Missing Data: Current Practice in Football Research and Recommendations for Improvement

David N Borg(1,2), Robert Nguyen(3), Nicholas J Tierney(4,5)

4 December 2020

Preprint (not peer reviewed)

Affiliations

1: Griffith University, Menzies Health Institute Queensland (MHIQ), The Hopkins Centre, Brisbane, Australia. 2: Griffith University, School of Allied Health Sciences, Brisbane, Australia. 3: University of New South Wales, Department of Statistics, School of Mathematics and Statistics, Sydney, Australia. 4: Monash University, Department of Econometrics and Business Statistics, Melbourne, Australia. 5: Australian Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), Melbourne, Australia.

Co-Author Agreement: We the authors agree to the sharing of this preprint on *SportRχiv*

Twitter: elborgo9, roberto_boberto, nj_tierney

Corresponding Author

Nicholas J Tierney
Monash University, Department of Econometrics and Business Statistics, Melbourne, Australia.
Email: nicholas.tierney@monash.edu

Data Availability

The code to reproduce the results and plots in the paper can be accessed at the GitHub repository: https://github.com/SciBorgo/Missing-data.

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 and accuracy of study

1

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 136 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 11.0% (95% CI: 6.3% to 17.5%); suggesting that missingness is seldom considered by researchers. We recommend that 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. Researchers should consider these recommendations, and pay greater attention to missing data and its influence on research results.

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 [1–3]. Data can go missing in 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 [1].

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. Similarly, if a study screens out individuals who do not meet inclusion criteria, these are not missing values, 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 [4]. Football relevant examples are provided in Supplement 1, 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, 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 can 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 [5]. While there has been significant interest in modelling missing data [6,7], the exploration and reporting of missings have received less attention [5]. The current article aimed to: (1) estimate the proportion of studies 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 [8]. Articles on football topics, or involving footballers, 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.

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 (Supplement 2). Two authors independently extracted the data (91.9% agreement). Disagreements were resolved by author DNB. Results are reported as the proportion and 95% confidence interval (CI).

Results

Of the 199 articles screened, 136 met the inclusion criteria, representing ~10% of the articles found (see Supplement 2, Figure 1). The proportion of articles that reported missing data was 11.0% (95% CI: 6.3% to 17.5%). Of the articles that did not mention missing data (121/136), only 7.4% (95% CI: 3.5% to 13.7%) 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.

Discussion

Our survey of 136 articles published on football related topics found that only one in nine papers (11.0%) 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 might be explained by a lack of awareness and education [9], similar to other sciences [10,11]. Below we discuss several practices to avoid in relation to missingness, and provide recommendations for exploring, visualising and reporting missing data.

Practices to Avoid

Unreported missing values. While not always the case, even if not reported, missing data can be obvious. For example, a study examined the relationship between match performance indicators and outcome, in Australian Football between the 2001 and 2016 [12], where at least one of the 91 performance indicators ('meters gained') included in the analysis was not available over the entire 2001-2016 period [13]. This was not noted anywhere in the paper, or in the supplementary materials. It is unknown whether imputation was used. 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. Authors should include a missing data statement, irrespective of whether data are missing, or not.

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 the 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. Authors sometimes choose to replace missing observations with mean values, such as a participant or group average [14,15]. 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 in the dataset, and can bias (underestimate) standard errors [16]. 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.

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 [17].

Common Causes of Missing Values

Broadly, there are two types of missing values in data: explicit and implicit missings. Explicit values are missing, but recorded; whereas implicit missings are not recorded in the dataset. For example, in Figure 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 should have been recorded as 0, rather than NA. Other places missing data can arise include: through joins when merging data without corresponding values, on surveys, an inability to collect a biological sample (e.g., venous blood), or due to equipment malfunction, failure, or not being worn.

Figure 1: Two tables demonstrating explicit and implicit missings. 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 column the quarter they played, and the third column the goals they scored. Note that Player, 'Koenen' has no entries for Quarters 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 [18]; for example, overview plots [19]). The panels in Figure 2 give an overall sense of the extent of missing and complete data, in a simulated injury dataset. To learn more about exploring missing values, we recommend the vignettes in *naniar* [20], and the methods in [18].
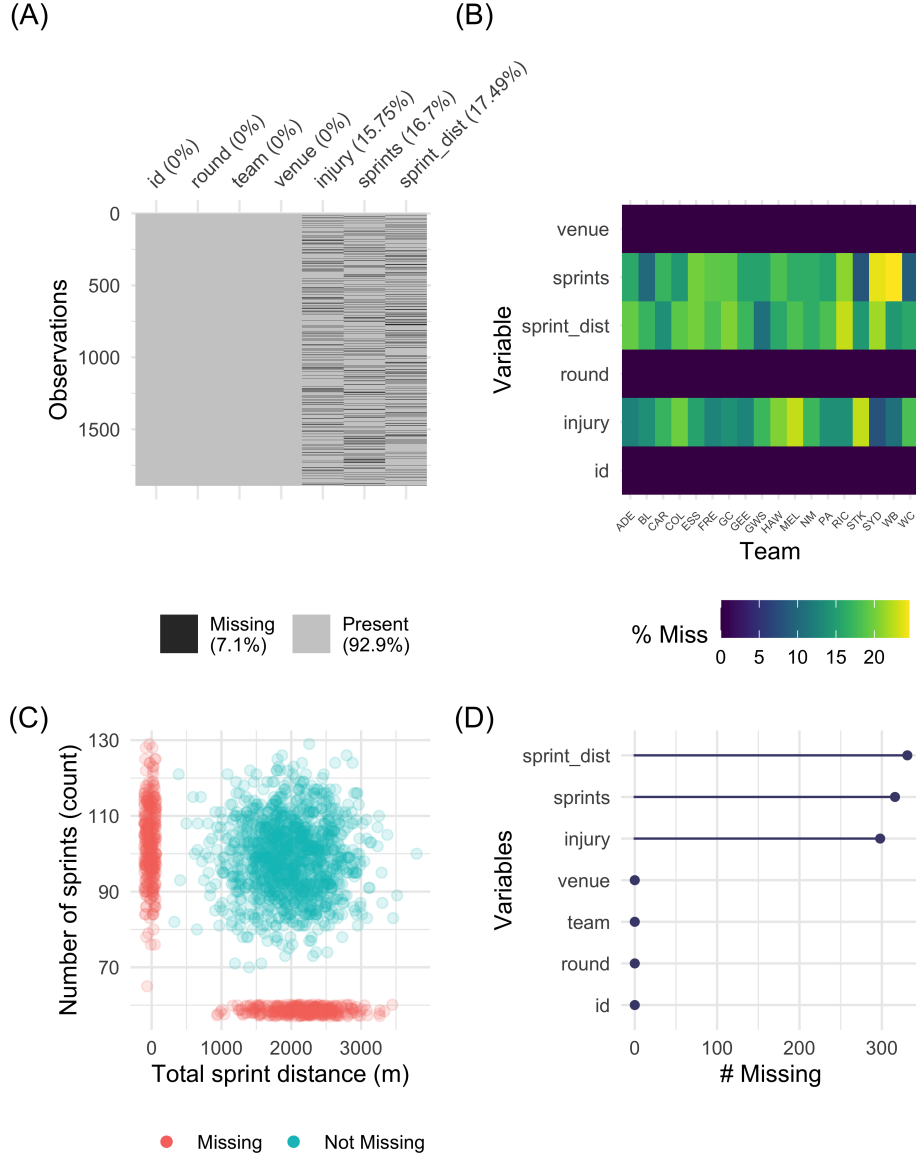
Figure 2: Overviews of missing values in simulated injury data from Australian Football matches played in August, 2020. The dataset contains observed player id, team, and match venue data, in addition to simulated data on injury incidence, and number of sprints per match, and total sprint distance per match. Panel A shows the location of missing and present cases, and the percentage of missing data in each variable (and overall). Panel B shows the percentage of missings for each team, in each variable. Panel C shows the relationship between two variables in the dataset, including missing values (red) which are shifted, so that they can be visualised on the same axes as the regular values. Panel D shows the number of missing values in each variable.

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 (e.g., 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 the response variable (the "dependent variables"). Care should be taken when imputing the response variable, and in general they should not be imputed with a single imputation, as this will bias the subsequent analysis. Methods that provide multiple imputed values for a response variable, such as multiple imputation, or using a Bayesian framework that generates a distribution of posterior values, can be appropriate for imputing outcome variables as they describe the uncertainty (see Chapter 15 in [21]). Missing values in the outcome can be dealt with specially for Randomised Control Trials [17,22].

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 [5,23,24].

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 [23].

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 [18].

Recommendations

The current state of reporting missing data in football research has room for improvement. In this section we recommend 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 authors 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).
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%)*"

2. State if missing data was explored. If there is missingness, provide a summary graphic (see Figure 2)

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

3. 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*"

4. State actions taken in handling missing data.

   Were they removed? Imputed? Another approach?

5. Describe and justify any imputation process used.

6. 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:

> *Twenty-five 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 articles in football research showed that only one in nine studies reported missingness. We speculate that this result is representative of practice in the wider sports science literature. The low reporting of missing data could suggest that this aspect of analysis receives little attention. To address this issue and assist researchers, we have provided recommendations for reporting and exploring missing data. Researchers 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. Physical Medicine and Rehabilitation. 2015;7:990–4.

2. Nakagawa S, Freckleton RP. Missing inaction: The dangers of ignoring missing data. Trends in Ecology and Evolution. 2008;23:592–6.

3. Sterne J a C, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ. 2009;338:b2393.

4. Newman DA. Missing data: Five practical guidelines. Organizational Research Methods. 2014;17:372–411.

5. Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychological Methods. 2002;7:147–77.

6. Tierney NJ, Harden FA, Harden MJ, Mengersen KL. Using decision trees to understand structure in missing data. BMJ Open. 2015;5:e007450.

7. Barnett AG, McElwee P, Nathan A, Burton NW, Turrell G. Identifying patterns of item missing survey

data using latent groups: An observational study. BMJ Open. 2017;7:e017284.

8. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. PLoS Medicine. 2009;6:e1000097.

9. Sainani KL, Borg DN, Caldwell AR, Butson ML, Tenan MS, Vickers AJ, et al. Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. British Journal of Sports Medicine. 2020;1–5.

10. Peugh JL, Enders CK. Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of Educational Research. 2004;74:525–56.

11. Jeličić H, Phelps E, Lerner RM. Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. Developmental Psychology. 2009;45:1195–9.

12. 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. 2019;22:467–71.

13. Young C, Luo W, Gastin P, Tran J, Dwyer D. Modelling match outcome in australian football: Improved accuracy with large databases. International Journal of Computer Science in Sport. 2019;18:80–92.

14. Sampson JA, Murray A, Williams S, Halseth T, Hanisch J, Golden G, et al. Injury risk-workload associations in ncaa american college football. Journal of Science and Medicine in Sport. 2018;21:1215–20.

15. Cresswell SL. Possible early signs of athlete burnout: A prospective study. Journal of Science and Medicine in Sport. 2009;12:393–8.

16. Scheffer J. Dealing with missing data. Research Letters in the Information and Mathematical Sciences. 2002;3:153–60.

17. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clinical trials. 2004;1:368–76.

18. Tierney NJ, Cook DH. Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. arXiv preprint arXiv:180902264. arxiv.org; 2018;

19. Tierney N. Visdat: Visualising whole data frames. The Journal of Open Source Software. 2017;2:355.

20. Tierney N, Cook D, McBain M, Fay C, O'Hara-Wild M, others. Naniar: Data structures, summaries, and visualisations for missing data. R Package version 0.6.0. 2019.

21. McElreath R. Statistical rethinking: A bayesian course with examples in r and stan. 2nd ed. CRC press; 2020.

22. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. BMC Medical Research Methodology. 2017;17:162.

23. Buuren S van. Flexible imputation of missing data. CRC Press; 2012.

24. Cheema JR. A review of missing data handling methods in education research. Review of Educational Research. 2014;84:487–508.

Supplement 1

Missing Data Examples Relevant to Australian Football

One way to describe patterns of missing values is through Rubin's [1] three missing data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). These mechanisms 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. Here, we provide an example of data MAR, MCAR and MNAR, relevant to Australian Football.

Data could be considered MCAR if an Australian Football data collector happened to take a lunch break during the third quarter of a match, resulting in data not being recorded for this time period. Data is MCAR, because the missings are not systematically related to other measured variables. For data to be MCAR, this lunch break related issue would need to occur completely at random, not systematically repeated each game.

Data MAR could arise if players' body mass measurements were taken with a scale that produced more missing values when placed on soft carpet, rather than hard tiles, in the team change room. If surface type was recorded, data would be MAR, because the missingness is related the surface, but not related to the value of body mass itself [2].

Data MNAR is when the missings are related to something unobservable or unobserved. The previous example of data MAR could be data MNAR if the surface type was not recorded in the dataset—it would then be unobserved. Data MNAR also arises when missing values are related to the values missing. This could take place if, for example, the scales used only measure up to 110 kg, and then produce missing values. This is a systematic issue—it happens each time, and means that heavier players are not recorded.

References

1. Rubin DB. Inference and missing data. Biometrika. Oxford University Press; 1976;63:581–92.

2. Buuren S van. Flexible imputation of missing data. CRC Press; 2012.

Supplement 2

Systematic Search Details

The systematic search was performed in accordance with the reporting requirements of the PRISMA statement [1]. The search strategy is described below. The search protocol was not registered, and no funding was provided to support the undertaking of this work.

Objective

We aimed to estimate the proportion of articles that reported missing data in football research.

Method

Football related articles published in 2019 were studied. Three major sports medicine databases were searched on May 20, 2020, using key terms and limits. The databases were: SPORTDiscus, Embase and Cinahl. The search terms were: 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.

Included articles were: (a) published in 2019; (b) in English; (c) had an accessible full-text; 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 sample were not eligible for inclusion (e.g., full-texts not in English, conference abstracts, qualitative studies), an additional 5% random sample was taken. It was anticipated that 30% of the additional 5% sample would also be ineligible. With this loss, it was expected that 10% of the articles that remained after title and abstract screening would undergo survey. Figure 1 summarises the search process.

Articles were surveyed to determine: whether a missing data statement was included, or if missing data were acknowledged in a figure/table; and if not, whether articles shared their data. The articles that included a missing data statement were surveyed to determine whether: (a) the type of missingness was reported; (b) missing data were explored; and (c) if data were imputed. Two authors independently extracted the data. Discrepancies were resolved by DNB. Findings are reported as the proportion and 95% confidence interval (CI). Confidence intervals were calculated using the Clopper-Pearson method for the binomial distribution [2] via the *binom* package [3] in R [4].

Results

The search returned a total of 3,863 articles. Of the 199 articles screened, 136 met the inclusion criteria, representing ~10% of the articles found (Figure 1). The proportion of articles that reported missing data was 11.0% (95% CI; 6.3% to 17.5%). Of the articles that did not mention missing data (121/136), only 7.4% (95%
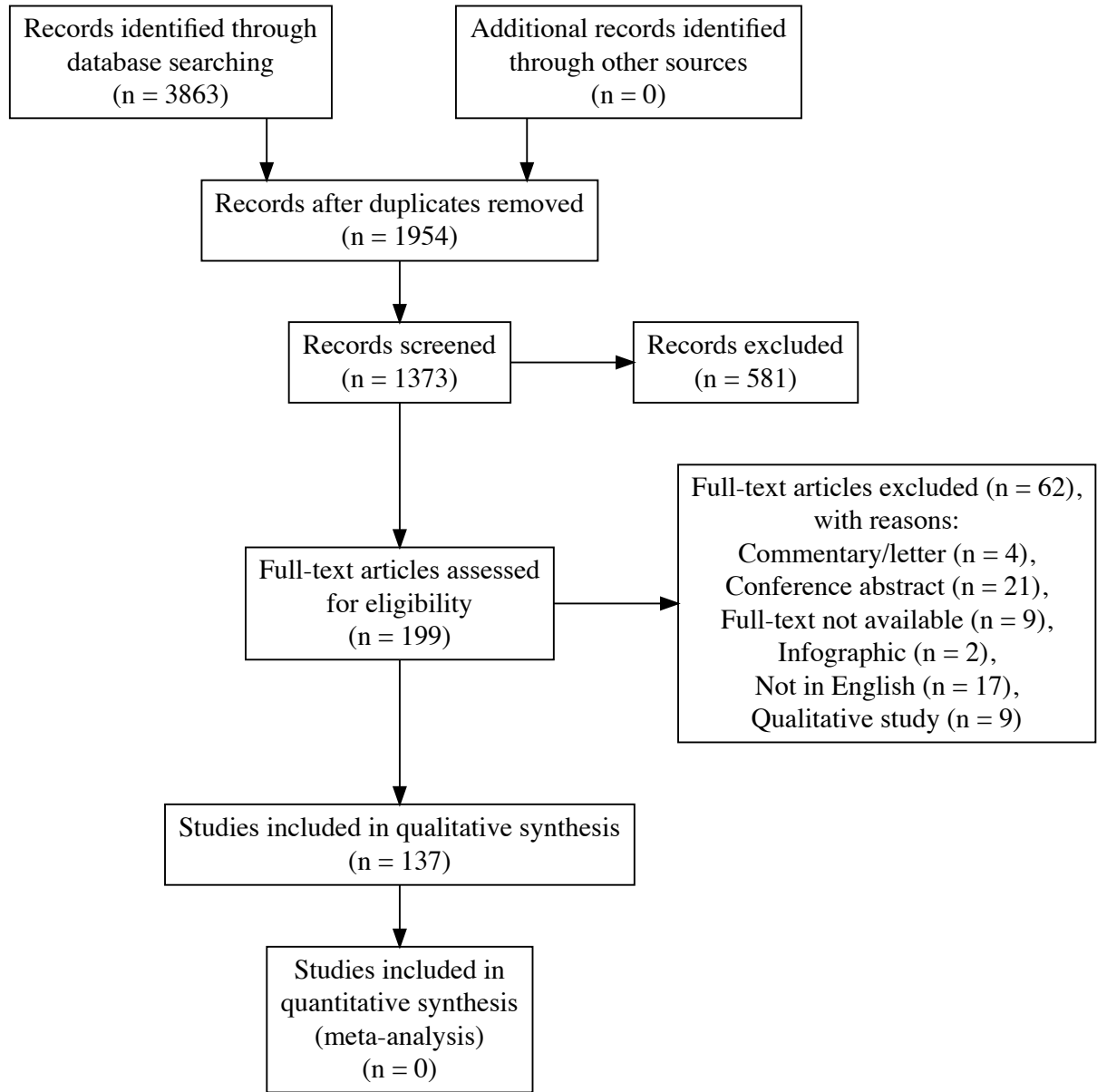
Figure 1: Flowchart of the article search and inclusion for the systematic search. The full-text of 137 articles was reviewed.

Table 1: Reporting of missingness, many papers did not explore missings in a meaningful way, and there was the same number of papers who imputed values as reported MCAR, MAR, or MNAR.

Reporting of Missingness	Mean and 95% CI
Type (i.e., MCAR, MAR or MNAR)	26.7% [7.8% - 55.1%]
Missing data were explored	40.0% [16.3% - 67.7%]
Missing data were imputed	26.7% [7.8% - 55.1%]

CI; 3.5% to 13.7%) 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. There were 15 studies that provided a statement about missing data. A breakdown of the information extracted from these studies is provided in Table 1.

Summary

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

References

1. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Medicine. 2009;6:e1000097.

2. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26:404–13.

3. Dorai-Raj S. Binom: Binomial confidence intervals for several parameterizations. 2014; Available from: https://CRAN.R-project.org/package=binom

4. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020; Available from: https://www.R-project.org/