A mine of information: can sports analytics provide wisdom from your data?

Journal:	*International Journal of Sports Physiology and Performance*
Manuscript ID	IJSPP.2016-0644.R1
Manuscript Type:	Invited Brief Review
Keywords:	training, physical activity, physical performance, exercise training

Title: A mine of information: can sports analytics provide wisdom from your data?

Louis Passfield and James G. Hopker
Endurance Research Group, School of Sport and Exercise Sciences
University of Kent. Chatham Maritime. UK. ME4 4AG

Corresponding Author
Louis Passfield
Endurance Research Group, School of Sport and Exercise Sciences
University of Kent. Chatham Maritime. UK. ME4 4AG

Email: l.passfield@kent.ac.uk
Twitter: trainalytics

Word Count: 4177~~4299~~

Abstract: 222

Figures: 4 Tables: 0

Abstract

This paper explores the notion that the availability and analysis of large datasets has the capacity to improve practice and change the nature of science in the sport and exercise setting. The increasing use of data and information technology in sport is giving rise to this change. Websites hold large data repositories and the development of wearable technology, mobile phone applications and related instruments for monitoring physical activity, training and competition, provide large data sets of extensive and detailed measurements. Innovative approaches conceived to exploit more fully these large datasets could provide a basis for more objective evaluation of coaching strategies and new approaches to how science is conducted. The emergence of a new discipline, sports analytics, could help overcome some of the challenges involved in obtaining knowledge and wisdom from these large datasets. Examples of where large datasets have been analyzed, to evaluate the career development of elite cyclists, and to characterize and optimize the training load of well-trained runners are discussed. Careful verification of large datasets is time consuming and imperative before useful conclusions can be drawn. Consequently, it is recommended that prospective studies are preferred to retrospective analyses of data. It is concluded that rigorous analysis of large datasets could enhance our knowledge in the sport and exercise sciences, inform competitive strategies, and allow innovative new research and findings.

In recent years there has been an explosion in the use of information technology within the sport and exercise fields. The data and ~~thus~~ information derived from these advances has long been recognized to have the potential for a profound impact[1]. Websites ~~now~~ accumulate large repositories of primary and secondary data that previously would have been impossible for sport and exercise scientists to access and collate by hand. The ~~instrumentation of equipment and~~ invention of wearable technology enables extensive measurements to be gathered during exercise, training and competition. Increasingly, athletes and coaches recognize that such detailed, high quality data can be used to inform objective decision making on aspects of training and performance. In this paper we discuss how rigorous analysis of large datasets may hold the potential to change ~~not only~~ sport, ~~but~~ and the nature of its related sciences too.

"Moneyball"[2], and "Big Data" style stories in high performance sport readily capture the public interest, but ~~there remains a question as to whether~~it's not clear that scientists are making the most of their available data. There is a risk that the unprecedented ~~capacity for obtaining~~volume of data is overwhelming and ~~prevents us from~~not used fully ~~ing it~~ to ~~obtain insight and~~ inform practice. Consequently, ~~it seems appropriate to ask if~~we suggest there is scope to advance by following other disciplines (such as business and economics), in developing methods to analyze more rigorously the extensive data sources available to us.~~?~~ Rowley[3], ~~suggests~~ proposes ~~that~~ a wisdom hierarchy of data processing ~~exists~~. This hierarchy ~~sees~~ describes how~~a mass of~~ raw data is converted into information, ~~the~~ information into knowledge, and ~~the~~ knowledge into wisdom. Gaining ~~this~~ knowledge and wisdom from data is challenging, but could spawn a new discipline in the sports sciences, that of sports analytics.

Thornton et al.[34] note that ~~the ubiquity of~~ mobile phones and wearable technology present simple methods to assess and promote physical activity but this area is still underdeveloped. Excellence in the nascent field of sports analytics ~~promises~~ will need to ~~help~~ sieve ~~the deluge of~~ data from repositories and ~~these~~ devices in order to filter out meaningful information. The benefits of this work could be wide-ranging for the coach and scientist, such as identifying new talent, optimizing training programs, informing team selection, and deriving and evaluating competition tactics. The success of sports analytics will be governed by whether its findings can be translated clearly and for the benefit of ~~its~~ users, such as exercisers, athletes, and ~~their~~ coaches. A further challenge for sports analytics is that ~~in order to conduct~~ effective data analysis requires~~,~~ a fusion of diverse expert knowledge ~~has to occur~~; for example, in training theory, sports psychology, data handling and analysis, statistics and mathematical modelling, determinants of performance, and competition strategies. ~~At the moment t~~This presents a genuine interdisciplinary challenge as few, if any, individuals are sufficiently well versed in such disparate areas. Thus for sports analytics to fully mature as a discipline, new opportunities for the development of its practitioners are needed ~~to be conceived~~. This will likely require universities to develop new courses that ~~enable students to~~ combine ~~and acquire~~ a deep understanding of the science of sport, alongside extensive skills for data handling and analysis.

~~In this paper~~Next we provide two examples of the kind of opportunities that can be found in tackling this challenge, and discuss some consequent issues. We present two preliminary studies from our endurance research group that illustrate different ways of mining and modelling data to look at talent development and optimization of training. Our aim is to promote wider ~~recognition and~~ discussion ~~of the evolving discipline~~ of sports analytics and its potential to influence research and practice in the sport and exercise sciences.

Obtaining large datasets for analysis

Once a research question has been established, one way of addressing it can be to evaluate existing data ~~that has already been gathered~~. Data mining is a method where raw data is translated into information by analyzing and interpreting its patterns ~~within the data set~~. Data mining may also involve mathematical or statistical modelling, particularly where some kind of predictive capacity is required. The ~~Information~~ information ~~might be~~ obtained ~~from data~~can be used to help coaches predict changes in sports performance[5], ~~find events that co-occur or their sequence of occurrence, and divide data into similar groups[6]. Data mining techniques have been used to obtain information by examining~~ examine the relationship between performance and its determinants~~attributes~~[6], and to interrogate athletes' existing performance ~~related~~ data to identify new strategies.[7,8,9] Ofoghi et al.[~~7~~-8] ~~show how~~used data mining ~~could be used~~ to inform strategic planning for rider selection and training prioritization in the multi-discipline ~~events such as the~~ omnium in track cycling, and Moffatt et al.[9] for identifying sprint race tactics. There is a cost though, as ~~in many instances~~ the amount or complexity of the data, and preparing it for analysis can challenge even the most determined, especially where each athlete, team, game or event, across a season is modelled. It is also very important that the research question and methods are established before analysis is begun[10]. The evaluation of an hypothesis formed *a priori* helps to reduce the chance of bias and false positives arising from the analytic process. Otherwise, data fishing or P-hacking in large datasets ~~is likely to~~may result in ~~many~~ spurious but statistically significant results.

Analyzing race results

Some websites provide the potential to exploit large datasets by analyzing their ~~information they hold~~data. With the website's permission it is possible to use web-spider or web-crawler software to extract data ~~from its databases~~ for subsequent analysis. We examined the career progression and success of elite cyclists by using this approach to conduct a retrospective analysis of ~~their~~ race results.[10]~~[11]~~ ~~It is~~Coaches and scientists generally accepted that athletes ~~have to~~ undertake many years of training to achieve elite status in endurance sports. Yet the development profile of the most successful senior athletes and ~~the likelihood that~~whether this involves performing well in ~~elite~~ junior competition~~s~~ remains unclear.[11]~~[12]~~ To explore this issue we extracted race results for major junior and senior elite cycling races from 1980 to 2014 from one of the freely accessible online databases documenting race results (www.procyclingstats.com). For the purposes of the study we focused upon 25 major races and were able to obtain 67,503 results for 5,561 cyclists from 75 countries. This data included the name, date of birth, nationality, race, finishing position (including general classification and individual stage results from multi-stage races) of all the cyclists competing. From this data we were able to establish that the cyclists' average career length for competing in these most prestigious races was 3 seasons. However, as the data was heavily skewed by a few highly prolific cyclists, we also used the semi-interquartile range (SIQR) as an alternative way of depicting cyclists' typical career length. The SIQR comprises of the 50% of data between 25th and 75th percentile and it showed that half of all cyclists' careers ranged between 1 and 7 years. Notably, a large proportion of cyclists (86%), never achieve a top 10 placing in the major races we studied in their career. Our data mining also revealed ~~findings with implications for long-term development of cyclists, and team selection. As shown in Figure 1, we identify~~ evidence of a relative age effect[12]~~[13]~~, sometimes referred to as the Matthew effect, within the population of world-class cyclists ~~.~~

~~***** Figure 1 near here *****~~

There ~~appears to be~~is an over-representation of cyclists at the World Tour level who were born early in the calendar year (January-March). This ~~analysis raises the issue of whether~~observation suggests there is an inappropriate bias in how cyclists are ~~being~~ identified and developed by their coaches. To avoid this coaches should encourage a later specialization and ~~prematurely, or on an inappropriate basis e.g. more~~ focus upon technical skills, rather than physiological parameters ~~be better for~~in developing young cyclists. V~~Varying~~ arying the ~~youth cyclists'~~ age group cut-off dates within the competition year (e.g. ~~should~~ 9 or 15 months ~~be used~~ rather than 12 months) could ~~also~~ be considered. ~~Or alternatively, y~~Youth teams could also have quotas based upon chronological age within a year. ~~This~~ Only interrogation of a large volume of race data allowed us to describe th~~is~~e evolution of successful cyclists' and ~~substantiate the presence of~~identify the "Matthew effect" within elite cycling.

There are several challenges with establishing the validity and reliability of large datasets, especially where the analysis is retrospective ~~that need to be considered prior to conducting a study~~. For this reason, a prospective study design is often preferable ~~in order~~so that the integrity of the data can be overseen as it is gathered. ~~Trying to verify~~Establishing the veracity of large numbers of observations retrospectively is often impractical. For example, in our study above[10]~~[11]~~ the collection of retrospective race results from 3rd party websites using web-crawlers assumed these were accurately reported to reflect the "official" finishing positions. Moreover, collecting data in this way brought with it ethical considerations when deciding where, and how fast to crawl. Prior permission was always obtained from the data or website owner.

Nonetheless, fast crawlers can have a crippling impact on the performance of a website as the server deals with multiple simultaneous requests. ~~Once the web-crawler finished gathering data, p~~Pre-processing of the data was imperative to check for errors in its structure, and for subsequent filtering and cleaning. Within the cycling results database ~~for example, some~~ race names ~~had~~ changed over the years, or were listed in both native and English languages across various~~ editions e.g. Tour de Pologne/Ronde van Polen/Polen-Rundfahrt/Tour of Poland~~. ~~In some instances, there w~~Where results were missing ~~results~~ we ~~that~~ needed to verif~~ication of~~y whether the race took place~~, or if its results were just absent from the database~~. Similarly, ~~where~~ misspelt cyclists' names were ~~misspelt they needed to be~~ corrected prior to analysis, ~~otherwise their~~to ensure their results w~~ould have been in~~ere correctly assigned. In ~~short~~summary, ~~the opportunity to~~ analyz~~ing~~e large data sets can provide ~~as~~ a means of answering ~~to pre-~~specified research questions ~~provides the chance to extract~~with novel findings. It does require substantial meticulous and time-consuming work though, and the approach should not be regarded as a surrogate for prospectively conducted studies. Furthermore, conducting prospectively designed studies will help reduce the chance of bias and false positives[10] as mentioned previously.

Analysis of exercise and training data

When athletes and coaches monitor~~ing~~ exercise, training and racing, large datasets are now generated routinely. Advances in training technology have resulted in portable devices (such as accelerometers and similar activity monitors, GPS, heart rate monitors, power output meters, and related mobile phone apps), being used habitually to gather data by a wide spectrum of users from recreational exercisers to elite athletes. These devices typically gather data on all the activity of their users with a level of accuracy and detail once unthinkable. Characteristically, this data has been used to describe~~ and recount~~ completed exercise or training bouts and races.[13~~14~~,14~~15~~,15~~16~~,16~~17~~] However, by exploiting these opportunities more fully, scientists could produce exciting and innovative ~~new~~ findings. With this technology, performance can ~~now~~ be evaluated directly in the field, rather than be inferred from laboratory trials and simulations. Accurate measurements that previously required specialised laboratory equipment ~~are~~ can be ~~now~~ gathered by the coach during normal training and competition (Figure 1~~2~~). Furthermore, patterns of daily activity and inactivity can be described to evaluate lifestyle interventions more objectively[17~~18~~]. As a consequence, more realistic and ecologically valid experimentation can be designed and questions addressed that were previously beyond the reach of the laboratory-based scientist. An ~~enticing~~ example of ~~this is~~ these insights ~~that could come from being able to~~is in ~~accurately quantify~~prescribing training.

***** Figure 1~~2~~ near here *****

To date the process of prescribing training has relied upon the experience and intuition of those involved (i.e. coaches and athletes), as the necessary research in this area is lacking[18~~19~~]. Over the past four decades, the scientific basis for prescribing training programs has advanced little beyond Banister and colleagues' seminal work[19~~20~~,21]. This is in marked contrast to the ~~tremendous~~ advances ~~that have been~~ made in our understanding of the adaptions that result from training[22]. However, this ~~situation~~ could change with the capability to measure individuals' training and racing accurately and in detail in the field. The resulting large volumes of field measurements ~~could present~~allows ~~the discipline of~~ sports analytics ~~with an early opportunity~~ to contribute to our understanding of effective training program prescription[23]. Furthermore, ~~this~~ detailed monitoring of training and performance in the field provides an opportunity to reverse the usual scientific paradigm for research on this topic. Specifically, instead of conducting experiments to compare the effects of specific (laboratory-based) training regimens, we can measure study participants' training, and track their resulting changes in performance. It may then be possible to determine which aspects of their monitored training is most effective, given sufficient data. With this scientific paradigm the method of enquiry consists of identifying which training led to the observed changes in performance, rather than trying to evaluate how performance changes in response to a ~~carefully~~ restricted laboratory-based training protocol. Here the bigger the data, the better the insight, as effective training ~~is likely to~~may be identified more clearly when the number of participants involved and the diversity of their training is greater. Exploring a wide range of training regimes with large numbers of participants is not a viable option for laboratory-based research, but in a field study it becomes quite plausible. Participants can be recruited to undertake their usual training program and compete in their preferred competitions, no longer restricted to ~~following~~ scientists' abstract training regimes ~~and~~ or evaluation~~g them~~ with contrived laboratory-based performance trials.

~~Studies involving our endurance research group have demonstrated the potential for extracting useful insights from carefully conducted field studies.~~ Galbraith et al.[24] used GPS devices to record all the training and performances of 14 highly-trained endurance runners for a year-long study. This study resulted in measurements for 2.5 million time-points. In ~~our~~ the original analysis we summarized and collapsed this data into 3 training zones, finding total distance, and percent time spent at the highest intensity related to performance. This kind of analysis is difficult to translate into future training prescription for athletes however. Therefore, in order to analyze this dataset more fully Kosmidis and Passfield[25] proposed the use of training distribution and training concentration profiles (Figures ~~3~~2 and 3 respectively). This training distribution profile is obtained by plotting the amount of time spent above the reference speed during the session. For example, at 0 km·h^{-1} all the training was completed above this speed and therefore the total number of observations for the session is plotted. In contrast, at 15 km·h^{-1} only a small fraction of the total observations is seen to occur above this speed. In effect the analysis assumes every possible speed is a training threshold and shows how the pattern of training time changes with speed. The training concentration profile is the derivative of the distribution curve or in statistical terms a concentration curve. It shows the cumulative time spent training at each speed during the session(s) analyzed. By comparing the training distribution profiles with resulting changes in performance, the~~se~~ researchers were able to identify the runners' training speeds that were significantly related to improvement. ~~Not only could they identify these significant speeds for training, but t~~T~~hey could also~~ his information was used to model how endurance performance would change in response to training. Notably, ~~the authors observed that~~ the significant training speeds could not be determined from laboratory test data, but only from the analysis of the runners' training and performances. These methods and findings indicate that ~~in the future~~ it may be possible to support the coach by identifying ~~the~~ optimal training sessions for athletes ~~to complete for specific race performances~~. Perhaps even more importantly, ~~people~~ those promoting ~~exercising~~ exercise for health could ~~specify their available training time, and~~ use the same

Formatted: Superscript

method to calculate the most efficient exercise regime that provides ~~the~~ maximum benefits.

***** Figures 2 and 3 near here *****

There were some theoretical issues that the training analysis highlighted. Kosmidis and Passfield[25] set out with the ambition to retain all of the available data, to minimize the number of assumptions they made, and still utilize a parsimonious model with as few predictor variables as possible. When a data set is summarized, ~~whether~~ such as with a mean and standard deviation ~~or something more complex~~, much of the information in the original dataset is compressed ~~in the process~~ too. An advantage of the training distribution and concentration profiles is that they retain all the available data from every session for analysis. Furthermore, ~~relatively assumption-less approach to modelling their data meant~~ the authors did not rely on existing models of physiology to make sense of the data. ~~Rather they made the data "talk" and checked subsequently to see if their analysis supported traditional physiological models of training.~~ As mentioned above, their findings did not support existing models used for training, as their traditional laboratory tests results could not be used to identify the training speeds that were related significantly to the changes in performance. If the training data had been described with reference to the laboratory test data (i.e. as percentages of maximum or lactate threshold) ~~at the outset~~, the analysis would not have succeeded. Finally, as with most modeling work, a key challenge is ensuring parsimony to keep the model as simple as is reasonable. The training distribution and concentration curves help this process by reducing the complexity of the underlying dataset whilst still retaining a simpler, yet comprehensive representation of it.

~~There are many challenges to be overcome before it will be possible to introduce a rigorous scientific method into the process of prescribing training. Nonetheless some important lessons were learned from the studies above.~~ Data cleaning and checking was an arduous process, as with the study of cyclists' development profiles discussed earlier[11]. Every training session was plotted and manually inspected for obvious errors. This process ~~quickly~~ highlighted ~~that~~ the ~~subsequent analysis would have~~ need to deal with unrealistic "spikes" in the recorded values, and calculations where the training speed was at, or close to, zero. In addition to ~~clear~~ visib~~ua~~le data spikes, we also had to identify unreasonable values e.g. where the apparent speed was ~~clearly~~ above world record pace for the observed distance. These observations were due to problems with the GPS signal, or runners forgetting to switch off their GPS ~~when cycling or driving home~~ after a training session or race. Most of these issues could be addressed within the analysis, but a particular ~~challenging~~ issue was how to proceed in the absence of data. ~~All the runners were asked to submit their training programs, as these were not specified by the research team.~~ By matching the observed training data to the runners' training program ~~provided,~~ gaps caused by missing training data were identified. ~~The athletes' training record could also be used to determine whe~~Other missing observations in a training session implied a rest period, a gap between successive sessions, or a runner moving ~~very~~ slowly~~, or simply missing data~~. However, as this was a retrospective analysis ~~of the data~~ of data from an earlier study[24], it was not always possible to confirm these assumptions ~~were not always possible to verify~~. ~~As discussed earlier in this paper v~~Verifying the dataset wa~~i~~s a time-consuming but critical part of the analysis. This ~~re-emphasiz~~underlines our earlier recommendation that scientists prefer conducting prospective studies, ~~as opposed~~ to retrospective analyses of large ~~training~~ data sets when~~ever~~ possible.

Summary

Technological advances in recent years have enabled large datasets to be gathered in sport and exercise settings. Examples of these large datasets are information held by websites, and data generated by people monitoring their regular exercise, training or competitions. Careful analysis of these large datasets can enhance our knowledge in the sport and exercise sciences, support the coach by informing competitive strategies, and allow innovative new research and findings. The interest in making more from the data in sport and exercise sciences appears to be spawning a new discipline of sports analytics. This discipline necessitates the fusion of a diverse range of knowledge in computing, mathematics, statistics and sports sciences, that may require new development opportunities before the discipline can develop fully. Examples of preliminary work exploring large datasets from websites and GPS devices have been discussed along with some of the issues that this work presents. A common theme for this kind of work is that careful quality checking of the large dataset is imperative and time-consuming. Identification of missing data and strategies for dealing with it is also critical. Accordingly, it is recommended that prospective studies are preferred to retrospective analyses of data.

References

1. Liebermann DG, Katz L, Hughes MD, Bartlett RM, McClements J, and Franks IM. Advances in the application of information technology to sport performance. *J Sports Sci*. 2002;20(10):755-769. PubMed doi: 10.1080/026404102320675611
2. Lewis M, *Moneyball: The Art of Winning an Unfair Game.* New York. Norton & Company; 2004.
3. Rowley J, The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*. 2007; 33:163–180. doi: 10.1177/0165551506070706
4. Thornton JS, Frémont P, Khan K, Poirier P, Fowles J, Wells GD, Frankovich RJ. Physical activity prescription: a critical opportunity to address a modifiable risk factor for the prevention and management of chronic disease: a position statement by the Canadian Academy of Sport and Exercise Medicine. *Br J Sports Med*. 2016;50(18):1109-14. PubMed doi: 10.1136/bjsports-2016-096291.
5. Cangley P, Passfield L, Carter H, and Bailey M. A model for performance enhancement in competitive cycling. *Movement & Sport Sciences*. 2012;1:59-71.

6. Ofoghi B, Zeleznikow J, MacMahon C, Raab M. Data mining in elite sports: A review and a framework. *Meas Phys Educ Exerc Sci*. 2013;17:171-186. doi: 10.1080/1091367X.2013.805137

7. Chen I, Homma H, Jin C, Yan HH. Identification of elite swimmers' race patterns using cluster analysis. *Int J Sports Sci Coach*. 2007;2, 293–303. doi: 10.1260/174795407782233083

8. Ofoghi B, Zeleznikow J, MacMahon C, Dwyer D. A machine learning approach to predicting winning patterns in track cycling omnium. In M. Bramer (Ed.) *Proceedings of the International Federation for Information Processing (IFIP).* Conference on Advances in Information and Communication Technology 2010; pp. 67–76. Brisbane, Australia: Springer Berlin Heidelberg.

9. Moffatt J, Scarf P, Passfield L, McHale IG. Zhang K. To lead or not to lead: analysis of the sprint in track cycling. *J Quant Anal Sports*. 2014;10(2): 161-172. doi: 10.1515/jqas-2013-0112

10. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of P-hacking in science, *PLoS Biol.* 2015; 13(3):1-15. PubMed doi:10.1371/journal.pbio.1002106

11. Hopker J, Dietz KC, Schumacher YO, Passfield L, Using retrospective analysis of race results to determine success in elite cycling. *Journal of Science and Cycling*. 2015;4:29.

12. Schumacher YO, Mroz R, Mueller P, Schmid A, Ruecker G. Success in elite cycling: A prospective and retrospective analysis of race results. *J Sports Sci.* 2006;24(11):1149-56. PubMed doi. 10.1080/02640410500457299

13. Armstrong N, Welsman J. Physiology of the child athlete, *Lancet*, 2005;366:s44-45.

14. Esteve-Lanao JO, San Juan AF, Earnest CP, Foster CA, Lucia AL. How do endurance runners actually train? Relationship with competition performance.

Med Sci Sports Exerc. 2005; 37(3):496-504.

15. Esteve-Lanao J, Foster C, Seiler S, Lucia A. Impact of training intensity distribution on performance in endurance athletes. *J Strength Cond Res*. 2007; 21(3):943-949. PubMed doi: 10.1519/R-19725.1
16. Muñoz I, Seiler S, Bautista J, España J, Larumbe E, Esteve-Lanao J. Does polarized training improve performance in recreational runners. *Int. J. Sports Physiol. Perform*. 2014;9(2):265-272. PubMed doi: 10.1123/ijspp.2012-0350
17. Neal CM, Hunter AM, Brennan L, O'Sullivan A, Hamilton DL, DeVito G, Galloway SD. Six weeks of a polarized training-intensity distribution leads to greater physiological and performance adaptations than a threshold model in trained cyclists. *J Appl Physiol.* 2013;114(4):461-471. PubMed doi: 10.1152/japplphysiol.00652.2012
18. Strath SJ, Kaminsky LA, Ainsworth BE, Ekelund U, Freedson PS, Gary RA, Richardson CR, Smith DT, Swartz AM. *Circulation*. 2013;128(20):2259-79. PubMed doi:10.1161/01.cir.0000435708.67487
19. Borresen J, Lambert MI. The quantification of training load, the training response and the effect on performance, *Sports Med.* 39(9):779-795. PubMed doi: 10.2165/11317780-000000000-00000
20. Banister EW, Calvert TW, Savage MV, Bach TM. A system model of training for athletic performance. *Australian Journal of Sports Medicine*, 1975;7: 57–61.
21. Calvert TW, Banister EW, Savage MV, Bach T. A systems model of the effects of training on physical performance. *IEEE Trans. Syst. Man Cybern.* 1976;6:94-102.

22. Mann T, Lamberts RP, Lambert MI. Methods of prescribing relative exercise intensity: Physiological and practical considerations. *Sports Med.* 2013;43(7): 613-625. PubMed doi: 10.1007/s40279-013-0045-x
23. Jobson SA, Passfield L, Atkinson G, Barton G, Scarf P. The analysis and utilization of cycling training data. *Sports Med.* 2009;39(10):833-844. PubMed doi: 10.2165/11317840-000000000-00000
24. Galbraith A, Hopker JG, Cardinale M, Cunniffe B, Passfield L. A 1-year study of endurance runners: Training, laboratory tests, and field tests. *Int J Sport Phys Perf*. 2014; 9:1019-25. PubMed doi: 10.1123/ijspp.2013-0508
25. Kosmidis I, Passfield L. Linking the performance of endurance runners to training and physiological effects via multi-resolution elastic net. *arXiv preprint 2015;arXiv*:1506.01388.

Figure Legends

Figure 1: ~~The percentage of riders placing in the top 10 of World Tour cycling races by birth month. Data is percentage normalized for month length. The horizontal line at 8.33 represents the uniform distribution over the 12-month period.~~[~~11~~]

Formatted: List Paragraph

~~Figure 2:~~ A training session for an endurance runner, showing running speed over time. Data were gathered by wrist-worn GPS recording every second for each variable measured.

Figure 2~~3~~: A training distribution profile for the training session shown in Figure 1~~2~~ as proposed by Kosmidis and Passfield[2425~~25~~] for analyzing large training datasets. The distribution profile shows the total session time spent training above the corresponding speed.

Figure 3~~4~~: A training concentration profile for the training session shown in Figure 1~~2~~ as proposed by Kosmidis and Passfield[2425~~25~~]. The concentration profile shows the session time spent training at the corresponding speed.

Figures

~~Figure 1~~

Formatted: Font: (Default) Times New Roman

Formatted: Normal

Formatted: Level 1

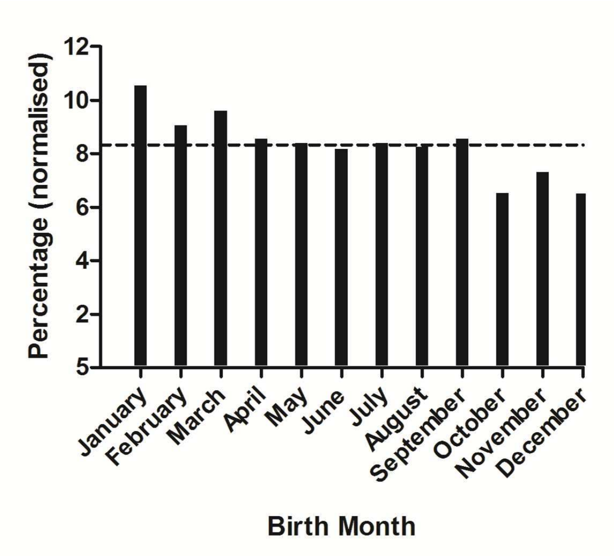

Figure 1~~2~~

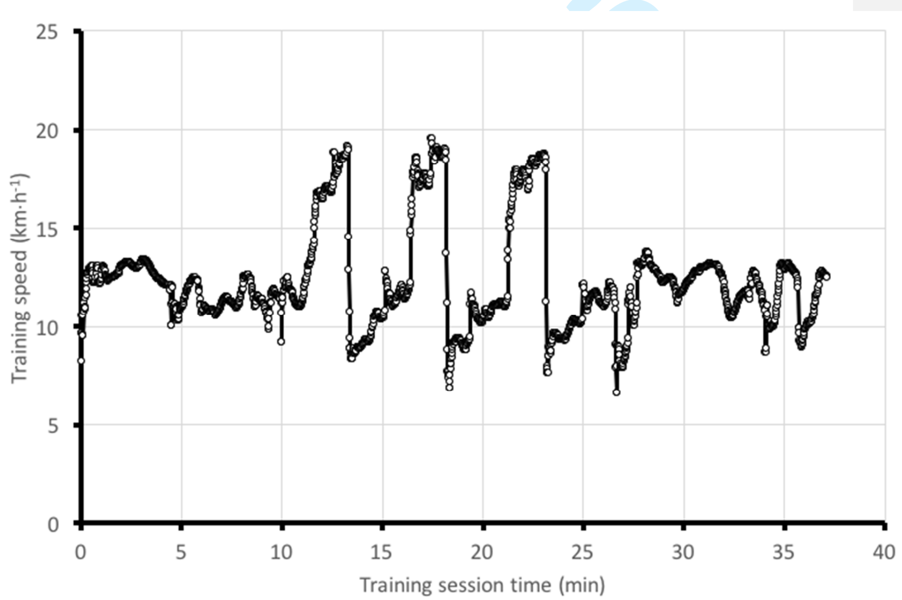

Figure 2~~3~~

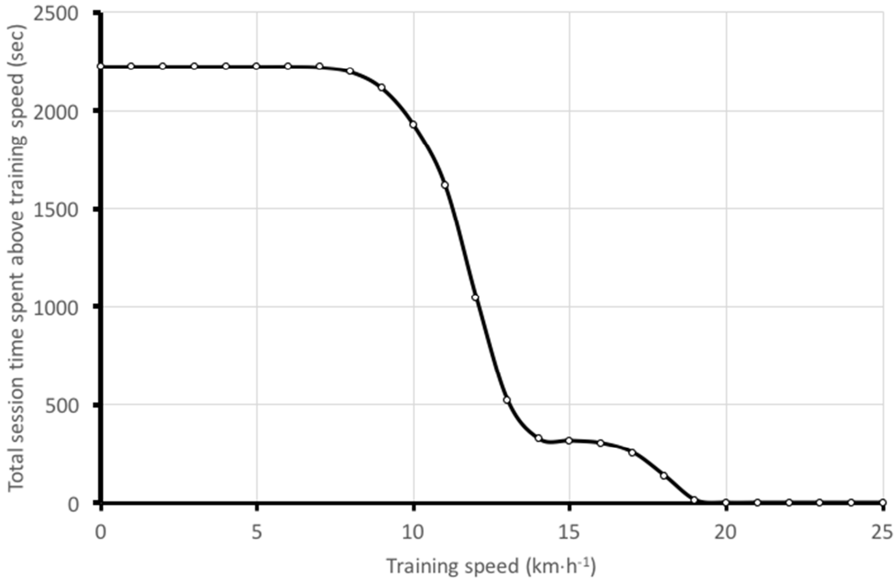

Figure 3~~4~~