substantially, which increased the standard error and rendered the interaction non-significant, though the coefficient remained negative and larger in magnitude. Gender adjustment left the interaction essentially unchanged.

Summary of key results

In summary, three complementary statistical approaches provided convergent evidence that injury attenuates the normal speed-dependent increase in stride length:

1. Frequentist LMM (Table 1): Significant interaction (p = 0.009), with a 0.042 m per 1 m/s reduction in slope for injured runners, indicating an attenuated speed-stride length control law in the injured group.
2. Bayesian model (Table 2): 99.5% posterior probability that the interaction is negative, with a 95% credible interval excluding zero, providing strong probabilistic evidence for the hypothesised protective adaptation.
3. Bootstrap (Fig. 1, Table 2): 90.2% of resampled interaction estimates were negative, with a 95% CI that included zero but was predominantly negative, supporting a small-to-moderate but consistent effect.
4. Random forest + SHAP (Table 3, Fig. 2): Speed and stride length were the top two features discriminating injured from healthy runners, and SHAP values confirmed that shorter stride length at a given speed is associated with injury, echoing the primary LMM/Bayesian pattern.

These results support the hypothesis that injured runners adopt a protective motor strategy that limits stride length when running faster, thereby potentially reducing impact loading on the injured tissue [39].

Discussion

Summary of principal findings

This secondary analysis of a large, open biomechanics dataset (N = 1,798) tested the hypothesis that running-related injury attenuates the normal speed-dependent increase in stride length. Using four complementary statistical methods, linear mixed modelling, bootstrap resampling, Bayesian hierarchical modelling, and random forest with SHAP values, we obtained convergent evidence supporting the hypothesis. The linear mixed model revealed a significant speed × injury interaction (β = − 0.0416, p = 0.009), indicating that injured runners increase stride length by 0.042 m less per 1 m/s increase in speed compared to healthy runners. Bootstrap resampling showed that 90.2% of interaction estimates were negative, and the Bayesian model yielded a 99.5% posterior probability that the true interaction is negative. Random forest variable importance identified speed and stride length as the top two features discriminating injured from healthy runners, and SHAP values confirmed that lower speed and shorter stride length increase the predicted probability of injury. These results provide strong, multi-method evidence that injury leads to a protective motor adaptation: a shallower stride length-speed slope that likely reduces impact loading at higher running speeds. Each method has limitations: LMM assumes correct random effect structure; bootstrap intervals can be conservative; Bayesian results depend on prior specification; SHAP values are approximations sensitive to correlated features. However, the convergence across methods reduces the likelihood that the finding is artefactual.

Interpretation within the motor adaptation to pain framework

The motor adaptation to pain model posits that pain or the threat of pain induces a reorganisation of motor commands to protect the injured tissue while maintaining task performance [8]. In running, vertical ground reaction forces and joint moments increase non-linearly with speed [40]. A straightforward protective strategy is to limit the increase in stride length at higher speeds, because shorter strides reduce vertical impact forces and peak joint moments [13]. Our finding that injured runners exhibit a reduced stride length-speed slope is therefore consistent with this protective strategy. Notably, the absence of a significant speed × injury interaction for step width (p = 0.973) suggests that the adaptation is specific to a parameter directly linked to vertical loading, rather than a generalised cautious gait. This specificity argues against a non-specific “fear of movement” explanation [41] and instead supports a functionally targeted reorganisation of the locomotor pattern.

More broadly, these findings illustrate a context-dependent, load-specific reorganisation of locomotor control rather than a uniform cautious pattern. In other words, the control law that links speed to stride length is selectively flattened, while parameters less directly tied to vertical loading (for example, step width) remain unchanged. This kind of selective remapping may be a general principle of motor adaptation to pain across tasks and body regions, whereby the nervous system modifies specific control relationships that most efficiently reduce load on threatened tissues.

Comparison with previous literature

Our results align with several smaller-scale studies that have reported altered stride length or stride frequency in injured runners. For example, runners with patellofemoral pain have been observed to adopt shorter strides and lower cadence compared to healthy controls [39, 42]. Similarly, individuals with Achilles tendinopathy exhibit reduced stride length and increased contact time, interpreted as a load-management strategy [43]. However, these previous studies typically compared groups at a single, fixed speed or over a narrow speed range, and most had sample sizes under 50 participants [44, 45]. Our study extends this literature by demonstrating that the *slope* of the speed-stride length relationship, not just the mean stride length is modified by injury. This is an important theoretical distinction: it suggests that injured runners do not simply adopt a fixed “shorter stride” pattern; rather, they alter the control law linking speed to stride length. Such a context-dependent adaptation is precisely what the motor adaptation to pain framework would predict [8].

Our use of a large, heterogeneous sample (1,272 unique runners) also addresses the long-standing criticism that gait analysis laboratories often operate in isolation, producing results with limited generalizability [7, 46]. The sample included runners aged 18–65 years (mean 36.2), with a range of running experience (1–30 years), and a variety of injury types. Demographic breakdown by sex, age, and injury location is available in the original dataset [4] and in the OSF repository By analysing data collected over several years across many injury types and running abilities, we increase confidence that the observed effect is not an artefact of a specific clinic or injury subtype. The open nature of the dataset and our own code further enhances reproducibility, a value increasingly emphasised in biomechanics [46].

Methodological strengths

The multi-method statistical approach is a major strength. Most biomechanical studies rely on a single frequentist model (e.g., ANOVA or simple linear regression), which can be sensitive to distributional assumptions and multiple comparisons. Here, we corroborated the LMM finding with a non-parametric bootstrap that makes no assumptions about normality [47]. We then added a Bayesian model that provides a direct probabilistic statement about the hypothesis, a 99.5% posterior probability that the interaction is negative, which is more intuitive than a p-value [48]. Finally, the random forest and SHAP analysis offered an interpretable machine learning perspective, confirming that speed and stride length are the dominant features separating injured from healthy runners [49]. SHAP explanations can be sensitive to feature correlations and noise; however, the dominance of speed and stride length was consistent across multiple runs and with the LMM results, increasing confidence. The convergence of four independent methods substantially reduces the likelihood that the finding is a statistical artefact.

Another strength is the use of an open, pre-registered analysis pipeline (OSF; https://doi.org/10.17605/OSF.IO/2EJ9K). All data (cleaned CSV files), analysis scripts (R code, numbered 00–

06), and outputs (model RDS objects, bootstrap distributions, Bayesian posterior draws, figures, and tables) are publicly available on the OSF (https://osf.io/2ej9k). The repository also includes a session-info.txt file listing package versions and a README with reproduction instructions. This transparency aligns with the journal's commitment to methodological soundness and reproducibility.

Limitations

Several limitations must be acknowledged. First, the original data were collected on a treadmill, which may not fully represent over-ground running due to differences in belt compliance, air resistance, and perceptual feedback [50]. However, treadmill running is widely used in clinical and research settings, and the relative differences between groups are likely preserved. Second, running speed was self-selected, not standardised across participants. This means that the observed interaction could partly reflect differences in preferred speed between injured and healthy groups. Nevertheless, our analysis explicitly modelled speed as a continuous covariate and tested the interaction, which accounts for differences in speed distributions. Importantly, the key result concerns the *slope* of stride length as a function of speed within individuals, not only differences in preferred speeds between groups. By focusing on the interaction term, we specifically quantify how the speed-stride length control law differs between injured and healthy runners over the shared range of speeds. Third, the dataset includes a heterogeneous mix of injury types (e.g., patellofemoral pain, Achilles tendinopathy, iliotibial band syndrome). While this increases generalizability, it also means we cannot determine whether specific injuries produce different adaptations. Subgroup analyses by injury type (e.g., comparing knee vs. ankle injuries) would be a valuable extension. Our aim in the present study was therefore to characterise a generic motor adaptation pattern associated with lower limb running-related injury, rather than to attribute distinct control laws to individual diagnostic entities. Future work that stratifies by diagnosis and pain severity will be needed to test whether specific conditions exhibit qualitatively different modifications of the speed-stride length control law.

Fourth, we used only discrete summary variables (stride length, step width, etc.) provided in the dataset, not the raw joint angle waveforms. Consequently, we could not perform waveform-based analyses such as functional principal component analysis (fPCA) or statistical parametric mapping (SPM), which might reveal more nuanced differences in the time course of joint coordination. Similarly, uncontrolled manifold (UCM) analysis, a direct measure of motor synergy, requires joint angle time series. Future work extracting the full waveforms from the raw marker data could address these questions. Fifth, the cross-sectional design precludes causal inferences. We cannot determine whether the observed gait pattern preceded the injury (a risk factor) or emerged as a consequence of injury (an adaptation). Longitudinal studies tracking runners before and after injury are needed to establish causality.

We included sex as a covariate in a sensitivity analysis; the speed × injury interaction remained virtually unchanged ($\beta = -0.0397$, $p = 0.010$) and sex itself was a significant predictor of stride length (males had longer strides). However, the dataset only recorded binary sex (53% female) and did not assess gender identity. Future research with more nuanced measures could explore whether motor adaptation differs between sexes or gender groups. Sixth, despite the large sample, the bootstrap confidence interval for the interaction term [−0.0706, 0.0147] included zero, indicating that the effect is not extremely large. This is consistent with a small-to-moderate effect size that might be obscured by heterogeneity in injury types and running speeds. Formal heterogeneity indices (e.g., I^2) are not directly applicable to a single study, but the variation in effect across injury types and speed ranges is evident from the sensitivity analysis that excluded extreme speeds. Nonetheless, the Bayesian posterior probability (99.5%) and the proportion of negative bootstrap estimates (90.2%) both point

strongly in the same direction. We therefore interpret the evidence as moderate to strong for a negative interaction, while acknowledging that the effect is not massive.

It is worth noting that our bootstrap confidence interval for the interaction term marginally overlaps zero, despite the strong evidence from the mixed model and Bayesian posterior. This pattern is consistent with a small-to-moderate effect size expressed in a heterogeneous sample, where resampling variability is substantial. Crucially, 90.2% of bootstrap estimates were negative and the Bayesian analysis yielded a 99.5% posterior probability that the interaction is negative, and the machine-learning model independently exploited the same speed-stride length pattern to distinguish injured from healthy runners. Bootstrap resampling assumes that the sample is representative of the population; residual heterogeneity in injury types and running speeds may cause the bootstrap distribution to be overly variable, which could explain the wide CI. Taken together, these convergent lines of evidence support a genuine attenuation of the speed-stride length control law in injured runners, even if the effect is not large in absolute terms.

Clinical and practical implications

Our findings suggest that injured runners spontaneously adopt a stride length-speed relationship that reduces mechanical loading at higher speeds. This has implications for rehabilitation and injury prevention. First, gait retraining interventions that explicitly target stride length and cadence, such as increasing step rate by 5–10%, have been shown to reduce impact forces and alleviate symptoms in conditions like patellofemoral pain [51, 52]. Our results provide a theoretical rationale for such interventions: they may reinforce or mimic a natural protective adaptation. Second, clinicians could use speed-stride length plots (as in Fig. 1) to monitor whether an injured runner is returning to a healthy scaling pattern during recovery. A return toward a steeper slope might indicate reduced protective adaptation and readiness for higher-intensity training.

Third, the null finding for step width suggests that injured runners do not widen their base of support, contrary to what might be expected if they were simply “cautious”. This reinforces that the adaptation is specific to loading-related parameters. Clinicians should primarily target sagittal plane mechanics (stride length, stride rate), but not neglect frontal plane control, which contributes to overall stability and load distribution [23].

Future directions

Several avenues for future research emerge. First, extracting the full joint angle waveforms from the raw marker data would enable fPCA, SPM, and UCM analyses, providing a more complete picture of how injury reorganises multi-joint coordination. Second, subgroup analyses by specific injury diagnosis (e.g., patellofemoral pain vs. Achilles tendinopathy) could reveal whether different injuries elicit different motor adaptations. Third, longitudinal studies that track runners before and after an injury would determine whether the observed pattern is a pre-existing risk factor or a consequence of injury. Fourth, While we used injury severity categories (e.g., ‘continuing to train in pain’) to define injury status, future work could incorporate continuous pain ratings (e.g., visual analogue scale) or repeated measurements of pain during running to examine dose-response relationships between pain intensity and stride length adaptation. Finally, extending the analysis to walking trials would examine whether the adaptation generalises to lower speeds.

Conclusion

Using a large, open biomechanics dataset and a multi-method statistical approach (linear mixed model, bootstrap, Bayesian inference, random forest with SHAP), we demonstrated that running-related injury attenuates the normal increase in stride length with speed. Injured runners increase stride length by 0.042 m less per 1 m/s increase in speed compared to healthy runners. This pattern is consistent with a protective motor adaptation that reduces impact loading at higher velocities, and it aligns with the motor adaptation to pain framework [8]. The convergence of evidence across four independent methods, combined with a time-stamped open-science project on OSF and full transparency of data and code, makes this a robust and reproducible finding. Our results advance the theoretical understanding of how the locomotor system reorganises under pathological constraints and provide a methodological template for future secondary analyses of large biomechanical datasets. Ultimately, they underscore the value of open science and multi-method inference in human movement science.

Abbreviations

Abbreviation	Full form
LMM	linear mixed model(s)
SHAP	SHapley Additive ExPlanations
BMI	body mass index
OSF	Open Science Framework
fPCA	functional principal component analysis
SPM	statistical parametric mapping
UCM	uncontrolled manifold
R^2	coefficient of determination
SD	standard deviation
CI	confidence interval
CrI	credible interval
MCMC	Markov chain Monte Carlo
OU-ISIR	Osaka University - Institute of Scientific and Industrial Research
OU-MVLP	Osaka University - Multi-View Large Population
CC-BY	Creative Commons Attribution licence
NSERC	Natural Sciences and Engineering Research Council
CRediT	Contributor Roles Taxonomy

Declarations

Ethics approval and consent to participate

Not applicable. This study is a secondary analysis of a de-identified, openly available biomechanics dataset; no new data were collected. All procedures for the original data collection were approved by the University of Calgary's Conjoint Health Research Ethics Board (E 21705, E 22194, E 24339), and all participants provided written informed consent, as detailed in Ferber et al. (2024). The present secondary analysis used de-identified public data and complied with applicable confidentiality and data protection standards.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analysed during the current study are available as part of the Running Injury Clinic Kinematic Dataset on Figshare+ https://plus.figshare.com/articles/dataset/Running_Injury_Clinic_Kinematic_Dataset/24255795 (Ferber et al., 2024). The complete data extraction and analysis scripts (R code, numbered 00_run_all.R through 06_figures_and_tables.R), the cleaned data (dv_full_clean.csv, analysis_metadata_clean.csv), all output files (model summaries, bootstrap distributions, Bayesian posterior draws, figures, tables), and a session-info.txt file are publicly available on the OSF (https://doi.org/10.17605/OSF.IO/2EJ9K). The OSF project is licensed under a CC BY 4.0 Attribution International license.

Competing interests

The authors declare that they have no competing interests.

Funding

The original data collection underlying this secondary analysis was supported by an NSERC Discovery Grant (Project No. 1030390). The present secondary analysis did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

KNA conceived and designed the secondary analysis, performed the statistical analyses, and drafted the main manuscript text. NA supervised the project, validated the findings, and reviewed and edited the manuscript. NS contributed to data curation, investigation, and critical revision of the manuscript. LRE contributed to methodological decisions, software support, and reviewed and edited the manuscript. All authors read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Acknowledgements

The authors gratefully acknowledge Dr. Reed Ferber and colleagues for making the Running Injury Clinic biomechanics dataset openly available, which made this secondary analysis possible. We also thank the University of Calgary's Running Injury Clinic staff and participants for their contributions to the original data collection.

References

1. Stergiou N, Decker LM. Human movement variability, nonlinear dynamics, and pathology: Is there a connection? Human Movement Science. 2011;30:869–88. https://doi.org/10.1016/j.humov.2011.06.002
2. Jakob V, Küderle A, Kluge F, Klucken J, Eskofier BM, Winkler J, et al. Validation of a Sensor-Based Gait Analysis System with a Gold-Standard Motion Capture System in Patients with Parkinson's Disease. Sensors. 2021;21:7680. https://doi.org/10.3390/s21227680
3. Yin J, Cong W, Wang Y, Zhou C. Three-dimensional gait analysis in spinal disorders: biomechanical insights and clinical applications for diagnosis, surgical planning, and rehabilitation. Front Neurol. 2025;16:1666267. https://doi.org/10.3389/fneur.2025.1666267
4. Ferber R, Brett A, Fukuchi RK, Hettinga B, Osis ST. A Biomechanical Dataset of 1,798 Healthy and Injured Subjects During Treadmill Walking and Running. Sci Data. 2024;11:1232. https://doi.org/10.1038/s41597-024-04011-7
5. Chakravorti T, Koneru S, Rajtmajer S. Reproducibility and replicability in research: What 452 professors think in Universities across the USA and India. Naudet F, editor. PLoS ONE. 2025;20:e0319334. https://doi.org/10.1371/journal.pone.0319334
6. He Z, Tang X, Yang X, Guo Y, George TJ, Charness N, et al. Clinical Trial Generalizability Assessment in the Big Data Era: A Review. Clinical Translational Sci. 2020;13:675–84. https://doi.org/10.1111/cts.12764
7. Knudson D. Confidence crisis of results in biomechanics research. Sports Biomechanics. 2017;16:425–33. https://doi.org/10.1080/14763141.2016.1246603
8. Hodges PW, Tucker K. Moving differently in pain: A new theory to explain the adaptation to pain. Pain. 2011;152:S90–8. https://doi.org/10.1016/j.pain.2010.10.020
9. Preece SJ, Brookes N, Williams AE, Jones RK, Starbuck C, Jones A, et al. A new integrated behavioural intervention for knee osteoarthritis: development and pilot study. BMC Musculoskelet Disord. 2021;22:526. https://doi.org/10.1186/s12891-021-04389-0
10. Lopes Ferreira C, Barroso FO, Torricelli D, Pons JL, Politti F, Lucareli PRG. Women with patellofemoral pain show altered motor coordination during lateral step down. Journal of Biomechanics. 2020;110:109981. https://doi.org/10.1016/j.jbiomech.2020.109981
11. Grooms D, Appelbaum G, Onate J. Neuroplasticity Following Anterior Cruciate Ligament Injury: A Framework for Visual-Motor Training Approaches in Rehabilitation. Journal of Orthopaedic & Sports Physical Therapy. 2015;45:381–93. https://doi.org/10.2519/jospt.2015.5549
12. Leacox A, Fashingbauer L, Ferguson T, Zajakowski A, Baum B, Reinking M. The Effect of Running Speed on Cadence and Running Kinetics. International Journal of Sports Physical Therapy [Internet]. 2025 [cited 2026 Apr 23];20. https://doi.org/10.26603/001c.140544
13. Lenhart RL, Thelen DG, Wille CM, Chumanov ES, Heiderscheit BC. Increasing Running Step Rate Reduces Patellofemoral Joint Forces. Medicine & Science in Sports & Exercise. 2014;46:557–64. https://doi.org/10.1249/MSS.0b013e3182a78c3a
14. Baček T, Xu Y, Peng L, Oetomo D, Tan Y. Gait adaptations in step length and push-off force during walking with functional asymmetry. Front Bioeng Biotechnol. 2025;13:1550710. https://doi.org/10.3389/fbioe.2025.1550710
15. O'Connor SM, Kuo AD. Direction-Dependent Control of Balance During Walking and Standing. Journal of Neurophysiology. 2009;102:1411–9. https://doi.org/10.1152/jn.00131.2009

16. Mota JA. Linear Mixed-Effect Modeling in Kinesiology Research: A Tutorial in R. BJSTR [Internet]. 2023 [cited 2026 Apr 23];49. https://doi.org/10.26717/BJSTR.2023.49.007794
17. Efron B, Tibshirani RJ. An Introduction to the Bootstrap [Internet]. 0 edn. Chapman and Hall/CRC; 1994 [cited 2026 Apr 23]. https://doi.org/10.1201/9780429246593
18. Ebrahim EA, Cengiz MA. Bayesian hierarchical models for multivariate mixed responses with repeated measures: A case study in arterial occlusive disease. Belay DB, editor. PLoS One. 2026;21:e0346331. https://doi.org/10.1371/journal.pone.0346331
19. Lemoine NP. Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. Oikos. 2019;128:912–28. https://doi.org/10.1111/oik.05985
20. Aghababa MP, Andrysek J. Exploration and demonstration of explainable machine learning models in prosthetic rehabilitation-based gait analysis. Naseer N, editor. PLoS ONE. 2024;19:e0300447. https://doi.org/10.1371/journal.pone.0300447
21. Eddelbuettel D, Knapp B, Lemire D. RcppSimdJson: 'Rcpp' Bindings for the 'simdjson' Header-Only Library for 'JSON' Parsing [Internet]. 2020 [cited 2026 Apr 23]. p. 0.1.15. https://doi.org/10.32614/CRAN.package.RcppSimdJson
22. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2025.
23. Ferber R, Osis ST, Hicks JL, Delp SL. Gait biomechanics in the era of data science. Journal of Biomechanics. 2016;49:3759–61. https://doi.org/10.1016/j.jbiomech.2016.10.033
24. Frömer R, Nassar MR, Ehinger BV, Shenhav A. Common neural choice signals can emerge artefactually amid multiple distinct value signals. Nat Hum Behav. 2024;8:2194–208. https://doi.org/10.1038/s41562-024-01971-z
25. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using **lme4**. J Stat Soft [Internet]. 2015 [cited 2026 Mar 7];67. https://doi.org/10.18637/jss.v067.i01
26. Kuznetsova A, Brockhoff PB, Christensen RHB. **lmerTest** Package: Tests in Linear Mixed Effects Models. J Stat Soft [Internet]. 2017 [cited 2026 Mar 9];82. https://doi.org/10.18637/jss.v082.i13
27. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statist Med. 2000;19:1141–64. https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9%3C1141::AID-SIM479%3E3.0.CO;2-F
28. Muth C, Oravecz Z, Gabry J. User-friendly Bayesian regression modeling: A tutorial with rstanarm and shinystan. TQMP. 2018;14:99–119. https://doi.org/10.20982/tqmp.14.2.p099
29. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner P-C. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. arXiv; 2019 [cited 2026 Apr 23]; https://doi.org/10.48550/ARXIV.1903.08008
30. Martínez-Gramage J, Albiach JP, Moltó IN, Amer-Cuenca JJ, Huesa Moreno V, Segura-Ortí E. A Random Forest Machine Learning Framework to Reduce Running Injuries in Young Triathletes. Sensors. 2020;20:6388. https://doi.org/10.3390/s20216388
31. Breiman L. Random Forests. Machine Learning. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324
32. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics. 2007;8:25. https://doi.org/10.1186/1471-2105-8-

25

33. Biecek P. DALEX: Explainers for Complex Predictive Models in R. Journal of Machine Learning Research. 2018;19:1–5.
34. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions [Internet]. arXiv; 2017 [cited 2026 Apr 23]. https://doi.org/10.48550/ARXIV.1705.07874
35. Wendt CJ, Anderson GB. Ten simple rules for finding and selecting R packages. Markel S, editor. PLoS Comput Biol. 2022;18:e1009884. https://doi.org/10.1371/journal.pcbi.1009884
36. Wickham H. ggplot2 [Internet]. Cham: Springer International Publishing; 2016 [cited 2026 Apr 23]. https://doi.org/10.1007/978-3-319-24277-4
37. Bogossian PM, Nattala U, Wong ASM, Morrice-West AV, Zhang GZ, Rana P, et al. A machine learning approach to identify stride characteristics predictive of musculoskeletal injury, enforced rest and retirement in Thoroughbred racehorses. Sci Rep. 2024;14:28967. https://doi.org/10.1038/s41598-024-79071-1
38. He Z, Yang Y, Fang R, Zhou S, Zhao W, Bai Y, et al. Integration of shapley additive explanations with random forest model for quantitative precipitation estimation of mesoscale convective systems. Front Environ Sci. 2023;10:1057081. https://doi.org/10.3389/fenvs.2022.1057081
39. Heiderscheit BC, Chumanov ES, Michalski MP, Wille CM, Ryan MB. Effects of Step Rate Manipulation on Joint Mechanics during Running. Medicine & Science in Sports & Exercise. 2011;43:296–302. https://doi.org/10.1249/MSS.0b013e3181ebedf4
40. Schache AG, Blanch PD, Rath DA, Wrigley TV, Starr R, Bennell KL. A comparison of overground and treadmill running for measuring the three-dimensional kinematics of the lumbo–pelvic–hip complex. Clinical Biomechanics. 2001;16:667–80. https://doi.org/10.1016/S0268-0033(01)00061-4
41. Karos K, Meulders A, Gatzounis R, Seelen HAM, Geers RPG, Vlaeyen JWS. Fear of pain changes movement: Motor behaviour following the acquisition of pain-related fear. European Journal of Pain. 2017;21:1432–42. https://doi.org/10.1002/ejp.1044
42. Noehren B, Scholz J, Davis I. The effect of real-time gait retraining on hip kinematics, pain and function in subjects with patellofemoral pain syndrome. Br J Sports Med. 2011;45:691–6. https://doi.org/10.1136/bjsm.2009.069112
43. Kim S, Yu J. Changes of gait parameters and lower limb dynamics in recreational runners with achilles tendinopathy. J Sports Sci Med. 2015;14:284–9.
44. Clemente F, Ramirez-Campillo R, Beato M, Moran J, Kawczynski A, Makar P, et al. Arbitrary absolute vs. individualized running speed thresholds in team sports: A scoping review with evidence gap map. bs. 2023;40:919–43. https://doi.org/10.5114/biolsport.2023.122480
45. Fukuchi CA, Fukuchi RK, Duarte M. Effects of walking speed on gait biomechanics in healthy participants: a systematic review and meta-analysis. Syst Rev. 2019;8:153. https://doi.org/10.1186/s13643-019-1063-z
46. Camargo J, Ramanathan A, Flanagan W, Young A. A comprehensive, open-source dataset of lower limb biomechanics in multiple conditions of stairs, ramps, and level-ground ambulation and transitions. Journal of Biomechanics. 2021;119:110320. https://doi.org/10.1016/j.jbiomech.2021.110320
47. Chambers R, Chandra H. A Random Effect Block Bootstrap for Clustered Data. Journal of Computational and Graphical Statistics. 2013;22:452–70. https://doi.org/10.1080/10618600.2012.681216

48. Ferguson J. Bayesian interpretation of p values in clinical trials. BMJ Evid Based Med. 2022;27:313–6. https://doi.org/10.1136/bmjebm-2020-111603
49. Han R, Qi F, Wang H, Yi M. Innovative machine learning approach for analysing biomechanical factors in running-related injuries. mcb. 2024;21:530. https://doi.org/10.62617/mcb530
50. Van Hooren B, Fuller JT, Buckley JD, Miller JR, Sewell K, Rao G, et al. Is Motorized Treadmill Running Biomechanically Comparable to Overground Running? A Systematic Review and Meta-Analysis of Cross-Over Studies. Sports Med. 2020;50:785–813. https://doi.org/10.1007/s40279-019-01237-z
51. Bramah C, Preece SJ, Gill N, Herrington L. A 10% Increase in Step Rate Improves Running Kinematics and Clinical Outcomes in Runners With Patellofemoral Pain at 4 Weeks and 3 Months. Am J Sports Med. 2019;47:3406–13. https://doi.org/10.1177/0363546519879693
52. Esculier J-F, Bouyer LJ, Roy J-S. Running gait modifications can lead to immediate reductions in patellofemoral pain. Front Sports Act Living. 2023;4:1048655. https://doi.org/10.3389/fspor.2022.1048655

Figures

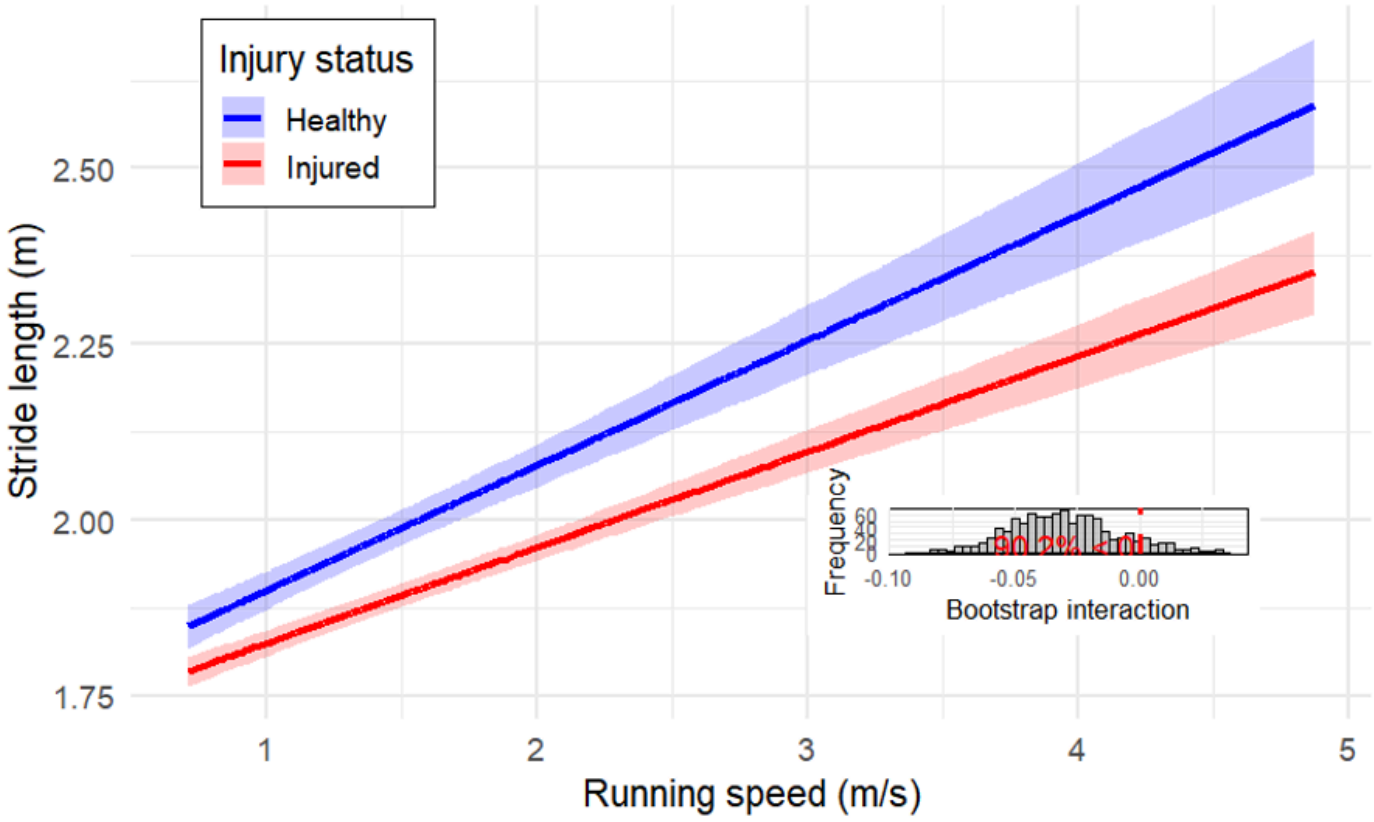


Figure 1

Predicted stride length versus running speed from the linear mixed model (solid lines, 95% confidence bands). Inset: bootstrap distribution of the speed × injury interaction coefficient (vertical dashed line = zero; 90.2% of estimates < 0).

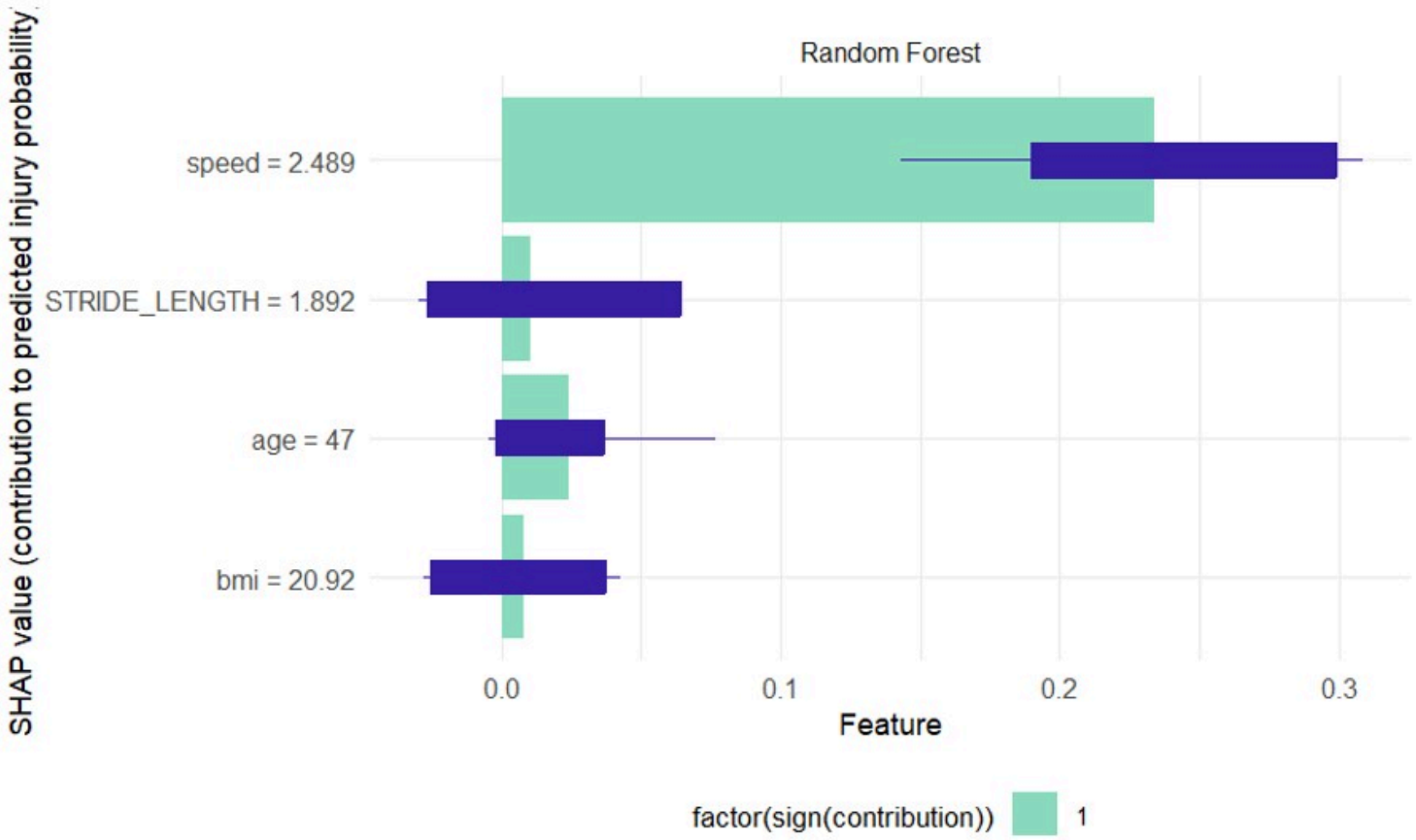


Figure 2

SHAP summary plot for the random forest model. Points = trials; colour = feature value (red high, blue low). Positive SHAP values indicate higher predicted injury probability. Lower speed and shorter stride length increase injury prediction.