www.elsevier.com/locate/jbiomech

Gait Biomechanics in the Era of Data Science

Reed Ferber, Sean T. Osis, Jennifer L. Hicks, Scott L. Delp

PII: S0021-9290(16)31133-2
DOI: http://dx.doi.org/10.1016/j.jbiomech.2016.10.033
Reference: BM7943

To appear in: *Journal of Biomechanics*
Accepted date: 21 October 2016

Cite this article as: Reed Ferber, Sean T. Osis, Jennifer L. Hicks and Scott L Delp, Gait Biomechanics in the Era of Data Science, *Journal of Biomechanics* http://dx.doi.org/10.1016/j.jbiomech.2016.10.033

Perspectives Article

Gait Biomechanics in the Era of Data Science

Reed Ferber (PhD)[1,2,3], Sean T. Osis (MSc)[1,3], Jennifer L. Hicks (PhD)[4], Scott L. Delp (PhD)[4,5,6].

[1] Faculty of Kinesiology, University of Calgary, Calgary, Alberta, Canada
[2] Faculty of Nursing, University of Calgary, Calgary, Alberta, Canada
[3] Running Injury Clinic, Calgary, Alberta, Canada
[4] Department of Bioengineering, Stanford University, Stanford, California, USA
[5] Department of Mechanical Engineering, Stanford University, Stanford, California, USA
[6] Department of Orthopaedic Surgery, Stanford University, Stanford, California, USA

Corresponding author:

Reed Ferber, PhD
Associate Professor, Director: Running Injury Clinic
University of Calgary
2500 University Drive NW
Calgary, Alberta, CANADA
T2N 1N4
E-mail: rferber@ucalgary.ca
Phone: +1 (403) 210-6468
Fax: +1 (403) 289-9117

Keywords: Biomechanics, Gait, Data Science, Machine Learning.

Word count: 182 words (Abstract), 1760 words; 9 double-spaced manuscript pages (main text)

Abstract

Data science has transformed fields such as computer vision and economics. The ability of modern data science methods to extract insights from large, complex, heterogeneous, and noisy datasets is beginning to provide a powerful complement to the traditional approaches of experimental motion capture and biomechanical modeling. The purpose of this article is to provide a perspective on how data science methods can be incorporated into our field to

advance our understanding of gait biomechanics and improve treatment planning procedures. We provide examples of how data science approaches have been applied to biomechanical data. We then discuss the challenges that remain for effectively using data science approaches in clinical gait analysis and gait biomechanics research, including the need for new tools, better infrastructure and incentives for sharing data, and education across the disciplines of biomechanics and data science. By addressing these challenges, we can revolutionize treatment planning and biomechanics research by capitalizing on the wealth of knowledge gained by gait researchers over the past decades and the vast, but often siloed, data that are collected in clinical and research laboratories around the world.

Biomechanical gait analysis is commonly used to analyse sport performance and evaluate pathologic gait. Significant advances in motion capture equipment, research methodologies, and data analysis techniques have enabled a host of studies that have advanced our understanding of the biomechanics of walking and running. Despite these advances, one can argue that the fundamental approach to clinical gait analysis, and gait biomechanics research, has not evolved at the same speed. Clinical gait analysis laboratories continue to operate separately from one another, and analyze each patient in isolation, lacking the tools or time to statistically compare a new patient to the thousands of patients seen in the same lab over the years. Moreover, the vast majority of gait biomechanics research investigates the factors associated with injury, pathology, and performance by focusing on a small number of variables

and experiments performed on a few subjects. The purpose of this article is to share the lessons we have learned about these challenges and promise of using modern data science methods (e.g., machine learning) to analyse the growing amounts of gait biomechanics data collected by researchers and clinicians.

In recent years, sophisticated data science methods have been adopted across disciplines including robotics, genetics, and economics, and increasingly, in biomechanics as well. Data science draws from mathematics, computer science, statistics, and signal processing to develop methods for gaining insight from heterogeneous, noisy, sparse, or otherwise complex data. These methods can process large quantities of data, without exclusively relying on *a priori* knowledge of predictive variables. For example, Facebook uses network analysis and graph mining to locate your friends from high school, and Google uses convolutional neural networks to achieve state-of-the-art performance in speech and image recognition. In the realm of biomechanics, researchers have characterized normal and pathological gait kinematics and kinetics using data science methods such as principal component analysis (PCA) and support vector machines (SVM) (Deluzio et al., 1997; Eskofier et al., 2013; Federolf et al., 2013). As another example, Mansi and colleagues (2012) used a machine learning approach to build finite-element models from medical imaging data to study the biomechanical impact of mitral valve repair.

Applying data science methods with careful attention to validation and knowledge of biomechanics research is providing new insights about how to design effective gait biomechanics studies and how to treat gait pathology. One example comes from a paper recently published in this Journal by the Calgary group (Ferber, Osis), which analysed running

patterns. Many biomechanical investigations examine groups based on injury criteria (e.g., patellofemoral pain), demographic factors (e.g., age, height, sex), or gait speed, under the assumption that movement patterns will be homogeneous within these groups. By contrast, we used a data science approach to identify two distinct kinematic running gait patterns within a single group of healthy runners (Phinyomark et al., 2015). The dataset was collected over a period of two years as part of an initiative to build a database of running and walking biomechanics. We identified two kinematic running gait patterns using a hierarchical cluster analysis without *a priori* knowledge of how many groups would be identified or what variables would best separate the groups. The results revealed that selection of a small number of subjects without accounting for subjects' running type can result in a biased outcome when making clinical comparisons in gait kinematics (e.g., if the treatment arm, by random chance, includes a greater number of group 1 runners than the control arm of a study). Caution is therefore advised when assuming that running gait patterns within a sample will be representative of a population - a data science approach was able to reveal this insight.

Additional examples come from the Stanford group (Delp, Hicks) who, together with collaborators at Gillette Children's Specialty Healthcare, combined biomechanical modeling and data science methods to better understand the causes and improve treatment of gait abnormalities in individuals with cerebral palsy. Musculoskeletal simulation has demonstrated that muscle activity in the period leading up to the swing phase of gait restricts normal knee motion (Goldberg et al., 2003; Goldberg et al., 2004; Reinbolt et al., 2008). By integrating this biomechanical knowledge into a statistical model they were able to predict, from gait analysis and other clinical data, which patients would benefit from rectus femoris muscle transfer

surgery with 88% accuracy (Reinbolt et al., 2009). Similarly, biomechanical modeling has shown that correcting short or slow hamstrings with hamstring-lengthening surgery can lead to improved knee flexion in stance (i.e., crouch gait: Arnold et al., 2006a; Arnold et al., 2006b). This information helped develop a statistical model that was able to predict improvement in crouch gait with 73% accuracy, while in practice only 48% of patients improve after surgery (Hicks et al., 2011). This approach demonstrates that integrating data science methods with experimental and computational biomechanical data can yield robust predictions to improve clinical practice.

Two important challenges must be met to advance the use of data science approaches in biomechanical studies. First, in gait research, data are often collected at different time points on scales from seconds to years. These time points might correspond to a real-time activity like the three-dimensional acceleration of a sensor on the wrist during running, or a clinical measurement like muscle strength taken at six, eight and ten months post-injury, at 36 years of age, or simply as a function of calendar time. In fact, data like these abound in biomechanical studies, particularly when combining different sources of data. To address this challenge, and to build accurate classification and prediction models for gait research, new tools are needed.

Statistical and machine learning tools are being developed to make predictions and identify trends, correlations, and clusters in large-scale, sparse, and irregular time-varying measurements. For example, functional data analysis (e.g., Ramsay, 2006; James et al., 2000; James et al., 2001), is a statistical tool that accounts for the fact that discrete time point data are generated from an underlying, time-varying function. The predictive features (e.g., to identify injury risk) parameterize the time-varying curve(s) and are chosen automatically from

the data based on the ability to discriminate between subjects or predict the outcome of interest. These time-series data can be collected at different sampling points for different subjects and the selected features typically encode more information than the minima, maxima, or other discrete features chosen subjectively by the researcher building the model.

Mixed effects models are another powerful tool for analysing longitudinal data that can separate the fixed effects of interest (e.g., the effect of an injury or pathology on the deterioration of gait over time) from random effects (e.g., the effect of fatigue on a given measurement day). Researchers are also developing advanced machine learning systems that use layers of neural networks or graphical models to automatically learn a set of features or a hierarchical structure to make predictions (Sen et al., 2015; De Sa et al., 2016; Lasko et al., 2013; Razavian and Sontag, 2015).

A second challenge faced by the biomechanics research community is over-fitting of data. Over-fitting artificially inflates the accuracy of a model or significance of a predictive variable and can arise from assessing a model's accuracy on the same data used to train a model or select predictive features. The risk of overfitting is larger if the ratio of the number of subjects to the number of variables is low. Over-fitting reduces the generalisability of results and, even more troubling, may promote errors in the selection of which statistical models best characterize clinical populations. To avoid overfitting, one must employ techniques such as cross-validation (Stone, 1974), which can help give estimates of the ability of a model to make predictions for a new set of subjects. Moreover, overfitting can be avoided by testing with larger and independent datasets.

Thus, more data that comes from a variety of labs and clinics is needed. To apply data science methods, an adequate number of samples must be obtained, and the number grows with the number of variables used in the analysis. While data about movement abounds in biomechanics labs and clinics around the world, the majority of these data are siloed. One approach the Calgary group has taken to overcome this problem is to create a worldwide infrastructure of clinical and research partners linked through an automated 3-dimensional biomechanical gait data collection system. This system continually aggregates data, without relying on *a priori* knowledge to direct data collection modes and analyses. This approach has resulted in the accumulation of walking and running data from over 4000 individuals, and work has just begun to unlock the potential of applying data science methods to help answer clinical and biomechanical questions. Researchers have begun to share their data publicly on Simtk.org and other platforms such as ODHSI.org and CrowdSignals.io. Data sharing is a necessary step to enable new and more comprehensive biomechanical studies to be performed.

Data science will be most successful in the field of gait biomechanics if education and method development occurs as a close collaboration between the two disciplines. The National Institutes of Health (NIH) is supporting this cross-fertilization through its Big Data to Knowledge Initiative (BD2K). For example, the Mobilize Center is a NIH BD2K Center that is developing and applying data science methods in collaboration with gait researchers, with the mission of sharing these resources with the biomechanics community (Ku et al., 2015). The Center is creating new methods for applications in movement biomechanics, including osteoarthritis, cerebral palsy, and running-related injuries, motivated by the unique challenges of analyzing complex movement data. For example, Chris Re and colleagues are pioneering machine

learning systems to efficiently and transparently encode expert knowledge from biomedical researchers when processing and analyzing large, heterogeneous, and noisy data typical of biomechanics (Sen et al., 2015; De Sa et al., 2016). By developing these new approaches, sharing data and validated software tools, and training thousands of researchers, the Mobilize Center collaborates with laboratories around the world, like the Calgary group and other researchers with biomechanical databases seeking to transform human movement research.

In summary, as data science continues to develop, and technology provides the opportunity for gait laboratories to combine and share data, the field of clinical gait analysis will advance. In the near future we will begin to analyze large quantities of data, explore unstructured or complex data, ask open-ended questions based on these data, and develop predictive models that produce new insights. We encourage the biomechanics research community to openly share data between laboratories, learn more about how to employ data science methods, develop data research networks, and join our collective efforts to advance the field of gait biomechanics research.

Conflict of interest statement.

The authors declared that there are no conflicts of interest.

Acknowledgements.

Ferber and Osis are funded primarily by Alberta Innovates: Health Solutions (Grant no: 200700478) and a Discovery Grant (Grant no: 1028495) and Accelerator Award (Award no: 1030390) through the Natural Sciences and Engineering Research Council of Canada (NSERC).

Delp and Hicks are funded through the National Institutes of Health (NIH) Big Data to Knowledge (BD2K) Initiative, which supports the Mobilize Center (Grant no: U54 EB020405) and

the NIH Medical Rehabilitation Research Resource (MR3) Network, which supports the National Center for Simulation in Rehabilitation Research (Grant no: P2C HD065690).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Arnold, A.S., Liu, M.Q., Schwartz, M.H., Ounpuu, S., and Delp, S.L. 2006. The role of estimating muscle-tendon lengths and velocities of the hamstrings in the evaluation and treatment of crouch gait. Gait Posture 23:273-281.

Arnold, A.S., Liu, M.Q., Schwartz, M.H., Ounpuu, S., Dias, L.S., and Delp, S.L. 2006. Do the hamstrings operate at increased muscle-tendon lengths and velocities after surgical lengthening? J Biomech 39:1498-1506.

De Sa, C., Ratner, A., Ré, C., Shin, J., Wang, F., Wu, S., Zhang, C. 2016. DeepDive: Declarative Knowledge Base Construction. SIGMOD Record.

Deluzio KJ, Wyss UP, Zee B, Costigan PA, Sorbie C. Principal component models of knee kinematics and kinetics: normal vs. pathological gait patterns. J Hum Mov Sci 1997;16:201–17.

Eskofier BM, Federolf P, Kugler PF, Nigg BM. Marker-based classification of young-elderly gait pattern differences via direct PCA feature extraction and SVMs. Comput Methods Biomech Biomed Engin. 2013

Federolf PA, Boyer KA, Andriacchi TP. Application of principal component analysis in clinical gait research: identification of systematic differences between healthy and medial knee-osteoarthritic gait. J Biomech. 2013 Sep 3;46(13):2173-8.

Goldberg, S.R., Ounpuu, S., and Delp, S.L. 2003. The importance of swing-phase initial conditions in stiff-knee gait. J Biomech 36:1111-1116.

Goldberg, S.R., Anderson, F.C., Pandy, M.G., and Delp, S.L. 2004. Muscles that influence knee flexion velocity in double support: implications for stiff-knee gait. J Biomech 37:1189-1196.

Hicks, J.L., Delp, S.L., and Schwartz, M.H. 2011. Can biomechanical variables predict improvement in crouch gait? Gait Posture 34:197-201.

James, G., Hastie, T., and Sugar, C. 2000. A Principal Component Models for Sparse Functional Data. Biometrika 87: 587-602.

James, G., Hastie, T. 2001. Functional Linear Discriminant Analysis for Irregularly Sampled Curves. J Royal Statistical Society, Series B 63: 533-550.

Ku JP, Hicks JL, Hastie T, Leskovec J, Ré C, Delp SL. The mobilize center: an NIH big data to knowledge center to advance human movement research and improve mobility. J Am Med Inform Assoc. 2015 Nov;22(6):1120-5.

Lasko, Thomas A., Joshua C. Denny, and Mia A. Levy. "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data." PloS one 8.6 (2013): e66341.

Mansi T, Voigt I, Georgescu B, Zheng X, Mengue EA, Hackl M, Ionasec RI, Noack T, Seeburger J, Comaniciu D. An integrated framework for finite-element modeling of mitral valve biomechanics from medical images: application to MitralClip intervention planning. Med Image Anal. 2012 Oct;16(7):1330-46.

Phinyomark A, Osis S, Hettinga BA, Ferber R. Kinematic gait patterns in healthy runners: A hierarchical cluster analysis. J Biomech. 2015 Nov 5;48(14):3897-904.

Ramsay, J. O. 2006. Functional Data Analysis. Encyclopedia of Statistical Sciences.

Razavian, Narges, and David Sontag. "Temporal Convolutional Neural Networks for Diagnosis from Lab Tests." arXiv preprint arXiv:1511.07938 (2015). http://arxiv.org/pdf/1511.07938v4.pdf

Reinbolt, J.A., Fox, M.D., Arnold, A.S., Ounpuu, S., and Delp, S.L. 2008. Importance of preswing rectus femoris activity in stiff-knee gait. J Biomech.

Reinbolt, J.A., Fox, M.D., Schwartz, M.H., and Delp, S.L. 2009. Predicting outcomes of rectus femoris transfer surgery. Gait Posture 30:100-105.

Sen W., Zhang, C., De Sa, C., Shin, J., Wang, F., and Ré, C. 2015. Incremental Knowledge Base Construction Using DeepDive. VLDB.

Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. 1974. J Royal Statistical Society. Series B (Methodological) 36 (2): 111-147.