Missing data: Current practice in football research and recommendations for improvement

Table of Contents

## Warning: package 'tidyr' was built under R version 3.5.2

## Warning: package 'stringr' was built under R version 3.5.2

## Warning: package 'forcats' was built under R version 3.5.2

## Warning: package 'visdat' was built under R version 3.5.2

Affiliations:

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

2: University of New South Wales, Department of Statistics, School of Mathematics and Statistics, Sydney, Australia.

3: Monash University, Department of Econometrics and Business Statistics, Melbourne, Australia.

4: Australian Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), Melbourne, Australia.

**Acknowledgement/conflict of interest**

The authors declare no conflicts of interest.

**Corresponding author**

Nicholas J Tierney Monash University, Department of Econometrics and Business Statistics, Melbourne, Australia. Address: Email:

**Data availability**

**Word count**

## For information on available language packages for 'koRpus', run  
##   
## available.koRpus.lang()  
##   
## and see ?install.koRpus.lang()

##   
## Attaching package: 'koRpus'

## The following object is masked from 'package:readr':  
##   
## tokenize

|  |  |  |
| --- | --- | --- |
| Method | koRpus | stringi |
| Word count | 2802 | 2739 |
| Character count | 18487 | 18480 |
| Sentence count | 222 | Not available |
| Reading time | 14 minutes | 13.7 minutes |

# Abstract

Missing data are often unavoidable. The reason values go missing, along with decisions made of how missing data are handled (deleted or imputed), can have a profound effect on the validity, accuracy and usability 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 practices to avoid 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 statement was included. As expected, the proportion of studies in football research that included a missing data statement was low, at only X% (95% confidence interval; X% to X%); suggesting that missingness is seldom considered by researchers. It is suspected that this result is consistent with practice in 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:** Exercise, imputation, missingness, missings, naniar, sport

# Introduction

Missing data are values that should have been observed, but were not. Since the values are unobserved, this can mean the population of interest might not be properly sampled, inducing bias into a study’s results. Left undetected or poorly handled, missing data can undermine the validity and accuracy of research results (Sainani 2015; Nakagawa and Freckleton 2008; Sterne et al. 2009). 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. Contextual factors–such as, the study design and objective, and pattern of missingness–determine how missings should be handled, on a case-by-case basis (Sainani 2015).

For clarity, we describe an example of missingness. An Australian Football data collector was taking lunch during the third quarter, so no data was recorded for this time period. This would be considered missing data. Compare this to the game being cancelled due to a pandemic. There is no missing data because there is no data to be observed. Similiarly, if a study screens out individuals who do not meet inclusion criteria, this is not a missing value, because there is no intent for these values to be observed or measured.

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 the least harmful, and MAR and MNAR biased (Newman 2014). Football relevant examples are provided in Supplement 1, and more detailed explanations of MCAR, MAR, MNAR can be found elsewhere (Sainani 2015; Nakagawa and Freckleton 2008; Sterne et al. 2009).

The three categorisations of missingness types provide a useful framework to describe bias arising from missing data. If you strongly suspect data is MCAR, your results are less likely to be biased. If you suspect MAR or MNAR, there is bias in your results. This does not mean data MAR or MNAR make results invalid, instead it identifies potential bias, which can be used to 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, are reported (Schafer and Graham 2002). While there has been significant interest in modelling missing data (**???**), the exploration and reporting of missings have received less attention (Schafer and Graham 2002). This article aimed to: (1) estimate the proportion of articles that report missing data in football research; (2) highlight several practices that should be avoided; 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 (**???**). 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 2). Included articles were: (a) written in English; (b) had an accessible full-text; and (c) included quantitative data. The PRISMA figure in Supplement 2 summarises the search process, and further details of the search can also be found in Supplement 2.

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 was reported or acknowledged; and if not, whether a dataset was shared. Where applicable, information relating to how missing values were treated was also extracted (see Supplement 2). Two authors independently extracted the data (X% agreement). Findings are reported as the proportion and 95% confidence interval.

# Results

Of the 199 articles screened, 136 met the inclcusion criteria, representing ~10% of the articles found (see Supplement 2, Figure 1). 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 practice in the wider sports science literature. The low proportion of papers reporting missingness could be explained by a lack of awareness and education on missing data (<https://bjsm.bmj.com/content/early/2020/08/19/bjsports-2020-102607.abstract>), similar to other areas of sciences (refs-education and ecology examples)(2)**references needed**. Below we discuss several practices to avoid in relation to missingness, and provide recommendations for exploring, visualising and reporting missing data.

## Practices to avoid across the analysis pathway

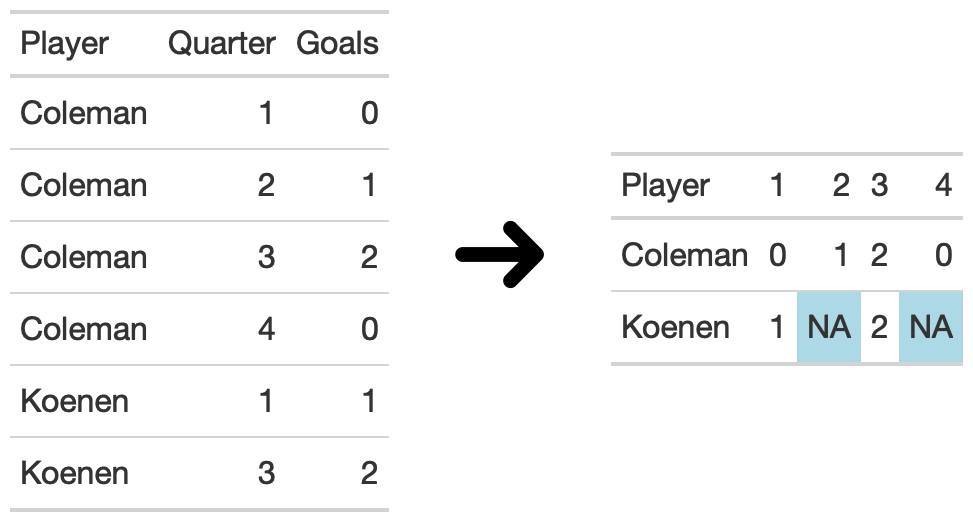
Practice to avoid: *Unreported missing values.* While not always the case, even if not reported, missing data can be obvious. A recent paper examined the relationship between match performance indicators and outcome, in Australian Football between the 2001 and 2016. Of the 91 team performance indicators included in the analysis, at least one variable (i.e., ‘meters gained’) was not available over the entire 2001-2016 period (Young et al. 2019)**Check this reference**. **We might need to add a sentence about ‘question mark’ in the figures?** 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. The absence of a missing data statement when data are missing is not unique to the study above (Young et al. 2019)**Check this reference**. Authors should include a missing data statement, irrespective of whether there is data missing, or not.

Practice to avoid: *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) (Young et al. 2019)**Check this reference**. 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). 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.

Practice to avoid: *Not evaluating the effect, or choice, of imputation.* The method used to impute missing data has the potential to affect the outcome of a study. 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–in terms of 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 @ref(tab:implicit-missings), 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: through joins when merging data without corresponding values, surveys, an inability to collect a biological sample (e.g., venous blood), equipment malfunction, failure, or not being worn.

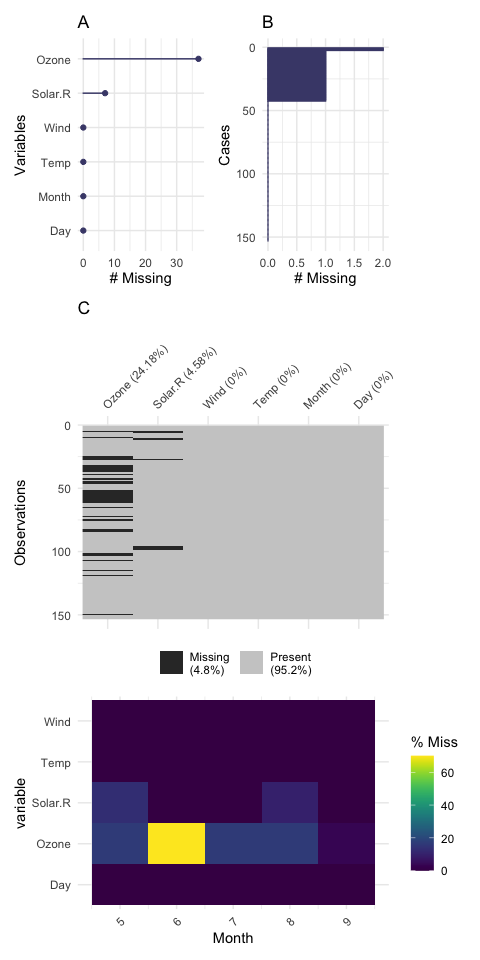


Two tables demonstrating explicit and implicit missing. The first table shows the number of goals scored for a player in a given quarter of an AFL match with the first column showing the player name, the second the quarter they played, and the third the goals they scored. Note that Player, ‘Koenen’ has 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.

## Visualising missing values

We recommend that researchers use missing data overview graphics (see Graphics section in (Tierney and Cook 2018); for example, overview plots (Tierney 2017).

(Figure @ref(fig:missing-overview)) give an overall sense of the extent of missing and complete data. The simualted data set contains … **Add short description (1-2 sentences) here of the simulated data set**



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 (Tierney et al. 2019), and the methods in (Tierney and Cook (2018)).

## Understanding imputation

Imputing data might feel wrong, as we are ‘making up data’. The truth is, we can (generally) 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). We deliberately do not suggest a ‘rule of thumb’ for the amount of missingness that should be imputed, and caution against looking for a specific threshold. Rather, we recommend: values should generally be imputed, avoiding methods imputing the same value (such as the mean or median), and suggest using imputation methods such as: linear regression, k-nearest neighbours, or expectation maximisation.

Missing values can occur in the both the predictors (the “independent variables”) and response variable (the “dependent variables”). Care should be taken when imputing data, as typically only predictors should be imputed. For discussion on imputing the response value, see … (**???**). **(note: take a look again)**

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. Or it could mean predicting responses using linear or tree models. 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 (**???**; Schafer and Graham 2002; Cheema 2014).

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, (**???**).

Irrespective of 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 Tierney and Cook (2018).

# Recommendations

The current state of reporting missing data in the sports literature has room for improvement. 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

Our survey of 2019 articles on football related topics showed that current practice of reporting missing data is poor, with only about one in 20 studies reported missing data. This could suggest that this aspect of data anlaysis receives little attention. To address this issue and assist researchers, we have provided recommendations for reporting and exploring missing data. Research should consider these recommendations, and pay greater attention to missing data and its influence on research results.

# References

# Supplement 1

# Supplement 2

# Supplement 3

Cheema, Jehanzeb R. 2014. “A Review of Missing Data Handling Methods in Education Research.” *Review of Educational Research* 84 (4). American Educational Research Association: 487–508.

Nakagawa, Shinichi, and Robert P Freckleton. 2008. “Missing Inaction: The Dangers of Ignoring Missing Data.” *Trends in Ecology & Evolution* 23 (11): 592–96.

Newman, Daniel A. 2014. “Missing Data: Five Practical Guidelines.” *Organizational Research Methods* 17 (4). SAGE Publications Inc: 372–411.

Sainani, Kristin L. 2015. “Dealing with Missing Data.” *PM & R: The Journal of Injury, Function, and Rehabilitation* 7 (9): 990–94.

Schafer, Joseph L, and John W Graham. 2002. “Missing Data: Our View of the State of the Art.” *Psychological Methods* 7 (2): 147–77.

Scheffer, Judi. 2002. “Dealing with Missing Data.” *Research Letters in the Information and Mathematical Sciences* 3. Massey University: 153–60.

Sterne, Jonathan a C, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. 2009. “Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls.” *BMJ* 338: b2393.

Tierney, N, D Cook, M McBain, C Fay, M O’Hara-Wild, and others. 2019. “Naniar: Data Structures, Summaries, and Visualisations for Missing Data.” *R Package*.

Tierney, Nicholas. 2017. “Visdat: Visualising Whole Data Frames.” *The Journal of Open Source Software* 2 (16): 355.

Tierney, N J, and D H Cook. 2018. “Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations.” *arXiv Preprint arXiv:1809.02264*. arxiv.org.

Young, Christopher M, Wei Luo, Paul Gastin, Jacqueline Tran, and Dan B Dwyer. 2019. “The Relationship Between Match Performance Indicators and Outcome in Australian Football.” *Journal of Science and Medicine in Sport / Sports Medicine Australia* 22 (4): 467–71.