BARÇA INNOVATION HUB

SPORTS TOMORROW

Predicting and Individualising Training Load using historical GPS data in Elite Soccer.

Dr. Kenny McMillan[3], Dr. Andrew Simpkin[1,2,4], Dr. Brian Moore[4], Prof. John Newell[1,2,4].

1. School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland.
2. Insight Centre for Data Analytics, National University of Ireland, Galway, Ireland.
3. Aspire Academy of Sports, Sports Department, Doha, Qatar.
4. Orreco.

There is a fine balance when optimising training load to maximize performance while minimizing injury. Accurate prediction of individual training loads for a planned training session is clearly beneficial. Annotated drills based on GPS data collected over several seasons on a professional soccer team were used to build an interactive Drill Planner. Linear and non-linear mixed models were used to create a connected statistical model that accounted for the number of drills, drill duration and rest between sets that provided reliable and robust estimates of expected training load metrics. Unsupervised learning techniques were used to combine data between similar drills. Non-parametric regression and smoothing methods were incorporated into the mixed model framework to model the correlation structure over time and to account for outcomes where the functional form was better modelled as a non-linear relationship (e.g. maximum velocity). Random effects were used to account for within player variability, player position, drill similarity and game schedules. Suitably weighted residuals were used to identify potential outliers; players that will find certain combinations difficult or easy) in order to aid in injury prevention and to prescribe individualised load adjustment. The Drill Planner achieved strong out of sample predictive performance where the model achieves correlations over 0.95 in out-of-sample testing, with median differences of below 1% of GPS outcomes. This paper demonstrates that a reliable Drill Planner can be built to create personalised training plans that are updated as more data become available in order to design fit for purpose individualised training sessions.

Maximise performance, training drills, mixed models, smoothing splines.

1. Introduction

One of the most popular and effective monitoring devices in elite sports is the Global Positioning System (GPS) (Larsson et al, 2003). They are widely used in professional soccer (Bush et al., 2015, Salvo et al., 2007, Barnes et al, 2014, Bradley et al., 2009), rugby and other sports (Dwyer et al., 2012, Varley et al., 2017, Wisbet et al., 2010, White 2015 et al., 2016, Hausler et al. 2016) to summarize a player's training load (i.e. metrics such as total distance covered, high intensity distance, number of accelerations) across the different drills in a session. Accurate 'prescription' of individual training loads in a training session is crucial in order to optimise training while minimizing the risk of injury and increasing player availability.

Soccer players participate in a wide range of training drills during a training session in order to induce adaptations needed to succeed in competitive match-play (Dwyer et al., 2012). A typical soccer training session consists of a combination of warm-up, technical/tactical drills, fitness related drills, and cool down activities. The choice and duration of each drill has to be planned in advance of training. The choice may depend on the proximity of the next upcoming game and the manager's preference. Each prescribed drill will induce a specific physiological effect and workload and these effects may differ greatly between drills and possibly interact with each other (certain combinations may elicit higher workloads than their sum).

Once a training session is completed the resulting GPS data are typically used to assess each component of the session separately by generating visualisations and tables summarising each metric at the squad and individual player level.

In this paper, an alternative and novel use of GPS training load data is presented where statistical models, using data from retrospective drills, have been built to predict future training workload in order to optimize training. The emphasis moves from 'past tense' to 'future tense'. Coaches can now use historical data to instantly predict all workload metrics of interest before a session. In this way, each day or week can be planned effectively. The estimated workload metrics can be updated quickly if there are any last minute changes to a session or changes occur during a session. Key metrics like the total distance and number of accelerations (Gabbett T.J, 2018) are automatically calculated allowing player and positon specific changes to be made. Players that have found certain drills atypically hard (or easy) can be identified and suitable modifications made. Training sessions can be saved for future use and 'ideal' sessions created. Additionally, a search can be made to find the most similar previous session to the one planned in order to audit outcomes (e.g. injury occurrence, rate of perceived exertion).

Our approach was to use linear and non-linear mixed models to create a set of connected statistical models that accounted for the number of drills, drill duration and rest between sets that provided reliable estimates of expected training load metrics and corresponding estimates of uncertainty. Unsupervised learning techniques such as k-nearest neighbours were used to combine data between similar drills to reduce uncertainty. Non-parametric regression and smoothing methods were incorporated into the mixed model framework (e.g. cubic splines, kernel smoothers) to model the correlation structure over time and to account for outcomes where the functional form was better modelled as a non-linear relationship (e.g. maximum velocity). Random effects were used to account for within player variability, player position, drill similarity (e.g. 7v7 and 8v8 small-sided games) and game schedules. Suitably weighted residuals were used to identify potential outliers.

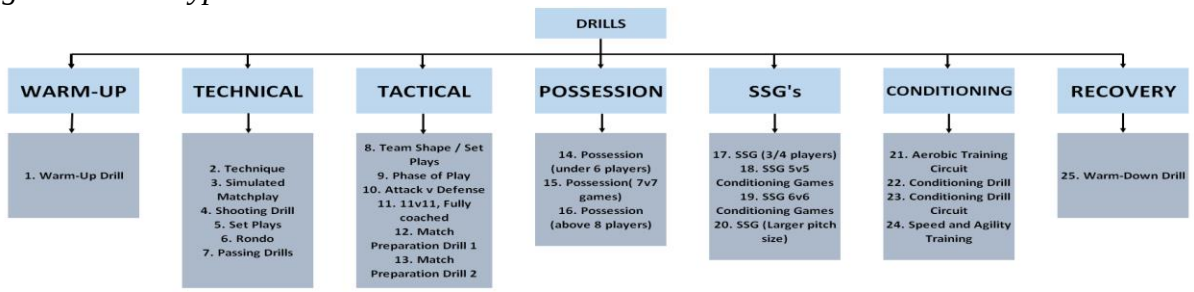
This research is important as it demonstrates that with correctly annotated GPS data, a reliable Drill Planner can be built in order to create personalized training plans using statistical learning algorithms that are updated as more data become available. It is applicable to all sports where annotated GPS data are collected (e.g. soccer, NFL, rugby, basketball). Such a planner allows coaches to better plan, prescribe and tailor training drills in advance by using previous data in an objective manner to design fit for purpose individualized training sessions.

2. Data

In the analysis presented here, training session data (n=4514 observations) collected from a European professional soccer team using an industry standard GPS device were used. Player names and characteristics are not provided and generic drill names have been used to protect anonymity.

Drills were classified by Type (e.g. warm-up, possession, tactical, possession, Small sided games, conditioning and recovery) with 25 different drills in total. Each drill will induce a specific physiological effect and workload, and these effects may greatly differ between drills.

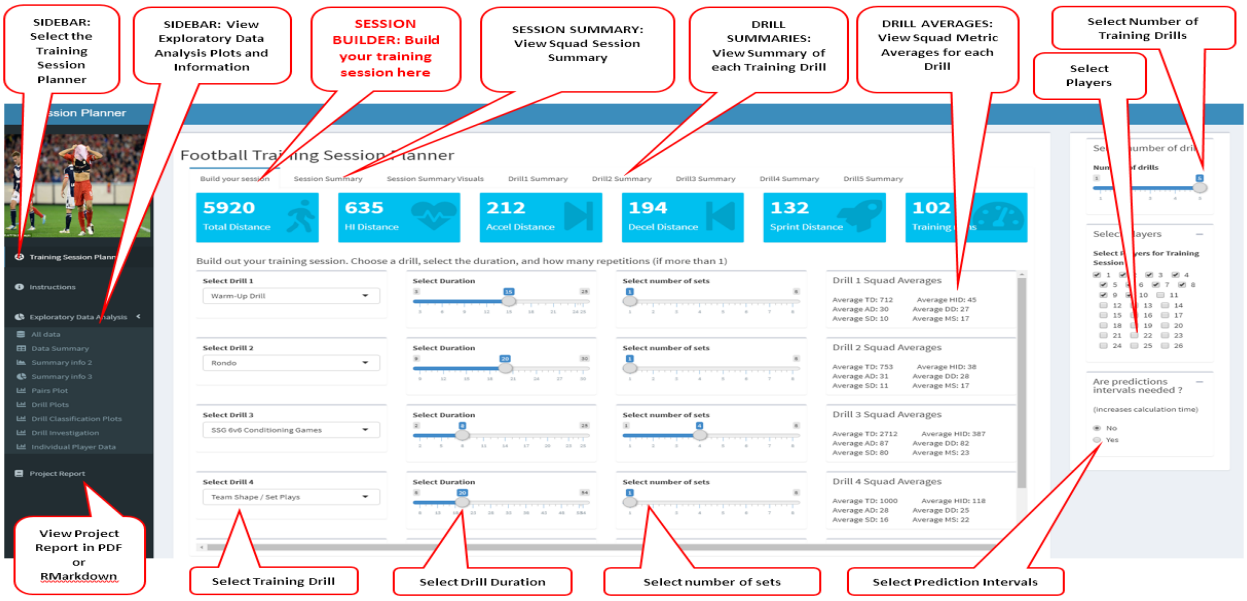
Figure 1. Drill Types.



3. Statistical Analysis

The drill data have a complex hierarchical structure (i.e. players are nested within drills across time) and the relationship between GPS data and drill type is not necessarily linear over time. A variety of statistical models were needed depending on the functional form required to model the relationship (i.e. linear, non-linear), the longitudinal component and the multivariate responses (e.g. distance covered, number of accelerations, high-intensity running). A brief technical description of the models used to build the Drill Planner (Figure 2) is now given.

Figure 2. Drill Planner screen-shot.



3.1. Linear Mixed Models

Linear mixed models [12] are an attractive framework to model the serial correlation in each player's GPS data over time by incorporating a random effect in order to model each player's serial correlation in a flexible way.

Let y_{ij} be a repeatedly measured GPS outcome variable (such as distance) where $i = 1, \dots, n$ index the n players in the squad and $j = 1, \dots, n_i$ index the n_i observations for player i, such that players can have different numbers of repeated measurements. The x_{ij} represent the drill duration in minutes for player i on occasion j. As observations within the same individuals are dependent, the between and within individual variance should be considered separately. This can be achieved through the linear mixed model as follows

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \epsilon_{ij}$$

where $\beta_{0i} = \beta_0 + b_{0i}$ and $\beta_{1i} = \beta_1 + b_{1i}$ are individual intercept and slope terms, combining the population intercept and slope parameters β_0 and β_1 with individual level residuals b_{0i} and b_{1i}. The residuals are not estimated directly while the variance and covariance of these random effects are.

In this study the variance and covariance components were all estimated using restricted maximum likelihood (REML). Details of estimation of random effects through REML are described in detail in Durbán et al. (2005) and Ruppert et al. (2003).

3.2. Polynomial Mixed Models

Where linear relationships are not present, polynomial regression can be extended to the mixed model framework by allowing for random effects b_{pi} for some or all polynomial terms x^p with $p = 0, \dots, P$ as

$$y_{ip} = \sum_{p=0}^{P} (\beta_p + b_{pi}) x^p + \epsilon_{ij}$$

where

$\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, $b_{pi} \sim N(0, \Sigma)$

and

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{b0}^2 & \sigma_{b01} & . & . & \sigma_{b0P} \\ \sigma_{b01} & \sigma_{b1}^2 & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \sigma_{b0P} & . & . & . & \sigma_{bP}^2 \end{pmatrix}.$$

3.3. Spline Mixed Models

The polynomial mixed model proposed above might not be flexible enough if the shapes of the individual curves are complex. Moreover, using a global polynomial basis means that local changes in the data will have global effects across the range of data collected. Local modelling approaches such as splines do not suffer from this issue. A spline framework is given in the method described by Durban et al. (2005) in which individual curves are modelled as the sum of a population curve and subject-specific departures from the mean, modelled as penalized splines with random coefficients:

$$y_{ij} = f(x_{ij}) + g_i(x_{ij}) + \epsilon_{ij}$$

Both the mean function (f) and the individual trajectories (g_i) are estimated using truncated polynomials (which break at predefined knots κ_k, $\kappa = 1, \dots, K$) and have their own set of random effect coefficients (u_k and v_{ki} respectively) that allow for these individual trajectories to be obtained. The u_k allow for a flexible mean trend estimate and are the same for each individual; the v_{ki} allow for this flexible mean estimate to vary between individuals

$$y_{ij} = \sum_{p=0}^{P} (\beta_p + b_{pi}) x_{ij}^p + \sum_{k=1}^{K} u_k (x_{ij} - \kappa_k)_+^P + \sum_{k=1}^{K^*} v_{ki} (x_{ij} - \kappa_{k*})_+^P + \epsilon_{ij}$$

where $a_+ = a$ if $a > 0$ or 0 otherwise, $b_{pi} \sim N(0, \Sigma)$, $u_k \sim N(0, \sigma_v^2)$, $v_{ki} \sim N(0, \sigma_v^2)$ with the variance covariance matrix

$$\mathbf{G} = \begin{pmatrix} \sigma_u^2 \boldsymbol{I} & 0 & 0 \\ 0 & \boldsymbol{\Sigma} & 0 \\ 0 & 0 & \sigma_v^2 \boldsymbol{I} \end{pmatrix}.$$

The use of spline mixed models can lead to an incorrect population mean estimate and incorrect pointwise standard errors, due to the effect of covariance structure of the individual effects on the population mean estimate. To address this a simple approach was used by selecting the knots for the individual curves (κ_{k^*}) from a subset of the knots of the population curves (κ_k). As discussed in Durban et al. (2005), *K* should be large enough to ensure the flexibility of the curve, with the knots chosen as quantiles of x_{ij} with probabilities *1/(K +1), … K/(K +1).*

3.4. Clustering Drills

As the number of drills *D* can be large, the number of parameters to estimate can become unwieldy. One solution is to combine similar drills into drill 'clusters' and use these to reduce the number of dummy variables in the mixed models. Clusters can be based on Type (e.g. warm-up, recovery, tactical), Drill (e.g. small-sided game 7 vs 7, small-sided game 8 vs 8) or by the GPS training load metric. In the latter case, we used k-nearest neighbours to combine drills with similar relationships between GPS outcome metric and drill duration.

3.5. Missing data

Given the amount of data collected as part of a season-long training monitoring program, missing data are an expected consequence (e.g. potential failure of wearable devices, lack of player availability) where 'gaps' will appear in a player's data and the occurrence of these gaps may differ from player to player. Such missing data may be uninformative (unlikely to add bias to subsequent inference) or informative (likely to add bias as missingness is masking an important effect).

Uninformative missing data may occur completely at random Little et al, (2002) where there are no systematic differences between the missing values and the observed values. For example, training load measurements may be missing for a particular player because of a GPS device malfunction. If the data are missing completely at random a complete case analysis will provide unbiased estimates of model parameters, albeit with an increase in uncertainty due to the smaller sample size (White et al., 2011).

If missing data occur where the differences between missing values and observed values can be explained by the observed data (i.e. covariates) such data are referred to as missing at random [15]. If the observed data cannot explain the presence of missing data such data are referred to as missing not at random e.g. a player not recording rate of perceived exertion after a training session due to an underlying condition that the sports scientist is not aware of and not captured in the data.

Given the level of monitoring that occurs for soccer players, an assumption that missing data are missing at random is plausible. When data are missing at random, multiple imputation (using chained equations) is a popular approach for imputing such missing data [17] as the distribution of observed data can be used to simulate several different plausible imputed data sets (i.e. each imputed dataset has a different estimate for the missing data). A separate analysis is then run on each imputed dataset and model averaging used (with suitable weighting) to appropriately combine the results from each imputed dataset [18]. If the data can be assumed to be missing at random and linear mixed models (using the observed data only) are used (as in this analysis) a multiple imputation step is unnecessary as mixed models account for such missing data as part of parameter estimation (Schafer et al., 2002).

3.6. Identifying 'at risk' players

Different players will react to training drills in different ways. A player can be considered 'at risk' of potential injury if the prescribed training is too intense, or 'at risk' of not being challenged physiologically if the prescribed training is not intensive enough. Given the large number of training drill types available and drill order combinations it may be a considerable challenge for a coach to identify potential 'at risk' players for a given set of prescribed drills.

For example, a 9 v 9 training drill on a full size pitch may induce a greater running load on a midfield player compared to a central defender. Conversely, a small-sided game (e.g. 4 v 4 on a 30m by 20m pitch size) might manifest a similar workload on each player. Training drills may be designed to overload a certain workload metric, for example, an intermittent running drill may be designed to accumulate time above a certain speed threshold. Tactical drills may also stress players

in a specific way, an example being a striker having to make repeated maximal accelerations when trying to score from a winger's cross into the penalty box.

The residuals from the statistical models used can be exploited to identify players that have predicted training workloads that are atypically high (or low) for the combination of drills prescribed (Figure 3). This information can be used to personalize training at the individual level in order to prescribe individualized load adjustment, crucial in a dynamic high performance environment where plans can change quickly.

Figure 3. Players identified as 'at risk' for the prescribed drills.

3.7. Model Performance and Predictive Accuracy

Model choice (i.e. fixed effects, random effects, variance structure and functional form) was based on Root Mean Squared Error (RMSE) and the Akaike Information Criterion (AIC) and assumptions checked using suitable residual plots.

Predictive performance was evaluated by comparing the predicted versus actual training load metrics in holdout test data. The available data were split into a training (70%) and validation set (30%). The performance of the final models was assessed by comparing predicted versus actual training load across commonly used training load metrics such as total distance, high load distance, accelerations, decelerations, explosive distance and dynamic stress load. Model performance was assessed using Pearson's correlation coefficient and quantification of the median difference between predicted and actual values for each GPS variable.

4. Results

Summary statistics are given in Table 1 on how well the model performed when predicting key GPS training load metrics using the test sets. The models underlying the Drill Planner achieved strong predictive performance where the model achieves correlations of 0.79 to 0.97 in out-of-sample testing. The typical differences between the actual and predicted training load is quite small for each metric. For example, the correlation between the actual and predicted total distance

was 0.97 and the corresponding plot (Figure 4) highlights the good agreement overall and within drill type.

Figure 4. Scatterplot of the observed and predicted total distance (in meters) colored by drill type.

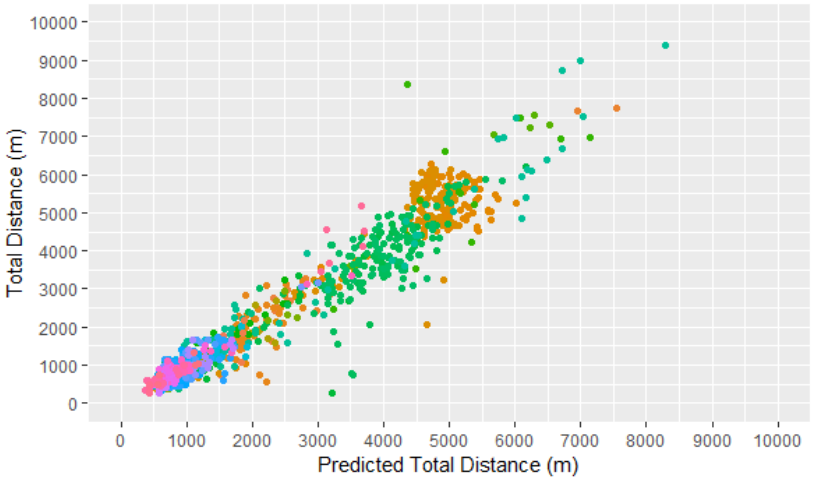


Table 1. Model Performance for a selection of key Workload Metrics.

Workload Metric	**Absolute Median difference between actual and predicted training load**	**Correlation between actual and predicted training load**
Total Distance covered (m)	5.10	0.97
Total number of Accelerations	1.8	0.88
Max speed attained (m.s-1)	0.21	0.82
Number of sprints	0.57	0.79

5. Discussion

There is a fine balance when optimizing training load to maximize performance while minimizing injury (Arnason et al., 2004). The impact of sports and exercise science on maximizing a player's potential while minimizing injury risk is well-established and a crucial component of successful elite sporting organizations. Data from player monitoring devices are one of the key resources on which decisions are made.

We have demonstrated in this research that a reliable Drill Planner can be built in order to create personalized training plans using statistical models that are updated as more data become available. It is applicable to all sports where annotated GPS data are collected (e.g. soccer, NFL, rugby, basketball) and is applicable to all makes of GPS devices. Such a planner allows coaches to better plan, prescribe and tailor training drills in advance by using previous data in an objective manner to design fit for purpose individualized training sessions. It is also a vital tool for coach education.

Once accurate predictions of training load are available, the Drill Planner can be used in reverse to recommend a set of drills that best match the training load outcomes and drills required e.g. find a set of drills that last 30 minutes that will result in total distance covered of 2500m with 120 accelerations and 30 impacts. Details of when a similar set of drills were used can be called up and an investigation of relevant outcomes in the acute phase undertaken.

Annotating each drill by the number of days before the game (e.g. MD -1), the training conditions (e.g. hard surface) and team schedule and including this information as additional explanatory variables is likely to further explain variability in workload for identical drills of similar durations. Bayesian regression models are another useful framework to consider when building a drill planner as priors can be elicited from domain knowledge not captured by the data.

Monitoring systems should be intuitive, provide efficient data analysis and interpretation, and enable efficient reporting of simple, yet scientifically valid, feedback (Halson, 2014). The Drill Planner proposed in this paper achieves these aims by providing a coach with a validated tool to design and implement training sessions in an efficient and safe manner by better use of GPS data already collected.

Such a tool is clearly beneficial for a player also as an individualized training session designed to make them physiologically prepared for competitive matches has the potential to improve their performance and reduce their risk of injury due to inappropriate training loads.

References

Arnason A, Sigurdsson SB, Gudmundsson A, Holme I, Engebretsen L, Bahr R. (2004). Physical fitness, injuries, and team performance in soccer. Med. Sci. Sports Exerci, 36, 278–285.

Barnes C, Archer D.T, Hogg B, Bush M, Bradley P.S. (2014). The evolution of physical and technical performance parameters in the English Premier League. Int. J. Sports. Med., 35, 1095–1100.

Bradley P.S, Sheldon W, Wooster B, Olsen P, Boanas P, Krustrup P. (2009). High-intensity running in English FA Premier League soccer matches. J. Sports Sci., 27, 159–168.

Bush M, Barnes C, Archer D.T, Hogg B, Bradley P.S. (2015). Evolution of match performance parameters for various playing positions in the English Premier League. Hum. Mov. Sci., 39, 1–11.

Durbán M, Harezlak J, Wand M, Carroll R. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine, 24*(8), 1153-1167.

Dwyer D.B, Gabbett T.J. (2012). Global Positioning System Data Analysis: Velocity Ranges and a New Definition of Sprinting for Field Sport Athletes. J. Strength Cond. Res. 26, 818–824.

Gabbett T.J. (2018). Debunking the myths about training load, injury and performance: empirical evidence, hot topics and recommendations for practitioners. Br J Sports Med. 2018. Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 963-974.

Halson S. (2014) Monitoring Training Load to Understand Fatigue in Athletes. Sports Med. 44 (Suppl 2):S139–S147.

Hausler J, Halaki M, Orr R. (2016). Application of Global Positioning System and Microsensor Technology in Competitive Rugby League Match-Play: A Systematic Review and Meta-analysis. Sports Med, 46, 559–588.

Ibrahim, Joseph G, Molenberghs G. (2009). "Missing data methods in longitudinal studies: a review" Test (Madrid, Spain) vol. 18,1: 1-43.

Larsson P. (2003). Global positioning system and sport-specific testing. Sports Med., 33, 1093–1101.

Little RJA, Rubin DB. (2002). Statistical analysis with missing data. New York: Wiley.

Rubin DB. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley: New York.

Ruppert D, Wand M. P, Carroll R. J. (2003). *Semiparametric regression* (Vol. 12): Cambridge University Press.

Schafer JL, Yucel RM (2002). Computational strategies for multivariate linear mixed-effects models with missing values. Journal of Computational and Graphical Statistics. 11:437-457.

Salvo V.D, Baron R, Tschan H, Montero F.J.C, Bachl N, Pigozzi F.(2007). Performance Characteristics according to Playing Position in Elite Soccer. Int. J. Sports. Med., 28, 222–227.

Van Buuren S, Boshuizen HC, Knook DL. (1999).Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med;18:681-94.

Varley M.C, Jaspers A, Helsen W.F, Malone JJ. (2017). Methodological Considerations When Quantifying High-Intensity Efforts in Team Sport Using Global Positioning System Technology. Int. J. Sports Physiol. Perform, 12, 1059–1068.

White I. R, Royston P, Wood A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. Statist. Med., 30: 377-399.

White A.D, Macfarlane N.G. (2015). Analysis of International Competition and Training in Men's Field Hockey by Global Positioning System and Inertial Sensor Technology. J. Strength Cond. Res, 29, 137–143.

Wisbey B, Montgomery P.G, Pyne D.B, Rattray B. (2010). Quantifying movement demands of AFL football using GPS tracking. J. Sci. Med. Sport, 13, 531–536.