Challenges with Confidence Intervals for Sport Injury Burden, other Ratio Measures and Clustered data

Supplementary materials:
https://osf.io/shrnj/files/osfstorage
For correspondence:
ian.shrier@mcgill.ca

Ian Shrier[1], Franco M. Impellizzeri[2], Avinash Chandran[3], Sean Williams[4], Joseph W. Shaw[5], and Russell J. Steele[6]

[1] Centre for Clinical Epidemiology, Lady Davis Institute, McGill University, [2] Faculty of Health, School of Sport, Exercise and Rehabilitation, University of Technology Sydney, 3 Datalys Center for Sports Injury, Research and Prevention, 6151 Central Ave., Indianapolis, IN 46220, 4 Centre for Health, Injury and Illness Prevention in Sport, University of Bath, Bath, UK, 5 Faculty of Sport, Technology and Health Sciences, St. Mary's University Twickenham, London, UK; High Performance Unit, Irish Rugby Football

All authors have read and approved this version of the manuscript. This article was last modified on Month XX, YEAR.

Authors X @*xtwitter* and Y @*ytwitter* can be reached on Twitter.

Union, Dublin, Ireland, 6 Department of Mathematics and Statistics, McGill University

Please cite as: Shrier, I et al. (2025). Re-establishing confidence in Confidence Intervals: An evaluation of recent practices. *SportRχiv*.

ABSTRACT

Several studies investigating injury burden have used "standard" formulae for injury rates when calculating their 95% confidence intervals. However, this may have led to artificially narrow confidence intervals because the authors' calculations did not account for (1) violations of important underlying assumptions, and (2) the same athlete having multiple injuries. Although previous authors have recommended appropriate methods such as bootstrapping to solve the first challenge, there is little guidance for sport medicine researchers on how to implement bootstrapping given the complexity of data in our field. The purposes of this article are (1) to illustrate when the "standard" formulae for injury rate confidence intervals can be used in sport medicine research, (2) why the standard formulae for injury rate confidence intervals are inappropriate when estimating injury burden and (3) provide more detailed instructions on how to use bootstrapping for confidence intervals in the context of any sport medicine study that includes repeated measures.

Manuscript

In 2005, Vickers wrote "A mistake in the operating room can threaten the life of one patient; a mistake in statistical analysis or interpretation can lead to hundreds of early deaths. So it is perhaps odd that, while we allow a doctor to conduct surgery only after years of training, we give SPSS® (SPSS, Chicago, IL) to almost anyone."[1]

When we make inferences and recommendations based on statistical analyses, we must also include a measure of uncertainty.[2] One common measure is the confidence interval (CI).[3 4] There are several excellent articles and books explaining what CI are, and what they are not.[4 5] In brief, CI measure the uncertainty about our results due to random sampling of participants from a larger population of interest, which is referred to as imprecision. Additional uncertainty occurs in real-world studies with issues that might affect validity. If CI are to be helpful, researchers need to calculate them using appropriate methods or readers may believe the results are more (or less) precise than they really are.

Williams et al. recently highlighted that many authors were incorrectly implementing a formula for the calculation of CI for "injury burden" in sport medicine research (i.e. mean days

lost per 1000 h).[6] In the Williams et al. example with 20 injuries and injury burden of 300 days/1000 hours, one inappropriate method suggested a 95% CI of 277 to 325, whereas one appropriate method suggested the true 95% CI was as large as 106 to 545.[6] This suggests that researchers often apply methods without fully understanding their suitability for the context. In the Appendix, we briefly explain why those implementations were incorrect, and provide greater detail with more appropriate alternatives in the Supplementary Material.

For a typical operationalisation of injury burden (injury rate x mean days lost per injury), Williams et al. proposed "bootstrapping" as a method to properly calculate CI[6]. Further, they briefly noted in the editorial that accounting for "clustering" (e.g. multiple injuries to the same participant) in regression models requires further considerations. Clustering occurs when there is data dependency, which simply means there are correlations between the individual observations within the data. For example, we expect some athletes are more prone to injury than others creating data dependency when estimating the overall injury rate. Similarly, data dependency occurs when estimating the mean time-loss because we expect (1) some athletes heal more quickly than others, and (2) health care professionals from different teams have different skills and strategies for prevention and treatment and make different return-to-play decisions.[7] Williams et al.[6] have been cited several times where clustering was present, but appropriate bootstrap methods to account for this were not described or followed. Therefore, the 95% CI were likely too narrow, leading to overconfidence about the precision of the estimated injury rate, injury burden or time-loss.

The purpose of this article is to help sport medicine researchers avoid repeating what appear to be common errors in calculating CI. As an example, consider the objective is to estimate the mean injury burden for a particular sport and we have two hypothetical studies, each with 100 observations. In one study, the authors measured one hundred athletes with one injury each. In the other study, the authors measured 10 athletes, with 10 injuries each. Clearly, we would be more confident in our injury burden estimate for the sport in the study with 100 athletes compared to 10 athletes. This is because we expect the injury burden for each of the 10 injuries per athlete to be more closely correlated (clustered together, creating an intra-class correlation) than the injury burden between the 10 different athletes. Therefore, even though both studies have 100 observations, we have more independent information in the study where values are not correlated with one another (i.e., the study on 100 athletes) compared to the study where values are correlated with one another (i.e., the study measuring 10 athletes, with 10 injuries each). When we have clustered data, we must properly account for the reduced amount of independent information due to the dependence, or we risk making

inaccurate conclusions and invalid recommendations. The same principles apply to other outcomes about which we are making inferences from the results (e.g., location of injuries, types of injuries, injury rates). In the next sections, we describe how to implement some of the methods falling under the term "bootstrapping".

General Principles of Bootstrapping

"Bootstrapping" refers to a group of (parametric and non-parametric) methods used for different purposes, one of which is to determine appropriate CIs.[8] A significant benefit of bootstrapping is that it enables the calculation of robust CI without making assumptions about the underlying data distributions. This makes bootstrapping particularly useful when dealing with complex or non-normally distributed data, which are common in injury count data. However, each bootstrap method (including the Bayesian bootstrap) must respect the data structure (e.g., research question), data dependency (e.g. clustering), and assumptions about the data distribution (e.g., heteroskedasticity). Therefore, the method is not a panacea,[9] and statistical expertise is necessary to choose the most appropriate method given the research context.[8]

In 2009, Shrier et al.[10] explained how to implement the relatively simple non-parametric percentile method for bootstrapping. In brief, consider an original data set of 100 rows of non-clustered observations where each row is one non-team sport athlete with one injury. In bootstrapping, we sample one of the rows and place it in a new data set. We then return and resample a second row of the original data set and add it to our new data set. We repeat this process for each row of data, which is 100 times for our data. The key to bootstrapping is that each time we resample the data, it is from the entire original 100 rows of data. This is called resampling with replacement. Because we resample with replacement, our new data set is very likely to contain duplicates of some rows from the original data, and to be missing other rows. Therefore, when we calculate the mean injury burden for the new data set, the number will be different from the original data set. We now repeat the process many times (we will use n=1000 here) to generate new data sets. For each of the new data sets, we calculate the injury burden by whatever analysis is most appropriate. In the percentile bootstrap method explained by Shrier et al. [10], we sort our 1000 estimates of injury burden in ascending order. The lower 95% CI would be the 2.5 percentile (i.e. the 25th ranked value of our 1000 estimates) and the upper 95% CI would be the 97.5 percentile (i.e. the 975th ranked value of our 1000 estimates).

In the original bootstrapping R code provided by Williams et al[6] to obtain CI for injury burden, the CI will be correct if each athlete is injured once, and provided other assumptions about data structure, data dependency and data distributions hold. Note that their R code does not provide the methods to check these assumptions and authors will likely need to collaborate with appropriate statisticians. In the next section, we describe how the same principles are followed to calculate CI and account for the common issue of repeated injuries on the same athletes or other forms of clustering.

Clustering on one variable

In this section, we consider the data to be clustered by athlete only, where some athletes will have more than one injury. Consider a study where there were 20 athletes without any injury, and 80 athletes with a total of 100 injuries. For each athlete, we have time to injury when injured, time to end of study following the last injury or if never injured, and the number of days lost due to injury.

When we need to account for clustering with bootstrapping, we cannot simply bootstrap on the raw data above or we would replicate the clustering problem. Rather, we must bootstrap on the clustered variable. In our example, the athlete is the clustered variable and we have two methods for bootstrapping. The first method is simpler and appropriate for some questions related to injury rate and injury burden. The second method is more general and could be used to answer a much wider range of questions.

In the first method, we collapse all rows for an individual athlete so there is one row per athlete, with columns representing their (1) total number of injuries, (2) total time of exposure in the study, (3) total number of days lost across all injuries, and (4) a calculated mean injury burden across all injuries for each athlete. We now bootstrap the data the same as we did with non-clustered data above, since we have accounted for clustering by collapsing all the information for one athlete into one row. Our new "starting data set" has 100 rows of athletes with injury burden calculated for each one (which includes 0 for rows where no injury occurred). Our bootstrap data sets will have 100 rows of athletes, where some athletes are duplicated and some are omitted. We generate 1000 bootstrap data sets. For each bootstrap, we can calculate mean injury rate, mean injury burden using total number of days lost and total exposures, or mean injury burden from the injury burden calculated from each athlete. The advantages and disadvantages of these two calculations for injury burden (which may yield different results because the underlying assumptions are different) is beyond the scope of this article. Finally, using the numerical results for injury rate or injury burden from each of the

1000 bootstraps, we then find the 2.5 and 97.5 percentiles to obtain our 95% confidence intervals. The R code provided by Williams et al has now been updated to provide examples for clustered data.[11]

The above method will provide appropriate CI for injury rate and injury burden. However, we lose information about individual injuries (e.g. time loss, time-to-injury, body part, etc.) because we collapsed the data across all injuries. What if we are interested in questions that require us to calculate rate ratios (or rate differences) for weighted means or weighted medians, or recurrent injuries, or injury rates from regression models as a function of some exposure of interest (e.g. type of injury)?[10] To address these questions, we need to know specific details about each injury and cannot collapse the data as we did above. An alternative implementation of bootstrapping is to create a data set that includes only the athlete identifier. In our example, this would represent 100 athletes. We bootstrap on the athlete identifier to create a new data set of 100 athlete identifiers that has some duplicates and omissions of the original athlete identifiers. We then extract and merge all of the injury observations from the original data set that correspond to these athlete identifiers to create a new data set. Since some athletes will have 1 injury, and others will have 2 or more injuries, these different bootstrap data sets will have different numbers of injuries, which means they will have different numbers of rows or observations. By preserving the same structure as the original data set, this bootstrap implementation allows one to account for clustering when answering questions that go beyond injury rate and burden, and that could have been answered by the original data set had there not been any clustering.

Clustering on more than 1 variable

As above, clustering may occur at several levels. For example, injury burden may be clustered by team in addition to athlete. The approach in this case depends on several factors, the most important being the research question.

If we are interested in the "average injury burden for each team", we would bootstrap only on the team and not the individual. In fact, bootstrapping on the individual may give the wrong answer. For example, the injury rate expressed per game (and hence injury burden per game) will be different for players who play a lot of the game compared to those who play less. Since our question is about team injury burden, we must take this into account. Because bootstrapping on the individual ignores this, it will likely provide an invalid estimate.

Consider a different example where one wants to know the average injury burden across all participants in a league, and we expect clustering by both athlete and team. This

hierarchical clustering can theoretically be accommodated by bootstrapping on both athlete and team. This is more complicated than it seems. There are numerous assumptions that have to be verified before bootstrapping should be used, and this should not be done without the help of experienced statisticians.[12] [13] Although the required assumptions and further details are beyond the scope of this paper, the principles are the same. In brief, we set up our data so it only includes both the athlete and team identifiers. We first bootstrap on the team, and then later bootstrap on the athletes within a team.[12]

An alternative to the nonparametric procedure presented in the previous section is a parametric bootstrap method that uses the residuals from a statistical model, rather than the individual observations.[14]As with the methods described above, our objective is to perturb the data to properly account for the additional uncertainty due to clustering (e.g., repeated measures on the same athlete). However, in complex statistical models, obtaining the correct level of error to add can be difficult.[15] One proposed method is to bootstrap on the residuals from the model. In brief, the difference between each individual's predicted value from the statistical model and their observed value is known as a "residual". When we expect the errors (residuals) to be randomly distributed, we can increase our imprecision by taking each observed value and adding a residual that is resampled from the full distribution of residuals. With this bootstrap method, athletes are expected to have different residuals in each of the data sets, which will yield different results when we conduct our analyses on each of the bootstrapped data sets. As before, we can bootstrap this process to create 1000 values and take the appropriate percentile values to obtain our 95% CI. Unlike the non-parametric resampling with replacement method, this method requires that the statistical model is correct. If the model is not correct, the bootstrap values will not be correct. While this might be considered an "unnecessary" assumption, the assumption is already part of the statistical model that is used to calculate the point estimate. If the model is incorrect for the CI, it is also incorrect for the point estimate.

Alternative Methods

One alternative to bootstrapping that is easier to implement is "jackknifing".[13] Historically, jackknifing required much less computational time and could be done easily by hand. With improvements in computational processing speeds, this is no longer an issue. Like bootstrapping, the jackknife procedure can be used to create new data sets from the original data set. However, instead of resampling with replacement from the original collapsed data set (one row for each of the 100 athletes), each data set is created by simply deleting one

observation (leave-one-out). The jackknife process is considered more appropriate when there are few observations.[16] For example, if there are only 5 participants, bootstrapping may sometimes sample the same individual five times. Therefore, bootstrapping is more likely to produce inappropriately narrow confidence intervals compared to jackknifing. Readers should understand that the definition of "few observations" is context specific. It can occur with large numbers of participants if the statistical model includes several covariates. As a rule-of-thumb (for which there will be exceptions), "few" can be defined as <10 participants for each of the cells created by the statistical model (e.g., a model with a 2-level covariate and 3-level covariate has 6 unique combinations of cells).[16-18] Alternatively, researchers might be interested in the injury burden per team. If there are only 20 teams in a league and researchers want to include two covariates in the statistical model, bootstrapping may not be appropriate.

Williams et al also mentioned the delta method[19] to calculate confidence intervals for injury burden.[6] The delta method is a generic method to calculate variances for functions of one or multiple random variables (e.g., a ratio of two sample means) and therefore can be used to calculate CI. We provide more detail in the Supplementary Material. It is often used for variables that are ratios, which would be applicable to injury burden. In brief, the variance of a ratio is not simply the variance of the numerator divided by the variance of the denominator. It must account for the correlation between these two variables to avoid assuming we have more information than we do. Therefore, the delta method is sometimes necessary to calculate appropriate CI even when there is no clustering. If there is clustering, one would use bootstrapping methods to generate the appropriate 1000 data sets, and then use the delta method on each of the data sets to calculate the appropriate CI.

In summary, CI are essential to help us understand the precision of our estimates if we want to make inferences and recommendations from our study results.[3] Although some formulae appear simple, researchers need to properly implement the formulae or bootstrap methods to avoid misleading readers. This article is only an introductory overview to bootstrapping for calculating CIs. While applied researchers can independently perform basic calculations in straightforward scenarios, we strongly recommend that researchers collaborate with qualified and experienced statisticians to avoid the errors that are occurring in the sport medicine literature. Failure to do so may lead to more articles in statistical journals showing deficiencies in our field,[20] and to inappropriate conclusions that could lead to sub-optimal decisions and recommendations about prevention or treatment programs.

Contributions

Contributed to conception, and writing: All authors
Drafted and/or revised the article: All authors
Simulations and programming code: JS and IS

Acknowledgements

None

Funding information

None

Data and Supplementary Material Accessibility

The simulations and programming code for the simulations and analyses are available on OSF through the link in the manuscript.

REFERENCES

[1] A. Vickers. Interpreting data from randomized trials: the Scandinavian prostatectomy study illustrates two common errors. *Nat Clin Pract Urol* 2005;2(9):404-5. doi:doi: 10.1038/ncpuro0294

[2] S. Senn. Error point: The importance of knowing how much you don't know. In: Mayo DG, editor. *Error statistics philosophy*, 2020.

[3] J.P. Vandenbroucke, E. von Elm, D.G. Altman et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *PLoS Medicine* 2007;4(10):e297. doi:doi: 10.1371/journal.pmed.0040297

[4] S. Greenland, S.J. Senn, K.J. Rothman et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016;31(4):337-50. doi:doi: 10.1007/s10654-016-0149-3

[5] D. Altman, D. Machin, T. Bryant et al. Statistics with confidence. 2nd ed: John Wiley & Sons, 2006. pages.

[6] S. Williams, J.W. Shaw, C. Emery et al. Adding confidence to our injury burden estimates: is bootstrapping the solution? *British journal of sports medicine* 2024;58(2):57-58. https://doi.org/10.1136/bjsports-2023-107496

[7] I. Shrier. The strategic assessment of risk and risk tolerance (StARRT) framework for return to play decision making. *British Journal of Sports Medicine* 2015;49:1311-1315. https://doi.org/10.1136/bjsports-2014-094569

[8] G.T. LaFlair, J. Egbert, L. Plonsky. A practical guide to bootstrapping descriptive statistics, correlations, t tests, and ANOVAs. *Advancing quantitative methods in second language research*: Routledge, 2015:46-77.

[9] D.B. Rubin. The bayesian bootstrap. *The annals of statistics* 1981:130-134. https://www.jstor.org/stable/2240875

- [10] I. Shrier, R.J. Steele, B. Rich et al. Analyses of injury count data: some do's and some don'ts. *Am J Epidemiol* 2009;170:1307-1315. DOI: 10.1093/aje/kwp265

[11] S. WilliamsJ.W. Shaw. Bootstrapping Burden CIs. https://osf.io/shrnj/files/osfstorage . Last accessed: 2025-03-13

[12] B. EfronR. Tibshirani. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Stat Sci* 1986;1(1):54-75.

[13] B. Efron. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* 1981;68(3):589-599.

[14] H.T. Thai, F. Mentré, N.H. Holford et al. A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. Pharmaceutical Statistics 2013;12(3):129-140. https://doi.org/10.1002/pst.1561

[15] D. Altman, S.M. Gore, M.J. Gardner et al. Statistical guidelines for contributors to medical journals. In: Altman D, Machin D, Bryant T, Gardner M, editors. *Statistics with confidence*. 2nd ed: John Wiley & Sons, 2006:171-190.

[16] A. Severiano, J.A. Carrico, D.A. Robinson et al. Evaluation of jackknife and bootstrap for defining confidence intervals for pairwise agreement measures. *PLoS One* 2011;6(5):e19539. doi:10.1371/journal.pone.0019539

[17] C.-F.J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann Stat* 1986;14(4):1261-1295. DOI: 10.1214/aos/1176350142

[18] R.J. TibshiraniB. Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability* 1993;57(1):1-436.

[19] K. Knight. Mathematical statistics. Unites States: Chapman & Hall/CRC, 1999. pages.

[20] K.L. Sainani, D.N. Borg, A.R. Caldwell et al. Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. *British Journal of Sports Medicine* 2021;55(2):118-122. https://doi.org/10.1136/bjsports-2020-102607