12-5-2023

NONAN GaitPrint: An IMU gait database of healthy young adults

Tyler M. Wiles
Madhur Mangalam
Joel H. Sommerfeld
Seung Kyeom Kim
Kolby J. Brink

See next page for additional authors

Follow this and additional works at: https://digitalcommons.unomaha.edu/biomechanicsarticles

Part of the Biomechanics Commons
Please take our feedback survey at: https://unomaha.az1.qualtrics.com/jfe/form/SV_8cchtFmpDyGfBLE
Authors
Tyler M. Wiles, Madhur Mangalam, Joel H. Sommerfeld, Seung Kyeom Kim, Kolby J. Brink, Anaëlle Charles, Alli Grunkemeyer, Marilena Kalaitzi Manifrenti, Spyridon Mastorakis, Nicholas Stergiou, and Aaron Likens
NONAN GaitPrint: An IMU gait database of healthy young adults

Tyler M. Wiles1, Madhur Mangalam1, Joel H. Sommerfeld1, Seung Kyeom Kim1, Kolby J. Brink1, Anaelle Emeline Charles1, Alli Grunkemeyer1, Marilena Kalaitzi Manifrenti1, Spyridon Mastorakis2, Nick Stergiou1,3 & Aaron D. Likens1 ✉

An ongoing thrust of research focused on human gait pertains to identifying individuals based on gait patterns. However, no existing gait database supports modeling efforts to assess gait patterns unique to individuals. Hence, we introduce the Nonlinear analysis Core (NONAN) GaitPrint database containing whole body kinematics and foot placement during self-paced overground walking on a 200-meter looping indoor track. Noraxon Ultium Motion™ inertial measurement unit (IMU) sensors sampled the motion of 35 healthy young adults (19–35 years old; 18 men and 17 women; mean ± 1 s.d. age: 24.6 ± 2.7 years; height: 1.73 ± 0.78 m; body mass: 72.44 ± 15.04 kg) over 18 4-min trials across two days. Continuous variables include acceleration, velocity, position, and the acceleration, velocity, position, orientation, and rotational velocity of each corresponding body segment, and the angle of each respective joint. The discrete variables include an exhaustive set of gait parameters derived from the spatiotemporal dynamics of foot placement. We technically validate our data using continuous relative phase, Lyapunov exponent, and Hurst exponent—nonlinear metrics quantifying different aspects of healthy human gait.

Background & Summary

An ongoing thrust of research focused on human gait pertains to identifying individuals based on gait patterns using various analytical techniques on various data types (e.g., images, videos, and radar signals)1–9. The motivation behind this approach is that if the model is complicated enough (e.g., the neural network used has enough layers), then analyzing large amounts of gait data will yield the expected results. Even without any time spent a priori using expertise in human movement science to decide on features to include in building such model. However, these attempts are thwarted by the fact that we still do not know whether some gait features can be attributed uniquely to individuals. This knowledge is fundamental to understanding whether each individual has a unique “gaitprint” just as each individual has a unique fingerprint.

The distinctiveness of one’s gaitprint readily lends itself to personal identification. Ongoing research from our laboratory offers compelling evidence that an individual’s manner of walking is as unique and singular as a fingerprint itself10,11. This revelation has profound implications, spanning diverse fields. In the realm of biometrics, the utilization of gait analysis could revolutionize the identification of perpetrators at crime scenes or enable seamless access to authorized areas without disrupting the individual’s activities. Similarly, in the domain of athletics, an individual’s distinctive running style or specialized techniques geared towards achieving specific sports objectives could greatly benefit from gaitprint recognition5. This extremely personal identification enables the creation of tailored training regimens, honing in on the subtle nuances that differentiate athletes from their peers. However, the most transformative applications may lie within the medical arena. Personalized health monitoring has long been championed as a superior approach to rehabilitation when contrasted with symptom-based treatment plans applied broadly to patient populations. The intricacies of transient or persistent gait characteristics used for distinguishing individuals hold immense potential for crafting nuanced rehabilitation programs. These programs could offer more effective and precise interventions, warranting thorough investigation and exploration.

Stability in human performance goes hand in hand with fluctuation—even the most skilled musicians and athletes bring their lifetime of reliable practice and tune it based on the unique moment of each live performance. Perfect replication is rare and might hamper performance to meet the present circumstance or task.

1Division of Biomechanics and Research Development, Department of Biomechanics, and Center for Research in Human Movement Variability, University of Nebraska at Omaha, Omaha, NE, 68182, USA. 2College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE, 68182, USA. 3Department of Physical Education and Sport Science, Aristotle University, Thessaloniki, Greece. ✉e-mail: alikens@unomaha.edu
Variation is necessary and pervasive in well-practiced movements. In all human performance, there is a stability that thrives on and coexists with natural variability. Even for upright bipedal-walking gait, regular though our stride may be, no two footfalls are exactly the same. There is in all human performance a stability that thrives on and coexists with natural variability. Human gait variability has been specifically quantified to predict cognitive and physiological declines and prevent life-threatening consequences (e.g., identification of older adults at great risk of falling)\textsuperscript{12,13}. For instance, variability has been studied for decades in clinical research on heart rate irregularities\textsuperscript{14-18}, congestive heart failure, arterial blood pressure irregularities\textsuperscript{17,18}, cerebral ischemia\textsuperscript{19,20}, epileptic seizures\textsuperscript{21,22}, and many other conditions\textsuperscript{23-26}, to comprehend their complexity and eventually create prognostic and diagnostic tools. Likewise, the study of human gait variability can provide a window for understanding whether each individual has a unique “gaitprint” just as each individual has a unique fingerprint. For instance, natural fluctuations in walking (e.g., stride-to-stride fluctuations) follow a certain state of variability, which refers to the idea that human steps never exactly replicate themselves\textsuperscript{27,28}. Therefore, we contend that the variability of human gait is the key to answering this question\textsuperscript{30,31}. Assessing gait characteristics unique to individuals could then be used as an early indicator and predictor of disease and disability.

Movement fluctuations in healthy individuals show long-range correlations closely resembling fractional Gaussian noise (fGn) whose autoregressive coefficient \( f \) decays with lag \( k \) in a power-law fashion:
\[
\rho_k = \frac{1}{2} (|k|^{-2 + f} - 2 |k|^{-2 + f} + |k|^{-1 + f})^{22}.
\]
The Hurst exponent, \( H \), can then be used as an early indicator and predictor of disease and disability. Pink noise is a fractal, or statistically self-similar with fluctuations scaling invariantly with time, as suggested by the above autocorrelation equation\textsuperscript{33}. The lack of a pink noise structure is referred to as a “white noise” structure, or lacking long-range correlations (i.e., \( H_{fGn} = 0.5 \)). Finding a pink noise structure in a measurement series suggests that the underlying causal processes operate simultaneously across several timescales and provide a “persistent,” long-range correlated structure in the time-evolving behavior. So far, nonlinear analytical methods quantifying long-range correlations have been used to characterize the healthy structure of variability in different gait parameters\textsuperscript{32-34}, and to distinguish healthy temporal structures of variability from those found during aging and disease\textsuperscript{37-40}. These nonlinear analytical methods have never been used to identify gait characteristics unique to individuals. Any attempt to identify gait characteristics unique to individuals by leveraging variability inherent to the human gait would require a large gait dataset that, at the very least, fulfills the following conditions:

- reflects walking in everyday life (overground as opposed to walking on a treadmill)
- includes nuances necessary to provide sufficient variability in gait parameters inherent to daily walking (e.g., when taking turns, varying walking speed, etc.)
- contains huge quantities of data per individual—preferably spanning across trials to hours to days to weeks—
- to identify which gait parameters remain consistent within the individual and which parameters vary
- yields a variety of gait parameters—lying in the spatial, temporal, and spatiotemporal domains—to increase
- the chances of identifying the latent relationships among different gait parameters

However, to our knowledge, no gait dataset exists that fulfills those conditions, as evident by a sampling of existing gait datasets in Supplementary Table 1. Supplementary Table 1 provides a survey of existing gait datasets and supports the above claim. We review only datasets published after 2010 as motion capture technology has significantly advanced in the last decade, including rapid improvements to inertial measurement unit (IMU)-based motion recording devices. Note that this survey is not exhaustive, and we may have missed several important gait datasets, but the sole purpose of this exercise is to put the present gait dataset in context and to provide interested readers with a glimpse of state-of-the-art datasets. These datasets provide critical gait data in multitudinous walking conditions from multiple countries spanning all around the globe and are fueling gait research in meaningful ways, as affirmed by the thousands of citations they have accumulated. However, these datasets also have several limitations that preclude their utility for assessing gait characteristics unique to individuals using the most advanced nonlinear analytical methods:

- **A singular focus on straight walking.** With a few exceptions\textsuperscript{41}, almost all existing gait datasets focus on straight walking over an even surface. Straight walking is not fully representative of real-life walking, which involves curved paths, sharp turns, uneven surfaces, and countless obstacles. During curved walking, for example, a combination of motor strategies come into play to stabilize the head in space and orient the entire body in the desired direction\textsuperscript{42-44}. Consequently, curved walking challenges spatiotemporal, stability, and symmetry-related gait patterns\textsuperscript{44-46}. A gait dataset representative of real-world walking must have at least a few elements of curved walking.

- **Lack of dynamic optic flow.** Existing gait datasets mostly constitute data collected while walking on the treadmill in a gait laboratory. As opposed to walking in a natural environment where optic flow is highly dynamic, optic flow while walking on a treadmill is relatively static. Treadmill walking has additional constraints such as limited space to walk, possible dangers if the subject is too far back or too close to the side, and a velocity vector from the belt that is continuously added to the subject’s velocity vector. Given the critical role of optic flow in the perceptuomotor control of walking\textsuperscript{45,46} and the role of these constraints on moderating walking dynamics\textsuperscript{47-54}, these datasets miss a very important component influencing gait patterns. A more useful gait dataset must be based on walking in a dynamically changing visual environment.

- **Reliance on laboratory-grade movement registration devices.** Many existing gait datasets have been collected using expensive, laboratory-grade movement registration devices (e.g., instrumented treadmills, and
optical motion trackers with multiple cameras). Often, this gold standard gait research may cost upwards of a hundred thousand dollars. There is no doubt that these devices provide high-precision and high-accuracy recordings with potentially high impact. However, not all institutions and investigators can receive consistent and substantial funding to be able to have access to these facilities. At the same time, a reliance on those laboratory-grade devices affects practitioners whose subjects may not be able to enter the laboratory due to injury, disease, or disability. These factors severely limit the expansion and utility of these datasets. To circumvent these challenges, a gait dataset, collected using a portable and more easily accessible yet accurate motion registration device, allows systematic expansion of the previous datasets to subjects restricted to in home or assistive-care settings and comparison with future datasets.

- **Emphasizing sample size over an individual.** The typical approach to existing gait datasets has been to focus on breadth (e.g., a large number of subjects) at the cost of depth (e.g., a large amount of data per subject). The motivation behind this approach is a habit of blindly analyzing a wide array of shallow data without using human movement expertise to decide on included features to build appropriate models a priori. Nonetheless, a gait dataset appropriate for assessing gait characteristics unique to individuals must also involve a large amount of data per individual and domain expertise (e.g., biomechanics, human movement variability) to train models with carefully curated features of gait that have proven track records of characterizing human movements.

- **Limited scope for analyses.** A major drawback of the existing gait datasets is the small number of gait parameters used or calculated. For instance, joint angle trajectories are not enough to determine step width variability, which are significant predictors of fall risk in older adults, and can differentiate older adult fallers from non-fallers after a slipping event. Likewise, a dataset singularly focused on foot placement variables (e.g., step length, stride length, step width) is not equipped to investigate the coordination patterns underlying specific features of imperceptible gait pattern fluctuations. On the other hand, a comprehensive study of gait variability can incorporate as many as 16 gait parameters at the scale of individual steps in addition to the joint angle amplitudes, velocities, and accelerations belying these gait parameters. The literature is hungry for a gait dataset that collates all these kinematic variables and gait parameters for the same set of individuals to produce a well-rounded investigation of unique gait characteristics—an individual’s “gaiprint.”

- **Test-retest reliability.** Test-retest reliability is a measure of reliability obtained by administering the same test twice over a period of time. The mean and standard deviation may not vary, but the temporal structure may vary across trials. Some previous studies have assessed the retest reliability of gait parameters and found them to be high for many investigated parameters. Notwithstanding, existing gait datasets fail to repeatedly sample subjects across a sufficient timeframe (e.g., days and weeks) to assess test-retest reliability. Furthermore, these datasets cannot identify which gait patterns are unique to the individual and how robust gait patterns are after a retest has been completed. A gait dataset with many trials for each subject collected over two sessions with a rest period of multiple days is necessary to distinguish between reliable and unreliable gait parameters.

The present dataset—which we call the Nonlinear Analysis Core (NONAN) GaitPrint dataset—overcomes all the above-mentioned limitations, providing a valuable resource for assessing gait characteristics unique to individuals based on established principles of human movement variability. The database provides whole-body kinematics and foot placement variables during self-paced overground walking on a 200-meter looping indoor track. Noraxon Ultium Motion IMUs sampled the motion of 35 healthy young adults (19–35 years old) at 200 Hz over 18 4-min trials across two days. Many discrete and continuous variables such as velocity, acceleration, and cadence are outlined in the section “Data processing and extraction of foot placement variables.”

**Methods**

**Subjects and ethical requirements.** Thirty-five adults (18 men and 17 women; mean ± 1 s.d. age: 24.6 ± 2.7 years; height: 1.73 ± 0.78 m; body mass: 72.44 ± 15.04 kg) participated in the present study. Subjects were identified and recruited by word of mouth, campus-wide emails, and Facebook/Twitter posts that link to webpages owned by the University of Nebraska at Omaha. This study was also allowed to use recruiting flyers distributed throughout the community at universities, clinics, health centers, gyms, fitness classes, libraries, cafes/ coffee shops, retirement homes, stores, and community bulletin boards. Subjects were paid $20 per session for participation. Interested individuals were recruited only if they (i) were able to have access to these facilities. At the same time, a reliance on those laboratory-grade devices affects practitioners whose subjects may not be able to enter the laboratory due to injury, disease, or disability. These factors severely limit the expansion and utility of these datasets. To circumvent these challenges, a gait dataset, collected using a portable and more easily accessible yet accurate motion registration device, allows systematic expansion of the previous datasets to subjects restricted to in home or assistive-care settings and comparison with future datasets.

**Test-retest reliability.** Test-retest reliability is a measure of reliability obtained by administering the same test twice over a period of time. The mean and standard deviation may not vary, but the temporal structure may vary across trials. Some previous studies have assessed the retest reliability of gait parameters and found them to be high for many investigated parameters. Notwithstanding, existing gait datasets fail to repeatedly sample subjects across a sufficient timeframe (e.g., days and weeks) to assess test-retest reliability. Furthermore, these datasets cannot identify which gait patterns are unique to the individual and how robust gait patterns are after a retest has been completed. A gait dataset with many trials for each subject collected over two sessions with a rest period of multiple days is necessary to distinguish between reliable and unreliable gait parameters.

The present dataset—which we call the Nonlinear Analysis Core (NONAN) GaitPrint dataset—overcomes all the above-mentioned limitations, providing a valuable resource for assessing gait characteristics unique to individuals based on established principles of human movement variability. The database provides whole-body kinematics and foot placement variables during self-paced overground walking on a 200-meter looping indoor track. Noraxon Ultium Motion IMUs sampled the motion of 35 healthy young adults (19–35 years old) at 200 Hz over 18 4-min trials across two days. Many discrete and continuous variables such as velocity, acceleration, and cadence are outlined in the section “Data processing and extraction of foot placement variables.”

**Sample size justification.** We expect that many of the statistical analyses performed on the present dataset will involve either statistical comparisons distinguishing between clinical groups (e.g., young vs. older adults, healthy adults vs. stroke survivors) or machine learning to identify distinguishing gait patterns among individuals. Excitingly, this dataset constitutes the first batch of subjects from an ongoing project that involves five other populations: healthy middle-aged adults (36–55 years old); healthy older-aged adults (56 + years old); lower-limb amputees; post-stroke patients; and patients with peripheral arterial disease (PAD). Hence, the issue of statistical power can be addressed by first considering the magnitude of differences between any of these groups that can be detected with reasonable probability. To address this issue within the context of mixed-effects models—which constitute the most common method for linear modeling, we simulated datasets ($n = 1,000$) based on effect sizes reported in a recent meta-analysis that identified systematic differences in young and old adults in terms 1/f
for 2.5 seconds with their arms at their side and their feet close to shoulder-width apart. They then walked for 15 seconds at a self-selected pace, made a 180° turn, walked back to the starting position, and stood still due to age were smaller than those due to clinical group. Based on these simulations, a sample size of 30 healthy adults will provide 98% power to detect a standardized effect size of 0.20 (a small effect defined by Cohen), assuming a 5% type 1 error rate. Hence, the current sample size of 35 healthy adults should allow reliable comparison with previously published datasets. Finally, to fully characterize each individual’s gait patterns, instead of collecting a limited amount of data for a larger number of subjects, we decided to collect a huge amount of data per subject over multiple days for a modest number of subjects to confer benefits to future analyses.

Setup and procedures. Kinematic data were collected during self-paced overground walking on a 200 m indoor track at the University of Nebraska at Omaha on two separate days spread one week apart (one subject—identified as S017—did not return for the second session). Each session lasted for a maximum of two hours and thirty minutes. Subjects were requested to wear semi-tight athletic clothing and shoes that they were comfortable walking in for no longer than 2.5 hours. Subjects were screened and consented at the Center for Research in Human Movement Variability (MOVCENTR), housed within the Biomechanics Research Building at the University of Nebraska at Omaha. Per session, subjects performed nine four-minute overground walking trials at their self-selected walking speed on an indoor running/walking track at the adjacent Health & Kinesiology Building. The trials were spread over three blocks of three trials each, with a self-chosen resting period between consecutive blocks of up to five minutes.

Kinematic data during walking were collected at 200 Hz using Noraxon Ultium Motion™ IMUs (Noraxon, Inc., Scottsdale, AZ). Sensors were placed on the following parts of the subject’s body, as recommended by Noraxon: head (middle of the back of the head), upper thoracic (below C7 in line with the spinal column, but high enough to not be affected by upper trapezius muscle movement), lower thoracic (in line with the spinal column at L1/T12; strap belt was positioned on lower ribs on the front side of the body), pelvis (body area of the sacrum), upper arm (midway between the shoulder and elbow joints, lateral to the bone axis), forearm (posterior and distal, where there is a low amount of muscle tissue), hand (dorsal), thigh (frontal and distal half, where there was a lower amount of muscle displacement during motion), shank (front and slightly medial to be placed along the tibia medial surface of the tibias), and foot (upper foot, slightly below the ankle; Fig. 1). All sensors, except those attached to the foot, upper thoracic, lower thoracic, and pelvic, were clipped onto a Velcro strap tightened around the respective body segment to ensure minimal motion artifact but loose enough to promote movement. At the foot, the sensor was clipped into a rubber strap going over the dorsal side of the foot wrapped around the back of the heel and under the arch of the foot. The pelvis and thoracic sensors were clipped into a small plastic clip before placement. The upper thoracic sensor was taped to the upper thoracic area on subjects’ clothing or, when possible, on the subject’s skin. At the lower thoracic and pelvic area, the sensor and platform were secured to a clipped strap tightened around the torso.

Subjects’ body mass, height, and the following anthropometric measures were collected before the first experimental session to reconstruct and scale the biomechanical gait model in the MyoResearch 3.18.126 software: skull height (mental protuberance to skull vertex), shoulder width (inter-acromion joint distance), lumbar + thoracic (C7 to S1), pelvis width (inter-anterior superior iliac spine distance), upper arm (acromion process to lateral humeral epicondyle), forearm length (lateral humeral epicondyle to radial styloid process), hand length (ulnar notch to the tip of the third phalange), thigh length (greater trochanter to lateral femoral epicondyle), shank length (lateral femoral epicondyle to lateral malleolus), and foot length (length of show). Several validation studies have shown that the kinematic data obtained from the Noraxon IMU sensors closely match the kinematic data obtained using traditional infrared motion capture systems.

A functional walking calibration procedure was completed before each trial. First, subjects stood still for 2.5 seconds with their arms at their side and their feet close to shoulder-width apart. They then walked for 15 seconds at a self-selected pace, made a 180° turn, walked back to the starting position, and stood still.
for another 2.5 seconds in the same initial posture. Next, subjects were instructed to walk for four minutes at a self-selected pace that they could maintain for at least 30 minutes and to maintain that same speed throughout the four-minute trial. Subjects walked on the track either clockwise or counterclockwise, with the order of the direction decided by the rules of the track (e.g., trials completed on Monday, Wednesday, Friday, or Sunday were walked clockwise; in 423 trials (69.12%) the participants walked clockwise, 423 trials (30.88%) they walked counterclockwise). Two investigators always walked behind subjects with a cart holding the Ultium Motion receiver connected to a laptop computer while simultaneously monitoring data quality and subject safety. At the end of each four-minute trial, subjects were verbally instructed to stop walking. They then returned to the initial position for calibration, following which the subsequent trial began. All subjects started to walk from the same position on the track. A five minute rest period was given every three trials if needed.

**Data processing and extraction of foot placement variables.** Before exporting the data for the present database, the following post-processing options were applied by the Noraxon MyoResearch 3.18.126 software. First, the processing fusion mode was set to “standard”—the default fusion setting that uses an adaptive filtering technique to optimize sensor tracking while considering the cleanliness of the magnetometer data while recording. Course Stabilization was set to "Foot, Shank, Thigh, Spine," which acts as a high-pass filtering operation on the secondary joint angles to remove sensor drift using a ten-second sliding window. This applied stabilization is recommended for severe magnetic interference affecting the whole body and stabilizes the foot, shank, and thigh from the top down. It also stabilizes the upper and lower spine segments relative to the pelvis. The final setting, progression, was set to "translation." Secondary knee angles were turned on to determine the left and right knee abduction and external knee rotation. Acceleration was set to “sensor-based” to capture acceleration data with respect to the coordinate frame of each sensor. The software detected heel contacts using the gyroscope and accelerometer data of the foot sensors to determine stance and swing as an on/off signal. The Noraxon MyoResearch 3.18.126 software also provides an anti-wobbling correction that was applied to smooth the data at 5 Hz with a 300 ms residual to remove soft tissue artifacts.

For each trial, the following variables were exported by the Noraxon MyoResearch 3.18.126 software: the acceleration, velocity, position, and orientation of each sensor and the acceleration, velocity, position, orientation, and rotational velocity of each corresponding body segment, and the angle of each respective joint. For each trial, the following gait parameters, units in parentheses, were determined using the filtered time series data, as defined previously:

**Spatiotemporal parameters**

- **Step length (cm)—**the distance from one heel strike to the next heel strike of the opposite foot.
- **Stride length (cm)—**the distance between two consecutive heel strikes of the same foot.
- **Step width (cm)—**the lateral distance between the heel center of one heel strike and the line joining the heel center of two consecutive heel strikes of the opposite foot.
- **Distance traveled (m)—**the distance the subject traveled as tracked by the pelvis sensor.

**Temporal parameters**

- **Cadence (steps/min)—**the number of steps per minute, also referred to as step rate.
- **Step time (s)—**the time elapsed from the initial contact of one foot to the initial contact of the opposite foot.
- **Stride time (s)—**the time elapsed between the initial contacts of two consecutive footfalls of the same foot.
- **Stance time (s)—**the time elapsed between the first and last contacts of a single footfall (the stance phase starts at heel contact and ends at toe off of the same foot).
- **Swing time (s)—**the time elapsed between the last contact of the current footfall and the first contact of the following footfall of the same foot (the swing phase starts with toe off and ends with the first contact of the same foot).
- **Single support time (s)—**the total time one foot is in contact with the ground throughout the gait cycle.
- **Double support time (s)—**the total time both feet are simultaneously in contact with the ground throughout the gait cycle.

**Temporophasic parameters**

- **Stance time (%Stride time)—**stance time normalized to stride time.
- **Swing time (%Stride time)—**swing time normalized to stride time.
- **Single support time (%Stride time)—**single support time normalized to stride time.
- **Double support time (%Stride time)—**double support time normalized to stride time.

**Spatiotemporal parameters**

- **Gait speed (m/s)—**the ratio of distance walked and trial time.
- **Stride speed (m/s)—**the ratio of stride length and stride time.

The MATLAB (Mathworks, Inc., Natick, MA) script used to determine these gait parameters—GaitPrint_Spatiotemporal_Calculation.m—is provided as part of the database on figshare.
Data Records

All data is made available using Figshare\textsuperscript{33}. The deidentified subject information is stored in the file named GaitPrint_Subject_Characteristics.csv, describing for all subjects the screening outcomes, age, sex, body mass, height, hand and food dominance, shoe make and model, and anthropometrics measures: skull height, shoulder width, lumbar + thoracic, pelvis width, upper arm, forearm length, hand length, thigh length, shank length, and foot length. An additional file, GaitPrint_Trial_Characteristics.csv, described for all subjects the session date and start time, shoe make and model, walking direction, start time, first step taken, and any notes taken during the trial. All kinematic data have been grouped into subject-specific folders and .csv files, each file bearing the subject’s alphanumeric code (e.g., S001\_G01\_D01\_B01\_T01). All data files have a tabular structure with the following headers:

<table>
<thead>
<tr>
<th>ID</th>
<th>Group</th>
<th>Day</th>
<th>Block</th>
<th>Trial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For each subject, the ID field codes the subject number. It is labeled “S001,” “S002,” “S003,” … for the 35 subjects. The Group field codes the study population. It is labeled “G01” for the current dataset. The Day field codes whether the data was collected during the first or second day. It is labeled “D01” and “D02” for Day 1 and Day 2, respectively. The Block field codes the block number within a session. It is labeled as “B01,” “B02,” and “B03” for Block 1, Block 2, and Block 3, respectively. The Trial field codes the trial number within a block. It is labeled as “T01,” “T02,” and “T03” for Trial 1, Trial 2, and Trial 3, respectively.

S###.zip (Raw Data). These folders contain one .csv file for each trial completed by the participant. For example, S001.zip contains 18 .csv files each as a table providing all the raw data from that subject’s trial. The data are arranged in a matrix of 48,000 rows by 321 columns, with one row per timestamp. The first column, “Time” provides the timestamps, in milliseconds. The next 320 columns each provide the kinematic variables exported by the Noraxon MyoResearch 3.18.126 software: the acceleration, velocity, position, and orientation of each sensor, and the acceleration, velocity, position, orientation, and rotational velocity of each corresponding body segment, and the angle of each respective joint.

Spatiotemporal_variables.zip. This folder contains one .csv per trial where each file contains a table providing all the spatiotemporal variables listed above. The data are arranged in 26 columns, with one row per sample in the time series of each gait parameter: cadence (steps/min), step time (s), left step length (cm), right step length (cm), right step width (cm), right stride length (cm), right stride time (s), right stride time (s), right stance time (s), right swing time (s), single support time (s), double support time (s), left pct stance (%GC), right pct stance (%GC), left pct swing (%GC), right pct swing (%GC), pct single (%GC), pct double (%GC), average speed (m/s), left stride speed (m/s), right stride speed (m/s), and distance traveled (m).

We acknowledge that numerous researchers, particularly those affiliated with teaching institutions, may encounter obstacles when delving into the analyses of spatiotemporal variables or raw data. These hurdles may include limited access to paid software tools or a potential deficiency in coding, particularly in languages such as Python or R. To foster accessibility and facilitate secondary research endeavors involving this data, we have taken the initiative to offer all spatiotemporal variables and raw data for each subject and trial as distinct .csv files. Spatiotemporal files are conveniently bundled within a compressed folder titled “Spatiotemporal_Variables.zip” and each subject’s raw data within compressed folders titled “S###.zip.” In addition, we have provided a compressed folder titled “template_scripts.zip” containing basic code examples with MATLAB, Python, and R extensions. This approach serves to lower the barrier to entry, enabling a wider range of researchers to engage with the data effectively.

Technical Validation

Continuous phase relationship between the right and left limb segments reveals highly consistent inter-limb coordination over multiple gait cycles. The lower extremity segments of a walking individual can be considered a coupled system, and the interaction between the segments effectively moves the body forward. The behavior of such a dynamical system can be described by plotting a variable versus its first derivative—the phase portrait—that quantifies human movement\textsuperscript{94}. The phase portraits of limb segments resemble a continuous relative phase, $\Phi$, describes the phase space relation between two segments in a form of a low-dimensional variable that reflects changes in phase relationship throughout the movement. Hence, this makes continuous relative phase an appealing variable to assess inter- and intra-limb coordination\textsuperscript{96–98}. $\Phi$ was defined for the following joint pairs: (i) right and left thigh pitch angle, (ii) right and left shank pitch angle, and (iii) right and left foot pitch angle, for each gait cycle\textsuperscript{100}. Angular position data were first normalized to 100 data points, and centered so that the calculated segment phase angles are oriented around 0° $\left(\Psi_{segment} = x(t) - \min(x(t)) - (\max(x(t)) - \min(x(t))) / 2\right)$. Then, to minimize the error caused by the phase portrait method, analytical signals were obtained via Hilbert transform $l(t) = H_{\Psi_{segment}}(t) + iH(t)$. The segment phase angle of each data point was calculated using the real part and the imaginary part that are computed through the Hilbert transform $\left(\phi_{segment}(t) = \arctan(H(t) / x(t))\right)$. $\Phi$ was calculated as the difference between the right
and left segment phase angles at each time point \( \Phi(t_i) = \Phi_{\text{right segment}}(t_i) - \Phi_{\text{left segment}}(t_i) \). When \( \Phi \) is 0°, it indicates that the phase space relationship between the segments is symmetrical and is called in-phase. On the other hand, when \( \Phi \) is 180°, the two segments are in an anti-phase relation, i.e., out of phase. \( \Phi \) can capture changes in limb coordination across the gait cycle of healthy adults\(^{106,107} \), the effects of inter- and intra-limb limb coordination due to cognitive load and walking speed\(^{103,104} \), changes associated with development\(^{106} \) and aging\(^{106} \), neuropathy\(^{107} \), injury\(^{108} \), recovery post-injury\(^{109,110} \), and amputation\(^{108} \). \( \Phi \) was calculated across all gait cycles within each self-paced overground walking trial.

Figure 2b–d show violin plots of \( \Phi \) between the right and left segment pitch angles for thigh, shank, and foot for five representative subjects. The obtained \( \Phi \) for all three segment pairs hover around 180°, indicating an expected out-of-phase relationship between the right and left limb segments characteristic of healthy gait. Additionally, the high intra-individual consistency in \( \Phi \) between the two days further strengthens our assertion about the reliability of our data. This session-to-session reliability of the obtained \( \Phi \) point toward the uniqueness of individual gait patterns across both sessions\(^{111} \).

While most subjects showed \( \Phi \) close to 180°, we found a total of 13 trials across four subjects with \( \Phi \) deviating considerably (e.g., Subject 031, as indicated in Fig. 2d). To understand the cause of these anomalous \( \Phi \), we explored these trials in further detail. We found altered mean joint angle trajectories across all gait cycles within these trials, as well as a greater standard deviation across all gait cycles (see joint angle trajectories from one such anomalous trial—for Subject 031—contrasted with a valid trial; Fig. 3). Hence, the continuous phase relationship between the right and left limb segments was also diagnostic of data quality.

**Lyapunov exponent (\( \lambda_1 \)) values reveal highly consistent trajectories over multiple gait cycles.** Walking, in part, results from the repetitive rotation of limb segments around a joint (e.g., ankle, knee, hip) which is a cyclical process. This repetition creates trajectories in three-dimensional space that can be tracked, recorded, and plotted by motion capture equipment and software. One way to study the variability of walking trajectories is to inspect how the repetitive rotation of these joints changes over many gait cycles\(^{27} \). That is, do limb segments trace out the exact same trajectory over multiple gait cycles? Or, do movement trajectories deviate over time? That is, do movements converge to a common trajectory, or do they diverge? The (Largest) Lyapunov Exponent (\( \lambda_1 \)) provides a direct measure of the divergence of trajectories by examining trajectory behavior within the so-called reconstructed “state space attractor”—a set of states that a dynamical system tends to evolve toward over various initial conditions\(^{112–115} \) (Fig. 4a,b). As such, it is one aspect of measuring the stability of system's time-evolving dynamics. Technically, \( \lambda_1 \) measures the rate at which trajectories diverge over gait cycles. This rate reflects the unique functional organization of each person's neuromuscular system\(^{115} \). Changes in this tendency to converge or diverge should reflect how a walker adapts to perturbations during locomotion and hence, qualitative changes in movement trajectory dynamics could reveal the onset of pathology\(^{117–119} \). We used Wolf et al.'s algorithm\(^{112} \) to assess \( \lambda_1 \).

Figure 4c–h show violin plots of \( \lambda_1 \) for segment pitch angles at the thigh, shank, and foot for five representative subjects. The obtained \( \lambda_1 \) is rather small and within close approximation to the values reported in several previous studies\(^{122} \), indicating highly predictable gait cycles characteristic of healthy walking. We would like to also note that while some previous studies reported even smaller \( \lambda_1 \)\(^{123–125} \), these studies used joint angles for estimating \( \lambda_1 \) as opposed to segment pitch angles used in the present dataset. Additionally, we observe high inter-individual variation in \( \lambda_1 \), accompanied by high intra-individual consistency in \( \lambda_1 \) between the two sessions. This session-to-session test-retest reliability of the obtained \( \lambda_1 \) points toward the uniqueness of individual gait patterns across both sessions\(^{111} \).

Several subjects showed large variations in \( \lambda_1 \) across trials (e.g., \( \lambda_1 \) for the right thigh for Subject 001; Fig. 4d). To understand whether this variation reflects anomalous levels of divergence across trajectories, we explored such trials in further detail. We found relatively normal mean joint angle trajectories across all gait cycles within these trials (see joint angle trajectories from one such pair of trials; Fig. 5). Hence, we posit that these highly variable \( \lambda_1 \) reflect overground walking on a looping track, which imposes fewer constraints on gait than a treadmill, and evokes greater divergence among linked segment angles in high-dimensional space due to a wider walking space and the presence of turns. This heightened sensitivity of \( \lambda_1 \) to individuals’ responses to gait constraints also means that \( \lambda_1 \) could serve as a more reliable metric—or an important member of a set of metrics—for assessing gait patterns unique to individuals.

**Close-to-one hurst exponent, \( H_{fGn} \) values indicate the pink noise structure in the time series of stride length and stride time.** The optimal movement variability hypothesis suggests that human gait variability (i.e., the fact that human steps never repeat themselves exactly) exhibits certain characteristics and optimal forms typically found in healthy individuals\(^{27,28,126} \). Measurements of physiological, motor, and cognitive variability in healthy individuals show varying degrees of long-range correlations\(^{127–131} \), and their loss due to aging and disease\(^{25–26} \). The Hurst exponent \( H \) captures long-range correlations in terms of the moments of the autocorrelation, which diverge for 0.5 < \( H_{fGn} \) ≤ 1\(^{122,123} \) (Fig. 6a). An fGn time series with \( H_{fGn} < 0.5 \) is anti-persistent, i.e., an increase will most likely be followed by a decrease and vice-versa. This means that future values tend to return to a long-term mean. An fGn time series with \( H_{fGn} = 0.5 \) resembles a random walk, implying no correlation between the current and future values. An fGn time series with \( H_{fGn} > 0.5 \) is persistent, i.e., an increase will most likely be followed by another increase and vice-versa. The very notion of the mean is meaningless for persistent time series. As mentioned above, biologists and psychologists have encountered the mathematical structure of \( H_{fGn} > 0.5 \) under the label of “1/fnoise” or “pink noise” (e.g.,\(^{130,134} \), \( H_{fGn} \) or alternatively, the fractal scaling exponent \( \alpha \), has been used extensively to analyze the temporal structure of gait variability during walking.
Fig. 2 The right and left segment pitch angles showed an out-of-phase continuous phase relationship. (a) A representative sample of the right and left thigh pitch angle used to calculate the phase relationship between the right and left limbs. (b–d) Violin plots of the continuous relative phase, $\Phi$, across all gait cycles within a trial between the right and left segment pitch angles for five representative subjects. (b) Violin plots of $\Phi$ for thigh pitch angle. (c) Violin plots of $\Phi$ for shank pitch angle. (d) Violin plots of $\Phi$ for foot pitch angle. Each violin plot shows $\Phi$ for the two days (Day 1 on left, and Day 2 on right). As expected, $\Phi$ for all three segment pairs hover around 180°, indicating an out-of-phase relationship between the right and left limb segments. Additionally, the high intra-individual consistency in $\Phi$ between the two days further strengthens our assertion about the validity of our data. Vertical lines represent the interquartile range of relative phase, and colored horizontal bars represent the mean of $\Phi$ ($n=9$ trials) for the respective subject/day.
and running, while locomoting overground or on a treadmill\textsuperscript{135–137}. The values of $H_{fGn}$ depend on task constraints such as the walking speed\textsuperscript{138,139} and can distinguish between healthy gait in young adults and altered gait due to aging and disease\textsuperscript{140–142}.

We specifically tested two gait parameters—spatial (Stride length) and temporal (Stride time)—to verify the presence of pink noise structure characteristic of a healthy, adaptable gait. We assessed the strength of long-range correlations in the time series of stride length and stride time following the Bayesian approach developed by Tyralis and Koutsoyiannis\textsuperscript{143}. We chose this method because it performs well on short time series, as compared to more common methods like detrended fluctuation analysis\textsuperscript{144,145}, which requires much longer time series data, and yields $H_{fGn}$ comparable to the more commonly used detrended fluctuation analysis\textsuperscript{146}.

Figure 6b,c represents violin plots of $H_{fGn}$ values for the time series of Stride length and Stride time, respectively. The obtained $H_{fGn}$ values indicate the pink noise structure in stride length and stride time, both of which are characteristic of a healthy gait\textsuperscript{40,147–150}. Shuffling the original series yields $H_{fGn}$ values resembling white noise, further confirming long-range correlations in the time series of Stride length and Stride time. Additionally, we observe high inter-individual variation in $H_{fGn}$ values, accompanied by high intra-individual consistency in $H_{fGn}$ values between the two sessions. This session-to-session test-retest reliability of the obtained $H_{fGn}$ values points toward the uniqueness of individual gait patterns across both sessions\textsuperscript{111}.

Strengths and limitations. The present dataset has several strengths that make it unique among existing gait datasets. First, the present dataset reproduces many characteristics established in the literature on healthy human gait variability. Second, the application of several nonlinear analytical methods require long time series that span several thousand samples. An unusually long series of each of these variables ($>40,000$ samples for continuous variables and $>2,000$ samples for discrete variables, as opposed to $\approx 2000$ and $\approx 200$ samples, respectively, typical of existing datasets) makes the present dataset amenable to nonlinear analyses of gait variability which would be impractical with short series typical of existing datasets. Third, multiple recordings ($n = 18$) for each subject imply that the present dataset will significantly aid the modeling efforts to assess the uniqueness of individual gait patterns. In other words, the present database will aid the exploration of whether each person has a distinct “gaitprint” (i.e., gait characteristics) similar to how each person has a distinct fingerprint. If individuals...
Fig. 4 The largest Lyapunov Exponent, $\lambda_1$, describes the limb segment's diverging trajectories over multiple gait cycles. The motion of thigh, shank, or foot over multiple strides can be depicted as a series of neighboring trajectories in a time-delay embedding. The embedding dimension is calculated from the False Nearest Neighbors algorithm, which detail the number of time delay copies required to accurately replicate movement trajectory dynamics (i.e., an attractor). This example shows an embedding dimension of 3 (original plus 2 copies), for viewing purposes. The time, $\tau$, is determined through the Average Mutual Information algorithm. (a) Each of the time series (original plus copies) are plotted to provide the reconstructed state space attractor. This is an example of the thigh segment angle time series embedded into 3 dimensions. (b) Inset from a showing the calculation of the Lyapunov exponent. Conceptually, a single point along a trajectory and its true nearest neighbor are selected, shown by a red line connecting two dots in the upper left corner of (b) and the Euclidean distance between these points is calculated ($d_t$). These points are then followed along their respective trajectories for a certain number of time points ($n$), followed by distance recalculation ($d_t'$, shown by red lines connecting red dots). A new true nearest neighbor is selected, shown by a red line connecting two dots in the bottom right corner of (b) and the process is repeated. The log of the ratio of these distances is calculated and then normalized to the time the points traversed through the trajectory to yield the Lyapunov exponent. A smaller $\lambda_1$ indicates higher predictability, or less divergence, of the dynamical system. (c–h) Violin plots of $\lambda_1$ for the segment pitch angles for five representative subjects. (c) Left thigh. (d) Right thigh. (e) Left shank. (f) Right shank. (g) Left foot. (h) Right foot. Each violin plot shows $\lambda_1$ for the two days (Day 1 on left, and Day 2 on right). As expected, these $\lambda_1$ indicate gait cycle sensitivity that is dependent on initial conditions and is characteristic of healthy walking. Additionally, the high inter-individual variation in $\lambda_1$ is accompanied by high intra-individual consistency in $\lambda_1$ between the two days, further strengthening our assertion about the validity of our data. Vertical lines represent the interquartile range of $\lambda_1$, and colored horizontal bars represent the mean of $\lambda_1$ ($n = 9$ trials) for the respective subject/session.
have a unique gaitprint, then they should show consistency in gaitprint-defining statistical features across multiple recordings and sessions. This way, the present database could produce fundamental knowledge critical for predicting disease, predicting physiological declines, and improving rehabilitation.

The present dataset also has one fundamental weakness: the motion capture system recording the subjects' motion. The sensors use the earth's magnetic field to detect motions in the horizontal plane; hence, small variations in the magnetic field lead to drift; "an unwanted sliding movement on the horizontal plane of the complete biomechanical model in the motion capturing software". Furthermore, metal and magnetic objects in the measurement space can also alter this magnetic field, leading to phantom changes in direction. However, no vertical drift is observed, as no external factors aid the calculations in the vertical plane. Although the drift associated with the Noraxon Ultium Motion is much smaller than observed while using the previous generations of inertial sensors, a robust drift correction algorithm is recommended to enhance data accuracy. Drift correction is an area of ongoing development, and we recommend implementing whatever state-of-the-art drift correction algorithm is available when analyzing the present dataset. However, we strongly believe that this weakness does not undermine the purported utility of the present dataset, as we have already shown the validity of the present dataset using several linear and nonlinear measures characteristic of healthy human gait variability.

Usage Notes
We provide the present dataset as zipped folders in the data repository and also provide the scripts supporting analyses in MATLAB, Python, and R working environments; existing MATLAB and Python open-source software packages focused on gait could be used to analyze the raw data. For example, GaitPy is a Python library that provides functions to read and estimate the clinical characteristics of gait from accelerometer data. Likewise, the Kinematics and Inverse Dynamics toolbox for MATLAB provides functions for analyzing joint kinematics and dynamics. Other open-source software packages like biomechZoo could also be used to process, analyze, and visualize the present dataset. MATLAB and Python functions needed to perform nonlinear time series analyses on foot placement variabilities—such as those presented in the

Fig. 5  Mean joint angle trajectories across all gait cycles within two representative self-paced overground walking trials. Both subjects show λ close and far, from the mean as shown in Fig. 4. Shaded areas indicate standard deviation across all gait cycles. Note that despite significant anomaly in λ, the mean joint angle trajectories and standard deviation across all gait cycles can be atypical or irregular across trials (i.e., left vs. right panels).
“Technical validation” section—can be obtained from the GitHub repository of the Nonlinear Analysis Core at the University of Nebraska Omaha (https://github.com/Nonlinear-Analysis-Core/NONANLibrary).

To facilitate the use of these software packages, the present dataset includes a MATLAB script entitled GaitPrint.m, which contains the function that computes all the spatiotemporal variables.

Finally, as part of an ongoing project, we plan to make publicly available gait datasets on five other populations:

- Group 2: Healthy middle-aged adults (36–55 years old)
- Group 3: Healthy older-aged adults (56+ years old)
- Group 4: Lower-limb amputees
- Group 5: Post-stroke patients
- Group 6: Patients with peripheral arterial disease (PAD)

Together these gait datasets will support and standardize the decision of researchers, clinicians, and therapists when assessing gait abnormalities or tracking the outcomes of rehabilitation interventions in older adults.
and clinical populations. They will also provide the reference database necessary for investigating whether healthy adults show unique individual gait patterns and what happens to these unique gait patterns with frailty associated with aging and disease. Users of our work will notice that the currently available dataset includes nonconsecutive subject IDs because our six groups were recruited and collected simultaneously. The inclusion of Groups 2–6 are to be released in future work.

Code availability

The Matlab code GaitPrint.m, template scripts (MATLAB, Python, R), and all associated functions used for post-processing of all raw data are available as part of the database on figshare.

Received: 8 February 2023; Accepted: 31 October 2023; Published online: 05 December 2023

References


**Acknowledgements**

This work was supported by the Center for Research in Human Movement Variability at the University of Nebraska at Omaha, the University of Nebraska Collaboration Initiative, the NSF award 212491, the NIH awards P20GM109090 and R01NS114282, the NASA EPscOR mechanism, and the IARPA BRIAR WatchD award.

**Author contributions**


**Competing interests**

The authors declare no competing interests.