. . . Published ahead of Print

Predicting Musculoskeletal Loading at Common Running Injury Locations using Machine Learning and Instrumented Insoles

Bas Van Hooren, Lars van Rengs, and Kenneth Meijer

NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University Medical Centre+, Department of Nutrition and Movement Sciences, Maastricht, THE NETHERLANDS

Accepted for Publication: 13 May 2024

Medicine & Science in Sports & Exercise, Publish Ahead of Print
DOI: 10.1249/MSS.0000000000003493

Predicting Musculoskeletal Loading at Common Running Injury Locations using Machine Learning and Instrumented Insoles

Bas Van Hooren, Lars van Rengs, and Kenneth Meijer

NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University Medical Centre+, Department of Nutrition and Movement Sciences, Maastricht, THE NETHERLANDS

Address for correspondence: Bas Van Hooren, NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University Medical Centre+, Department of Nutrition and Movement Sciences, Universiteitssingel 50, Maastricht, The Netherlands, 6229 ER;
E-mail: b.vanhooren@maastrichtuniversity.nl

Conflict of Interest and Funding Source: BVH was funded by an Eurostars grant (ID 12912) awarded by Eureka. LvR and KM declare that they have no conflict of interest. BVH was (part-time) employed by Ato-Gear during a period in which this manuscript was written.

ABSTRACT

Introduction: Wearables have the potential to provide accurate estimates of tissue loads at common running injury locations. Here we investigate the accuracy by which commercially available instrumented insoles (ARION) can predict musculoskeletal loading at common running injury locations. **Methods:** 19 runners (10 males) ran at five different speeds, four slopes, with different step frequencies, and forward trunk lean on an instrumented treadmill, while wearing instrumented insoles. The insole data was used as input to an artificial neural network that was trained to predict the Achilles tendon strain, and tibia and patellofemoral stress impulses and weighted impulses (damage proxy) determined with musculoskeletal modelling. Accuracy was investigated using leave-one-out cross-validation and correlations. The effect of different input metrics was also assessed. **Results:** The neural network predicted tissue loading with overall relative percentage errors of 1.95 ± 8.40, -7.37 ± 6.41, and -12.8 ± 9.44% for the patellofemoral joint, tibia and Achilles tendon impulse, respectively. The accuracy significantly changed with altered running speed, slope, or step frequency. Mean (95% confidence interval) within-individual correlations between modelled and predicted impulses across conditions were generally nearly perfect, being 0.92 (0.89 to 0.94); 0.95 (0.93 to 0.96); and 0.95 (0.94 to 0.96) for the patellofemoral, tibial, and Achilles tendon stress/strain impulses, respectively. **Conclusions:** This study shows that commercially available instrumented insoles can predict loading at common running injury locations with variable absolute, but (very) high relative accuracy. The absolute error was lower than methods that measure step-count only, or assume a constant load per speed or slope. This developed model may allow for quantification of in-field tissue loading and real-time tissue loading-based feedback to reduce injury risk.

Key Words: WEARABLES, ACHILLES TENDON, ARTIFICIAL NEURAL NETWORK, BIOFEEDBACK, BIOMECHANICS

INTRODUCTION

Running is one of the most popular sporting activities but also an activity with high dropout rates. The primary reason for dropout is a running-related injury (1, 2), with patellofemoral pain, tibial medial stress syndrome, and Achilles tendinopathy being among the most common injuries (3-5). Although the cause of running injuries is multifactorial with both biological and mechanical causes, it is generally accepted that repetitive cyclical loading with insufficient recovery is an important contributing factor (6). Specifically, repetitive loading leads to microdamage, which can accumulate and eventually lead to macroscopic damage (i.e., a running-related overuse injury) when accompanied by insufficient time for remodeling and adaptation.

Due to the importance of mechanical load in the development of running injuries, an increasing number of wearables are being used to estimate biomechanical loading variables such as the ground reaction force or other external loading variables (e.g., tibial shock, vertical loading rate) based on the assumption that these external loading measures provide useful estimates of the internal forces experienced by tissues such as bones or tendons (7, 8). However, these surrogate measures do not always accurately reflect the force experienced by tissues within the body (7, 9-12), and use of these surrogate measures to predict or prevent injuries can therefore be misleading (7, 9). Alternative approaches are thus required to provide more accurate estimations of tissue loading.

Musculoskeletal modelling is a popular non-invasive approach to estimate tissue loading (13-16). However, this approach typically requires three-dimensional marker positions and external forces obtained with motion capture and force plates, respectively, thus making it

unsuitable for in-field applications. However, by combining musculoskeletal modelling with wearable sensors and statistical methods, it may become possible to estimate tissue loading in-field. Specifically, statistical methods such as linear regression or machine learning algorithms can be used to map the output from a wearable sensor (e.g., accelerations from an accelerometer) to the tissue loads estimated in a lab setting (9, 17-20) and the established relationship between the wearable output(s) and tissue load can in turn be used in-field to estimate tissue loading using only the wearable output. Elstub and colleagues (19) for example used musculoskeletal modelling to estimate tibial compressive forces, and subsequently mapped the output from an instrumented insole to the musculoskeletal estimates using a regression method. A limitation to this study as well as other previous studies (9) is that they only modelled the axial tibial force. However, this approach may not accurately inform on actual tibial loading during sloped running (21, 22), thus limiting in-field use where runners are likely to encounter slopes. For example, the axial force may decrease during downhill running, whereas the force due to bending increases, thereby collectively increasing total tibial stress due to the larger magnitude of the bending forces (22). Moreover, most previous studies attempted to predict load at just one anatomical location, which this often also being a location that is a less commonly injured by runners (e.g., patella tendon (18, 20) or tibiofemoral joint (17) instead of patellofemoral joint). Quantification of the load specifically at common injury locations is likely to provide more useful information for training guidance. Moreover, simultaneous quantification of the load at multiple common running injury locations is important as advice to change running technique to offload one tissue can increase load and thus injury risk at another tissue (22).

The impulse (i.e., aera under a waveform curve) is often considered a useful metric to quantify biomechanical (tissue) loading because it captures both the magnitude and the duration of an applied load. However, it has been argued that the impulse does not accurately reflect tissue damage and thus injury risk due to the non-linear relationship between a given load and the resulting damage (15, 22). Indeed, *ex vivo* experiments show that the fatigue life (i.e., number of cycles to failure) of tendon and bone tissue follows an inverse power-law relationship (23-26), whereby small increases in loading magnitudes result in large decreases in fatigue life. Wren and colleagues (24) for instance found the human Achilles tendon to fail on average after 1400 cycles when stretched to 6% strain, while reducing the strain magnitude by 50% (i.e., to 3%) increased the number of cycles to failure to 93,000 (i.e., a 6600% increase). This inverse power-law relationship between an applied load and the resulting damage can be modelled by applying a weighting factor (*b*) to a given load (e.g., stress or strain value), with *b* denoting the slope of the power function between fatigue life and the tissue-specific stress or strain (6, 13, 15). Previous studies did not consider this non-linear relationship when attempting to predict tissue loading. However, this is important as the impulse for some tissues decreases with higher speeds, while the weighted impulse (and peak loading metrics) increases (15, 22). Importantly, only the latter relationship is consistent with the higher risk of injuries at higher running speeds seen in experimental studies (27, 28), thus highlighting the importance of accounting for this non-linear relationship using weighted impulse measures to accurately inform on injury risk.

The primary aim of the current study was therefore to investigate how accurate a machine learning algorithm can quantify the patellofemoral stress, tibial stress, and Achilles strain

impulses and weighted impulses (proxy of damage) based on instrumented insole data collected during running when compared to musculoskeletal modelling. As a secondary aim we compared this accuracy to a method that assumed that the load on each tissue was constant across *all* conditions, or to a method that assumed an average load per condition. The former approach is similar to measuring only step count to monitor injury risk (9), whereas the latter approach reflects a wearable that attempts to predict tissue loading from the average load reported in the literature at a specific speed or slope, without using any metrics to personalize this estimate. Finally, as tertiary aims, we also explored 1) how the prediction error changed with different dataset sizes used to train the machine learning algorithm, and 2) the sensitivity of the prediction accuracy to exclusion of wearable (input) metrics (e.g., exclusion of contact time).

METHODS

General design of the study

All participants completed a single test session and were instructed to avoid strenuous activity for 36 h, alcohol for 24 h, caffeine for 6 h, and a heavy meal 1 h before the session. When entering the lab, anthropometric measurements were taken using standardized procedures and the participants were then equipped with instrumented insoles and retroreflective markers as described below. After subject calibration and a familiarization period, the participants completed short (1 min) runs while ground reaction forces and lower body and trunk kinematics were collected.

Participants

All data was part of a larger project (22, 29) whereby 19 participants (10 males, 9 females, mean ± SD age 23.6 ± 3.7 years, body height 174.9 ± 9.2 m; body mass 67.2 ± 10.4 kg) that were free of any moderate (for previous 3 months) or minor (for previous 1 month) musculoskeletal injuries, were comfortable with treadmill running, had a body mass index (BMI) of <26, and were aged 18-45 volunteered to participate. The study was approved by the local ethics committee (nr. 2019-1138), was conducted according to the declaration of Helsinki and all participants signed an informed consent form prior to the measurements.

Instruments

The computer assisted rehabilitation environment (CAREN, Motek, The Netherlands) system combines an instrumented split-belt treadmill (belt length and width 2.15 x 0.5 m, 6.28-kW motor per belt, 60 Hz belt speed update frequency and 0-18 $km \cdot h^{-1}$ speed range) with a 12-camera three-dimensional motion capture system (VICON NEXUS v2.1, Oxford Metrics Group, Oxford, UK, 100 Hz) and was used to measure 3D marker positions and ground reaction forces. Each participant ran in their own shoes that were fitted with an appropriate sized instrumented insole (Arion, ATO-GEAR, Eindhoven, The Netherlands) as described previously (29). The insoles comprised of a tri-axial accelerometer, gyroscope, and eight spatially distributed force-resistive pressure sensors, with only the pressure data (150 Hz) being used to compute outcomes of interest (see later). The participants' own shoes, rather than standardized shoes were used to maximize the applicability of the findings to in-field conditions. The microprocessor was clipped

to the lateral side of the shoe and transmitted the collected data to a mobile phone from which it was transferred to a computer for further analyses.

Data collection

After placement of the insole and calibration of the systems, 26 retroreflective skin markers with a diameter of 14 mm were attached to the skin with double-sided tape using a modified lower-limb and trunk marker set (Human Body Model v2) (30).

Prior to data collection, the participants were instructed to run for 8 min at a fixed-paced speed of 2.78 $m \cdot s^{-1}$ to familiarize themselves with treadmill running (31). This was followed by 4 minutes of running at 3.33 $m \cdot s^{-1}$. The participants then completed a series of 1-minute runs at different fixed-paced speeds and treadmill slopes (Table 1), with the order of conditions being randomized by an online research randomizer (https://www.randomizer.org/). These speeds and slopes were selected to reflect conditions that recreational runners could encounter when running in-field. After these conditions, the participants ran another three trials at 3.33 $m \cdot s^{-1}$, but with a higher and lower step frequency (± 10 $steps \cdot min^{-1}$) compared to their step frequency during the 4-minute trial at 3.33 $m \cdot s^{-1}$, and they ran with self-selected forward trunk lean. The order of these conditions was also randomized. These latter conditions were included to assess the validity of the predicted tissue loads during common gait-retraining methods (e.g., (32, 33)). This may in turn inform on the suitability of the wearable to quantify the effectiveness of gait-retraining at modifying tissue loading. The participants were allowed to take rest periods between trials when required and were instructed to run as if they were running outside and focus on the simulated virtual environment.

Lab-based data analysis

Musculoskeletal modelling. Musculoskeletal modelling was used to determine the load and damage at three common injury locations as detailed previously (22). Briefly, kinematic and kinetic data were filtered using a zero-lag 4th order Butterworth filter with a low-pass filter of 20 Hz and exported to musculoskeletal simulation software (OpenSim SimTK 3.3, Stanford, USA) (34) using custom-made Matlab scripts. Footstrike and toe-off were identified when the vertical ground reaction force exceeded and dropped below 20 N.

Musculoskeletal simulation was performed using a modified full-body musculoskeletal model (22 rigid body segments, 37 degrees of freedom and 80 muscles) that was designed to produce more realistic moment arms during tasks involving large hip and knee flexions (35). The model's geometry and mass were subsequently scaled to the participant's individual segments length and total body mass. A weighted least squares minimization of markers on segments and joints was then used to match the virtual markers on the scaled model to the experimental markers during each frame using the "Inverse kinematics" option, while net joint moments for each degree of freedom were computed using the "Inverse dynamics" option. A dynamic optimization criterion that used a cost function that minimized the sum of the squared muscle activations was used to determine the muscle forces required to reproduce the joint moments (36).

Determination of load and damage at common injury locations. The "Joint Reaction analysis" tool (37) was used to calculate patellofemoral and ankle contact forces by summing the joint reaction/intersegmental forces from inverse dynamics and muscle forces from dynamic

optimization. Patellofemoral force was computed as the compressive component of the force between the femur and patella. The magnitude of this force depends on the force produced by all quadriceps muscles, and the knee flexion angle, which alters the angle of pull between the femur and patella, and thereby affects the shear and compressive magnitude (22); Figure 1. Patellofemoral contact stress was subsequently estimated by dividing the compressive component of the joint contact force by the contact area. The contact area was determined based on sex-specific data from Besier and colleagues (38) who estimated patellofemoral contact areas as a function of the knee flexion angle during squats using magnetic resonance imaging. Stress fractures often occur at the posteromedial distal 1/3 or middle 2/3 of the tibia (39), which is also the location with peak compressive stresses (13, 40). Modelling studies have shown that 72-83% of the peak normal tibia stress during running is caused by the bending moment, with the remaining stress being caused by axial compressive forces (21, 22, 40). We therefore estimated the combined tibial stress due to the bending moment and axial force as described previously (22). Achilles tendon force was calculated by summing the forces from the medial and lateral gastrocnemius and soleus. Achilles tendon length change was estimated by dividing the Achilles tendon force by tendon stiffness, with tendon stiffness being derived from ultrasound experiments (22). The length change was subsequently expressed as a percentage of resting length (i.e., strain). For all three structures, we computed the impulse by integrating the stress or strain over the stance phase. Further, we also determined the weighted impulse by raising the stress or strain values to the power of an empirically derived exponent before impulse calculation. The exponent used was 7 for patellofemoral and tibial bone damage, and 9.3 for Achilles tendon damage (6, 15, 24, 26). Justification for these exponents is provided in (22). The impulse and weighted impulse at each tissue was determined for both legs over a period of 50

seconds of steady state running. This duration was chosen to ensure enough steps (i.e., >25 (41, 42)) to achieve stable biomechanical data and ensure sufficient training data for the neural network.

Neural network modelling

An artificial neural network (Figure 1) implemented in Python version 3.9.12 (Python Software Foundation, USA) was used to predict the patellofemoral stress, tibial stress, and Achilles tendon strain impulse and weighted impulse at each step from 12 predictors derived from the instrumented insole. Separate neural networks were made for the impulse and weighted impulse. Section 2.6.2 describes how the neural network structure (e.g., number of layers) was developed, whereas section 2.7 describes how the final neural network was evaluated.

Predictors included in the neural network. The predictors included spatiotemporal outcomes such as contact time and step frequency derived from the processed raw sensor data from which noise and unwanted artifacts were removed through signal processing. The full list of predictors along with a short explanation of their meaning is provided in Supplemental Table 1 (Supplemental Digital Content, http://links.lww.com/MSS/D30). Spatiotemporal outcomes were determined using previously validated proprietary algorithms (29). All predictors and outcomes were normalized between 0 and 1 using a fixed range specific to each variable to ensure equal importance of each variable to the network and thereby avoid that variables with larger numeric values were given more relevance in the neural network than those with smaller values (43). The minimum or maximum values for each variable were selected so that they encompassed the most common range within our dataset, while at the same time excluding extreme values that may be

considered outliers. This normalization procedure also ensured no training data information was used for normalization of the test set.

The normalized 12 predictors for each step were mapped to the musculoskeletal model normalized stress/strain impulse or weighted impulse for each structure per step. This was done across all conditions, for both legs, and all participants. Specifically, the instrumented insole metrics served as the input, and patellofemoral joint, tibial or Achilles tendon stress/strain impulse or weighted impulse as the target/output (Figure 1). The wearable metrics and lab-based tissue loads for each step were synchronized using a jump performed at the start and end of each condition. Training was done separately for the impulse and weighted impulse.

Neural network development and validation. The structure of the ANN that produced the highest accuracy was developed using the following trial and error procedure. First, we defined a 'simple' model consisting of 1 input, 1 hidden and 1 output layer. The performance of this model was assessed (see below) and the complexity of the model was then increased by adding hidden layers until a total of 6 hidden layers as a further increase in the number of layers did not substantially improve the performance of the model (error decrease with >6 layers was <1%). The number of neurons per layer was set as [the number of neurons in the input layer + output layer] × 2/3. 10, 20, 30, 20 and 10 neurons were additionally added for hidden layer 2, 3, 4, 5, and 6 respectively. A rectified linear unit activation function was defined between the input and each hidden layer, while the output layer used a linear activation function. A previous study found that the activation function used did not impact the predictive ability of an ANN (44).

To evaluate each ANN structure during the development process, we randomly divided the complete dataset into a training and testing set with 90% and 10% of the data being allocated to each group. Note that in this division, data from the same participant could be present in all datasets. However, this did not impose any issues to generalization as this procedure was used only to establish the ANN structure that produced the highest accuracy, with a separate leave-one-out-validation being performed later to assess the performance of the final model on a participant not used in the training procedure.

Each neural network was trained (i.e., the weights between neurons were optimized) on the training set with the ADAM stochastic gradient-based optimization method (45). During this training process, we monitored the increase in loss in the validation group as an indicator of overfitting. Specifically, the training loss describes the mean absolute error between the modelled output and predicted values for each training iteration. If the model can perfectly predict the modelled output, the training loss would be zero. However, the ANN may also learn features that are specific to the training set, but not to the validation/testing set, a process known as overfitting. This reduces the performance of the ANN on 'unseen' data. To minimize both training and validation loss, we trained the network for 1000 iterations (46), and training was stopped if the gradient of the loss function did not decrease more than 0.5% for ten consecutive iterations. A uniform distribution was used to initiate learning and the learning rate was set to 0.01. Moreover, overfitting was prevented by implementing random drop-outs of neurons in the hidden layers (47) (5% in hidden layer 1, 1% in the other hidden layers). All input data (i.e., steps) was shuffled to avoid order effects. A 10-fold cross-validation was used to assess the robustness of the learned relationships for each subsequent ANN structure alteration including

the final structure. Briefly, this procedure involved re-shuffling and randomly splitting the original training datasets 10 times (9 folds for training, 1 fold for validation) to assess the variability in the error. The performance of the model on the 10% of data in the test set was then used to assess if the altered structure improved the performance of the model.

Final neural network structure. The final neural network structure established using the trial-and-error procedure described in section 2.6.2 included, 1 input, 6 dense hidden (layers where each of the neurons is connected to all the neurons of the previous layer and consequently to all the neurons of the next layer) and 1 output layer (Figure 1). The input layer had 12 variables (i.e., nodes/neurons) and the output layer 3 variables.

Neural network evaluation

The neural network predicted tissue stress/strain (weighted) impulses were evaluated using a leave-one-subject-out cross-validation (implemented in Python) to assess the prediction on a new individual, from which no other trials were previously included in the training set. For example, the final neural network was trained on the data from participant 1-18 and then predicted the impulse or weighted impulse for participant 19 (i.e., a participant who's data had not been used for training). This process was repeated for each participant. The mean error over all steps for each leave-one-subject-out cross-validation were calculated in original units (MPa·s and %·s for patellofemoral and tibial stress, and Achilles tendon strain impulses, respectively), relative percentage units, and absolute percentage error. These errors therefore reflect the average error across multiple (on average 134) steps. This was subsequently averaged over all validations across all trials to provide an indication of the overall error. The weighted impulse

was raised to the power of 1/*b* for reporting purposes to reduce the magnitude, hereby easing interpretation (15, 22), where *b* reflects the exponent used.

Objective assessment of absolute agreement. Absolute agreement between the predicted and modelled load may be important for models that aim to predict injury risk at a given time point using probabilistic functions of damage and adaptation (e.g., (13, 14, 16, 48-50)) and was assessed using a statistical approach proposed by Shieh (51) (implemented in RStudio v4.2.3) with the percentage difference as the unit for comparison. In this method the mean percentage difference and standard deviation of this difference between the modelled and predicted load is assessed in relation to an *a priori* determined threshold, whereby a specified proportion of the data should fall within the threshold to declare agreement. Errors for the predicted value were compared against three thresholds: of <10%, <20%, and <30%. These thresholds were chosen to reflect the range in errors reported in previous studies (17, 18, 20). The central null-proportion (reflecting the fraction of datapoints that should fall within this threshold) was set to 0.95 in line with the widely used 95% limits of agreement, and the alpha level to 0.05. Therefore, if the 95% confidence intervals of the 95% limits of agreement between the modelled and predicted value for the assessed outcome, fell within the specified threshold, we rejected the null-hypothesis that there is no agreement at the specified threshold.

Group mean approach. To assess if the developed model represented an improvement over existing methods (secondary aim), we compared our developed model with a method that assumed that the load on each tissue was constant across all conditions. To this purpose, the overall mean impulse or weighted impulse was imputed for each condition regardless of speed or

slope. As a secondary approach, we imputed the average impulse or weighted impulse per condition for each individual. This approach would reflect a wearable that uses average loading data for a given speed or slope published in the literature to estimate tissue loading per individual based on only their speed or surface slope.

Assessment of change in error with changes in speed, slope, and cadence. A linear-mixed model (implemented in SPSS Version 25 IBM Corporation, Chicago, IL) was used to assess if the relative percentage error changed with increases in running speed, changes in surface slope, and changes in step frequency. Models included speed, surface gradient, and step frequency as continuous co-variates (fixed effects) and a random intercept and slope per participant. We did not perform pairwise comparisons between all conditions (e.g., running speed 1 vs 2, 1 vs 3, etc.) since this would require 13 x (13-1)/2 = 39 *t*-tests per outcome, per tissue, thus drastically increasing potential for type I errors. Instead, the *p*-value for the slope obtained from the mixed model regression was used to interpret whether the dependent variable (i.e., relative percentage error) significantly changed with changes in the independent variable. A similar mixed model was used to compare the overall absolute percentage error between the three different structures. Statistical significance was set at 0.05. Assumptions of each model were assessed by visually inspecting the level 1 (repeated measures across speeds/slopes/step frequencies) and level 2 (participants) residuals on histograms, Q-Q plots and boxplots to verify that the residuals were approximately normally distributed. Influential outliers were assessed using a combination of visual inspection and Cook's distance at both levels, whereby datapoints with a Cook's distance of >1 were considered influential outliers and excluded from analysis (this was typically one

datapoint). As the exclusion of outliers, or log-transformation to improve the distribution, did not change the interpretation of the models, the original models were retained.

Assessment of relative agreement. Relative agreement may be important when using the wearable to compare changes in load between different running conditions and was assessed by computing a Pearson correlation coefficient between the modelled and predicted (weighted) impulse across all conditions within each individual. To this purpose, we first Fisher *z*-transformed the mean within-individual correlation to better approximate a normal distribution, then computed the mean correlation coefficient and 95% confidence intervals across all individuals and finally back-transformed the *z*-transformed correlation to aid interpretation. Correlations were interpreted as <0.1 trivial; 0.1-0.29 small; 0.30-0.49 moderate; 0.5-0.69 large; 0.7-0.89 very large; 0.9-0.99 nearly perfect (52).

We also assessed relative agreement by computing the change in the modelled/estimated (weighted) impulse at each condition relative to the modelled/estimated (weighted) impulse at 2.78 $m \cdot s^{-1}$. The difference in the change was subsequently expressed as a percentage.

Effect of number of steps and metrics on error. As a tertiary analysis, we explored how the prediction error of the neural network changed when an increasing number of steps per subject per condition (1, 5, 10, 20, 30, 40, 50, 60, or all steps) was used to train the network to predict the impulse. The performance for each model was again evaluated using a leave-one-subject-out cross-validation on an unseen subject. Finally, we also investigated the importance of each input variable to the predicted outcome by re-training the model while excluding one wearable-derived

variable (e.g., contact time) and re-computing the mean absolute percentage error over all leave-one-subject-out cross-validations.

RESULTS

Five trials among five different participants were excluded because the wearable insole data was missing (n=2), or because there was an error with the lab-based data processing (n=3). Therefore, 223 trials, providing a total of 29,894 steps were included. Each participant on average contributed 1573 steps across all trials, thus meaning that on average 28,321 steps were used for training of the neural network, and 1573 for testing (evaluation) in the leave-one-subject-out cross-validation.

Absolute and relative agreement

The modelled and predicted tissue stress/strain impulse values from the leave-one-subject-out cross-validation are provided in Table 2. The overall relative percentage errors were 1.95 ± 8.40, -7.37 ± 6.41, and -12.8 ± 9.44% for the patellofemoral joint, tibia and Achilles tendon, respectively. Figure 2 additionally shows the relative percentage error for the predicted impulse at each structure across the different conditions. The relative percentage error significantly changed with changes in speed, increasing uphill or downhill slope, or step frequency (Supplemental Table 2, Supplemental Digital Content; Figure 2, http://links.lww.com/MSS/D30).

The overall absolute percentage errors were 12.4 ± 3.35, 9.18 ± 4.92, and 15.4 ± 5.61% for the patellofemoral joint, tibia and Achilles tendon, respectively. Supplemental Table 3

(Supplemental Digital Content, http://links.lww.com/MSS/D30 susu) provides the absolute percentage errors obtained with the group mean approach. When imputing the mean lab-based impulse at each condition for each individual, the overall absolute percentage errors were 27.4 ± 15.2, 15.8 ± 11.6, and 29.3 ± 36.2% for the patellofemoral joint, tibia and Achilles tendon, respectively. When imputing the mean lab-based impulse at each condition for each individual, the overall absolute percentage errors were 22.4 ± 8.44, 14.6 ± 6.70, and 24.7 ± 20.1% for the patellofemoral joint, tibia and Achilles tendon, respectively.

Supplemental Table 4 (Supplemental Digital Content, http://links.lww.com/MSS/D30) provides the modelled and predicted weighted stress/strain impulses as well as their relative and absolute (percentage) differences, while Supplemental Table 5 (Supplemental Digital Content, http://links.lww.com/MSS/D30) details the absolute percentage errors obtained with the group mean approach for the weighted impulse. The median (interquartile range) overall absolute percentage errors between modelled and predicted weighted stress/strain impulses were 67.7 (61.1 to 78.8), 62.2 (48.7 to 68.8), and 96.2 (93.8 to 97.9)% for the patellofemoral joint, tibia and Achilles tendon, respectively.

Within-individual correlations between modelled and predicted impulse across conditions were generally nearly perfect, with the mean (95% confidence interval) correlations being 0.92 (0.89 to 0.94); 0.95 (0.93 to 0.96); and 0.95 (0.94 to 0.96) for the patellofemoral stress impulse, tibial stress impulse, and Achilles tendon strain impulse, respectively (Figure 3). The within-individual correlations for the weighted impulse ranged from large to nearly perfect, with the mean (95% confidence interval) correlations being 0.90 (0.80 to 0.96); 0.57 (0.38 to 0.71); and

0.97 (0.92 to 0.99) for the patellofemoral stress weighted impulse, tibial stress weighted impulse, and Achilles tendon strain weighted impulse, respectively (Figure 3). Supplemental Table 6 (Supplemental Digital Content, http://links.lww.com/MSS/D30) reported the percentage change in loading relative to loading at 2.78 $m \cdot s^{-1}$ captured by the machine learning model. Overall, the machine learning model predicted 32-57% of the increase in the impulse across all conditions, with more variable differences for the weighted impulse.

The overall absolute percentage error for the patellofemoral joint impulse was significantly larger than the absolute percentage error for the tibial impulse (difference of 3.21 ± 0.95 percent points; p <.001), but smaller than the absolute percentage error for the Achilles tendon impulse (difference of -3.07 ± 1.26 percent points; p <.001). There was no significant difference in the absolute percentage error between the tibial and Achilles tendon impulse (difference of -6.28 ± 0.82 percent points; p =.12). Similar were observed for the weighted impulse.

The effect of training data size

Supplemental Table 7 (Supplemental Digital Content, http://links.lww.com/MSS/D30) shows the mean number of steps used for training per condition for the overall neural network. Briefly, the average number of steps per subject per condition was 134 ± 15.0. Supplemental Table 8 (Supplemental Digital Content, http://links.lww.com/MSS/D30) provides the overall mean ± SD error for each structure with different number of steps in the training set, while Figure 4 depicts these errors. Briefly, the overall error did not change substantially with different number of steps in the training set.

The importance of predictor variables

Supplemental Table 9 (Supplemental Digital Content, http://links.lww.com/MSS/D30) shows the overall mean ± SD absolute percentage error for each structure (impulse) after removal of selected wearable metrics that were used as input to the neural network. Supplemental Figure 1 depicts these errors (Supplemental Digital Content, http://links.lww.com/MSS/D30). Briefly, the error did not change with removal of each individual input metric.

DISCUSSION

The primary aim of this study was to investigate the accuracy by which commercially available instrumented insoles can predict the patellofemoral stress, tibial stress, and Achilles strain impulse and weighted impulse during running when compared to musculoskeletal modelling. Our findings show that the wearable was able to predict tissue impulses with overall relative percentage errors of 1.95 ± 8.40, -7.37 ± 6.41, and -12.8 ± 9.44% for the patellofemoral joint, tibia and Achilles tendon, respectively. Moreover, while the slope of the individual regression lines differed from the line of identify, we show nearly perfect relative agreement between the predicted and modelled tissue impulse with mean correlations (95% confidence interval) being 0.92 (0.89 to 0.94); 0.95 (0.93 to 0.96); and 0.95 (0.94 to 0.96) for the patellofemoral, tibial, and Achilles tendon stress/strain impulses, respectively.

A separate neural network was trained to predict the weighted impulses. The relative percentage errors were considerably larger for the weighted impulse (i.e., a damage proxy), with values of -39.9 ± 48.6, -53.7 ± 13.6, and 56.9 ± 466% for the patellofemoral joint, tibia and

Achilles tendon, respectively. Nevertheless, while the slope of the individual regression lines again differed from the line of identify, there was still large to nearly perfect relative agreement between the predicted and modelled weighted impulses with mean correlations (95% confidence interval) being 0.90 (0.80 to 0.96); 0.57 (0.38 to 0.71); and 0.97 (0.92 to 0.99) for the patellofemoral joint, tibia, and Achilles tendon, respectively.

Accuracy differences between tissues and between conditions

The accuracy of the predicted impulse and weighted impulse differed between the different tissues (Table 2, Figure 2). Specifically, the overall absolute percentage error for the patellofemoral joint impulse was significantly higher than the absolute percentage error for the tibia (difference of 3.21 ± 0.95 percent points; p <.001), but smaller than the absolute percentage error for the Achilles tendon (difference of -3.07 ± 17.9 percent points; p <.001). There was no significant difference in the absolute percentage error between the tibia and Achilles tendon, although quantitatively the absolute error was higher for the Achilles tendon (difference of 6.28 ± 17.9 percent points; p =.12). Similar findings were obtained for the weighted impulse. The observation that the error at the patellofemoral joint was smaller than at the Achilles tendon, and only slightly larger than the error for the tibia is a notable finding as the wearable is located at the foot but was nevertheless able to predict loading at more proximal joints such as the patellofemoral joint with approximately similar accuracy as distal structures located closer to the insole. This is important for in-field application as may reduce the need to attach potentially cumbersome sensors at the legs, although such sensors can be added to further increase accuracy. In line with these findings, a previous study also found a lower correlation between modelled and predicted Achilles tendon load than modelled and predicted patella tendon load during running

(20). While the authors speculated that the absence of footstrike quantification may have explained their reduced ability to predict Achilles tendon force, our model did include footstrike index, yet showed a similar higher error for the Achilles tendon. A possible reason for these findings is that the Achilles tendon impulse changes relatively more with changes in speed, surface gradient and step frequency than the patellofemoral and tibial impulses (22), thus making it more difficult to accurately estimate the Achilles tendon impulse across all conditions.

The accuracy of the predicted impulse and weighted impulse also differed between conditions such that across all tissues there was typically a trend towards an underestimation of the predicted load when the modelled tissue load increased, and overestimation when the modelled load decreased (Figure 2, Figure 3). For example, uphill running typically decreases the load at the patellofemoral joint relative to level running (Table 2), and while the wearable also predicted a decrease (Table 2 and Supplemental Table 3, Supplemental Digital Content; Figure 3, http://links.lww.com/MSS/D30), the predicted decrease was smaller than the actual decrease. Indeed, the predicted increase was approximately 50% of the actual increase (Supplemental Table 6, Supplemental Digital Content, http://links.lww.com/MSS/D30). A similar effect occurred for the Achilles tendon where steeper uphill slopes increased the Achilles tendon impulse, with the wearable predicting a ~50% smaller increase than the actual increase. This effect can also be seen in Figure 3 where the slope of the regression line for each individual is less steep than the line of identity, and the regression lines also intersect with the line of identity, thus indicating overestimation at relatively lower loads and underestimation at relatively higher loads. These findings suggests that changes in the force vector directions may not fully be captured with the current set of spatiotemporal metrics. While the application of a generalized

correction factor that adjusts the predicted load as a function of its magnitude (i.e., towards lower values at relatively small impulse values or towards higher values at relatively larger impulse values) slightly improved the relative percentage error (i.e., agreement at a group level), it substantially increased the absolute percentage error (i.e., agreement at an individual level), thus offering no improvement over the original machine learning model outcome (Supplemental Figure 2, Supplemental Digital Content, http://links.lww.com/MSS/D30). This was expected as the machine learning model minimized the absolute error, not the relative error. Personalized correction factors may therefore be required to decrease the absolute error. Nevertheless, despite the smaller increase or decrease with changes in speed, surface gradient or step frequency, there were large to nearly perfect correlations between the modelled and predicted impulse and weighted impulse at an individual level (Figure 3). This indicates that relative *changes* in tissue load within an individual across different running conditions can be detected, despite a systematic bias in the absolute predicted load. For example, the wearable did predict a decrease in tibial stress in downhill running as compared to level running, similar to the modelled load, whereas patellofemoral stress was simultaneously predicted to increase with steeper downhill slopes, similar to the modelled load. This suggests the wearable may capture how changes in running technique with the aim to offload one tissue, can increase load and thus injury risk at another tissue (22). Similarly, the wearable predicted a decrease in load on all tissues when increasing step frequency, thus suggesting potential for in-field gait retraining. However, the wearable may not be able to accurately assess all gait retraining strategies. A slight forward trunk lean is for example expected to decrease patellofemoral joint loading as opposed to an upright trunk (53). While our modelled patellofemoral impulse also decreased with forward trunk lean in comparison with the self-selected trunk position (see Table 2, 3.33 $m \cdot s^{-1}$ vs trunk lean; $p = .033$),

the wearable predicted patellofemoral impulse trended to increase with forward trunk lean compared with the self-selected trunk position (Table 2; p = .46). This trend for an increase is likely because the forward trunk lean also slightly increased step length, contact time and footstrike index, all of which would normally be associated with higher patellofemoral joint loading (e.g., as seen when adopting a lower step frequency). An additional sensor at the trunk (e.g., embedded within a heart rate belt) may therefore be required to also accurately capture gait-retraining strategies that modify trunk position.

Despite the relatively high absolute percentage errors in particular during graded running, the model represents an improvement over methods that assume a constant load per step (i.e., as derived from a step counter), or methods that assume a constant load for a given speed or slope for each runner. For example, when we compared the modelled and predicted impulse or weighted impulse when imputing the average modelled impulse or weighted impulse for each condition for each individual, the absolute errors were substantially higher than the absolute percentage errors obtained with the neural network, with errors of 27.4%, 15.8%, and 29.3% for the patellofemoral joint, tibia and Achilles tendon impulse, respectively (Supplemental Table 3, Supplemental Digital Content, http://links.lww.com/MSS/D30). A similar effect was seen when assuming a constant load per step (Supplemental Table 3, Supplemental Digital Content, http://links.lww.com/MSS/D30). The improvement over existing methods becomes even more pronounced when comparing the errors obtained with the modelled and imputed weighted impulse (Supplemental Table 4 vs Supplemental Table 5, Supplemental Digital Content, http://links.lww.com/MSS/D30), with the average absolute errors ranging from 60% to >1000%.

These findings therefore indicate that the machine learning model can better predict tissue load than methods that are currently used to quantify loading.

Overall, the errors for the weighed impulse were larger compared to the (unweighted) impulse (Table 2 vs Supplemental Table 4, Supplemental Digital Content, http://links.lww.com/MSS/D30). This indicates that it is more challenging for the machine learning model to predict the weighted impulse compared to the (unweighted) impulse. The peak force on a structure has a larger relative contribution to the weighted impulse than the contact time, whereas the (unweighted) impulse is more equally determined by both the time and magnitude of the force (22). For example, re-analyses of our data (22) shows typically larger correlations between contact time and the impulse for each structure across all conditions (r = 0.01, 0.69, and 0.42 for the patellofemoral joint, tibia and Achilles tendon, respectively), than the corelations between contact time and the weighted impulse (r = -0.06, 0.05, and 0.31). The larger error for the weighted impulse therefore suggests that the wearable was better able to estimate the time component than the magnitude component of the tissue impulse. This can also be expected from the input metrics as the temporal metrics such as contact time likely provide a reasonable reflection of the time component of the tissue impulse, while being less directly related to the magnitude of the load. Future research may therefore explore if the incorporation of kinetic data such as the peak pressure or pressure-time integral improves the magnitude of the weighted impulse. Another potential reason for the larger error for the weighted impulse is that noise in the modelled tissue load is enlarged by the exponentiating procedure, thereby also reducing the ability to predict the weighted impulse. In this regard, one interesting observation was that the machine learning model predicted a constant weighted impulse for two individuals

at the each joint across all conditions, resulting in poor absolute and relative agreement for these individuals (e.g., correlations of r <0.1). The range within the training data at each condition was similar to the test data from these participants, thus suggesting there was no overfitting on extreme values. For example, Supplemental Figure 3 (Supplemental Digital Content, http://links.lww.com/MSS/D30) shows that the range in values within the test data for these subjects across all conditions was similar to the range within the training dataset. However, the data from the participants had a low variability in this metric as indicated by most values showing a smaller variability across conditions. A similar lower variability was seen for some other metrics. This low variability may reduce the ability of the neural network to differentiate between conditions and thus reduce the ability to accurately predict changes in load across conditions for these two individuals.

Comparison with previous wearable prediction models

The machine learning model developed in the present study can better predict tissue load than most previously developed wearable prediction models. For example, Brund and colleagues (18) used the output from a Garmin watch combined with an accelerometer embedded in a heart rate strap in a regression model to predict the patellar and Achilles tendon impulse with overall absolute errors of 14% and 25%, respectively. Similarly, Rasmussen and co-workers (20) observed a root-mean squared error of 15-20% for patella and Achilles tendon force prediction during running based on idealized inertial measurement units attached at various anatomical locations. These errors are (slightly) higher than the mean absolute percentage errors for the patellofemoral joint and Achilles tendon impulse with our wearable (12.4% and 15.4%, respectively). Notably, our study also included a larger variety of conditions such as higher

speeds and surface gradients, and the error was typically even smaller during level-only running at relatively slow speeds (Table 2). Another study used an artificial neural network to estimate knee joint forces during various sports movements including running, based on data obtained by two inertial measurement units attached to the right leg (17). In comparison with the inverse dynamics approach, the neural network relative root mean square error ranged from 20-40% for the vertical, anterior-posterior, and medial-lateral net knee joint forces, which is also higher than the overall absolute percentage errors for the patellofemoral joint impulse in our study (Table 2). Notably, the lower error at the patellofemoral joint in the present study was achieved despite the instrumented insole being further away from the patellofemoral joint than the two inertial measurement units. The accuracy of the predicted tibial impulse is slightly lower than a previous study that predicted tissue loading during running. Specifically, Elstub and co-workers (19) used LASSO regression to predict tibial axial force from data derived from pressure-sensitive insoles with a shoe-mounted inertial measurement unit and showed this to yield a mean absolute percentage error of 5.7%, as opposed to an absolute percentage error of 9.18% in our study. However, the researchers performed additional calibration to achieve this error, whereas our model did not require the participant to perform additional calibration procedures. Without additional calibration, the mean absolute percentage error observed previously was higher (11.3%) than the error in our study. Further, the axial tibial load may be more easily estimated from a pressure-sensitive insole than the bending moments and thus total tibial stress modelled in the present study, as the peak axial force likely has a more direct relationship with peak normal/vertical ground reaction force than the bending moment. Indeed, re-analyses of our data showed the axial tibial impulse showed a stronger (albeit still moderate) repeated-measures correlation with the peak vertical ground reaction force across all conditions (r_m = -0.44) than the

total tibial stress, that includes the contributions of the bending moment (r_m = 0.23). Further, whereas Elstub and co-workers (19) used 'raw' data (e.g., maximum normal plantar force, center of pressure) as input to the regression model, we used solely spatiotemporal metrics computed using previously validated proprietary algorithms (29). While some likely relevant raw data is discarded by using spatiotemporal metrics, the algorithms for computing these metrics employed various filtering techniques which may have reduced the influence of noise that is typically present in raw signals. This may in turn have improved the accuracy of the present model as compared to the model without additional calibration in (19). Similarly, the use of spatiotemporal metrics in the present study likely made the input less sensitive to differences in insole size or shoe lacing that all can influence the raw pressure recordings (54). Nevertheless, an idealized instrumented insole (i.e., simulated wearable signal without noise) can predict tibial axial force with a mean absolute percentage error of 2.6% (9), thus suggesting potential for further improvement. Note however, that this error reflects only the axial tibial force component, and not the total tibial stress which includes bending as used in the present study. Further, this error was achieved only when using both pressure and foot inertial-measurement unit data, whereas an error 4.7% was achieved when using only pressure data. Because the metrics used as input to the neural network in the present study were computed solely using pressure data (as detailed in (29)), and because the absolute percentage error for the tibial impulse in the present study is larger than 4.7% (i.e., 9.18%), further improvements in accuracy may be achieved by a) also using foot position data from the foot-mounted inertial measurement unit already present in the wearable, b) by further improvements in noise reductions in raw signals, and c) by calibrating the sensor for each individual.

The differences in accuracy between studies that estimated tissue loading from wearables might also reflect differences in the number of datapoints used to train the model and the exact statistical model used. For example, some previous studies used linear (multiple) regression analysis (e.g., (18)), and it could be argued that this analysis is not able to capture potential non-linear relationships between the input/predictors and output variables as accurately as the artificial neural network employed in the present study. Yet, other studies showed substantial errors in tissue load predictions even with a neural network (17), thus suggesting that other factors are also/more important. One of these factors could be related to the differences in the size of the dataset. For example, Brund and colleagues (18) used data from 576 running strides as input to the regression model, whereas we used on average 28,321 steps. This number of datapoints is also higher than Elstub and co-workers (19) who used data from 10 strides across 41 different conditions in 9 participants, thus yielding 3280 (3690 minus data from one participant) steps for training purposes. However, when we re-trained our model with fewer steps, the error did not substantially change, even when only one step per participant, per condition (i.e., 216 in total), was used (Figure 4). This suggests that the size of the training set was not the primary reason for the observed differences.

The absence of substantial changes in the error with smaller training datasets may reflect the relatively large number of total steps included even when including only one step per subject per condition (i.e., $1 \times 18 \times 12 = 216$ steps). However, the error remained also largely similar when including only 100 randomly selected steps across all conditions. Modelling studies have shown that more noise in the input data reduces the training data size requirements because there is less useful data from which the neural network can learn (55). The absence of changes in error

with a larger number of steps may therefore indicate the presence of "noise" in the input metrics, which in this scenario is likely introduced by the indirect relationship between spatiotemporal metrics and tissue loading. Specifically, some variability in tissue loading caused by for example changes in knee flexion angle may not be completely reflected in spatiotemporal metrics, thus reducing the ability to learn. A second potential explanation for the absence of changes in error with a larger number of steps is the use of data from multiple steps to compute spatiotemporal metrics at a given step. Specifically, some input metrics (i.e., cadence, inter-step variability) required data from multiple steps and the wearable also employed a moving average filter to remove outliers in single-step spatiotemporal data (29). Therefore, spatiotemporal metrics from a single-step contained information from multiple steps, thus potentially also contributing to the absence of changes in error with a different number of steps.

Importance of input metrics

We explored the importance of different input variables to the obtained accuracy by removing one input variable (e.g., contact time) and re-computing the mean difference between the modelled and predicted load over all leave-one-out validations. This analysis revealed that the accuracy was largely maintained when removing each metric (Supplemental Table 8, Supplemental Figure 1, Supplemental Digital Content, http://links.lww.com/MSS/D30). The low sensitivity of the model's performance to the removal of temporal metrics such as contact, flight, and swing time may be because these metrics are providing partially redundant information such that the removal of one of these metrics can still partly be compensated by the information provided by the remaining metrics. Similarly the low sensitivity to surface slope and running speed as input metrics suggests that plantar pressure data alone (and the spatiotemporal metrics

computed from this) likely contains sufficient information to discriminate between running on different slopes and speeds. In line with this, a previous study also found that removing speed and slope as input to a machine learning model had a negligible effect on its ability to predict ground reaction forces from pressure insoles (44). Consistent with the small changes in the prediction error with different number of steps as input, the low sensitivity to removal of input metrics may also reflect noise in the input signal, or variability in the output that was not captured by the input, thus reducing the sensitivity of the neural network to one particular variable for learning. The low sensitivity of the model's performance to the removal of a single input variable suggests that the number of inputs may be reduced. While it is beyond the scope of this study to comprehensively investigate the minimum input requirements, we explored the prediction error with just three inputs that may be easily obtained using various wearable sensors: contact time, step frequency, and footstrike angle. With these inputs, the absolute percentage errors increased from 12.4%, 9.18%, and 15.4% to 29.0%, 11.5%, and 21.6% for the patellofemoral joint, tibia, and Achilles tendon, respectively (Supplemental Table 10, Supplemental Digital Content, http://links.lww.com/MSS/D30). This indicates that removal of a larger number of inputs *does* reduce the model's performance, in particular for more proximal structures such as the patellofemoral joint.

Strengths and limitations

Strengths of this study include the simultaneous prediction of tissue loading at three common injury locations as opposed to one location in most previous studies, and the addition of weighted loading metrics to better reflect tissue fatigue-life and thus injury risk. A further strength is the large range of experimental conditions and diverse sample that both improve the

generalizability to in-field conditions among various running populations. A final major strength is the use of a commercially available wearable insole that allows the developed model to be directly implemented in-field using a mobile phone application (Figure 5).

This study also has several limitations that should be kept in mind when interpreting the results. First, the modelling approach relied on numerous assumptions (e.g., maximum isometric force (56), muscle moment arms (56), tendon slack length, optimization function (57)) that can influence the accuracy of the modelled tissue load, and therefore also the predicted load. Nevertheless, our previous study (22) showed close agreement between modelled and experimentally measured Achilles tendon strain, between modelled and measured muscle activation, and between measured and modelled vastus lateralis fascicle behavior, thus overall increasing the confidence in the modelled loads. Yet, there are several simplifications to the model that should be considered. The method used to determine contact area for the patellofemoral joint as a function of knee angle does for example not consider differences in contact area due to internal knee rotation (e.g. 58). Similarly, the tibial loading model does only consider axial and bending forces, while *in vivo* tibial loading is likely characterized by a more complex combination of loads (e.g., shear forces). Such information may be incorporated in future models to improve the accuracy of the estimated load and damage.

A second limitation relates to the individuality of the predicted load. Specifically, our modelling approach assumed all individuals are using the same muscle activation strategy for muscle force production. However, this may be altered for example in the presence of fatigue or pain. Future wearables may incorporate wearable muscle activation recording methods to further

individualize the muscle activation strategy to capture these effects (e.g., (59, 60)). Moreover, we used a similar generic tibial cross-sectional area and Achilles tendon stiffness for all individuals and used a generic knee contact area that differed only between sexes. Yet, these characteristics are known to vary between individuals and this can affect the stress/strain and thus damage for a given load. Therefore, the stress/strain predicted with the machine learning algorithm is only partly specific to the individual and may overestimate load (and damage) for individuals with a large cross-sectional area or stiff tendons, while underestimating load and damage for those with smaller cross-sectional area and relatively compliant tendons. Similarly, females have a smaller transverse plane tibial cross-sections and lower resistance to bending (61), which increases their injury risk for a given load. Future work is therefore needed to further individualize the predicted stress/strain. This can be done by 1) (periodically) measuring subject-specific information (e.g., MRI scan at a local medical facility to obtain bone geometry and microarchitecture (62)), 2) continuously measuring relevant parameters (e.g., tendon stiffness using wearable ultrasound (59, 63)), or 3) using statistical shape modelling (61, 64) (but see also (65)) or regression equations to estimate relevant parameters based on other characteristics that might be provided as user information in the wearable application.

A final limitation relates to the neural network approach used. First, neural networks are known to not always generalize well to conditions beyond the training data conditions. Because we used only periods of constant-speed unidirectional running at one specific surface (albeit at various speeds, slopes and with varying step frequencies), it is unknown how accurate the estimated tissue load are during periods of accelerations, during turns, or speeds, slopes and

surfaces beyond those tested in this study. Nevertheless, distance running is characterized by a quasi-constant speed, and the slopes investigated in this study reflect the slopes typically encountered in-field. Further, we did not standardize shoe wear and the differences in shoe cushioning between participants may partly mimic differences in surface stiffness, thus also increasing generalizability. Finally, uphill and downhill running have some similarity to acceleration and deceleration (66), thus potentially also improving generalizability to these conditions. The generalizability of the model may however be further improved by including a physics-based approach in future studies. Such an approach explicitly models important biomechanical principles within the neural network (e.g., (67)) and can therefore use these biomechanical principles to predict tissue loading in a wider variety of conditions, and in different tasks where spatiotemporal metrics may not be useful (e.g., squatting). The accuracy and predictive ability may also be further improved by changes in the specific neural network used. For example, a recurrent neural network (such as a long short-term memory network (68-70)) can use information from prior inputs along with the current input to calculate an output and has been shown to better predict biomechanical time-series data than other neural network approaches (71). Similarly, a convolutional neural network can learn to exploit the correlation between different data points in a time-series and has been shown to also outperform the multilayer perceptron network used in our study to predict kinematics during walking (70). Finally, we used stance-only impulses as the output. Although the stance phase is characterized by the highest load on most tissues, it may also be important to include the load acting on each tissue during the flight phase for a more comprehensive assessment of tissue loading.

Implications

The findings of this study have several implications for both researchers, clinicians, runners, and wearable tech developers. The wearable may be used by clinicians to provide more detailed information on tissue loading with different gait retraining strategies than can be obtained from standard 2D video analysis. For example, the wearable may be used to assess the effect of step frequency modifications on (cumulative) patellofemoral, tibial or Achilles tendon load. The instrumented insoles and developed machine learning model may also be used by researchers in large-scale in-field studies to establish relations between (changes in) cumulative tissue load and injuries. This information may in turn be used to develop interventions that reduce injury risk by modifying tissue load, for example using real-time feedback on running technique or running speed. Moreover, a combination of (individualised) tissue loads obtained from the wearable and periodic measurements of tissue integrity or biomarkers related to tissue repair and adaptation can allow for more detailed study of the tissue-specific repair time after an applied mechanical load *in vivo* in humans. Up to now, such studies typically required invasive animal studies that allow estimation of the mechanical load and damage (e.g., (72, 73)). Finally, by combining the present model with probabilistic functions for injuries and models of repair and adaptation (e.g., (13, 14, 16, 48-50)), the wearable may also be used to inform runners on an appropriate training program, for example by suggesting rest days or training modifications. It is important to note in this context that the weighted impulse may be better able to capture injury risk than the non-weighted impulse because it accounts for the non-linear relationship between a given load and the damage incurred (6, 15, 22). The absolute percentage errors for the weighted impulse were however considerably higher (in the order of 58-92%; Supplemental Table 4, Supplemental Digital Content, http://links.lww.com/MSS/D30). While previous research (and

our study, see later) has shown that errors in damage of at least 41% can still capture trends in higher vs lower distance workouts (9), it remains unknown if this level of accuracy is sufficient for specific applications. Nevertheless, while the errors in the impulse and in particular weighted impulse may seem large, these estimates are substantially more accurate than currently employed external load metrics such as the distance ran or number of steps (9, 18), as also shown in our study. Moreover, we also show large to nearly perfect relative agreement between the modelled and predicted weighted impulse (Figure 3), suggesting that relative *changes* in damage within an individual can still be captured.

To give some indication on the performance of the machine learning model in a real-world context, Figure 6 provides an example of how the predicted patellofemoral cumulative weighted impulse agrees with the modelled patellofemoral cumulative weighted impulse across different simulated training sessions for a recreational runner. The patellofemoral joint was chosen in this example because this is the most common injury location in recreational runners (3-5). Further details on the simulation approach can be found in Supplemental Table 11 (Supplemental Digital Content, http://links.lww.com/MSS/D30). Most importantly, the machine learning predicted cumulative weighted impulse captured increases and decreases in cumulative load across some, but not all sessions, and overall underestimated the cumulative load in this specific example. For example, session three represented an interval running session with a total of 4 km ran at 4 $m \cdot s^{-1}$, and 3 km at 2.78 $m \cdot s^{-1}$. While the musculoskeletal modelling approach predicted an increase in the cumulative weighted impulse for this session as compared to session one and two (which reflected steady-state runs at 2.78 $m \cdot s^{-1}$), the machine learning model predicted a slight decrease compared to the same sessions, and thus not captured the increase in

damage and injury risk (27, 28) with higher speed interval running. A partly similar effect was observed for session 6 (which reflected a simulated 5-km race at 5 $m \cdot s^{-1}$, with 4 km warm-up/cool-down at 2.78 $m \cdot s^{-1}$), with the modelled cumulative load increasing compared to most other sessions involving a steady-state run, but the machine learning predicted cumulative load being largely similar to other sessions. The machine learning model did however predict the cumulative load to increase during session eight, which involved repeated uphill and downhill intervals, and during session 10, which involved a 15-km long-run. Importantly, the machine learning approach also performed substantially better across all sessions than a method that measures only step count/distance (Figure 6).

Finally, our findings also have implications for wearable developers. Specifically, spatiotemporal metrics as opposed to 'raw' data such as pressure or accelerations were used as input to the artificial neural network and some of these metrics can also be obtained from other wearables such as an inertial measurement unit attached to the shoe, integrated within a chest belt, or even within a sports watch. Future studies may therefore explore if tissue loading at these injury locations can also be predicted from other wearables. However, the wearable used in the present study has been shown to measure spatiotemporal metrics with an overall difference of -0.09% as compared to an instrumented treadmill during the conditions employed in this study (29). This high accuracy likely contributed to the relatively high accuracy for the predicted tissue load compared to other studies, and it is likely that a lower accuracy in spatiotemporal metrics obtained with most other wearables presently on the market would introduce additional noise in the input, thereby reducing the accuracy of the predicted tissue load. Moreover, further research is required to determine which spatiotemporal metrics are most important to optimize the

accuracy. Finally, future studies may also explore the utility of incorporating kinetic predictor variables (e.g., peak pressure, pressure-time integral) to improve the tissue load prediction.

CONCLUSIONS

We show that a machine learning algorithm that uses spatiotemporal metrics obtained from an instrumented insole as input, can predict loading at common running injury locations with overall absolute percentage errors of 12.4%, 9.18%, and 15.4%, for the patellofemoral joint, tibia, and Achilles tendon impulses, respectively. Further, relative agreement was nearly perfect with mean correlations being 0.92, 0.95, and 0.95 for the patellofemoral, tibial, and Achilles tendon stress/strain impulses, respectively. While the errors were larger for the weighted impulse, the relative agreement was also large to nearly perfect. The developed model may be used to quantify in-field tissue load at multiple common running injury locations, to assess the effects of gait retraining on tissue loading, and it may be used in combination with real-time feedback in large-scale studies to reduce injury risk more effectively. Finally, the model also opens opportunities to study the association between mechanical load and tissue repair and adaptation in humans *in vivo*. Further work is required to incorporate subject-specific parameters for tissue properties and remodelling.

Acknowledgements

BVH was funded by an Eurostars grant (ID 12912) awarded by Eureka.

Author contributions

BVH conceived the study, collected and analysed the data, and wrote the first draft of the manuscript, LvR assisted in data processing. KM provided comments and edits. All authors approved the final version.

Conflicts of interest

LvR and KM declare that they have no conflict of interest. BVH was (part-time) employed by Ato-Gear during a period in which this manuscript was written. The results of the study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation. The results of the present study do not constitute endorsement by the American College of Sports Medicine.

Data availability

All data is available upon request.

REFERENCES

1. Clough PJ, Dutch S, Maughan RJ, Shepherd J. Pre-race drop-out in marathon runners: reasons for withdrawal and future plans. *Br J Sports Med.* 1987;21(4):148-9.
2. Koplan JP, Rothenberg RB, Jones EL. The natural history of exercise: a 10-yr follow-up of a cohort of runners. *Med Sci Sports Exerc.* 1995;27(8):1180-4.
3. van Gent RN, Siem D, van Middelkoop M, van Os AG, Bierma-Zeinstra SM, Koes BW. Incidence and determinants of lower extremity running injuries in long distance runners: a systematic review. *Br J Sports Med.* 2007;41(8):469-80; discussion 80.
4. Buist I, Bredeweg SW, Bessem B, van Mechelen W, Lemmink KA, Diercks RL. Incidence and risk factors of running-related injuries during preparation for a 4-mile recreational running event. *Br J Sports Med.* 2010;44(8):598-604.
5. 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.
6. Edwards WB. Modeling overuse injuries in sport as a mechanical fatigue phenomenon. *Exerc Sport Sci Rev.* 2018;46(4):224-31.
7. Matijevich ES, Branscombe LM, Scott LR, Zelik KE. Ground reaction force metrics are not strongly correlated with tibial bone load when running across speeds and slopes: implications for science, sport and wearable tech. *PLoS One.* 2019;14(1):e0210000.
8. Van Hooren B, Goudsmit J, Restrepo J, Vos S. Real-time feedback by wearables in running: Current approaches, challenges and suggestions for improvements. *J Sports Sci.* 2020;38(2):214-30.

9. Matijevich ES, Scott LR, Volgyesi P, Derry KH, Zelik KE. Combining wearable sensor signals, machine learning and biomechanics to estimate tibial bone force and damage during running. *Hum Mov Sci.* 2020;74:102690.

10. Sasimontonkul S, Bay BK, Pavol MJ. Bone contact forces on the distal tibia during the stance phase of running. *J Biomech.* 2007;40(15):3503-9.

11. Scott SH, Winter DA. Internal forces of chronic running injury sites. *Med Sci Sports Exerc.* 1990;22(3):357-69.

12. Zandbergen MA, Ter Wengel XJ, van Middelaar RP, Buurke JH, Veltink PH, Reenalda J. Peak tibial acceleration should not be used as indicator of tibial bone loading during running. *Sports Biomechanics.* 2023:1-18.

13. Edwards WB, Taylor D, Rudolphi TJ, Gillette JC, Derrick TR. Effects of running speed on a probabilistic stress fracture model. *Clin Biomech (Bristol, Avon).* 2010;25(4):372-7.

14. Sinclair J, Huang G, Taylor PJ, Chockalingam N, Fan Y. Effects of running in minimal and conventional footwear on medial tibiofemoral cartilage failure probability in habitual and non-habitual users. *J Clin Med.* 2022;11(24):7335.

15. Firminger CR, Asmussen MJ, Cigoja S, Fletcher JR, Nigg BM, Edwards WB. Cumulative metrics of tendon load and damage vary discordantly with running speed. *Med Sci Sports Exerc.* 2020;52(7):1549-56.

16. Paul E, Pant A, George S, Willson J, Meardon S, Vahdati A. In silico modeling of tibial fatigue life in physically active males and females during different exercise protocols. *Biomed Phys Eng Express.* 2022;8(3):035019.

17. Stetter BJ, Ringhof S, Krafft FC, Sell S, Stein T. Estimation of knee joint forces in sport movements using wearable sensors and machine learning. *Sensors (Basel).* 2019;19(17):3690.

18. Brund RB, Waagepetersen R, O Nielsen R, et al. How precisely can easily accessible variables predict Achilles and patellar tendon forces during running? *Sensors (Basel).* 2021;21(21):7418.

19. Elstub LJ, Nurse CA, Grohowski LM, Volgyesi P, Wolf DN, Zelik KE. Tibial bone forces can be monitored using shoe-worn wearable sensors during running. *J Sports Sci.* 2022;40(15):1741-9.

20. Rasmussen J, Skejø S, Waagepetersen RP. Predicting tissue loads in running from inertial measurement units. *Sensors (Basel).* 2023;23(24):9836.

21. Baggaley M, Derrick TR, Vernillo G, Millet GY, Edwards WB. Internal tibial forces and moments during graded running. *J Biomech Eng.* 2022;144(1):011009

22. Van Hooren B, Van Rengs L, Meijer K. Per-step and cumulative load at three common running injury locations: the effect of speed, surface gradient and cadence. *Scand J Med Sci Sports.* 2024;34(2):e14570.

23. Carter DR, Caler WE, Spengler DM, Frankel VH. Fatigue behavior of adult cortical bone: the influence of mean strain and strain range. *Acta Orthop Scand.* 1981;52(5):481-90.

24. Wren TA, Lindsey DP, Beaupre GS, Carter DR. Effects of creep and cyclic loading on the mechanical properties and failure of human Achilles tendons. *Ann Biomed Eng.* 2003;31(6):710-7.

25. Firminger CR, Edwards WB. Effects of cyclic loading on the mechanical properties and failure of human patellar tendon. *J Biomech.* 2021;120:110345.

26. Haider IT, Lee M, Page R, Smith D, Edwards WB. Mechanical fatigue of whole rabbit-tibiae under combined compression-torsional loading is better explained by strained volume than peak strain magnitude. *J Biomech.* 2021;122:110434.

27. Junior LCH, Costa LOP, Lopes AD. Previous injuries and some training characteristics predict running-related injuries in recreational runners: a prospective cohort study. *J Physiother.* 2013;59(4):263-9.

28. McCrory JL, Martin DF, Lowery RB, et al. Etiologic factors associated with Achilles tendinitis in runners. *Med Sci Sports Exerc.* 1999;31(10):1374-81.

29. Van Hooren B, Willems P, Plasqui G, Meijer K. The accuracy of commercially available instrumented insoles (ARION) for measuring spatiotemporal running metrics. *Scand J Med Sci Sports.* 2023;33(9):1703-15.

30. van den Bogert AJ, Geijtenbeek T, Even-Zohar O, Steenbrink F, Hardin EC. A real-time system for biomechanical analysis of human movement and muscle function. *Med Biol Eng Comput.* 2013;51(10):1069-77.

31. Van Hooren B, Fuller JT, Buckley JD, 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(4):785-813.

32. Dos Santos AF, Nakagawa TH, Serrão FV, Ferber R. Patellofemoral joint stress measured across three different running techniques. *Gait Posture.* 2019;68:37-43.

33. Heiderscheit BC, Chumanov ES, Michalski MP, Wille CM, Ryan MB. Effects of step rate manipulation on joint mechanics during running. *Med Sci Sports Exerc.* 2011;43(2):296-302.

34. Delp SL, Anderson FC, Arnold AS, et al. OpenSim: open-source software to create and analyze dynamic simulations of movement. *IEEE Trans Biomed Eng.* 2007;54(11):1940-50.

35. Catelli DS, Wesseling M, Jonkers I, Lamontagne M. A musculoskeletal model customized for squatting task. *Comput Methods Biomech Biomed Engin.* 2019;22(1):21-4.

36. De Groote F, Kinney AL, Rao AV, Fregly BJ. Evaluation of direct collocation optimal control problem formulations for solving the muscle redundancy problem. *Ann Biomed Eng.* 2016;44(10):2922-36.

37. Steele KM, Demers MS, Schwartz MH, Delp SL. Compressive tibiofemoral force during crouch gait. *Gait Posture.* 2012;35(4):556-60.

38. Besier TF, Draper CE, Gold GE, Beaupre GS, Delp SL. Patellofemoral joint contact area increases with knee flexion and weight-bearing. *J Orthop Res.* 2005;23(2):345-50.

39. Milgrom C, Zloczower E, Fleischmann C, et al. Medial tibial stress fracture diagnosis and treatment guidelines. *J Sci Med Sport.* 2021;24(6):526-30.

40. Derrick TR, Edwards WB, Fellin RE, Seay JF. An integrative modeling approach for the efficient estimation of cross sectional tibial stresses during locomotion. *J Biomech.* 2016;49(3):429-35.

41. Oliveira AS, Pirscoveanu CI. Implications of sample size and acquired number of steps to investigate running biomechanics. *Sci Rep.* 2021;11(1):3083.

42. Riazati S, Caplan N, Hayes PR. The number of strides required for treadmill running gait analysis is unaffected by either speed or run duration. *J Biomech.* 2019;97:109366.

43. Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification. Taipei, Taiwan; 2003.

44. Honert EC, Hoitz F, Blades S, Nigg SR, Nigg BM. Estimating running ground reaction forces from plantar pressure during graded running. *Sensors (Basel).* 2022;22(9):3338.

45. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv*. arXiv:14126980. 2014.

46. Liew BXW, Rugamer D, Mei Q, et al. Smooth and accurate predictions of joint contact force time-series in gait using over parameterised deep neural networks. *Front Bioeng Biotechnol.* 2023;11:1208711.

47. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res.* 2014;15(1):1929-58.

48. Pyles CO, Dunphy M, Vavalle NA, et al. Longitudinal tibia stress fracture risk during high-volume training: a multiscale modeling pipeline incorporating bone remodeling. *J Biomech Eng.* 2022;144(10):101002.

49. Young SR, Gardiner B, Mehdizadeh A, Rubenson J, Umberger B, Smith DW. Adaptive remodeling of Achilles tendon: a multi-scale computational model. *PLoS Comput Biol.* 2016;12(9):e1005106.

50. Dimnik JM, Haider IT, Edwards WB. A continuum damage model of fatigue and failure in whole bone. *J Mech Behav Biomed Mater.* 2023;143:105907.

51. Shieh G. Assessing agreement between two methods of quantitative measurements: exact test procedure and sample size calculation. *Stat Biopharm Res.* 2020;12(3):352-9.

52. Hopkins WG. A scale of magnitudes for effect statistics: NewStats; 2002 [cited 15 1-3]. Available from: newstats.org/effectmag.html.

53. Teng HL, Powers CM. Sagittal plane trunk posture influences patellofemoral joint stress during running. *J Orthop Sports Phys Ther.* 2014;44(10):785-92.

54. Harrison K, Vladika B, Feeney D. Use of in-shoe pressure to quantify running shoe fit. *Footwear Sci.* 2021;13(Suppl 1):S71-2.

55. Alwosheel A, van Cranenburgh S, Chorus CG. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *J Choice Model.* 2018;28(C):167-82.

56. Nasab SHH, Smith CR, Maas A, et al. Uncertainty in muscle-tendon parameters can greatly influence the accuracy of knee contact force estimates of musculoskeletal models. *Front Bioeng Biotechnol.* 2022;10:808027.

57. Wagner H, Bostrom KJ, de Lussanet MHE, de Graaf ML, Puta C, Mochizuki L. Optimization reduces knee-joint forces during walking and squatting: validating the inverse dynamics approach for full body movements on instrumented knee prostheses. *Motor Control.* 2023;27(2):161-78.

58. Liao TC, Keyak JH, Powers CM. Runners with patellofemoral pain exhibit greater peak patella cartilage stress compared with pain-free runners. *J Appl Biomech.* 2018;34(4):298-305.

59. Frey S, Kartsch V, Leitner C, et al. Ultrasound system for long-term muscle activity monitoring. *arXiv.* 2023;arXiv:2309:06851.

60. Lyons NR, Worsey MTO, Devaprakash D, et al. Washable garment-embedded textile electrodes can measure high-quality surface EMG data across a range of motor tasks. *IEEE Sens J.* 2023;23(17):20150-8.

61. Bruce OL, Baggaley M, Khassetarash A, Haider IT, Edwards WB. Tibial-fibular geometry and density variations associated with elevated bone strain and sex disparities in young active adults. *Bone*. 2022;161:116443.

62. O'Leary TJ, Rice HM, Greeves JP. Biomechanical basis of predicting and preventing lower limb stress fractures during arduous training. *Curr Osteoporos Rep.* 2021;19(3):308-17.

63. Hu H, Huang H, Li M, et al. A wearable cardiac ultrasound imager. *Nature*. 2023;613(7945):667-75.

64. Bruce OL, Baggaley M, Welte L, Rainbow MJ, Edwards WB. A statistical shape model of the tibia-fibula complex: sexual dimorphism and effects of age on reconstruction accuracy from anatomical landmarks. *Comput Methods Biomech Biomed Engin.* 2022;25(8):875-86.

65. Bruce OL, Tu J, Edwards WB. Predicting tibia-fibula geometry and density from anatomical landmarks via statistical appearance model: influence of errors on finite element-calculated bone strain. *J Biomech Eng.* 2024;146(9):091005.

66. di Prampero PE, Fusi S, Sepulcri L, Morin JB, Belli A, Antonutto G. Sprint running: a new energetic approach. *J Exp Biol.* 2005;208(Pt 14):2809-16.

67. Ma T, Xu X, Chai Z, Wang T, Shen X, Sun T. A wearable biofeedback device for monitoring tibial load during partial weight-bearing walking. *IEEE Trans Neural Syst Rehabil Eng.* 2023:31:3428-36.

68. Donahue SR, Hahn ME. Estimation of gait events and kinetic waveforms with wearable sensors and machine learning when running in an unconstrained environment. *Sci Rep.* 2023;13(1):2339.

69. Mundt M, Koeppe A, Bamer F, David S, Markert B. Artificial neural networks in motion analysis-applications of unsupervised and heuristic feature selection techniques. *Sensors (Basel).* 2020;20(16):4581.

70. Mundt M, Johnson WR, Potthast W, Markert B, Mian A, Alderson J. A comparison of three neural network approaches for estimating joint angles and moments from inertial measurement units. *Sensors (Basel).* 2021;21(13):4535.

71. Burton WS, 2nd, Myers CA, Rullkoetter PJ. Machine learning for rapid estimation of lower extremity muscle and joint loading during activities of daily living. *J Biomech.* 2021;123:110439.

72. Sample SJ, Hao Z, Wilson AP, Muir P. Role of calcitonin gene-related peptide in bone repair after cyclic fatigue loading. *PLoS One.* 2011;6(6):e20386.

73. Liu C, Carrera R, Flamini V, et al. Effects of mechanical loading on cortical defect repair using a novel mechanobiological model of bone healing. *Bone*. 2018;108:145-55.

FIGURE LEGENDS

Figure 1. Artificial multilayer perceptron neural network structure. The input layer consisted of 12 neurons corresponding to 12 predictors derived from the instrumented insole at each step. The neurons in the input layer forwarded the normalized wearable data to the neurons in the next hidden layer and multiplied this by a weight factor. Each receiving neuron summed the data from all incoming neurons and added a bias term. Subsequently, the summed input plus bias was processed by a rectified linear activation function that returned a value of zero for negative values or returned the same value for positive values (see enlargement). This output was then forwarded to the next layer. Hidden layer 1-6 consisted of 21, 31, 51, 81, 101, and 111 neurons, respectively. There was one output neuron for each structure, which was defined by a linear activation function.

The schematic models on the right side depict the approach used to model tissue loading at the three common running injury locations. A) Patellofemoral joint where the compression stress between the femur and patella depends on the force produced by all quadriceps muscles (F_Q), and the knee flexion angle (Θ_{knee}), which alters both the angle of pull between the quadriceps and patella tendon and the contact area. F_{PT} depicts the force vector of the patella tendon, and F_{PFJ} depicts the contact force between the femur and patella. B) Ankle contact force vector and bending moment on the distal third of the tibia. F_G represents the gravitational force acting on the tibia center of mass, and F_{ACF} the ankle contact force vector, which results from both the gravitational force and forces caused by muscle contraction (i.e., medial and, lateral gastrocnemius, soleus, tibialis posterior, flexor digitorum longus, flexor hallucis longus, tibialis anterior, peroneus brevis, longus, and tertius, extensor digitorum longus, and extensor hallucis

longus). The vector of the ankle/tibia contact force will bend the tibia posteriorly, which will create compressive stress at the posterior side of the tibia (green arrows in the box), which is further increased by the gravitational force. C) Achilles tendon force vector (blue arrow) resulting from the sum of all individual triceps surae muscles (i.e., soleus and gastrocnemius lateralis and medialis as depicted by the green arrows).

Figure 2. Relative percentage errors for the patellofemoral joint (left), tibia (center) and Achilles tendon (right) (unweighted) impulse across different conditions. Positive values indicate an overestimation of the stress/strain impulse by the neural network, while negative values indicate an underestimation. *p*-values depict the effect of speed, downhill or uphill slope, or step frequency on the change in the relative percentage error.

Figure 3. Repeated-measures correlations for the patellofemoral joint (left), tibia (center) and Achilles tendon (right) (unweighted) impulse (top rows) and weighted impulse (bottom row) across all participants. Symbols depict individual datapoints and solid lines the best-fitted regression line. The depicted correlation represents the mean correlation magnitude with 95% confidence intervals across all within-subject correlations. The dashed black line presents the line of identity that corresponds to perfect agreement between the lab-based and wearable-based tissue load. Values above the line of identity are overestimated, while values below the line of identity are underestimated relative to the modelled stress or strain impulse or weighted impulse.

Figure 4. Overall relative percentage error per joint impulse as a function of the number of steps used to train the neural network. The thick line presents the mean error over all leave-one-out

validations, while the shaded area depicts the standard deviation. Similar results were observed for the absolute percentage error.

Figure 5. Screenshot from the wearable application showing the tissue load in a model of the human body after a running session, with darker red colors depicting a higher tissue load. Note that the app also predicts tissue load at other anatomical locations based on the muscle forces computed for modelling the tissue loading in the present study. However, the load at these other anatomical locations has not been explicitly validated in this study. The machine learning model employed within the application is the model trained on all 19 individuals.

Figure 6. Example of the agreement between the musculoskeletal model derived (green), machine learning predicted (red), and step meter predicted (grey) patellofemoral weighted impulse across different training sessions (e.g., steady-state runs, race, hill intervals) for a typical recreational runner expressed in original units (left), or expressed as a percentage of the modelled load (right).

SUPPLEMENTAL DIGITAL CONTENT

SDC 1: Suppl file V3.docx

Figure 1

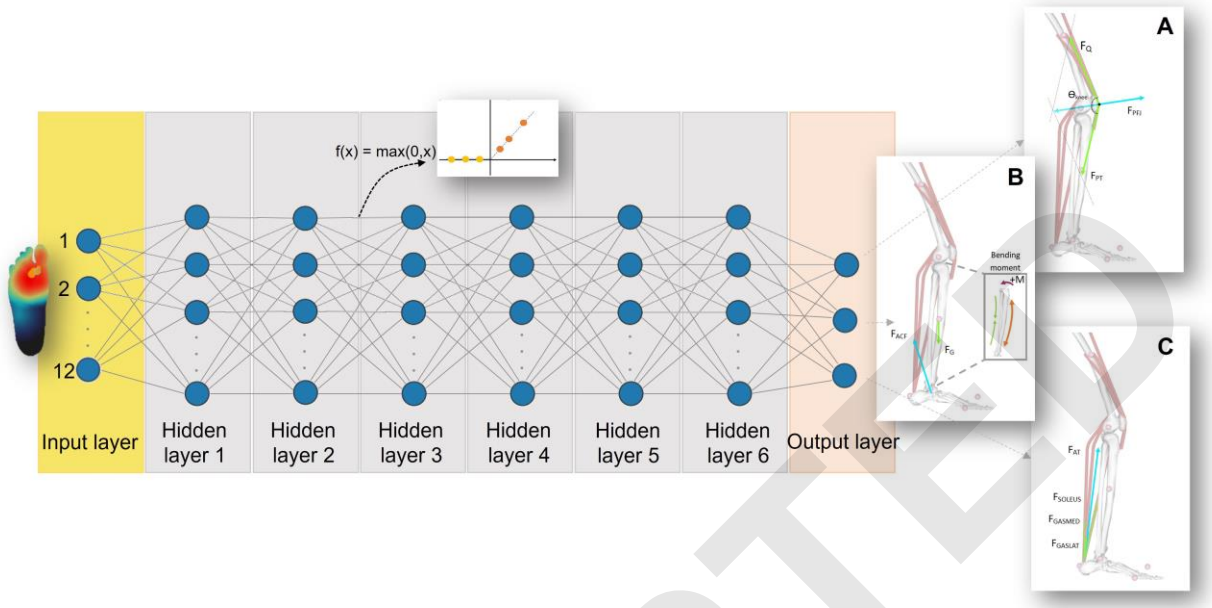

Figure 2

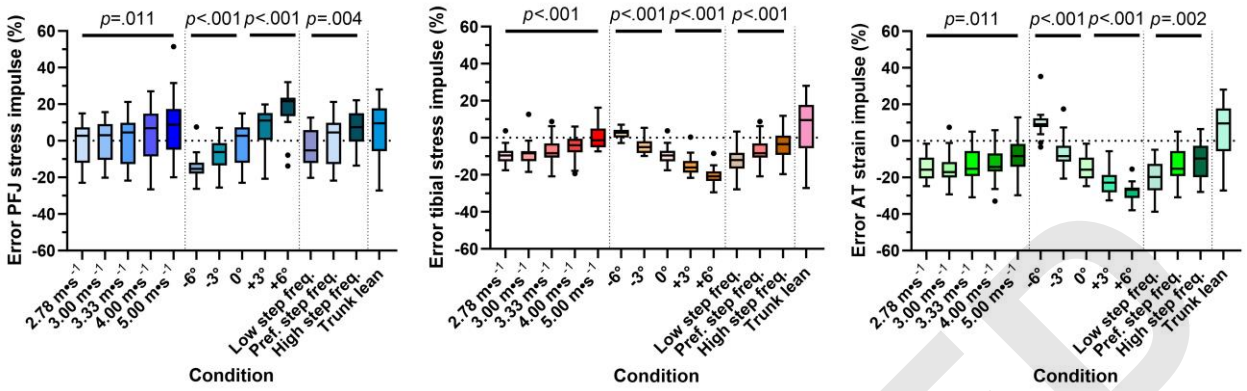

Figure 3

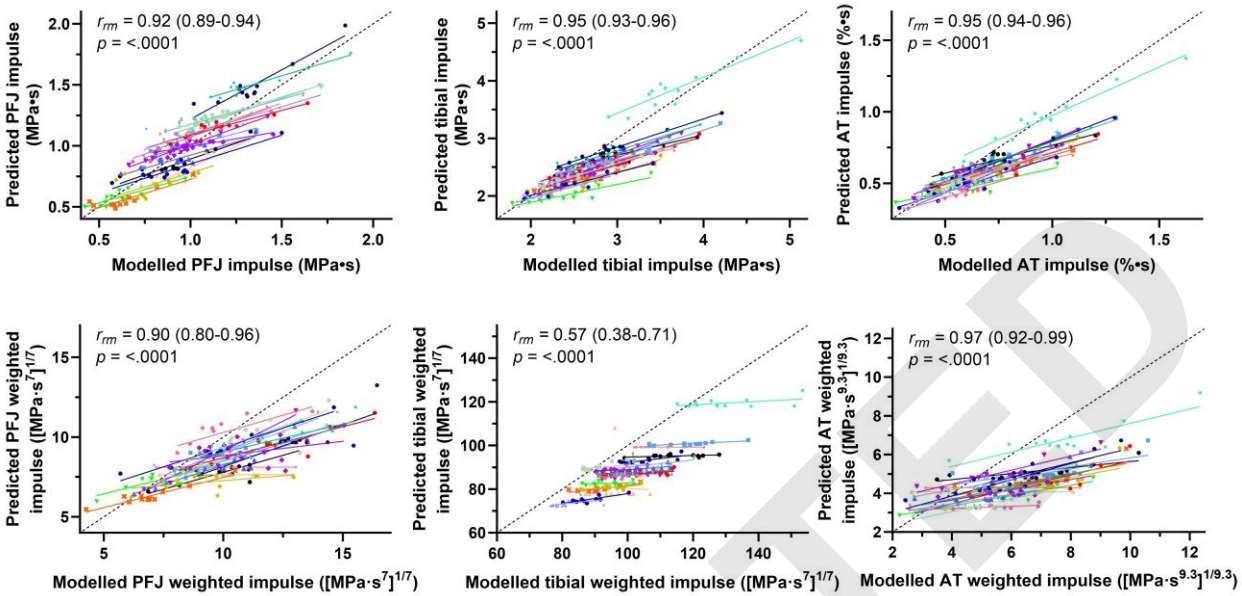

Figure 4

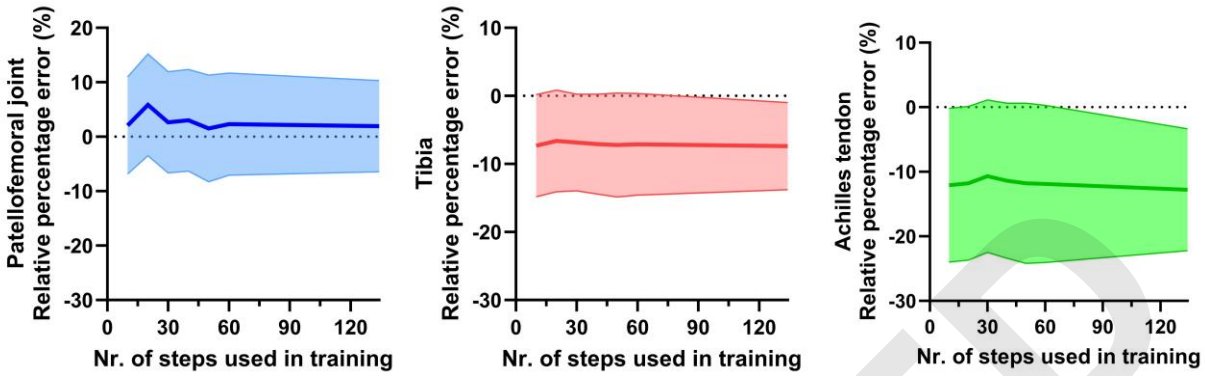

Figure 5

Figure 6

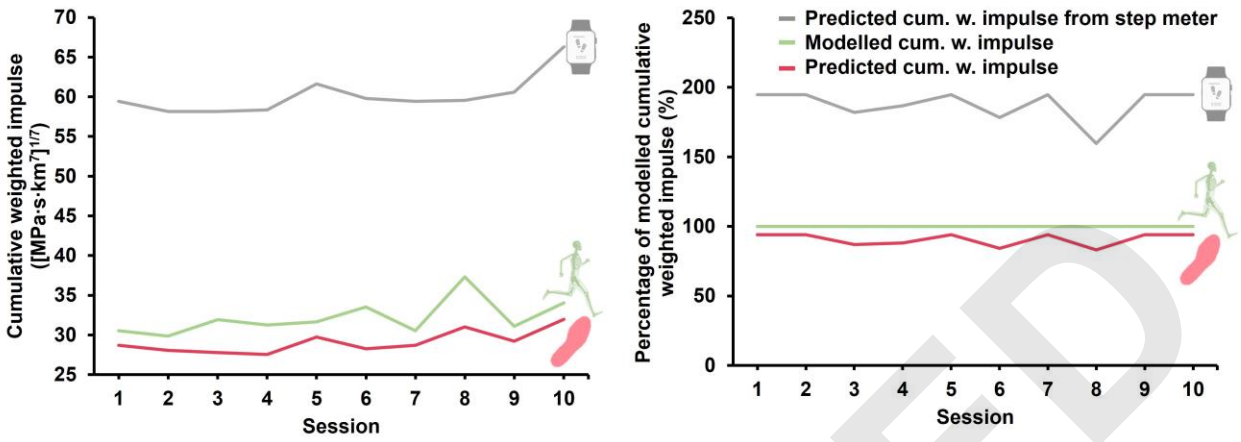

Table 1. Running conditions

		Treadmill slope (degrees)					Lower step frequency	Higher step frequency	Forward trunk lean
		-6	-3	0	3	6			
Speed ($m \cdot s^{-1}$)	2.78	X	X	X	X	X			
	3			X					
	3.33			X			X	X	X
	4			X					
	5			X*					

*all conditions were performed for 1 minute, but some (untrained) participants could not always complete 1 minute and shorter periods were therefore used in these situations.

Table 2. Mean ± SD modelled and predicted stress/strain impulse values per tissue and condition and their relative and absolute percentage differences

Condition	Lab-based impulse	Wearable-based impulse	Relative difference	Relative percentage difference*	Absolute percentage difference
Patellofemoral stress (MPa·s)					
2.78 m·s^{-1}	1.02 ± 0.19	1.01 ± 0.27	-0.01 ± 0.12	-1.56 ± 12.4	10.5 ± 6.39
3.00 m·s^{-1}	0.99 ± 0.20	1.00 ± 0.27	0.01 ± 0.11	0.26 ± 11.5 *	9.30 ± 6.46
3.33 m·s^{-1}	0.99 ± 0.17	1.01 ± 0.26	0.02 ± 0.14	1.19 ± 13.4	11.8 ± 5.77
4.00 m·s^{-1}	0.95 ± 0.20	0.99 ± 0.28	0.04 ± 0.14	3.22 ± 15.1	12.8 ± 8.02
5.00 m·s^{-1}	0.87 ± 0.21	0.95 ± 0.28	0.08 ± 0.14	9.28 ± 17.7	15.2 ± 12.5
6 deg down	1.47 ± 0.25	1.26 ± 0.31	-0.21 ± 0.11	-15.1 ± 7.74	15.9 ± 5.84
3 deg down	1.21 ± 0.24	1.13 ± 0.28	-0.08 ± 0.09	-7.25 ± 7.94 *	8.14 ± 6.98
3 deg up	0.86 ± 0.17	0.93 ± 0.25	0.07 ± 0.10	6.81 ± 12.1	12.7 ± 5.14
6 deg up	0.76 ± 0.19	0.89 ± 0.26	0.13 ± 0.10	17.3 ± 11.6	19.7 ± 6.41
Lower step freq.	1.07 ± 0.20	1.04 ± 0.26	-0.03 ± 0.09	-3.64 ± 9.77 *	8.76 ± 5.30
Higher step freq.	0.89 ± 0.21	0.96 ± 0.27	0.07 ± 0.09	6.86 ± 10.6	10.5 ± 6.79
Trunk lean	0.97 ± 0.19	1.03 ± 0.26	0.06 ± 0.14	5.93 ± 14.8	13.4 ± 7.97
Overall	**1.00 ± 0.19**	**1.02 ± 0.10**	**0.01 ± 0.09**	**1.95 ± 8.40 ***	**12.4 ± 3.35**
Tibial stress (MPa·s)					
2.78 m·s^{-1}	2.82 ± 0.37	2.57 ± 0.43	-0.25 ± 0.15	-9.25 ± 4.95 *	9.64 ± 4.08
3.00 m·s^{-1}	2.78 ± 0.31	2.54 ± 0.41	-0.24 ± 0.21	-8.86 ± 6.78 *	10.2 ± 4.37
3.33 m·s^{-1}	2.71 ± 0.35	2.52 ± 0.41	-0.18 ± 0.19	-6.79 ± 7.06 *	8.36 ± 4.96
4.00 m·s^{-1}	2.55 ± 0.37	2.41 ± 0.35	-0.14 ± 0.19	-5.06 ± 6.91 *	6.70 ± 5.23
5.00 m·s^{-1}	2.31 ± 0.25	2.33 ± 0.32	0.02 ± 0.16	0.64 ± 6.53 **	5.30 ± 3.66
6 deg down	2.21 ± 0.29	2.26 ± 0.33	0.05 ± 0.07	2.23 ± 2.88 ***	2.96 ± 2.07
3 deg down	2.45 ± 0.30	2.36 ± 0.37	-0.10 ± 0.11	-4.21 ± 4.50 ***	5.33 ± 2.98
3 deg up	3.28 ± 0.41	2.78 ± 0.48	-0.50 ± 0.18	-15.5 ± 5.34 *	15.5 ± 5.21
6 deg up	3.71 ± 0.48	2.97 ± 0.52	-0.75 ± 0.16	-20.4 ± 4.63	20.4 ± 4.63
Lower step freq.	2.87 ± 0.34	2.53 ± 0.42	-0.34 ± 0.21	-12.0 ± 7.11 *	12.4 ± 6.41
Higher step freq.	2.55 ± 0.32	2.46 ± 0.40	-0.10 ± 0.20	-3.81 ± 7.38 *	6.33 ± 5.24
Trunk lean	2.70 ± 0.30	2.56 ± 0.40	-0.14 ± 0.20	-5.42 ± 6.62 *	7.03 ± 4.73
Overall	**2.75 ± 0.41**	**2.52 ± 0.19**	**-0.22 ± 0.22**	**-7.37 ± 6.41 ***	**9.18 ± 4.92**
Achilles tendon strain (%·s)					
2.78 m·s^{-1}	0.71 ± 0.13	0.61 ± 0.13	-0.10 ± 0.06	-14.5 ± 7.27	14.5 ± 7.27
3.00 m·s^{-1}	0.71 ± 0.10	0.60 ± 0.12	-0.11 ± 0.06	-15.4 ± 8.35	16.20 ± 6.6
3.33 m·s^{-1}	0.69 ± 0.12	0.59 ± 0.12	-0.09 ± 0.07	-13.3 ± 9.96	14.1 ± 8.75
4.00 m·s^{-1}	0.65 ± 0.13	0.56 ± 0.10	-0.09 ± 0.07	-13.2 ± 9.89	14.10 ± 8.4
5.00 m·s^{-1}	0.58 ± 0.08	0.54 ± 0.09	-0.05 ± 0.07	-8.06 ± 11.0	10.8 ± 8.24
6 deg down	0.40 ± 0.08	0.43 ± 0.08	0.04 ± 0.02	9.77 ± 7.68 *	10.2 ± 7.04
3 deg down	0.54 ± 0.08	0.50 ± 0.10	-0.03 ± 0.05	-6.58 ± 9.49	9.76 ± 5.94
3 deg up	0.92 ± 0.14	0.72 ± 0.15	-0.21 ± 0.07	-22.8 ± 6.81	22.8 ± 6.81
6 deg up	1.13 ± 0.17	0.82 ± 0.17	-0.31 ± 0.06	-27.9 ± 5.41	27.9 ± 5.41
Lower step freq.	0.75 ± 0.11	0.59 ± 0.12	-0.15 ± 0.07	-20.6 ± 9.36	20.6 ± 9.36
Higher step freq.	0.64 ± 0.11	0.58 ± 0.11	-0.07 ± 0.07	-9.94 ± 10.4	11.6 ± 8.33
Trunk lean	0.68 ± 0.09	0.60 ± 0.11	-0.07 ± 0.07	-11.1 ± 9.54	12.5 ± 7.61
Overall	**0.70 ± 0.18**	**0.59 ± 0.10**	**-0.10 ± 0.09**	**-12.8 ± 9.44**	**15.4 ± 5.61**

*** agreement at <10% threshold; ** agreement at <20% threshold; * agreement at <30% threshold. No star indicates no agreement at any of the specified threshold. For example, relative percentage differences with one asterisk could be interpreted as having statistical agreement at a 30% threshold.

The relative percentage difference reflects the relative difference expressed as a percentage of the modelled load.

The absolute percentage difference reflects a measure of the difference whereby the absolute (i.e., positive) value of a negative difference is taken to compute the percentage difference relative to the modelled load. This therefore represents a more conservative estimate of the difference whereby positive and negative differences cannot cancel out.

Supplemental file S1. Additional data

Table S1. Input metrics and a brief description of their interpretation

Predictor	Description
1. Cadence	Number of steps per minute
2. Contact time	Duration of foot-ground contact
3. Swing time	Duration between foot toe-off and same foot ground contact
4. Flight time	Duration between two foot contacts
5. Stability	Variability in the medio-lateral displacement of the center of pressure during the stance phase
6. Inter-step variability	Variation in the anterior-posterior position of the center pressure at initial contact between steps
7. Footstrike index anterior-posterior	Anterior-posterior position of the center pressure in relation to the length of the foot at initial contact
8. Footstrike index medio-lateral	Medio-lateral position of the center pressure in relation to the width of the foot at initial contact
9. Toe-off index anterior-posterior	Anterior-posterior position of the center pressure in relation to the length of the foot at toe-off
10. Toe-off index medio-lateral	Medio-lateral position of the center pressure in relation to the width of the foot at toe-off
11. Running speed*	The speed at which the individual was running
12. Inclination*	The surface gradient of the treadmill

Note that the inter-step variability and cadence metrics require data from multiple steps, but are updated on a per-step basis.

***** These metrics were obtained from the laboratory set-up, but could also be computed using e.g. the step length and step frequency and angular position of the magnetometer in the insole.

Table S2. Intercept, slope, *p*-value for slope, for the change in relative percentage error for the impulse and weighted impulse with changes in speed, uphill or downhill slope, and step frequency

Effect	Patellofemoral joint	Tibia	Achilles tendon
Impulse			
Speed	-14.2 + 4.60x, p =.011	-21.9 + 4.42x, p <.001	-23.6 + 2.97x, p =.011
Uphill slope	-1.98 + 3.21x, p <.001	-9.46 – 1.87x, p <.001	-15.0 – 2.29x, p <.001
Downhill slope	-1.21 + 2.25x, p <.001	-9.49 – 1.91x, p <.001	-15.9 – 4.05x, p <.001
Step frequency	-99.2 + 0.59x, p =.004	-70.7 + 0.38x, p <.001	-110 + 0.57x, p =.002
Weighted impulse*			
Speed	12.0 – 18.6x, p =.002	-1.10 – 14.2x, p <.001	-76.4 – 3.50x, p =.178
Uphill slope	-54.8 + 24.3x, p =.008	-36.1 – 6.74x, p =.001	-86.7 – 2.03x, p =.007
Downhill slope	-40.4 + 7.10x, p =.001	-28.5 + 4.99x, p =.910	-308 – 269x, p <.001
Step frequency	-322 + 1.61x, p =.001	-248 + 1.16x, p =.038	-131 + 0.25x, p =.350

* Note that most weighted impulse outcomes were not normally distributed and the intercept and slope should therefore be interpreted with caution. As transformation or exclusion of outliers did not substantially alter the overall outcome of the model (i.e., direction of the effect and *p*-value), the original model was retained and reported here.

Table S3. Mean ± SD absolute percentage error for each structure when imputing the mean lab-based impulse at each condition for each individual, or when imputing the overall mean impulse

Condition	Patellofemoral joint (%)	Tibia (%)	Achilles tendon (%)
Mean lab-based impulse at each condition imputed for each individual			
2.78 m·s^{-1}	15.3 ± 16.1	9.10 ± 7.81	12.4 ± 10.9
3.00 m·s^{-1}	18.5 ± 13.5	16.1 ± 8.62	16.7 ± 10.8
3.33 m·s^{-1}	14.9 ± 16.6	10.4 ± 10.5	14.6 ± 14.1
4.00 m·s^{-1}	62.7 ± 41.7	12.9 ± 9.82	36.9 ± 10.7
5.00 m·s^{-1}	48.6 ± 39.2	10.5 ± 9.02	12.3 ± 7.74
6 deg down	39.3 ± 11.9	50.6 ± 17.8	141 ± 46.7
3 deg down	21.6 ± 12.2	13.9 ± 8.30	32.1 ± 13.0
3 deg up	20.0 ± 20.2	21.0 ± 9.47	28.1 ± 10.3
6 deg up	22.4 ± 21.0	9.17 ± 7.55	10.2 ± 8.67
Lower step freq.	19.9 ± 11.4	12.0 ± 8.38	15.4 ± 10.5
Higher step freq.	30.7 ± 30.0	15.7 ± 11.6	21.9 ± 16.5
Trunk lean	15.1 ± 16.8	8.13 ± 7.79	9.98 ± 9.30
Overall	**27.4 ± 15.2**	**15.8 ± 11.6**	**29.3 ± 36.2**
Mean overall lab-based impulse imputed for each individual (step counter)			
2.78 m·s^{-1}	15.0 ± 15.6	8.66 ± 7.81	12.2 ± 10.4
3.00 m·s^{-1}	16.6 ± 19.4	8.50 ± 6.67	12.5 ± 7.42
3.33 m·s^{-1}	15.0 ± 17.2	9.90 ± 10.3	14.4 ± 13.6
4.00 m·s^{-1}	18.9 ± 23.7	14.4 ± 8.03	18.8 ± 10.7
5.00 m·s^{-1}	28.8 ± 28.3	20.6 ± 12.4	22.6 ± 16.6
6 deg down	29.9 ± 13.1	27.5 ± 12.2	82.0 ± 35.4
3 deg down	20.9 ± 12.6	15.4 ± 8.40	34.3 ± 13.5
3 deg up	23.8 ± 23.2	15.6 ± 9.12	22.7 ± 11.1
6 deg up	43.1 ± 39.1	25.0 ± 10.3	36.7 ± 11.7
Lower step freq.	15.2 ± 14.0	8.50 ± 8.38	11.0 ± 11.4
Higher step freq.	25.0 ± 26.6	12.7 ± 9.60	17.8 ± 12.1
Trunk lean	16.3 ± 18.3	8.10 ± 8.37	10.8 ± 10.1
Overall	**22.4 ± 8.44**	**14.6 ± 6.70**	**24.7 ± 20.1**

Table S4. Median (interquartile range) modelled and predicted stress/strain weighted impulse values per tissue and condition

Condition	Lab-based weighted impulse	Wearable-based weighted impulse	Relative difference	Relative percentage difference	Absolute percentage difference
Patellofemoral stress ($[MPa \cdot s^{7}]^{1/7}$)					
2.78 $m \cdot s^{-1}$	9.55 (9.30 to 10.5)	8.98 (7.91 to 9.71)	-8.72 (-9.46 to -7.84)	-50.9 (-75.8 to -28.3)	66.7 (32.0 to 77.6)
3.00 $m \cdot s^{-1}$	10.0 (9.30 to 10.2)	8.78 (8.02 to 9.56)	-8.98 (-9.51 to -7.99)	-56.1 (-75.1 to -33.9)	65.1 (43.1 to 78.5)
3.33 $m \cdot s^{-1}$	10.0 (9.70 to 10.8)	8.94 (8.07 to 9.44)	-9.39 (-10.4 to -8.03)	-67.1 (-78.8 to -43.2)	68.7 (51.2 to 78.8)
4.00 $m \cdot s^{-1}$	10.7 (10.1 to 11.6)	8.79 (7.85 to 9.22)	-10.3 (-11.5 to -9.49)	-78.6 (-88.9 to -63.6)	78.6 (63.6 to 88.9)
5.00 $m \cdot s^{-1}$	11.1 (9.76 to 12.2)	8.52 (7.77 to 8.93)	-10.8 (-12.0 to -8.99)	-82.9 (-94.6 to -67.2)	82.9 (67.2 to 94.6)
6 deg down	13.7 (13.0 to 14.7)	11.0 (9.28 to 11.7)	-13.0 (-14.2 to -12.8)	-81.2 (-88.9 to -74.3)	81.2 (74.3 to 88.9)
3 deg down	11.4 (10.7 to 11.9)	9.75 (8.18 to 10.4)	-10.5 (-11.4 to -10.1)	-77.7 (-87.7 to -51.4)	77.7 (51.4 to 87.7)
3 deg up	8.34 (7.80 to 9.00)	8.05 (7.06 to 8.21)	-6.94 (-8.09 to -0.35)	-36.0 (-55.6 to 3.68)	49.1 (32.3 to 64.4)
6 deg up	7.18 (6.21 to 8.06)	7.65 (6.49 to 7.94)	4.48 (-6.86 to 6.05)	10.7 (-27.7 to 85.7)	36.1 (19.6 to 95.5)
Lower step freq.	10.9 (9.90 to 11.5)	9.11 (7.98 to 9.58)	-10.33 (-11.0 to -8.92)	-77.5 (-85.5 to -52.6)	79.5 (62.2 to 85.5)
Higher step freq.	9.34 (8.62 to 10.3)	8.73 (7.83 to 9.34)	-8.67 (-9.82 to -6.59)	-49.7 (-68.7 to -22.7)	52.3 (37.5 to 69.6)
Trunk lean	9.63 (9.05 to 10.2)	8.41 (7.71 to 9.44)	-8.69 (-9.92 to -5.93)	-55.3 (-79.1 to -5.43)	64.0 (15.2 to 80.7)
Overall	**10.0 (9.50 to 11.0)**	**8.79 (8.50 to 9.01)**	**-9.19 (-10.4 to -8.69)**	**-61.6 (-77.9 to -50.6)**	**67.7 (61.1 to 78.8)**
Tibial stress ($[MPa \cdot s^{7}]^{1/7}$)					
2.78 $m \cdot s^{-1}$	96.0 (90.2 to 99.6)	88.7 (84.3 to 94.0)	-85.5 (-89.7 to -79.3)	-46.5 (-53.1 to -32.5)	46.9 (35.2 to 56.0)
3.00 $m \cdot s^{-1}$	96.5 (92.0 to 102)	87.7 (81.5 to 92.9)	-88.6 (-91.4 to -80)	-53.1 (-62.9 to -45.2)	53.1 (45.2 to 62.9)
3.33 $m \cdot s^{-1}$	101 (91.7 to 103)	87.6 (81.3 to 92.2)	-90 (-97.7 to -82.7)	-59.9 (-65.6 to -44.6)	59.9 (44.6 to 65.6)
4.00 $m \cdot s^{-1}$	98.5 (96.4 to 108)	86.8 (81.6 to 91.0)	-92.3 (-99.6 to -90.4)	-64.5 (-73.1 to -58.2)	64.5 (58.2 to 73.1)
5.00 $m \cdot s^{-1}$	106 (97.5 to 113)	87.5 (81.1 to 90.8)	-101 (-110 to -92.9)	-72.2 (-79.1 to -62.6)	72.2 (62.6 to 79.1)
6 deg down	102 (94.4 to 110)	86.8 (81.0 to 93.9)	-98 (-105 to -87.4)	-70.5 (-79.6 to -49.7)	70.5 (49.7 to 79.6)
3 deg down	92.2 (87.7 to 99.3)	87.0 (80.8 to 93.2)	-78.9 (-82.6 to -74.5)	-36.9 (-42.5 to -29.0)	36.9 (29.0 to 42.5)
3 deg up	103 (96.4 to 110)	89.8 (82.3 to 92.2)	-99.1 (-105 to -88.9)	-68.3 (-73.8 to -57.7)	68.3 (57.7 to 73.8)
6 deg up	112 (103 to 115)	90.4 (83.1 to 94.3)	-108 (-111 to -98.7)	-78.1 (-81.3 to -71.0)	78.1 (71.0 to 81.3)
Lower step freq.	100 (95.6 to 109)	87.7 (81.3 to 91.1)	-93.4 (-105 to -89.3)	-66.4 (-75.4 to -53.6)	66.4 (53.6 to 75.4)
Higher step freq.	94.9 (90.3 to 99.5)	87.7 (82.8 to 91.0)	-85.7 (-91.7 to -77.8)	-49.3 (-56.8 to -34.2)	49.3 (35.8 to 56.8)
Trunk lean	96.0 (90.7 to 104)	87.3 (79.9 to 91.5)	-84.9 (-90.3 to -71.1)	-36.7 (-59.9 to -20.8)	36.7 (27.9 to 59.9)
Overall	**99.2 (96.0 to 102)**	**87.6 (87.2 to 88)**	**-91.2 (-98.3 to -85.7)**	**-62.2 (-68.8 to -48.6)**	**62.2 (48.7 to 68.8)**
Achilles tendon strain ($[\%\cdot s^{9.3}]^{1/9.3}$)					
2.78 $m \cdot s^{-1}$	5.67 (5.13 to 6.21)	4.29 (3.91 to 4.81)	-5.52 (-6.13 to -4.81)	-94.2 (-96.7 to -82.3)	94.2 (82.3 to 96.7)
3.00 $m \cdot s^{-1}$	5.76 (5.16 to 6.66)	4.29 (3.92 to 4.82)	-5.75 (-6.56 to -5.10)	-95.9 (-97.9 to -86.1)	95.9 (86.1 to 97.9)
3.33 $m \cdot s^{-1}$	5.77 (5.29 to 6.82)	4.29 (3.87 to 4.78)	-5.77 (-6.79 to -5.23)	-95.5 (-97.7 to -85.3)	95.5 (85.3 to 97.7)
4.00 $m \cdot s^{-1}$	5.95 (5.22 to 7.15)	4.32 (3.93 to 4.64)	-5.90 (-7.14 to -5.10)	-96.5 (-98.6 to -92.8)	96.5 (92.8 to 98.6)
5.00 $m \cdot s^{-1}$	6.75 (5.44 to 7.32)	4.33 (3.96 to 4.68)	-6.70 (-7.30 to -5.42)	-97.8 (-99.1 to -92.1)	97.8 (92.1 to 99.1)
6 deg down	2.96 (2.67 to 3.58)	3.81 (3.11 to 4.33)	3.68 (2.98 to 4.17)	907 (72.1 to 2064)	907 (124 to 2064)
3 deg down	4.12 (3.81 to 4.69)	4.13 (3.75 to 4.38)	-3.51 (-4.10 to 3.88)	-23.0 (-81.0 to 131)	82.9 (70.7 to 131)
3 deg up	7.20 (6.80 to 8.12)	4.82 (4.24 to 5.46)	-7.20 (-8.09 to -6.73)	-98.5 (-99.1 to -94.4)	98.5 (94.4 to 99.1)
6 deg up	8.75 (8.13 to 9.78)	5.87 (4.66 to 6.39)	-8.75 (-9.75 to -8.12)	-98.3 (-98.8 to -96.1)	98.3 (96.1 to 98.8)
Lower step freq.	6.28 (5.86 to 7.29)	4.30 (3.77 to 4.73)	-6.19 (-7.28 to -5.82)	-97.8 (-99.0 to -89.3)	97.8 (89.3 to 99.0)
Higher step freq.	5.35 (4.91 to 6.86)	4.26 (4.02 to 4.78)	-5.35 (-6.83 to -4.82)	-92.8 (-97.0 to -79.0)	92.8 (79.0 to 97.0)
Trunk lean	5.69 (5.03 to 6.27)	4.33 (3.83 to 4.76)	-5.62 (-6.09 to -4.95)	-88.9 (-95.2 to -78.0)	88.9 (78.0 to 95.2)
Overall	**5.77 (5.59 to 6.40)**	**4.30 (4.28 to 4.33)**	**-5.76 (-6.32 to -5.48)**	**-95.7 (-97.8 to -91.8)**	**96.2 (93.8 to 97.9)**

Note that the percentage errors are computed using the original units (i.e., not raised to the power of 1/b).

Table S5. Mean (interquartile range) absolute percentage error for each structure when imputing the mean lab-based weighted impulse at each condition for each individual, or when imputing the overall mean weighted impulse

Condition	Patellofemoral joint (%)	Tibia (%)	Achilles tendon (%)
Mean lab-based weighted impulse at each condition imputed for each individual			
2.78 $m \cdot s^{-1}$	76.7 (48.8 to 108.5)	57.8 (37.8 to 129)	213 (81.1 to 697)
3.00 $m \cdot s^{-1}$	273 (225.4 to 504.2)	137 (65.5 to 236)	468 (53.8 to 1481)
3.33 $m \cdot s^{-1}$	47.0 (30.2 to 71.9)	53.3 (11.4 to 103.5)	164 (65.1 to 498)
4.00 $m \cdot s^{-1}$	643 (337 to 1042)	157 (61.7 to 198)	99.2 (97.4 to 99.9)
5.00 $m \cdot s^{-1}$	91.4 (53 to 374.2)	56.4 (37.6 to 76.1)	96.2 (71.5 to 98.2)
6 deg down	95.8 (93.9 to 97.4)	65.9 (42.4 to 181)	1063497 (184100 to 2876723)
3 deg down	46.1 (29.0 to 78.7)	122 (47.0 to 216)	7953 (2306 to 16553)
3 deg up	804 (432 to 1339)	55.9 (21.9 to 127)	72.1 (40.6 to 109)
6 deg up	64.3 (36.6 to 242)	62.0 (38.8 to 160)	188 (27.6 to 420)
Lower step freq.	46.3 (30.8 to 76.9)	43.4 (18.2 to 78.6)	84.9 (58.8 to 178)
Higher step freq.	324 (114 to 646)	156 (83.1 to 261)	1203 (55.8 to 2833)
Trunk lean	80.2 (44.8 to 166)	38.8 (16.9 to 63.6)	84.1 (32.2 to 555)
Overall	**85.8 (60.0 to 286)**	**59.9 (55.2 to 126)**	**176 (93.4 to 652)**
Mean overall lab-based weighted impulse imputed for each individual (step counter)			
2.78 $m \cdot s^{-1}$	268 (94.6 to 342)	112 (78.9 to 235)	1883 (771 to 4947)
3.00 $m \cdot s^{-1}$	175 (139 to 345)	105 (49 to 191)	1605 (348 to 4645)
3.33 $m \cdot s^{-1}$	161 (54.6 to 231)	70.4 (40.1 to 196)	1587 (265 to 3721)
4.00 $m \cdot s^{-1}$	78.9 (36.4 to 156)	81.6 (54.5 to 106)	1172 (131 to 4178)
5.00 $m \cdot s^{-1}$	83.7 (37.4 to 216)	46.8 (16.9 to 92.2)	294 (85.2 to 2823)
6 deg down	74.2 (58.3 to 83.7)	56.1 (37 to 142)	824004 (142623 to 2228938)
3 deg down	40.4 (18.3 to 81.3)	183 (67.7 to 302)	38923 (11560 to 80596)
3 deg up	850 (459 to 1413)	59.1 (24.5 to 114)	114 (56.3 to 269)
6 deg up	2257 (932 to 5737)	36.7 (25.6 to 63.8)	73.2 (52.3 to 91.0)
Lower step freq.	58.3 (27.4 to 166)	62.5 (15.9 to 91.3)	483 (55.2 to 1226)
Higher step freq.	330 (117 to 657)	131 (65.4 to 226)	3283 (236 to 7518)
Trunk lean	182 (62.0 to 349)	60.4 (12.7 to 157)	1234 (344 to 4648)
Overall	**168 (77.7 to 283)**	**66.4 (58.3 to 107)**	**1410 (435 to 2233)**

Note that the percentage errors are computed using the original units (i.e., not raised to the power of 1/b).

Table S6. Percentage change in loading relative to loading at 2.78 m·s^{-1} captured by the machine learning model

Condition	Patellofemoral joint (%)	Tibia (%)	Achilles tendon (%)
Impulse			
2.78 m·s^{-1}	-	-	-
3.00 m·s^{-1}	37.9	67.0	189
3.33 m·s^{-1}	-5.95	39.0	62.7
4.00 m·s^{-1}	33.4	56.6	81.3
5.00 m·s^{-1}	40.9	46.9	58.0
6 deg down	55.1	50.3	55.8
3 deg down	60.5	57.2	61.5
3 deg up	50.3	46.2	50.4
6 deg up	46.8	44.5	49.7
Lower step freq.	54.9	-97.5	-43.9
Higher step freq.	42.6	41.0	48.1
Trunk lean	-26.8	5.06	18.9
Overall	**35.4 ± 27.2**	**32.4 ± 45.8**	**57.4 ± 54.6**
Weighted impulse			
2.78 m·s^{-1}	-	-	-
3.00 m·s^{-1}	238	-2339	2083
3.33 m·s^{-1}	1173	-104	-114
4.00 m·s^{-1}	543	-38.6	-30.7
5.00 m·s^{-1}	387	-33.0	-48.2
6 deg down	-209	-18.8	26.1
3 deg down	-298	54.9	17.2
3 deg up	-180	-14.3	40.8
6 deg up	-307	7.09	51.2
Lower step freq.	-912	-24.5	-34.6
Higher step freq.	54.2	169	-120
Trunk lean	535	98.8	110
Overall	**93.1 ± 562**	**-204 ± 712**	**180 ± 635**

Values <100 indicate that the machine learning model underestimated the change in load, values >100 that the model overestimated the change in load, and negative values reflect a directionally different predicted change in load (e.g., predicted lower load relative to 2.78 m·s^{-1}, while the actual load increased).

Table S7. Mean ± SD number of steps used for training per condition per subject

Condition	Number of steps
2.78 $m \cdot s^{-1}$	135 ± 5.69
3.00 $m \cdot s^{-1}$	137 ± 5.87
3.33 $m \cdot s^{-1}$	139 ± 6.23
4.00 $m \cdot s^{-1}$	137 ± 25.8
5.00 $m \cdot s^{-1}$	128 ± 47.3
6 deg down	132 ± 7.74
3 deg down	134 ± 6.84
3 deg up	134 ± 9.86
6 deg up	132 ± 19.6
Lower step freq.	123 ± 18.0
Higher step freq.	142 ± 17.9
Trunk lean	134 ± 9.63
Overall	**134 ± 15.0**

Table S8. Mean ± SD relative impulse prediction error (%) as a function of the number of steps per subject per condition used for training of the neural network

Nr. of steps	Patellofemoral joint (%)	Tibia (%)	Achilles tendon (%)
1	-0.57 ± 13.2	-7.04 ± 8.74	-7.73 ± 16.9
5	1.54 ± 9.64	-5.36 ± 6.34	-6.00 ± 11.2
10	2.05 ± 8.93	-7.32 ± 7.50	-12.1 ± 11.9
20	5.87 ± 9.35	-6.62 ± 7.48	-11.8 ± 11.9
30	2.65 ± 9.28	-6.86 ± 7.11	-10.7 ± 11.8
40	3.04 ± 9.36	-7.07 ± 7.35	-11.4 ± 12.0
50	1.51 ± 9.79	-7.22 ± 7.62	-11.8 ± 12.4
60	2.33 ± 9.39	-7.11 ± 7.48	-11.9 ± 12.2
All (avg. 134)	**1.95 ± 8.40**	**-7.37 ± 6.41**	**-12.8 ± 9.44**

Table S9. Overall mean ± SD relative percentage error for each structure after removal of each predictor

Predictor variable removed	Patellofemoral joint (%)	Tibia (%)	Achilles tendon (%)
Cadence	10.3 ± 9.72	10.3 ± 9.46	11.5 ± 9.89
Contact time	10.4 ± 9.83	10.3 ± 9.64	11.5 ± 9.88
Swing time	10.3 ± 9.63	10.3 ± 9.57	11.5 ± 9.99
Flight time	10.3 ± 9.62	10.3 ± 9.48	11.5 ± 9.91
Footstrike index x	10.4 ± 9.98	10.4 ± 9.82	11.5 ± 9.93
Footstrike index y	10.4 ± 9.87	10.3 ± 9.60	11.5 ± 9.92
Stability	10.3 ± 9.70	10.3 ± 9.60	11.5 ± 9.94
Variability	10.3 ± 9.75	10.3 ± 9.56	11.5 ± 9.87
Toe-off index x	10.3 ± 9.61	10.3 ± 9.58	11.5 ± 9.96
Toe-off index y	10.3 ± 9.71	10.3 ± 9.48	11.5 ± 9.95
Slope	10.3 ± 9.61	10.2 ± 9.39	11.5 ± 9.98
Speed	10.4 ± 9.83	10.3 ± 9.72	11.5 ± 10.0

Footstrike and toe-off x represent the medial-lateral direction of the center of pressure relative to the whole foot, while y represents the anterior-posterior direction.

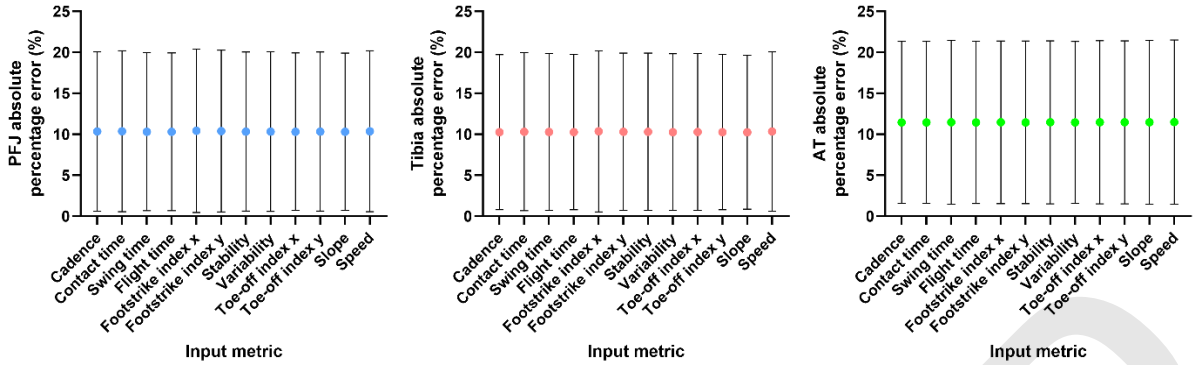

Figure S1. Mean and standard deviation absolute percentage error for each structure after removal of each predictor. Note that the error remains relatively similar across all metrics.

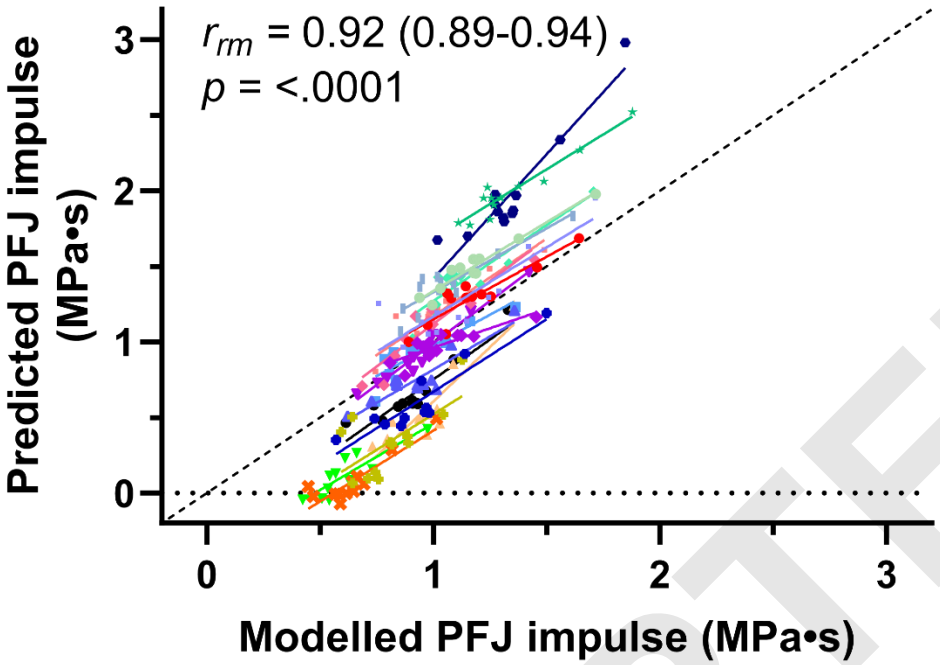

Figure S2. Repeated-measures correlations for the patellofemoral joint impulse across all participants after the application of a generalized correction factor that reduced the magnitude of relatively lower values, and increased the magnitude of relatively higher values in an attempt to better match the modelled load. While the relative percentage error decreased from 1.95% to 0.36%, the absolute percentage error increased from 12.4% to 29.6%.

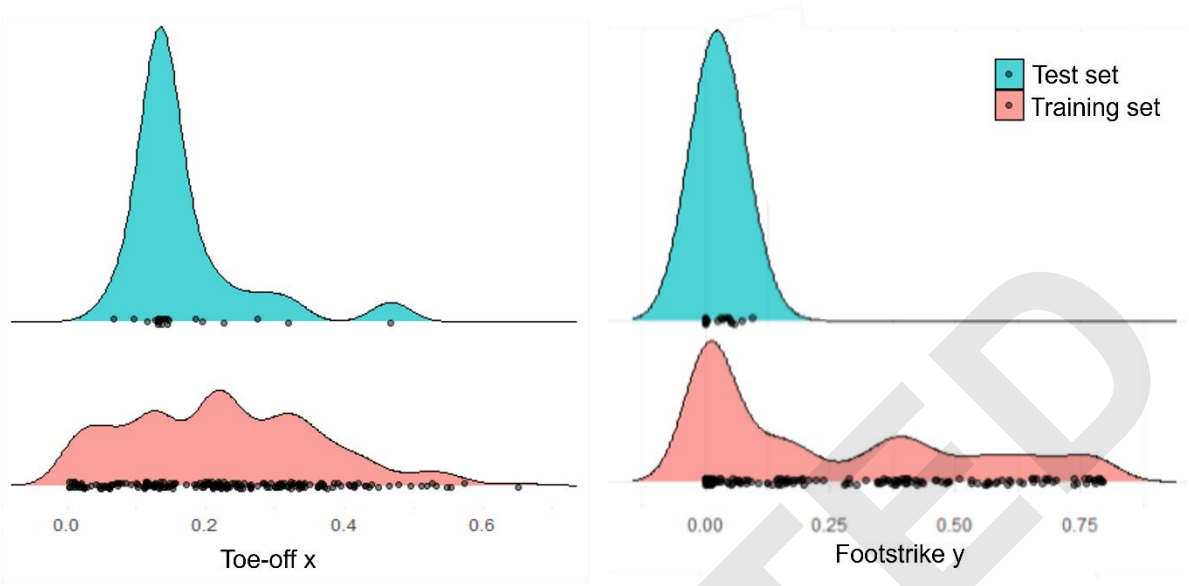

Figure S3. Example of mean spatiotemporal metrics in training data across all conditions (red) and the mean spatiotemporal metrics (only shown for toe-off x and footstrike y as examples) for each condition for the two individuals where the neural network predicted a constant impulse across all conditions (blue).

Table S10. Mean ± SD modelled and predicted stress/strain impulse values per tissue and condition when using three input metrics (contact time, step frequency and footstrike angle)

Condition	Lab-based impulse	Wearable-based impulse	Difference	Percentage difference*	Absolute percentage difference
Patellofemoral stress (MPa·s)					
2.78 m·s^{-1}	1.02 ± 0.19	1.04 ± 0.24	0.03 ± 0.26	5.17 ± 27.6	21.5 ± 17.4
3.00 m·s^{-1}	0.99 ± 0.20	1.00 ± 0.25	0.01 ± 0.28	4.33 ± 31.7	24.5 ± 19.7
3.33 m·s^{-1}	0.99 ± 0.17	1.06 ± 0.25	0.07 ± 0.28	9.88 ± 31.1	25.4 ± 19.7
4.00 m·s^{-1}	0.95 ± 0.20	0.95 ± 0.25	0.00 ± 0.28	4.04 ± 33.1	26.0 ± 12.0
5.00 m·s^{-1}	0.87 ± 0.21	1.05 ± 0.25	0.19 ± 0.26	26.8 ± 38.2	34.8 ± 30.6
6 deg down	1.47 ± 0.25	1.07 ± 0.37	-0.45 ± 0.39	-25.7 ± 22.0	28.9 ± 17.9
3 deg down	1.21 ± 0.24	1.06 ± 0.26	-0.14 ± 0.30	-9.04 ± 26.6	20.4 ± 18.9
3 deg up	0.86 ± 0.17	1.04 ± 0.24	0.18 ± 0.26	24.0 ± 33.8	34.1 ± 22.9
6 deg up	0.76 ± 0.19	1.03 ± 0.32	0.28 ± 0.25	43.4 ± 42.1	48.8 ± 36.1
Lower step freq.	1.07 ± 0.20	1.06 ± 0.49	-0.19 ± 0.44	-0.42 ± 22.7	18.8 ± 16.8
Higher step freq.	0.89 ± 0.21	1.06 ± 0.34	0.12 ± 0.35	24.8 ± 42.0	36.0 ± 33.5
Trunk lean	0.97 ± 0.19	1.05 ± 0.41	0.09 ± 0.28	13.7 ± 38.6	28.8 ± 30.8
Overall	**1.00 ± 0.19**	**1.04 ± 0.03**	**0.01 ± 0.20**	**10.1 ± 18.3**	**29.0 ± 8.40**
Tibial stress (MPa·s)					
2.78 m·s^{-1}	2.82 ± 0.37	2.57 ± 0.45	-0.26 ± 0.19	-9.26 ± 6.04	10.0 ± 4.59
3.00 m·s^{-1}	2.78 ± 0.31	2.53 ± 0.41	-0.25 ± 0.22	-9.03 ± 7.24	10.4 ± 5.01
3.33 m·s^{-1}	2.71 ± 0.35	2.50 ± 0.40	-0.21 ± 0.20	-7.65 ± 6.82	8.96 ± 4.86
4.00 m·s^{-1}	2.55 ± 0.37	2.42 ± 0.36	-0.13 ± 0.15	-4.97 ± 5.96	6.43 ± 4.22
5.00 m·s^{-1}	2.31 ± 0.25	2.37 ± 0.33	0.06 ± 0.16	2.38 ± 6.25	5.25 ± 3.99
6 deg down	2.21 ± 0.29	2.54 ± 0.42	0.33 ± 0.17	14.8 ± 6.09	14.8 ± 6.09
3 deg down	2.45 ± 0.30	2.56 ± 0.43	0.11 ± 0.18	3.94 ± 6.41	5.60 ± 4.95
3 deg up	3.28 ± 0.41	2.62 ± 0.54	-0.66 ± 0.26	-20.4 ± 6.98	21.0 ± 5.09
6 deg up	3.71 ± 0.48	2.67 ± 0.84	-1.04 ± 0.33	-28.5 ± 8.99	28.5 ± 8.99
Lower step freq.	2.87 ± 0.34	2.49 ± 0.69	-0.38 ± 0.21	-13.4 ± 6.92	13.6 ± 6.57
Higher step freq.	2.55 ± 0.32	2.48 ± 0.40	-0.08 ± 0.19	-3.15 ± 6.65	5.63 ± 4.61
Trunk lean	2.70 ± 0.30	2.54 ± 0.90	-0.16 ± 0.17	-6.20 ± 5.73	7.58 ± 4.11
Overall	**2.75 ± 0.41**	**2.52 ± 0.08**	**-0.22 ± 0.36**	**-6.79 ± 11.24**	**11.5 ± 7.08**
Achilles tendon strain (%·s)					
2.78 m·s^{-1}	0.71 ± 0.13	0.61 ± 0.14	-0.10 ± 0.07	-14.8 ± 8.20	15.4 ± 6.89
3.00 m·s^{-1}	0.71 ± 0.10	0.60 ± 0.12	-0.11 ± 0.07	-15.9 ± 9.65	17.0 ± 7.48
3.33 m·s^{-1}	0.69 ± 0.12	0.58 ± 0.12	-0.10 ± 0.07	-14.6 ± 10.3	15.3 ± 9.22
4.00 m·s^{-1}	0.65 ± 0.13	0.56 ± 0.11	-0.09 ± 0.06	-13.1 ± 9.04	14.2 ± 7.12
5.00 m·s^{-1}	0.58 ± 0.08	0.55 ± 0.10	-0.03 ± 0.07	-5.16 ± 11.0	9.21 ± 7.65
6 deg down	0.40 ± 0.08	0.60 ± 0.12	0.20 ± 0.07	50.8 ± 15.1	50.8 ± 15.1
3 deg down	0.54 ± 0.08	0.60 ± 0.13	0.07 ± 0.07	12.4 ± 13.6	15.3 ± 10.1
3 deg up	0.92 ± 0.14	0.63 ± 0.18	-0.29 ± 0.10	-32.3 ± 9.70	32.4 ± 9.26
6 deg up	1.13 ± 0.17	0.66 ± 0.25	-0.46 ± 0.13	-42.0 ± 12.4	42.0 ± 12.4
Lower step freq.	0.75 ± 0.11	0.58 ± 0.18	-0.17 ± 0.08	-22.7 ± 10.6	22.7 ± 10.6
Higher step freq.	0.64 ± 0.11	0.58 ± 0.12	-0.06 ± 0.07	-9.60 ± 10.2	11.9 ± 7.23
Trunk lean	0.68 ± 0.09	0.60 ± 0.22	-0.08 ± 0.06	-12.5 ± 9.20	13.4 ± 7.96
Overall	**0.70 ± 0.18**	**0.60 ± 0.03**	**-0.10 ± 0.16**	**-9.95 ± 23.34**	**21.6 ± 13.1**

Further details on the data presented in Figure 6

We assumed a typical running distance of 7 km per session for a recreational runner, and an average running speed of 2.78 $m \cdot s^{-1}$ for the average session based on on-field data collected in a large randomized-controlled trial (1).

Table S11. Information for the data presented in Figure 6

Session no.	Session type	Speed ($m \cdot s^{-1}$)	Total distance (km)	Modelled weighted impulse per step ($[MPa \cdot s^7]^{1/7}$)	Predicted weighted impulse per step ($[MPa \cdot s^7]^{1/7}$)	Average step length at speed (m)‡	no. of strides to complete 1 km	Modelled cumulative weighted impulse ($[MPa \cdot s \cdot km^7]^{1/7}$)	ML predicted cumulative weighted impulse ($[MPa \cdot s \cdot km^7]^{1/7}$)	SM predicted cumulative weighted impulse ($[MPa \cdot s \cdot km^7]^{1/7}$)
1	Steady-speed run	2.78	7	9.55	8.98	1.03	485	31	29	59
2	Steady-speed run	2.78	6	9.55	8.98	1.03	485	30	28	58
3	Interval session*		7					32	28	58
4	Steady-speed run	3	6.5	10	8.78	1.09	459	31	28	58
5	Steady-speed run	2.78	9	9.55	8.98	1.03	485	32	30	62
6	Race*		9					34	28	60
7	Steady-speed run	2.78	7	9.55	8.98	1.03	485	31	29	59
8	Uphil sprints at 6 deg*		7					37	31	60
9	Steady-speed run	2.78	8	9.55	8.98	1.03	485	31	29	61
10	Long-run	2.78	15	9.55	8.98	1.03	485	34	32	66

* Interval session 1 consisted of 4 km at 4 $m \cdot s^{-1}$ and 3 km warm-up/active recovery between intervals at 2.78 $m \cdot s^{-1}$

* Race consisted of 5 km at 5 $m \cdot s^{-1}$ and 4 km warm-up and cool-down at 2.78 $m \cdot s^{-1}$

* Uphill sprints consists of 2 km uphill running at 6 deg, 2 km downhill running at 6 deg and 3 km level running at 2.78 $m \cdot s^{-1}$

‡ The average step length at each condition was taken from Van Hooren et al (2).

ML = machine learning; SM = step meter

References

1. Van Hooren B, Plasqui G, Meijer K. The Effect of Wearable-Based Real-Time Feedback on Running Injuries and Running Performance: A Randomized Controlled Trial. Am J Sports Med. 2024;Epub ahead of print:3635465231222464. Epub 20240129. doi: 10.1177/03635465231222464. PubMed PMID: 38287728.

2. Van Hooren B, Van Rengs L, Meijer K. Per-step and cumulative load at three common running injury locations: the effect of speed, surface gradient and cadence. Scand J Med Sci Sports. 2024;34(2):e14570.