Evaluation and Design of Non-Invasive, Wearable Musculoskeletal Monitoring Tools for Research, Occupational and Sport Applications

By

Emily S. Matijevich

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University in partial

fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Mechanical Engineering

January 31, 2021

Nashville, Tennessee

Approved:

Karl Zelik, Ph.D.

Michael Goldfarb, Ph.D.

Leon Scott, M.D.

Kevin Galloway, Ph.D.

Allison Gruber, Ph.D.

Acknowledgements

I would like to thank my advisor, Karl Zelik, for challenging me to think broadly and deeply about how I approach research questions, communicate scientific findings, and envision future innovations. Thank you for creating a research environment that was scientifically rigorous, inclusive and collaborative.

I also thank my committee (Michael Goldfarb, Leon Scott, Kevin Galloway and Allison Gruber) for their valuable feedback. Additional thanks to Peter Volgyesi and Gerasimos Bastas for their mentorship.

A huge thank you to the support and friendship from the CREATe family: Matt Yandell, Eric Honert, Erik Lamers, Maura Eveld, Shane King, Rachel Teater, David Ziemnicki, Anna Wolfe, Kirsty McDonald, Laura Judson, Cameron Nurse, Andres Martinez, Ben Gasser, and all of the BAT lab undergraduate and Masters students. It has been a pleasure to conduct experiments together, engage in research discussions, travel to conferences, and support each other in our research endeavors.

An additional thank you the numerous Vanderbilt graduate students who were always there for an afternoon coffee break or evening workout session to take a break from the lab: Katy Riojas, Margaret Rox, Colette Abah, Janna Eaves, Kirsty Walsh, Bryn Pitt, Kelsay Neely, Steve Culver, Jantzen Lee, Joe Howard, Darren Tinker, Chris Nash, Dom Ropella, Garrett Marshall, Matt Fitzgerald.

My journey to graduate school was made possible by many amazing researchers and educators who inspired my interest in math, science, research and biomechanics: Liz Hsiao-Wecksler, Jennifer Amos, Brad Sutton, Angela Wolters, Eric Barth, Ms. Gilmore, Mr. Mshar, Ms. Harbin, Ms. Szczesniak, Mr. Witt and Mr. Adamovic.

A special thanks to my forever supportive parents, Sue and Terry, who have advocated for my success every step of the way. Additional thanks to my brother Tyler, my sister Sara, and my grandmother Pat.

Table of Contents

Table of Contents

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Monitoring musculoskeletal dynamics: the scientific, societal, and clinical opportunities

Multi-scale biomechanics involves connecting our understanding of whole-body movement to underlying musculoskeletal dynamics inside the body. The ability to estimate musculoskeletal dynamics (e.g., forces, stress and motion of individual muscles, bones and tendons) provides numerous scientific, societal, and clinical opportunities. For instance, a musculoskeletal-level understanding of the mechanisms underlying human movement helps scientists gain insight on how and why our bodies move the way they do, connecting the form of different musculoskeletal structures to the function they serve as we walk, run, go up and down stairs, lift objects, etc. [1]–[3]. These scientific insights can then be translated into injury prevention and performance enhancement tools and devices, improving the health and mobility of society and pushing the capabilities of human movement [4]–[8]. Furthermore, the ability to monitor musculoskeletal dynamics offers many opportunities for gaining clinical insights on recovery and treatment strategies, helping individuals return to work, play, and life more efficiently [9]. There is significant value to being able to measure and monitor human movement at the musculoskeletal level.

Traditionally, musculoskeletal dynamics are measured in two ways: 1) using implantable sensors [10]–[12] or 2) using motion analysis laboratory-based measurement systems (e.g., motion capture, force plates, and electromyography (EMG)) in combination with physics-based musculoskeletal modeling [13]–[15]. While implantable sensors offer value as a direct measurement on the musculoskeletal structure, they are invasive and typically impractical for wide scale or long term monitoring. While the motion analysis and musculoskeletal modeling approach is non-invasive, it is typically constrained to a lab environment and this indirect estimate involves a number of modeling limitations and assumptions. While each approach (implantable sensors and musculoskeletal modeling) has benefits and drawbacks, neither is particularly well suited for longitudinal (long-term) testing, for widespread daily monitoring or for applied situations outside the lab.

Over the past few decades, additional measurement modalities have enhanced capabilities for monitoring human movement and musculoskeletal dynamics. For example, imaging modalities allow for personalized anatomical and physiological parameters to be incorporated into musculoskeletal models or for the dynamics of musculoskeletal structures to be estimated using image processing [16]–[18]. Additionally, non-invasive wearable sensors allow for human movement to be estimated unconstrained and outside the lab. Collectively, these non-invasive imaging and wearable measurement modalities open new opportunities for understanding musculoskeletal dynamics, providing tools for long-term and/or widespread monitoring in applied situations [19]–[21].

However, the excitement and desire to apply these new imaging and wearable measurement modalities, in both research and commercial settings, has in some cases outpaced the careful validation work that is needed to ensure they are being used and interpreted properly within the context of monitoring musculoskeletal dynamics [22]–[24]. There is risk that measurements from imaging modalities and wearable sensors may be misapplied or may lead to misinterpretation of what dynamics are being estimated, which can hamper our ability to infer bone, muscle, and tendon function. There are key knowledge gaps regarding the validation, application, and interpretation of measurements from these non-invasive imaging and wearable measurement modalities. Further, there are opportunities to use imaging and wearables in novel ways to expand on the capabilities of

these measurement modalities for monitoring musculoskeletal dynamics.

In this dissertation, I first elucidate the limitations of two musculoskeletal monitoring approaches (using ultrasound to monitor muscle-tendon dynamics and using ground reaction force metrics to infer bone force and associated overuse injury risk). Second, I develop new approaches from combining wearables, musculoskeletal biomechanics and machine learning to more accurately estimate loading on structures inside the body. By both identifying limitations of current musculoskeletal monitoring tools and inventing new tools to overcome these limitations, I have expanded our capabilities for monitoring internal loading in applied situations. These new capabilities provide research opportunities to gain a deeper understanding of internal loading during movement tasks, with applications in improved workplace productivity, enhanced sport performance, and overuse injury prevention.

1.2 Dissertation Contributions

1.2.1 Chapter 2 Contributions

Chapter 2 explores using ultrasound imaging to non-invasively estimate tendon dynamics *in vivo* during human movement. Biomechanics researchers typically use ultrasound imaging methods to track a feature of a muscle tendon unit (MTU), map this experimental data onto a conceptual model of the MTU, and finally use tendon dynamic estimates to draw conclusions about the function of tendon. However, ultrasound-based tendon estimates remain challenging to validate, and it was previously unknown which ultrasound-based estimation methods are most accurate, and under what MTU force or displacement conditions do these methods yield reliable estimates of tendon dynamics.

The primary contribution of my work in this chapter is the discovery that two commonly-used ultrasound based methods for estimating tendon dynamics yielded unrealistic shortening of tendon when the ankle plantarflexes beyond neutral. Specifically, I used B-mode ultrasound to image the gastrocnemius muscle fascicles and the gastrocnemius Achilles tendon muscle-tendon junction of eight healthy participants, while synchronously collecting motion capture and ground reaction forces, during three tasks involving different high/low MTU force and high/low MTU displacement. Next, I compare experimental muscle and tendon length changes to our expectations, using a common MTU model of a linear extension spring (tendon) acting in series with an actuator (muscle). During ankle dorsiflexion / plantarflexion with the foot in the air and during a standing heel raise, we unexpectedly estimated the Achilles tendon to shorten (by an average of 20 and 9 mm, respectively), inconsistent with expectations. Further, for the heel raise tasks, Achilles tendon length change estimates were different between the two ultrasound methods.

My discovery that current ultrasound based methods yielded unexpected shortening of tendon when the ankle plantarflexes exposes a key knowledge gap in how the scientific community experimentally estimates and/or conceptually models MTU dynamics *in vivo*. Identifying this knowledge gap is particularly important because many locomotor tasks of interest involve ankle plantarflexion beyond neutral (e.g. walking, running). As first author, I published “Ultrasound estimates of Achilles tendon exhibit unexpected shortening during ankle plantarflexion” in Journal of Biomechanics in 2018.

1.2.2 Chapter 3 Contributions

Chapter 3 explores the relationship between ground reaction force (GRF) metrics and tibia bone loading during running across speeds and slopes. It is common for both researchers and wearable device developers to assume, report, or interpret GRF metrics to signify increased bone loading or associated injury risk. While it is established that the GRF is not equivalent to the force experienced by structures inside the body, it was unknown if running is a special case, such that increases in GRF are strongly correlated with increases in tibial bone load.

The primary contribution of my work in this chapter is identifying that GRF metrics should not be assumed to be a surrogate for tibia bone loading or associated overuse injury risk while running across speeds and slopes. Specifically, 10 runners participated in a motion analysis study in which four vertical GRF metrics (impact peak, loading rate, active peak, impulse) and two tibial force metrics (peak, impulse) are estimated across a range of running speeds and slopes. Most correlations between GRF metrics and tibial metrics were negligible, weak, or moderate. For instance, correlating GRF impact peak and loading rate (two metrics commonly used in the biomechanics / sports / wearable device communities to infer injury risk) with peak tibial load resulted in average correlations of $r = -0.29 \pm 0.37$ and $r = -0.20 \pm 0.35$ (inter-subject mean and standard deviation), respectively.

My discovery that GRF metrics should not be assumed to be an indicator of tibia bone load highlights the need for more targeted attempts to monitor loading on specific internal structures in order to more effectively monitor overuse injury risks in daily life. As first author, I published “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” in PLoS ONE in 2019.

1.2.3 Chapter 4 Contributions

Chapter 4 explores using wearable sensors to monitor tibia bone loading when running across speeds and slopes. As highlighted in Chapter 3, many metrics from wearable sensors are generally not indicative of, or strongly correlated with, the loads experienced by most musculoskeletal structures in the body. Given the exciting opportunities wearables provide for daily non-invasive monitoring of human movement, a key question to address is: How can we combine data from a few wearable sensors to more accurately estimate tibia bone loading?

The primary contribution of my work in this chapter is that I developed and validated a novel method for fusing wearable sensor data to more accurately estimate peak tibial bone force, compared to conventional methods using a single wearable metric. Using idealized wearable data, or lab-based data distilled to the types of signals feasibly measured with wearables, I develop a physics-based algorithm that reduces force estimation error by a factor or two, and a machine learning algorithm that reduced force estimation error by a factor of four, compared to approaches that use GRF loading rate of other single variable metrics. These algorithms would require just three pieces of wearable sensor hardware: a pressure sensing insole and inertial measurement units on the foot and shank.

By creating these algorithms that combine wearables, musculoskeletal biomechanics and machine learning, I provide a foundation for developing more accurate tools for monitoring musculoskeletal

loading in applied situations. As first author, I published **"**Combining wearable sensor signals, machine learning and biomechanics to estimate tibial bone force and damage during running" in Human Movement Science in 2020. A patent related to this work titled "wearable device to monitor musculoskeletal loading, estimate tissue microdamage and provide injury risk biofeedback" was filed in 2019.

1.2.4 Chapter 5 Contributions

Chapter 5 explores developing a practical and accurate wearable solution for monitoring low back loading during manual lifting tasks. To supplement time-consuming ergonomic assessment and provide tools for personalized daily monitoring of injury risk, there remains a need for automated, unconstrained, and widespread monitoring of low back loading in workplace environments.

The primary contribution of my work in this chapter is that I identify, develop, and validate a practical and accurate tool for monitoring low back loading with wearable sensors. Using idealized wearable data, I identify a promising reduced subset of wearable sensors (trunk inertial measurement unit and pressure-sensing insoles) that when combined with a machine learning algorithm can capture key trends in lumbar moment across a broad range of manual lifting tasks. This solution accurately captures changes in low back loading when lifting objects of varying masses, providing added capabilities over single sensor solutions that emphasize posture and motion. Additionally, this chapter provides benchmark results using real wearable sensors and I identify that additional signal processing or hardware development may be necessary to leverage signals from pressure sensing insoles for this application.

The algorithms I create for monitoring low back loading offer a unique, wearable tool for building our understanding of low back loading during strenuous and repetitive tasks, for monitoring overuse injury risk in real world environments, and for developing new interventions. These findings will be submitted to Sensors Special Issue "Advances in Design and Integration of Wearable Sensors for Ergonomics" in late December 2020. A provisional patent related to this work titled "System and Method for Monitoring Back Loading or Damage, and Providing Biofeedback or Injury Risk Assessment" was submitted in 2020.

1.2.5 Summary of Contribution Deliverables

Manuscripts

Matijevich, E. S., Branscombe, L. M. and Zelik, K. E. (2018). Ultrasound estimates of Achilles tendon exhibit unexpected shortening during ankle plantarflexion. J Biomech. 72, 200-206. https://doi.org/10.1016/j.jbiomech.2018.03.013

Matijevich E. S., Branscombe L. M., Scott L. R., Zelik K. E. (2019). 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 14(1): e0210000.

Matijevich, E. S., Scott, L. R., Volgyesi, P., Derry, K. H., & Zelik, K. E. (2020). Combining wearable

sensor signals, machine learning and biomechanics to estimate tibial bone force and damage during running. Human Movement Science. 74: 102690.

Matijevich, E. S., Volgyesi, P., & Zelik, K. E. Estimating low back loading over a range of manual lifting tasks using a small number of wearable sensors. Planned Submission.

Conference Abstracts

Matijevich, E. S., Branscombe, L. M. and Zelik, K. E. Are ultrasound-based estimates of Achilles tendon kinematics consistent with the expected behavior of a passive elastic tissue in series with muscle? International Society of Biomechanics, July 2017, Brisbane, Australia & American Society of Biomechanics, August 2017, Boulder, CO, USA.

Matijevich, E. S., Branscombe, L. M., Scott, L. R., and Zelik, K. E. Beyond Ground Reaction Forces: Towards Wearable Tech to Monitor Bone Loading & Prevent Injury. American Society of Biomechanics Annual Conference, August 2018, Rochester, MN, USA.

Matijevich, E. S., Branscombe, L. M., Scott, L. R., and Zelik, K. E. Beyond Ground Reaction Forces: Towards a Wearable Device for Monitoring Bone Stress, Preventing Stress Fractures. American Congress of Rehabilitation Medicine, September 2018, Dallas, TX, USA.

Matijevich, E. S., Branscombe, L. M., Scott, L. R., and Zelik, K. E. Wearables and injury prevention: the pitfalls and opportunities for monitoring musculoskeletal loading. ISB/ASB Conference, August 2019, Calgary, Canada & Footwear Biomechanics Symposium, July 2019, Calgary, Canada.

Matijevich, E. S., Scott, L. R., Volgyesi, P., Derry, K. H., & Zelik, K. E., Multi-sensor fusion algorithms offer a promising approach for monitoring tibia bone damage over multiple workouts. Mid-South Movement Biomechanics Conference, February 2020, Memphis, TN, USA.

Patent Applications

Wearable device to monitor musculoskeletal loading, estimate tissue microdamage and provide injury risk biofeedback. PCT/US2019/029790

System and Method for Monitoring Back Loading or Damage, and Providing Biofeedback or Injury Risk Assessment. U.S. provisional patent application no. 63/058,066 (2020).

2 Ultrasound estimates of Achilles tendon exhibit unexpected shortening during ankle plantarflexion

2.1 Abstract

The purpose of this study was to investigate Achilles tendon (AT) length changes during a series of tasks that involved combinations of higher/lower force, and larger/smaller length changes of the medial gastrocnemius muscle-tendon unit (MTU). We sought to determine if common ultrasound-based estimates of AT length change were consistent with expectations for a passive elastic tendon acting in series with a muscle. We tested 8 healthy individuals during restricted joint calf contractions (high force, low displacement), ankle dorsi-/plantar-flexion (DF/PF) with the foot in the air (low force, high displacement), and heel raises (high force, high displacement). We experimentally estimated AT length change using two ultrasound methods, one based on muscle-tendon junction (MTJ) tracking and one based on muscle fascicle (MF) tracking. Estimates of AT length change were consistent with model expectations during restricted calf contractions, when the MTU underwent minimal length change. However, estimates of AT length changes were inconsistent with model expectations during the ankle DF/PF and heel raise tasks. Specifically, the AT was estimated to shorten substantially, often 10-20 mm, when the ankle plantarflexed beyond neutral position, despite loading conditions in which a passive, stiff spring would be expected to either lengthen (under increasing force) or maintain its length (under low force). These unexpected findings suggest the need for improvements in how we conceptually model and/or experimentally estimate MTU dynamics in vivo during motion analysis studies, particularly when the ankle plantarflexes beyond neutral.

2.2 Introduction

The Achilles tendon (AT) is a passive elastic structure that facilitates safe [25], [26] and economical [27]–[32] locomotion, and which informs the development of assistive and rehabilitative interventions [33], [34]. Ultrasound provides a means to non-invasively estimate AT kinematics *in vivo* during human movement. AT kinematics have also been combined with estimates of AT force to compute tendon energy storage and return, which provides additional insights on the functional benefits of tendons, and their interplay with muscle mechanics [35]–[38]. Although these ultrasound-based tendon estimates have been employed for decades, estimating AT kinematics and kinetics *in vivo* remains challenging to validate [39], [40], which can confound scientific interpretation of movement biomechanics. Critical questions remain, including: which (of several) ultrasound-based estimation methods are most accurate, and under what circumstances do these methods yield reliable estimates of tendon dynamics.

Multiple ultrasound tracking methods have been developed and employed on humans and animals to study the AT and other MTU dynamics. Certain methods estimate the distance from the muscle-tendon junction (MTJ) to the tendon's distal insertion [36], [41]–[44]. Other methods seek to track muscle fascicle (MF) length, then subtract muscle length from estimates of overall muscle-tendon unit (MTU) length [45], [46] to approximate tendon kinematics [35], [40]–[42], [47]–[51]. Yet other methods quantify local elongations of the tendon [52]–[55]. While the tracking methods themselves rely on slightly different methodological assumptions, researchers typically map experimental data onto a similar conceptual model of the MTU and adopt a similar set of assumptions when interpreting

the results. Several common model assumptions include: the AT acts longitudinally in series with muscle [30], [49], [56], [57], individual MTUs can be analyzed in isolation [56], and MTU length is primarily a function of joint angle [45], [46].

A variety of methodological factors can affect ultrasound estimates of tendon dynamics [40]. Experimental inconsistencies (e.g., in stiffness and hysteresis, [39]) and oddities (e.g., AT estimated to return more energy than it stores, despite being a passive structure, [47], [58]) in the published literature suggest that ultrasound estimates may be prone to errors. Ultrasound methods, or model assumptions, may begin to breakdown under different mechanical loading conditions or movement tasks, but this has not explicitly been tested. Therefore, the purpose of this study was to investigate a series of movement tasks that involved combinations of higher or lower MTU force, and larger or smaller MTU length changes of the medial gastrocnemius (MG), to determine if common ultrasound-based estimates of AT kinematics were consistent with expectations for a passive spring-like tendon acting in series with a muscle. In this study, we estimated AT kinematics using two common ultrasound-based tracking methods, MTJ and MF (summarized above, and detailed in *Methods*); thus, a secondary aim was also to compare the consistency of results between these methods.

2.3 Methods

2.3.1 Subjects

Eight healthy subjects participated (5 M / 3 F, age = 21 ± 2 years, mass = 77 ± 12 kg, height = 1.79 ± 0.04 m). Each subject performed a series of movement tasks while lower-body kinematics and ground reaction forces, as well as B-mode ultrasound of the MG MFs and MG-AT MTJ were collected (**Figure 2.1**). Electromyography (EMG) of the MG, lateral gastrocnemius (LG), soleus (SOL), and tibialis anterior (TA) was also recorded (further detailed in **Chapter 2 Supplemental Material**). All subjects gave informed consent to the protocol, which was approved by the Institutional Review Board at Vanderbilt University.

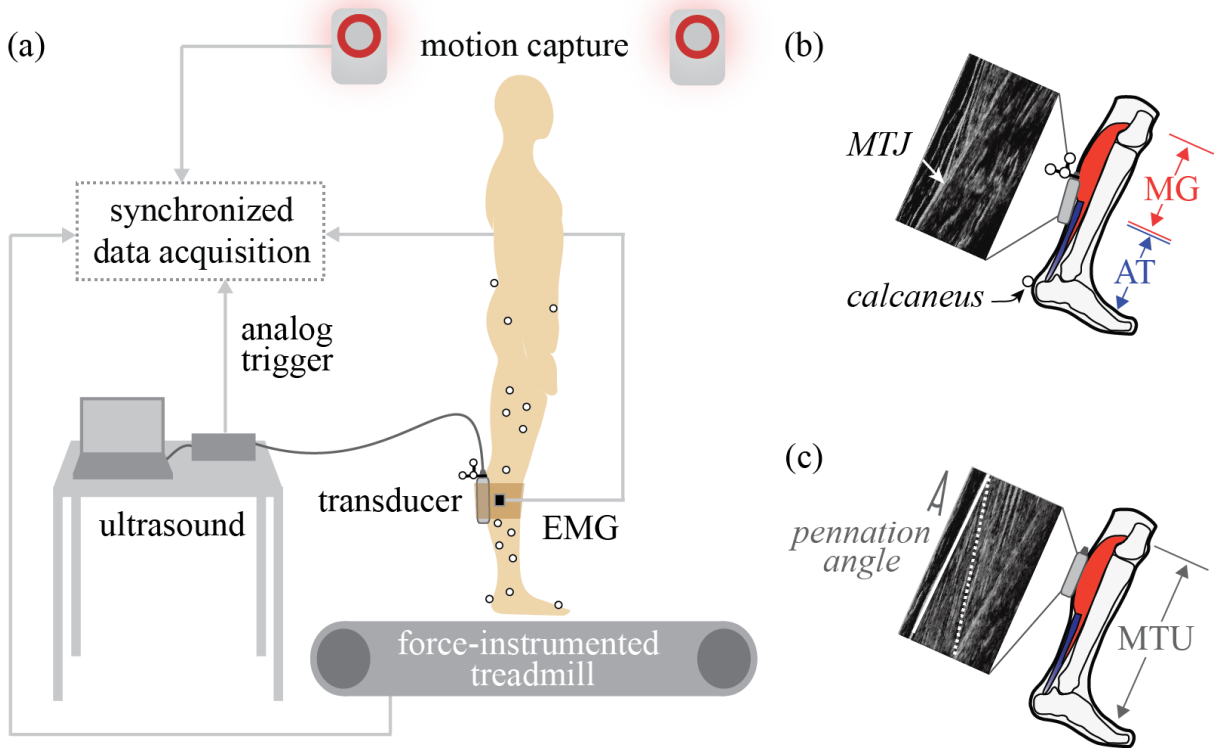


Figure 2.1 Experimental methods. (a) B-mode ultrasound was collected synchronously with motion capture, electromyography (EMG), and ground reaction force data. (b) MTJ tracking. The ultrasound transducer (gray box) was placed over the MTJ of the MG muscle and AT. AT length change was estimated as changes in the straight-line distance from the MTJ to the calcaneus. (c) MF tracking. The transducer was placed over the MG muscle belly. Longitudinal MG length change was calculated from MF length (dotted white line) corrected by the cosine of the pennation angle. Pennation angle was defined as the angle between MF and the superficial fascia (solid white line). MTU length change was estimated from a regression equation based on joint angles.

2.3.2 Experimental Protocol

Subjects performed tasks that involved combinations of high or low force on the MTU, and large or small MTU length changes. The terms “high” or “low” force and “large” or “small” length change are used to signify magnitudes relative to other tasks tested. Restricted joint calf contractions involved high force and small length changes of the MTU (**Figure 2.2a**). Ankle dorsiflexion/plantarflexion (DF/PF) with the foot in the air involved low force and large length changes of the MTU (**Figure 2.2b**). Heel raises involved high force and large length changes of the MTU (**Figure 2.2c**). Prior to data collection, each subject walked ~300 steps to pre-condition their AT [43]. For each task, about 10 cycles were performed to a metronome, paced at 40 beats per minute, to ensure a slow and smooth motion with minimal soft tissue dynamics.

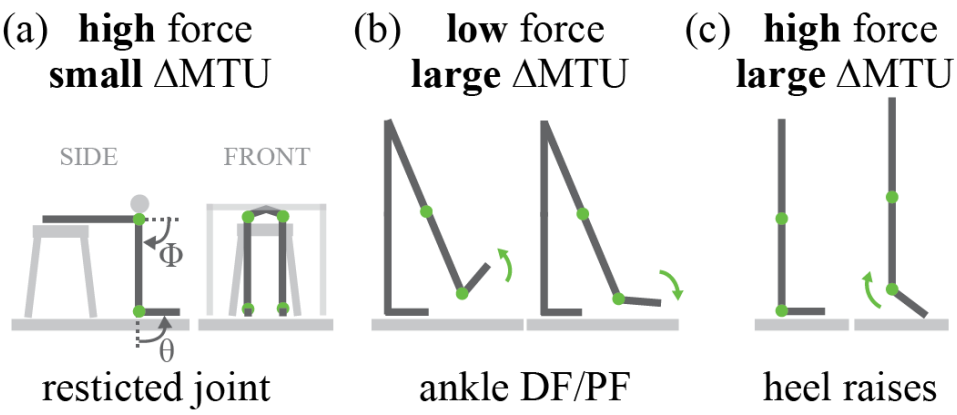


Figure 2.2 Experimental tasks. **(a)** Restricted joint calf contractions, involving high force and small MTU length change (ΔMTU). Subjects were seated on a stool with a rigid bar affixed above their knees, restricting both ankle and knee rotation. The stool and rigid bar were positioned such that the knee was flexed about 90° and the ankle was in the neutral (90°) position. The subject was relaxed, then contracted their calf muscles to push upward against the bar, then returned to a relaxed state. Gray arrows indicate ankle joint angle (θ) and knee joint angle (φ) conventions **(b)** Ankle DF/PF with foot in the air, involving low force and large ΔMTU. Subjects stood on their left foot with their right foot off the ground. Subjects began with their ankle in neutral position, then dorsiflexed their ankle, then fully plantarflexed their ankle, then returned to neutral. **(c)** Heel raises, involving high force and large ΔMTU. Subjects stood flat-footed with normal posture, then contracted their calf muscles to rise up off their heels, then relaxed to return to the flat-footed posture. Green arrows indicate motion.

2.3.3 Kinematics

Kinematics were collected at 100 Hz (Vicon), then low pass filtered at 6 Hz (3rd order, zero-lag Butterworth) prior to computing joint angles. 4 markers were placed bilaterally on the pelvis, and 2 bilaterally on the greater trochanters. Additional markers were placed unilaterally on the right limb: 4 on each segment (thigh, shank, foot), 2 on the lateral and medial femoral epicondyles, and 2 on the lateral and medial malleoli. Functional joint centers were computed using C-Motion Visual3D software and joint angles were calculated using the convention described in [45]; neutral ankle position (foot orthogonal to shank) was defined as 90°, and fully extended knee was defined as 0°. Increasing ankle angles indicated increasing dorsiflexion, and increasing knee angles indicated increasing flexion (**Figure 2.2a**).

2.3.4 Kinetics

Ground reaction forces were collected independently under each foot at 2000 Hz during heel raises only, using a force-instrumented split-belt treadmill (Bertec). Forces were low-pass filtered at 15 Hz (3rd order, zero-lag Butterworth).

2.3.5 Ultrasound

B-mode ultrasound was used to image the MG-AT MTJ or the MG MFs (**Figure 2.1b,c**). Images were collected at approximately 60 Hz in B-mode with a 60 mm field of view and 50 mm depth (Echo Blaster 128, LV7.5/60/128Z-2 transducer, Telemed). Ultrasound data were synchronized with the other measurement modalities via an analog trigger (**Figure 2.1a**), using time stamp data and a synchronization time delay that was quantified in preliminary experiments (similar to Rousseau et al., 2006). The ultrasound transducers were localized (position and orientation) in the lab reference frame using a custom 3D-printed fixture with motion capture markers [60]. For subjects 1-4, a single ultrasound transducer was positioned to track the MTJ, and each task was performed. Directly afterwards, tasks were repeated with the transducer positioned to track the MG MFs. Prior to data collections on subjects 5-8, a second (identical) ultrasound system was acquired, and for these subjects, MFs and MTJ were imaged simultaneously.

2.3.6 Data Analysis

Muscle, AT and MTU length changes of the MG were estimated using established MTJ tracking methods [36], [41], [42], and MF tracking methods [35], [40]–[42], [47]–[51]. The MG MTU length change was estimated from a regression equation based on ankle and knee kinematics [45]. Via the MTJ tracking method, AT length change was estimated using the straight-line distance from the MTJ (tracked in the ultrasound images, then localized in the motion capture reference frame) to the calcaneus (tracked with motion capture, **Figure 2.1b**). MG muscle length change was then estimated by subtracting AT length change from the overall MTU length change. Via the MF tracking method, MG muscle length change was estimated from MF length changes corrected by pennation angle. AT length change was then estimated by subtracting MG length change from MTU length change (**Figure 2.1c**). See **Chapter 2 Supplemental Material** for further details on these tracking methods. AT force was estimated for the heel raise task using standard inverse dynamics to estimate ankle moment, then assuming a constant AT moment arm to estimate AT force, similar to [35], [38].

To address the primary aim of this study, AT length change waveforms were qualitatively compared against expectations for a passive tendon acting in series with a muscle (expectations detailed below). Ultrasound and force data were each resampled to 100 Hz to match motion data. Data from each task cycle were normalized to 1000 data points (representing 0-100% cycle). For each task, on a subject-specific basis, data were averaged over five sequential cycles. Muscle, AT and MTU length changes were non-dimensionalized (divided by subject-specific shank length to account for size differences between subjects) before computing inter-subject means and standard deviations. For reporting purposes, length change results were re-dimensionalized by multiplying by average subject shank length (424 mm). Maximum AT lengthening and maximum AT shortening over an average cycle were computed as summary metrics for each task.

To address the secondary aim, we computed the Pearson correlation between average AT length change waveforms estimated from MTJ and MF tracking methods. Correlation coefficients were computed for each task and averaged across subjects.

2.4 Model Expectations

A common MTU model was used to determine the expected AT behaviors during each task. This simple model is comprised of a passive linear extension spring (representing tendon) acting in series

with an actuator (representing muscle). Here we briefly summarize expectations (**Figure 2.3**). See **Chapter 2 Supplemental Material** for detailed rationale. During the restricted calf contraction (high MTU force, small MTU length change), the AT was expected to lengthen, with magnitude roughly equal to longitudinal muscle shortening. During the ankle DF/PF task (low MTU force, large MTU length change), the AT was expected to exhibit negligible length change. During heel raises (high MTU force, large MTU length change), the AT was expected to lengthen as MTU force increased; albeit with a double-peak length change profile that followed the expected AT loading profile.

2.5 Results

2.5.1 Restricted Joint Calf Contractions

As expected, we estimated AT lengthening and MG muscle shortening for all subjects and both tracking methods (**Figure 2.3, top row**). On average, the AT lengthened a maximum of 5.4 $\pm$ 3.1 mm for the MTJ method and 8.4 $\pm$ 3.1 mm for the MF method (**Table 2.1**, *N=7*). Muscle shortening magnitude was typically slightly larger than AT lengthening (**Figure 2.3, top row**). One subject was omitted from analysis of this task due to negligible MG activation and length change. AT length changes were strongly correlated between MTJ and MF methods (r = 0.90$\pm$ 0.07, with min = 0.83, max = 0.99, *N=7*).

2.5.2 Ankle DF/PF with Foot in the Air

Contrary to expectations, we estimated substantial AT shortening for all subjects (**Figure 2.3**, **middle row**). Maximum shortening was 19.4 $\pm$ 8.8 mm for the MTJ method, and 19.7 $\pm$ 5.6 mm for the MF method (**Table 2.1**, *N=8*). AT shortening was roughly proportional to ankle angle for angles <90° (Figure 2.4**, left column**). For ankle angles >90°, AT length changes were small, typically less than a few millimeters, as expected. AT length changes were strongly correlated between MTJ and MF tracking methods (r = 0.91 $\pm$ 0.10, with min = 0.71, max = 0.98, *N*=8).

2.5.3 Heel Raises

For the MTJ method, the AT was estimated to lengthen slightly (3.4 $\pm$ 1.7 mm) at the beginning of the movement cycle, and then to shorten substantially (9.3 $\pm$ 5.6 mm, **Table 2.1,** *N=8*). This large shortening was inconsistent with the expected behavior of the AT under these loading conditions (**Figure 2.3, bottom row** and **Supplemental Figure 2.5**). For most subjects, we again observed a roughly linear relationship between AT length change and ankle angles for angles <90° (Figure 2.4**, right column**). For the MF method, AT length change estimates varied greatly across subjects. Some subjects exhibited primarily tendon lengthening and others primarily tendon shortening (**Supplemental Figure 2.6**). Comparing MTJ vs. MF methods, we found high inter-subject variability and that AT length changes were only weakly correlated, on average (r = 0.44 $\pm$ 0.45, with min = -0.60, max = 0.80, *N=8*, Supplemental **Figure 2.6**).

See **Chapter 2 Supplemental Material** for kinematics and EMG results.

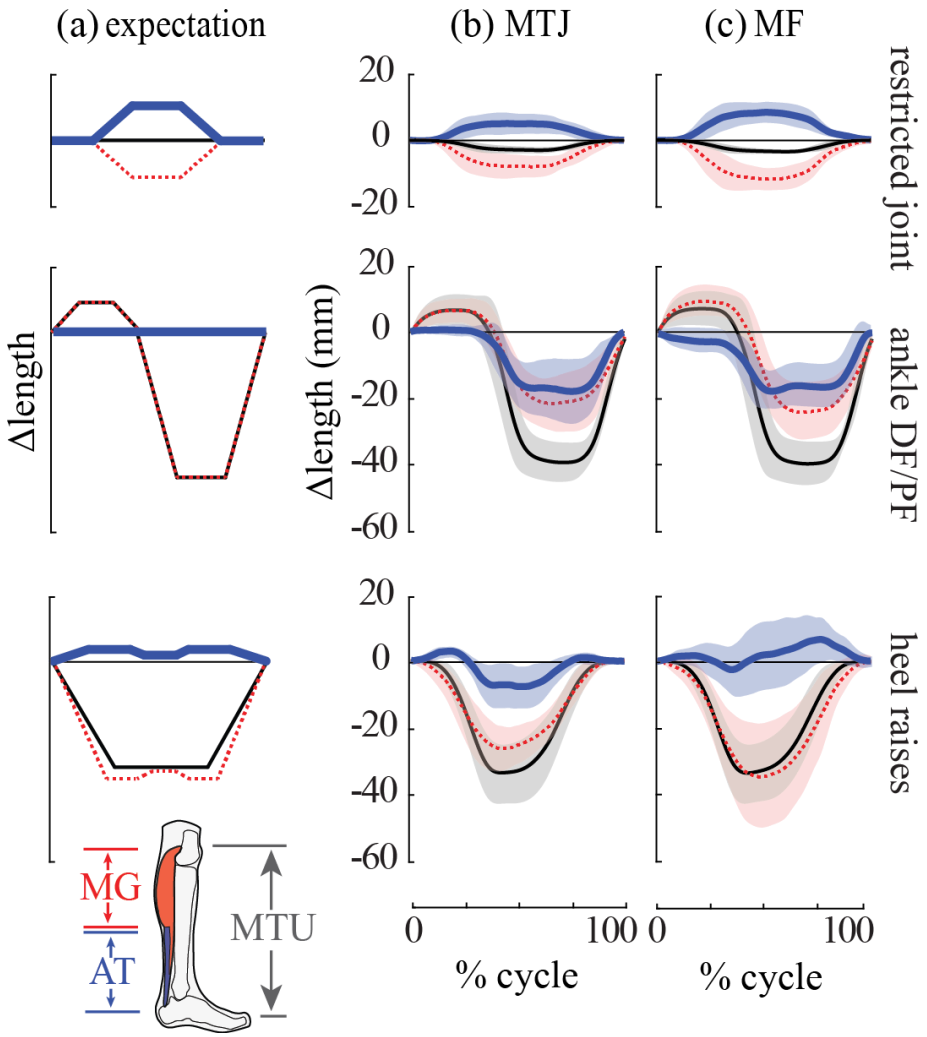


Figure 2.3 Expected vs. subject-averaged results. Rows indicate each task. (a) Expected behaviors of the MG muscle (dashed red), AT (thick blue) and MTU (thin black) based on a simple model of the MTU, consisting of a passive extension spring (tendon) acting in series with an actuator (muscle). Expectations are qualitative, so no units are provided on axes. Experimental results from the (b) MTJ tracking method and (c) MF tracking method. Depicted are inter-subject means (lines) and standard deviations (shaded regions). Length change waveforms (in mm) are plotted as a percentage of each movement cycle.

Table 2.1 Maximum and minimum AT length changes for each task.Positive values indicate tendon lengthening and negative values indicate tendon shortening compared to length at 0% cycle. Data were non-dimensionalized (divided by subject-specific shank length), then averaged across subjects, and finally re-dimensionalized (multiplied by average shank length across subjects) for reporting purposes. All values are in mm, mean ± standard deviation.

	restricted joint (*N*=7)		ankle DF/PF (*N*=8)		heel raises (*N*=8)	
	MTJ	MF	MTJ	MF	MTJ	MF
AT max	5.4 ± 3.1	8.4 ± 3.1	1.4 ± 1.0	1.8 ± 1.9	3.4 ± 1.7	9.5 ± 7.1
AT min	-0.5 ± 0.4	-0.4 ± 0.5	-19.4 ± 8.8	-19.7 ± 5.6	-9.3 ± 5.6	-5.3 ± 5.5

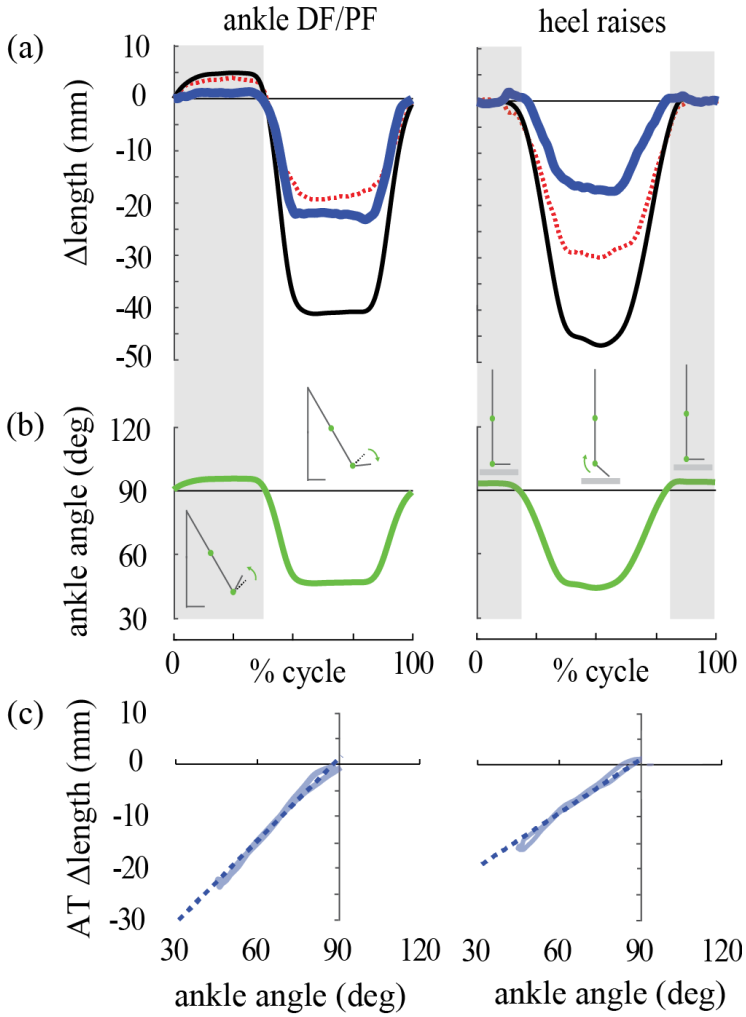


Figure 2.4. AT length change and ankle angle data for a representative subject. Results for the MTJ tracking method are shown for ankle DF/PF task (left column) and heel raises (right column). **(a)** AT (thick blue), muscle (dashed red) and MTU (thin black) length changes vs. movement cycle. **(b)** Ankle angle vs. movement cycle. 90° signifies neutral position, and decreasing angles indicate plantarflexion. **(c)** AT length change vs. ankle angle plotted only when the ankle was plantarflexed beyond neutral (<90°, white background of **(b)**). During these periods, the AT shortened proportionally (roughly linearly, dashed line) with decreasing ankle angle.

2.6 Discussion

We found that AT kinematics were highly inconsistent with model expectations when the MTU underwent large length changes due to ankle plantarflexion beyond neutral. For instance, the AT was

estimated to shorten substantially during the ankle DF/PF task (by an average of about 20 mm, **Figure 2.3, middle row**), despite being a stiff passive structure under low force. Although ultrasound-based estimates of AT kinematics were consistent with our model-based expectations when the MTU underwent minimal length changes, scenarios in which the ankle joint is restricted are of limited utility for studying locomotion. The observed incongruence between the experimental estimates of AT length change and the model-based expectations represents a key obstacle to the research field in terms of confidently interpreting tendon function during movement tasks that involve non-negligible ankle plantarflexion. These non-intuitive findings call attention to a pressing need, and important opportunity, to develop improved experimental estimation methods and/or conceptual models of MTU dynamics.

2.6.1 Unexpected Tendon Shortening

The most striking and unexpected observation was that substantial AT shortening was estimated as the ankle plantarflexed beyond the neutral position. At ankle angles <90°, there was a surprisingly linear relationship between ankle angle and AT length change, such that with more plantarflexion the AT shortened proportionally (Figure 2.4). This relationship was observed consistently for all subjects during the ankle DF/PF task, and for about half of the subjects during the heel raise task. Our observations were qualitatively consistent with [61] who estimated AT shortening of up to 13 mm when the ankle plantarflexed 20° beyond neutral. At ankle angles >90° (i.e., more dorsiflexed), the AT length change estimates in our study were more consistent with model-based expectations (i.e., small length changes under low force, **Figure 2.3, middle row**).

These findings have potential implications for the interpretation of AT function during locomotor tasks that involve substantial ankle plantarflexion. During walking, energy storage occurs primarily at ankle angles >90°, i.e., ankle angles in this study when we observed expected AT elongations. However, energy return occurs over a larger range of motion that includes angles <90° [35], [47], [50], i.e., ankle angles when we observed unexpected AT shortening. It may be that AT shortening (and therefore energy return) is overestimated when employing the ultrasound methods described in this study. It is often presumed that the AT accounts for the vast majority of positive ankle work near the end of stance in human gait (e.g. [27], [29], [62], [63]). However, it has also been documented that common ultrasound-based methods can estimate 2-5 times more AT energy return than energy storage, which is implausible for a passive tendon [58]. Based on findings our current study, and also observations from other researchers (using separate lines of reasoning and observation, e.g. [64]), it may be necessary to further explore the relative contribution of the AT to ankle work generation in human gait.

It is challenging to pinpoint the precise reason(s) for the unexpected AT shortening estimates, though primary source(s) of error may reside in the assumed MTU model and/or experimental methods. Collecting data and mapping results onto a model requires many choices (e.g., which features of the musculoskeletal system to track and how to associate measurements with model features). A simplified model of the MTU must also be assumed to fuse ultrasound, motion and force data. A variety of methodological choices and/or model assumptions could contribute to unexpected AT length change estimates. The impact of these choices may be highly dependent on the task, MTU range of motion or loading pattern. Interestingly, the ankle DF/PF task had consistent results between methods (**Figure 2.3**), suggesting errors may reside in a shared assumption between methods. On the

contrary, inconsistent results in the heel raise task (**Figure 2.3**) suggest errors may be in assumptions unique to each method. Unexpected AT shortening in both the ankle DF/PF and the heel raise task (Figure 2.4) may further suggest errors in the shared conceptual model. Follow-up studies are needed to unmask the primary culprit(s).

2.6.2 Where Might the Simplified Conceptual Model Go Wrong?

A foundational assumption of the MTU model is that a tendon is loaded in series with a muscle. However, some researchers have argued against this assumption, noting that the actual loading behavior may be complex and highly non-intuitive [65], [66]. Thus, the 1-dimensional, in-series model may fail to capture important dynamics. Another implicit model assumption is that MTUs can be studied in isolation, ignoring forces from adjacent MTUs. There is some evidence suggesting that muscle loading is borne primarily along individual tendon fascicles [55], [67]. However, there also exist intermuscular interactions via epimuscular linkages [68], as well as sheet-like aponeuroses and other connective tissues that interconnect the plantarflexors, and thus transverse forces and biaxial strain may affect tendon dynamics [69], [70].

2.6.3 Where Might the Experimental Methods Go Wrong?

Each method provides a partial snapshot of the MTU. Various assumptions in relating features of the ultrasound images to actual tissue motion could introduce errors. For the MF method, correcting localized muscle fascicle length change by pennation angle is assumed to provide a reasonable approximation of overall muscle length change, however this may not always be valid [71], [72]. For the MTJ method, it is assumed that the junction feature tracked in each ultrasound frame is representative of the overall MTJ displacement. However, the MTJ is a complex 3D interweaving of muscle and tendinous tissues [72], making it difficult to precisely track the same feature in each frame, or to comprehensively capture tissue dynamics. For both methods, the AT is assumed to have a linear connection between the MTJ and calcaneus, but this assumption begins to break down if the tendon curves due to muscle bulging or wrapping around the calcaneus, or if the tendon becomes slack. See **Chapter 2 Supplemental Material** for extended discussion of these potential confounds and how much of the unexpected tendon shortening they might explain.

2.6.4 MTJ vs. MF Tracking Methods

As evident in **Figure 2.3**, and confirmed via correlation analysis, the MTJ and MF methods yielded similar AT length change estimates during restricted joint calf contraction and ankle DF/PF tasks. However, notable differences in AT length change were observed for most subjects during heel raises (**Figure 2.3, bottom row** and **Supplemental Figure 2.6**). The inconsistent results during the heel raise task (high force, large MTU displacement) suggest that using different methods to estimate AT behavior during other tasks involving high force and large MTU displacement may lead to disparate results. Indeed, prior studies of walking [58], running [42] and other plantarflexor tasks [73] provide more direct evidence supporting this implication.

2.6.5 Limitations

This study evaluated the consistency between simple model expectations and experimental estimates of AT length change. We did not collect ground truth tendon length change or force data, as these would require invasive sensor implantation, and non-invasively partitioning individual muscle forces remains a grand challenge in the field. Experimental tasks were carefully selected to avoid the need for precise ground truth measures. Results were interpreted in the context of whether the AT lengthened or shortened under higher or lower loading conditions. Methodological limitations related to the MTJ and MF methods were discussed above and have been well-documented in prior literature [40], [58]. For Subjects 1-4, the MFs and MTJs were not imaged simultaneously; therefore the task performance varied slightly.

2.7 Conclusion

Our ability to correctly infer muscle and tendon function depends on using well-validated methods that can map experimental data onto a model, and employing a model that adequately captures the salient features and dominant dynamics of the physiological MTU being studied. The AT is often conceptualized as an extension spring acting in series with muscle, and ultrasound imaging provides a non-invasive means of peering underneath the skin at these tendon dynamics. However, here we observed simple movement tasks in which the AT was empirically estimated to shorten despite loading conditions in which a passive spring would be expected to either stretch (under increasing force) or maintain its length (under low force). These unexpected findings suggest the need for improvements in how we conceptually model and/or experimentally estimate MTU dynamics *in vivo* during motion analysis studies, particularly when the ankle plantarflexes beyond neutral.

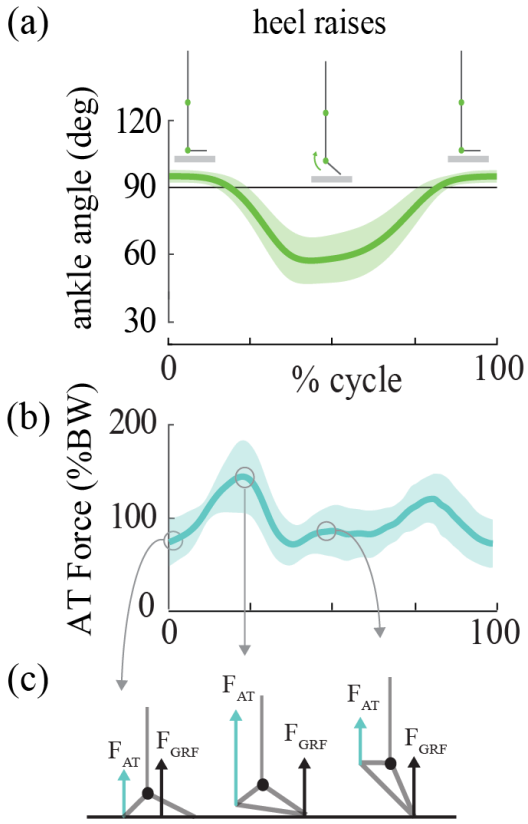
2.7.1 Acknowledgements

This work was supported by funding from the National Institutes of Health (K12HD073945) and from Vanderbilt University. The authors would like to thank Ethan Jones for his assistance with data collection and analysis.

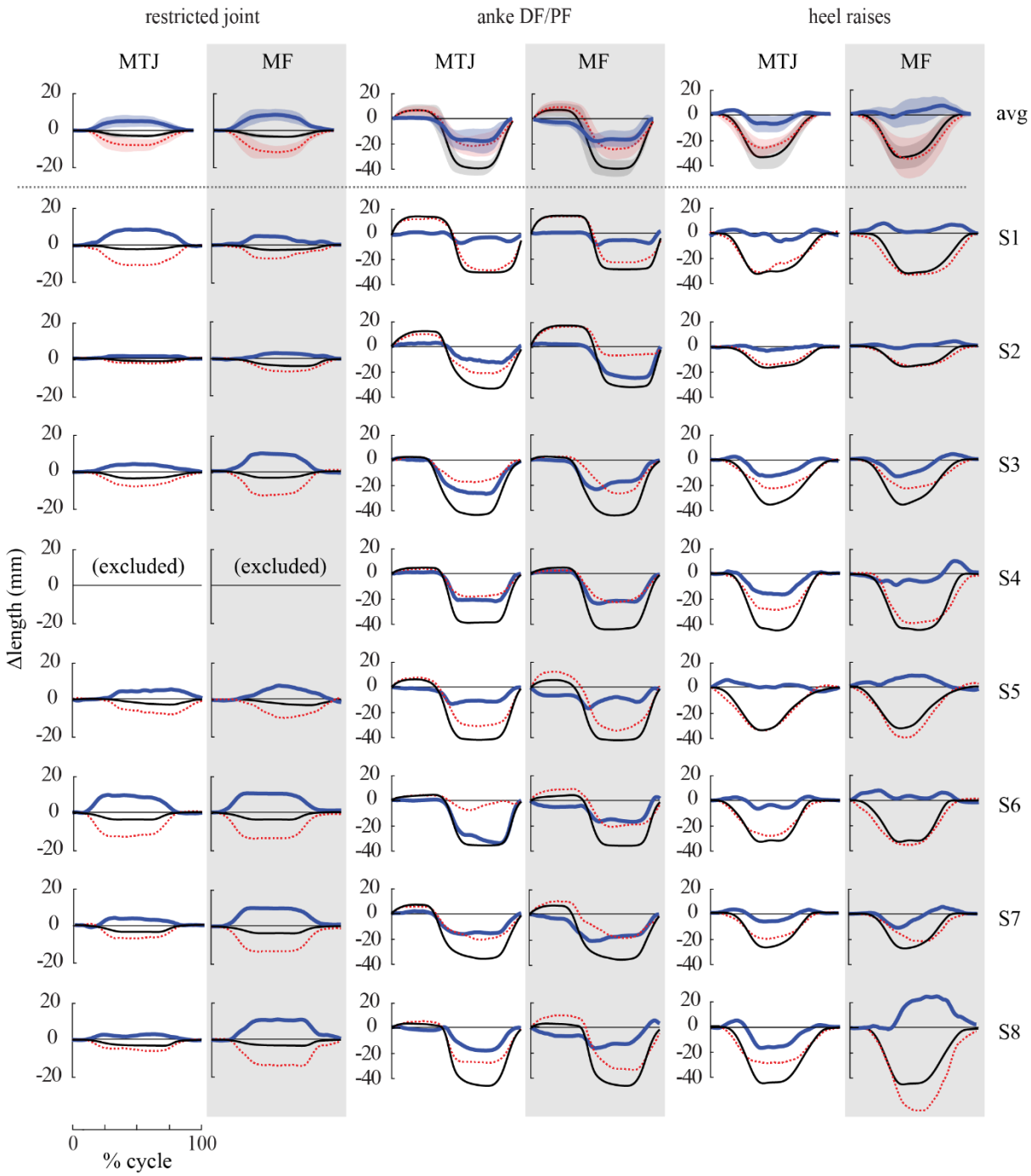
2.8 Abbreviations

AT – Achilles tendon
DF/PF – dorsiflexion/plantarflexion
MF – muscle fascicle
MG – medial gastrocnemius
MTJ – muscle-tendon junction
MTU – muscle-tendon unit
GRF – ground reaction force
EMG - electromyography
LG – lateral gastrocnemius
SOL – soleus
TA - tibialis anterior

2.9 Chapter 2 Supplemental Material



Supplemental Figure 2.5. Heel raise kinematics and kinetics. **(a)** Ankle angle vs. movement cycle. **(b)** Estimated AT force vs. movement cycle. AT force (F_{AT}) was estimated by dividing ankle moment (from standard inverse dynamics) by AT moment arm about the ankle and expressed as percentage of body weight (%BW). Depicted are inter-subject means and standard deviations (shaded regions). **(c)** Simplified moment balance about the ankle joint. The double-peaked AT force profile occurs because the AT has a roughly constant moment arm about the ankle whereas there is a dynamically-changing moment arm from the ankle to the center of pressure (the point of application of the ground reaction force, F_{GRF}).



Supplemental Figure 2.6. Subject-specific results. AT (thick blue), MG muscle (dashed red) and MTU (thin black) length changes vs. movement cycle. Major columns are tasks. Unshaded columns show the MTJ method results and shaded columns show the MF method results. The top row represents the mean and standard deviation across subjects. Remaining rows each represent subject-specific average results.

2.9.1 MTJ Tracking Methods

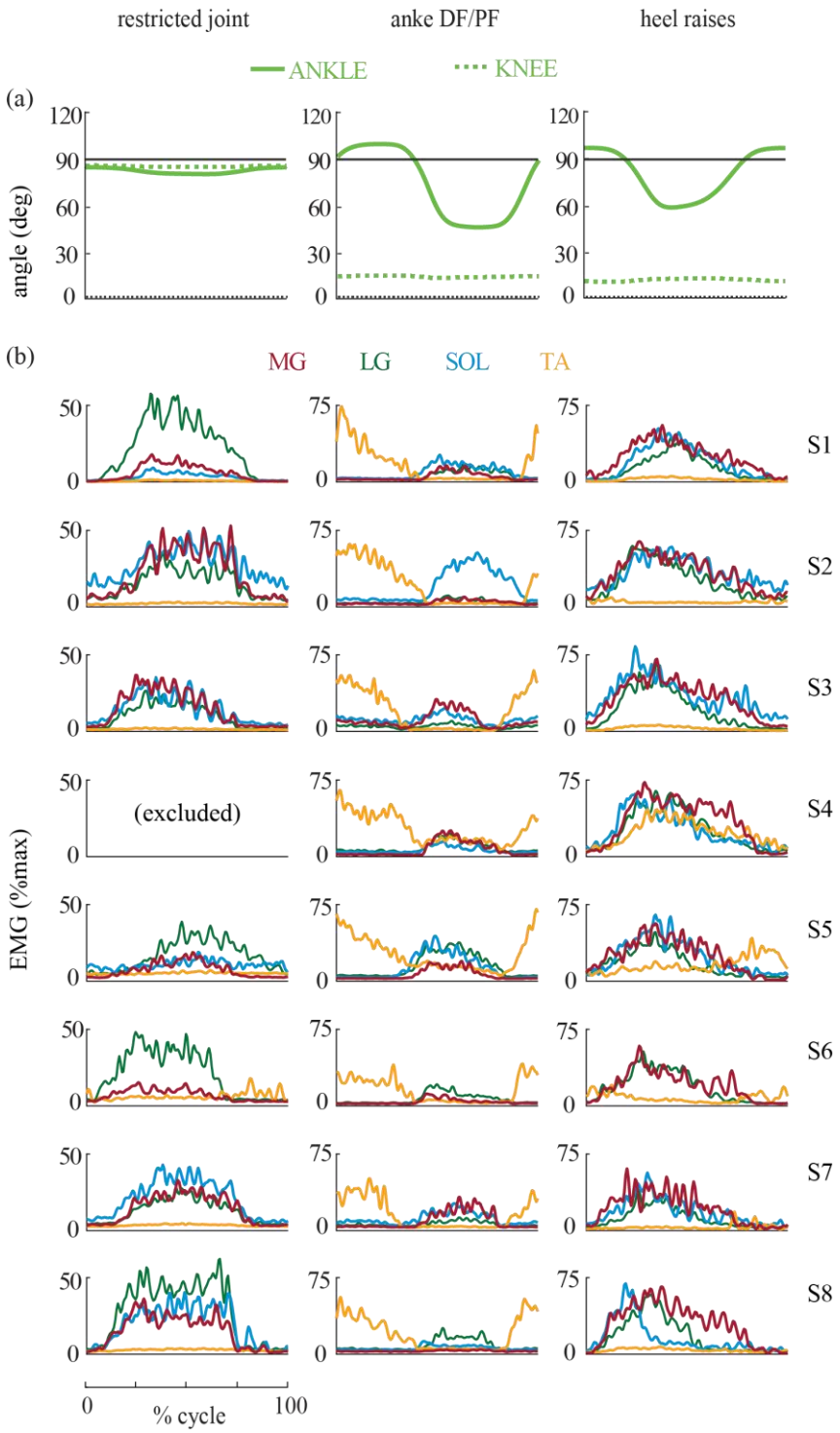
We estimated AT, muscle, and MTU length changes using previously published MTJ tracking methods [36], [41], [42]. The ultrasound transducer was placed over the MG-AT MTJ, then securely affixed to the shank using HypaFix® retention tape and ACE bandages (**Figure 2.1b**). To approximate AT origin, the characteristic "Y" of the MTJ was manually tracked in each ultrasound image. A calcaneus motion capture marker approximated AT insertion. A custom 3-D printed fixture attached to the transducer allowed tracking of its position and orientation with respect to the lab reference frame. The ultrasound field of view was assumed to be perpendicular to the longitudinal plane of the transducer. AT length was estimated as the straight-line 3D distance from the calcaneus to the MTJ. MG muscle length change was then approximated as the difference between MTU length change (from a regression equation based on ankle and knee angles, [45]) and AT length change.

2.9.2 MF Tracking Methods

We estimated AT, muscle, and MTU length changes using previously published MF tracking methods [35], [40]–[42], [47]–[51]. The transducer was placed over the MG muscle belly, then securely affixed to the shank using HypaFix® retention tape and ACE bandages (**Figure 2.1c**). An automated affine flow algorithm was used to estimate time-varying length of an individual MG muscle fascicle [74]. Manual corrections of the fascicle endpoints were made for frames in which the automated calculation did not track the fascicle well, based on visual inspection of a trained researcher. If the fascicle extended beyond the field of view, the endpoint locations were estimated via linear extrapolation (i.e., by extending the lines of the fascicle and aponeuroses beyond the field of view and finding the intersection point). Pennation angle was calculated for each frame as the angle between the superficial fascia and the muscle fascicle. Changes in fascicle length were then corrected for pennation angle to approximate changes in MG muscle length along the muscle's line of action. AT length change was then estimated as the difference between MTU length change (from a regression equation based on ankle and knee angles,
[45]), and MG length change.

2.9.3 EMG Data Collection & Analysis

EMG data were collected to supplement primary outcome measures, and used to confirm that muscle activity was consistent with loading expectations for each task. Surface EMG sensors (Delsys Trigno) were placed unilaterally on the medial gastrocnemius (MG), lateral gastrocnemius (LG), soleus (SOL), and tibialis anterior (TA). EMG signals were demeaned, high-pass filtered at 150 Hz, rectified, and low-pass filtered at 10 Hz, and then normalized based on the maximum muscle activation magnitude that occurred during the three recorded tasks [75], [76]. The resulting EMG envelopes from each cycle were normalized to 1000 data points (representing 0 to 100% of the cycle). For each task, on a subject-specific basis, EMG envelopes were averaged over five sequential cycles (**Supplemental Figure 2.7b**). For subjects in which MTJ and MF tracking trials were performed separately, EMG is only reported for the MF tracking trials.



Supplemental Figure 2.7. Kinematics and subject-specific EMG. (a) Ankle (solid green) and knee (dashed green) joint angles vs. movement cycle, averaged across eight subjects. **(b)** Medial gastrocnemius (MG, red), lateral gastrocnemius (LG, green), soleus (SOL, blue), and tibialis anterior (TA, orange), muscle activation vs. movement cycle. EMG magnitudes are reported as a percentage of the maximum activation observed during these three tasks. Each column is a task. Each row represents subject-specific average results.

2.9.4 Explanation of Model Expectations for Each Task

Below we summarize the expected behaviors of a common MTU model: a passive linear extension spring (representing the AT and associated aponeurosis) acting in series with an actuator (representing the muscle). In this study the model represents the MG MTU, which spans distally from the calcaneus (heel), posteriorly across the ankle and knee joints, and then connects proximally to the thigh segment. Since experimental tasks were performed slowly (quasi-statically) dynamic effects due to inertia were ignored for this simple model. It was assumed that the extension spring remained linear and did not become slack (see Extended Discussion below on slack length).

The restricted joint calf contraction task is akin to fixing each end of the MTU model, then ramping up/down the actuator force. As force increased, we expected that the actuator would shorten and the series spring would lengthen by an equal magnitude (**Figure 2.3a**, **top row**). Experimentally this task was achieved by affixing a rigid bar above the knee and thigh. All subjects, except one, showed activation of the MG, LG, and SOL that was in accordance with muscle forces ramping up and then down (i.e., the loading pattern assumed in order to derive model expectations). All subjects, except one (**Supplemental Figure 2.7b, column 1**), showed negligible TA activation, indicating that antagonistic muscle contraction was not a significant confounding factor. Subjects also exhibited minimal change in ankle angle (**Supplemental Figure 2.7a, column 1**), as expected due to the rigid bar. However, due to soft tissue deformation against the rigid bar, it was not possible to completely restrict joint motion [44], [77]. Nonetheless, the simplified model and experiment task were qualitatively consistent, each capturing the same main muscle and tendon length change behaviors.

The ankle DF/PF with foot in air task was modeled by fixing the MTU at the proximal end, hanging a small mass at the distal end, then slowly driving the actuator to lift and lower the mass. With the foot in the air, relatively low AT force was expected, i.e., only force needed to counteract torque due to the mass of the foot about the ankle joint and to overcome any passive resistance from antagonistic MTUs. The mass of the biological foot is about 1 kg and the center of mass of the foot is about 60 mm anterior to the ankle in neutral position. Assuming an anthropometric AT moment arm of about 50 mm [38], [78], <<1 Nm of torque is needed to support the mass of the foot. Passive joint moments have been estimated to be relatively small across a range of ankle angles. For example, passive ankle moments are expected to be <5 Nm at ankle angles <90°, and <10 Nm for the maximum dorsiflexion angles reached during this task (based on [79]). Again assuming an anthropometric AT moment arm, AT forces would typically be <100 N, and at most be about 200 N (which is only about 7% of the force experienced by the AT during walking at moderate speed, Bogey et al., 2005). Consistent with the expectation of low loading, EMG from the MG and other plantarflexors was relatively low throughout the entire DF/PF ankle range of motion (**Supplemental Figure 2.7, column 2**). Also, the TA exhibited activity during dorsiflexion beyond neutral, but was inactive during plantarflexion beyond neutral (**Supplemental Figure 2.7b, column 2**). Additional forces due to joint friction were assumed to be negligible, since experiments involved young, healthy subjects. AT stiffness values in literature are on the order of 130-470 N/mm [36], [63], [77], [81, p. 200], [82]. Due to these relatively low forces acting on a relatively stiff tendon, we expected AT length change to be small (i.e., less than a few mm) throughout the movement cycle of this DF/PF task (**Figure 2.3a, middle row)**.

Heel raises were modeled by fixing the MTU at one end, with the distal end free to move as the actuator shortened/lengthened under relatively large force. We assumed that there was low loading on the actuator and series spring at the beginning of the cycle, then forces increased. One might intuit

that the MTU force profile over the heel raise cycle would increase monotonically (as one lifts up from the flat footed posture), then plateau at some higher force while the heel was off the ground, then decrease force as a person returned back to foot-flat. The AT force profile is actually somewhat more complex: it is double-peaked due to the dynamically-changing moment arm from the ankle joint to the center of pressure (COP), the point of application of the ground reaction force (GRF). Beginning from foot-flat, as the calf muscles increase their force and lift the heels off the ground, the COP shifts anteriorly towards the distal end of the foot, increasing the ankle-to-COP moment arm and causing a peak in AT force. Then, as the foot rotates about the metatarsophalangeal joints, the ankle shifts anteriorly, decreasing the ankle-to-COP moment arm, and reducing the AT force. As the heel lowers, the cycle is reversed creating a second AT force peak. We confirmed this double-peaked force profile experimentally using motion and ground reaction force data **(Supplemental Figure 2.5**). AT moment arm varies slightly with ankle angle and muscle activation level, but for the ~30° of ankle plantarflexion during the heel raise task, the AT moment arm is expected to increase by less than 10 mm [78], [83]. Even with a 10 mm increase in moment arm, the AT force at peak plantarflexion (~50% cycle) would only be about 15% (~110 N) less than that estimated with a constant moment arm. Based on AT stiffness values in literature [36], [63], [77], [81, p. 200], [82], this reduction in tendon force would reduce tendon length by <1 mm. Regardless of the precise force profile, the key model takeaway is that we would expect the spring to lengthen proportionally with increasing MTU force (**Figure 2.3a, bottom row**). For the heel raise task, EMG results confirmed that all subjects increased activation of the MG, LG, and SOL as the heel began lifting off the ground (**Supplemental Figure 2.7b, column 3**), consistent with the model loading expectations summarized above.

2.9.5 Post-Hoc Corrections to AT Length Change Estimate

For most subjects, as they plantarflexed their ankle beyond neutral in the ankle DF/PF task we observed a roughly linear decrease in AT length (**Figure 2.3, middle row**). It was initially tempting to use this DF/PF task to derive an AT length change correction factor based on ankle angle to remove this trend; however, this approach may be ill-advised until the source of error is better understood; for reasons summarized below. Numerically, we could approximate a proportional relationship between ankle angle and AT length change (Figure 2.4**, left column**) during this low force task, and treat it as a correction factor. We could apply this correction factor, a function of ankle angle, by subtracting it from the AT estimates. As a result, AT length changes would become small throughout the entire DF/PF cycle (qualitatively consistent with model expectations). The correction factor derived from the low force task could also be used to adjust AT length change estimates from the heel raise task (Figure 2.4**, right column**). This would shift the AT behavior towards lengthening, which would again make the experimental results more consistent with model expectations. However, this type of post-hoc adjustment to AT length change would also necessitate adjusting our estimate of either the muscle or MTU length change to compensate, in order to ensure that muscle length change plus tendon length change was still equivalent to overall MTU length change. Presently it is not clear if, or justified why, either the muscle or MTU estimate should be adjusted *post hoc*. Therefore, before applying any such correction factors we advise that it would be prudent to first discern why this AT length change vs. ankle angle trend exists at all. We anticipate that this deeper understanding will inform our conceptual model of MTU dynamics, or motivate specific refinements in the empirical methods used to estimate AT length change.

2.9.6 Extended Discussion of Slack Length & Straight-Line Tendon Assumption

Slackening of the MG and/or AT ([49], [84], [85]) have been estimated to occur at about 20-40° of plantarflexion beyond neutral (i.e., at 50-70° in Figure 2.4), when the ankle is passively rotated. Thus, AT slackening may degrade the accuracy of tendon length estimates during the DF/PF task, a low force ankle rotation task, when ankle angle is less than about 50-70°; highlighting the importance of considering slack length for certain tasks. However, slack length does not seem to explain the substantial tendon shortening (~10 mm) as the ankle plantarflexed from 90° to 70° in the DF/PF task, nor similar shortening reported in prior literature on passive ankle rotation [61]. Furthermore, slack length would not seem to explain the shortening observed during heel raises when the MG was actively contracting (**Supplemental Figure 2.7b, column 3**). Regression-based estimates of overall MTU length [45], [46] are based on cadaver studies, assuming straight-line approximations and that MTU length is only a function of joint angles [86]. Prior studies [87], [88], found that when using a linear approximation, the AT appeared to shorten by 3 mm more than when using an estimate that accounted for curvature (for 30° of plantarflexion beyond neutral). However, this 3 mm error was small compared to the 15 mm of AT shortening we observed for a similar 30° of plantarflexion (during ankle DF/PF, **Figure 2.3, middle row**). To approximate the effect of curvature on our estimates, we added a virtual waypoint 50 mm posterior to the ankle joint center that rotated with the foot segment. We then used the calcaneus, waypoint, and MTJ positions to calculate a piecewise tendon length. Using this piecewise tendon estimate accounted for less than 40% of the total shortening for both the ankle DF/PF and heel raise task at peak plantarflexion, indicating that curvature alone does not appear to explain the majority of the tendon shortening. Potential confounds discussed above do not reflect a comprehensive list. Other potential estimation issues have been described in prior literature (e.g., [40], [58], [66], [89]).

3 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

3.1 Abstract

Introduction: Tibial stress fractures are a common overuse injury resulting from the accumulation of bone microdamage due to repeated loading. Researchers and wearable device developers have sought to understand or predict stress fracture risks, and other injury risks, by monitoring the ground reaction force (GRF, the force between the foot and ground), or GRF correlates (e.g., tibial shock) captured via wearable sensors. Increases in GRF metrics are typically assumed to reflect increases in loading on internal biological structures (e.g., bones). The purpose of this study was to evaluate this assumption for running by testing if increases in GRF metrics were strongly correlated with increases in tibial compression force over a range of speeds and slopes. Methods: Ten healthy individuals performed running trials while we collected GRFs and kinematics. We assessed if commonly-used vertical GRF metrics (impact peak, loading rate, active peak, impulse) were strongly correlated with tibial load metrics (peak force, impulse). Results: On average, increases in GRF metrics were not strongly correlated with increases in tibial load metrics. For instance, correlating GRF impact peak and loading rate with peak tibial load resulted in r=-0.29±0.37 and r=-0.20±0.35 (inter-subject mean and standard deviation), respectively. We observed high inter-subject variability in correlations, though most coefficients were negligible, weak or moderate. Seventy-six of the 80 subject-specific correlation coefficients computed indicated that higher GRF metrics were not strongly correlated with higher tibial forces. Conclusions: These results demonstrate that commonly-used GRF metrics can mislead our understanding of loading on internal structures, such as the tibia. Increases in GRF metrics should not be assumed to be an indicator of increases in tibial bone load or overuse injury risk during running. This has important implications for sports, wearable devices, and research on running-related injuries, affecting >50 scientific publications per year from 2015-2017.

3.2 Introduction

Tibial stress fractures are a common type of overuse injury, associated with the accumulation of bone microdamage due to repeated submaximal loading that causes mechanical fatigue [90], [91]. There is a high prevalence of tibial stress fractures in military recruits [92], recreational and elite runners [93], [94], and other athletes [95]–[97]. Tibial stress fractures result in pain, healthcare costs and reduced physical activity [97], [98]. Moreover, because recovery from tibial stress fracture typically requires rest and/or ankle immobilization (often for 6-12 weeks), this injury commonly results in missed work, decreased productivity, and physiological distress [99].

Factors that influence bone stress injury risk include the bone load intensity (magnitude, direction and duration of load), the rate of bone remodeling (influenced by length of activity and length of rest), and intrinsic factors (age, gender, bone density, geometry, mineral content, etc.) [9]. One potential way to reduce the incidence of bone stress injuries may be to monitor one or more of these risk factors in daily life, use bone fatigue models to estimate the damage accumulation, and then preemptively alert individuals of excessive damage accumulation. This approach might empower individuals, for instance runners, to modify training and allow the bone time to remodel and recover before an injury occurs. The challenge lies in how to implement this preventative solution, since direct measurements of bone

load intensity, bone remodeling, and intrinsic factors are impractical in daily life. In the context of tibial stress fractures, monitoring load intensity might be realized via indirect estimates: using wearable sensors that are capable of estimating tibial bone force.

In the scientific literature, a number of lab-based motion analysis studies have sought to understand and predict overuse injury risks (to the tibia and other internal structures) by monitoring ground reaction force (GRF), as measured by a force plate under the foot [100], [101]. Increases in GRF metrics are routinely assumed to reflect increases in internal structure loading (e.g., tibial bone loading). In an attempt to apply this approach outside of the laboratory, a growing number of consumer wearable devices – targeted largely towards runners and athletes – have been developed that use sensors capable of capturing features or correlates of the GRF. Commonly, wearable devices use one or more of the following: (i) pressure-measuring insoles, which capture localized forces acting normal to the surface of each sensor, and can be summed to estimate a component of the GRF, (ii) accelerometers mounted on the foot or shank, which can provide a correlate of GRF impact peaks [102], [103] or loading rates [104], or (iii) accelerometers mounted on the pelvis, which can be analyzed to approximate the GRF active peak that occurs in midstance of running [20], [105]. Commercial wearable devices then attempt to use these GRF-correlated signals to provide musculoskeletal loading or injury risk feedback to the user.

One limitation of current wearable devices (i.e., research and consumer wearables), as well as with the scientific running literature motivating them, is that they aim to understand, predict or prevent overuse injury risks solely by monitoring GRFs (or GRF correlates). However, GRF (the force between the shoe and the ground) is not the force experienced by structures inside the body, such as bones, muscles or joints [10], [13], [14], [106], [107]; and therefore GRF is not necessarily reflective of the actual repetitive loading that causes overuse injury to these internal structures. From a biomechanical perspective, there are several reasons why monitoring GRF to understand tibial bone loading or risk of tibial stress fracture is potentially problematic, a few of which are summarized below.

First, the load on the tibial bone is generally much larger than the GRF. This is because the vast majority of bone loading is due to muscle contractions during locomotion, not due to GRF; a fundamental insight derived from the work of Giovanni Borelli in the 17th century [106]. During running, peak GRFs are typically 2-3 times body weight, whereas peak forces on the distal end of the tibia are typically 6-14 times body weight, as evidenced by gait analysis ([10], [13], **Figure 3.1a**) and modeling studies [14], [108]. Likewise, a cadaver study that simulated walking using a robotic gait simulator found peak GRFs of 1.1 times body weight, measured with a force platform under the foot, and peak tibial compression force of 4.1 times body weight, measured with a force transducer directly in series with the tibia [107].

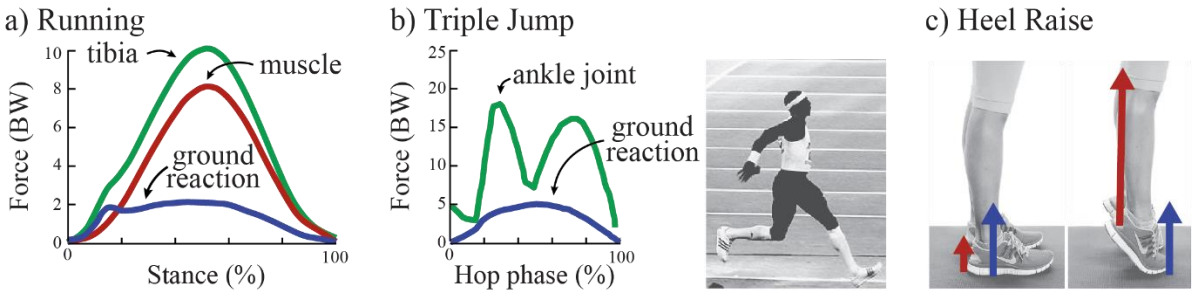


Figure 3.1 Examples of GRFs vs. tibial bone loading. (a) Tibial bone compression force (green) is much larger than GRF (blue) during running due to forces from muscle contractions (red); adapted from [13]. Forces are reported in body weights (BWs). **(b)** Peaks in tibial force (at the ankle joint, green) do not temporally coincide with peaks in GRF (blue) during the triple jump; adapted from [109]. Note, the GRF impact peak is not depicted here because it was not reported in this prior study, but it would have occurred at 0% of the cycle. **(c)** Standing flat footed vs. standing on one's toes results in the same GRF (blue), but different tibial forces, due to calf muscle contraction force (red) [110].

Second, peaks in GRF often do not coincide temporally with peaks in bone force. A GRF peak in running and jump landing often occurs at foot contact (impact peak), but tibial bone loading is typically small at this time. This is evidenced by *in vivo* bone stress and strain measurements [11], instrumented cadavers [107], musculoskeletal models [14], gait analysis studies [13], [111] and data from instrumented joint implants [112]. Peak tibial load in running generally occurs later in the movement cycle, near midstance, and is closer in timing to (though not necessarily coincident with) the second peak of the GRF (often termed *active peak*). In the triple jump, two peaks in ankle joint (distal tibia) contact force have been estimated to happen at very different times than the impact and active peaks in the vertical GRF during the hop phase, due to muscle forces around the joint (**Figure 3.1b**, [109]).

Third, increases in tibial bone forces can occur without increases in GRF [113]. For example, standing flat-footed vs. standing on one's toes results in the same GRF magnitude, but the latter can have much higher bone force due to calf muscle forces (**Figure 3.1c,** [110]). The GRF and tibial force are related through equations of motion [13], [14] which depend on other time-varying factors such as the center-of-pressure under the foot, segment orientations, muscle contraction forces, and the direction of the GRF vector. There may be a subset of activities when increases in GRF metrics are indicative of increases in tibial bone loading; however, this is only expected in very special cases (e.g., if all the other terms in the equation of motion are constant, or nearly constant, or if changes in terms uniquely offset each other as to have negligible effect on total bone loading for a given subset of activities).

Despite these limitations, the use of GRF metrics (e.g., peaks, loading rates) or correlates from wearable sensors (e.g., tibial shock) remains popular amongst researchers and commercial device developers aimed at identifying and reducing overuse injury risks. The advantage of using GRF metrics is that they are easy to measure non-invasively in the lab using force plates, or outside the lab with portable wearable devices (which are relatively cheap and easy to integrate into shoes and clothing). However, a key question remains unanswered: is running a special case, such that increases in GRF are strongly correlated with increases in tibial bone load? If so, then GRF metrics (or GRF-correlates from pressure-insoles or accelerometers) may indeed serve as a useful tool for monitoring tibial bone loading changes during running, supporting the approaches currently used in scientific research and commercial wearable devices. If not, then it would dissuade the use of GRFs as a surrogate for tibial bone loading, and suggest the need to move beyond GRF measures (and GRF-correlates) alone in order to effectively monitor overuse injury risks in daily life. The purpose of this study was to determine if higher GRFs were indicative of (i.e., strongly correlated with) higher tibial bone loads when running over a range of speeds and ground slopes. Because of the complex relationship between GRF and internal bone loading, we hypothesized that increases in common GRF metrics (impact peak, loading rate, active peak, impulse) would not be strongly correlated with increases in tibial bone load metrics (peak force and impulse) across this range of running conditions (i.e., $r<0.8$).

3.3 Methods

Ten healthy subjects participated who each reported that they run a minimum of 10 miles per week (5 male, 5 female; age: 24±2.5 years; height: 1.7±0.1 m; mass: 66.7±6.4 kg). All subjects gave written informed consent to the protocol, which was approved by the Institutional Review Board at Vanderbilt University.

We selected a subset of running conditions that a recreational runner might encounter on a daily run. A fully comprehensive condition set (i.e., all plausible combinations of speed, ground slope, step frequency, footstrike pattern, footwear, terrain stiffness, fatigue level, etc.) was not feasible to test in lab. Thus, we had to select a subset of conditions to explore. We note that typical speeds and slopes will be different for each individual runner and across different runs/days (based on their fitness, training environment, etc.). Since there are no definitive criteria by which to select the subset of conditions, we chose a range of speeds and slopes that we felt was reasonable, practical and relevant based on the recreational runners we planned to test (**Figure 3.2a**). In an effort to maximize generalizability, each runner performed the same set of conditions. At slower speeds (2.6-3.0 m/s), we swept across the broadest range of slopes, from -9 to +9 degrees (**Figure 3.2a**). The highest speeds (3.4-4.0 m/s) were only performed on level ground to ensure all runners could complete the same conditions, to help limit the total number of conditions and to mitigate confounds due to fatigue (which we considered an interesting but separate investigation). Subjects wore their own personal running shoes. Each condition was performed on a treadmill for at least 30 seconds; ~10 seconds to adjust to the speed and slope, then data were recorded for 20 seconds. Breaks were taken between trials to adjust the treadmill slope, or if the subject requested a break for any reason.

a)

		treadmill slope (degrees)						
		-9	-6	-3	0	3	6	9
speed (m/s)	2.6	x	x	x	x	x	x	x
	2.8	x	x	x	x	x	x	x
	3.0	x	x	x	x	x	x	x
	3.2		x	x	x	x	x	
	3.4				x			
	3.6				x			
	3.8				x			
	4.0				x			

b)

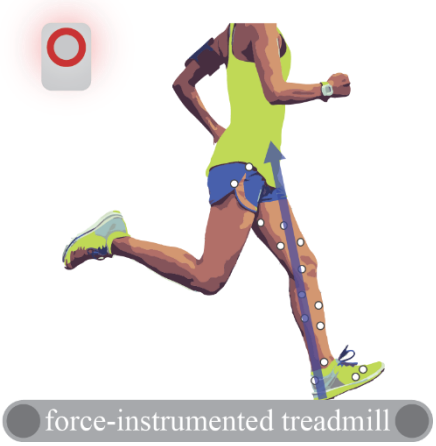


Figure 3.2 Summary of methods. (a) Each subject performed 30 running trials at a combination of speeds and slopes. **(b)** Experimental protocol involved subjects running on a force-instrumented treadmill while GRFs (blue vector) and lower-limb kinematics were recorded (white circles represent motion capture markers).

We collected lower-limb kinematics and GRFs (**Figure 3.2b**). Kinematics were collected at 100 Hz (Vicon), then low pass filtered at 10 Hz (3rd order, zero-lag Butterworth). Four markers were placed on each segment (thigh, shank, foot), 2 on the lateral and medial femoral epicondyles, and 2 on the

lateral and medial malleoli. Functional joint centers, segment angles, and joint moments were computed using C-Motion Visual3D software. The GRFs under each foot were collected at 1000 Hz using a force-instrumented treadmill (Bertec). The GRFs were low-pass filtered at 15 Hz (3rd order, zero-lag Butterworth) prior to computing inverse dynamics, similar to [114]. However, to avoid smoothing out GRF impact transients, these data were low-pass filtered at 45 Hz (3rd order, zero-lag Butterworth) for extracting GRF metrics, similar to [115]. For each trial, individual stance phases were parsed out, outcome metrics (as detailed below) were computed on a step-by-step basis, and then averaged.

We computed four vertical GRF metrics that are commonly reported in the running literature, with vertical defined with respect to the absolute lab reference frame (i.e., parallel to the gravity vector): $F_{vgrf,active}$ (vertical GRF active peak), $F_{vgrf,impact}$ (vertical GRF impact peak), VALR (vertical GRF average loading rate) and J_{vgrf} (vertical GRF impulse) (**Figure 3.3a**). $F_{vgrf,active}$ was defined as the maximum vertical GRF during 40-60% stance. $F_{vgrf,impact}$ was defined as the local maximum peak of vertical GRF between foot contact and $F_{vgrf,active}$. Foot contact was defined as when vertical GRF increased above 20 N. If an impact peak was absent in more than half of the gait cycles for a trial, then average $F_{vgrf,impact}$ was not calculated for that running condition. The number of running conditions for which a subject did display an impact peak in more than half the gait cycles was also recorded. The VALR was estimated as the change in vertical GRF for the first 25 ms after reaching a threshold of 50 N, a method that does not rely on the presence of an impact peak [116], [117]. J_{vgrf} was calculated as the time integral of the vertical GRF over stance.

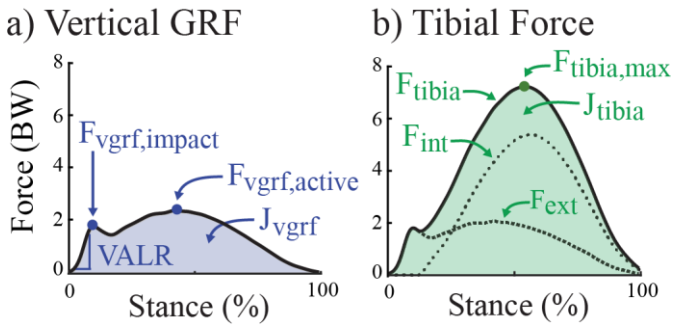


Figure 3.3 Outcome metrics. (a) Four commonly-used vertical GRF metrics: $F_{vgrf,impact}$: impact peak; VALR: vertical average loading rate; $F_{vgrf,active}$: active peak; J_{vgrf}: total vertical impulse. **(b)** Two tibial bone force metrics: $F_{tibia,max}$: maximum tibial compression force; J_{tibia}: tibial compression force impulse. Two additional force estimates are shown for reference: F_{ext}: the contribution of the external GRF to tibial compression; F_{int}: the contribution of internal muscle force to tibial compression.

Tibial bone load over the stance phase of gait was calculated as the longitudinally compressive force on the distal end of the tibia, a common location for stress fractures in runners [118]. The total force on the ankle (i.e., distal tibia) was calculated using a lower limb model, similar to prior studies (e.g., [13]): by summing the net force on the ankle (F_{ext}) plus an estimate of force from the calf muscles generating torque about the ankle (F_{int}). Ankle force was assumed to be indicative of tibial bone loading [119]. Net force on the ankle was estimated as the 3D GRF projected onto the long axis of the tibia, estimated as the vector connecting the ankle joint to the knee joint. In this calculation, foot mass and inertia were assumed to be negligible to avoid underestimating contributions from the GRF due to measurement errors in modeling or tracking foot segment motion; though net ankle force estimates using an anthropometric foot mass are very similar (typically within ~0.1 body weight based on an

informal sensitivity analysis of our own data). Calf muscle force contribution was estimated as the sagittal plane ankle moment divided by the Achilles tendon moment arm, assumed constant (5 cm, [1], [120]). We used this simplified model because it has been previously shown to yield Achilles force estimates during running that were similar to Achilles force estimates from a musculoskeletal model using 300 muscles with static optimization that minimized sum of muscle forces (e.g., peak tendon forces within ~4%, [121]). Studies that included other muscle groups when estimating tibial compression found a small tibial compressive force contribution from dorsiflexors (<0.5 body weight, [14], [108]), but this force only existed for 0-20% and 90-100% stance, and a small compressive force contribution from other plantarflexors (<0.35 body weight, [14]). The lower limb model used has produced results that are qualitatively consistent with *in vivo* tibial strain measurements [11] and direct tibial bone load measurements in cadavers [107].

We then computed two tibial bone load summary metrics: $F_{tibia,max}$ (maximum tibial compression force), J_{tibia} (tibial compression force impulse) (**Figure 3.3b**). $F_{tibia,max}$ was defined as the peak tibial bone force over stance phase. J_{tibia} was calculated as the time integral of the tibial bone force over stance. These metrics were selected because of their relevance to the load intensity (magnitude and time duration of loading): maximum force magnitude is relevant to cyclic fatigue and force impulse is relevant to creep damage accumulation or cumulative load over time or distance [90], [122]–[128].

The Pearson correlation coefficient (r) was computed for each GRF metric versus each tibial force metric on a subject-by-subject basis across all running conditions. The inter-subject range of correlation coefficients was identified and average correlation coefficients across subjects were computed using Fisher's z transformation [129]. Force data were normalized by subject body weight for reporting purposes. A *strong positive* correlation was defined here as $r \geq 0.8$, *moderate positive* correlation as $0.5 \leq r < 0.8$, *weak positive* correlation as $0.3 \leq r < 0.5$, *negligible* correlation as $-0.3 < r < 0.3$, *weak negative* correlation as $-0.5 < r \leq -0.3$, *moderate negative* correlation as $-0.8 < r \leq -0.5$, and a *strong negative* correlation as $r \leq -0.8$.

3.4 Results

On average, none of the GRF metrics were strongly correlated to tibial force metrics (**Table 3.1**, **Figure 3.4**); nor were there any GRF metrics for which the majority of subjects exhibited strong positive correlations with either of the two tibial load metrics.

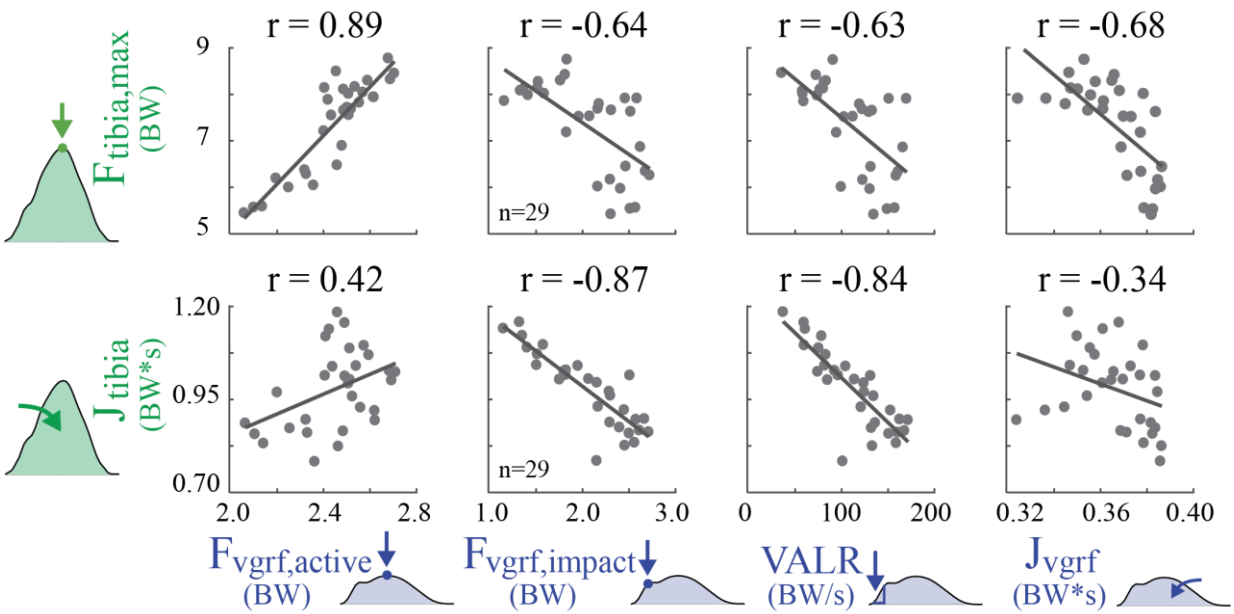


Figure 3.4 Regression results for GRF metrics vs. tibial bone load metrics across 30 running trials. Results **for a single subject (Subject 1)** Each gray dot represents a single condition (i.e., a given speed and slope from **Figure 3.2**), and *n* indicates number of running conditions that exhibited a measurable GRF impact peak for Subject 1. The correlation coefficient (r) was computed for Subject 1 across all running speeds and slopes. Note that no single subject should be considered representative given the large inter-subject variability observed. For instance, the one strong correlation shown for this subject (r = 0.89) was as low as r = 0.16 for another subject. The correlation coefficients for each individual subject are reported in **Table 3.1**.

Table 3.1 Correlation coefficients (r) between GRF metrics and tibial bone load metrics Results *across all trials within a subject*.Ten rows represent the 10 subjects (F=female, M=male). Within a subject, (n) indicates the number of running conditions (of 30 total conditions) that exhibited a measurable GRF impact peak (i.e., evident in more than half the gait cycles). Mean and standard deviation (std) were computed using Fisher's z transformation.

Subject	$F_{vgrf,active}$	$F_{vgrf,impact}$	VALR	J_{vgrf}	$F_{vgrf,active}$	$F_{vgrf,impact}$	VALR	J_{vgrf}
	$F_{tibia,max}$				J_{tibia}			
1 (F)	0.89	-0.64, n=29	-0.63	-0.68	0.42	-0.87, n=29	-0.84	-0.34
2 (F)	0.84	-0.47, n=27	0.23	-0.02	0.58	-0.78, n=27	-0.10	0.12
3 (F)	0.72	-0.06, n=10	0.01	-0.36	-0.17	-0.62, n=10	-0.80	-0.17
4 (M)	0.90	0.27, n=19	0.13	-0.20	0.60	-0.14, n=19	-0.27	-0.20
5 (F)	0.72	-0.33, n=24	-0.30	-0.66	0.13	-0.87, n=24	-0.80	-0.48
6 (F)	0.58	-0.45, n=19	0.13	-0.49	-0.60	0.30, n=19	-0.91	0.07
7 (M)	0.85	-0.24, n=22	-0.44	-0.68	0.47	-0.78, n=22	-0.81	-0.53
8 (M)	0.26	0.34, n=15	-0.24	-0.36	-0.54	-0.13, n=15	-0.72	-0.07
9 (M)	0.16	-0.46, n=16	-0.65	-0.84	-0.65	-0.10, n=16	-0.80	-0.34
10 (M)	0.63	-0.64, n=17	0.00	0.19	0.09	-0.14, n=17	-0.62	0.74
mean±std	0.72 ± 0.42	-0.29 ± 0.37	-0.20 ± 0.35	-0.46 ± 0.40	0.03 ± 0.51	-0.51 ± 0.53	-0.72 ± 0.41	-0.11 ± 0.41
[min max]	[0.16 0.90]	[-0.64 0.34]	[-0.65 0.23]	[-0.84 0.19]	[-0.65 0.60]	[-0.87 0.30]	[-0.91 -0.10]	[-0.53 0.74]

3.4.1 Active Peak

$F_{vgrf,active}$ was positively or negligibly correlated with $F_{tibia,max}$ in all subjects (0.72 ± 0.42): two exhibited negligible correlations, four individuals exhibited moderate correlations, and four exhibited strong correlations. $F_{vgrf,active}$ had an inconsistent relationship with J_{tibia} (0.03 ± 0.51): four subjects showed a moderate positive correlation, three showed a moderate negative correlation, and three showed negligible correlation.

3.4.2 Impact Peak

$F_{vgrf,impact}$ was on average negatively correlated with $F_{tibia,max}$ (-0.29 ± 0.37): three subjects exhibited a negligible correlation, four exhibited a weak negative correlation, one exhibited a weak positive correlation, and two exhibited moderate negative correlations. $F_{vgrf,impact}$ was on average negatively correlated with J_{tibia} (-0.51 ± 0.53): four subjects showed a negligible correlation, one showed a positive moderate correlation, three showed a moderate negative correlation, and two showed a strong negative correlation. On average, measurable GRF impact peaks were only observed for 20 ± 6 conditions for each subject. The majority of subjects were rearfoot strikers and had measurable GRF impact peaks during level and decline running; however, most individuals changed their footstrike pattern on more inclined slopes and the impact peaks tended to disappear. Across all subjects, measurable GRF impact peaks were present during 71 of 80 level running trials and during 98 of 110 decline trials, but only during 29 of 110 incline trials.

3.4.3 Loading Rate

The correlation between VALR and $F_{tibia,max}$ was generally weak or negligible, but varied considerably between subjects (-0.20 ± 0.35): six subjects exhibited a negligible correlation, two exhibited a weak negative correlation, and two exhibited a moderate negative correlation. The VALR and J_{tibia} were negatively or negligibly correlated in all subjects (-0.72 ± 0.41): two subjects showed a negligible correlation, two showed a moderate negative correlation, and six showed a strong negative correlation.

3.5 Impulse

The correlation between J_{vgrf} and $F_{tibia,max}$ varied across subjects (-0.46 ± 0.40): three subjects exhibited a negligible correlation, two a weak negative correlation, four a moderate negative correlation, and one a strong negative correlation. The correlation between J_{vgrf} and J_{tibia} also varied across subjects (-0.11 ± 0.41): five subjects showed a negligible correlation, four showed a moderate negative correlation and one showed a moderate positive correlation.

3.6 Discussion

We found that increases in GRF metrics were not strongly correlated with increases in tibial bone loading metrics during running across speeds and slopes (**Table 3.1**); rather most correlations were

negligible, weak, or moderate. Although there was high inter-subject variability in the strength of correlation, 76 of the 80 subject-specific correlation coefficients supported our hypothesis. The only strong positive correlations were in 4 of 10 subjects between $F_{tibia,max}$ and $F_{vgrf,active}$. Also of note, 2 subjects showed a strong *negative* correlation between J_{tibia} and $F_{vgrf,impact}$, and 6 showed a strong *negative* correlation between J_{tibia} and VALR.

Isolating effects due to speed and slope provides some insights into why GRF metrics were not strongly correlated with tibial loading (**Figure 3.5**). When runners encountered a change in ground slope, relationships between GRF and tibial loading often changed drastically. For instance, $F_{vgrf,impact}$ and VALR were often positively correlated as speed increased on level ground, but typically switched to having a negative correlation across a range of slopes when speed was held constant (**Figure 3.5**). $F_{vgrf,impact}$ and VALR metrics both decreased with increasing ground slope (similar to [130]), however $F_{tibia,max}$ increased due to higher muscle forces. Likewise, J_{vgrf} and J_{tibia} were positively correlated as speed increased on level ground, but were negatively correlated when slope changed at a single fixed speed.

These findings highlight that there are only limited special cases when GRF metrics are strong indicators of tibial bone load. When ground slope was held constant at zero degrees (level), then as speed increased all subjects showed a strong positive correlation between $F_{vgrf,active}$ and $F_{tibia,max}$ (r=0.97, inter-subject mean computed in post-hoc analysis). Further, $F_{tibia,max}$ was also moderately to strongly correlated to $F_{vgrf,impact}$ (r=0.90) and VALR (r=0.90) in all subjects during running on the zero-degree slope. However, even at this fixed slope, some runners exhibited weak relationships between GRF metrics and J_{tibia}. When looking at level and decline running conditions together, the correlation between $F_{tibia,max}$ and $F_{vgrf,impact}$ (r=-0.19) and between $F_{tibia,max}$ and VALR (r=0.45) became negligible and weak; though $F_{vgrf,active}$ and $F_{tibia,max}$ remained strongly correlated (r=0.94). Once level, decline, and incline conditions were all analyzed together the correlation between $F_{vgrf,active}$ and $F_{tibia,max}$ became moderate (r=0.72). Note that the special cases when GRF metrics were strongly correlated to tibial bone load are not necessarily well-suited for real-world outdoor running, in which a runner who encounters a decline generally also encounters an incline if they aim to start and end at the same location.

From this very simple set of running conditions (i.e., varying only speed and slope), it is evident why increases in GRF metrics generally should not be assumed to be a surrogate for, or indicator of, increases in tibial bone loading. During real-world training, additional confounds such as changes in muscle coordination or running pattern (e.g., due to fatigue, soreness, terrain, shoe properties, footstrike pattern, etc.), may further alter the relationship between GRF and tibial bone load.

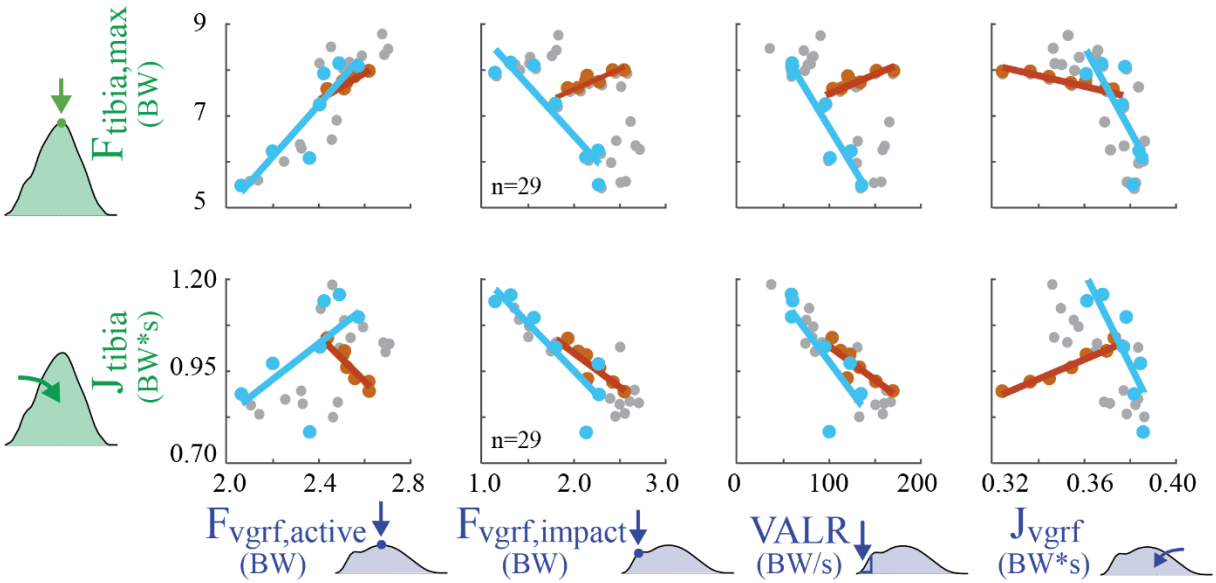


Figure 3.5 Force trends due to changing speed vs. changing slope. Results for a single subject (Subject 1). Lines represent regression results for GRF metrics vs. tibial bone load metrics when only speed or slope was varied. Dark orange dots represent conditions when speed was varied while running on a fixed slope (level ground). Light blue dots represent conditions when ground slope was varied while speed is held constant (at 2.6 m/s). Small gray dots are all remaining parameter sweep conditions.

3.6.1 Key implications & and discussion related to scientific research

The lack of strong correlations in this study suggest that GRFs provide limited insight into tibial bone loading during running across speeds and slopes, and therefore may provide limited utility for understanding or predicting overuse injury risk associated with this repetitive internal structure loading. Running GRFs have been analyzed in many ways in the scientific literature over the last several decades – extracting impact peaks, loading rates, active peaks, and impulses from vertical GRF, analyzing peaks from fore-aft GRFs, quantifying GRF frequency content, etc. – in the hopes that some feature of this force between the foot and ground might indicate injury risk, such as tibial stress fracture risk. However, GRFs provide an incomplete perspective on musculoskeletal loading and may be the wrong signal to be monitoring/analyzing if we seek to understand tibial bone load or assess overuse injury risk. This is of broad concern to the research field given the substantial time and resources invested into gait analysis and epidemiological studies that seek to relate bone stress injury to GRF metrics (or GRF correlates, e.g., based on accelerometers or pressure insoles), and given the pervasive influence these GRF metrics have had on sport science (e.g., comparisons of footwear or running patterns), training, coaching and product development (e.g., running shoes).

Results from this study reinforce previous experimental evidence and theoretical arguments against using or interpreting GRF impacts and loading rates to identify overuse injury risks [111], [131]–[134]. Although some epidemiological studies have observed an association between high impacts or loading rates and running-related injuries [116], [135]–[138], there are many studies that have failed to find such a correlation [113], [139]–[144]. These conflicting results also exist between prospective studies (e.g., [135], [136], [143]–[146]) . Several textbooks, review articles and commentaries further highlight this conflicting evidence [100], [101], [111], [134], [147], [148]. Nevertheless, the use of higher GRF

impact peaks and loading rates to infer injury risk remains extremely common in the scientific literature. Based on a literature search of articles published between 2015-2017, we discovered that during this period more than 50 peer-reviewed publications per year assume, report or interpret GRF impact peaks or loading rates to signify increased injury risk.

The reason for the sustained popularity of GRF metrics is likely multifaceted, but may stem partly from measurement convenience, or from people's intuition based on how we as humans perceive external (vs. internal) loads on the body (though such intuitions can often be misleading in biomechanics, e.g., [106]), or simply from the GRF impact paradigm being deeply embedded in the running literature over recent decades. Another contributing factor may be related to misinterpretation and misapplication of prior bone mechanics studies [149], [150]. Studies on rabbits and guinea pigs have shown that repeated impulsive loading can cause bone microdamage [151], [152, p. 197]. However, this finding has been applied in running-related injury studies to support a subtly but significantly different contention: that higher impact peaks (impulsive loading) are associated with higher injury risk (e.g., due to more microdamage accumulation) [115], [153], [154]. The critical thing to highlight is that this contention is only valid if we also assume that during running the impact loading is *the primary cause* of bone microdamage. To our knowledge, there is no evidence to support this assumption. During running, the majority of the tibial force is due to muscle contraction (**Figure 3.1a**). Note that the commonly-cited impulsive loading studies [151], [152, p. 197] did not compare damage due to impact forces relative to damage due to muscle forces, and therefore provide no experimental evidence that GRF impacts are the primary cause of bone microdamage. The relationship between the magnitude of force on a biological structure and the microdamage caused by the force is non-linear [90], [155]. According to Miner's rule, microdamage is roughly proportional to force to the *C* exponent (i.e., force ^{C}) [9], where *C* is a bone-specific constant found experimentally via mechanical fatigue studies. We can apply this well-established relationship to estimate the relative amount of damage caused by forces at different parts of the gait cycle (e.g., impact vs. midstance). As depicted in **Figure 3.1**, the GRF impact peak is ~2 body weights, whereas the peak tibial force in midstance is ~8 body weights (which includes muscle contraction forces). Using the empirically-derived exponent for cortical bone of *C*=7 [122], [125], [133], we estimate that on every step the peak tibial force at midstance would be expected to cause about sixteen thousand ($8^7/2^7 \approx 16{,}000$) times more microdamage to bone than the GRF impact peak. Thus, the relative damage due to the impact forces may be trivially small regardless of whether impacts are slightly larger or smaller (e.g., 1.8 vs. 1.55 body weights, as reported in [156]). This evidence contradicts the prevailing belief that impacts are the source of overuse injuries, and further highlights why it generally should not be assumed that increases in GRF impact peak are reflective of increased injury risk.

A similar misinterpretation issue may underlie the use of loading rate (and strain rate) findings from bone mechanics studies. One commonly-cited study by Schaffler et al. [157] concluded that "cyclic loading at a higher physiological strain rate causes more damage than cyclic loading at a lower strain rate." However, upon careful inspection of the results we discovered that this conclusion was not substantiated by the statistical analysis performed. The study compared the effects of high strain rate vs. an unloaded control, and the effects of low strain rate vs. an unloaded control; however, it did not directly compare the high vs. low strain rate groups. When we performed statistical analyses using the study results presented in the paper we failed to find significant differences between high vs. low strain rate groups for any of the reported outcome metrics; namely, bone stiffness loss ($p=0.07$), number of microcracks ($p=0.32$), density of microcracks ($p=0.28$) and length of microcracks ($p=0.48$). We compared bone stiffness loss using the Mann Whitney U test (the primary statistical test employed

in the original study) since a table of specimen-specific results were provided. For the remaining metrics we performed a two-sample t-test using the mean ± standard error results reported in the publication (since specimen-specific results were not provided to perform a Mann Whitney U test). Inspection of the estimated 95% confidence interval for the difference in means of each outcome metric further substantiated our take-away that from these published data one cannot conclude that the higher loading rate caused more damage than the lower loading rate. Significance level of 0.05 was used for our interpretation, consistent with the threshold set in the original publication. Meanwhile, a more recent cyclic loading study on bone specimens found that high loading rates associated with GRF impacts in running had little effect on bone fatigue [133]. Likewise, bone samples loaded at higher rates have been observed to take more cycles to failure, suggesting less damage accumulation per cycle [90], [122], [128]. Collectively, this evidence seems to call into question the common assumption that higher loading rates indicate higher bone damage accumulation (or injury risk).

In summary, there are substantive concerns about how impulsive bone loading studies are commonly interpreted [149], [150] , and how this may misguide the use of GRF metrics like impact peak and loading rate. The field would benefit from a clear and careful synthesis of bone mechanics studies, to ensure this knowledgebase is being appropriately interpreted and applied in the assessment of overuse injury risks.

3.6.2 Key implications & and discussion related to wearable devices

While motion analysis and musculoskeletal modeling methods have allowed researchers in the lab to estimate forces on certain internal structures, recreating these estimates outside the lab with non-invasive, low-cost, and portable sensors remains a grand challenge in the biomechanics field. Most commercial devices use GRF-correlated metrics (e.g., tibial shock) from accelerometers and/or pressure insoles to provide loading or injury risk feedback to the user. A key underlying assumption of these devices is that increases in GRF metrics reflect increases in loading inside the body. For instance: (i) IMeasureU outputs a “bone load” metric that increases with “the size of the [ground] impact derived from each individual step,” based on the stated rationale that impact peaks “can function as a surrogate measure of the loads experienced by the underlying musculoskeletal tissue” [158], [159], (ii) Runscribe outputs “Impact Gs” and indicates that lower “Impact Gs” at footstrike may help prevent injuries [160], (iii) MileStone states that a “low rate of impact… is optimal and can help prevent injury” [161], (iv) Stridalyzer states that “ ‘Pounding’ [the foot] during landing… can increase impact forces, which over time leads to injuries” [162], and (v) Sensoria provides an "Impact Score,” stating it is “a quantitative relative measure, on a scale from 1 to 10, driven by the impact forces generated when your foot hits the ground while your run. In order to reduce likelihood of injury, you want to keep your impact score as low as possible [163]" However, the key assumption underlying each of these statements/claims remains unsubstantiated for running, both experimentally and from a theoretical standpoint. Tibial acceleration may be correlated with tibial load around impact (i.e., over the first ~40 ms after foot-ground contact [164]) when forces are relatively low (**Figure 3.1a**). However, as shown in this study, commonly-use GRF metrics (e.g., impact peak, loading rate) are not necessarily reflective of the much larger tibial bone forces experienced later in the gait cycle, nor is there a biomechanical rationale or consistent epidemiological evidence to support interpretations of these GRF metrics or correlates like tibial shock as indicators of injury risk.

Presently there is a lack of transparency and validation amongst consumer wearables [22]. Many

commercial devices selectively cite studies that support their chosen outcome metric, while omitting published counter evidence. Others simply fail to provide scientific evidence that their outcomes are indicative of increased loading or injury risk to specific internal structures. As a result, biofeedback from existing wearable devices may be misleading users. Wearable devices often employ ambiguous terminology such as "limb load", "step intensity" or "biomechanical load" – terms which do not specify which individual structure in the body, if any, experience the purported load. Some devices employ misleading terminology such as "bone load," which has been defined as a weighted sum of impact peaks [158] . As seen in this study, footstrike impact peaks are not the main source of bone loading, and cannot be used as a surrogate to infer the peak force or force impulse experienced by the tibia. Some commonly-used metrics – like impact peaks and loading rates – may even be negatively correlated with bone loading during running (**Figure 3.4**, **Table 3.1**). This means the current interpretation of these values in wearable devices may be leading to the wrong conclusions about the accumulation of microdamage to a bone such as the tibia. Further ambiguity is introduced when commercial device metrics refer or allude to overall injury risk (i.e., to structures throughout the body). The idea of having a single or small number of output metrics that can capture overall injury risk is appealing to consumers, clinicians, researchers and wearable device manufacturers alike. However, there is no guarantee, nor theoretical basis, that this global injury metric is embedded within the GRF waveform (i.e., hidden within this relatively small force magnitude between the foot and ground).

The wearable device field would benefit from more deliberate and targeted attempts to monitor loading on specific internal structures at high risk of injury, with less emphasis on GRF metrics. There have been a number of innovative advances in sensing that provide estimates of loading on a given muscle or internal structure (e.g., [165]). These more targeted approaches offer the opportunity to better understand structure-specific loading, such that we can more confidently associate forces on a given structure (bone, muscle, tendon, etc.) with overuse injuries that may eventually develop in said structure. Moreover, estimating the time-varying force experienced by specific structures (as opposed to only computing discrete summary metrics related to peaks or impulses) may offer a more promising avenue of identifying and understanding specific injury risks in running and other activities. Given the complexity of human movement, and difficulty of measuring internal forces non-invasively, data from multiple wearable sensors may need to be fused in order to monitor load on certain internal structures *in situ*.

3.6.3 Potential utility of GRF metrics

There may be situations when GRFs still provide some utility for understanding bone loading or overuse injuries. For instance, a given GRF metric might be useful in situations when the metric has been validated to be a strong indicator of loading on a specific internal structure (e.g., tibial bone) for a given individual (or subset of individuals) and for a given subset of activities (e.g., running over a specified range of speeds, slopes, etc.). However, at present, the majority of published studies seem to use and interpret GRF metrics without adequate subject- or activity-specific validation.

One potential use of GRF metrics might be for lab-based experiments performed only on level ground (or potentially a flat track, though we did not assess curvilinear running in this study), given that we found strong correlations with $F_{tibia,max}$ for all 10 subjects (as detailed earlier in Discussion). However, in post-processing of our data we found that running speed itself was also strongly correlated with $F_{tibia,max}$ on level ground (r=0.97, similar to the correlation between $F_{vgrf,active}$ and $F_{tibia,max}$). In many cases

it may be preferable and easier to monitor speed than GRF.

A second potential use of GRF metrics might be for an individual runner who is studied extensively to establish that over a specified set of running conditions (speeds, slopes, terrains, levels of fatigue, etc.), a given GRF metric is a good indicator of load on a specific bone. For example, two subjects in our study (Subjects 1 and 7) showed a strong positive correlation between $F_{vgrf,active}$ and $F_{tibia,max}$ and a strong *negative* correlation between VALR and J_{tibia} (**Table 3.1**). For these two subjects, increases in $F_{vgrf,active}$ strongly indicated *increases* in $F_{tibia,max}$, and increases in VALR strongly indicated *decreases* in J_{tibia} over the range of speeds and slopes tested. These two subjects demonstrate a scenario when specific GRF metrics might be utilized to indicate bone load, but only when (i) the relationship was validated for that subject and that subset of conditions, and (ii) the relationship was identified as being strongly positive or strongly negative for a given GRF metric. Subject-specific validation may be feasible for elite runners, but would likely be impractical or prohibitively expensive for many recreational runners due to the amount of instrumented gait analysis required to perform this validation.

A third potential use of GRF metrics would be for testing hypotheses in which there is an indirect relationship between GRFs and loading on specific internal structures. Here we summarize one example: Studies suggest that larger GRF impacts cause more energy to be dissipated through wobbling of muscles in the legs [166]. To maintain a given running speed, when energy is dissipated then it must be offset by positive work performed through active muscle contraction. This additional muscle work could be achieved by either higher peak muscle forces and/or applying muscle forces over longer periods of time. It therefore might be hypothesized that increases in GRF impacts lead to more energy dissipation, which then leads to increased loading magnitude or duration of certain muscles or bones. Note that in this example the GRF impact peak is not the source of high internal structure loading, as commonly assumed. Rather, large GRF impacts might conceivably help explain a mechanism by which higher internal structure loading could result (potentially at a different time in the stride cycle). As such, GRF metrics could be useful in situations when they are a core part of a testable hypothesis.

One additional consideration worth noting is that impact peaks were not present in about one third of the running conditions in our study (mostly on inclines), which may limit the utility of this metric in comparing across a broad range of conditions. Relatedly, estimating GRF loading rate from running strides with vs. without impact peaks may be capturing slightly different aspects of the gait dynamics [167]. This may also limit practical applications and interpretations of this metric across different running conditions, particularly those in which footstrike pattern changes. Finally, we remind that this discussion is in relation to overuse injury risk, specifically tibial stress fracture risk. There are of course many other situations and applications when GRFs are extremely useful (e.g., computing inverse dynamics), and more broadly, GRFs remain one of the most important measurements in the field of biomechanics.

3.6.4 Limitations

The scope of this study was limited to estimating changes in bone load within each subject. As discussed in the introduction, additional factors also affect stress fracture risk. To assess injury risk between subjects, additional subject-specific information on bone remodeling and intrinsic factors

(age, gender, bone density, geometry, nutrition, mineral content, etc.) may also be necessary. A limitation of this study (and nearly all gait analysis studies) is that we are unable to directly measure tibial bone loading. However, there is strong converging evidence from cadaver [107], implanted sensor [10], [11] and musculoskeletal modeling studies [14], [108] that the non-invasive estimates used provide a reasonable approximation of tibial bone loading. In particular, we have confidence in the trends predicted with changing speed and slope, and would not expect imperfect bone load estimates to alter any of the major conclusions or interpretations. Another limitation is that we only performed linear, univariate regression analysis because this appears to be the most common way that prior scientific studies and current wearable devices are using GRF metrics (or correlates) to infer musculoskeletal loads or injury risks. Another limitation is that we only quantified tibial compression load. In future studies, it would also be interesting to estimate bending, shear or torsional loads [13], [14], [168] which contribute to the stress/strain of the tibia, or to use advanced modeling techniques to estimate local stress/strain concentrations [90], [124], [168], [169]. These loading patterns are all highly influenced by muscle forces; therefore, GRF metrics should not be assumed to reflect these loads either, unless validated for a given subset of activities. Because there are currently no wearable devices that can track tibial bone load (or localized stress) longitudinally in daily life it is not yet known (e.g., from prospective studies) which of these bone loading directions or metrics might be most useful. Given the large magnitude of the tibial compression load this seems like a reasonable candidate to explore, and more informative of bone loading than GRFs. A final limitation is that in order to make this study and analysis tractable we focused on a single bone and a single overuse injury. The specific conclusions drawn are all related to whether changes in GRF metrics are reflective of changes in loading on this particular bone. However, stress fracture and overuse injuries commonly occur in several other internal structures in the lower limb as well (e.g., calcaneus bone, Achilles tendon). The broader implication of our study is that loading on these other bones, muscles and tendons in the body -- perhaps even loading on the vast majority of structures in the body -- may be poorly understood by monitoring changes in GRF or GRF correlates alone.

3.7 Conclusion

In summary, increases in GRF metrics should not be assumed to correlate with increases in bone loading, nor assumed to signify increased risk for tibial stress fractures. The high inter-subject variability in correlations further strengthens this general conclusion. For any individual, both subject- and task-specific validation would be needed to assess if GRFs provide useful insight on loading of the tibial bone, or other internal structures. This study has important implications for scientific research on running-related injuries, and for the development and validation of current and future wearable devices. Specifically, these findings demonstrate that the way GRF metrics are commonly interpreted as indicators of musculoskeletal loading and injury risk in literature is often flawed, and the application of these GRF metrics in scientific research, sports and wearable devices can be highly misleading. Although GRF metrics may be convenient to measure and may seem to intuitively "make sense" as a way to monitor loading on the musculoskeletal system, the results here and in prior literature reveal that commonly-used GRF metrics may provide very limited insight on, and may even mislead our understanding of, internal structure loading such as to the tibial bone. The GRF may simply be the wrong signal to be monitoring/analyzing if we seek to understand or predict overuse injury risk to the tibia, or to other bones, joints, muscles and tendons in the body. Summarized poetically:

You go for a run down the street.

You feel the ground force on your feet.
  You may think these reveal
  The bone loads that you'll feel,
But this thinking is just incomplete.

The force due to ground reaction
May be a stress fracture distraction.
  Don't assume force on shoe
  To mean tibia load too
Since bone load's mostly from muscle contraction.

3.8 Abbreviations

GRF – ground reaction force
VALR – vertical average loading rate

4 Combining wearable sensor signals, machine learning and biomechanics to estimate tibial bone force and damage during running

4.1 Abstract

There are tremendous opportunities to advance science, clinical care, sports performance, and societal health if we are able to develop tools for monitoring musculoskeletal loading (e.g., forces on bones or muscles) outside the lab. While wearable sensors enable non-invasive monitoring of human movement in applied situations, current commercial wearables do not estimate tissue-level loading on structures inside the body. Here we explore the feasibility of using wearable sensors to estimate tibial bone force during running. First, we used lab-based data and musculoskeletal modeling to estimate tibial force for ten participants running across a range of speeds and slopes. Next, we converted lab-based data to signals feasibly measured with wearables (inertial measurement units on the foot and shank, and pressure-sensing insoles) and used these data to develop two multi-sensor algorithms for estimating peak tibial force: one physics-based and one machine learning. Additionally, to reflect current running wearables that utilize running impact metrics to infer musculoskeletal loading or injury risk, we estimated tibial force using a commonly measured impact metric, the ground reaction force vertical average loading rate (VALR). Using VALR to estimate peak tibial force resulted in a mean absolute percent error of 9.9%, which was no more accurate than a theoretical step counter that assumed the same peak force for every running stride. Our physics-based algorithm reduced error to 5.2%, and our machine learning algorithm reduced error to 2.6%. Further, to gain insights into how force estimation accuracy relates to overuse injury risk, we computed bone damage expected due to a given loading cycle. We found that modest errors in tibial force translated into large errors in bone damage estimates. For example, a 9.9% error in tibial force using VALR translated into 104% error in estimated bone damage. Encouragingly, the physics-based and machine learning algorithms reduced damage errors to 41% and 18%, respectively. This study highlights the exciting potential to combine wearables, musculoskeletal biomechanics and machine learning to develop more accurate tools for monitoring musculoskeletal loading in applied situations.

4.2 Introduction

Monitoring loads on musculoskeletal structures inside the body is challenging in the laboratory, and remains an unsolved grand challenge in applied situations outside the lab. Inside the lab, musculoskeletal loads (e.g., forces, moments, stresses, strains) are typically estimated in one of two ways: (i) using implantable or percutaneous instrumentation to directly monitor tissue stress or strain (e.g., [10], [11]) or (ii) using non-invasive instrumentation such as motion capture and force plates along with musculoskeletal modeling to indirectly estimate tissue loads (e.g., [13], [14], [108]). While these two approaches have fueled decades of musculoskeletal biomechanics research, they can be impractically invasive and/or require sophisticated instrumentation and expertise. Thus, these approaches are typically limited to small sample sizes and infrequent data collections, which severely restricts collection of movement data in applied situations (i.e., real-world or daily activities).

There are many practical applications if non-invasive monitoring of musculoskeletal loads was possible in applied situations [19], [22]–[24], [170]–[172]. Musculoskeletal load monitoring in combination with other information could provide transformative new health monitoring tools. For instance, an injury

risk monitoring tool that monitors musculoskeletal loading in combination with other risk factors could help athletes coordinate training and rest days accordingly. Similarly, a clinical tool that monitors musculoskeletal loading in combination with other patient specific physiological measures (e.g., tissue quality, diet, menstrual cycles, sleep/rest information) could help researchers investigating overuse injury etiology. A biofeedback tool that combines musculoskeletal monitoring and specified thresholds or targets could help clinicians, trainers or ergonomists deliver objective biofeedback training to help ensure safe and efficient return to play or return to work after an injury. Finally, a solely musculoskeletal load monitoring tool could enable researchers, assistive device developers, or footwear/sports accessory designers to evaluate the effect of interventions on specific musculoskeletal loads outside the lab and over much longer time scales than is currently practical.

Small, portable, inexpensive wearable sensors along with accompanying software algorithms and user interfaces (together termed *wearables*) provide exciting opportunities for non-invasive monitoring of human movement in applied situations without the need for sophisticated lab equipment. Wearables often contain sensors such as inertial measurement units (IMUs), global positioning systems (GPS), or pressure or strain sensors. These wearables typically track spatiotemporal metrics (such as cadence, speed, number of steps taken, or time spent being active), body segment motions (such as accelerations or orientations), or interaction forces between the person and the environment (such as insole pressures or ground reaction forces, GRF). These metrics, particularly GRF metrics, are then often used to evaluate injury risk to structures *inside* the body. However, GRF metrics are generally not indicative of loads experienced by musculoskeletal structures in the body (e.g., tibial force), nor is there consistent epidemiological evidence to support the use of GRF metrics as indicators of bone stress injury [173]. See Matijevich et al., 2019 and Vigotsky et al., 2019 for extended discussions on the widespread misuse of GRF metrics to infer internal tissue loading and associated injury risk in both scientific research and commercial wearables. To date there are only a very limited number of wearables (mostly non-commercial, in-development) that have demonstrated validity or reported accuracy in estimating the mechanical loading on specific structures *inside* the body (e.g., [165]).

There are numerous knowledge gaps surrounding how to more accurately estimate musculoskeletal loads with wearables. Many inter-related choices must be made in order to develop a new wearable, for instance: What type(s) of wearable sensor(s) to use? What algorithm(s) to employ to fuse multi-sensor data? Which musculoskeletal loading metrics to track, and what level of accuracy is desired? There are innumerable approaches for tackling the many intertwined and open-ended questions, making this an interesting opportunity for exploratory research that combines wearables, musculoskeletal biomechanics and machine learning. To manageably explore these choices, we targeted a specific musculoskeletal structure and movement: the tibial (shin) bone during running. We were motivated to monitor tibial force because tibial overuse injuries (e.g., bone stress fractures) are a common debilitating injury amongst runners, athletes and cadets [92]–[94], [96], [97], [175]; however, we believe the risk of injury can be reduced if tibial forces can be monitored in daily life. Further, there are established and validated methodologies for non-invasively estimating tibial compression force using lab-based instrumentation and modeling [13], [14], [108], giving us confidence in our target bone force metric.

The objective of this study was to develop multi-sensor algorithms for estimating tibial force using idealized wearable sensor signals, quantify the accuracy of each algorithm, and then compare the accuracy of these algorithms to approaches that rely on ground reaction force metrics or other single variable surrogates to try to gain insight on internal loading or injury risk. Here we summarize the

development of two promising multi-sensor algorithms (one physics-based and one using machine learning) for estimating peak tibial force across various running speeds and slopes. For reference, we also report two other algorithms (single variable linear regression and group mean), which were intended to reflect the current state-of-the-art for commercial wearables.

4.3 Methods

4.3.1 Data collection

In this study we re-analyzed an existing dataset from our prior publication [173], in which ten recreational runners performed 30 running conditions where for each condition the running speed and slope was varied. These data are publicly archived [173]. The dataset includes 5 males and 5 females (age: 24±2.5 years; height: 1.7±0.1 m; mass: 67±6 kg) all of whom reported running a minimum of 10 miles per week. All participants gave written informed consent to the original protocol, which was approved by the Institutional Review Board at Vanderbilt University. To briefly summarize methods, 30-s trials were collected across a range of speeds (2.6 to 4.0 m/s) and slopes (-9 to +9 degrees), while runners wore their own personal running shoes. Ground reaction forces were collected at 1000 Hz on a force-instrumented treadmill (Bertec), synchronously with unilateral lower-limb kinematics at 100 Hz (Vicon). For further details on the experimental protocol or rationale underlying the running conditions see [173].

4.3.2 Lab-based data analysis

Data were analyzed to estimate tibial compression force as detailed in [173], thus only briefly summarized below. Lower-body segmental kinematics were estimated based on optical motion capture data and rigid-body inverse kinematics. GRF and kinematics were combined via rigid-body inverse dynamics to estimate joint kinetics (C-Motion, Visual3D). The time-series compression force acting on the distal end of the tibia during the stance phase of running was then estimated using a simple musculoskeletal model [173]; by summing estimates of the net force on the ankle and the compression force from plantarflexor calf muscles (**Figure 4.1**). We chose peak tibial force of each running condition as our target metric because maximum force is a key factor that contributes to cyclic fatigue of bone and resulting damage accumulation (see [173] for additional discussion). Peak tibial force was estimated as the maximum force across stance on a step-by-step basis, and then averaged to compute the mean for each condition (F_{max}), reported in units of body weight (BW).

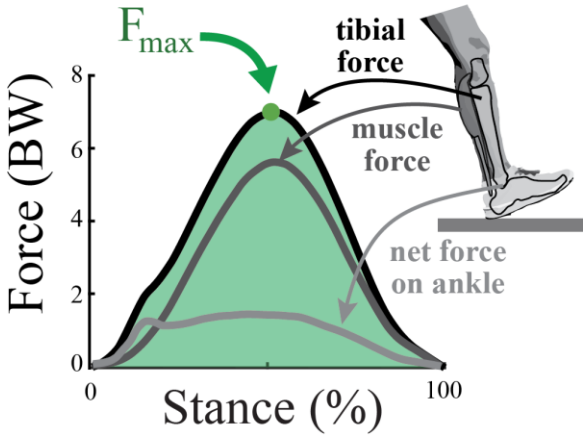


Figure 4.1 Tibial compression force was computed from motion capture and ground reaction force data using musculoskeletal modeling: by summing the net force on the ankle and the compression force from plantarflexor calf muscles. Peak tibial force (F_{max}) in bodyweights (BW) during stance was computed for each running condition. The lab-based estimate of F_{max} is the target load metric that we are interested in monitoring with wearable sensors.

4.3.3 Candidate wearable sensor signal identification and data preparation

Before developing and training algorithms to estimate our target load metric (peak tibial force, F_{max}) we first identified candidate wearable sensor signals to use as algorithm inputs. Candidate wearable sensor signals were selected because they could feasibly be collected with existing commercial wearable sensors (e.g., pressure-sensing insoles, IMUs) and we expected them to be important to the estimation of tibial force based on mechanics (e.g., signals typically used in inverse dynamics and musculoskeletal models of the lower limb). The four candidate sensor signals we selected were foot and shank orientations, vertical GRF and center of pressure (CoP, the point of application of the GRF vector under the foot).

We considered two options for using wearable sensor signals to develop tibial force estimation algorithms: (i) performing new experiments where we collect data from multiple wearable sensors synchronously with lab-based instrumentation, or (ii) converting lab-based data into the types of signals that could be feasibly obtained with existing commercial wearables (e.g., simplifying the three dimensional GRF vector into a one dimensional normal force, which represents the type of signal that can be estimated from a pressure-sensing insole). We chose the second option and termed these signals *idealized wearable signals*. Idealized wearable signals were beneficial for this feasibility study because they allowed us to not be limited by the quality or accuracy of existing wearables. In reality, sensor hardware, algorithms and signal quality are constantly improving; for example, sophisticated filtering has improved the quality of absolute angle estimates from IMUs [176]. Further, idealized wearable signals allowed us to explore a broader combination of sensor signals from our pre-existing dataset, and eliminated the need to select and integrate a specific sensor set and carry out new experiments before knowing which subset of sensors were actually needed.

In **Table 4.1,** we summarize how we converted lab-based data into idealized wearable signals. Algorithms used either time-series idealized wearable signals as inputs or discrete features extracted

from these signals as inputs. Discrete features were computed as follows: Vertical GRF features – vertical average loading rate (VALR), active peak, impulse – were extracted using established methodologies as detailed in [173]. Foot/shank minimum and maximum angles were the minimum and maximum sagittal plane angles during the stance phase of running. Foot/shank angles at midstance were the sagittal plane angles at the time of the vertical GRF active peak. CoP at midstance was the medial-lateral and anterior-posterior position of the CoP at the time of vertical GRF active peak. Stance time was the time from foot contact to toe off. Stride time was the time from foot contact to the next ipsilateral foot contact. All idealized wearable signals and discrete features were computed on a step-by-step basis, and then averaged to compute the mean for each running condition. Condition means were used in algorithm development and evaluation, and were computed from approximately 25-30 strides for each condition.

Table 4.1 Idealized wearable signals. Summary of our selected idealized wearable signals, existing wearable sensors that can estimate these types of signals, how lab-based data were converted into idealized wearable signals, and which discrete features were extracted from these signals.

Idealized wearable signal	**Wearable sensors** examples of existing sensor hardware that can be used to estimate each signal	**Data conversion** how lab-based signals were converted into idealized wearable signals	**Discrete features** features extracted from the idealized wearable signals
Shank angle (sagittal)	IMU on the shank	Inverse kinematics processed to output shank orientation in the lab frame	Min angle, max angle, angle at midstance
Foot angle (sagittal)	IMU on the foot/shoe	Inverse kinematics processed to output foot orientation in the lab frame	Min angle, max angle, angle at midstance
Vertical GRF	Pressure-sensing insole	3D GRF[†] transformed into foot's coordinate frame & projected into 1D GRF normal to the ground	VALR, active peak, impulse, stance time, stride time
CoP	Pressure-sensing insole	GRF and motion capture used to transform treadmill CoP into the foot's coordinate frame	Medial-lateral & anterior-posterior CoP at midstance
Speed	IMU or GPS	Used treadmill speed	Mean running speed
Slope	IMU	Used treadmill slope	Ground slope

[†]Prior to extracting GRF metrics, the data were low-pass filtered at 45 Hz (3rd order, zero-lag Butterworth) to avoid filtering out foot impact and loading rate dynamics.

4.3.4 Algorithm development

We developed two multi-sensor algorithms for more accurately estimating peak tibial force using idealized wearable sensor signals: one physics-based and one using machine learning. Additionally, we report two other algorithms: single variable linear regression and group mean. These were intended to reflect the current state-of-the-art for commercial wearables, which typically rely on VALR (or correlated metrics like impact peaks and lower limb peak accelerations) and/or step counters to provide musculoskeletal loading and injury risk feedback. See [173] for an extended discussion (and specific examples) on the widespread misuse of GRF metrics to signify bone loading and associated

injury risk in both scientific research and commercial wearables. For each algorithm, one or more of the idealized wearable signals or discrete features were used as inputs, and the output was an estimate of peak tibial force (in units of BW). To develop each algorithm, we used k-fold validation by runner, a commonly-used technique to assess the generalizability of an algorithm [170]. In other words, we used data from nine runners to train the model (i.e., optimize model parameters), and then tested the model on the remaining runner. This train-test process was repeated for all ten combinations. Summary statistics were then calculated on accuracy. Below we first overview the algorithms themselves, followed by a description of how algorithm estimates were evaluated vs. lab-based estimates of tibial force. An overview of the lab-based data analysis and algorithm evaluation is provided in **Figure 4.2**.

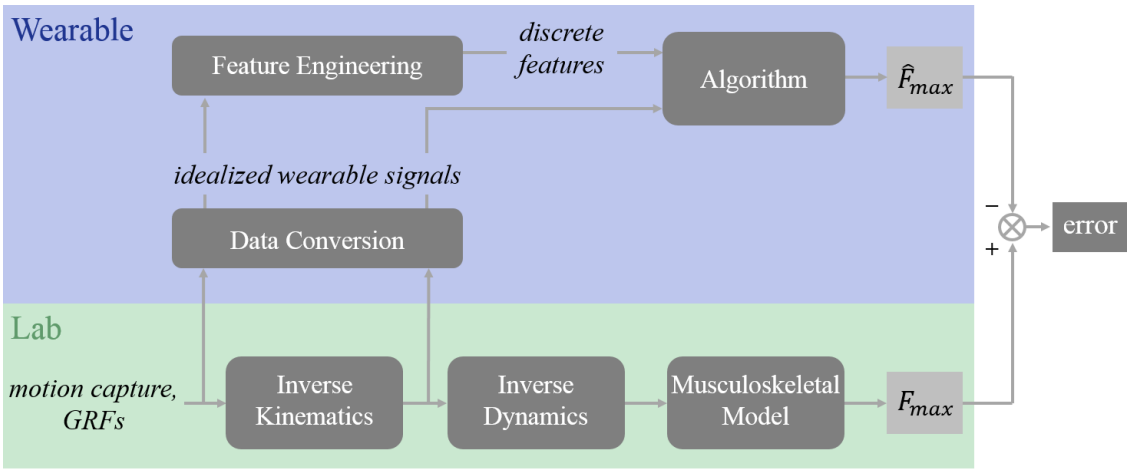


Figure 4.2 Lab-based data analysis and algorithm evaluation overview. (Lower, Lab) Lab-based data were collected (*motion capture* and *GRFs*) and standard inverse kinematics and inverse dynamics analyses were performed. A musculoskeletal model was used to calculate peak tibial force (F_{max}), which was treated as ground truth bone load for the purposes of this study. (Upper, Wearable) Lab-based signals were converted to *idealized wearable signals*, representing the types of signals that can be obtained with existing commercial wearable sensors. Algorithms used idealized wearable signals and *discrete features* that were extracted from these signals as inputs. Algorithms output estimates of peak tibial force ($\hat{F}_{max}$). Errors between the algorithm estimates and lab-based estimates of peak tibial force were computed.

4.3.4.1 Single variable linear regression

To reflect the current state-of-the-art in commercial wearables, we report an estimate of peak tibial force ($\hat{F}_{max}$) based on the common assumption/misconception that VALR can serve as a surrogate for tibial force and/or injury risk [100], [101]. VALR and lab-based estimates of peak tibial force (F_{max}) were used to find linear scaling coefficients by k-fold validation to estimate peak tibial force ($\hat{F}_{max}$). We selected VALR, in part, because unlike impact peak, it can be computed for every step regardless of the presence of a transient peak in the GRF [173]. Based on our own curiosity, we also generated additional single variable linear regression models using other commonly tracked metrics: GRF active peak, GRF impulse, running speed and ground slope.

4.3.4.2 Group mean approach

As another point of reference, we assumed that peak tibial force was constant for every step, regardless of speed or slope. Practically speaking, this approach represents what we would expect if we tried to use only a step counter (pedometer) to monitor bone injury risk. With this naïve modeling approach, we assumed that peak tibial force ($\hat{F}_{max}$) for all conditions for a given runner was equal to the mean peak tibial force across the remaining runners and conditions.

4.3.4.3 Physics-based algorithm

We implemented a physics-based algorithm (**Supplemental Material 4.7.1**) using four idealized wearable sensor signals (**Table 4.1**). This algorithm is similar to how we estimate musculoskeletal forces in the laboratory using inverse dynamics and biomechanical modeling. However, the key difference is that wearable sensors only provide limited signals (e.g., one dimensional normal GRF rather than three-dimensional GRF), which are insufficient to compute the full equations of motion. There are innumerable ways to deal with incomplete data, including simplifying the equations of motion themselves (e.g., removing terms that are expected to have small effects or terms that are impractical to collect with wearables), or using available data to estimate the missing terms (e.g., using regression equations to approximate the complete 3D GRF vector using data from multiple insole pressure sensors [177], [178]). Here, we used the former approach to provide a physics-based estimate of peak tibial force.

4.3.4.4 Machine learning algorithm

For the machine learning algorithm, we used 13 discrete features from our idealized wearable sensor signals to estimate peak tibial force (**Table 4.1,** discrete features). This set of discrete features was based on preliminary exploration of data – by analyzing time series signals, scatter plots and correlation values. Feature pruning was necessary to avoid overfitting the model with a limited dataset [170] and to speed up the hyperparameter search process. Features were normalized to z-scores prior to model training. We explored a variety of supervised machine learning techniques for multi-variable regression, including generalized linear models, ensemble methods, neural networks and support vector regression. Ultimately, we selected LASSO (Least Absolute Shrinkage and Selection Operator) regression, a least squares linear model with a L_1 regularization. This regularization penalizes the sum of the absolute values of the coefficients and forces a subset of learned coefficient weights to zero, effectively selecting a sparser subset of features. In addition to providing estimates of peak tibial force ($\hat{F}_{max}$), the resulting feature weights are used to rank the relative importance of each discrete feature.

4.3.5 Algorithm evaluation

To evaluate each algorithm, we computed the mean absolute percent error (MAPE) between F_{max} and $\hat{F}_{max}$ across all conditions for each runner, and report the inter-runner mean, standard deviation, and range. In some instances, we also computed root mean square error (RMSE) in units of BW for reference.

4.4 Results

MAPE results for the single variable linear regression and multi-sensor algorithms are shown in **Figure 4.3** and **Table 4.2**. Results in RMSE for reference: The single variable linear regression algorithm that used VALR to estimate peak tibial force ($\hat{F}_{max}$) resulted in a RMSE of 0.97 ± 0.32 BWs (**Figure 4.3A**, **Table 4.2**). The physics-based algorithm resulted in a RMSE of 0.48 ± 0.09 BWs (**Figure 4.3B**, **Table 4.2**). The multivariable machine learning algorithm resulted in a RMSE of 0.25 ± 0.07 BWs (**Figure 4.3C**, **Table 4.2**).

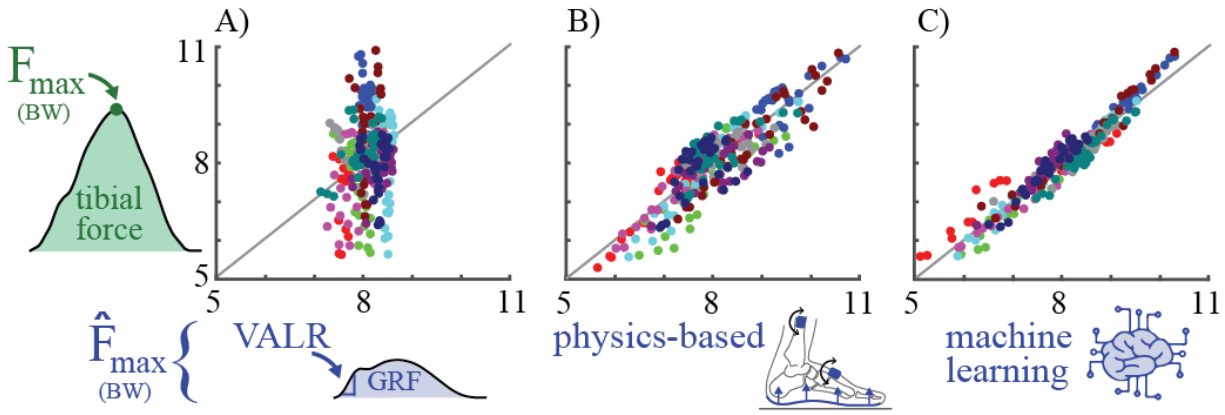


Figure 4.3 Comparison of force estimation algorithms. The plots depict the lab-based estimate of peak tibial force (F_{max}) vs. wearable sensor algorithm estimates of peak tibial force ($\hat{F}_{max}$), in units of BW. Estimates are from A) single variable linear regression using VALR, B) physics-based algorithm, and C) machine learning algorithm. Colors represent 10 runners. Each point represents a single running speed-slope combination (30 conditions per runner). A line with a slope of one is added to visualize a perfect correspondence between lab-based and wearable estimates.

Table 4.2 Average errors of force estimation algorithms. This table reports MAPE for algorithm estimates of peak tibial force ($\hat{F}_{max}$) relative to lab-based estimates of peak tibial force (F_{max}). Shown are the inter-runner means, standard deviations and ranges (*N*=10). The physics-based algorithm and machine learning algorithm exhibited lower force estimation errors than single variable linear regression algorithm using VALR (p=0.004 and p<0.001 respectively, based on Wilcoxon signed-rank test on the k-fold cross-validation error results).

	Single variable linear regression (VALR)	Physics-based	Machine learning
mean ± std	9.9 ± 3.5 %	5.2 ± 1.3 %	2.6 ± 0.8 %
[min max]	[4.5% 15.0%]	[3.8% 8.1%]	[1.5% 4.4%]

Other single variable linear regression estimates resulted in average MAPE of about 8-11% (**Figure 4.4** and **Table 4.3**). Results in RMSE for reference: Using GRF active peak resulted in a RMSE of 0.77 ± 0.12 BWs (**Figure 4.4A, Table 4.3**). Using GRF impulse resulted in a RMSE of 1.03 ± 0.38 BWs (**Figure 4.4B**, **Table 4.3**). Using ground slope resulted in a RMSE of 0.79 ± 0.33 BWs (**Figure 4.4C**, **Table 4.3**). Using running speed resulted in a RMSE of 0.91 ± 0.38 BWs (**Figure 4.4D**, **Table 4.3**). Using the group mean algorithm approach (i.e., assuming peak tibial force for a given runner was the same for every running condition) resulted in a RMSE of 0.99 ± 0.38 BWs (**Figure 4.4E**, **Table 4.3**).

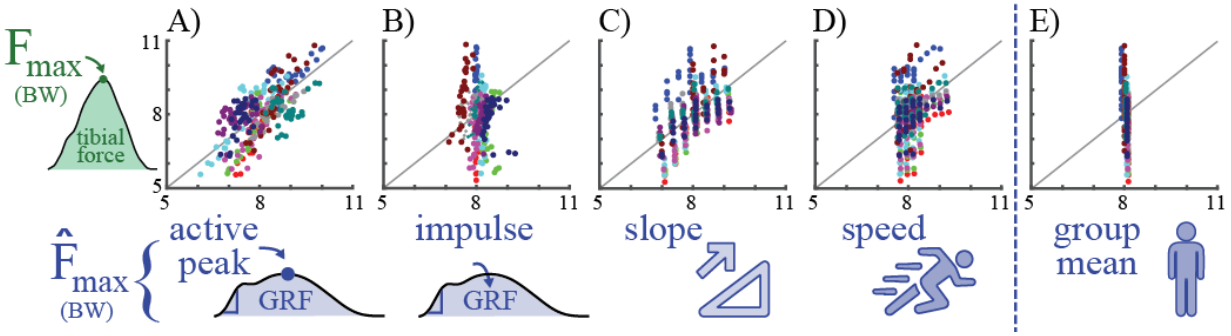


Figure 4.4 Comparison of additional single variable linear regression force estimation algorithms, and the group mean approach. The plots depict the lab-based estimate of peak tibial force (F_{max}) vs. single variable linear regression estimates of peak tibial force ($\hat{F}_{max}$), in units of BW. Single variable linear regression results used A) vertical GRF active peak, B) vertical GRF impulse, C) ground slope, and D) running speed. Peak tibial force estimates using E) the group mean force (i.e., a constant peak tibial force) are also plotted for reference. Colors represent 10 runners. Each point represents a single running speed-slope combination (30 conditions per runner). A line with a slope of one is added to visualize a perfect correspondence between lab-based and wearable estimates.

Table 4.3 Average errors of additional single variable linear regression force estimation algorithms, and the group mean approach. The table reports MAPE for algorithm estimates of peak tibial force ($\hat{F}_{max}$) relative to lab-based estimates of peak tibial force (F_{max}). Shown are the inter-runner means, standard deviations and min/max ranges (*N*=10).

	Single variable linear regression				Group mean force
	Vertical GRF active peak	Vertical GRF impulse	Slope	Speed	
mean ± std	8.2 ± 1.6 %	10.5 ± 3.9 %	8.5 ± 3.4 %	9.6 ± 4.2 %	10.3 ± 4.0 %
[min max]	[5.6% 10.4%]	[5.3% 15.2%]	[4.0% 16.1%]	[4.1% 15.9%]	[5.1% 15.8%]

4.5 Discussion

Here we demonstrate two multi-sensor algorithms that provide improved estimates of peak tibial force during running across speeds and slopes, as compared to conventional approaches used by current wearables. On average, the physics-based algorithm had 5% error and the machine learning algorithm has a 3% error, compared to the single variable linear regression approach using VALR or other GRF and single variable metrics that had average errors ranging from 8-11%. These findings

highlight that multi-sensor algorithms offer a promising and feasible approach for more accurately estimating musculoskeletal loads, like tibial force, using wearable sensors in applied situations.

4.5.1 Encouraging implications for using existing wearable sensors with improved algorithms

Fortunately, all the sensor signals and discrete features used in these algorithms can be obtained from existing wearable sensor hardware. This suggests that there are opportunities to achieve more accurate musculoskeletal load estimates solely by selecting appropriate sensors and accompanying software algorithms. Specifically, the physics-based algorithm used time series shank angle, foot angle, GRF magnitude, and anterior-posterior CoP under the foot, which could be obtained with IMUs placed on the shoe and shank and a pressure-sensing insole/shoe. For the machine learning algorithm, all discrete features could be estimated from signals obtained from the same IMUs and pressure-sensing insole/shoe.

While our reported machine learning algorithm used discrete features, which could be obtained from *three* wearable sensors, we also explored how reducing the number of sensors used in the algorithm to just *one* or *two* would influence algorithm estimates (**Table 4.4**). Three interesting insights we observed were: (i) Adding the shank IMU features to the pressure insole and foot IMU features had minimal effect on the tibial force estimation error. This was corroborated by outputs from the LASSO algorithm as after training the original machine learning model, the shank features were all ranked as low importance. (ii) Relatedly, the models including both the pressure insole and foot IMU resulted in the lowest errors. This was also consistent with LASSO outputs, as the original model ranked three features from these sensors as the highest importance: GRF active peak, anterior-posterior CoP at midstance, and the foot angle at midstance. (iii) Each of these iterations used more than a single feature to estimate peak tibial force and three of the combinations resulted in an error less than our single variable linear regression estimates or group mean approach, and three of the combinations resulted in an error equal to our single variable linear regression estimates or group mean approach (**Table 4.3** vs. **Table 4.4**). While it would be extremely convenient if a single metric/feature was a strong indicator of tibial force, we show that by combining just a few data features we are able to improve our peak tibial force estimates (average error of <3% compared to ~10% using a single metric). While not reported here, a similar exploration of resulting algorithm estimates using different sensor combinations could be repeated for the physics-based approach.

Table 4.4 Errors in tibial force estimates for machine learning algorithms using different sensor combinations. The table reports MAPE of wearable algorithm estimates of peak tibial force ($\hat{F}_{max}$) compared to lab-based estimates of peak tibial force (F_{max}). Each iteration used a different combination of wearable sensors providing a different number of discrete features. Refer to **Table 4.1** for an explanation of which discrete features were available from each idealized wearable sensor. Reported are the inter-runner means and standard deviations (*N*=10). Note: The first combination is equivalent to the subset used in our presented machine learning mode (equivalent to **Figure 4.3** and **Table 4.2**) and is included here for reference.

Wearable sensors used (f = total number of discrete features used)	**MAPE mean $\pm$ std (*N*=10)**
pressure insole + foot IMU + shank IMU (f=13)	2.6 ± 0.8 %
pressure insole + foot IMU (f=10)	2.6 ± 0.6 %
pressure insole (f=7)	4.7 ± 1.7 %
foot IMU + shank IMU (f=9)	8.3 ± 4.9 %
shank IMU (f=5)	8.0 ± 2.9 %
foot IMU (f=5)	7.9 ± 2.3 %

4.5.2 Opportunities for advancement of wearables that monitor musculoskeletal loads

There are numerous opportunities for further exploration and improvement of the two algorithms presented here. First, now that candidate sensors and signals/features have been identified in this study, a fully portable/wearable system could be developed and its accuracy could be assessed. It is important to remind that idealized wearable signals allowed us to identify the most promising signals and features; however, in many instances a given signal or feature could be estimated from multiple different types of sensors. For example, GRF active peak could be estimated either with a pressure insole, or with a waist-mounted accelerometer [20], [105]. Thus, there may be opportunities to strategically select sensor hardware to accommodate application constraints (e.g., a footwear developer may opt to measure GRF active peak with an insole, whereas an apparel developer may opt to integrate an accelerometer into the waistband of shorts).

Second, an additional algorithm approach is to fuse the physics-based and machine learning algorithms, each of which has its own benefits and challenges. The physics-based algorithm is based on equations of motion that describe the system (the runner's tibial bone), which theoretically should estimate force accurately beyond the running conditions tested here. However, as detailed in methods, current wearable sensors provide limited signals that are insufficient to compute the full equations of motion. The LASSO machine learning algorithm is advantageous because it uses a small set of idealized wearable sensor features and develops a model that minimizes tibial force estimation error in the training data. However, the final model is dependent on the training data and should not be assumed to extrapolate to running conditions beyond the training set. Potentially, one could first estimate tibial force using the physics-based algorithm and then use these results, in addition to other wearable features, as inputs to the machine learning model. This would leverage the strengths of each algorithm. Another interesting and yet unexplored research direction is to build and train non-linear neural network models with internal structures influenced by the physics-based equations – effectively embedding the domain knowledge (biomechanics) into the network architecture. For example, if from inverse dynamics we know that CoP will be multiplicatively related to GRF (to estimate ankle moment), we could add that non-linearity into the neural network model structure.

Third, as new sensor hardware is developed, it is expected to further enhance musculoskeletal load estimation accuracy. For instance, future commercial pressure insoles may be able to measure the anterior-posterior GRF, providing more complete input signals for physics-based algorithms, and additional candidate features for machine learning. Ultimately, the desired estimation accuracy will vary based on the intended application as there is some tradeoff between hardware complexity and estimation accuracy.

4.5.3 Discouraging implications for VALR and other single variable surrogates for musculoskeletal load

VALR was a poor predictor of tibial bone force. The average force errors from VALR (9.9%, 0.97 BWs) were nearly the same as using the naive group mean approach (10.3%, 0.99 BWs), which simply assumes a constant peak tibial force for every step. These findings suggest that if our goal is to monitor peak force on the tibial bone, then tracking VALR may be no more accurate than simply using a pedometer (step counter) and assuming a constant mean force for every step. Of note, for the group mean approach presented here to emulate a step counter, we chose to use the group mean of peak tibial force, but any other average metrics (median, mode, etc.) would similarly result in estimates that do not capture trends in tibial force (**Figure 4.4E**).

The machine learning algorithm results provide further evidence of the limited utility of VALR. VALR was the one and only discrete feature for which the LASSO machine learning algorithm drove the weighting factor to zero, indicating that VALR was not needed or used in model predictions. These findings may be surprising to some readers given the fact that VALR (or correlated signals like GRF impact peak, peak foot/shank accelerations ([103], [104], [179], etc.) is commonly assumed in sport science, biomechanics and related disciplines to signify athlete workload, internal loading or injury risk. However, as discussed in [173] and [174], we believe these assumptions/interpretations reflect fundamental misconceptions in the field regarding the relationship between external forces (between the foot and the ground), the loads experienced inside the body by musculoskeletal structures (e.g. **Figure 4.3A**), and the resulting damage and injury risk. Thus, we continue to urge vigilance in how VALR and other impact-related metrics are used and interpreted within the context of sport science and injury risk assessment.

4.5.4 Important insights gained from estimating tibial damage

The methods above describe how we assessed the bone force estimation accuracy of each algorithm, which enables us to evaluate relative performance of algorithms. However, an open question in the research field is: what level of musculoskeletal force estimation accuracy is needed for assessing overuse injury risk? It is impossible to know definitively at this stage and deciding when a model is accurate enough is a known challenge in the field (see [170] and **Section 4.5.5** below). Nevertheless, one way to gain insight is to consider the relationship between bone force and bone microdamage, since overuse injuries are believed to result from an accumulation of microdamage [9]. Here we use the word *damage* to refer to the mechanical fatigue-induced microstructural damage to bone tissue [9].

There is extensive experimental evidence that material (e.g., bone) damage is exponentially related to

the force applied (Damage α Force^7, [9], [90], [123], see **Supplement 4.7.2** for additional details). As an example of how even modest errors in force estimates result in enormous errors in bone damage estimates, 8-11% errors in tibial force estimates based on single variable linear regressions would translate, on average, to bone damage errors of about 70-120% (**Supplemental Figure 4.2 & Supplemental Figure 4.3**, **Supplemental Table 4.1** & **Supplemental Table 4.2**). Alternatively, 5% errors in tibial force estimated using the physics-based algorithm would translate to about 41% error in bone damage and 3% errors in tibial force estimated using the machine learning algorithm would translate to about 18% error in bone damage (**Supplemental Figure 4.2, Supplemental Table 4.1**). See **Supplement 4.7.2** for full bone damage results.

To gain intuition into how errors in our force and damage estimates may accrue over a series of loading cycles (i.e., over multiple running strides), we also estimated *cumulative* damage, using both our lab-based estimate of peak tibial force and our algorithm estimates of peak tibial force, over simulated five-mile running sessions. For brevity of the main text, these cumulative damage methods and full results are presented in **Supplement 4.7.2**, with one finding presented here:

When comparing 25 simulated running workouts of differing intensity (varying number of running strides at different combinations of speeds and slopes), our lab-based estimates of cumulative damage resulted in different amounts of cumulative bone damage per run (**Figure 4.5**). If the runner used the VALR linear regression based estimates of cumulative damage, they would be misguided about how much damage their tibial bone experiences during each workout, with some workouts being largely overestimated and some being largely underestimated. There are important consequences to severely over- or under-estimating damage to the bone. Large overestimates of bone damage may result in wearables that advise to sideline an athlete unnecessarily, while large underestimates of damage may result in a failure to notify an athlete of heightened risk before an injury occurs. Alternatively, estimating cumulative damage per run with the physics-based or machine learning algorithm more accurately estimates the bone damage per run (**Figure 4.5**); capturing key trends in higher vs. lower damage workouts. These two multi-sensor algorithms offer an exciting and powerful tool for capturing important trends in bone damage and potentially overuse injury risk (due to microdamage accumulation).

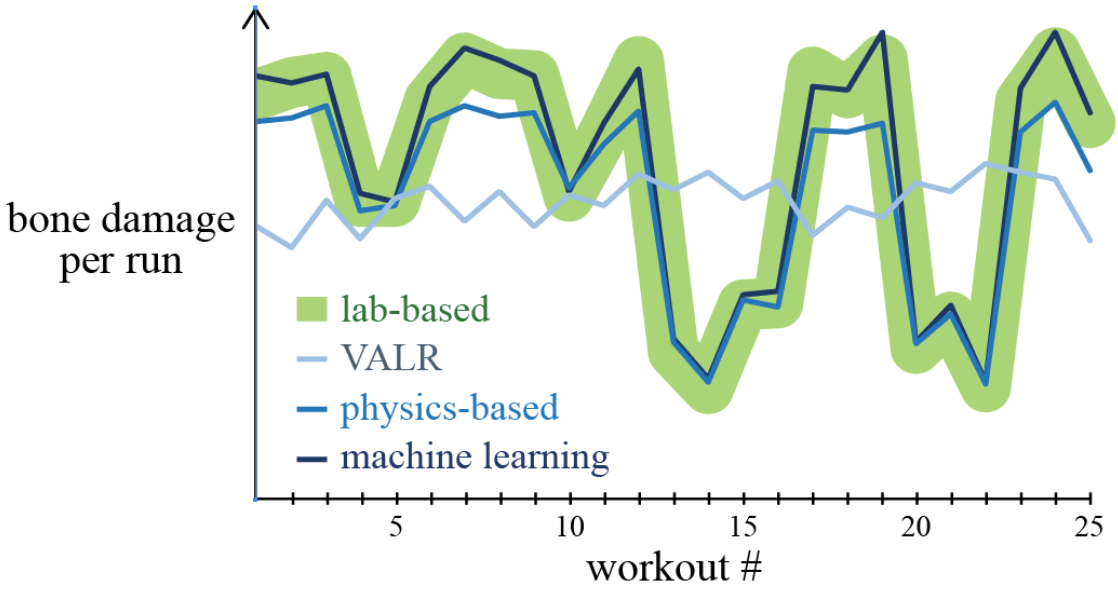


Figure 4.5 Example of how each algorithm performs across 25 simulated running workouts of varying intensity. This plot highlights how the physics-based (medium blue line) and machine learning (dark blue line) algorithms are able to track trends in tibial bone damage per workout over multiple running workouts. The green band represents lab-based estimates of bone damage +/- a small margin of error. Using VALR (light blue line) to track bone damage across multiple runs fails to provide the runner with accurate or useful information about which workouts expose them to higher vs. lower amounts of cumulative bone damage and associated injury risk.

Considering bone *damage* when developing musculoskeletal load monitoring tools theoretically takes us one step closer to the mechanical mechanism underlying bone overuse injuries. While it is not yet clear what level of accuracy is sufficient for different applications (e.g., longitudinal research studies, athlete injury risk monitoring), our findings highlight the importance of considering tissue damage (in this case bone damage), not only force, when developing musculoskeletal loading and injury prevention tools.

4.5.5 Path from new tools to scientific understanding and societal impact

There are a couple common questions regarding monitoring bone load, damage and overuse injury risk that we have received since sharing initial results and a preprint of this study, namely: How do we know that estimates of peak bone force or associated damage will be good indicators of injury risk? How do we know which specific bone loading metric(s) to track? How do we know how accurate estimates of bone loading or damage need to be? The simple answer: We don't. This would be akin to asking: Before the invention of the pulse oximeter, how did medical device developers know this tool could be used to screen neonates for Critical Congenital Heart Disease, know the precise oximetry metrics to track, and know what level of estimation accuracy of these metrics was necessary? They didn't. Often when envisioning a new screening, diagnostic or health monitoring tool, the first versions are developed based on the state of scientific and clinical evidence at that time. Pulse oximeter developers used existing knowledge that monitoring oxygen saturation could potentially provide useful insight about physiological function and health. After developing an oxygen saturation

estimation tool, and performing a series of research studies to refine the hardware, analytics and clinical insights, device developers were then able to use this new and validated tool to monitor neonatal (and other clinical) health and improve care, resulting in a huge societal impact.

Similarly, when envisioning a new bone stress injury risk monitoring tool, we were motivated by the current state of scientific evidence on the etiology of bone stress injuries. Specifically, we were motivated by evidence that bone stress injuries are consistent with mechanical fatigue processes due to tissue damage accumulation resulting from repetitive bone loading ([9], [91]). Interestingly, a risk assessment tool has previously been developed to estimate tissue damage to the low back due to repeated lifting, and this tool has been validated against two occupational epidemiological injury databases (i.e., shown to explain 72-95% of the deviance in low back disorders, [180]). Thus, there is both a strong scientific motivation and risk assessment precedence (from the occupational health field) for using tissue load monitoring in the way we propose and envision.

However, currently there is a critical technical gap: we lack the ability to monitor tibial bone loading outside the lab, which inhibits the evaluation of these mechanical fatigue processes in real world environments for runners or other individuals. Development of a novel wearable bone load (and damage) monitoring tool will enable researchers and clinicians to collect longitudinal prospective data on bone loading while also tracking injury incidence. Examining these data can then allow for testing hypotheses about the role of bone loading in the development of bone stress injuries, potentially identifying new bone health and injury risk predictive metrics. Proper measurement tools must be developed and deployed before injury risk predictive metrics will be discovered and validated. Of note, bone load and damage monitoring will likely not be used in isolation. Rather, these data may be combined with other physiologically relevant information (e.g., sleep duration, diet, bone geometry, demographics, bone bending moments) to develop deeper insights into bone health and overuse injury mechanisms. Much remains unknown about which metrics are key, how to combine them to most effectively monitor bone health or injury risk, and what accuracy of each metric is needed for different applications. What is known is that developing a wearable bone load and damage monitoring tool will provide researchers and clinicians with unique new capabilities to gain a deeper scientific understanding of bone health. Expanding knowledge related to bone stress injury mechanisms and prevention has huge potential for improving societal health, happiness, and productivity. This study represents a tangible step towards this goal.

4.5.6 Additional limitations and opportunities

First, we focused on one specific musculoskeletal structure (the tibia), direction of loading (tibial compression), and range of conditions which were all treated with equal weight in algorithm development (treadmill running across a range of speeds and slopes). The presented algorithms were developed within these bounds, and any additional musculoskeletal structure, direction of loading, or range of conditions would necessitate additional validation studies. Nevertheless, the same analysis could be used to investigate other musculoskeletal structures (e.g., lumbar loading for individuals at risk of back overuse injury). Our outlined process can help identify candidate sensors and signals, develop algorithms, and begin to establish expectations for levels of accuracy that might be achievable with existing types of sensors.

Second, algorithm development required us to choose which method of estimating tibial force (and

resulting damage) was considered our ground truth. As described previously [173], we have confidence in lab-based estimates of tibial compression force. There is less agreement in the literature about other loading directions (e.g., tibial bending, [13], [14], [168], [169], [181]–[183]). Nevertheless, there may be value to exploring algorithms for estimating these other loading directions, or for estimating damage due to the combined loading from axial compression and bending (see **Supplement 4.7.2** Limitations for additional discussion). However, it should be emphasized that these other loading directions are also highly dependent on muscles forces, and it is again unlikely that a single variable metric (e.g., VALR) would be an accurate predictor of internal loading.

Third, these algorithms were developed using idealized wearable signals. Actual wearable sensor signals are expected to exhibit additional noise, errors, or drift, and therefore the results presented here should be interpreted as best case expectations for wearables utilizing the sensor/signal sets described here. Future studies could use actual wearable sensor signals as model inputs, or apply noise, errors, or drift to current idealized wearable signals. However, previous validation experiments have shown good agreement between pressure insoles and the vertical GRF [184] and IMU-based kinematics and lab-based motion capture kinematics [176]. Additionally, sensor hardware, algorithms and signal quality are constantly improving, bringing wearable sensor measurements closer to lab-based measurements.

Fourth, with advances in sensor hardware and sensor integration techniques, the potential for development and adoption of musculoskeletal load monitoring tools into daily life will also evolve. Sensor integrated clothing, footwear, insoles, textiles, etc. are becoming increasingly common and the use of multiple sensors is becoming increasingly practical. In fact, there are already multiple commercial wearables available that contain a pressure-insole and an IMU (a combination of sensors yielding the lowest bone force estimation error in this study, **Table 4.4**). The results of this study demonstrate the potential pay-offs of leveraging a multi-sensor approach. It is important to remember that there was a time when walking around with a phone in your pocket seemed almost unimaginable. But now it is routine to walk around all day with a pocket full of sensors -- smartphones that fit seamlessly into our daily lives enhance our productivity, connectedness, health monitoring, etc. in ways that were previously unthinkable. Similarly, we expect that the limiting factor of musculoskeletal load monitoring tools will not be the number or cost of sensors, or the size of batteries, but rather the limiting factor will be our imagination, our vision, and the ability to combine data science (e.g., machine learning) and domain knowledge (e.g., biomechanics) to provide actionable and scientifically validated insights that improve societal health and well-being.

Finally, we note that the analysis here was performed on stride-averaged data. However, we expect all the same conclusions to hold true for stride-by-stride analysis, since the variability within a running condition was small relative to the variability across conditions.

4.6 Conclusion

Here we demonstrate two multi-sensor algorithms – one physics-based and one machine learning – that offer promising solutions for estimating peak tibial force with wearable sensors. We show that by harnessing signals from existing wearable sensors, and applying multi-sensor algorithms, it is feasible to drastically improve the estimation accuracy of peak tibial force (by two or four fold relative to the current state-of-the-art in commercial wearables). This study highlights the exciting potential to

combine wearables, musculoskeletal biomechanics and machine learning to develop more accurate tools for monitoring musculoskeletal loading in applied situations.

4.7 Chapter 4 Supplemental Material

4.7.1 Physics-based Algorithm Methods

Equation 4.1 describes the lab-based estimate of tibial force ($F(t)$) as the summation of the net force on the ankle and the force from the plantarflexor calf muscles (F_m).

$$F(t) = |\boldsymbol{GRF}| \cdot \cos(\beta) + F_m \quad \textbf{Equation 4.1}$$

Where β is the 3D angle between the $\boldsymbol{GRF}$ and long-axis of the lower leg segment, and F_m was estimated by dividing the net sagittal plane ankle moment by the Achilles tendon moment arm (assumed constant, 5 cm, (Matijevich et al., 2019)). From $F(t)$ we extracted F_{max}, the maximum tibial force during stance. Vectors are bolded, and scalars are non-bolded.

To estimate tibial force from the available idealized wearable signals ($\hat{F}(t)$) we created a modified version of **Equation 4.1**. We approximated both the net force on the ankle and the force from plantarflexor calf muscles (**Equation 4.2**). The net force on the ankle was estimated by projecting the normal force under the foot (GRF_{normal}) onto an estimate of the long axis of the tibia. The force from the plantarflexor calf muscles was estimated by approximating the net ankle moment, then dividing by Achilles tendon moment arm length (r_{AT}, assumed constant, 5 cm, [173]). To approximate the net ankle moment, we computed $CoP_{ap} \cdot GRF_{normal}$, where CoP_{ap} was the anterior-posterior center of pressure distance relative to the ankle joint position. The ankle joint position was assumed 5 cm anterior to the heel of the foot, based on a typical Achilles tendon moment arm length. Summing these approximations in **Equation 4.2** yields an unscaled estimate of tibial force $\hat{F}'(t)$. From $\hat{F}'(t)$ we extracted $\hat{F}'_{max}$, the maximum unscaled tibial force during stance.

$$\hat{F}'(t) = GRF_{normal} \cdot \cos(\hat{\beta}) + \frac{(CoP_{ap}) \cdot GRF_{normal}}{r_{AT}} \quad \textbf{Equation 4.2}$$

Where $\hat{\beta}$ is the approximate angle between GRF_{normal} and the long-axis of the lower leg segment. $\hat{\beta}$ is estimated as $90° - \theta_{foot} + \theta_{shank}$, where θ_{foot} and θ_{shank} are foot and shank angles in the sagittal plane relative to the horizontal.

Of note, **Equation 4.2** assumes that CoP_{ap} and GRF_{normal} are perpendicular to each other, neglects moment and reaction force contributions from the anterior-posterior and mediolateral GRFs and is devoid of a true ankle joint center estimate. We therefore anticipated this physics-based algorithm might systematically under- or over-estimate tibial force; thus why we refer to this initial estimate as

unscaled. To account for this, linear scaling coefficients were calculated using k-fold validation, then applied to arrive at a scaled estimate of peak tibial force ($\hat{F}_{max}$).

4.7.2 Extended methods and results for estimating bone damage and cumulative bone damage over simulated running sessions

Appendix B provides extended methods and results for estimating bone damage and cumulative bone damage over simulated running sessions. We evaluated both the bone damage per running stride and the cumulative bone damage accrued over simulated five-mile running sessions, to supplement our analysis of force estimation accuracy.

4.7.2.1 Methods: Bone damage per running stride

Bone damage per running stride was estimated using an established inverse power law for bone damage ($D = cF_{max}{}^{e}$ and $\widehat{D} = c\hat{F}_{max}{}^{e}$) that describes the life-stress relationship of materials under cyclic loading, where e is an empirically-derived exponential constant for bone of seven [122], [125], [133]. Here, c is set equal to unity such that our cumulative damage are dimensionless units of relative cumulative damage. A different constant c and additional modeling factors would be needed to translate D into physical representations of mechanical fatigue (e.g., number of cycles until bone failure or probability of failure, [123], [124]). For each algorithm, we computed the MAPE between D and $\widehat{D}$ across all conditions for each runner.

4.7.2.2 Methods: Cumulative damage and simulated running session

We estimated *cumulative* damage over a series of loading cycles (i.e., over multiple steps). We created 1000 *simulated runs* that were each approximately five miles. The series of loading cycles was constructed by randomly assigning a speed-slope combination in half mile increments, constrained by speed-slope combinations we collected in our experiment (one example in **Supplemental Figure 4.1**). To generate a plausible running session, for each subsequent increment, the slope was constrained to change by either +3°, 0°, or -3°, and the speed was randomly assigned. The number of steps taken at each half mile increment was estimated based on a previously reported speed-cadence relationship [185]. We estimated cumulative damage over each simulated run using Miner's rule of cumulative fatigue over a series of loading cycles at different peak forces (**Equation 4.3, Equation 4.4**, [9], [90]),

$$\Sigma D = \sum_{i=1}^{10} n_i \cdot F_{max,i}^{7} \qquad \textbf{Equation 4.3}$$

$$\Sigma \widehat{D} = \sum_{i=1}^{10} n_i \cdot \hat{F}_{max,i}^{7} \qquad \textbf{Equation 4.4}$$

Where i represents the half mile increment, n_i is the number of steps taken in that half mile increment based on the speed assigned, and $F_{max,i}$ and $\hat{F}_{max,i}$ represent the peak force for the speed-slope

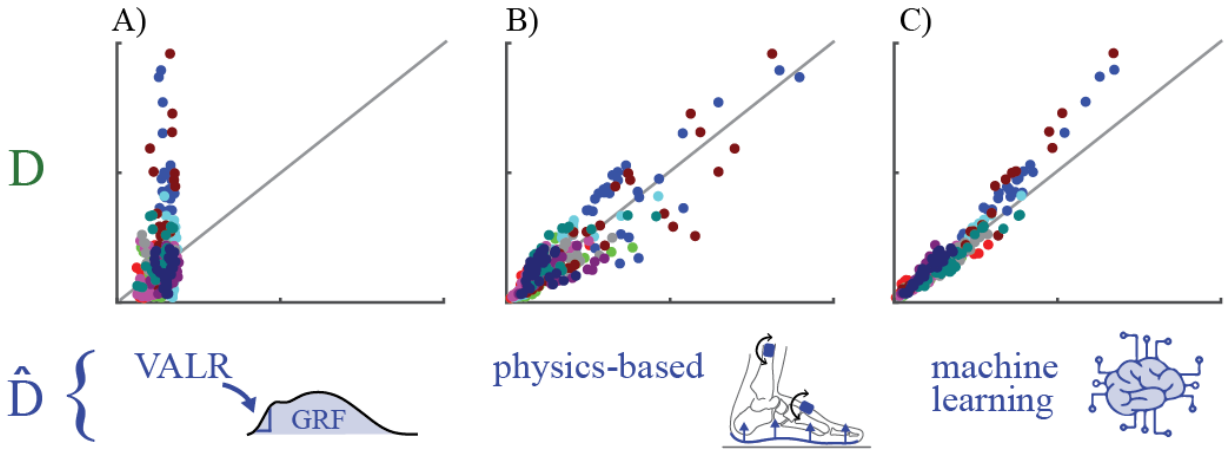
combination assigned to that half mile increment (see **Supplemental Figure 4.1** for one example). For each runner, the cumulative damage ΣD vs. $\Sigma \widehat{D}$ was plotted for all 1000 simulated runs. We then computed the MAPE for each runner. Finally, we calculated the inter-runner average and range of MAPE.

i	1	2	3	4	5	6	7	8	9	10
slope (˚)	0	0	3	3	0	-3	-6	-3	0	0
speed (m/s)	3.0	3.4	3.0	3.2	3.2	2.6	2.8	2.8	3.4	4.0
n (# steps)	750	650	750	750	750	850	850	850	650	550
$F_{max,i}$ (BW)	7.9	8.3	8.3	8.4	8.0	7.1	7.3	7.5	8.3	8.5

Supplemental Figure 4.1 Example of one five-mile simulated running session. The embedded table shows our lab-based estimates of peak tibial force ($F_{max,i}$) for our tenth runner.

4.7.2.3 Results & Discussion: Bone damage per running stride

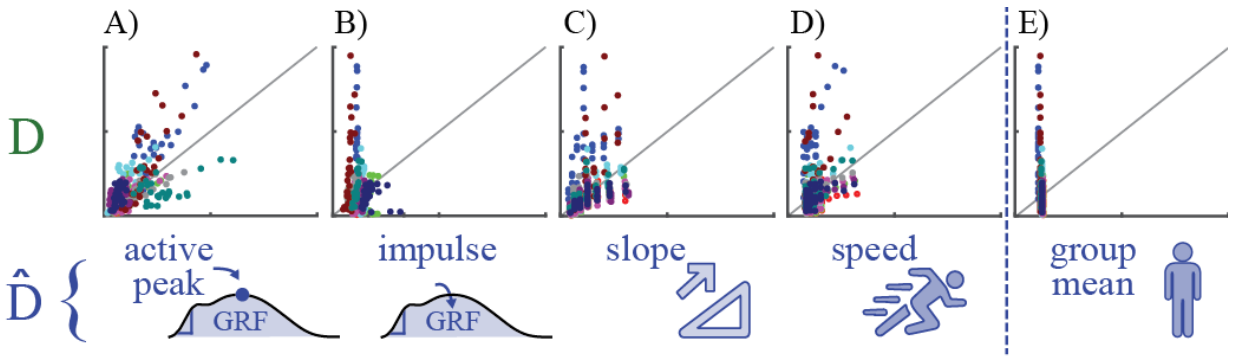
Below are the figures and tables that support the **Discussion section 4.5.4**.



Supplemental Figure 4.2 Comparison of bone damage based on force estimation algorithms. Bone damage estimated using lab-based peak tibial force (D) vs. damage estimated using wearable sensor algorithm peak tibial force ($\widehat{D}$). Damage is reported in dimensionless units. Algorithm estimates are from A) single variable linear regression using VALR, B) physics-based algorithm, and C) machine learning algorithm. Colors represent 10 runners. Each point represents a single running speed-slope combination (30 conditions per runner). A line with a slope is added to visualize a perfect correspondence between lab-based and wearable estimates. See **Figure 4.3** for accompanying force estimation results.

Supplemental Table 4.1 Average errors in bone damage estimates based on force estimation algorithms. The table reports MAPE of D vs. $\widehat{D}$. Reported are the inter-runner means and standard deviations (*N*=10). The physics-based algorithm and machine learning algorithm exhibited lower damage estimation errors than single variable linear regression algorithm using VALR (p=0.02 and p<0.001 respectively, based on Wilcoxon signed-rank test on the k-fold cross-validation error results translated into estimates of damage). See **Table 4.2** for accompanying force estimation results.

	Single variable linear regression (VALR)	Physics-based	Machine learning
mean ± std	104 ± 70 %	41 ± 16 %	18 ± 5 %
[min max]	[25% 242%]	[25% 81%]	[11% 27%]



Supplemental Figure 4.3 Comparison of bone damage based on additional single variable linear regression force estimation algorithms, and the group mean approach. The plots depict bone damage based on lab-based estimates of peak tibial force (D) vs. damage based on single variable linear regression estimates of peak tibial force ($\widehat{D}$). Damage is reported in dimensionless units. Single variable linear regression results used A) vertical GRF active peak, B) vertical GRF total impulse, C) ground slope, and D) running speed. Damage estimates based on E) the group mean approach are also plotted in for reference. Colors represent 10 runners. Each point represents a single running speed-slope combination (30 conditions per runner). A line with a slope of one is added to visualize a perfect correspondence between lab-based and wearable estimates. See **Figure 4.4** for accompanying force estimation results.

Supplemental Table 4.2 Errors in bone damage estimates based on additional single variable linear regression force estimation algorithms, and the group mean approach. The table reports MAPE of D vs. $\widehat{D}$. Reported are the inter-runner means and standard deviations (*N*=10). See **Table 4.3** for accompanying force estimation results.

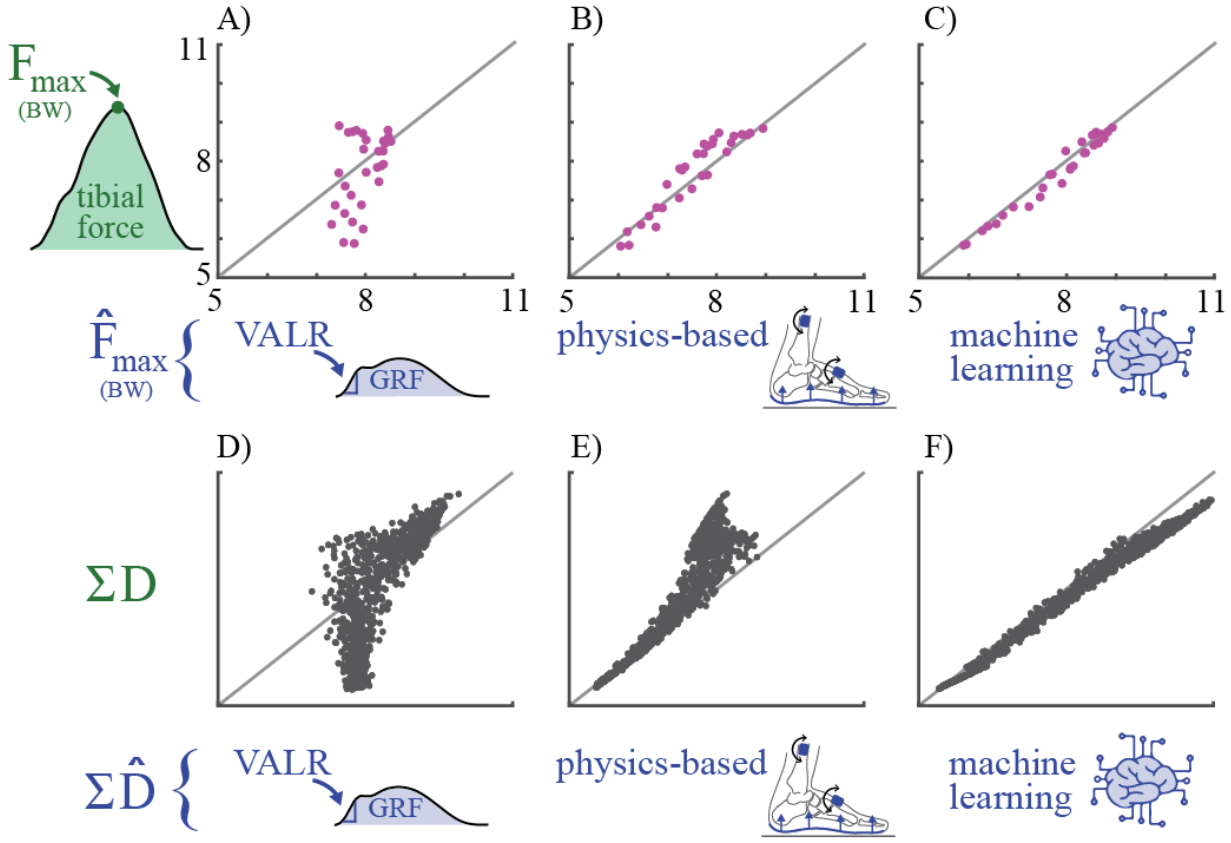
	Single variable linear regression				Group mean force
	Vertical GRF active peak	Vertical GRF impulse	Slope	Speed	
mean $\pm$ std	73 $\pm$ 36 %	127 $\pm$ 93 %	71 $\pm$ 35 %	103 $\pm$ 73 %	121 $\pm$ 84 %
[min max]	[39% 132%]	[33% 304%]	[24% 133%]	[28% 234%]	[38% 261%]

4.7.2.4 Results & Discussion: Cumulative damage and simulated running session

Cumulative damage calculated using the single variable linear regression that scales VALR to estimates of peak tibial force resulted in an average MAPE of 85 $\pm$ 59 % across runners (**Supplemental Figure 4.4D**, **Supplemental Table 4.3**). Cumulative damage estimated using our physics-based algorithm estimates of peak tibial force resulted in a MAPE of 31 $\pm$ 12 % (**Supplemental Figure 4.4E**, **Supplemental Table 4.3**). Cumulative damage estimated using our machine learning algorithm estimates of peak tibial force resulted in a MAPE of 13 $\pm$ 5 % (**Supplemental Figure 4.4F**, **Supplemental Table 4.3**).

To help visualize how errors in force translate to errors in damage, we plotted force and damage results together for a single runner (**Supplemental Figure 4.4**). We first plotted lab-based estimates of peak tibial force across 30 running conditions vs. wearable sensor algorithm estimates (**Supplemental Figure 4.4A-C**). We then plotted lab-based estimates of cumulative damage across 1000 simulated

runs vs. wearable sensor estimates (**Supplemental Figure 4.4D-F**) directly underneath for the same runner.

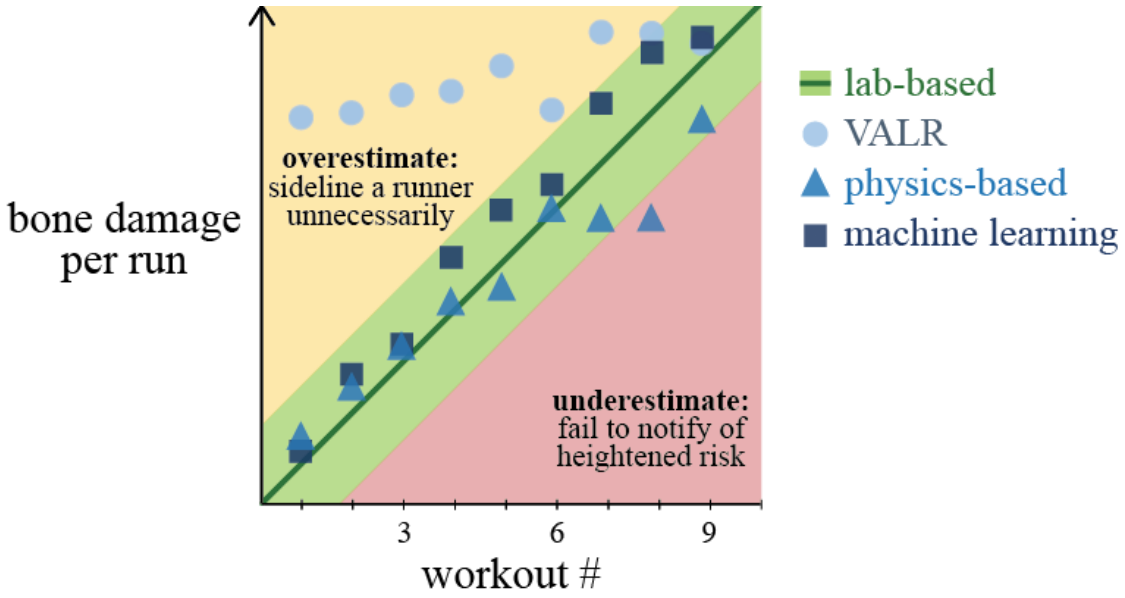


Supplemental Figure 4.4 Cumulative damage from simulated runs for a single example runner. A-C) These plots depict lab-based estimates of peak tibial force (F_{max}) vs. wearable sensor algorithm estimates of peak tibial force ($\hat{F}_{max}$) for an example runner for 30 running conditions. **D-F)** These plots depict lab-based estimates of cumulative damage over 1000 simulated runs ($\sum D$) vs. wearable sensor algorithm estimates of cumulative damage ($\sum \hat{D}$). Force is reported in units of BW and damage is reported in dimensionless units. Using VALR, a 8.5% error in tibial force translated to a 77% error in cumulative damage. Using the physics-based algorithm, a 6.8% error in force translated to a 38% error in cumulative damage. Using the machine learning algorithm, a 2.1% error in force translated to a 10% error in cumulative damage. A line with a slope of unity is added to visualize a perfect correspondence between lab-based and wearable estimates.

Supplemental Table 4.3 Average errors in relative cumulative damage from simulated runs. The table reports MAPE of $\sum D$ vs. $\sum \hat{D}$. Reported are the inter-runner means and standard deviations (*N*=10).

	Single variable linear regression (VALR)	Physics-based	Machine learning
mean ± std	85 ± 59 %	31 ± 12 %	13 ± 5 %
[min max]	[25% 213%]	[17% 57%]	[7% 22%]

To contextualize the consequences of misestimating bone damage, we imagined a runner with a history of tibial stress fractures who wanted to monitor bone damage across multiple days of running, in order to understand which workouts lead to more or less wear and tear on her tibia. This runner would need a wearable that could be worn daily and could track bone damage for each workout, so she could evaluate when rest and recovery days are needed to help avoid a reoccurrence of the overuse injury. While this runner doesn't need a perfect estimate of bone damage, she does need the estimate to be within some reasonable margin of error. This is because if a wearable largely *over*estimates the cumulative damage to the bone, she may think she is at heightened injury risk, and sideline herself unnecessarily (Supplemental Figure 4.5 Example of how each algorithm performs across nine workouts of increasing bone damage. **Supplemental Figure 4.5**). If a wearable *under*estimates the cumulative damage to the bone, she may not be notified of heightened injury risk in time, and fail to take rest days before an overuse injury occurs (**Supplemental Figure 4.5**).



Supplemental Figure 4.5 Example of how each algorithm performs across nine workouts of increasing bone damage. From our 1000 simulated 5-mile runs, we selected nine runs that spanned from low to high bone damage. The green line represents the lab-based estimates of bone damage per run. The green band is a visual reminder that for any application there would be some acceptable margin of error. Blue circles, triangles, and squares represent wearable estimates of bone damage using VALR, the physics-based, and the machine learning algorithms, respectively. For this subset of runs, VALR severely overestimates the damage to the bone for seven of the nine workouts. In contrast, the physics-based algorithm only underestimates the damage to the bone for one workout, and the machine learning algorithm did not yield any estimates outside the depicted margin of error.

To further illustrate how a wearable that misestimates cumulative damage may affect our imagined runner, we did the following: From our 1000 simulated five-mile running sessions, we selected a subset of 25 runs, each with a different amount of lab-based cumulative damage per run. This illustration is presented in Discussion of the main text and **Figure 4.5**. This example also provides an important reminder that even though runs may be the same distance (5 miles), they can result in hugely different amounts of cumulative bone damage per run.

4.7.2.5 Limitations

These cumulative damage estimates do not account for bone remodeling and adaptation [90] that would occur over time between these simulated workouts as our goal was only to compare damage accrued during each simulated running workout due to repeated bone loading. Interestingly, bone remodeling and adaptation are also influenced by bone load [9], and therefore using estimates based on single-variable linear regression (e.g., VALR) may also limit estimation and understanding of these processes. While a wearable tool for tracking long-term bone health will likely include a composite model that combines estimates of bone damage, remodeling, adaptation, and other processes and risk factors, a reasonable estimate of bone load is foundational to estimating many of these components.

Additionally, the damage estimate here was based solely on compressive (axial) force. Future algorithms could be developed to estimate other loading directions (e.g., bending moment) and/or resulting bone stresses, and future estimates of damage may be obtained using these combined loads and/or stresses.

We note that even if we did add additional remodeling processes and/or loading directions to our estimates of damage, it would not alter the conclusions drawn here: (i) that the use of single-variable linear regression (e.g., VALR) leads to inaccurate tibial force and damage estimates, and (ii) that more accurate estimates of tibial force (e.g. using the physics-based and machine learning algorithms introduced in this study) can improve our estimates of tibial damage.

5 A promising wearable solution for practical and accurate monitoring of low back loading in manual material handling

5.1 Abstract

(1) Background: Low back disorders are a leading cause of missed work and physical disability in manual material handling due to repetitive lumbar loading and overexertion. Ergonomic assessments are often performed to understand and mitigate the risk of musculoskeletal overexertion injuries. Wearable sensor solutions for monitoring low back loading have the potential to improve the quality, quantity, and efficiency of ergonomic assessments and to expand opportunities for the personalized, continuous monitoring of overexertion injury risk. However, existing wearable solutions using a single inertial measurement unit (IMU) are limited in how accurately they can estimate back loading when objects of varying mass are handled, and alternative solutions in the scientific literature require so many distributed sensors that they are impractical for widespread workplace implementation. We therefore explored new ways to accurately monitor low back loading using a small number of wearable sensors. (2) Methods: We synchronously collected data from laboratory instrumentation and wearable sensors to analyze 10 individuals each performing about 400 different material handling tasks. We explored dozens of candidate solutions that used IMUs on various body locations and/or pressure insoles. (3) Results: We found that the two key sensors for accurately monitoring low back loading are a trunk IMU and pressure insoles. Using signals from these two sensors together with a Gradient Boosted Decision Tree algorithm has the potential to provide a practical (relatively few sensors), accurate (up to r2 = 0.89), and automated way (using wearables) to monitor time series lumbar moments across a broad range of material handling tasks. The trunk IMU could be replaced by thigh IMUs, or a pelvis IMU, without sacrificing much accuracy, but there was no practical substitute for the pressure insoles. The key to realizing accurate lumbar load estimates with this approach in the real world will be optimizing force estimates from pressure insoles. (4) Conclusions: Here, we present a promising wearable solution for the practical, automated, and accurate monitoring of low back loading during manual material handling.

5.2 Introduction

Low back disorders are a leading occupational health problem, ranging from lumbar (low back) pain to muscle strains to herniated spinal discs. Physical pain, missed work, decreased productivity, healthcare costs, short- and long-term disability, and psychological distress due to these low back disorders are substantial and persistent burdens on our society. Back disorders account for about 40% of all work-related musculoskeletal disorders [186], and about one in four workers reports dealing with low back pain [187], [188]. Individuals working in manual material handling jobs (and other jobs with similar physical demands) are at particularly high risk for low back disorders due to repetitive lifting and bending, which can lead to musculoskeletal overexertion (overuse) injuries [186].

Overexertion injuries result from an accumulation of microdamage caused by repetitive loading to musculoskeletal tissues (e.g., muscles, tendons, ligaments, bones, discs). Overexertion injuries are consistent with a fatigue failure process: the weakening and eventual failure of a material due to repeated loading [9], [180], [189]. When modeling this fatigue failure process, both the number of loading repetitions and the magnitude of loading on the musculoskeletal tissues are important for

approximating the cumulative damage to the tissue. There are multiple opportunities to use musculoskeletal loading and fatigue failure insights to understand and reduce the risk of overexertion injuries, such as through ergonomic assessments or continuous, personal monitoring of injury risk.

5.2.1 Ergonomic assessments

Ergonomic risk assessment tools that evaluate low back loading and assess injury risk using fatigue failure principals have shown potential for predicting the incidence of low back disorders. For example, the Lifting Fatigue Failure Tool (LiFFT) estimates cumulative tissue damage to the low back using an estimate of lumbar moment. Cumulative damage across a series of lifting tasks estimated with LiFFT has been shown to explain 72–95% of the deviance in low back disorders from epidemiological databases [180]. Ergonomic risk assessments are traditionally performed via direct observation by a trained professional. For instance, to perform an ergonomic assessment using LiFFT (or other assessment tools like the NIOSH Lifting Equation [190]), an ergonomist or safety professional would monitor a single worker during their shift, or over a subset of representative job tasks, to manually record how much each lifted object weighed and how far away each lifted object was from the body, then input how many times each type of lift is performed during a shift. The time spent observing a worker depends on the variability of job tasks (e.g., short- vs. long-cycle jobs), but is often in the order of 1–8 hours per job.

While these valuable ergonomic assessments and injury risk profiles can inform the use of ergonomic controls to minimize the risk to workers, the assessments can be time-consuming and costly. Assessments can become prohibitively expensive when there are a large variety of jobs at a given workplace or when job functions are remote, unobservable, highly variable, or infrequent. Moreover, this kind of time-intensive professional observation is impractical for the personalized, continuous monitoring of injury risk over long durations or across an entire workforce. Video-based solutions that leverage advances in computer vision and machine learning have the potential to address some of these challenges by providing a semi-automated analysis of jobs. However, this approach is impractical for highly dynamic jobs (e.g., a construction worker moving all over a construction site), or jobs where visual obstructions occur (e.g., an aerial porter climbing in and out of arriving planes) and is not intended for personalized monitoring across an entire workforce. To efficiently evaluate ergonomic risk across a wide range of workers, high-risk jobs, and workplace environments, there remains a need for tools that enable the automated, unconstrained, and widespread monitoring of musculoskeletal loading and damage, particularly to the lower back.

5.2.2 Wearable sensors at a single body location for ergonomic assessment or continuous monitoring

Small, inexpensive, wearable sensors offer a promising solution for the unconstrained monitoring of job demands, including in confined spaces or during dynamic jobs. Wearable solutions could automate traditional job analysis or ergonomics assessments by replacing time-consuming observations and manual measurements with automated analytics from wearable sensor data, potentially improving the quality (e.g., consistency, accuracy) and quantity of data (e.g., the amount of assessment time per worker, the number of workers evaluated). Further, wearables can be practical for continuous monitoring, providing new opportunities to perform ergonomics assessments for remote and long-cycle duration jobs or for personalized, daily injury risk monitoring that could inform ergonomic controls.

Continuous monitoring also has the long-term potential to usher in a new era of preventative occupational safety and health that transforms how musculoskeletal risk is managed and insured.

While wearable sensors offer an exciting tool for monitoring low back loading and overexertion risks, current commercial and research technologies have some key limitations. Current commercial products (e.g., StrongArm Fuse, Soter Analytics Clip&Go, Kinetic Reflex, and Modjoul Smartbelt) use a single inertial measurement unit (IMU) mounted on the waist, back, or chest and analyze motion data (e.g., trunk orientation or acceleration) and the frequency of lifting/bending. We refer to these types of devices as wearable sensors at a single body location (or single wearable solutions, for short). We use this terminology because they each use hardware placed on one body location, although this hardware unit may contain multiple different sensors that measure numerous signals (e.g., IMUs are generally composed of accelerometers, gyroscopes, and magnetometers).

These single wearable solutions are relatively practical to implement in the workplace and may be most amenable to job analyses that characterizes postures and task frequency, but less well-suited for ergonomic assessments that quantitatively assess injury risk based on musculoskeletal loading and fatigue failure principals. This is because low back loading is dependent on factors beyond the kinematics of a single body segment, including the mass of the object being lifted and how far away the object is from the body. So, while these single wearable solutions can use segmental motion data to identify when a worker performs a deep forward bend, they are generally unable to distinguish, for instance, if the worker lifted a 5 lb vs. a 45 lb box. The heavier mass in this example is expected to result in 65× more tissue damage (based on LiFFT, and assuming boxes are located 25 inches anterior to the lumbar spine). There are some use cases where single wearable solutions are expected to estimate low back loading fairly well (e.g., if the objects lifted are of known mass and are in a fairly consistent location relative to the body). However, there are cases where single wearable solutions are likely to be insufficient because they do not account for varying object masses, object locations, or other external forces on the body. In these cases, single wearable solutions could potentially provide inaccurate or misleading information about loading and cumulative damage to the low back, or unreliable insight on low back injury risk for a specific job, task, or worker.

5.2.3 Distributed wearable sensors for ergonomic assessment or continuous monitoring

Using multiple wearable sensors at distributed locations on the body has the potential to provide better estimates of low back loading by capturing and integrating additional dynamics data (e.g., body segment motions or orientations, forces or moments, muscle activity). These distributed sensor solutions are conceptually similar to what is done in motion analysis labs when data from cameras, force plates, and/or other measurement modalities are combined with biomechanical models to compute the loading on the back. An example of a commercial distributed wearable sensor system is the Xsens system that uses up to 17 IMUs on different body segments to track motion. These data can then be passed through analytics software (e.g., Scalefit, AnyBody) to estimate musculoskeletal loading on the back. However, similar to single sensor solutions, distributed IMU systems cannot automatically distinguish the mass of the object being lifted. Often, this additional information must be entered manually, or additional sensor modalities must be added, which increases the complexity of data collection and analysis. Thus, distributed IMU systems may only partially automate ergonomic assessments, or they may provide inaccurate estimates of low back loading if analytics software simply assume a default object mass.

To fully automate back load monitoring, several research studies suggest that adding force-instrumented shoes or pressure insoles along with distributed IMUs over the upper and/or lower body is promising [21], [191]–[193]. For instance, [21] showed that by combining 8–17 IMUs and force-sensing shoes, lumbar moments could be estimated within 10–20% of peak extension moments. [191] found corroborating results, showing that with 12 IMUs and pressure insoles the peak axial load on the L5/S1 joint could be estimated with errors <5%.

However, there are a couple of critical limitations of these approaches. First, many of these solutions were developed and evaluated on a limited range of manual material handling tasks. For instance, [21] only evaluated four lifting tasks, all with a 10 kg box. It therefore remains unclear if these combinations of wearable sensors and/or algorithms are accurate and generalizable to a broad range of complex manual material handling motions performed in real world environments. Second, these wearable solutions require a large number of sensors distributed across the body, which introduces practical challenges related to technology implementation, ease of use, acceptance, and adoption. For scientific research or infrequent ergonomic assessment, the burden of distributed instrumentation may be an acceptable trade-off for increased accuracy. However, using numerous sensors requires longer donning and doffing times and more complexity, which presents a pragmatic barrier for workplace adoption. To enable more efficient and widespread ergonomic assessments or continuous monitoring of injury risk, there remains a need for a solution that requires a smaller number of wearable sensors (to be practical) and provides validated estimates of low back loading for a wide range of work-relevant tasks (to ensure accuracy).

5.2.4 Key requirements for fully automated ergonomics assessment or continuous monitoring

Based on our review of commercial technologies and scientific literature and our conversations and observations with manual material handlers and safety professionals across a range of industries (e.g., logistics, manufacturing, retail, agriculture, construction, military), we identified what we believe to be a key technological gap and unmet industry need related to ergonomic assessment and continuous personal monitoring of low back overexertion injury risk. Specifically, we found that a portable wearable sensor tool with the following characteristics and capabilities does not currently exist, but if it did we believe it could be game-changing for low back injury risk assessment, monitoring, and prevention in various industries:

1. The tool is **practical** to don, doff, and wear in unconstrained environments for prolonged periods of time by virtue of using only a small number of sensors at different body locations. This is important for industrial acceptance, adoption, and implementation. Of note, there is no simple limit for the maximum number of sensors or body locations that is practical, but this consideration helped motivate the approach we took in this research (as detailed in Methods).
2. The tool provides **accurate**, validated, and automated estimates of low back loading for a broad range of manual material handling tasks. This is important to ensure the system will be reliable during use in the real world and can distinguish differences in back loading that result from lifting objects of different weights without the need for professional observation or manually inputting object weights or other data.

The overarching question we sought to address in this study was: if we can only use a small number of wearable sensors to monitor low back loading, then which sensors should we use, where should we place them, what type of algorithm should we employ to fuse the sensor data, and how accurately can

we monitor low back loading during manual material handling tasks? To address this exploratory, multi-faceted, open-ended question, we collected synchronized data from laboratory instrumentation and wearable sensors across a broad range of lifting tasks and combined domain expertise in biomechanics with techniques from machine learning to develop musculoskeletal load estimation algorithms, similar to the approach we previously took to develop a wearable sensor system for monitoring bone loading and overexertion injury risks in the legs of runners [194]. The Methods section provides details on our exploratory approach and rationale.

5.3 Methods

5.3.1 Summary of approach

Here, we briefly summarize our exploratory research approach, followed by detailed methodology below:

First, we identified a candidate set of wearable sensors (number, type, and location of sensors). We bounded our candidate sensors based on biomechanical insight, prior literature [21], [191]–[193], and expected practicality for implementation in the real world. We selected IMUs placed on body segments (feet, shanks, thighs, pelvis, trunk) and pressure insoles placed inside the shoes (capable of estimating the interaction force and center of pressure between the foot and shoe) as our candidate sensors. These types of sensors are mature, and for years have been used in clinical and consumer devices that are worn daily; for instance, IMUs are ubiquitous in fitness trackers and phones, and pressure insoles are used for clinical screening (e.g., Orpyx) and to track running/sport performance (e.g., ARION, ReTiSense, NURVV). We elected not to use surface electromyography (EMG) due to practical challenges of implementing in the real world, such as their sensitivity to sweat, hair, and sensor placement, and reliability issues over days/weeks [195]. We also elected not to use any implanted or percutaneous sensors, or any emerging sensor technologies that have not yet been proven to be practical, reliable, affordable, and scalable in the real world. Focusing on mature, proven sensor technologies was with the hope and intention of arriving at a solution that would be feasible to translate into a product for real world use in the near future (e.g., next 2–5 years).

Second, we synchronously collected data from lab-based instrumentation and from real wearable sensors across 10 participants each performing about 400 different manual material handling tasks, which encompassed many different postures, movements, and object masses that a worker may encounter in the real world.

Third, we developed wearable sensor algorithms using various combinations of wearable sensor signals (algorithm inputs) and our lab-based gold-standard estimates of low back loading (algorithm target). We first used *idealized wearable sensor signals* [194], which consisted of lab-based data we converted into the types of signals reasonably obtained with wearables, to develop and evaluate algorithms. An example of an idealized wearable sensor signal is that we mapped the three-dimensional ground reaction force (GRF) vector from an in-ground force plate onto a one-dimensional force normal to the bottom of the foot to represent the type of signal that can be estimated from a pressure insole. This allowed us to explore algorithms for low back load estimation without worrying if the sensor or signal quality from a particular wearable sensor we used was a limiting factor. Next, we used *real wearable*

sensor signals to separately develop and evaluate algorithms, benchmark the accuracy of current wearable sensor technologies, and assess how these may or may not limit low back load monitoring tools. We use the terminology *idealized wearable sensor signals* and *real wearable sensor signals* to distinguish these two complementary approaches. Throughout, we also use the terms *idealized wearable sensors* and *real wearable sensors* to refer to physical sensors or sensor combinations, with *idealized wearable sensors* referring to the sensor that would be needed to measure the particular signals used. See [194] (as well as the Discussion of this paper) for more rationale on the value of using idealized wearable sensor signals when exploring new solutions for musculoskeletal load monitoring.

Finally, by applying various machine learning techniques to various subsets of idealized and real wearable sensor signals, we: (1) quantified how the number of sensors used influenced the algorithm estimation accuracy, (2) identified the most important types and locations of sensors for low back load estimation, and (3) benchmarked how much using real vs. idealized wearable sensor signals influenced the estimation accuracy. Below, we describe the human participant experiment and data analysis, followed by algorithm exploration, development, and evaluation.

5.3.2 Experiment overview

Ten healthy individuals participated in the study: 3 females and 7 males (age: 25 ± 3 years; height: 1.8 ± 0.1 m; mass: 79 ± 14 kg). All the participants gave written informed consent to the protocol, which was approved by the Institutional Review Board at Vanderbilt University (IRB # 141697).

This study involved participants each performing about 400 manual material handling tasks in a motion analysis lab. Tasks covered a broad range of bending, turning, twisting, squatting, stooping, and reaching postures while lifting and moving boxes of 5–23 kg, which were representative of tasks commonly performed by manual material handlers (e.g., case pickers in a warehouse, retail workers stocking shelves, or logistics workers at a sort facility). For instance, tasks involved moving boxes from high to low shelves, low to high shelves, from a lateral to a forward position, diagonally between shelves, and much more to obtain a rich, diverse, realistic, and work-relevant data set (see Video S1 for example videos of tasks). The data collection space was outfitted with various shelves at 3 heights with labeled locations (see Figure 1A for an example setup). Box masses, shelf heights, and actions were informed by manual lifting and ergonomics guidelines [196]. For each task, participants were given instructions such as “move the box from position 3 to 4” (Figure 1A) and were told to use any safe strategy to complete the task. Each task was performed once and the participants were given rest breaks intermittently throughout the protocol.

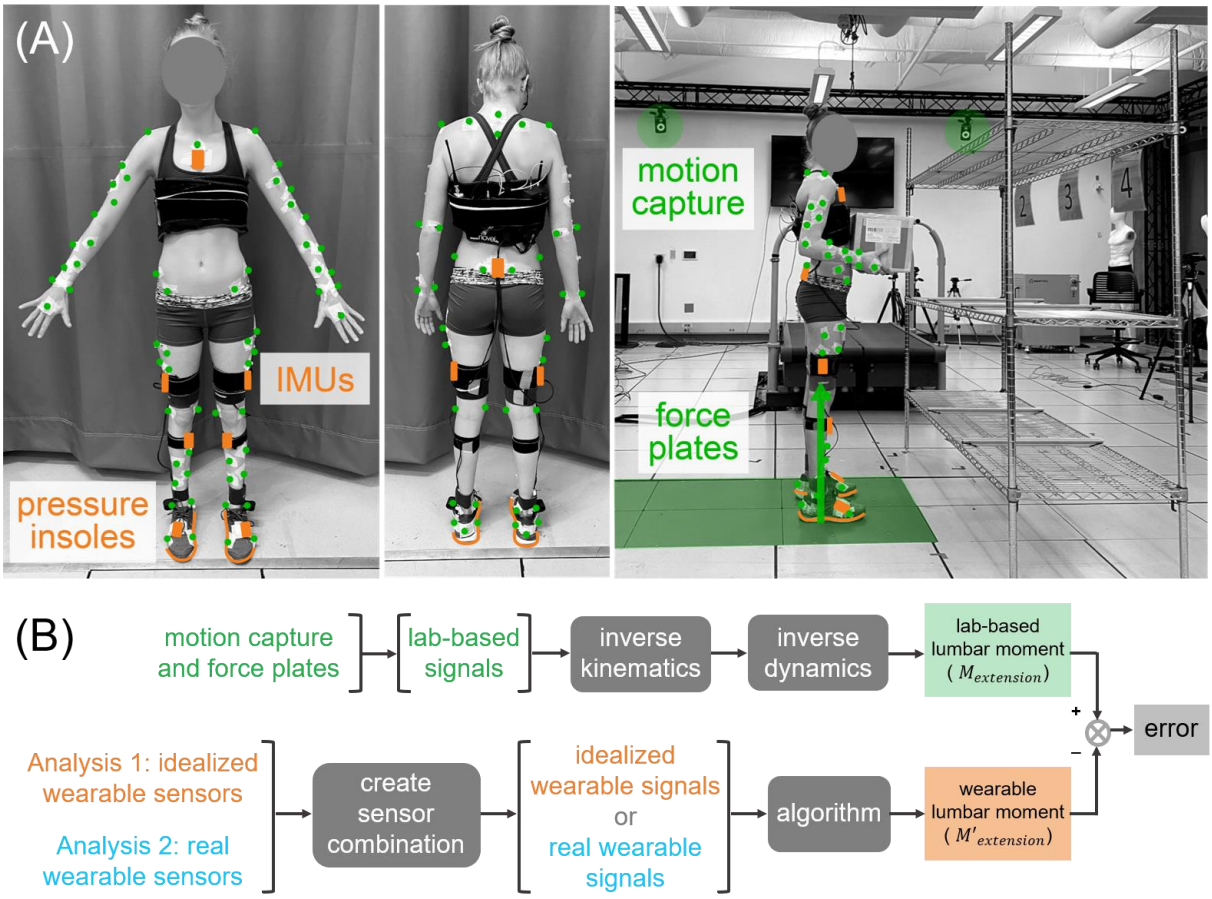


Figure 5.1 Experimentation and wearable algorithm development overview. A) Lab-based (green) and wearable sensor (real: orange, idealized: blue) measurements were collected synchronously in a motion analysis lab while participants performed about 400 manual lifting tasks. **B)** Lab-based analysis yielded a gold-standard estimate of lumbar extension moment ($M_{extension}$). Wearable analysis and algorithm development yielded wearable sensor estimates of lumbar extension moment ($M'_{extension}$). The wearable algorithm development was done twice, once using idealized wearable sensors as inputs (Analysis 1) and once using real wearable sensors as inputs (Analysis 2).

5.3.2.1 Lab-based measurement modalities

We collected full body kinematics and ground reaction forces (GRFs). Kinematics were collected at 200 Hz (Vicon), then low pass filtered at 6 Hz (3rd order, zero-lag Butterworth). Four markers were placed on each thigh, shank, arm, and forearm, 5 markers were placed on each foot, 6 markers were placed on the pelvis, and 4 were placed on the trunk. Additional markers were placed on the lateral and medial femoral epicondyles, the lateral and medial malleoli, each acromion, the lateral and medial humeral epicondyles, and the distal radius and ulna. The GRFs under each foot were collected at 1000 Hz using in-ground force plates (AMTI). The GRFs were low-pass filtered at 10 Hz (3rd order, zero-lag Butterworth).

5.3.2.2 Wearable measurement modalities

We synchronously collected IMU-based lower body and trunk kinematics (Xsens) and plantar pressures (Novel pedar-x, with 99 pressure sensors per insole). Kinematics were collected at 100 Hz using the standard Xsens "lower body + trunk" configuration. Scaling, calibration, and data pre-processing were performed by the Xsens software, providing a built-in anatomical model. Plantar pressures were collected bilaterally at 100 Hz and the total (normal) force and center of pressure were exported using the Novel software. Synchronization of all measurement modalities was achieved through recorded analog triggers, and any delays between measurement modalities were accounted for through temporal alignment/calibration algorithms based on pilot testing.

5.3.3 Wearable algorithm development

A visual overview of the lab-based data analysis and algorithm evaluation workflow is provided in **Figure 5.1B.**

5.3.3.1 Lab-based data analysis (algorithm target)

We selected the lumbar extension moment as our target musculoskeletal loading metric because it can be used to estimate cumulative tissue damage to the low back using a fatigue failure analysis [9], [180], [189]. We sought to estimate the time series lumbar extension moment (as opposed to just peak moment) because this enables us to identify bending/lifting frequency, to partition out individual movement cycles, and to better understand and distinguish cyclic lifts vs. prolonged bending. Time series data enables the assessment of loading and cumulative risk across all tasks, as well as the ability to perform task-specific load and risk assessment.

Lower-body segmental and joint kinematics were estimated based on optical motion capture data and rigid-body inverse kinematics. GRF and kinematics were combined via rigid-body inverse dynamics to estimate joint kinetics (C-Motion, Visual3D). Time series lab-based lumbar moment was estimated using bottom-up inverse dynamics in Visual3D. Moments are reported in units of body weight × body height (BW × BH).

5.3.3.2 Wearable sensor signal data preparation (algorithm inputs)

We used time series wearable sensor signals as inputs to the algorithm. *Idealized wearable sensor* signals are summarized in **Table 5.1**. *Real wearable sensor* signals are summarized in **Table 5.2**. The algorithm development workflow was completed twice, once using idealized wearable sensor signals as the inputs, and once using real wearable sensor signals as the inputs (Analysis 1 and Analysis 2, **Figure 5.1B**). The lab-based target, idealized wearable sensor signals, and real wearable sensor signals were all resampled to 100 Hz. Input signals were normalized to z-scores during algorithm development [170].

Table 5.1 Idealized wearable sensor signals.

Idealized wearable sensors	Idealized wearable sensor signals	# of signals
8 idealized IMUs (trunk, pelvis, R/L thigh, R/L shank. R/L foot) Segments (8): pelvis, trunk, R/L thigh, R/L shank. R/L foot Joints (7): lumbar, R/L hip, R/L knee, R/L ankle	XYZ motion capture segment kinematics (Euler angles)	24
	XYZ motion capture joint kinematics	21
Idealized pressure insoles	3D force plate GRF transformed into foot's coordinate frame and projected in 1D normal force	2
	force plate center of pressure transformed into foot's X/Y coordi nate frame	4
	Total	**51**

Table 5.2 Real wearable sensor signals.

Real wearable sensors	Real wearable sensor signals	# of signals
8 IMUs (sternum, pelvis, R/L thigh, R/L shank. R/L foot) Human model: Segments (11): pelvis, L5, L3, T12, T8, R/L thigh, R/L shank. R/L foot Joints (10): L5S1, L4L3, L1T12, T9T8, R/L hip, R/L knee, R/L ankle	XYZ segment kinematics (Euler angles)	33
	segment kinematics (quaternions)	44
	XYZ segment velocities	33
	XYZ segment accelerations	33
	XYZ joint kinematics	30
Pressure insoles	Total normal force	2
	X/Y center of pressure	4
	Total	**179**

5.3.3.3 Algorithm Development

We explored supervised machine learning algorithms (e.g., generalized linear models, support vector machines, neural networks) for multiple variable regression to predict the lumbar extension moment ($M_{extension}$) focusing on techniques that could provide instantaneous predictions, where wearable signals from a given time sample are used to estimate the target load metric for that same time sample. Ultimately, we achieved the most promising results with Gradient Boosted Decision Trees, a popular technique in machine learning and well-suited to handle missing values and redundant or non-predictive inputs [197], [198]. The number of input signals (tens or hundreds) also fits this approach. Furthermore, by using a histogram-based decision tree building algorithm influenced by LightGBM [199], we dramatically decreased the algorithm training time (to a few seconds with a few million time samples) without a noticeable degradation in the prediction accuracy. Briefly, this algorithm estimates the target load metric by building an ensemble of decision trees in a stage-wise fashion, where in each stage the new tree tries to estimate (and thus, remove) the residual error after combining the predictions of the previous trees. Our current results are based on ensembles of approximately 100 trees. We used the scikit-learn library and Amazon SageMaker, a cloud-based machine learning platform for algorithm development, model training, and evaluation.

To develop the algorithm, we used k-fold validation by participant (n = 10), a commonly used technique to assess the generalizability of an algorithm [170]. In other words, we used data from nine participants to train the algorithm (i.e., select hyperparameters), and then evaluated the algorithm accuracy on data from the remaining participant. This process was repeated for all ten participants to yield wearable algorithm estimates of the lumbar extension moment ($M'_{extension}$) for the entire dataset.

The algorithm workflow was first performed using all our candidate wearable sensor signals; we termed this the distributed sensor algorithm. Next, to evaluate the feasibility of using a reduced number of sensors for estimating lumbar moments, we developed additional algorithms using a reduced number of sensor signals (termed reduced sensor algorithms). While we explored 10 candidate wearable sensors (R/L pressure insoles, R/L foot IMUs, R/L shank IMUs, R/L thigh IMUs, pelvis IMU, trunk IMU), when iterating through potential reduced sensor algorithms we assumed that a final solution would have symmetrical bilateral sensors (e.g., if the wearable included a right insole, then it would also include a left insole). Thus, our 10 candidate wearable sensors actually corresponded to 6 candidate sensor locations: trunk, pelvis, thigh, shank, and foot IMUs, and the pressure insoles. The algorithm workflow was repeated to develop 62 additional algorithms that each used a reduced set of 1 to 5 sensor locations (see **Supplemental Figure 5.1** for an overview of all combinations).

5.3.3.4 Algorithm Evaluation

We evaluated the accuracy of different sensor combinations in two stages. First, we computed the coefficient of determination (r^2) to identify the most promising reduced sensor combinations and computed relative wearable sensor signal importance to identify the most important sensors. Then, we identified promising or interesting sensor combinations, reviewed wearable algorithm results using scatter plots and participant-specific results, and computed additional accuracy metrics to better understand the performance and limitations of each sensor combination.

We computed r^2 for each participant across all time samples [200] for all candidate sensor combinations. Based on our prior work on wearables for musculoskeletal load monitoring [173], [194], we have found r^2 to be useful for this initial sensor combination selection process (i.e., down selection from 62 sensor combinations here) because it provides insight into how well wearable estimates correlate with lab-based gold-standard estimates across the full range of lumbar moments observed. This research is early stage, so there is no precise r^2 threshold that we can define as the minimum viable, but to benchmark high algorithm accuracy we used $r^2 > 0.8$ as a threshold for promising solutions.

As a complementary analysis to evaluate which sensors were most important for algorithm estimates, we applied the permutation feature importance method [201]. Feature importance values represent the drop in model accuracy (Δr^2) when an input signal is randomly shuffled, with larger values indicating that the algorithm is more dependent on that signal. Of note, the permutation feature importance method was used rather than the impurity-based feature importance approach because the latter approach had some undesirable biases (e.g., favoring high cardinality features) and is not supported with histogram-based estimators.

Once a subset of promising sensor combinations was identified, we inspected participant-specific results with scatter plot data to understand the performance and limitations of each. We were particularly interested in how each sensor combination performed across the range of lumbar moment magnitudes observed (e.g., did certain sensor combinations perform better at low magnitudes vs. high magnitudes). We also computed the root mean square error (RMSE). In this data set, most samples are at relatively low lumbar moment magnitudes, but larger moments are the most damaging and dangerous to musculoskeletal tissues. We therefore also looked specifically at algorithm performance constrained to higher lumbar moments using mean absolute percent error (MAPE). We leveraged the benefits of both relative (r^2 and MAPE) and absolute (RMSE) accuracy metrics, along with biomechanics knowledge of

key factors that influence cumulative damage and overexertion injury risk, to make informed suggestions about using wearable sensors to monitor back loading across work-relevant lifting tasks.

5.4 Results

5.4.1 Results from idealized wearable sensors

As expected, maximum algorithm accuracy increased with the number of sensor locations (**Figure 5.2**). There were no single sensor solutions that yielded $r^2 > 0.8$ (i.e., coefficient of determination greater than 0.8 between idealized wearable sensor algorithm estimates and lab-based lumbar moment estimates). However, there was a noticeable jump in accuracy when moving from one to two sensor locations (maximum $r^2 = 0.74$ to $r^2 = 0.89$, **Figure 5.2**, **Table 5.3**). When increasing the number of sensors beyond two locations there were only small additional improvements in maximum algorithm accuracy (from $r^2 = 0.89$ using two sensor locations to $r^2 = 0.92$ using all six sensor locations, the maximum number of distributed sensor locations in this study).

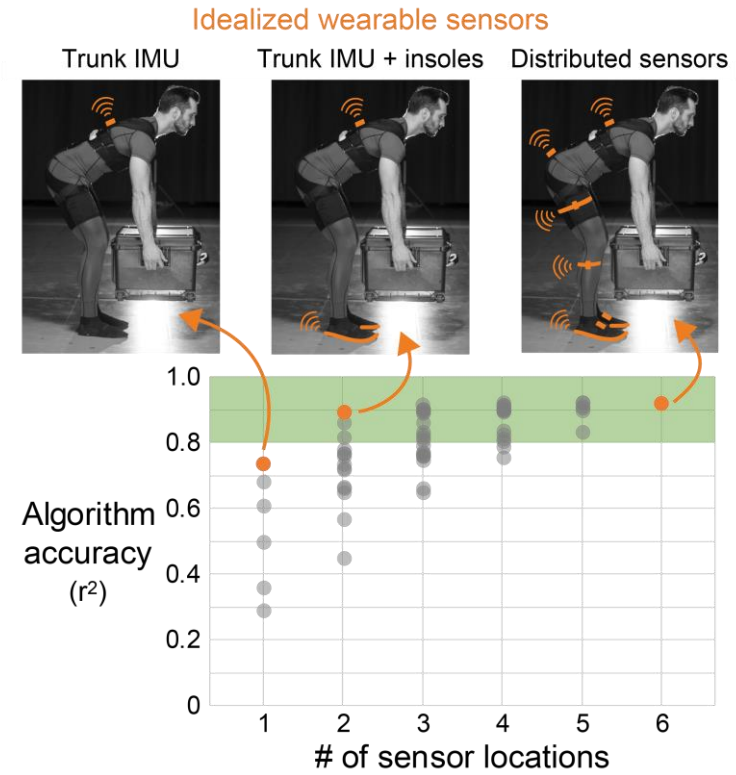


Figure 5.2 Maximum algorithm accuracy increased with number of sensor locations. Average accuracy using idealized wearable sensors summarized here using the average coefficient of determination (r^2) across all participants. Orange dots represent the distributed sensor algorithm (right) and the highest accuracy algorithms using 1 (left) and 2 (center) sensor locations. All algorithms here were developed with idealized wearable sensor signals and the target was lumbar extension moment. The top three algorithms using one and two sensor locations are reported in **Table 5.3**. A detailed summary of all algorithm accuracies and the exact sensor combinations for each algorithm is included in **Supplemental Figure 5.1**.

Table 5.3 Algorithm accuracy for a subset of idealized wearable algorithms. Average accuracy for the distributed sensor algorithm and the top three algorithms requiring one or two sensor locations. Accuracies reported here correspond to data points in **Figure 5.2** and **Supplemental Figure 5.1**.

# of sensor locations	idealized wearable sensor combination	algorithm accuracy (r^2)
6	distributed sensors	0.92
2	trunk IMU + insole	0.89
2	thigh IMUs + insole	0.86
2	pelvis IMU + insole	0.81
1	trunk IMU	0.74
1	thigh IMUs	0.68
1	pelvis IMU	0.61

The two most important signals for estimating lumbar extension moments identified during algorithm development were sagittal trunk angle and vertical GRFs (**Figure 5.3**). Consistent with this, the best solution using two sensor locations is the one that combined a trunk IMU and pressure insoles (r^2 = 0.74, **Figure 5.2**, **Table 5.3**). This combination was of highest interest to us because of its potential to be practical and accurate.

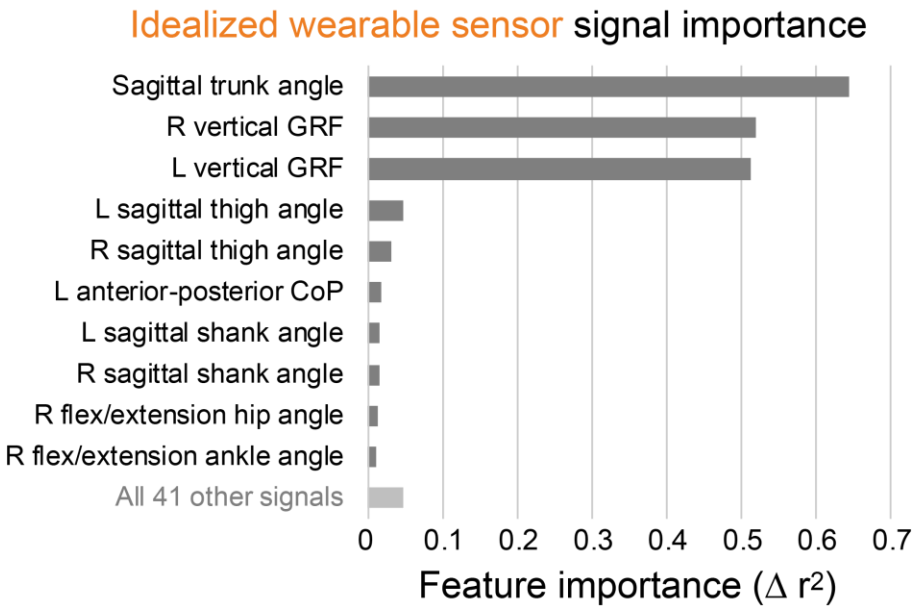
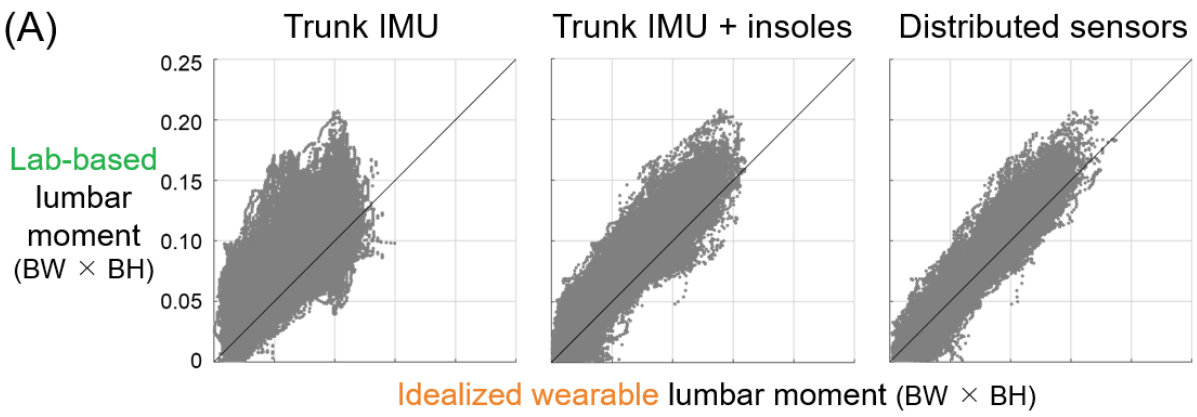


Figure 5.3 Sagittal trunk angle and vertical GRFs are the most important signals for estimating lumbar moments. Signal importances are from the idealized wearable sensor algorithm for estimating lumbar extension moments. R = right; L = left.

The trunk IMU (alone) and fully distributed sensor sets were also of interest for further analysis. The trunk IMU provides a point of reference for the potential accuracy of existing commercial wearables that use a single IMU to monitor lumbar loading, while the distributed sensor set provides insight on accuracy gains with higher instrumentation coverage. Therefore, we report participant-specific results and additional accuracy summary metrics (RMSE and MAPE) for these three different sensor combinations (**Figure 5.4**).



(B) Algorithm accuracy (r^2)

Participant	Trunk IMU	Trunk IMU + insoles	Distributed sensors
1	0.69	0.90	0.90
2	0.80	0.93	0.95
3	0.65	0.88	0.90
4	0.70	0.93	0.92
5	0.79	0.92	0.95
6	0.77	0.87	0.93
7	0.70	0.77	0.86
8*	0.75	0.89	0.93
9	0.75	0.93	0.94
10	0.76	0.90	0.94
avg ± std (avg RMSE)	**0.74 ± 0.05** (31 Nm)	**0.89 ± 0.05** (20 Nm)	**0.92 ± 0.03** (17 Nm)

Figure 5.4 Algorithm accuracies for three different idealized wearable sensor algorithms. A) Lab-based (gold standard) lumbar extension moment vs. idealized wearable algorithm estimates of lumbar moment for all time samples for an example participant (participant 8*). Positive moments correspond to lumbar extension moments. A line with a slope of one is added to visualize a perfect correspondence between lab-based and wearable estimates. BW × BH = body weight × body height. **B)** Coefficient of determination (r2) for each participant. Average results (avg, bottom) are equivalent to accuracies in Figure 2. The trunk IMU algorithm was less accurate than the trunk IMU plus pressure insoles algorithm, and then the distributed sensors algorithm ($p<0.001$ and $p<0.001$, respectively, based on Wilcoxon signed-rank test of the k-fold cross validation accuracy results). Comparing accuracy from the trunk IMU and pressure insoles algorithm vs. the distributed sensors algorithm yielded $p=0.054$. Average RMSE was converted into units of Nm (using mean participant height and weight) and included for reference.

The distributed sensor algorithm resulted in an average RMSE of approximately 17 Nm (**Figure 5.4B**), equivalent to about a 241 N (0.3 BW) error in spine compression force (assuming a 7 cm lumbar muscle extensor moment arm, [202]). The trunk IMU and pressure insole algorithm resulted in an average RMSE of approximately 20 Nm, equivalent to about a 282 N (0.4 BW) error in spine compression force. The trunk IMU algorithm resulted in an average RMSE error of approximately 31 Nm, equivalent to about a

444 N (0.6 BW) error in spine compression force. As one additional point of reference, the NIOSH Lifting Equation recommends limiting spine compression force to less than 3400 N (4.4 BW), so these RMSE values are about 7%, 9%, and 14% of this limit, respectively. Given the sensitivity of MAPE when target values are close to zero, we also computed the MAPE for all samples when the target load metric was greater than 0.05 BW*BH (which encompassed about half of all time samples of data for each participant). Average MAPE for the upper range of lumbar moments was 13%, 15% and 25% for the distributed sensor, trunk IMU and pressure insole, and trunk IMU algorithms, respectively.

We also observed that if the trunk IMU were substituted with thigh IMUs, then correlations only decreased slightly from $r^2 = 0.74$ to $r^2 = 0.68$ with a single sensor, and from $r^2 = 0.89$ to $r^2 = 0.86$ for the two sensor combination (**Table 5.3**). If the trunk IMU were substituted with a pelvis IMU, then correlations decreased slightly more from $r^2 = 0.74$ to $r^2 = 0.61$ with a single sensor, and from $r^2 = 0.89$ to $r^2 = 0.81$ for the two sensor combination (**Table 5.3**). All of the two sensor location solutions that achieved $r^2 > 0.8$ included GRFs from pressure insoles.

Participant-specific results (**Figure 5.4**) corroborated and strengthened the average results (**Figure 5.2**, **Figure 5.3**, **Table 5.3).** For instance, all ten participants exhibited high algorithm accuracies (r^2 ranging from 0.86 to 0.95) using the distributed (six sensor location) algorithm. When moving from a single trunk IMU to using a trunk IMU and pressure insoles, every participant exhibited an increase in r^2 value (**Figure 5.4**). Scatter plot data for each participant indicated that the improvement in r^2 going from one to two sensor locations was driven by both a decrease in the variation of data about the unity regression line and improved estimates at higher magnitude lumbar moments (see example participant data in **Figure 5.4A**). When moving from two to six sensor locations the variation of data about the regression line decreased more, but only slightly (**Figure 5.4A**). We also note that for two participants (numbers 1 and 4), going from two to six sensor locations did not increase r^2 at all (**Figure 5.4B**).

5.4.2 Results from real wearable sensors

Figure 5.5 is analogous to **Figure 5.3,** and **Figure 5.6** is analogous to **Figure 4**, except that **Figure 5.5** and **Figure 5.6** are based on real wearable sensors rather than idealized wearable sensors. Real wearable sensor results confirm that the most important sensor signals for estimating lumbar extension moments are sagittal trunk angle from a trunk IMU and vertical GRFs from pressure insoles (**Figure 5.5**). However, it is noteworthy that trunk angle signal importance was much higher than vertical GRFs in the analysis of the real wearable sensors (**Figure 5.5**), whereas with idealized signals these signal importances were of similar magnitude (**Figure 5.3**).

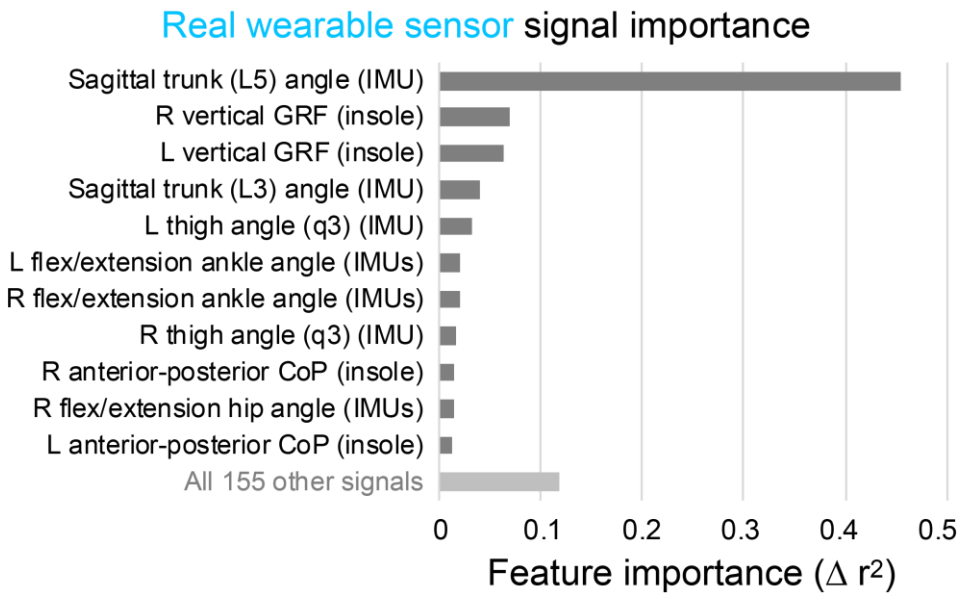
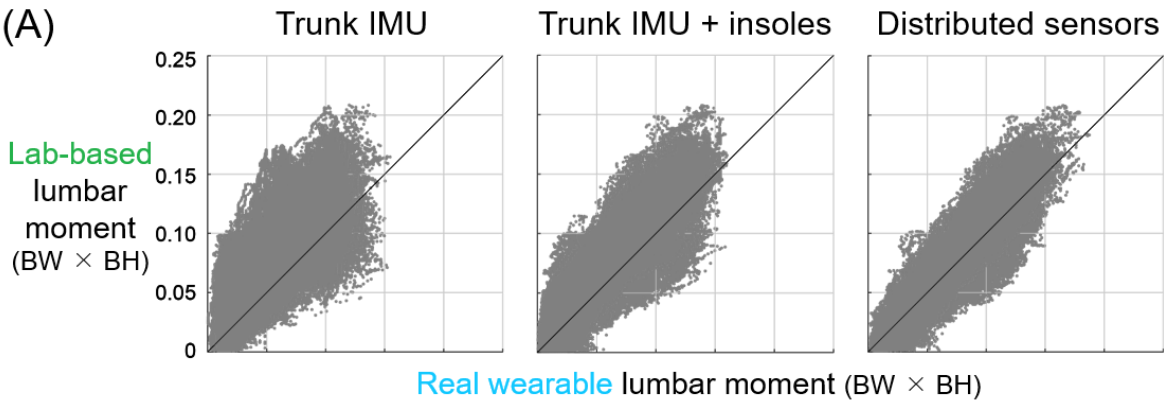


Figure 5.5 Sagittal trunk angle and vertical GRFs are the most important signals for estimating lumbar moments. Signal importances are from real wearable sensor algorithm for estimating lumbar extension moments. These results are consistent with the findings from idealized wearable sensor analysis in **Figure 5.3**. R = right; L = left.

Participant-specific results (**Figure 5.6**) again corroborated and strengthened the average results from real wearable sensor algorithms. Compared to idealized wearable sensor algorithms, there was no discernible increase in r^2 value when moving from one sensor location (trunk IMU, $r^2 = 0.79$) to two sensors locations (trunk IMU and pressure insoles, $r^2 = 0.80$). The increase in r^2 from two to six sensor locations also remained relatively small, similar to what was observed in the idealized wearable sensor analysis (**Figure 5.4**).



(B) Algorithm accuracy (r^2)

Participant	Trunk IMU	Trunk IMU + insoles	Distributed sensors
1	0.74	0.74	0.80
2	0.83	0.76	0.88
3	0.77	0.80	0.76
4	0.79	0.81	0.88
5	0.85	0.84	0.91
6	0.82	0.82	0.85
7	0.71	0.79	0.83
8*	0.80	0.83	0.88
9	0.79	0.74	0.84
10	0.79	0.81	0.84
avg ± std (avg RMSE)	**0.79 ± 0.04** (28 Nm)	**0.80 ± 0.04** (27 Nm)	**0.85 ± 0.04** (24 Nm)

Figure 5.6 Algorithm accuracies for three different real wearable sensor combinations. A) Lab-based (gold standard) lumbar moment vs. real wearable sensor algorithm estimates lumbar moment for all time samples for an example participant (participant 8*). Positive moments correspond to lumbar extension moments. Moment in units of bodyweight × bodyheight (BW × BH). A line with a slope of one is added to visualize a perfect correspondence between lab-based and wearable estimates. **B)** Coefficient of determination (r^2) for each participant. Both reduced sensor algorithms yielded accuracies that were lower than the distributed sensor combination (p=0.011 and p=0.014 for the trunk IMU, and trunk IMU and insole algorithms, respectfully; based on Wilcoxon signed-rank test of the k-fold cross validation accuracy results), but the accuracy of the two reduced sensor algorithms shown here were not different from each other (p=0.571).

5.4.3 Comparison of results from idealized vs. real wearable sensors

Figure 5.7 provides a side-by-side comparison of algorithm performance using idealized versus real wearable sensor signals. These plots are visualizations of the tabular results reported in **Figure 5.4B** and **Figure 5.6B**, and provided for clarity and to assist with interpretation. The key takeaway is that while the idealized wearable sensor analysis resulted in a noticeable jump in accuracy when moving from one to two sensors, a similar improvement was not observed in the real wearable sensor analysis (**Figure 5.7**). The Discussion section digs into why.

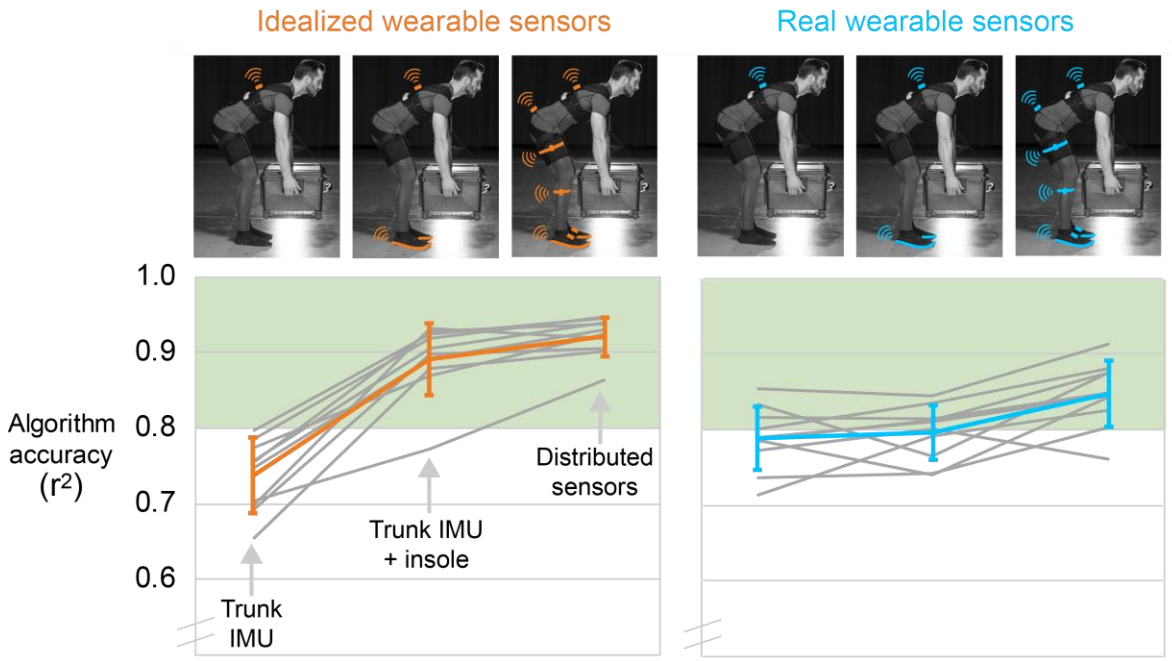


Figure 5.7 Side-by-side comparison of algorithm performance using idealized versus real wearable sensor signals. Gray lines are each of the 10 participants' accuracy results, and colored lines are the average (and standard deviation) across the 10 participants. Results using idealized wearable sensors are shown on the left (orange) and results using real wearable sensors are shown on the right (blue).

5.5 Discussion

These findings indicate that there is strong potential to use a small number of wearable sensors to create a portable tool for the practical and accurate monitoring of low back loading over a broad range of manual material handling tasks. We characterized the performance of over 60 different wearable sensor combinations and algorithms. The solution we found to be most promising combines signals from sensors at two body locations (an IMU on the trunk and pressure insoles under the feet) with a Gradient Boosted Decision Tree algorithm. While idealized wearable sensor results demonstrated promising proof-of-concept, the analysis of real wearable sensor signals revealed that to achieve accurate lumbar moment estimates in the real world, the key technological challenge will be to optimize force estimates and minimize variability from the pressure insoles. With further development and validation, we believe that this type of wearable solution has the potential to transform how ergonomic assessments are

performed in industry, to enhance the quality, quantity, and efficiency of occupational data collection, and to expand opportunities for personalized, continuous monitoring of low back injury risk. For example, time-series lumbar moments could be partitioned into individual lift/bend cycles and the magnitude and frequency of loading on the low back could be automatically input into ergonomic assessment tools like LiFFT to estimate injury risk. Below we discuss the major technical findings from this exploratory research, along with alternative solutions, key challenges, and new opportunities for advancement.

5.5.1 Which wearable sensors and locations are most important?

The trunk IMU and pressure insoles were identified in all analyses as together being the most important sensors for monitoring lumbar extension moments (**Figure 5.2, Figure 5.3, Figure 5.5, Table 5.3**). These results match our biomechanics intuition given that lumbar moment is strongly influenced by the weight of the object being lifted (which can be captured by pressure insoles) and by upper-body posture (which can be estimated with an IMU on the trunk).

Interestingly, the trunk IMU could be replaced with thigh IMUs or a pelvis IMU with relatively little degradation in accuracy (**Table 5.3**). Of note, the reason that thigh and pelvis IMU signals appear to have low importance in **Figure 5.3** and **Figure 5.5**, but can actually be useful substitutes for the trunk IMU, is because they are highly correlated with other signals and because of how the feature importance method works (see Methods). It is valuable to acknowledge these other alternatives because some sensor locations may be preferred for certain applications; for instance, a fall protection harness manufacturer may be able to integrate an IMU more easily on the trunk near the D-ring or on the thighs using the leg loops, whereas for a tool belt manufacturer it may be preferable to integrate the IMU at the waist. In contrast, there was no substitute for the pressure insoles, which provide unique force data that helps to distinguish if the person is lifting a heavy object vs. a light object vs. no object at all and just bending forward. In theory, object mass could be obtained using sensors beyond those we tested (e.g., measured directly using force-instrumented gloves, or estimated indirectly via muscle EMG), but these again introduce added complexity and practical implementation challenges may be barriers to adoption for many applications.

We observed that using two sensor locations (trunk IMU and pressure insoles) sacrificed minimal accuracy compared to using more sensor locations (e.g., all six distributed sensor locations, **Figure 5.4** and **Figure 5.6**). This supports the idea that it may be possible to use a relatively small subset of sensors to make workplace implementation more practical, while still obtaining accurate estimates of back loading. These findings also demonstrate that more sensors, or more widely distributed sensors, should not be assumed to result in substantially more accurate musculoskeletal load monitoring tools. For monitoring lumbar loading during manual material handling there appears to be a sweet spot for accuracy and practicality that involves using pressure insoles and a single IMU.

5.5.2 What types of algorithms work well for this sensor data fusion?

All results presented here were developed using Gradient Boosted Decision Tree algorithms. We found this type of algorithm to work well during early exploration of the data. Within the Gradient Boosted Decision Tree framework, we utilized the histogram-based decision tree building algorithm, as it significantly reduces the training time with larger datasets (>10k samples), but did not noticeably

degrade prediction performance of our algorithms, compared to traditional Gradient Boosted Decision Trees. Using this approach, input signal values are separated into bins, reducing the computational complexity of splitting decisions and efficiently leveraging parallel computational resources [203].

In pilot data analysis we also explored other categories of algorithms/models including generalized linear models, ensemble methods (random forests), shallow neural networks (2 hidden layers), and support vector regression. While most of these methods (linear, support vector regressions, forests) resulted in comparable prediction results to each other, Gradient Boosting consistently provided more accurate estimates in our preliminary data sets. Also, some of these methods (most notably, support vector regression) did not scale well with a large number of data points and became prohibitive to train.

We did not have success with traditional neural-network models. This may have been because of insufficient number of layers, nodes, or the chosen activation functions. We note that the hyperparameter space for neural networks is significantly larger than for the other methods we tried. We provide this brief commentary on the explored set of machine learning algorithms for this problem domain to share our initial experiences. Our review and evaluation of alternative algorithm approaches is not exhaustive and there are certainly other applicable AI-based or statistical methods beyond this initial study. Such promising candidates include convolutional neural layers and recurrent neural networks, which may be interesting to explore in the future.

5.5.3 How accurately can we monitor low back loading during manual material handling tasks?

The idealized results demonstrate the potential for a small number of sensors to provide accurate estimates of low back loading. Using a trunk IMU and pressure insoles resulted in lumbar moment estimates that were strongly correlated with lab-based lumbar moments (r^2 = 0.89, **Figure 5.4B**). And this solution performed well across the broad range of tasks and lumbar moment magnitudes captured (**Figure 5.4A**). The RMSE and MAPE accuracy results corroborated that this wearable sensor approach is very promising. The RMSE corresponds to less than 10% of the peak lumbar moments during heavy lifting. For context, we found that using just two sensor locations (trunk IMU and pressure insoles) during about 400 different material handling tasks exhibited similar levels of accuracy (r^2 = 0.89 and RMSE = 20 Nm) as those reported in [21] which combined 8-17 IMUs and force-sensing shoes to estimate lumbar moments during 4 tasks that involved lifting and carrying a 10 kg box (r^2 = 0.93 and RMSE < 20 Nm).

Real wearable results highlighted the technological key to realizing accurate estimates of back loading in the real world. Combining a real wearable trunk IMU and pressure insoles resulted in lower average accuracy than with the idealized wearable sensors (e.g., r^2 = 0.80 vs. r^2 = 0.89), and only marginal benefits over a real trunk IMU alone (r^2 = 0.80 vs. r^2 = 0.79). This appears to be due to variability in insole force estimates compared to vertical forces estimated from idealized wearable sensors (i.e., from lab-based force plates, **Supplemental Figure 5.5B**). In contrast, we found that trunk orientation from the real wearable sensor (trunk IMU) was a very strong indicator of trunk orientation from idealized wearable sensors (lab-based optical motion capture), with low variability (**Supplemental Figure 5.5A**). Together, this seems to explain why GRFs were of similar importance as the trunk IMU when using the idealized wearable sensors (**Figure 5.3**) but of much lower importance when using the real wearable sensors (**Figure 5.5**). A key technological priority should be to reduce the variability in insole force estimates. The good news is that there are various ways to improve these force estimates through advances in signal processing, calibrations, and sensor hardware, or via optimization of sensors for pressure/force magnitudes expected in certain tasks such as material handling. As variability in insole

forces is reduced, then the accuracy of algorithms developed using real wearable sensors will approach that observed from the idealized wearable sensors. We confirmed this to be true by replacing the real pressure insole data with idealized pressure insole data during algorithm development and finding lumbar moment estimates to have similar accuracy to our idealized wearable sensor algorithms.

These insights highlight the benefits of using idealized signals when exploring new wearable sensor solutions. If we used real wearable sensors alone, we may have concluded that pressure insoles do not improve back loading estimates compared to a single wearable trunk IMU. In actuality the pressure insoles provide unique and highly valuable force data (**Figure 5.2, Figure 5.3, Table 5.3**) that can help distinguish when someone is lifting a heavy object vs. simply bending forward, and that can greatly improve capabilities for monitoring trends in low back loading (particularly at higher magnitudes). Overall, our complementary analyses, evaluating accuracies across a range of reduced sensor algorithms for both idealized and real wearable sensors, and ranking signal importances, provides a systematic and effective approach to identifying key sensor signals and promising wearable sensor combinations.

5.5.4 Benefits and drawbacks of single wearable sensor solutions

The results demonstrate that a single IMU solution can perform reasonably well for estimating lumbar moments (**Table 5.3**). The practical benefits were described in the Introduction (e.g., relative simplicity for workplace implementation). The trunk IMU, and to a slightly lesser extent thigh and pelvis IMUs, provided moderately high correlation coefficients up to $r^2 = 0.74$ in idealized wearable sensor analysis, and up to $r^2 = 0.79$ in the real wearable sensor analysis. The reason for the slightly stronger correlations with real wearable sensors for the trunk IMU algorithm is unknown, but may be due to a richer set of candidate signals that we input into the real vs. idealized wearable algorithms (see **Table 5.1** vs. **Table 5.2**), which included additional spine segment and joint angle estimates from the Xsens functional skeleton calibration, and IMU accelerations and velocities. These results suggest that commercial wearables that place an IMU on one of these segments (trunk, pelvis, or thighs) are at least monitoring the types of signals that can be correlated with lumbar moments (with proper algorithm development and training).

The critical drawback of single IMU wearables is that they fail to capture increases in lumbar loading when different objects are lifted, and as a result they tend to perform worse for higher lumbar moments -- which unfortunately are the instances of highest ergonomic interest since these are most damaging to musculoskeletal tissues. This accuracy limitation at higher magnitudes is evident in plots of time series lumbar moments. For example, the trunk IMU algorithm does not capture the increase in low back loading peaks when a participant is picking and placing a 10 kg box (gray areas in **Figure 5.8**). In contrast, these elevated back loads from the handheld mass are captured by distributed sensor solutions that include pressure insoles along with at least one IMU (**Figure 5.8**). The time-series plots show a representative lifting task, while the scatter plots and ta-bles presented in the Results provide comprehensive results from all the participants and across all the manual material handling tasks collected.

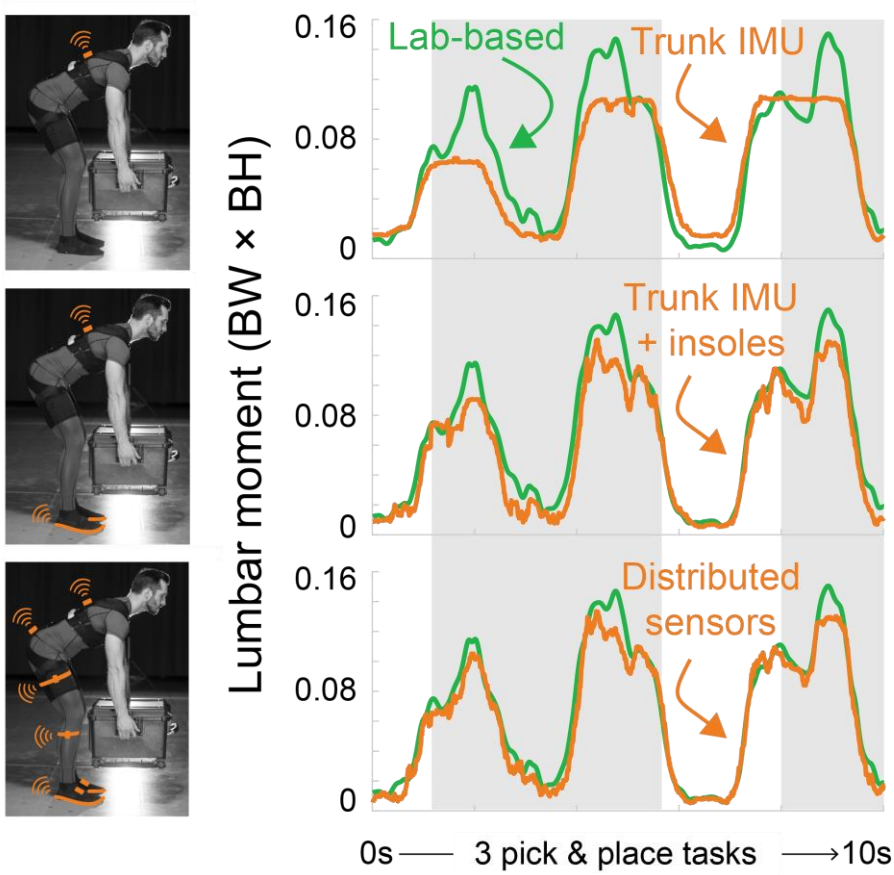


Figure 5.8 Single IMU wearable does not capture key trends and peaks in lumbar loading when objects are lifted. Time-series lab-based lumbar extension moment (green) and idealized wearable algorithm moments developed with three different idealized wearable sensor combinations (orange). Shown is a subset of the hundreds of manual material handling tasks performed for an example participant; 3 pick and place task cycles with a 10 kg box shown. Gray areas are approximately when the participant was holding the 10 kg box, white areas are when the participant had no object in their hands. The trunk IMU tends to perform worse when the box is being held or lifted, whereas the trunk IMU plus pressure insoles, and distributed sensors, are able to better track key lumbar loading trends (gray areas). BW = body weight*body height.

As another example, lifting objects of increasing mass with similar body posture causes an increase in peak lumbar moments (**Figure 5.9**). Using a trunk IMU alone completely misses the trend of increasing low back loading when individuals adopt similar trunk orientations (i.e., postures) for each lift, while using a solution that includes both pressure insoles and an IMU captures these increasing back load trends (**Figure 5.9**). These results confirm our expectations from the Introduction: while a single IMU (on the trunk, or elsewhere) may provide a reasonable estimate of back loading (or trends in loading) due to changes in general body posture, the estimation accuracy is compromised when objects of differing mass are handled or when other external forces are applied to the body (e.g., during pushing, pulling, or leaning). It may also be possible to use the trunk IMU plus pressure insoles combination during initial assessment of each worker, or intermittently over time, in order to better calibrate the trunk IMU (alone) for each worker -- in effect supplementing the minimal single senor solution to improve accuracy

and personalization. Of note, using the pressure insoles alone yielded fairly poor accuracy (**Supplemental Figure 5.1**), again highlighting the benefits of fusing data from multiple sensor locations.

In short, caution should be taken when using a single wearable (on any body segment) to monitor low back loading, particularly in situations where external forces are variable, or when object masses being handled are not manually input (or otherwise accounted for) in algorithms. Further exploration is warranted to understand the implications of single IMU sensor accuracy within the context of the sensitivity of risk assessment tools like LiFFT. Wearable solutions that fuse data from multiple sensor locations (e.g., trunk IMU and pressure insoles) are expected to provide more accurate and reliable ways to automate ergonomic assessments or provide continuous daily risk monitoring for material handling jobs that involve lifting objects of varying weight; albeit with slightly more implementation complexity due to more sensor modalities, and presuming the variability in pressure insole force estimates can be adequately reduced.

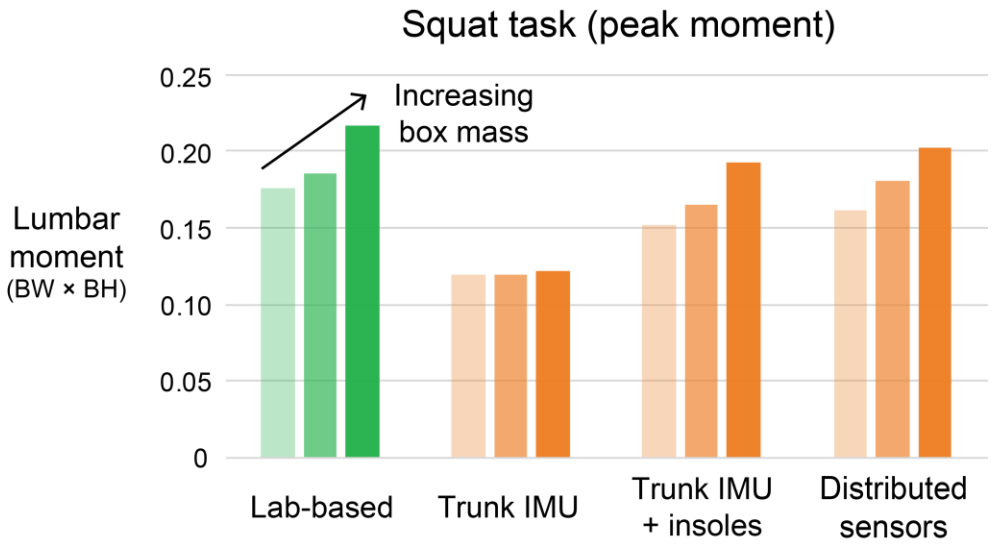


Figure 5.9 Single IMU wearable does not capture increases in lumbar loading when heavier objects are lifted. Shown is an illustrative example from one participant: peak lumbar moment of squat tasks when increasing box masses are lifted (10 kg, 15 kg, 23 kg are shown). The trunk IMU and insole algorithm, and also the distributed sensor algorithms capture the trend of increasing lumbar moment with heavier object mass. However, the trunk IMU algorithm does not; it predicts a similar peak moment with each lift regardless of the mass being lifted. BW = body weight × body height.

5.5.5 Lateral bending lumbar moment can also be estimated with trunk IMU and insoles

Lumbar extension moments have been shown to be a key metric for monitoring cumulative damage to the low back and resulting injury risk (see Introduction). However, there are also opportunities to provide a broader, multifactorial assessment of injury risk by monitoring other musculoskeletal loading metrics with wearables. One additional metric of interest to us was lumbar lateral bending moment, as increases in lateral bending moment contribute to increases in back muscle and disc compression forces [204], which influence cumulative damage to the low back. We therefore repeated the same algorithm development and evaluation process using the idealized wearable sensor data from this study, but using time series lumbar lateral bending moment as the target metric (**Supplemental Figure 5.2**,

Supplemental Figure 5.3, **Supplemental Figure 5.4**). Encouragingly, signals from the same set of wearable sensors (trunk IMU and pressure insoles) that were identified as most important for estimating lumbar extension moments were also the most important for estimating lateral bending moments (**Supplemental Figure 5.4**). Similar to our analysis of the lumbar extension moment, a single trunk IMU algorithm did not capture all trends in lateral bending moment, namely when the user held and moved objects of differing mass lateral to their body (**Supplemental Figure 5.3**). The trunk IMU algorithm resulted in an average accuracy of $r^2 = 0.65$ (**Supplemental Table 5.1**). Combining the pressure insoles with the trunk IMU increased accuracy to $r^2 = 0.83$ (**Supplemental Figure 5.3**). This again demonstrates how a small set of wearable sensors (trunk IMU and pressure insoles) could provide a practical and accurate tool for monitoring low back loading (due to both lateral and extension moments), with only a relatively small reduction in accuracy compared to the full set of distributed sensors tested ($r^2 = 0.88$, **Supplemental Figure 5.3**).

5.5.6 Limitations and future opportunities

Given the exploratory nature of developing next generation wearables, there were many interesting additional areas of research that were beyond the scope we chose to evaluate in this study. Numerous other candidate wearable sensors and emerging technologies, signal processing techniques, machine learning algorithms, and musculoskeletal metrics of interest could be explored in future studies. Additionally, while we focus on evaluating a tool for monitoring low back loading in a workplace environment, there are many other exciting research and clinical applications of a low back monitoring tool. For example, a similar wearable solution might be used in a clinical setting to monitor patients during post-injury or post-surgery rehabilitation, track their progress, or assist with return-to-work decisions.

Within the scope of this study we note some limitations of our approach. First, real wearable sensors used were research-grade instrumentation. Implementing algorithms on consumer-grade hardware, or any other hardware platform not tested here, would require additional algorithm calibration, validation and evaluation. Second, the number of participants tested was informed by our prior studies combining wearable sensors and machine learning [173], [194], but this kind of exploratory (non-hypothesis-driven) research is not amenable to traditional sample size calculations. The consistency of results for individual participants using the k-fold validation analysis suggests our sample was adequate, but we acknowledge that our understanding of how much data is enough to identify promising wearable monitoring tools using diverse machine learning techniques is continuing to evolve. Third, we did not use sensors to monitor the location of the object being lifted relative to the body (e.g., spine). Although this distance could be estimated by tracking multiple segments of the arms, we choose not to do this for reasons of simplicity and practicality. For now, adding this complexity seems unnecessary given that the simpler trunk IMU plus pressure insole solution presented here already shows strong potential for estimating lumbar moments. Fourth, we focused on load monitoring as a key risk factor for low back disorders, but it is worth reminding that sensors like the trunk IMU capture other data such as twisting (spine rotation) and trunk acceleration/deceleration, which can also be useful and complementary for injury risk assessment. Fifth, algorithms were developed and evaluated on a broad range of movement tasks we identified as representative of many manual material handling tasks performed in workplace environments. The efficacy of using a trunk IMU and pressure insoles to monitor low back loading for other tasks or jobs outside of those tested would require additional validation. To our knowledge, this is one of the largest databases ever collected of synchronized laboratory and wearable sensors in this ergonomics and material handling domain. As such, we plan to use this dataset for future secondary analysis, and to make it available to other researchers interested in exploring additional research

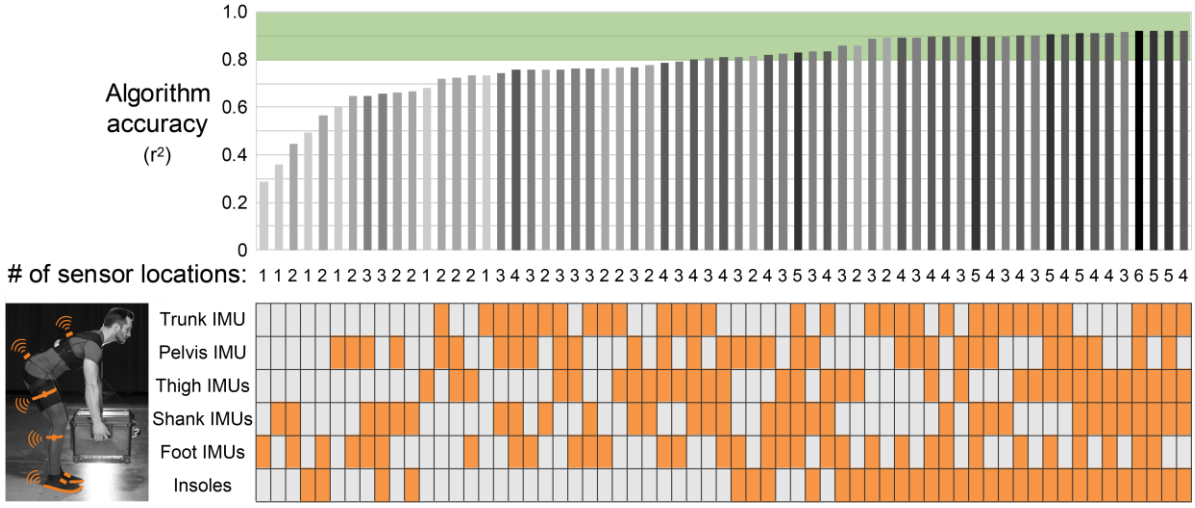
questions.

5.6 Conclusion

Here, we present a promising wearable solution for the practical, automated, and accurate monitoring of low back loading during manual material handling. We found that two key sensors for accurately monitoring low back loading are a trunk IMU and pressure insoles. Using signals from these two sensors together with a Gradient Boosted Decision Tree algorithm has the potential to provide a practical (relatively few sensors), accurate (up to $r^2 = 0.89$), and automated way (using wearables) to monitor time series lumbar moments across a broad range of material handling tasks. The trunk IMU could be replaced by thigh IMUs or a pelvis IMU without sacrificing much accuracy, but there was no practical substitute for the pressure insoles. The key to realizing accurate lumbar load estimates with this approach in the real world will be optimizing force estimates from pressure insoles. This promising wearable solution has the potential to transform low back injury risk assessment, monitoring, and prevention in various industries.

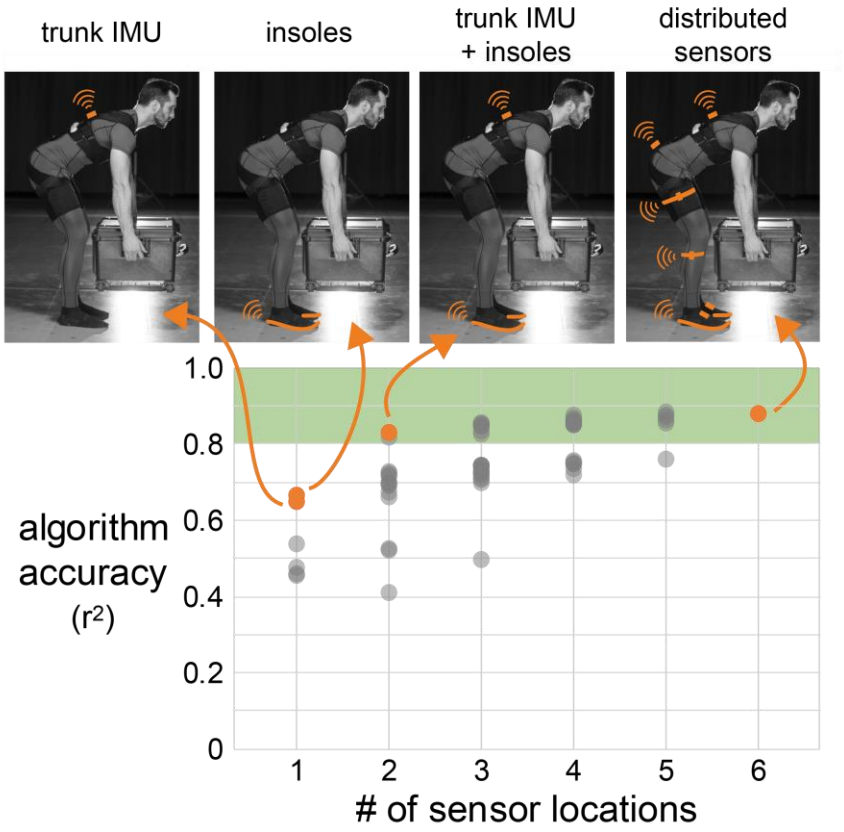
5.7 Supplemental Material

5.7.1 Algorithm accuracies for all wearable sensor combinations

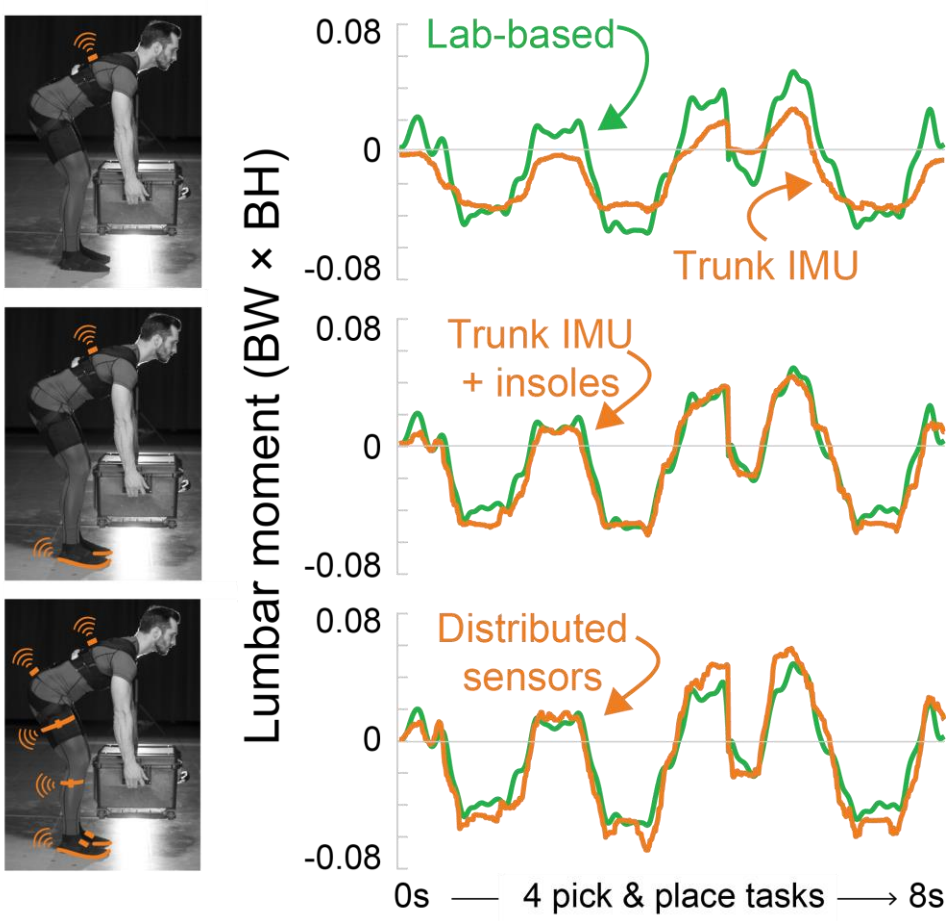


Supplemental Figure 5.1 Number of sensor locations and specific sensor combination influences algorithm accuracy. (Top) Average accuracy using idealized wearable sensors summarized here using the average coefficient of determination (r^2). Darker color bars correspond to an increasing number of sensor locations used in the algorithm. **(Bottom)** Summary of which sensor locations were used in each reduced sensor combination. Orange grid boxes indicate that signals from that sensor location were used for the algorithm. These results are equivalent to **Figure 5.2** but depicted here to visualize the performance of each sensor combination.

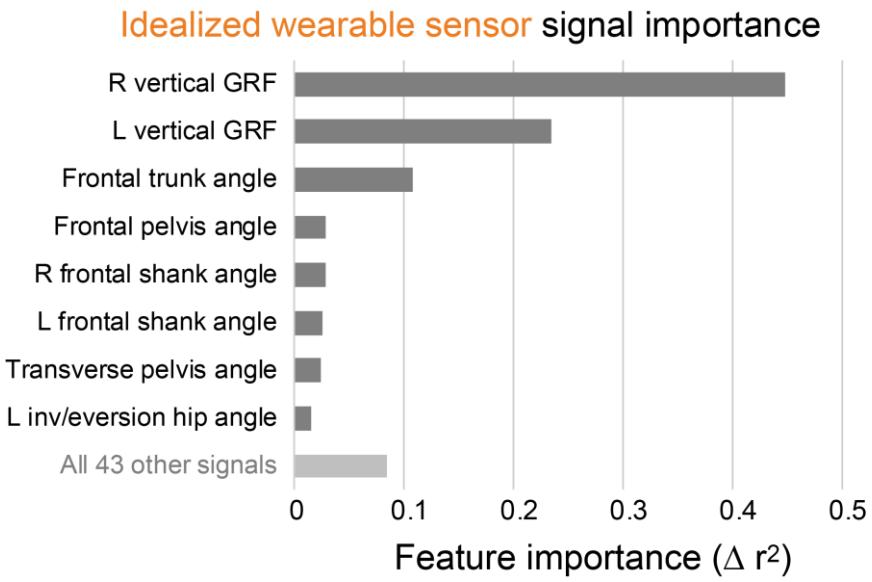
5.7.2 Lateral Bending Lumbar Moment



Supplemental Figure 5.2 Combining a trunk IMU and pressure insoles is also promising for estimating lumbar lateral bending moment. Average accuracy using idealized wearable sensors summarized here using the average coefficient of determination (r^2). As with estimating lumbar extension moments (Figure 2), the maximum algorithm accuracy increased with number of sensor locations. Orange dots represent the distributed sensor algorithm (6 sensor locations), and a subset of algorithms using 1 and 2 sensor locations. All algorithms here were developed with idealized wearable sensor signals and the target was lumbar lateral bending moment.



Supplemental Figure 5.3 Representative example of lumbar lateral bending moment estimates based on different wearable sensor combinations. Shown is lab-based (green) and algorithm-estimated (orange) lateral bending moments for three different sensor combinations. Depicted is a subset of the hundreds of manual material handling tasks performed for an example participant; 4 pick and place task cycles with a 5 kg box shown. These results were similar to those observed when estimating the lumbar extension moment (**Figure 5.8**): the single IMU wearable did not well estimate the higher magnitude lateral bending moments, but combining a pressure insole with at least one IMU improved these estimates. All algorithms here were developed with idealized wearable sensor signals and the target was lumbar lateral bending moment. BW = body weight × body height.

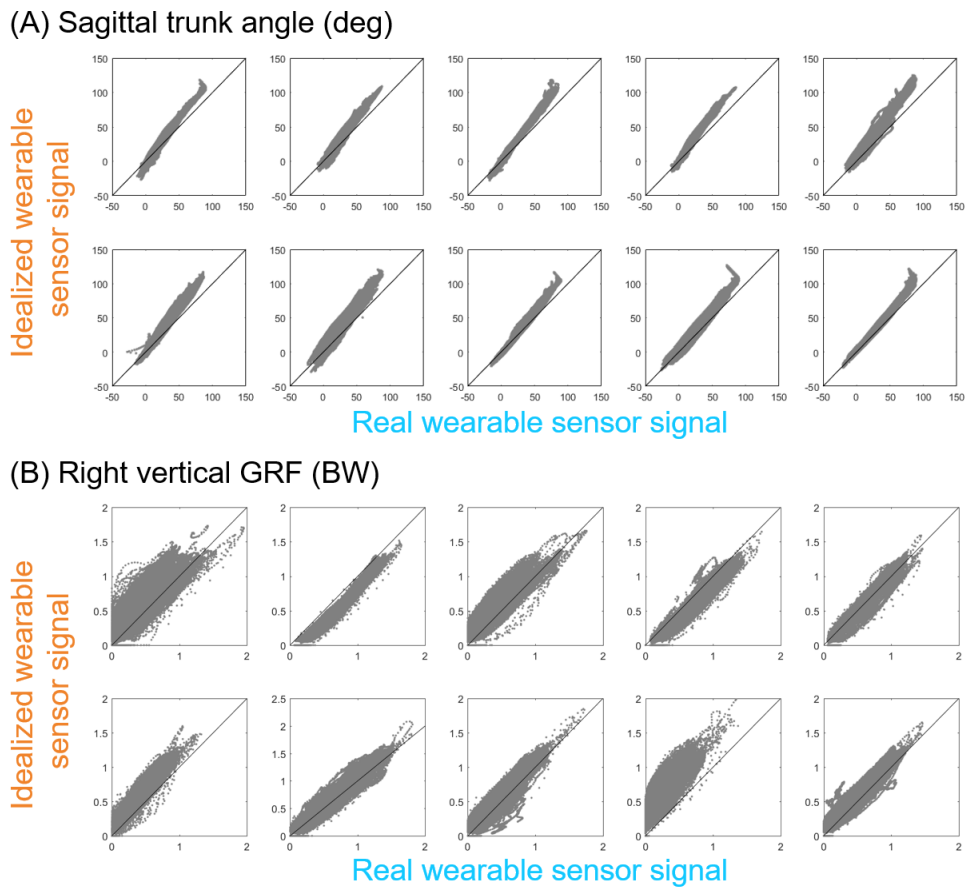


Supplemental Figure 5.4 Vertical GRFs and frontal trunk angle are the most important signals for estimating lumbar lateral bending moments. Signal importances are from the idealized wearable sensor algorithm for estimating lumbar lateral bending moment. Note that these signals can be obtained from the same two sensors (trunk IMU and pressure insoles) that we identified as being the most important for estimating lumbar extension moment. R = right; L = left.

Supplemental Table 5.2 Algorithm accuracies when estimating lumbar lateral bending moment. Shown are participant-specific and average (avg) accuracy results from three different subsets of idealized wearable sensors: trunk IMU, trunk IMU and pressure insoles, and all distributed sensors. Accuracy is reported as the coefficient of determination (r^2) across all time samples for a given participant.

	Algorithm accuracy (r^2)		
Participant	**Trunk IMU**	**Trunk IMU + insoles**	**Distributed sensors**
1	0.68	0.86	0.90
2	0.76	0.88	0.91
3	0.44	0.80	0.88
4	0.66	0.84	0.89
5	0.67	0.83	0.90
6	0.67	0.82	0.89
7	0.62	0.81	0.79
8	0.72	0.84	0.88
9	0.70	0.88	0.92
10	0.59	0.74	0.84
avg ± std	**0.65 ± 0.09**	**0.83 ± 0.04**	**0.88 ± 0.04**

5.7.3 Idealized wearable sensor signals vs. real wearable sensor signals



Supplemental Figure 5.5 Trunk IMUs provided a relatively precise estimate of trunk orientation while pressure insoles provided more variable estimate of vertical force. Ten scatter plots represent each participant, and gray dots represent each time sample. Real wearable trunk orientation from an IMU correlated well with idealized wearable trunk angle from lab-based motion capture (A), whereas real wearable vertical force from pressure insoles did not correlate as well with idealized wearable vertical force from lab-based force plates, and exhibited higher variability (B). A line with a slope of one is added to visualize a perfect correspondence between idealized and real wearable sensor signals. See **Table 5.1** and **Table 5.2** for details on idealized and real wearable sensor signals. deg = degrees; BW = body weight.

6 Conclusions and Future Work

This dissertation has presented several research contributions to the field of monitoring musculoskeletal dynamics with imaging and wearable measurement modalities. These research projects demonstrate the exciting potential to expand our capabilities for long-term, widespread, and unconstrained monitoring of musculoskeletal dynamics in applied situations outside the lab. However, many areas related to the validation, translation, and application of these wearable monitoring tools remain unexplored. The following section details possible directions for future work.

6.1 Future Work for Chapter 2: Ultrasound estimates of Achilles tendon dynamics

Ultrasound and image processing techniques offer an exciting opportunity to peer beneath the skin during human locomotion tasks and evaluate the dynamics of muscles and tendons. Given the many potential reasons I may have observed unexpected Achilles tendon shortening in these experiments, there are opportunities for additional targeted research studies that aim to evaluate if errors reside in the assumed musculoskeletal model and/or experimental methods. For example, to evaluate choices and assumptions in the musculoskeletal model, more sophisticated models could be utilized that account for interactions between adjacent muscle tendon units and/or connective tissues. Another area of future research is to identify which ultrasound method(s) are best suited for high muscle-tendon unit force and displacement tasks, given our inconsistent results between methods and scientific interest in these types of tasks (e.g. walking and locomotion).

6.2 Future Work for Chapter 3: Ground reaction forces and tibial force are not strongly correlated

The work presented here focused on evaluating the correlation between ground reaction force metrics and peak tibial force over a range of running speeds and slopes. The relationship between other single variable metrics and/or other musculoskeletal metrics of interest (e.g. loading on other tissues at risk of overuse injury) could be evaluated in future studies. Findings could motivate or dissuade the use of single variable metrics in research studies and/or commercial devices aimed at evaluating overuse injury risk on specific musculoskeletal tissues.

6.3 Future Work for Chapter 4: Estimating tibial force with wearable sensors

The work in Chapter 4 provides a promising and feasible approach for monitoring tibial loading with wearables. To translate this solution, additional research is needed to calibrate and validate the algorithm using specific device hardware and characterize the accuracy and repeatability across runners and across days. Once a robust, wearable tibial load monitoring tool is available for real world, repeated day use, there are many exciting areas for future scientific, sport and clinical research. From a clinical perspective, future research could use this portable tool to conduct the first prospective study that measures tibia bone loading continuously alongside injury incidence to gain new insights about bone stress injury etiology. These clinical insights can then inform scientific advancements in modeling longitudinal bone damage, adaptation, and remodeling, improving how we simulate, model, and predict bone health. In parallel, this tool can help take biomechanics and sports research outside the lab, allowing for widespread and unconstrained monitoring of bone loading during running, aiding the

evaluation of sports interventions (e.g. footwear, performance enhancing exoskeletons, etc.) in real world environments.

6.4 Future Work for Chapter 5: Estimating lumbar moments with wearable sensors

The work in Chapter 5 provides a promising and feasible approach for monitoring low back loading with wearables. Similar to Chapter 4, additional research is needed to translate this technology into a robust solution capable of being deployed for real world use. Based on our benchmark evaluation using real wearables, there is a particular need for improvements in how we monitor force under the foot with wearables (e.g. with pressure insoles), suggesting a need for research that improves the sensor hardware technology and/or develops biosignal analysis techniques for improving the signal reliability (e.g. strategic calibrations to routinely rescale the signals). Again similar to Chapter 4, once a robust, wearable low back load monitoring tool is available for real world, repeated day use, there are many exciting areas for future scientific, occupational and clinical research. From a clinical and ergonomics perspective, future research could use this portable tool to conduct prospective studies that measures low back loading continuously alongside injury incidence to supplement previous insights on the relationship between low back cumulative damage and injury risk. Future occupational research could use this tool for automated ergonomic assessments, helping to efficiently evaluate how various ergonomics factors influence worker safety. Future research could also adapt this tool to evaluate injury prevention interventions, such as exoskeletons, and quantify how much assistance is provided to users while wearing the exoskeleton in real world environments. It extremely exciting the breadth and depth of potential research that could be fueled with accurate and practical tools for monitoring musculoskeletal loading.

6.5 Conclusions

This dissertation has provided four main contributions. Chapter 2 presents the discovery that two commonly-used ultrasound based methods for estimating tendon dynamics yielded unrealistic shortening of tendon when the ankle plantarflexes beyond neutral. Chapter 3 identifies that GRF metrics should not be used assumed to be a surrogate for tibia bone loading, dissuading the use GRF metrics in both scientific research and commercial wearables as an indicator of injury risk. Chapter 4 presents the design and validation of a novel method for combining biomechanics, wearables, and machine learning to more accurately estimate tibia bone force during running, compared to conventional methods using a single wearable metric. Chapter 5 presents the design and validation of a practical and accurate tool for monitoring low back loading with wearables across a broad range of manual lifting tasks.

Imaging and wearable measurement modalities provide enhanced capabilities for monitoring human movement and musculoskeletal dynamics. My hope is that this dissertation provides an evaluation of some limitations of current methodologies for monitoring musculoskeletal dynamics, while also suggesting exciting new approaches for accurately monitoring musculoskeletal dynamics in real world environments. Together, I hope improvements in portable and wearable monitoring tools will improve the health, happiness, safety, and productivity of our society.

7 Bibliography

[1] E. C. Honert and K. E. Zelik, “Inferring Muscle-Tendon Unit Power from Ankle Joint Power during the Push-Off Phase of Human Walking: Insights from a Multiarticular EMG-Driven Model,” *PLOS ONE*, vol. 11, no. 10, p. e0163169, Oct. 2016, doi: 10.1371/journal.pone.0163169.

[2] K. E. Zelik and J. R. Franz, “It’s positive to be negative: Achilles tendon work loops during human locomotion,” *PLOS ONE*, vol. 12, no. 7, p. e0179976, Jul. 2017, doi: 10.1371/journal.pone.0179976.

[3] K. E. Zelik, K. Z. Takahashi, and G. S. Sawicki, “Six degree-of-freedom analysis of hip, knee, ankle and foot provides updated understanding of biomechanical work during human walking,” *J Exp Biol*, vol. 218, no. 6, pp. 876–886, Mar. 2015, doi: 10.1242/jeb.115451.

[4] E. P. Lamers, A. J. Yang, and K. E. Zelik, “Feasibility of a Biomechanically-Assistive Garment to Reduce Low Back Loading during Leaning and Lifting,” *IEEE Transactions on Biomedical Engineering*, pp. 1–1, 2018, doi: 10.1109/TBME.2017.2761455.

[5] M. B. Yandell, J. R. Tacca, and K. E. Zelik, “Design of a Low Profile, Unpowered Ankle Exoskeleton That Fits Under Clothes: Overcoming Practical Barriers to Widespread Societal Adoption,” *IEEE TNSRE*, in review 2019.

[6] H. L. Bartlett, B. E. Lawson, and M. Goldfarb, “Design, Control, and Preliminary Assessment of a Multifunctional Semipowered Ankle Prosthesis,” *IEEE/ASME Transactions on Mechatronics*, vol. 24, no. 4, pp. 1532–1540, Aug. 2019, doi: 10.1109/TMECH.2019.2918685.

[7] F. Sup, A. Bohara, and M. Goldfarb, “Design and Control of a Powered Transfemoral Prosthesis,” *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 263–273, Feb. 2008, doi: 10.1177/0278364907084588.

[8] R. J. Farris, H. A. Quintero, S. A. Murray, K. H. Ha, C. Hartigan, and M. Goldfarb, “A Preliminary Assessment of Legged Mobility Provided by a Lower Limb Exoskeleton for Persons With Paraplegia,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 3, pp. 482–490, May 2014, doi: 10.1109/TNSRE.2013.2268320.

[9] W. B. Edwards, “Modeling Overuse Injuries in Sport as a Mechanical Fatigue Phenomenon.,” *Exerc Sport Sci Rev*, Jul. 2018, doi: 10.1249/JES.0000000000000163.

[10] P. V. Komi, “Relevance of in vivo force measurements to human biomechanics,” *Journal of Biomechanics*, vol. 23, pp. 27–34, Jan. 1990, doi: 10.1016/0021-9290(90)90038-5.

[11] L. E. Lanyon, W. G. J. Hampson, A. E. Goodship, and J. S. Shah, “Bone Deformation Recorded in vivo from Strain Gauges Attached to the Human Tibial Shaft,” *Acta Orthopaedica Scandinavica*, vol. 46, no. 2, pp. 256–268, Jan. 1975, doi: 10.3109/17453677508989216.

[12] H. Wilke, P. Neef, M. Caimi, T. Hoogland, and L. E. Claes, “New In Vivo Measurements of Pressures in the Intervertebral Disc in Daily Life:,” *Spine*, vol. 24, no. 8, pp. 755–762, Apr. 1999, doi: 10.1097/00007632-199904150-00005.

[13] S. H. Scott and D. A. Winter, “Internal forces at chronic running injury sites,” *Medicine & Science in Sports & Exercise*, vol. 22, no. 3, pp. 357–369, Jun. 1990.

[14] S. Sasimontonkul, B. K. Bay, and M. J. Pavol, “Bone contact forces on the distal tibia during the stance phase of running,” *Journal of Biomechanics*, vol. 40, no. 15, pp. 3503–3509, Jan. 2007, doi: 10.1016/j.jbiomech.2007.05.024.

[15] J. Cholewicki and S. M. McGill, “EMG assisted optimization: A hybrid approach for estimating muscle forces in an indeterminate biomechanical model,” *Journal of Biomechanics*, vol. 27, no. 10, pp. 1287–1289, Oct. 1994, doi: 10.1016/0021-9290(94)90282-8.

[16] J. R. Franz, L. C. Slane, K. Rasske, and D. G. Thelen, “Non-uniform in vivo deformations of the human Achilles tendon during walking,” *Gait & Posture*, vol. 41, no. 1, pp. 192–197, Jan. 2015, doi: 10.1016/j.gaitpost.2014.10.001.

[17] G. Lichtwark, “In vivo mechanical properties of the human Achilles tendon during one-legged hopping,” *Journal of Experimental Biology*, vol. 208, pp. 4715–4725, 2005.

[18] R. List *et al.*, “A moving fluoroscope to capture tibiofemoral kinematics during complete cycles of free level and downhill walking as well as stair descent,” *PLoS One*, vol. 12, no. 10, Oct. 2017, doi: 10.1371/journal.pone.0185952.

[19] W. Wang and P. G. Adamczyk, “Analyzing Gait in the Real World Using Wearable Movement Sensors and Frequently Repeated Movement Paths,” *Sensors*, vol. 19, no. 8, p. 1925, Jan. 2019, doi: 10.3390/s19081925.

[20] D. Kiernan *et al.*, “Accelerometer-based prediction of running injury in National Collegiate Athletic Association track athletes,” *Journal of Biomechanics*, vol. 0, no. 0, 2018, doi: 10.1016/j.jbiomech.2018.04.001.

[21] G. S. Faber, I. Kingma, C. C. Chang, J. T. Dennerlein, and J. H. van Dieën, “Validation of a wearable system for 3D ambulatory L5/S1 moment assessment during manual lifting using instrumented shoes and an inertial sensor suit,” *Journal of Biomechanics*, p. 109671, Jan. 2020, doi: 10.1016/j.jbiomech.2020.109671.

[22] J. M. Peake, G. Kerr, and J. P. Sullivan, “A Critical Review of Consumer Wearables, Mobile Applications, and Equipment for Providing Biofeedback, Monitoring Stress, and Sleep in Physically Active Populations,” *Front. Physiol.*, vol. 9, 2018, doi: 10.3389/fphys.2018.00743.

[23] R. W. Willy, “Innovations and pitfalls in the use of wearable devices in the prevention and rehabilitation of running related injuries,” *Physical Therapy in Sport*, vol. 29, pp. 26–33, Jan. 2018, doi: 10.1016/j.ptsp.2017.10.003.

[24] A. Ranavolo, F. Draicchio, T. Varrecchia, A. Silvetti, and S. Iavicoli, “Wearable Monitoring Devices for Biomechanical Risk Assessment at Work: Current Status and Future Challenges—A Systematic Review,” *IJERPH*, vol. 15, no. 9, p. 2001, Sep. 2018, doi: 10.3390/ijerph15092001.

[25] T. J. Roberts and E. Azizi, “The series-elastic shock absorber: tendons attenuate muscle power during eccentric actions,” *Journal of Applied Physiology*, vol. 109, no. 2, pp. 396–404, Aug. 2010, doi: 10.1152/japplphysiol.01272.2009.

[26] N. Konow, E. Azizi, and T. J. Roberts, “Muscle power attenuation by tendon during energy dissipation,” *Proceedings of the Royal Society of London B: Biological Sciences*, p. rspb20111435, Sep. 2011, doi: 10.1098/rspb.2011.1435.

[27] R. M. Alexander, “Energy-saving mechanisms in walking and running,” *Journal of Experimental Biology*, vol. 160, no. 1, pp. 55–69, Oct. 1991.

[28] R. M. Alexander, “Tendon elasticity and muscle function,” *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, vol. 133, no. 4, pp. 1001–1011, Dec. 2002, doi: 10.1016/S1095-6433(02)00143-5.

[29] T. Fukunaga, K. Kubo, Y. Kawakami, S. Fukashiro, H. Kanehisa, and C. N. Maganaris, “In vivo behaviour of human muscle tendon during walking,” *Proceedings of the Royal Society of*

London B: Biological Sciences, vol. 268, no. 1464, pp. 229–233, Feb. 2001, doi: 10.1098/rspb.2000.1361.

[30] G. A. Lichtwark and A. M. Wilson, "Is Achilles tendon compliance optimised for maximum muscle efficiency during locomotion?," *Journal of Biomechanics*, vol. 40, no. 8, pp. 1768–1775, 2007, doi: 10.1016/j.jbiomech.2006.07.025.

[31] G. S. Sawicki, C. L. Lewis, and D. P. Ferris, "It pays to have a spring in your step," *Exerc Sport Sci Rev*, vol. 37, no. 3, p. 130, Jul. 2009, doi: 10.1097/JES.0b013e31819c2df6.

[32] K. E. Zelik, T.-W. P. Huang, P. G. Adamczyk, and A. D. Kuo, "The role of series ankle elasticity in bipedal walking," *Journal of Theoretical Biology*, vol. 346, pp. 75–85, Apr. 2014, doi: 10.1016/j.jtbi.2013.12.014.

[33] D. J. J. Bregman, M. M. van der Krogt, V. de Groot, J. Harlaar, M. Wisse, and S. H. Collins, "The effect of ankle foot orthosis stiffness on the energy cost of walking: A simulation study," *Clinical Biomechanics*, vol. 26, no. 9, pp. 955–961, Nov. 2011, doi: 10.1016/j.clinbiomech.2011.05.007.

[34] S. H. Collins and A. D. Kuo, "Recycling Energy to Restore Impaired Ankle Function during Human Walking," *PLOS ONE*, vol. 5, no. 2, p. e9307, Feb. 2010, doi: 10.1371/journal.pone.0009307.

[35] D. J. Farris and G. S. Sawicki, "Human medial gastrocnemius force–velocity behavior shifts with locomotion speed and gait," *PNAS*, vol. 109, no. 3, pp. 977–982, Jan. 2012, doi: 10.1073/pnas.1107972109.

[36] G. Lichtwark, "In vivo mechanical properties of the human Achilles tendon during one-legged hopping," 2005.

[37] D. J. Farris, G. Trewartha, and M. Polly McGuigan, "Could intra-tendinous hyperthermia during running explain chronic injury of the human Achilles tendon?," *Journal of Biomechanics*, vol. 44, no. 5, pp. 822–826, Mar. 2011, doi: 10.1016/j.jbiomech.2010.12.015.

[38] E. C. Honert and K. E. Zelik, "Inferring Muscle-Tendon Unit Power from Ankle Joint Power during the Push-Off Phase of Human Walking: Insights from a Multiarticular EMG-Driven Model," *PLOS ONE*, vol. 11, no. 10, p. e0163169, Oct. 2016, doi: 10.1371/journal.pone.0163169.

[39] T. Finni, J. Peltonen, L. Stenroth, and N. J. Cronin, "Viewpoint: On the hysteresis in the human Achilles tendon," *Journal of Applied Physiology*, vol. 114, no. 4, pp. 515–517, Feb. 2013, doi: 10.1152/japplphysiol.01005.2012.

[40] N. J. Cronin and G. Lichtwark, "The use of ultrasound to study muscle–tendon function in human posture and locomotion," *Gait & Posture*, vol. 37, no. 3, pp. 305–312, Mar. 2013, doi: 10.1016/j.gaitpost.2012.07.024.

[41] M. Hoffrén, M. Ishikawa, J. Avela, and P. V. Komi, "Age-related fascicle–tendon interaction in repetitive hopping," *Eur J Appl Physiol*, vol. 112, no. 12, pp. 4035–4043, Dec. 2012, doi: 10.1007/s00421-012-2393-x.

[42] G. A. Lichtwark and A. M. Wilson, "Interactions between the human gastrocnemius muscle and the Achilles tendon during incline, level and decline locomotion," *Journal of Experimental Biology*, vol. 209, no. 21, pp. 4379–4388, Nov. 2006, doi: 10.1242/jeb.02434.

[43] D. Hawkins, C. Lum, D. Gaydos, and R. Dunning, "Dynamic creep and pre-conditioning of the Achilles tendon in-vivo," *Journal of Biomechanics*, vol. 42, no. 16, pp. 2813–2817, 2009, doi: http://dx.doi.org/10.1016/j.jbiomech.2009.08.023.

[44] A. L. Bryant *et al.*, “Effects of estrogen on the mechanical behavior of the human Achilles tendon in vivo,” *Journal of Applied Physiology*, vol. 105, no. 4, pp. 1035–1043, Oct. 2008, doi: 10.1152/japplphysiol.01281.2007.

[45] Hawkins and M. L. Hull, “A Method For Determining Lower Extremity Muscle-Tendon Lengths During Flexion/Extension Movements,” 1990, Accessed: Aug. 12, 2016. [Online]. Available: http://ac.els-cdn.com/002192909090304L/1-s2.0-002192909090304L-main.pdf?_tid=f8a0199c-609f-11e6-84e4-00000aacb35f&acdnat=1471015267_717cdaca0be5291e5d92d77aa495aa03.

[46] D. W. Grieve, “Prediction of gastrocnemius length from knee and ankle joint posture,” 1978.

[47] J. Sakuma, H. Kanehisa, T. Yanai, T. Fukunaga, and Y. Kawakami, “Fascicle–tendon behavior of the gastrocnemius and soleus muscles during ankle bending exercise at different movement frequencies,” *Eur J Appl Physiol*, vol. 112, no. 3, pp. 887–898, Jun. 2011, doi: 10.1007/s00421-011-2032-y.

[48] G. A. Lichtwark, K. Bougoulias, and A. M. Wilson, “Muscle fascicle and series elastic element length changes along the length of the human gastrocnemius during walking and running,” *Journal of Biomechanics*, vol. 40, no. 1, pp. 157–164, 2007, doi: 10.1016/j.jbiomech.2005.10.035.

[49] P. D. Hoang, R. D. Herbert, G. Todd, R. B. Gorman, and S. C. Gandevia, “Passive mechanical properties of human gastrocnemius muscle–tendon units, muscle fascicles and tendons in vivo,” *Journal of Experimental Biology*, vol. 210, no. 23, pp. 4159–4168, Dec. 2007, doi: 10.1242/jeb.002204.

[50] M. Ishikawa, P. V. Komi, M. J. Grey, V. Lepola, and G.-P. Bruggemann, “Muscle-tendon interaction and elastic energy usage in human walking,” *Journal of Applied Physiology*, vol. 99, no. 2, pp. 603–608, Aug. 2005, doi: 10.1152/japplphysiol.00189.2005.

[51] M. Ishikawa, E. Niemelä, and P. V. Komi, “Interaction between fascicle and tendinous tissues in short-contact stretch-shortening cycle exercise with varying eccentric intensities,” *Journal of Applied Physiology*, vol. 99, no. 1, pp. 217–223, Jul. 2005, doi: 10.1152/japplphysiol.01352.2004.

[52] J. R. Franz, L. C. Slane, K. Rasske, and D. G. Thelen, “Non-uniform in vivo deformations of the human Achilles tendon during walking,” *Gait & Posture*, vol. 41, no. 1, pp. 192–197, Jan. 2015, doi: 10.1016/j.gaitpost.2014.10.001.

[53] L. A. Chernak and D. G. Thelen, “Tendon motion and strain patterns evaluated with two-dimensional ultrasound elastography,” *Journal of Biomechanics*, vol. 45, no. 15, pp. 2618–2623, Oct. 2012, doi: 10.1016/j.jbiomech.2012.08.001.

[54] J.-W. H. Korstanje, R. W. Selles, H. J. Stam, S. E. R. Hovius, and J. G. Bosch, “Development and validation of ultrasound speckle tracking to quantify tendon displacement,” *Journal of Biomechanics*, vol. 43, no. 7, pp. 1373–1379, May 2010, doi: 10.1016/j.jbiomech.2010.01.001.

[55] J. R. Franz and D. G. Thelen, “Depth-dependent variations in Achilles tendon deformations with age are associated with reduced plantarflexor performance during walking,” *Journal of Applied Physiology*, vol. 119, no. 3, pp. 242–249, Aug. 2015, doi: 10.1152/japplphysiol.00114.2015.

[56] M. F. Bobbert, P. A. Huijing, and G. J. van Ingen Schenau, “A model of the human triceps surae muscle-tendon complex applied to jumping,” *Journal of Biomechanics*, vol. 19, no. 11, pp. 887–898, Jan. 1986, doi: 10.1016/0021-9290(86)90184-3.

[57] F. E. Zajac, “Muscle and tendon: properties, models, scaling, and application to biomechnics and motor control,” *Critical Reviews in Biomedical Engineering*, vol. 17, no. 4, 1989.
[58] K. E. Zelik and J. R. Franz, “It’s positive to be negative: Achilles tendon work loops during human locomotion,” *PLOS ONE*, vol. 12, no. 7, p. e0179976, Jul. 2017, doi: 10.1371/journal.pone.0179976.
[59] F. Rousseau, P. Hellier, and C. Barillot, “A novel temporal calibration method for 3-D ultrasound,” *IEEE Transactions on Medical Imaging*, vol. 25, no. 8, pp. 1108–1112, Aug. 2006, doi: 10.1109/TMI.2006.877097.
[60] E. Matijevich, “Ultrasound Probe Marker Holder,” *Zelik Lab for Biomechanics and Assistive Technology*, 2016, Accessed: Jul. 04, 2017. [Online]. Available: https://my.vanderbilt.edu/batlab/resources/motion-analysis/.
[61] R. Csapo, J. Hodgson, R. Kinugasa, V. R. Edgerton, and S. Sinha, “Ankle morphology amplifies calcaneus movement relative to triceps surae muscle shortening,” *Journal of Applied Physiology*, vol. 115, no. 4, pp. 468–473, Aug. 2013, doi: 10.1152/japplphysiol.00395.2013.
[62] A. L. Hof, “In vivo measurement of the series elasticity release curve of human triceps surae muscle,” *Journal of Biomechanics*, vol. 31, no. 9, pp. 793–800, Sep. 1998, doi: 10.1016/S0021-9290(98)00062-1.
[63] C. N. Maganaris and J. P. Paul, “Tensile properties of the in vivo human gastrocnemius tendon,” *Journal of Biomechanics*, vol. 35, no. 12, pp. 1639–1646, Dec. 2002, doi: 10.1016/S0021-9290(02)00240-3.
[64] B. Nigg, *Biomechanics of sport shoes*. University of Calgary, 2010.
[65] W. Herzog and B. Nigg, *Biomechanics of the musculo-skeletal system*. John Wiley & Sons., 2007.
[66] M. Epstein and W. Herzog, “Aspects of skeletal muscle modelling,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 358, no. 1437, pp. 1445–1452, Sep. 2003, doi: 10.1098/rstb.2003.1344.
[67] A. Arndt, A.-S. Bengtsson, M. Peolsson, A. Thorstensson, and T. Movin, “Non-uniform displacement within the Achilles tendon during passive ankle joint motion,” *Knee Surg Sports Traumatol Arthrosc*, vol. 20, no. 9, pp. 1868–1874, Sep. 2012, doi: 10.1007/s00167-011-1801-9.
[68] H. Maas and T. Finni, “Mechanical Coupling Between Muscle-Tendon Units Reduces Peak Stresses.,” *Exercise and Sport Sciences Reviews*, vol. Publish Ahead of Print, Oct. 2017, doi: 10.1249/JES.0000000000000132.
[69] E. Azizi and T. J. Roberts, “Biaxial strain and variable stiffness in aponeuroses,” *The Journal of Physiology*, vol. 587, no. 17, pp. 4309–4318, Sep. 2009, doi: 10.1113/jphysiol.2009.173690.
[70] C. J. Arellano, N. J. Gidmark, N. Konow, E. Azizi, and T. J. Roberts, “Determinants of aponeurosis shape change during muscle contraction,” *Journal of Biomechanics*, vol. 49, no. 9, pp. 1812–1817, Jun. 2016, doi: 10.1016/j.jbiomech.2016.04.022.
[71] Z. J. Domire and J. H. Challis, “A critical examination of the maximum velocity of shortening used in simulation models of human movement,” *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 13, no. 6, pp. 693–699, Dec. 2010, doi: 10.1080/10255840903453082.
[72] V. Zatsiorsky, *Biomechanics of Skeletal Muscles*. Human Kinetics, 2012.

[73] A. Fukutani, J. Misaki, and T. Isaka, “Relationship between joint torque and muscle fascicle shortening at various joint angles and intensities in the plantar flexors,” *Scientific Reports*, vol. 7, no. 1, p. 290, Mar. 2017, doi: 10.1038/s41598-017-00485-1.

[74] D. J. Farris and G. A. Lichtwark, “UltraTrack: Software for semi-automated tracking of muscle fascicles in sequences of B-mode ultrasound images,” *Computer Methods and Programs in Biomedicine*, vol. 128, pp. 111–118, May 2016, doi: 10.1016/j.cmpb.2016.02.016.

[75] J. R. Potvin and S. H. M. Brown, “Less is more: high pass filtering, to remove up to 99% of the surface EMG signal power, improves EMG-based biceps brachii muscle force estimates,” *Journal of Electromyography and Kinesiology*, vol. 14, no. 3, pp. 389–399, Jun. 2004, doi: 10.1016/j.jelekin.2003.10.005.

[76] K. E. Zelik, V. L. Scaleia, Y. P. Ivanenko, and F. Lacquaniti, “Coordination of intrinsic and extrinsic foot muscles during walking,” *Eur J Appl Physiol*, vol. 115, no. 4, pp. 691–701, Apr. 2015, doi: 10.1007/s00421-014-3056-x.

[77] S. P. Magnusson, P. Aagaard, S. Rosager, P. Dyhre-Poulsen, and M. Kjaer, “Load-displacement properties of the human triceps surae aponeurosis in vivo,” *The Journal of Physiology*, vol. 531, no. 1, pp. 277–288, Feb. 2001, doi: 10.1111/j.1469-7793.2001.0277j.x.

[78] M. B. A. McCullough, S. I. Ringleb, K. Arai, H. B. Kitaoka, and K. R. Kaufman, “Moment Arms of the Ankle Throughout the Range of Motion in Three Planes,” *Foot & Ankle International*, vol. 32, no. 3, pp. 300–306, Mar. 2011, doi: 10.3113/FAI.2011.0300.

[79] A. Silder, B. Whittington, B. Heiderscheit, and D. G. Thelen, “Identification of passive elastic joint moment–angle relationships in the lower extremity,” *Journal of Biomechanics*, vol. 40, no. 12, pp. 2628–2635, 2007, doi: 10.1016/j.jbiomech.2006.12.017.

[80] R. A. Bogey, J. Perry, and A. J. Gitter, “An EMG-to-force processing approach for determining ankle muscle forces during normal human gait,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 13, no. 3, pp. 302–310, Sep. 2005, doi: 10.1109/TNSRE.2005.851768.

[81] K. Kubo, M. Morimoto, T. Komuro, N. Tsunoda, H. Kanehisa, and T. Fukunaga, “Influences of tendon stiffness, joint stiffness, and electromyographic activity on jump performances using single joint,” *Eur J Appl Physiol*, vol. 99, no. 3, pp. 235–243, Feb. 2007, doi: 10.1007/s00421-006-0338-y.

[82] T. Muraoka, T. Muramatsu, T. Fukunaga, and H. Kanehisa, “Elastic properties of human Achilles tendon are correlated to muscle strength,” *Journal of Applied Physiology*, vol. 99, no. 2, pp. 665–669, Aug. 2005, doi: 10.1152/japplphysiol.00624.2004.

[83] C. N. Maganaris and J. P. Paul, “In vivo human tendinous tissue stretch upon maximum muscle force generation,” *Journal of Biomechanics*, vol. 33, no. 11, pp. 1453–1459, Nov. 2000, doi: 10.1016/S0021-9290(00)00099-3.

[84] F. Hug, L. Lacourpaille, O. Maïsetti, and A. Nordez, “Slack length of gastrocnemius medialis and Achilles tendon occurs at different ankle angles,” *Journal of Biomechanics*, vol. 46, no. 14, pp. 2534–2538, Sep. 2013, doi: 10.1016/j.jbiomech.2013.07.015.

[85] T. Muraoka, T. Muramatsu, D. Takeshita, Y. Kawakami, and T. Fukunaga, “Length Change of Human Gastrocnemius Aponeurosis and Tendon during Passive Joint Motion,” *CTO*, vol. 171, no. 4, pp. 260–268, Jul. 2002, doi: 10.1159/000063128.

[86] T. J. Yuen and M. S. Orendurff, “A comparison of gastrocnemius muscle–tendon unit length during gait using anatomic, cadaveric and MRI models,” *Gait & Posture*, vol. 23, no. 1, pp. 112–117, Jan. 2006, doi: 10.1016/j.gaitpost.2004.12.007.

[87] A. Fukutani, S. Hashizume, K. Kusumoto, and T. Kurihara, “Influence of neglecting the curved path of the Achilles tendon on Achilles tendon length change at various ranges of motion,” *Physiol Rep*, vol. 2, no. 10, Oct. 2014, doi: 10.14814/phy2.12176.
[88] J. Stosic and T. Finni, “Gastrocnemius tendon length and strain are different when assessed using straight or curved tendon model,” *Eur J Appl Physiol*, vol. 111, no. 12, pp. 3151–3154, Dec. 2011, doi: 10.1007/s00421-011-1929-9.
[89] R. D. Herbert *et al.*, “In vivo passive mechanical behaviour of muscle fascicles and tendons in human gastrocnemius muscle–tendon units,” *The Journal of Physiology*, vol. 589, no. 21, pp. 5257–5267, Nov. 2011, doi: 10.1113/jphysiol.2011.212175.
[90] J. D. Currey, *Bones: Structure and Mechanics*. Princeton University Press, 2013.
[91] S. Gallagher and M. C. Schall, “Musculoskeletal disorders as a fatigue failure process: evidence, implications and research needs,” *Ergonomics*, vol. 60, no. 2, pp. 255–269, Feb. 2017, doi: 10.1080/00140139.2016.1208848.
[92] F. Cosman *et al.*, “Determinants of stress fracture risk in United States Military Academy cadets,” *Bone*, vol. 55, no. 2, pp. 359–366, Aug. 2013, doi: 10.1016/j.bone.2013.04.011.
[93] P. D. Brukner, C. Bradshaw, K. Khan, S. White, and K. M. Crossley, “Stress Fractures: A Review of 180 Case,” *Clinical Journal of Sport Medicine*, 1996.
[94] S. Yagi, T. Muneta, and I. Sekiya, “Incidence and risk factors for medial tibial stress syndrome and tibial stress fracture in high school runners,” *Knee Surg Sports Traumatol Arthrosc*, vol. 21, no. 3, pp. 556–563, Mar. 2013, doi: 10.1007/s00167-012-2160-x.
[95] K. L. Bennell *et al.*, “Risk factors for stress fractures in track and field athletes. A twelve-month prospective study,” *Am J Sports Med*, vol. 24, no. 6, pp. 810–818, Dec. 1996, doi: 10.1177/036354659602400617.
[96] A. Hulkko and S. Orava, “Stress Fractures in Athletes,” *Int J Sports Med*, vol. 08, no. 3, pp. 221–226, Jun. 1987, doi: 10.1055/s-2008-1025659.
[97] G. O. Matheson, D. B. Clement, D. C. Mckenzie, J. E. Taunton, D. R. Lloyd-Smith, and J. G. Macintyre, “Stress fractures in athletes: A study of 320 cases,” *Am J Sports Med*, vol. 15, no. 1, pp. 46–58, Jan. 1987, doi: 10.1177/036354658701500107.
[98] C. Johansson, I. Ekenman, and R. Lewander, “Stress fracture of the tibia in athletes: diagnosis and natural course,” *Scandinavian Journal of Medicine & Science in Sports*, vol. 2, no. 2, pp. 87–91, 1992, doi: 10.1111/j.1600-0838.1992.tb00326.x.
[99] C. S. Chan and H. Y. Grossman, “Psychological Effects of Running Loss on Consistent Runners,” *Perceptual and Motor Skills*, vol. 66, no. 3, pp. 875–883, Jun. 1988, doi: 10.2466/pms.1988.66.3.875.
[100] H. van der Worp, J. W. Vrielink, and S. W. Bredeweg, “Do runners who suffer injuries have higher vertical ground reaction forces than those who remain injury-free? A systematic review and meta-analysis,” *Br J Sports Med*, p. bjsports-2015-094924, Jan. 2016, doi: 10.1136/bjsports-2015-094924.
[101] A. A. Zadpoor and A. A. Nikooyan, “The relationship between lower-extremity stress fractures and the ground reaction force: A systematic review,” *Clinical Biomechanics*, vol. 26, no. 1, pp. 23–28, Jan. 2011, doi: 10.1016/j.clinbiomech.2010.08.005.
[102] H. P. Crowell, “Reducing Impact Loading During Running With the Use of Real-Time Visual Feedback,” *J Orthop Sports Phys Ther*, vol. 40, no. 4, pp. 206–213, Apr. 2010, doi: 10.2519/jospt.2010.3166.

[103] J. Hamill, T. R. Derrick, and K. G. Holt, “Shock attenuation and stride frequency during running,” *Human Movement Science*, vol. 14, no. 1, pp. 45–60, Jun. 1995, doi: 10.1016/0167-9457(95)00004-C.
[104] E. M. Hennig, T. L. Milani, and M. A. Lafortune, “Use of Ground Reaction Force Parameters in Predicting Peak Tibial Accelerations in Running,” p. 9, 1993.
[105] J. M. Neugebauer, K. H. Collins, and D. A. Hawkins, “Ground Reaction Force Estimates from ActiGraph GT3X+ Hip Accelerations,” *PLOS ONE*, vol. 9, no. 6, p. e99023, Jun. 2014, doi: 10.1371/journal.pone.0099023.
[106] G. A. Borelli, *On the Movement of Animals*. Berlin Heidelberg: Springer-Verlag, 1608.
[107] N. A. Sharkey and A. J. Hamel, “A dynamic cadaver model of the stance phase of gait: performance characteristics and kinetic validation,” *Clinical Biomechanics*, vol. 13, no. 6, pp. 420–433, Sep. 1998, doi: 10.1016/S0268-0033(98)00003-5.
[108] R. G. Burdett, “Forces predicted at the ankle during running.,” *Med Sci Sports Exerc*, vol. 14, no. 4, pp. 308–316, 1982.
[109] K. Dziewiecki, Z. Mazur, and W. Blajer, “Assessment of external and internal loads in the triple jump via inverse dynamics simulation,” *Biol Sport*, vol. 30, no. 2, pp. 103–109, Jun. 2013, doi: 10.5604/20831862.1044225.
[110] E. S. Matijevich, L. M. Branscombe, and K. E. Zelik, “Ultrasound estimates of Achilles tendon exhibit unexpected shortening during ankle plantarflexion,” *Journal of Biomechanics*, vol. 72, pp. 200–206, Apr. 2018, doi: 10.1016/j.jbiomech.2018.03.013.
[111] B. Nigg, M. Mohr, and S. R. Nigg, “Muscle tuning and preferred movement path - a paradigm shift,” *Current Issues in Sport Science (CISS)*, vol. 0, no. 0, Nov. 2017, Accessed: Feb. 18, 2018. [Online]. Available: https://webapp.uibk.ac.at/ojs2/index.php/ciss/article/view/2391.
[112] D. D. D’Lima, S. Patil, N. Steklov, J. E. Slamin, and C. W. Colwell, “Tibial Forces Measured In Vivo After Total Knee Arthroplasty,” *The Journal of Arthroplasty*, vol. 21, no. 2, pp. 255–262, Feb. 2006, doi: 10.1016/j.arth.2005.07.011.
[113] R. H. Miller and J. Hamill, “Computer simulation of the effects of shoe cushioning on internal and external loading during running impacts,” *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 12, no. 4, pp. 481–490, Aug. 2009, doi: 10.1080/10255840802695437.
[114] A. J. van den Bogert and J. J. de Koning, “On Optimal Filtering For Inverse Dynamics Analysis,” *Proceedings of the IXth Biennial Conference of the Canadian Society for Biomechanics, Vancouver*, 1996.
[115] I. S. Davis, B. J. Bowser, and D. R. Mullineaux, “Reduced vertical impact loading in female runners with medically diagnosed injuries: a prospective investigation,” *Br J Sports Med*, p. bjsports-2015-094579, Dec. 2015, doi: 10.1136/bjsports-2015-094579.
[116] C. E. Milner, R. Ferber, C. Pollard, J. Hamill, and I. S. Davis, “Biomechanical Factors Associated with Tibial Stress Fracture in Female Runners,” 2006.
[117] J. R. Yong, A. Silder, K. L. Montgomery, M. Fredericson, and S. L. Delp, “Acute Changes in Foot Strike Pattern and Cadence Affect Running Parameters Associated with Tibial Stress Fractures,” *Journal of Biomechanics*, vol. 0, no. 0, May 2018, doi: 10.1016/j.jbiomech.2018.05.017.
[118] A. M. McBryde, “Stress Fractures in Runners,” *Orthopedics; Thorofare*, vol. 5, no. 7, p. 913,915,918,920-928, Jul. 1982.

[119] J. H. Calhoun, F. Li, B. R. Ledbetter, and S. F. Viegas, “A Comprehensive Study of Pressure Distribution in the Ankle Joint with Inversion and Eversion,” *Foot Ankle Int.*, vol. 15, no. 3, pp. 125–133, Mar. 1994, doi: 10.1177/107110079401500307.
[120] D. J. Farris and G. S. Sawicki, “Human medial gastrocnemius force–velocity behavior shifts with locomotion speed and gait,” *PNAS*, vol. 109, no. 3, pp. 977–982, Jan. 2012, doi: 10.1073/pnas.1107972109.
[121] T. Kernozek, N. Gheidi, and R. Ragan, “Comparison of estimates of Achilles tendon loading from inverse dynamics and inverse dynamics-based static optimisation during running,” *Journal of Sports Sciences*, vol. 35, no. 21, pp. 2073–2079, Nov. 2017, doi: 10.1080/02640414.2016.1255769.
[122] W. E. Caler and D. R. Carter, “Bone creep-fatigue damage accumulation,” *Journal of Biomechanics*, vol. 22, no. 6, pp. 625–635, Jan. 1989, doi: 10.1016/0021-9290(89)90013-4.
[123] D. R. Carter and W. E. Caler, “A cumulative damage model for bone fracture,” *Journal of Orthopaedic Research*, vol. 3, no. 1, pp. 84–90, 1985, doi: 10.1002/jor.1100030110.
[124] W. B. Edwards, D. Taylor, T. J. Rudolphi, J. C. Gillette, and T. R. Derrick, “Effects of running speed on a probabilistic stress fracture model,” *Clinical Biomechanics*, vol. 25, no. 4, pp. 372–377, May 2010, doi: 10.1016/j.clinbiomech.2010.01.001.
[125] W. B. Edwards, D. Taylor, T. Rudolphi, J. Gillette, and T. Derrick, “Effects of Stride Length and Running Mileage on a Probabilistic Stress Fracture Model,” *Medicine & Science in Sports & Exercise*, pp. 2177–2184, Jan. 2009, doi: 10.1249/MSS.0b013e3181a984c4.
[126] R. H. Miller, W. B. Edwards, S. C. E. Brandon, A. M. Morton, and K. J. Deluzio, “Why Don’t Most Runners Get Knee Osteoarthritis? A Case for Per-unit-distance Loads,” *Medicine & Science in Sports & Exercise*, vol. 46, no. 3, pp. 572–579, Mar. 2014, doi: 10.1249/MSS.0000000000000135.
[127] R. O. Ritchie, J. H. Kinney, J. J. Kruzic, and R. K. Nalla, “A fracture mechanics and mechanistic approach to the failure of cortical bone,” *Fatigue & Fracture of Engineering Materials & Structures*, vol. 28, no. 4, pp. 345–371, 2004, doi: 10.1111/j.1460-2695.2005.00878.x.
[128] P. Zioupos, J. D. Currey, and A. Casinos, “Tensile Fatigue in Bone: Are Cycles-, or Time to Failure, or Both, Important?,” *Journal of Theoretical Biology*, vol. 210, no. 3, pp. 389–399, Jun. 2001, doi: 10.1006/jtbi.2001.2316.
[129] D. M. Corey, W. P. Dunlap, and M. J. Burke, “Averaging Correlations: Expected Values and Bias in Combined Pearson rs and Fisher’s z Transformations,” *The Journal of General Psychology*, vol. 125, no. 3, pp. 245–261, Jul. 1998, doi: 10.1080/00221309809595548.
[130] M. D. Wells, D. C. Dickin, J. Popp, and H. Wang, “Effect of downhill running grade on lower extremity loading in female distance runners,” *Sports Biomechanics*, vol. 0, no. 0, pp. 1–9, Oct. 2018, doi: 10.1080/14763141.2018.1510538.
[131] T. L. Chen, W. W. An, Z. Y. S. Chan, I. P. H. Au, Z. H. Zhang, and R. T. H. Cheung, “Immediate effects of modified landing pattern on a probabilistic tibial stress fracture model in runners,” *Clinical Biomechanics*, vol. 33, pp. 49–54, Mar. 2016, doi: 10.1016/j.clinbiomech.2016.02.013.
[132] J. Hamill, K. A. Boyer, and G. Weir, “A paradigm shift is necessary to relate running injury risk and footwear design – comment on Nigg et al.,” *Current Issues in Sport Science (CISS)*, vol. 0, no. 0, Apr. 2018.

[133] L. L. Loundagin, T. A. Schmidt, and W. B. Edwards, “Mechanical Fatigue of Bovine Cortical Bone Using Ground Reaction Force Waveforms in Running,” *J Biomech Eng*, vol. 140, no. 3, pp. 031003-031003–5, Jan. 2018, doi: 10.1115/1.4038288.
[134] B. M. Nigg, *Biomechanics of Sports Shoes*. Calgary, Alta.: Topline Printing Inc., 2010.
[135] S. W. Bredeweg, B. Kluitenberg, B. Bessem, and I. Buist, “Differences in kinetic variables between injured and noninjured novice runners: A prospective cohort study,” *Journal of Science and Medicine in Sport*, vol. 16, no. 3, pp. 205–210, May 2013, doi: 10.1016/j.jsams.2012.08.002.
[136] I. S. Davis, B. J. Bowser, and D. R. Mullineaux, “Greater vertical impact loading in female runners with medically diagnosed injuries: a prospective investigation,” *Br J Sports Med*, vol. 50, no. 14, pp. 887–892, Jul. 2016, doi: 10.1136/bjsports-2015-094579.
[137] R. Ferber, I. M. F. Davis, J. F. Hamill, C. D. Pollard, and K. A. McKeown, “Kinetic variables in subjects with previous lower extremity stress fractures,” *Medicine & Science in Sports & Exercise*, vol. 34, no. 5, May 2002.
[138] A. Hreljac and P. A. Hume, “Evaluation of lower extremity overuse injury potential in runners:,” *Medicine & Science in Sports & Exercise*, pp. 1635–1641, Sep. 2000, doi: 10.1097/00005768-200009000-00018.
[139] K. Bennell *et al.*, “Ground reaction forces and bone parameters in females with tibial stress fracture.,” *Med Sci Sports Exerc*, vol. 36, no. 3, pp. 397–404, Mar. 2004, doi: 10.1249/01.MSS.0000117116.90297.E1.
[140] M. W. Creaby and S. J. Dixon, “External Frontal Plane Loads May Be Associated with Tibial Stress Fracture,” *Medicine & Science in Sports & Exercise*, vol. 40, no. 9, pp. 1669–1674, Sep. 2008, doi: 10.1249/MSS.0b013e31817571ae.
[141] K. M. Crossley, K. L. Bennell, T. V. Wrigley, and B. W. Oakes, “Ground reaction forces, bone characteristics, and tibial stress fracture in male runners.,” *Medicine & Science in Sports & Exercise*, 1999.
[142] S. J. Dixon, M. W. Creaby, and A. J. Allsopp, “Comparison of static and dynamic biomechanical measures in military recruits with and without a history of third metatarsal stress fracture,” *Clinical Biomechanics*, vol. 21, no. 4, pp. 412–419, May 2006, doi: 10.1016/j.clinbiomech.2005.11.009.
[143] R. I. Dudley, D. N. Pamukoff, S. K. Lynn, R. D. Kersey, and G. J. Noffal, “A prospective comparison of lower extremity kinematics and kinetics between injured and non-injured collegiate cross country runners,” *Human Movement Science*, vol. 52, pp. 197–202, Apr. 2017, doi: 10.1016/j.humov.2017.02.007.
[144] D. J. Kuhman, M. R. Paquette, S. A. Peel, and D. A. Melcher, “Comparison of ankle kinematics and ground reaction forces between prospectively injured and uninjured collegiate cross country runners,” *Human Movement Science*, vol. 47, pp. 9–15, Jun. 2016, doi: 10.1016/j.humov.2016.01.013.
[145] C. Napier, C. L. MacLean, J. Maurer, J. E. Taunton, and M. A. Hunt, “Kinetic risk factors of running-related injuries in female recreational runners,” *Scandinavian Journal of Medicine & Science in Sports*, vol. 28, no. 10, pp. 2164–2172, Oct. 2018, doi: 10.1111/sms.13228.
[146] S. P. Messier *et al.*, “A 2-Year Prospective Cohort Study of Overuse Running Injuries: The Runners and Injury Longitudinal Study (TRAILS),” *Am J Sports Med*, vol. 46, no. 9, pp. 2211–2221, Jul. 2018, doi: 10.1177/0363546518773755.
[147] B. M. Nigg, J. Baltich, S. Hoerzer, and H. Enders, “Running shoes and running injuries: mythbusting and a proposal for two new paradigms: ‘preferred movement path’ and

'comfort filter,'" *Br J Sports Med*, p. bjsports-2015-095054, Jul. 2015, doi: 10.1136/bjsports-2015-095054.
[148] M. R. Paquette and R. H. Miller, "Reconciling new with old injury paradigms and the need to dig deeper – comment on Nigg et al.," *Current Issues in Sport Science (CISS)*, vol. 0, no. 0, Apr. 2018.
[149] B. M. Nigg, "The Role of Impact Forces and Foot Pronation: A New Paradigm," *Clinical Journal of Sport Medicine*, vol. 11, no. 1, p. 2, Jan. 2001.
[150] J. J. Collins and M. W. Whittle, "Impulsive forces during walking and their clinical implications," *Clinical Biomechanics*, vol. 4, no. 3, pp. 179–187, Aug. 1989, doi: 10.1016/0268-0033(89)90023-5.
[151] E. L. Radin, H. G. Parker, J. W. Pugh, R. S. Steinberg, I. L. Paul, and R. M. Rose, "Response of joints to impact loading — III," *Journal of Biomechanics*, vol. 6, no. 1, pp. 51–57, Jan. 1973, doi: 10.1016/0021-9290(73)90037-7.
[152] S. R. Simon and E. L. Radin, "The response of joints to impact loading — II In vivo behavior of subchondral bone," *Journal of Biomechanics*, vol. 5, no. 3, pp. 267–272, May 1972, doi: 10.1016/0021-9290(72)90042-5.
[153] I. S. Davis, H. M. Rice, and S. C. Wearing, "Why forefoot striking in minimal shoes might positively change the course of running injuries," *Journal of Sport and Health Science*, vol. 6, no. 2, pp. 154–161, Jun. 2017, doi: 10.1016/j.jshs.2017.03.013.
[154] M. Giandolini, N. Horvais, J. Rossi, G. Y. Millet, P. Samozino, and J.-B. Morin, "Foot strike pattern differently affects the axial and transverse components of shock acceleration and attenuation in downhill trail running," *Journal of Biomechanics*, vol. 49, no. 9, pp. 1765–1771, Jun. 2016, doi: 10.1016/j.jbiomech.2016.04.001.
[155] S. Suresh, *Fatigue of Materials*. Cambridge University Press, 1998.
[156] C. D. Pollard, J. A. Ter Har, J. J. Hannigan, and M. F. Norcross, "Influence of Maximal Running Shoes on Biomechanics Before and After a 5K Run," *Orthop J Sports Med*, vol. 6, no. 6, Jun. 2018, doi: 10.1177/2325967118775720.
[157] M. B. Schaffler, E. L. Radin, and D. B. Burr, "Mechanical and morphological effects of strain rate on fatigue of compact bone," *Bone*, vol. 10, no. 3, pp. 207–214, Jan. 1989, doi: 10.1016/8756-3282(89)90055-0.
[158] T. F. Besier, "The importance of measuring lower limb cumulative load in sport: a mechanobiological approach," *White Paper*, 2018, [Online]. Available: https://imeasureu.com/2018/02/26/measuring-lower-limb-cumulative-load-sport/.
[159] "IMU Research Papers," *IMeasureU*. https://imeasureu.com/imu-research-papers/ (accessed Sep. 05, 2018).
[160] "RunScribe." https://runscribe.com/ (accessed Sep. 05, 2018).
[161] "Milestone Sports," *MilestonePod*. https://www.milestonepod.com/explore/metrics/ (accessed Sep. 05, 2018).
[162] "Stridalyzer- Improve your Running Form.," *Stridalyzer by RetiSense*. http://www.retisense.com/ (accessed Sep. 05, 2018).
[163] "Sensoria - FAQ." http://www.sensoriafitness.com/support/faq/ (accessed Dec. 07, 2018).
[164] W. B. Edwards, E. D. Ward, S. A. Meardon, and T. R. Derrick, "The Use of External Transducers for Estimating Bone Strain at the Distal Tibia During Impact Activity," *J Biomech Eng*, vol. 131, no. 5, pp. 051009-051009–6, Apr. 2009, doi: 10.1115/1.3118762.
[165] J. A. Martin *et al.*, "Gauging force by tapping tendons," *Nature Communications*, vol. 9, no. 1, p. 1592, Apr. 2018, doi: 10.1038/s41467-018-03797-6.

[166] S. Schmitt and M. Günther, “Human leg impact: energy dissipation of wobbling masses,” *Arch Appl Mech*, vol. 81, no. 7, pp. 887–897, Jul. 2011, doi: 10.1007/s00419-010-0458-z.
[167] K. P. Clark, L. J. Ryan, and P. G. Weyand, “A general relationship links gait mechanics and running ground reaction forces,” *Journal of Experimental Biology*, vol. 220, no. 2, pp. 247–258, Jan. 2017, doi: 10.1242/jeb.138057.
[168] T. R. Derrick, W. B. Edwards, R. E. Fellin, and J. F. Seay, “An integrative modeling approach for the efficient estimation of cross sectional tibial stresses during locomotion,” *Journal of Biomechanics*, vol. 49, no. 3, pp. 429–435, Feb. 2016, doi: 10.1016/j.jbiomech.2016.01.003.
[169] S. A. Meardon, J. D. Willson, S. R. Gries, T. W. Kernozek, and T. R. Derrick, “Bone stress in runners with tibial stress fracture,” *Clinical Biomechanics*, vol. 30, no. 9, pp. 895–902, Nov. 2015, doi: 10.1016/j.clinbiomech.2015.07.012.
[170] E. Halilaj, A. Rajagopal, M. Fiterau, J. L. Hicks, T. J. Hastie, and S. L. Delp, “Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities,” *Journal of Biomechanics*, vol. 81, pp. 1–11, Nov. 2018, doi: 10.1016/j.jbiomech.2018.09.009.
[171] S. Lim and C. D’Souza, “A narrative review on contemporary and emerging uses of inertial sensing in occupational ergonomics,” *International Journal of Industrial Ergonomics*, vol. 76, p. 102937, Mar. 2020, doi: 10.1016/j.ergon.2020.102937.
[172] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, “A review of wearable sensors and systems with application in rehabilitation,” *Journal of NeuroEngineering and Rehabilitation*, vol. 9, no. 1, p. 21, Apr. 2012, doi: 10.1186/1743-0003-9-21.
[173] E. S. Matijevich, L. M. Branscombe, L. R. Scott, and K. E. Zelik, “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*, vol. 14, no. 1, p. e0210000, Jan. 2019, doi: 10.1371/journal.pone.0210000.
[174] A. Vigotsky, K. E. Zelik, J. Lake, and R. N. Hinrichs, “Mechanical misconceptions: Have we lost the ‘mechanics’ in ‘sports biomechanics’?,” *Journal of Biomechanics*, vol. 93, pp. 1–5, Aug. 2019, doi: 10.1016/j.jbiomech.2019.07.005.
[175] K. L. Bennell and P. D. Brukner, “Epidemiology and site specificity of stress fractures.,” *Clin Sports Med*, vol. 16, no. 2, pp. 179–196, Apr. 1997, doi: 10.1016/S0278-5919(05)70016-8.
[176] P. Picerno, “25 years of lower limb joint kinematics by using inertial and magnetic sensors: A review of methodological approaches,” *Gait & Posture*, vol. 51, pp. 239–246, Jan. 2017, doi: 10.1016/j.gaitpost.2016.11.008.
[177] D. T.-P. Fong, Y.-Y. Chan, Y. Hong, P. S.-H. Yung, K.-Y. Fung, and K.-M. Chan, “Estimating the complete ground reaction forces with pressure insoles in walking,” *Journal of Biomechanics*, vol. 41, no. 11, pp. 2597–2601, Aug. 2008, doi: 10.1016/j.jbiomech.2008.05.007.
[178] D. A. Jacobs and D. P. Ferris, “Estimation of ground reaction forces and ankle moment with multiple, low-cost sensors,” *J Neuroeng Rehabil*, vol. 12, Oct. 2015, doi: 10.1186/s12984-015-0081-x.
[179] H. P. Crowell and I. S. Davis, “Gait retraining to reduce lower extremity loading in runners,” *Clin Biomech (Bristol, Avon)*, vol. 26, no. 1, pp. 78–83, Jan. 2011, doi: 10.1016/j.clinbiomech.2010.09.003.
[180] S. Gallagher, R. F. Sesek, M. C. Schall, and R. Huangfu, “Development and validation of an easy-to-use risk assessment tool for cumulative low back loading: The Lifting Fatigue

Failure Tool (LiFFT)," *Applied Ergonomics*, vol. 63, pp. 142–150, Sep. 2017, doi: 10.1016/j.apergo.2017.04.016.
[181] A. Haris Phuah, A. G. Schache, K. M. Crossley, T. V. Wrigley, and M. W. Creaby, "Sagittal plane bending moments acting on the lower leg during running," *Gait & Posture*, vol. 31, no. 2, pp. 218–222, Feb. 2010, doi: 10.1016/j.gaitpost.2009.10.009.
[182] H. Rice, G. Weir, M. B. Trudeau, S. Meardon, T. Derrick, and J. Hamill, "Estimating Tibial Stress throughout the Duration of a Treadmill Run," *Medicine & Science in Sports & Exercise*, vol. Publish Ahead of Print, May 2019, doi: 10.1249/MSS.0000000000002039.
[183] P.-F. Yang *et al.*, "Torsion and Antero-Posterior Bending in the In Vivo Human Tibia Loading Regimes during Walking and Running," *PLOS ONE*, vol. 9, no. 4, p. e94525, Apr. 2014, doi: 10.1371/journal.pone.0094525.
[184] M. Koch, L.-K. Lunde, M. Ernst, S. Knardahl, and K. B. Veiersted, "Validity and reliability of pressure-measurement insoles for vertical ground reaction force assessment in field situations," *Applied Ergonomics*, vol. 53, pp. 44–51, Mar. 2016, doi: 10.1016/j.apergo.2015.08.011.
[185] W. W. K. Hoeger, L. Bond, L. Ransdell, J. M. Shimon, and S. Merugu, "ONE-MILE STEP COUNT AT WALKING AND RUNNING SPEEDS," *ACSM's Health & Fitness Journal*, vol. 12, no. 1, p. 14, Feb. 2008, doi: 10.1249/01.FIT.0000298459.30006.8d.
[186] "Back injuries prominent in work-related musculoskeletal disorder cases in 2016," Bureau of Labor Statistics, U.S. Department of Labor, 2018. Accessed: Oct. 09, 2020. [Online]. Available: https://www.bls.gov/news.release/archives/osh_11092017.htm.
[187] S. E. Luckhaupt *et al.*, "Prevalence, Recognition of Work-Relatedness, and Effect on Work of Low Back Pain Among U.S. Workers," *Ann Intern Med*, vol. 171, no. 4, pp. 301–304, May 2019, doi: 10.7326/M18-3602.
[188] H. Yang, S. Haldeman, M.-L. Lu, and D. Baker, "Low Back Pain Prevalence and Related Workplace Psychosocial Risk Factors: A Study Using Data From the 2010 National Health Interview Survey," *J Manipulative Physiol Ther*, vol. 39, no. 7, pp. 459–472, Sep. 2016, doi: 10.1016/j.jmpt.2016.07.004.
[189] S. Gallagher and M. Schall, "Musculoskeletal disorders as a fatigue failure process: evidence, implications and research needs," *Ergonomics*, vol. 60, no. 2, pp. 255–269, 2016, doi: 10.1080/00140139.2016.1208848.
[190] T. R. Waters, M.-L. Lu, and E. Occhipinti, "New procedure for assessing sequential manual lifting jobs using the revised NIOSH lifting equation," *Ergonomics*, vol. 50, no. 11, pp. 1761–1770, Nov. 2007, doi: 10.1080/00140130701674364.
[191] I. Conforti *et al.*, "Validation of a novel wearable solution for measuring L5/S1 load during manual material handling tasks," in *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*, Roma, Italy, Jun. 2020, pp. 501–506, doi: 10.1109/MetroInd4.0IoT48571.2020.9138259.
[192] F. G. Larsen, F. P. Svenningsen, M. S. Andersen, M. de Zee, and S. Skals, "Estimation of Spinal Loading During Manual Materials Handling Using Inertial Motion Capture," *Ann Biomed Eng*, vol. 48, no. 2, pp. 805–821, Feb. 2020, doi: 10.1007/s10439-019-02409-8.
[193] A. S. Koopman, I. Kingma, G. S. Faber, J. Bornmann, and J. H. van Dieën, "Estimating the L5S1 flexion/extension moment in symmetrical lifting using a simplified ambulatory measurement system," *Journal of Biomechanics*, vol. 70, pp. 242–248, Mar. 2018, doi: 10.1016/j.jbiomech.2017.10.001.

[194] E. S. Matijevich, L. R. Scott, P. Volgyesi, K. H. Derry, and K. E. Zelik, “Combining wearable sensor signals, machine learning and biomechanics to estimate tibial bone force and damage during running,” *Human Movement Science*, vol. 74, p. 102690, Dec. 2020, doi: 10.1016/j.humov.2020.102690.
[195] J. Li, P. Wang, and H. J. Huang, “Dry Epidermal Electrodes Can Provide Long-Term High Fidelity Electromyography for Limited Dynamic Lower Limb Movements,” *Sensors*, vol. 20, no. 17, Art. no. 17, Jan. 2020, doi: 10.3390/s20174848.
[196] D. Colombini, E. Occhipinti, E. Alvarez-Casado, and T. R. Waters, *Manual Lifting: A Guide to the Study of Simple and Complex Lifting Tasks*. CRC Press, 2012.
[197] J. H. Friedman, “Stochastic gradient boosting,” *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, Feb. 2002, doi: 10.1016/S0167-9473(01)00065-2.
[198] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Front. Neurorobot.*, vol. 7, 2013, doi: 10.3389/fnbot.2013.00021.
[199] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in *Advances in Neural Information Processing Systems 30*, 2017, pp. 3146–3154, Accessed: Jun. 15, 2020. [Online]. Available: http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf.
[200] R. Steel and J. Torrie, *Principles and procedures of statistics*. McGraw-Hill, 1960.
[201] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[202] G. Nemeth and H. Ohlsen, “Moment Arm Lengths of Trunk Muscles to the Lumbosacral Joint Obtained In Vivo with Computed Tomography,” *Spine*, vol. 11, no. 2, pp. 158–160, Mar. 1986.
[203] A. Guryanov, “Histogram-Based Algorithm for Building Gradient Boosting Ensembles of Piecewise Linear Decision Trees,” in *Analysis of Images, Social Networks and Texts*, Cham, 2019, pp. 39–50, doi: 10.1007/978-3-030-37334-4_4.
[204] W. S. Marras and K. P. Granata, “Spine loading during trunk lateral bending motions,” *Journal of Biomechanics*, vol. 30, no. 7, pp. 697–703, Jul. 1997, doi: 10.1016/S0021-9290(97)00010-9.