www.nature.com/scientificdata

scientific data

OPEN

DATA DESCRIPTOR

A Biomechanical Dataset of 1,798 Healthy and Injured Subjects During Treadmill Walking and Running

Reed Ferber[1,2] ✉, Allan Brett[1], Reginaldo K. Fukuchi[3], Blayne Hettinga[1] & Sean T. Osis[1,4]

Quantitative biomechanical gait analysis is an important clinical and research tool for injury and disease diagnosis and treatment. However, one major criticism is that gait analysis laboratories largely operate in isolation and there is a lack of benchmark datasets, which can be used to advance research and statistical methodologies. To address this, we present an open biomechanics dataset of n = 1798 healthy and injured, young and older adults during treadmill walking and/or running at a range of gait speeds. The full dataset is available on Figshare+ and data files are contained within a series of zipped folders with folder names representing the subject ID. Each subject ID folder contains walking and/or running data containing raw marker trajectory data along with metadata for each participant. Five tutorials are also provided, demonstrating aspects such as loading data files, sample analyses of discrete variables, and calculating joint angles from code along with covering more complex topics such as principal component analysis for dimensionality reduction, statistical parametric mapping, and conducting unsupervised clustering.

Background & Summary

Biomechanical gait analysis is an important tool for determining the aetiology of walking and running musculoskeletal injuries as well as for determining treatment options, informing surgical procedures, and evaluating rehabilitation progression[1–4]. Minimal equipment allows for the calculation of basic spatiotemporal gait parameters such as stride length and time, cadence, and gait speed. However, the gold standard for biomechanical gait analysis is multi-camera 3D motion capture (MoCap) equipment[2,3].

Despite advances in MoCap equipment, methodologies, and data analysis techniques, one can argue that the fundamental approach to gait analysis has not evolved much over the past few decades. One major criticism is that gait analysis laboratories continue to operate separately from one another and lack the ability to statistically compare a new patient or cohort of patients to the thousands of patients seen in the same lab over the years[5]. We have also recently proposed that there is a lack of benchmark datasets to allow researchers the opportunity to work with large subject numbers and advance research and statistical methodologies[6]. Most importantly, the vast majority of gait biomechanics research lacks reporting transparency and focuses on a small number of variables with experiments performed on only a few subjects[5,7]. These factors certainly limit the validity of inferences and the replication of studies in biomechanics research.

In response to these limitations, some groups have begun to publish public datasets in the hopes that other research groups will use these data to provide further insights and advance the field of gait analysis[8–10]. Unfortunately, these public datasets are limited to only a few dozen participants, which limits the generalizability to the broader community. In contrast, while there are large-scale data sets that provide massive amounts of gait data for healthy and impaired individuals, these datasets are limited to ground reaction forces only[11,12].

Here, we present a dataset of walking and running gait kinematic data along with metadata concerning injury status, gender, age and running history for 1,798 individuals during treadmill walking and/or running on healthy and injured, young and older adults at a range of gait speeds[13].

[1]Faculty of Kinesiology, University of Calgary, Calgary, Alberta, Canada. [2]Running Injury Clinic, Calgary, Alberta, Canada. [3]Biomedical Engineering Program, Federal University of ABC, Sao Bernardo do Campo, Brazil. [4]Mosaic Data Science, Leesburg, VA, USA. ✉e-mail: rferber@ucalgary.ca

Injury Status	Male	Female	Age (yrs)	Height (cm)	Body Mass (kg)	No. Sessions	Walk Speed (m/s)	Run Speed (m/s)
No Injury (age 18–49)	137	171	32.52	172.19	69.54	558	1.21	2.80
No Injury (age 50+)	39	49	55.80	165.33	69.89	130	1.18	2.58
Achilles tendonitis	30	22	42.62	190.19	77.20	68	1.30	2.68
Iliotibial band syndrome	39	61	35.21	171.99	67.76	128	1.26	2.64
Osteoarthritis	91	156	56.36	167.40	76.36	422	1.11	2.41
Patellofemoral pain	61	76	35.78	178.20	69.95	142	1.23	2.62
Plantar fasciitis	20	34	45.76	170.93	77.79	59	1.22	2.50

Table 1. Demographic information for non-injured participants, divided into younger (age 18–49 yrs) and older (age 50+ yrs) age groups, as well as the top 5 injuries present in the database. Note: Number of Sessions indicates the total number of data collections for the identified group (i.e., some individuals were part of a research study and presented for baseline and follow-up sessions while others had only 1 data collection).

Methods

Our primary motivation was to create a large dataset and data were collected at the University of Calgary Running Injury Clinic either as part of research studies or as part of clinical practice between 2009 and 2017. Specifically, the data were collected to curate a uniform dataset and all participants provided informed consent under approval from the University of Calgary's Conjoint Health Research Ethics Board (Ethics ID: E–21705). In total, n = 1197 (67%) can be considered unique datasets and have not been published in previous scientific manuscripts. However, 33% of the dataset (n = 601) were recruited for specific research studies and as such, have been used in previously published works including comparisons between recreational and competitive runners[14,15], healthy and knee osteoarthritis patients[16–18], developing novel methods for MoCap marker placement[19–22], and determination of subgroups in healthy and injured runners[23–26].

There were no exclusion criteria based on pain or injury as some participants were pain-free at the time of testing ($n = 396$) while others were experiencing a lower extremity running-related injury ($n = 1402$) at the time of testing. However, these injured participants did not experience any pain during treadmill running or the testing procedure. All injured subjects were assessed and diagnosed by a licensed medical professional. Details for the non-injured subjects as well as those of the top 5 injuries are presented in Table 1.

Experimental protocol.

While the experimental protocol/data collection has been described previously[14–26], here we reiterate the work in brief. Either eight or three high-speed optoelectronic infrared-based motion capture cameras (MX3/Bonita, Vicon, Oxford, UK) were used to film treadmill-running at either 120 Hz or 200 Hz.

Spherical retro-reflective markers (9 mm diameter, Mocap Solutions, Huntington Beach, USA) were attached to the following anatomical landmarks bilaterally and can be considered the 'core' marker set for all 1,798 subjects in the dataset (see Fig. 1): medial and lateral malleoli; medial and lateral femoral condyles; greater trochanters; in the manner described previously[19,22]. For 1,082 of the 1,798 total participants, additional anatomical markers were collected consisting of the following: bilateral 1st and 5th metatarsal heads, distal aspect of the shoe, tibial tuberosity, anterior superior iliac spines, and iliac crests (see Fig. 1). For all 1,798 subjects, marker clusters on rigid shells were used to track motion trials; these consisted of three or four markers and were placed over the following segments: sacrum, bilateral thigh and shank, and posterior aspect of both shoes. This technical marker-set therefore consisted of seven rigid segments and has been reported to produce reliable kinematic waveforms[19]. Each participant wore the same shoes (Pegasus, Nike, Beaverton, USA) to standardize the footwear condition.

Following placement of all the anatomical and segment markers, each participant was asked to stand on a motorized treadmill (Bertec Corporation, Columbus, OH) for a 1-second static trial. Standing position was controlled using a graphic template placed on the treadmill with their feet positioned 0.3 m apart and pointing straight ahead. Upon completion of the static trial, the markers on the anatomical landmarks were removed while the segment markers remained. The participants were instructed to warm-up on the treadmill for 2–3 minutes, and then they walked on the treadmill at a comfortable self-selected pace, between 0.4–1.9 m/s, for 20–60 seconds in which consecutive walking strides were collected. Subjects then ran on the treadmill at a comfortable self-selected pace, between 1.1–4.9 m/s, for 20–60 seconds in which consecutive running strides were collected for processing and analysis. It should be noted that all participants were experienced treadmill users and were permitted as much time as they required to familiarize themselves with treadmill running before beginning an experiment. As well, it is important to note that 396 subjects have only walking data, 112 have only running data, and 1290 subjects have both walking and running data.

Data analysis.

Segment kinematics were calculated based on cluster movement, using a singular-value decomposition approach[27] and a joint coordinate system[28]. Joint angles were calculated using 3D GAIT custom software (Gait Analysis Systems Inc., Calgary, Alberta, Canada). Based on residual analysis, three-dimensional marker co-ordinate data were filtered at 10 Hz using a fourth order Butterworth filter. The pelvis anatomical coordinate system (CS) was deemed aligned with the laboratory CS when the subjects were in standing anatomical neutral position. The thigh and shank anatomical CSs were defined as follows: the vertical axis as the line connecting the proximal and distal joint centres; the anterior-posterior (A-P) axis is mutually perpendicular to this vertical axis and a line connecting the lateral and medial joint markers (lateral and medial condyles or malleolus); and finally, the medial-lateral (M-L) axis is mutually perpendicular the A-P and vertical axes. The foot anatomical CS A-P axis was parallel to the laboratory CS A-P axis, the foot M-L axis was mutually orthogonal to the foot A-P axis and a vertical line connecting two markers in the shoe heel counter; and the foot vertical axis is mutually perpendicular to foot M-L and A-P axes. 3D hip, knee, and ankle angles were calculated using Cardan angles, with the distal segment expressed relative to the proximal segment; and adopting the following convention: the first rotation described occurred in the medial-lateral axis, which defines the flexion/extension movement; the third rotation described was around the vertical axis, which defines the internal/external rotations; and the second rotation described was around the A-P axis, perpendicular to the previous two axes, where abduction/adduction occurs.

The gait events of foot strike and toe off were identified using a Principal Component Analysis (PCA) approach, using previously published event detection methods[21]. Data were partitioned for the stance phase of gait and time normalized to 100 data points per gait cycle.

Data Records

The full dataset is available on Figshare+[13]. Data files are contained within a single zipped folder containing a series of folders with names representing the subject ID. Each subject ID folder contains timestamped datafile(s) in ".json" format, each containing walking or running data from a single collection session.

Metadata. In addition to the Raw Data, which are stored in .json data files, Metadata, which are in .csv files, present demographic and injury information about each subject. The .csv files contain the following column headers:

sub_id:	Subject identification number
datestring:	Date and time of data collection
filename:	Raw data filename
speed_w(r):	Calculated walking or running speed (m/s)
age:	Subject age in years at time of collection
Height:	Subject height (cm) measured at time of collection
Weight:	Subject weight (kg) measured at time of collection
Gender:	Subject self-reported gender
DominantLeg:	Subject self-reported dominant leg
InjDefn:	Injury definition as self-selected from 1 of 4 options: (1) No Injury, (2) Continuing to train in pain, (3) Training volume/intensity affected, (4) Minimum of two workouts missed in a row
InjJoint:	Joint location of injury
InjSide:	Injury side (right/left)
SpecInjury:	Specific injury diagnosis from medical professional (if applicable)
InjDuration:	Injury duration in days
InjJoint2:	Secondary injury joint location (if applicable)
InjSide2:	Secondary injury side (if applicable)
SpecInjury2:	Specific secondary injury diagnosis (if applicable)
Activities:	Other self-reported athletic activities performed on a regular basis
Level:	Self-reported level of athletic activity (recreational/competitive)
YrsRunning:	Number of years subject has been running on a regular basis
RaceDistance:	Preferred race distance
RaceTimeHrs:	Preferred race distance best time (hr)
RaceTimeMins:	Preferred race distance best time: (min)
RaceTimeSecs:	Preferred race distance best time: (secs)
YrPR:	Year of preferred race distance personal best time
NumRaces:	Self-reported number of races completed per year

Raw data. Raw marker trajectory data are stored in .json format and can be parsed using most common programming languages such as Python, MATLAB or R. These files contain fields such as sampling frequency, 3D coordinates of the anatomical markers, 3D coordinates of the tracking markers, and descriptive variables calculated from the gait data (see MATLAB Processing Code)[13].

Technical Validation

All participants were asked to follow the experimental procedure and were always supervised by a researcher. The motion capture system was calibrated using tools and techniques recommended by the manufacturer, and calibration errors were less than 0.5 mm. During marker data collection, the quality of the marker reconstructions were observed by the researcher and any issues were corrected. If any unrecoverable errors were made (e.g., camera jostled, markers fell off during movement, etc.) the entire calibration and collection procedure was repeated.

Data were processed through an extract, transform, load (ETL) pipeline to add each dataset to the database. As part of ETL, the majority of the data were inspected for calculated discrete variables that fell outside

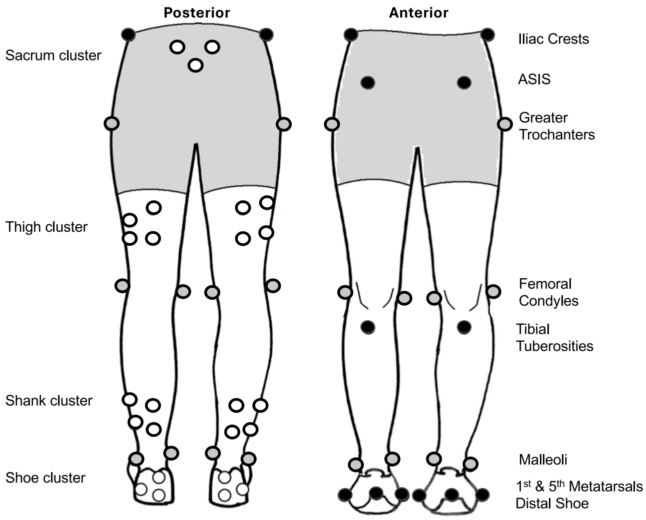

Fig. 1 The 'core' anatomical marker set (grey circles) and additional markers (black circles) for all 1,798 subjects in the dataset including the technical marker clusters on rigid shells (white circles) that were used to track motion trials.

of predefined quantiles, and datasets that fell outside of these ranges were reprocessed, if possible, or rejected if not. In addition, the processing code, included within the FigShare+ folder, contains several quality checks during calculation of kinematics. Two examples are: (1) removal of strides from analysis if any data points in the timeseries are more than 3 standard deviations away from the subject's mean waveform, (2) extraction/analysis of the largest block of contiguous valid strides.

All data included within the FigShare+ folder passed the above processing checks, indicating that the raw data are of sufficient quality to produce reasonable kinematics. However, there are several warnings that may be produced when processing the data, including using a less-robust method for event detection, or the rejection of a significant number of strides. The user will need to decide whether these warrant exclusion of data from their analyses. The sample code and tutorials in the FigShare+ folder provide examples of how to process and quality control the data.

Usage Notes

We have provided five tutorials within the FigShare+ folder[13]. The first tutorial demonstrates how to effectively load .json data files and the .csv metadata and also provides sample analyses of some discrete variables to examine fundamental principles of human locomotion. The second tutorial demonstrates how to determine matched samples when splitting data, examine demographic data, calculate joint angles, visualize joint angles, as well as generating a data set for further analysis. The third tutorial demonstrates the application of various techniques for quantitatively determining differences in joint angles including principal component analysis for dimensionality reduction as well as identifying significant differences based on statistical parametric mapping to determine significantly different sections of joint angle time-series. The fourth tutorial demonstrates assessing normality and removing outliers from discrete kinematic variables, conducting unsupervised clustering, and interpreting group membership using statistical analysis. The fifth tutorial demonstrates calculating dynamical metrics from time series data, using rejection sampling to select subsets of data, and applying statistical analyses to derived metrics.

Data limitations. We provided a standardized and detailed description of the data collection procedures and experimental protocol within the manuscript. However, we acknowledge that the data collection procedures may differ from other laboratories in the following, including but not limited to, the marker set protocol, the running shoes, and the selected gait speeds. Therefore, we advise caution when combining the current dataset with others. Moreover, we have included the data processing software, which can be reused or reproduced in other data sets so long as the same marker set protocol is used. We also acknowledge that the data were collected in a laboratory setting while subjects ran on a treadmill. Laboratory settings limit the ability to generalize to other conditions (e.g., outdoor overground running) and caution should be taken when comparing the results even with different treadmill models and protocols. Finally, although significant effort was made to collect and prepare the present dataset, there are likely deviations and/or errors present, as would be expected in any dataset.

Code availability

Software is available on FigShare+[13]. All code contains a permissive MIT license for unrestricted usage.

Received: 21 November 2023; Accepted: 17 October 2024;
Published online: 14 November 2024

References

1. Saleh, M. & Murdoch, G. In defence of gait analysis. Observation and measurement in gait assessment. *Bone Joint J.* **672**, 237–241 (1985).
2. Whittle, M. W. Clinical gait analysis: A review. *Hum Mov Sci.* **153**, 369–87 (1996).
3. Simon, S. R. Quantification of human motion: gait analysis—benefits and limitations to its application to clinical problems. *J Biomech.* **3712**, 1869–80 (2004).
4. Dugan, S. A. & Bhat, K. P. Biomechanics and Analysis of Running Gait. *Phys Med Reh Clin.* **163**, 603–21 (2005).
5. Ferber, R., Osis, S. T., Hicks, J. L. & Delp, S. L. Gait biomechanics in the era of data science. *J Biomech.* **49**(16), 3759–3761 (2016).
6. Fuller, D., Ferber, R. & Stanley, K. Why machine learning (ML) has failed physical activity research and how we can improve. *BMJ Open Sport Exerc Med.* **8**(1), e001259 (2022).
7. Knudson, D. Confidence crisis of results in biomechanics research. *Sports Biomech.* **16**(4), 425–433 (2017).
8. Fukuchi, C. A., Fukuchi, R. K. & Duarte, M. A public dataset of overground and treadmill walking kinematics and kinetics in healthy individuals. *PeerJ* **6**, e4640 (2018).
9. Fukuchi, R. K., Fukuchi, C. A. & Duarte, M. A public dataset of running biomechanics and the effects of running speed on lower extremity kinematics and kinetics. *PeerJ* **5**, e3298 (2017).
10. Theunissen, K., Van Hooren, B., Plasqui, G. & Meijer, K. Dataset of energetics and biomechanics of self-paced and fixed speed treadmill walking at multiple speeds. *Data Brief.* **41**, 107915 (2022).
11. Horsak, B. *et al.* GaitRec, a large-scale ground reaction force dataset of healthy and impaired gait. *Sci Data.* **7**, 143 (2020).
12. Horst, F. *et al.* Gutenberg Gait Database, a ground reaction force database of level overground walking in healthy individuals. *Sci Data.* **8**, 232 (2021).
13. Ferber, R., Brett, A., Fukuchi, R. K., Hettinga, B. & Osis, S. T. A biomechanical dataset of 1798 healthy and injured subjects during treadmill walking and running. *Figshare+* https://doi.org/10.25452/figshare.plus.24255795.v1 (2023).
14. Clermont, C. A., Osis, S. T., Phinyomark, A., Ferber, R. Kinematic Gait Patterns in Competitive and Recreational Runners. *J Applied Biomech* 1–26 (2017).
15. Kobsar, D., Osis, S. T., Hettinga, B. A. & Ferber, R. Classification accuracy of a single tri-axial accelerometer for training background and experience level in runners. *J Biomech.* **4710**, 2508–11 (2014).
16. Kobsar, D., Osis, S. T., Boyd, J. E., Hettinga, B. A. & Ferber, R. Wearable sensors to predict improvement following an exercise intervention in patients with knee osteoarthritis. *J Neuroeng Rehabil.* **14**(1), 94 (2017).
17. Rokhmanova, N., Kuchenbecker, K. J., Shull, P. B., Ferber, R. & Halilaj, E. Predicting knee adduction moment response to gait retraining with minimal clinical data. *PLoS Comput Biol.* **18**(5), e1009500 (2022).
18. Phinyomark, A., Osis, S. T., Hettinga, B. A., Kobsar, D. & Ferber, R. Gender differences in gait kinematics for patients with knee osteoarthritis. *BMC Musculoskelet Disord.* **17**, 157 (2016).
19. Pohl, M. B., Lloyd, C. & Ferber, R. Can the reliability of three-dimensional running kinematics be improved using functional joint methodology? *Gait Posture.* **32**(4), 559–563 (2010).
20. Phinyomark, A., Hettinga, B. A., Osis, S. & Ferber, R. Do intermediate- and higher-order principal components contain useful information to detect subtle changes in lower extremity biomechanics during running? *Hum Mov Sci.* **44**, 91–101 (2015).
21. Osis, S. T., Hettinga, B. A., Leitch, J. & Ferber, R. Predicting timing of foot strike during running, independent of striking technique, using principal component analysis of joint angles. *J Biomech.* **47**(11), 2786–9 (2014).
22. Osis, S. T., Hettinga, B. A., Macdonald, S. L. & Ferber, R. A novel method to evaluate error in anatomical marker placement using a modified generalized Procrustes analysis. *Comput Methods Biomech Biomed Engin.* **18**(10), 1108–1116 (2015).
23. Watari, R., Kobsar, D., Phinyomark, A., Osis, S. & Ferber, R. Determination of patellofemoral pain sub-groups and development of a method for predicting treatment outcome using running gait kinematics. *Clin Biomech.* **38**, 13–21 (2016).
24. Phinyomark, A., Hettinga, B. A., Osis, S. & Ferber, R. Do intermediate- and higher-order principal components contain useful information to detect subtle changes in lower extremity biomechanics during running? *Hum Mov Sci* **44**, 91–101 (2015).
25. Fox, A. S., Ferber, R. & Bonacci, J. Kinematic and Coordination Variability in Individuals With Acute and Chronic Patellofemoral Pain. *J Appl Biomech.* **37**(5), 463–470 (2021).
26. Phinyomark, A., Hettinga, B. A., Osis, S. T. & Ferber, R. Gender and Age-Related Differences in Bilateral Lower Extremity Mechanics during Treadmill Running. *PLoS One.* **98**, e105246 (2014).
27. Söderkvist, I. & Wedin, P. A. Determining the movements of the skeleton using well-configured markers. *J Biomech.* **26**, 1473–1477 (1993).
28. Cole, G. K., Nigg, B. M., Ronsky, J. L. & Yeadon, M. R. Application of the joint coordinate system to three-dimensional joint attitude and movement representation: a standardization proposal. *J. Biomech Eng.* **115**, 344–349 (1993).

Acknowledgements

This work was primarily funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant (Project No. 1030390) and Accelerator Award (Award No: 1030390).

Author contributions

R.F., A.B., R.K.F., B.H. and S.O. contributed to writing the initial draft, final manuscript, and data analysis. R.F. contributed to supervision, project administration, and funding acquisition. A.B., B.H. and S.O. contributed to data curation as well as writing the software and tutorials.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.