**Missing data in sports science: Current practice in football research and recommendations for reporting missingness**

David N Borg4, Robert Nguyen3, Nicholas J Tierney1,2

1 Monash University, Department of Econometrics and Business Statistics, Melbourne, Australia

2 Australian Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), Melbourne, Australia

3 University of New South Wales, Department of Statistics, School of Mathematics and Statistics, Sydney, Australia

4 Griffith University, Menzies Health Institute Queensland, The Hopkins Centre, Brisbane, Australia.

**Acknowledgement/conflict of interest**

NJT naniar package?

**Corresponding author**

Nicholas J Tierney

Monash University, Department of Econometrics and Business Statistics, Melbourne, Australia

**Data availability**

**Abstract**

Missing data are often unavoidable. The reason values go missing, along with decisions made in how missing data are handled (deleted or imputed), can have a profound effect on the validity, usability and accuracy of a study’s results. In this article, we aimed to: estimate the proportion of studies in football research that included a missing data statement, highlight several improvable practices in relation to missing data, and provide recommendations for exploring, visualising and reporting missingness. Football related articles, published in 2019 were studied. A survey of 137 articles, sampled at random, was conducted to determine whether a missing data was included. As expected, the proportion of studies in football research including a missing data statement was low, at only X%; which suggests that this aspect of data analysis is seldom considered by researchers. We suspect that this result is consistent with the wider sports science literature. We recommend researchers describe the number and percentage of missing values, including when there are no missing values. Exploratory analysis should be conducted to explore missing values, and visualisations describing missingness overall should be provided in the paper or at least supplementary materials. Missing values should almost always be imputed, and imputation methods should be explored to ensure they are appropriately representative.

**Keywords:** Imputation, missingness, missings, naniar, sport, football

**Introduction**

Missing data can undermine the validity of research results [[1–3]](https://www.zotero.org/google-docs/?TjtzG7). Data can go missing many ways. For example, accidentally skipping a survey question, equipment failure, or intentionally not recording values. Before analysis, missing data must be handled, with values typically deleted or imputed. There is no universal approach to handling missing data with contextual factors--such as, the study design and objective, and pattern of missingness--determining how to handle missings on a case-by-case basis [[1]](https://www.zotero.org/google-docs/?JCe60i). The effect of missing data on the validity and accuracy of a study’s results is largely determined by the reasons for missingness, and how missing data are handled.

Data does not go missing the same way every time. There are three broad categorisations describing why data can be missing: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). These categorisations help describe whether the missingness occurs by chance (MCAR), due to some observed variable (MAR) or an unobserved variable (MNAR), with MCAR being benign and the latter two biased. Football relevant examples are provided in Supplement X, and more detailed explanations of MCAR, MAR, MNAR can be found elsewhere [1–3].

The three categorisations of missingness types provide a useful framework to describe bias arising from missing data. If you strongly suspect data is MCAR, you can be more certain your data isn’t biased, otherwise there may be bias in your sample. This does not mean data MAR or MNAR make data invalid, instead it identifies potential biases, which can help improve future research design. However, these categorisations do not necessarily guide the user in their subsequent actions in analysis, or describe specifically how values go missing. To identify possible mechanisms for missingness, the data must be explored. It can be challenging to identify the missingness mechanism in data, but it is critical, since missing values could change the outcome of a study. This could mean the work cannot be reproduced, and worse, may lead scientists and practitioners to the wrong outcome.

It is imperative that missing data (or lack thereof--Absence of evidence is not evidence of absence) are reported. However, outside of sports science, researchers seldom consider missing data [5](<https://sci2s.ugr.es/keel/pdf/specific/articulo/Schafer_Graham02.pdf>). Moreover, while there has been significant interest in modelling missing data (refs), the exploration and reporting of missing data have received less attention [5]. This article aimed to: (1) estimate the proportion of articles that report missing data in football research; (2) highlight several bad missing data practices; and (3) provide recommendations for exploring and reporting missingness.

**Methods and materials**

To estimate the proportion of articles that report missing data in football research, we conducted a systematic search [[5]](https://www.zotero.org/google-docs/?FEfuSJ). Football related articles published in 2019 were studied. Three major sports medicine databases (SPORTDiscus, Embase and Cinahl) were searched, using key terms and search limits (Supplement 1). Included articles were: (a) published in 2019; (b) written in English; (c) had an accessible full-text; and (d) included quantitative data. The search process is summarised in Figure 1.

require(PRISMAstatement)

prisma()

<https://cran.r-project.org/web/packages/PRISMAstatement/vignettes/PRISMA.html>

library(PRISMAstatement)

e.g.,

prisma(found = x,

found\_other = x,

no\_dupes = x,

screened = x,

screen\_exclusions = x,

full\_text = x,

full\_text\_exclusions = x,

qualitative = x,

quantitative = x,

width = 800, height = 800)

|  |  |
| --- | --- |
| Imported | 3,863 |
| After duplicates removed | 1,954 |
| After title and abstract screen | 1,373 |
| Random sample full-text | 199 (needed for 10% sample) |
| Full-text included | 137 |
| Full-text excluded | Full-text papers excluded (n=62) with reasons:  Qualitative study (n=)  Not in english (n=)  Unable to access full-text (n=) |

**Figure 1.** PRISMA flow chart.

Our intention was to survey, at random, 10% of the total articles found (after title and abstract screening). We were primarily interested in whether missing data were reported or acknowledged; and if not, whether a dataset was shared. Where applicable, information about how missing values were handled was also extracted (see Supplement 1). Two authors independently extracted the data (X% agreement). Findings are reported as the proportion and 95% confidence interval. Supplement 1 provides further details of the search.

**Results**

Of the 199 articles screened, 136 were deemed eligible, representing ~10% of the articles found (Figure X). XX articles (of the 199) were excluded due to not meeting the inclusion criteria. The proportion of articles that reported missing data was 5.9% [95% confidence interval; X% to X%] or 8/136. Of the articles that did not mention missing data (128/136), 7.1% [3.3% to 13.1%] or 9/128 shared their dataset, making it difficult to determine whether there was any missingness. While the absence of a missing data statement does not necessarily mean missing data are not being reported, it does suggest that this aspect of the analysis process receives little consideration--especially in contrast to many of the papers that included (potentially meaningless) tests of normality (e.g., Kruskall-Wallis test).

**Discussion**

Our survey of 136 articles published on football related topics found that only about one in 20 papers reported missing data. It is reasonable to assume that this result is representative of current practices in the wider sports science literature. This finding may be largely explained by a lack of awareness and education on missing data practices (Kristin BJSM paper), similar to other areas of sciences (refs-education and ecology examples)(2). Below we discuss several poor practices in relation to missing data, and provide recommendations for exploring, visualising and reporting missingness.

*Missing data across the analysis pathway*

Problem: *Unreported missing values*. While not always the case, even if not reported, missing data can be obvious. A recent paper examined the relationship between match performance indicators and outcome, in Australian Football between the 2001 and 2016. Problematically, at least one of the 91 team performance indicators included in the analysis was not available over the entire 2001-2016 period [[6]](https://www.zotero.org/google-docs/?pAyeWc). The variable ‘meters gained’ was not collected over the specified duration (ref). This was not noted anywhere in the paper, or in the supplementary materials. It is unknown whether imputation was undertaken. Without these details, it is difficult for other analysts and researchers to use and/or extend the ideas presented in the paper, or reproduce the analysis. It is also difficult for a reader to evaluate how the missingness could bias the results. Unfortunately, the absence of a missing data statement when data are missing is not unique to the study above (Young, C., Luo, W., Gastin, P., Tran, J. and Dwyer, D., 2019. Modelling match outcome in Australian Football: Improved accuracy with large databases. International Journal of Computer Science in Sport, 18(1), pp.80-92.). Authors should include a missing data statement, irrespective of whether data is missing, or not.

Problem: *Mean imputation.* The default of most statistical software for handling missing data in modelling is listwise deletion. Listwise deletion removes entire rows that contain missing observations from the analysis. At best, listwise deletion reduces the power in analysis by reducing the sample size. At worst, it introduces bias. For example, if injured participants are removed from a study on injury prevention. Imputing values removes the need to conduct listwise deletion. However, it is critical how these values are imputed, and that the imputation method is documented. One imputation method is imputing the mean value (e.g., participant or group average) [5–7]. While imputing the mean preserves the study sample size (and point estimates), it also reduces the variance, can alter the relationship between the variable with missing observations and other variables, and can bias (underestimate) standard errors (Scheffer 2002 - <https://mro.massey.ac.nz/handle/10179/4355>). Smaller standard errors typically reduce *p*-values, which may lead to incorrect inference. Mean imputation should be avoided. An imputation strategy that incorporates information from other related variables in the data--such as linear regression, or K nearest neighbours--should be considered. This is discussed in the recommendations section.

Problem: *Not evaluating the effect, or choice, of imputation.* The imputation method used to impute missing data has the potential to affect the outcome of a study. Therefore, it is important that authors understand, and document, how imputation affected the goal of the analysis (i.e., inference, prediction, or both). For example, understanding the implications in using mean imputation, compared to K-nearest neighbour imputation--the effect on parameter estimates (‘significant’ versus ‘not significant’), and the (un)certainty of the coefficients (i.e., less/more).When using imputation, authors need to perform a sensitivity analysis (ref circulation article-Nick). This is discussed below.

*Common causes of missing values*

Broadly, there are two types of missing values in data: implicit and explicit missings. Explicit values are missing, but recorded; whereas with implicit missinges, their presence is implied based on other information in the data. For example, in Table 1, player Koenen has missing values for quarters two and four. Sometimes values like these can be logically imputed, as it might be known that these values are recorded as 0 in this format, rather than NA. Other places missing data can arise include: bad joins when merging data without corresponding values, surveys, an inability to collect a biological sample (e.g., venous blood), equipment malfunction or failure, or or not being worn.

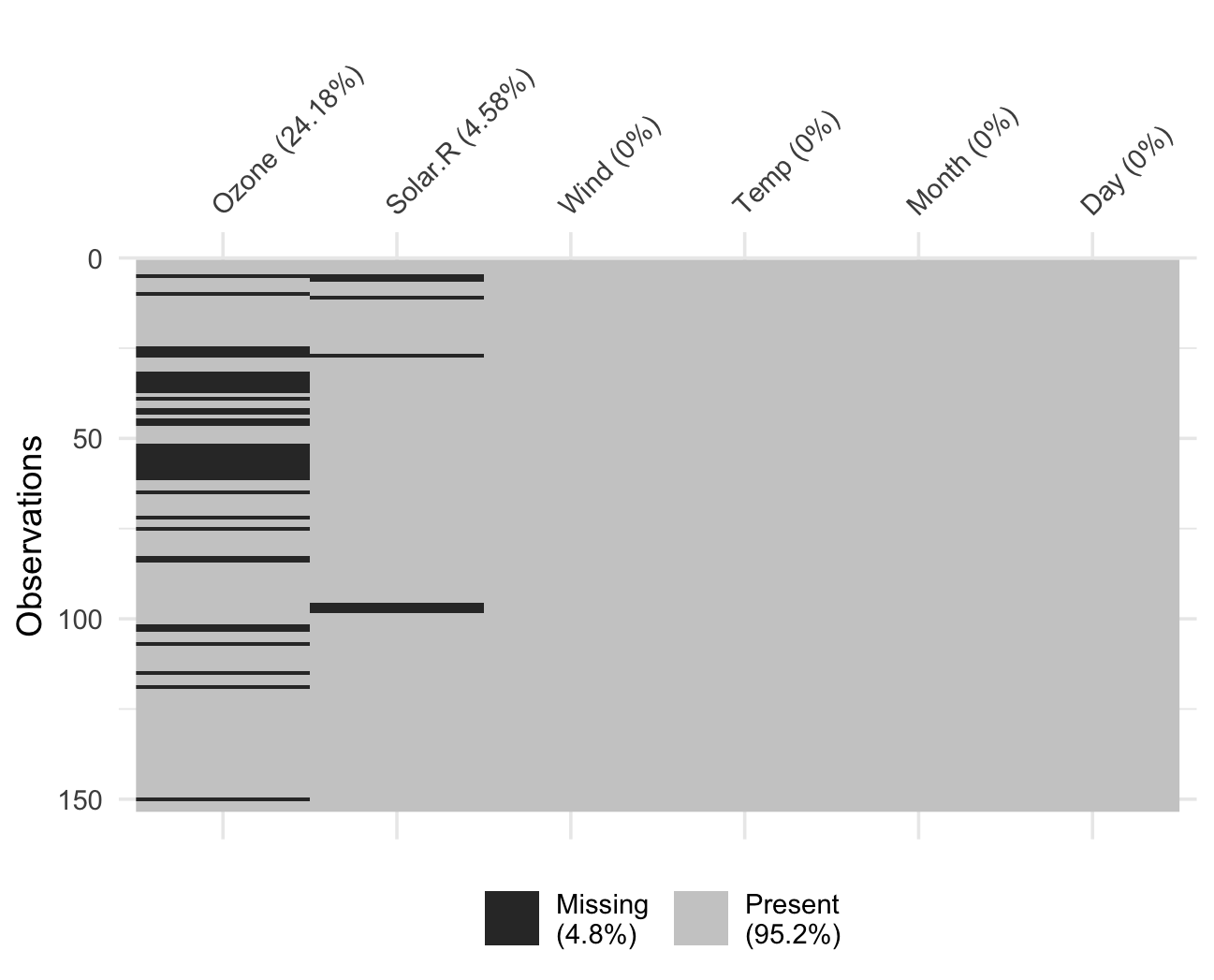
**Table 1. Two tables demonstrating explicit and implicit missing values**. The first table shows the number of goals scored for a player in a given quarter of an AFL match with the first column showing the player name, the second the quarter they played, and the third the goals they scored. Note that Player, “Koenen” has no entries for Quarter 2 and 4. The second table shows the same information from the first table pivoted, with each row being a player and the number of goals they scored in each quarter, with each quarter being a column. We notice that in the second form of the data, we can clearly see that Koenen has missing values. These types of missing values have a name, *implicit missing values*. The first table has implicit missing values, meaning they are implied, and the second table has those implicit missing values explicitly expressed.

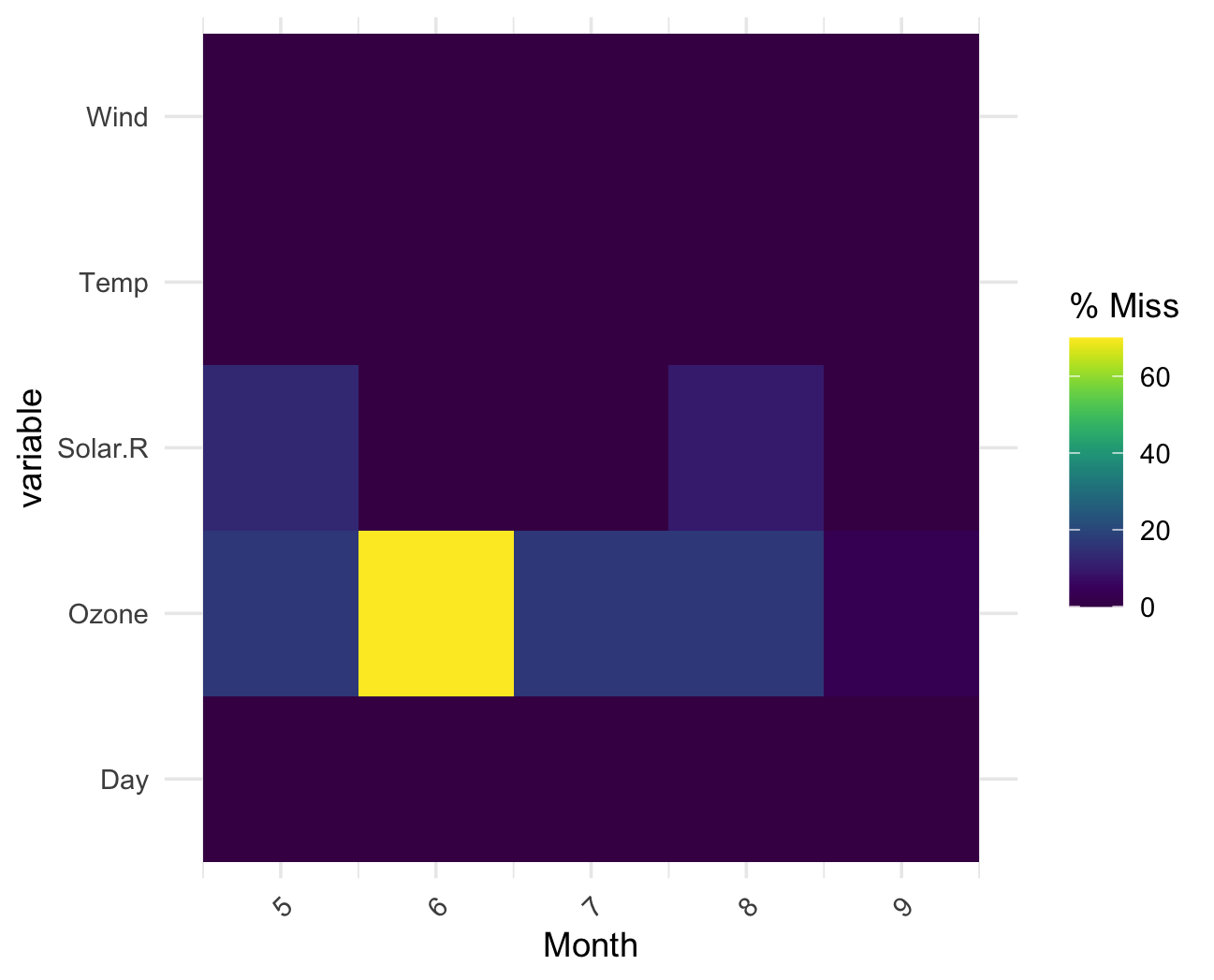
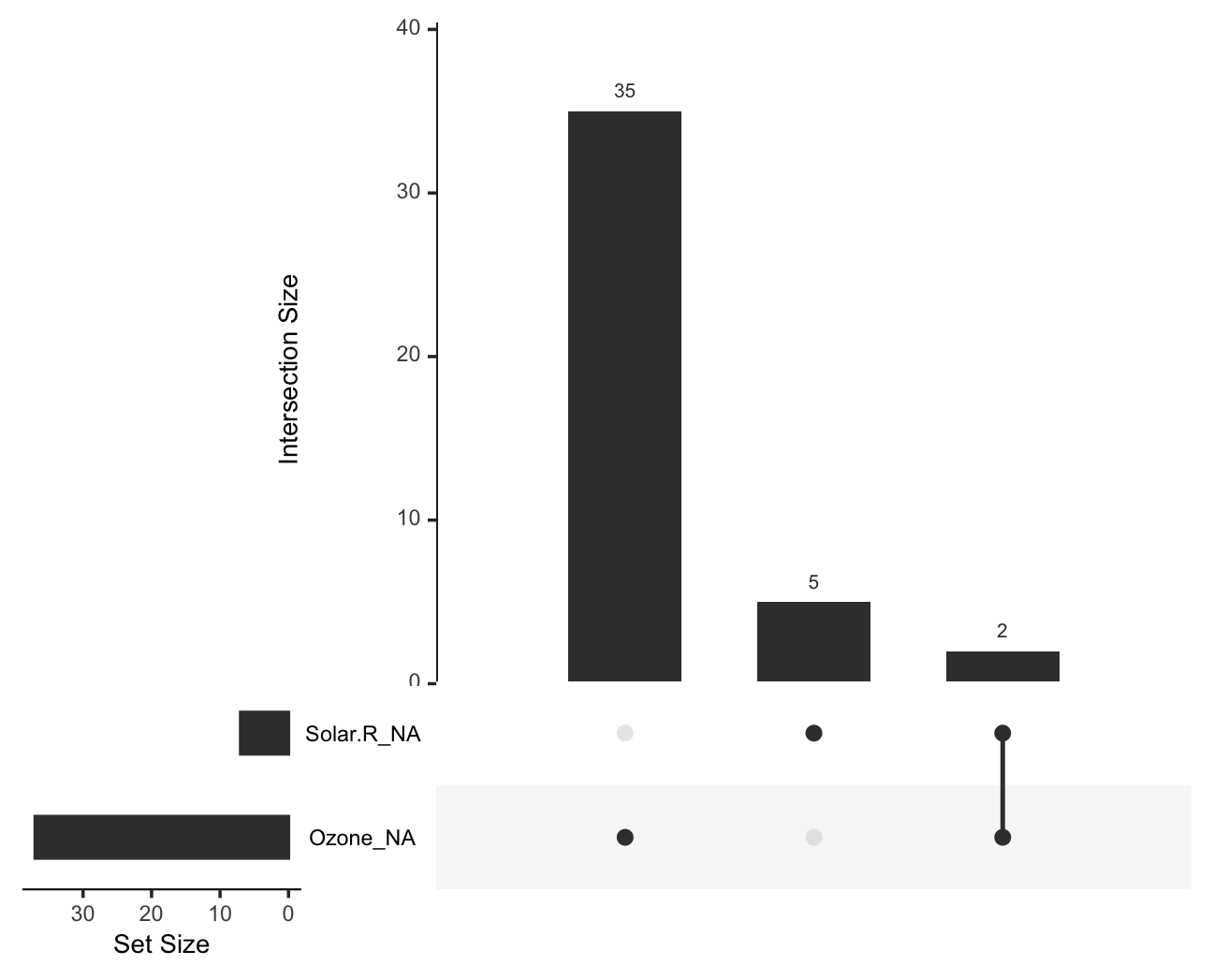
|  |  |  |
| --- | --- | --- |
| Player | Quarter | Goals |
| Coleman | 1 | 0 |
| Coleman | 2 | 1 |
| Coleman | 3 | 2 |
| Coleman | 4 | 0 |
| Koenen | 1 | 1 |
| Koenen | 3 | 2 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Player | 1 | 2 | 3 | 4 |
| Coleman | 0 | 1 | 2 | 0 |
| Koenen | 1 | NA | 2 | NA |

*Visualising missing values*

We recommend that researchers use missing data overview graphics (see Graphics section in <https://arxiv.org/abs/1809.02264>*)*. For example, overview plots (from visdat - <https://joss.theoj.org/papers/10.21105/joss.00355>*)* (Figure 2 A, B, C) give an overall sense of the extent of missing and complete data.





**Figure 2.** Overviews of missing values in airquality data. Panel A shows… Panel B shows … Panel C shows … (simulate data from football study, provide write up in the supplementary materials)

To learn more about exploring missing values, we recommend the vignettes in naniar [[7]](https://www.zotero.org/google-docs/?QHeloi), and the methods in (<https://arxiv.org/abs/1809.02264>*.*)

*Understanding imputation*

Imputing data might feel wrong, as we are ‘making up data’. The truth is, we (generally) can never know what the missing values were. The goal of imputing data is to make the best possible inference from the data. We recommend data is imputed, with a few caveats. Consider removing variables with a high proportion of missing data (e.g., variables with the majority or more missing than not). With respect to a ‘rule of thumb’ for when data should be imputed, we deliberately do not suggest any specific threshold (e.g., impute values when >5% of data are missing otherwise delete rows with <5% missing) and caution against such practice. Rather, we recommend values should generally be imputed.. We recommend avoiding imputation methods that impute the same value, such as mean or median imputation, and instead suggest using imputation methods such as: linear regression, k nearest neighbours, or expectation maximisation.

Much in the way of there is no single ‘best’ statistical method, there is no perfect, one-size-fits-all approach for imputing data. The goal is to generate similar values that might have been otherwise recorded - sometimes this means using a neighborhood approach of finding similar values, other times this means predicting responses using linear or tree models, and other times the most likely value might have been 0 or the last or first value carried forward. For detailed descriptions and a summary of these methods, we suggest (Stefan van Buuren (<https://stefvanbuuren.name/fimd/>)- Schafer (<https://sci2s.ugr.es/keel/pdf/specific/articulo/Schafer_Graham02.pdf>), Cheema 2014 (<https://study.sagepub.com/sites/default/files/cheema.pdf>)).

The imputation methods discussed so far impute a single value for each missing value, and have the eponymous name, “single imputation”. Multiple imputation is a method where multiple values are imputed for each missing value, creating “m” datasets, which are then specially combined during analysis. Multiple imputation is generally seen as the best method to get the most reasonable inference from the data, as it reflects the uncertainty in the missing values. For more information on using multiple imputation, we recommend Stefan van Buuren’s book, “Flexible imputation of missing data”.

No matter the imputation method used, it is essential to compare results of different missing data handling to understand how they may bias the results. For example, comparing analysis results from applying listwise deletion, compared to mean imputation, compared to linear regression imputation. This can reveal bias occurring in imputation methods. An example of this approach is described in the Case Study in (<https://arxiv.org/abs/1809.02264>).

*Recommendations*

The current state of reporting missing data in the sports literature is poor. In this section we discuss recommended practices for exploring and reporting missing data. Our intention is not to provide a ‘cookbook’ style approach to missing data, but rather broad recommendations to help researchers when writing the methods and results sections of a study, and assist researchers when evaluating a study during the peer-review process.

In the methods section of a study, we recommend the following points are addressed:

1. Describe screening procedures (define or example--think about screened out?).
2. State if any observations were dropped (if any).
3. Provide the number and percentage of observations dropped.
4. State if any potential bias was incurred as a result of the screening.
5. If screening is complex, consider providing a workflow diagram explaining how observations were kept or dropped.

In the results section, we recommend addressing the following points:

1. State the number and percentage of observations missing.

Examples: “*There were no missing values*” or “*25 of 280 values were missing (8.92%)*”

1. State if missing data was explored. If there is missingness, provide a summary graphic (see figure XX)

Examples: “*Missing data was explored, revealing a relationship between missingness and age*”, “*Missing data was not explored, as the reasons and impacts were known*”

1. State any reasons known or unknown for missing values and if bias have occurred as a result. For example, a study on injuries, removing players who get injured during the study seems potentially problematic.

Examples: “*Values were missing due to faults in GPS tracking*”, “*Values were missing in a survey as participants did not complete the section*”

1. State actions taken in handling missing data.

Were they removed? Imputed? Another approach?

1. Describe and justify any imputation process used.
2. Explore how imputation or lack of imputation impacts results.

Missing values of ZZ were imputed with a linear regression using terms XX.

Missing values were imputed with the mean values

An example writeup of missing data for a results section is provided below:

*25 of the 280 values in the dataset (~8.92%) were missing. These were due to faults in GPS tracking as participants ran through sections of track covered by forest, a known issue with some GPS tracking. These speed values were interpolated using a nearest neighbours approach, taking inputs of speed, and altitude. The track was imputed using the known track in the area for those sections.*

**Conclusion**

More often than not, researchers will be faced with missing data. The reasons for the missingness, along with the decision of how missing data are treated, can have a profound effect on the validity and accuracy of a study's results. A survey of 2019 articles on football related topics showed that current practice is poor. Only about one in 20 studies reported missing data. To help address this issue, we have discussed recommendations for exploring and reporting missing data, to assist researchers during this important process in the analytical pipeline. Research should consider these recommendations, and pay greater attention to missing data and its influence on research results.

**References**

[1. Sainani KL. Dealing With Missing Data. PM&R. 2015;7:990–4.](https://www.zotero.org/google-docs/?DtxqPJ)

[2. Nakagawa S, Freckleton RP. Missing inaction: the dangers of ignoring missing data. Trends in Ecology & Evolution. 2008;23:592–6.](https://www.zotero.org/google-docs/?DtxqPJ)

[3. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls | The BMJ [Internet]. [cited 2020 Jun 28]. Available from: https://www.bmj.com/content/338/bmj.b2393](https://www.zotero.org/google-docs/?DtxqPJ)

[4. The Age. https://www.theage.com.au/sport/afl/hair-test-afl-suspicious-after-cousins-shaved-head-20100827-13us5.html. 2010;](https://www.zotero.org/google-docs/?DtxqPJ)

[5. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS med. Public Library of Science; 2009;6:e1000097.](https://www.zotero.org/google-docs/?DtxqPJ)

[6. Young CM, Luo W, Gastin P, Tran J, Dwyer DB. The relationship between match performance indicators and outcome in Australian Football. Journal of science and medicine in sport. Elsevier; 2019;22:467–471.](https://www.zotero.org/google-docs/?DtxqPJ)

[7. Tierney N, Cook D, McBain M, Fay C, O’Hara-Wild M, Hester J, et al. naniar: Data Structures, Summaries, and Visualisations for Missing Data [Internet]. 2020 [cited 2020 Aug 9]. Available from: https://CRAN.R-project.org/package=naniar](https://www.zotero.org/google-docs/?DtxqPJ)

**Supplement 1.** [fix ordering]

For example, if a players’ player ratings--a statistic derived from many variables in the dataset--are unavailable for one, or several, Australian Football matches, despite the individual participating in the match and the other variables being recorded.

MCAR refers to missingness that is unrelated to any observed or unobserved data. Missing Australian Football player ratings could be considered MCAR, if the missing ratings are not systematically related to other variables. MAR is where missingness is related to some observed data. Data MAR could arise if... ^^Rob’s out on the full example--players who are injured are less likely to fill out wellness surveys than those who are fully fit.^^. MNAR is when missingness is related to something unobservable. For example, data MNAR could arise when, if studying doping, a player was to remove their head and body hair, to avoid providing a hair sample for testing [[4]](https://www.zotero.org/google-docs/?4Ildjf). In this case, data is missing (the hair sample), and this information is not captured by any other variable that was collected (hair length was not recorded).

**Supplement 2. Systematic search**

The systematic search was performed in accordance with the reporting requirements of the PRISMA statement (ref). The search strategy is described below. The search protocol was not registered, and no funding was provided to support the undertaking of the search.

**Objective**

This review aimed to estimate the proportion of articles published on the topic area of football in 2019 that reported on missing data.

**Method**

Football related articles, published in 2019 were studied. Three major sports medicine databases were searched using the key terms and limits, on May 20, 2020. The databases were: SPORTDiscus, Embase and Cinahl. A list of search terms and complete search statements are provided in Table X.

Included articles were: (a) published in 2019; (b) had an accessible English abstract; (c) had an accessible full text, in English; and (d) included quantitative data. Studies containing original data, including case studies (with time-series data) and n=1 trials, were considered. The search results were imported into EndNote (version 9.3.2). Duplicates were removed, and titles and abstracts were screened against the inclusion criteria. Of the articles identified for full text screening, 10% were sampled, at random, for survey. Because ~30% of the papers from the initial random sample were not eligible for inclusion (e.g., full-texts not in English, conference abstracts, qualitative studies), an additional 5% random sample was taken--with the expectation that 30% of the articles from the additional sample would also be ineligible--so to preserve a 10% screen of the total papers published on football related topics in 2019. Figure X summarises the search process.

Articles were surveyed to determine: (a) whether a missing data statement was included, or if missing data were acknowledged in a figure/table; (b) where data were missing, whether data were considered MCAR, MAR or MNAR; (c) whether imputation was used, and the imputation method; and (d) whether articles that did not include a missing data statement shared their data. Two authors independently extracted the data. Agreement was X%; discrepancies were resolved by NJT. Findings are reported as the proportion and 95% confidence interval (CI). Confidence intervals were calculated using the Clopper-Pearson method for the binomial distribution (ref),via the ‘binom’ package (ref) in R (version 3.5.1).

**Results**

The search returned a total of 3,863 articles (Figure X). Of the 199 articles screened, 136 were deemed eligible, representing ~10% of total articles published on football related studies in 2019 (Figure X). The proportion of articles that included a missing data statement (or acknowledged missing data in a table/figure) was 5.9% [95% confidence interval; X% to X%] or 8/136. A breakdown of the extracted information is provided in Table X.

**Summary**

A survey of 136 articles published in 2019, on football related topics showed that only one in 20 reported missing data. This result suggests that missing data are rarely considered by sport science researchers.

**References**

[Dave use plos one reference] Moher D. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. Ann Intern Med. 2009;151(4):264–269.<https://doi.org/10.7326/0003-4819-151-4-200908180-00135>.

Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26(4):404–413. <https://doi.org/10.1093/biomet/26.4.404>

Dorai-Raj S. binom: Binomial confidence intervals for several parameterizations. R Package ‘binom’ version 1.1-1. 2014.<https://cran.r-project.org/web/packages/binom/binom.pdf>

Search details

|  |  |  |
| --- | --- | --- |
| Sportdiscus | Embase | Cinahl |
| exercise science OR sports science OR physiology OR biomechanics OR nutrition OR training OR testing OR sports medicine OR performance analysis OR performance OR sport psychology OR coaching | (exercise AND science OR (sports AND science) OR physiology OR biomechanics OR nutrition OR training OR testing OR (sports AND medicine) OR 'performance analysis' OR (('performance'/exp OR performance) AND ('analysis'/exp OR analysis)) OR 'performance'/exp OR performance OR 'sport psychology'/exp OR 'sport psychology' OR (('sport'/exp OR sport) AND ('psychology'/exp OR psychology)) OR 'coaching'/exp OR coaching) AND [2019-2019]/py | exercise science OR sports science OR physiology OR biomechanics OR nutrition OR training OR testing OR sports medicine OR performance analysis OR performance OR sport psychology OR coaching |
| football OR soccer OR indoor soccer OR futsal OR rugby league OR rugby union OR gaelic football OR australian rules football OR AFL OR american football OR gridiron football OR touch football | (football OR soccer OR (indoor AND soccer) OR futsal OR rugby OR (rugby AND league) OR (rugby AND union) OR (gaelic AND football) OR 'australian rules footbal' OR (('australian'/exp OR australian) AND rules AND footbal) OR afl OR 'american football'/exp OR 'american football' OR (('american'/exp OR american) AND ('football'/exp OR football)) OR 'gridiron football' OR (gridiron AND ('football'/exp OR football)) OR 'touch football' OR (('touch'/exp OR touch) AND ('football'/exp OR football))) AND [2019-2019]/py | football OR soccer OR indoor soccer OR futsal OR rugby league OR rugby union OR gaelic football OR australian rules football OR AFL OR american football OR gridiron football OR touch football |
| Published Date: 20190101-20191231; English Abstract Available; Peer Reviewed | #1 AND #2 AND [english]/lim AND 2019:py | Published Date: 20190101-20191231; English Abstract Available; Peer Reviewed |
| Articles: 1,372 | Articles: 1,756 | Articles: 735 |
| https://gateway.library.qut.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=sph&bquery=(((football)+OR+(soccer)+OR+(indoor+soccer)+OR+(futsal)+OR+(rugby+league)+OR+(rugby+union)+OR+(gaelic+football)+OR+(australian+rules+football)+OR+(AFL)+OR+(american+football)+OR+(gridiron+football)+OR+(touch+football)))+AND+(((exercise+science)+OR+(sports+science)+OR+(physiology)+OR+(biomechanics)+OR+(nutrition)+OR+(training)+OR+(testing)+OR+(sports+medicine)+OR+(performance+analysis)+OR+(performance)+OR+(sport+psychology)+OR+(coaching)))&cli0=DT1&clv0=201901-201912&cli1=AA1&clv1=Y&cli2=RV&clv2=Y&authtype=sso&custid=qut&type=1&searchMode=Standard&site=ehost-live&scope=site |  | https://gateway.library.qut.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=c8h&bquery=(+exercise+science+OR+sports+science+OR+physiology+OR+biomechanics+OR+nutrition+OR+training+OR+testing+OR+sports+medicine+OR+performance+analysis+OR+performance+OR+sport+psychology+OR+coaching+)+AND+(+football+OR+soccer+OR+indoor+soccer+OR+futsal+OR+rugby+league+OR+rugby+union+OR+gaelic+football+OR+australian+rules+football+OR+AFL+OR+american+football+OR+gridiron+football+OR+touch+football+)&cli0=DT1&clv0=201901-201912&authtype=sso&custid=qut&type=1&searchMode=Standard&site=ehost-live&scope=site |

Search results

|  |  |
| --- | --- |
| Imported | 3,863 |
| After duplicates removed | 1,954 |
| After title and abstract screen | 1,373 |
| Random sample full-text | 199 (needed for 10% sample) |
| Full-text included | 137 |
| Full-text excluded | Full-text papers excluded (n=62) with reasons:  Qualitative study (n=)  Not in english (n=)  Unable to access full-text (n=) |

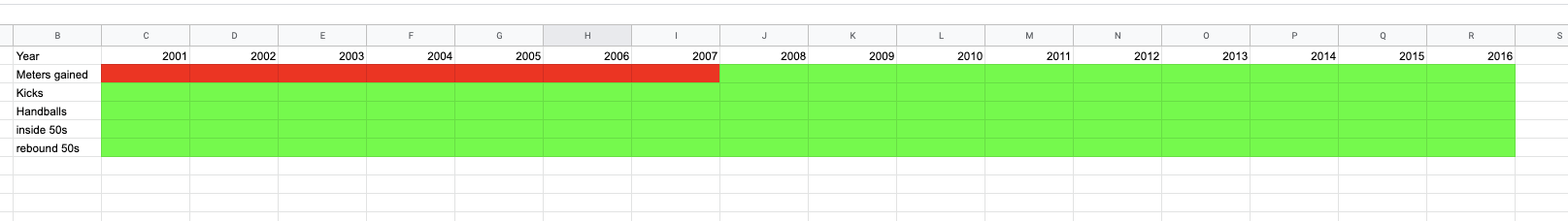
**Figure 1.** PRISMA flow chart.

Rob’s idea [case study section]:

* Including a spectrum of worst practice - best practice
  + Worst
    - Missing data can be observed in how statistical methods are reported
      * e.g., missing meters gained
  + less worse
    - Missing data reported, but no handling stated
      * e.g., 17% of data is missing, nothing mentioned about deletion/imputation  
        <https://www.sciencedirect.com/science/article/pii/S0167945719301939?casa_token=QfaCiQJaBQMAAAAA:tL3pKZG7s8-umTgU-cVZ53ZuPLchja5OExOEnkohWhjLH4NPGtqeFkSWlZq0EiPG9ip-ZnQ>
      * e.g., N = 18, and statistical analysis reports N = 14 (Robin, tweet, Dave to request paper)
  + neutral?
    - There is missing data, and this is reported?
  + good
    - Missing data reported, handling is reported, might be bad imputation like mean imputation / data not shared.
  + best
    - Missing data reported, handling is reported, appropriate imputation, data shared.

Appendix of techniques for dealing with missing values

* scripts for exploring missingness
* How to unpack the missingness in the paper text
* Ways to visualise missing data for reviewers, and readers of paper to get a feel of how much to take away.
  + So the meters gained example, I think a vis could be something like the figure below. Also a way to visualise missing values in athlete studies over time, if for whatever reason on X1,x7,x35 day couldn’t do a measurement on athlete 1. ^^instead of google sheet something ggplot^^
    - vis\_miss from visdat/naniar:



^^Background in sports and exercise science^^

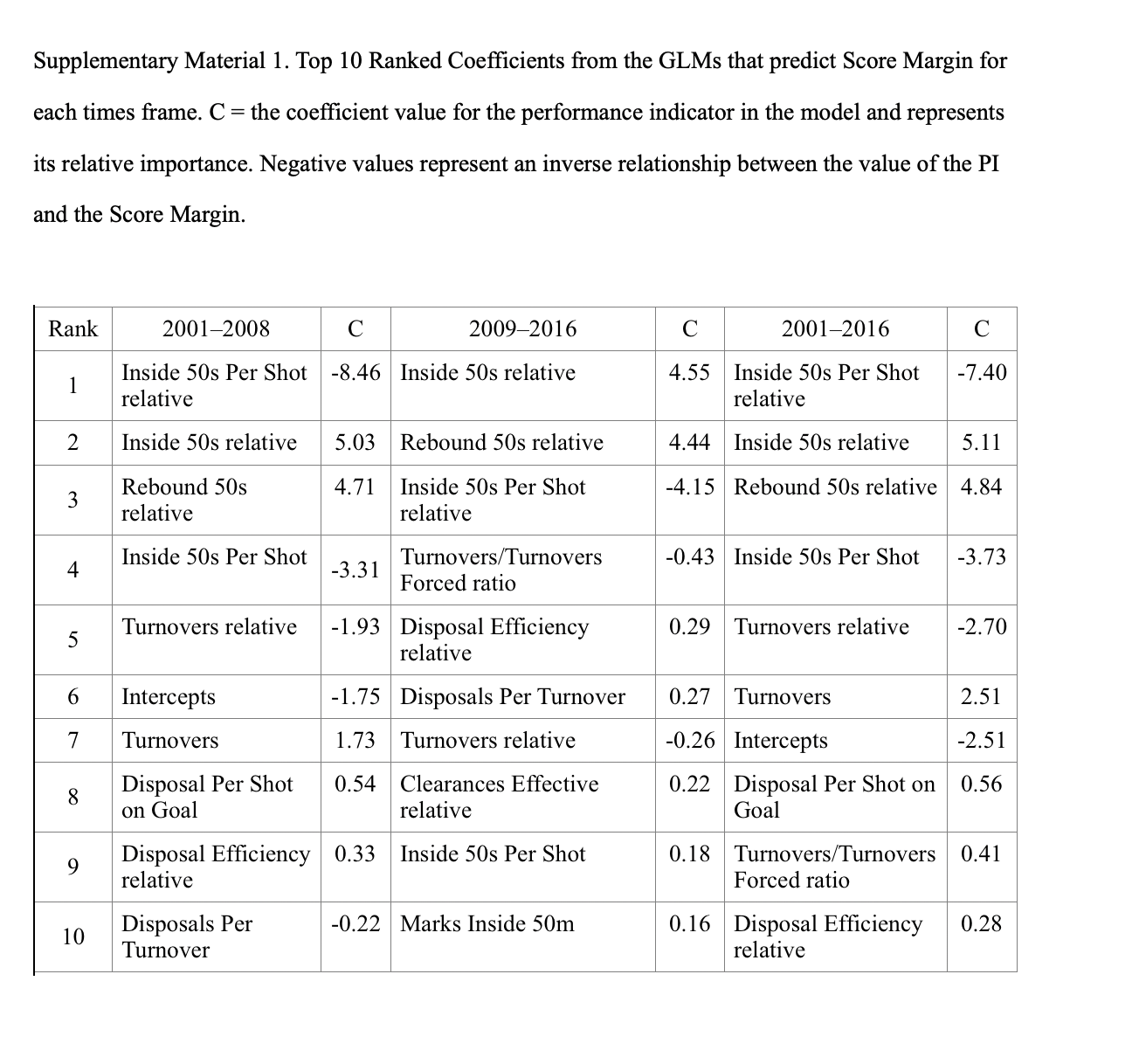
* Missing data (or the absence of missing data) is often not reported
* In general, imputation is not used
* Missing data is never really discussed
* Data are usually fit with ANOVA’s, therefore values are subbed in, e.g., mean. (really common practice)

**Notes/examples**

Example 1: <https://www.sciencedirect.com/science/article/abs/pii/S1440244018303141>

Problem—The most important predictor, meters gained relative isn't available in the vast majority of the dataset (2001-2008 for example) (which is where the ''prediction accuracy'' was the worst. This isn't mentioned anywhere in the paper, that their most important predictor isn't available. So as a reviewer, unless data is provided, or code shared. You are relying on domain knowledge to know this.

The supplementary materials which provides coefficients for the GLMS fitted

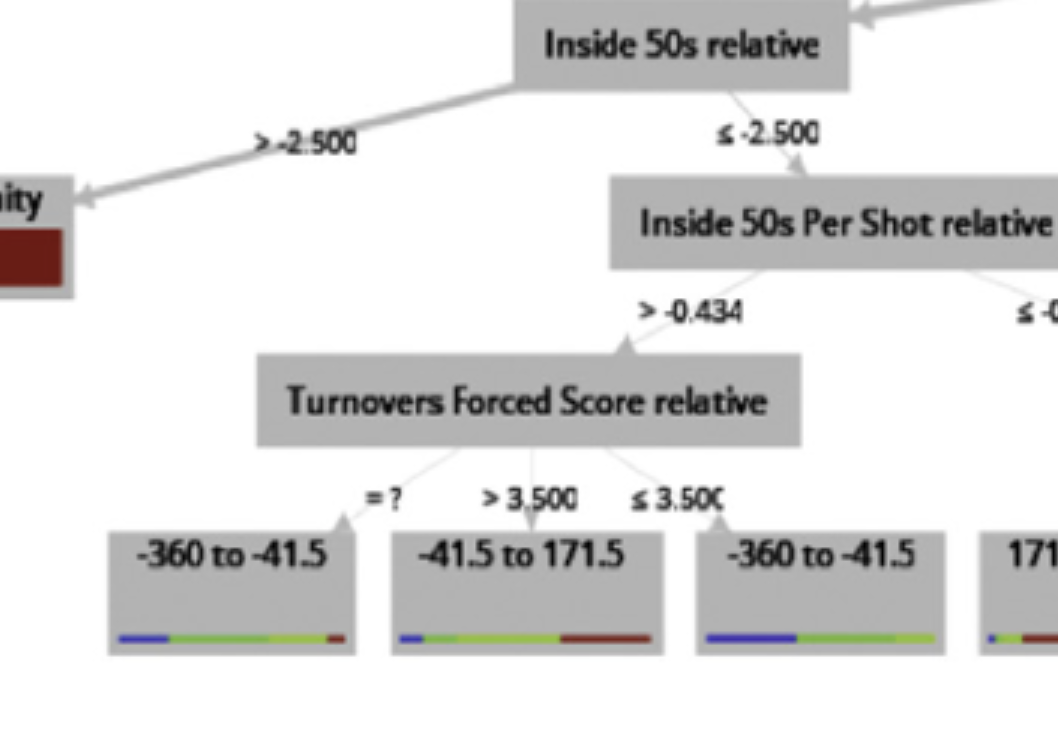


No meters gained? This is weird right, talk about a circle, we have meters gained most important, but GLM perform similar to decision tree but the GLM doesn’t have meters gained (the most important variable) according to the decision trees?

Oh and meters gained isn’t available throughout the dataset.

“GLMs were also created to predict Score Margin in each timeframe and the prediction accuracies are similar for each time frame but are higher than the decision tree prediction accuracies.”

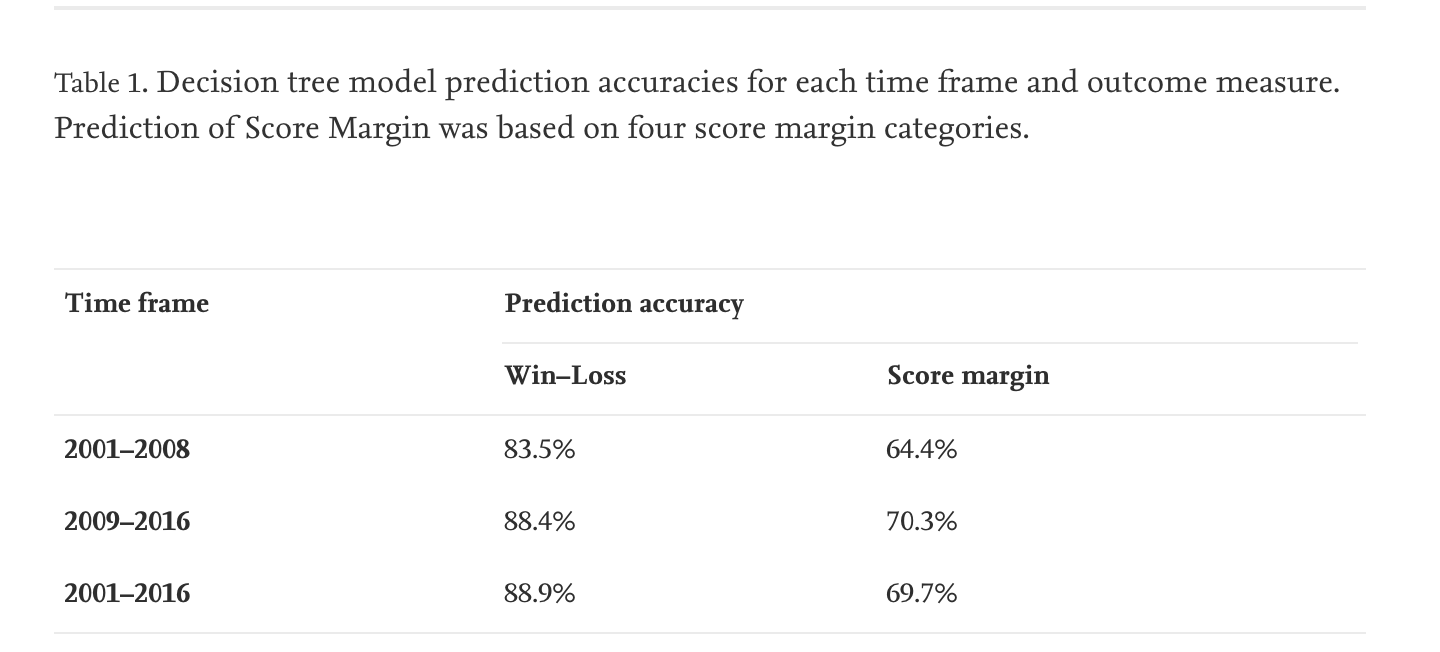
And whats with figure 2?



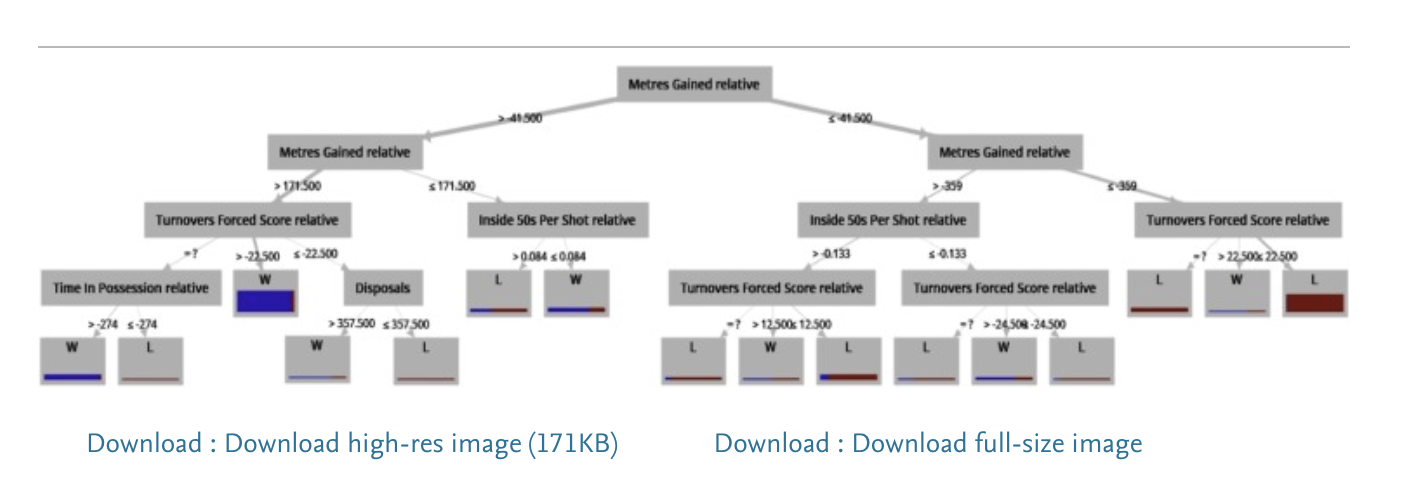
The end buckets are supposed to be meters gained difference, but we have a ?

Example 2: <https://twitter.com/championdata/status/472174368059957248>





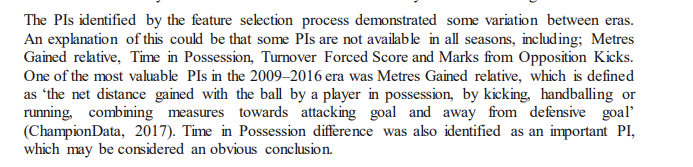
So if meters gained was first recorded in 2007, what does this mean for the time frames being compared?



Like OK that is 2009-2016, but what about the 2001-2016 was meters gained only used for 2007-2016 so what was inputted 2001-2007 and if we compared it to 2009-2016 where meters gained was FULLY available we have a worse win/loss pred and a in the dataset with the MOST IMPORTANT VARIABLE AVAILABLE ALL THE WAY THROUGH IT I am really confused.

**Modelling Match Outcome in Australian Football: Improved accuracy with large databases**

(similar authors as above—Tran, Young etc)

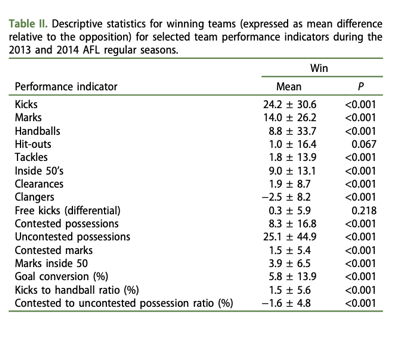


Problem: The most important variable, in 2019 published without mentioning its availability, nor are results presented in a way that gives the reader an idea of what was done? Seriously your trying to predict an outcome and literally the variable doesn't EXIST what was entered in?

Data sharing/code sharing would have made this available to reviewers and perhaps changed the tone of the papers.

Example 3:  
[https://shapeamerica.tandfonline.com/doi/abs/10.1080/02640414.2015.1066026#.XpkfaFMzZhE](https://protect-au.mimecast.com/s/FqW3CZYMqwSk7AkMsKELku?domain=shapeamerica.tandfonline.com)

These results on data available to public (basic stuff like counting of kicks isn't reproducible)—i.e., the table here:



Means should be available?

Example 4: <https://journals.sagepub.com/eprint/Pv3bEcp6EQMnPAa24cuP/full>

Problem: <https://twitter.com/RankingSw/status/961502281874092032>

A screenshot of a cell phone

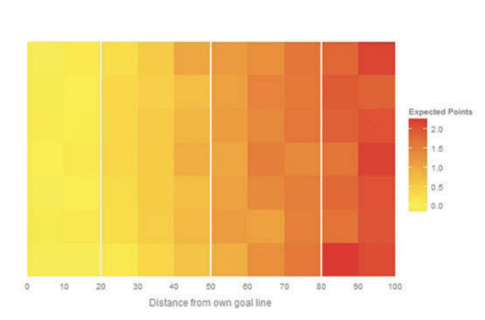
Description automatically generated



So why didn't the authors of the paper know this? Was it because the underlying model (AFLPR) wasn't available at the time? etc But certainly as a reviewer this is a thing someone should know before reviewing?

Example 5: <https://www.tandfonline.com/doi/abs/10.1080/02640414.2015.1066511>

Figure below seems odd:



* Progression of science, self-correction and the need for ‘self-correction to not be strike against your name’
* Speak about the content in the papers wanting to push the knowledge, not just get rushed for the purpose of a publication.

Recommendations – Ask the question: what am I being provided here to review?

* Barriers
  + Data privacy agreements (e.g., NDA), the reviewer would have to sign one, maybe this is something that’s not can occur, so what’s the minimum.
  + Betting privacy, players union—you couldn’t ever share who the slowest player in the NRL same with the AFL
* Framed from the reviewer’s point of view, needs to be a minimum understanding of what is being done in the paper.
* Exploratory and missingness (common problem in sports, maybe not able to get access to the player)
* Minimum requirements: please give the summary stats, i.e., the variable name, description, the dates collect, number of missing values, etc.
  + Describe the data in some type of meaningful sense
  + If there’s definitions, over what period does the definition hold for.
* Idea scenario: Start data, end date.

Example from Twitter message

Dave sent a twitter message about this paper here.

<https://sci-hub.tw/https://doi.org/10.1080/02640414.2020.1740490>

From the Paper

This study used commercially available data from the 2015 AFL season supplied by Champion Data, the official statistics provider for the AFL. An exemption from ethics review was granted by the institutional ethics committee. The original data set included all 149,449 possession events from all matches played during the 2015 AFL season (18 teams, 665 players, 206 regular season and finals series games). A data subset was created to focus the analysis on regular season matches in order to explore wholeof-league patterns (i.e., finals matches were omitted as they are only played by teams in the top 8 end-of-season ladder positions). This subset was filtered further to include only those sequences of possession events that i) started with a kick-in following a behind, and ii) ended with the offensive team losing possession through a score, stoppage, or turnover. These sequences are referred to herein as “kick-in possession chains”, where a sequence begins with the player who first receives the ball from a kick-in and excludes the initial disposal out of the goal square, unless the player starting with possession “plays on” to himself. Possession chains that involved only one player or were incomplete due to end-of-quarter time restrictions were excluded. Chains from 2 drawn matches were also excluded. The final data set used for analysis comprised a total of 1,720 kick-in possession chains from 194 games.

Now thankfully we can use fitzRoy to check out how many behinds there were in 2015 as per the subset.

fitzRoy::get\_afltables\_stats(start\_date="2015-01-01", end\_date="2016-01-01")%>% #Season chosen

select(Season,Home.score, Away.score, Round, Home.team, Away.team, Round, AQ4B, HQ4B)%>%# variables needed

filter(!Round %in% c("EF", "SF","QF","PF", "GF"))%>% #excluding finals

filter(Home.score != Away.score)%>% #excluding draws

distinct()%>%

mutate(totalbehinds=AQ4B+HQ4B)%>% #away behinds + home behinds

summarise(totalbehindsin2015=sum(totalbehinds))

totalbehindsin2015

<int>

1 4275

>

So if each quarter ended with a invalid observation (didn’t result in a score, turnover etc) this would be pretty high I’d imagine but help illustrate the point

4\*194

So we have 4275 kick ins from behinds, of which at MOST 194\*4 (776) would be excluded say if every quarter had a kick after the siren that resulted in a behind.

So that leaves us with 4275-776 total kick ins which is 3499 kick in ins total

**So this data suggests that only 1720/3499 or 0.491569 of kick ins resulted in a chain? in the study.**

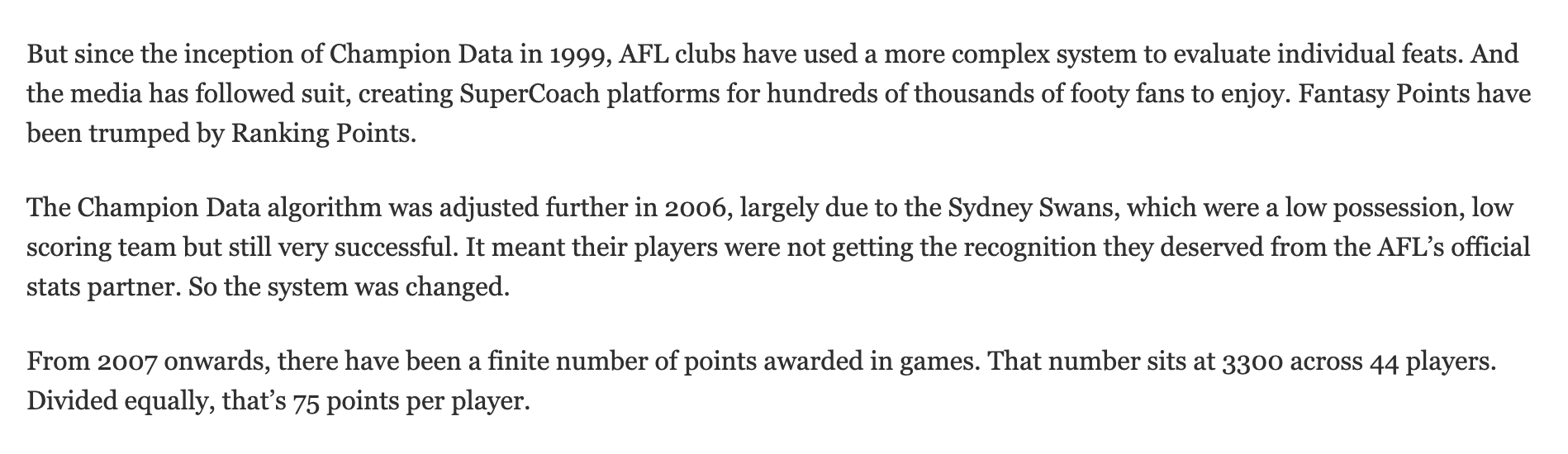
This seems not good.

To put another way,

out of the total behinds or ‘’opportunities for chains to start’’ if we take the exclusion rule. **“ Possession chains that involved only one player or were incomplete due to end-of-quarter time restrictions were excluded.”**

**Then**

**(4275-1720 )/ 4275 or 0.5976608 of chains are EXCLUDED that seems massive!**



New Paper: Recommending how to share data in sports science

* Simplest approach
  + - * If you report the summary statistics for each variable
      * Then recreate the data from that
        + The min-max and sample evenly
        + the distribution