

The Practical Value of Bayesian Inference in Describing the Epidemiology of Sports Injuries

Avinash Chandran, PhD, MS, FACSM¹

Ben Clarsen, PhD²

Ian Varley, PhD³

Karim Khan, PhD, MBBS^{4, 5}

Stephen W Marshall, PhD⁶

Roald Bahr, PhD^{7, 8}

¹Datalys Center for Sports Injury Research and Prevention, Indianapolis, IN, USA

²FIFA Medical, Fédération Internationale de Football Association, Zurich, Switzerland

³Nottingham Trent University, Nottingham, UK

⁴Department of Family Practice, Faculty of Medicine, The University of British Columbia, Vancouver, BC, Canada

⁵School of Kinesiology, Faculty of Medicine, The University of British Columbia, Vancouver, BC, Canada

⁶Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

⁷Aspetar Orthopaedic and Sports Medicine Hospital, FIFA Medical Centre of Excellence, Doha, Qatar

⁸Oslo Sports Trauma Research Center, Department of Sports Medicine, Norwegian School of Sports Sciences, Oslo, Norway

Corresponding author:

Avinash Chandran, PhD, MS, FACSM

Bayesian injury rates in sport injury epidemiology

Chief Science Officer | Managing Director, NCAA Injury Surveillance Program

Datalys Center for Sports Injury Research and Prevention

Indianapolis, IN 46220

avinashc@datalyscenter.org

Manuscript Type: Leading Article

The Practical Value of Bayesian Inference in Describing the Epidemiology of Sports Injuries

Abstract

Sports injury surveillance programs have been vital in advancing the understanding of injury epidemiology across various athlete populations. Surveillance-based epidemiological measures of injury occurrence are ubiquitous in the sports medicine literature, and the injury rate is one such commonly used measure. Traditional approaches to calculating injury rates have predominantly relied on frequentist methods, which, while informative, have limitations in addressing certain practical questions. We explore an alternative Bayesian framework for analyzing injury rates, highlighting its potential to enhance sports medicine practice. We delineate the practical implications of adopting a Bayesian approach, contrasting key analytical outputs such as credible intervals with their frequentist counterparts. Through simulated and real-world examples, we demonstrate the types of analyses and inferences that are only possible within this framework. We particularly discuss how Bayesian methods allow for direct calculation of probabilities for specific outcomes and provide intuitive interpretations of uncertainty. We discuss the computational and inferential advantages of the Bayesian approach, illustrating how it can offer more nuanced insights into injury incidence in sport injury epidemiology.

Key Points:

- Traditional methods for calculating injury rates in sports medicine have predominantly relied on frequentist approaches, which have limitations in addressing certain practical questions about future outcomes.
- Bayesian approaches to calculating injury rates enable probabilistic inferences and intuitive interpretations of uncertainty, allowing for direct answers to questions such as

Bayesian injury rates in sport injury epidemiology

the probability of exceeding specific injury thresholds or expected injury counts in future scenarios.

- Bayesian methods offer analytical advantages in sports injury epidemiology, providing a complementary framework to frequentist approaches and offering more nuanced insights into injury occurrence and risk assessment.

Background

Sports injury surveillance programs have proven invaluable for understanding the epidemiology of injuries in various athlete populations.[1–3] Calculations of injury occurrence (such as the injury rate) based on data captured through sports injury surveillance have been a cornerstone in sports medicine literature and have been used to inform policy decisions related to athlete health and welfare.[3–6] Such calculations typically involve a systematic data collection effort, followed by the application of established formulae to obtain estimates based on observed data (e.g., when injuries occurring during competitive events are recorded and divided by aggregated competition exposure time measurements to calculate competition-specific injury rates).[2,7] While this approach is straightforward, the nuances and limitations of traditional injury rate and associated interval calculations are often overlooked. We suggest that this can profoundly influence the inferences drawn from the results.

In diving further into this topic, it is first important to recognize that an *injury rate* is not directly observable; *injuries* are observed over some period of follow-up in a sample, and the *rate* of injury is inferred upon by a calculation. The injury rate calculation commonly achieved by dividing total injuries by total exposure time represents an average injury rate. However, the average injury rate is a summarized statistical reflection of an unobserved, and potentially variable, incidence process. The calculation from observed data produces an average injury rate but offers little insight into how the underlying process may have varied across the exposure period- because this process is not observable and cannot be reconstructed deterministically.

Current Practices in Injury Epidemiology

To better appreciate why we may reconfigure how commonly used epidemiological measures are obtained, let's consider a one-week football training camp involving multiple teams.

Suppose that 12 injuries are observed over 1,000 training minutes during this camp (for ease of

Bayesian injury rates in sport injury epidemiology

interpretation, we may consider these team-level exposures so training time may represent 1,000 team-minutes). The injury rate may be calculated as 12 per 1,000 minutes, and a measure of precision such as an interval around this estimate may even be obtained using established methods (i.e., a 95% confidence interval for this rate may be estimated to range from 5.2 to 18.8). Calculating the (average) injury rate by summing all injuries and exposures at the end of the camp provides a concise summary of the observed data, and there are well-established methods for calculating the injury rate and an associated interval in this context. Note that this traditional approach to calculating the injury rate would treat the rate as a *fixed* summary measure (neglecting the underlying incidence process), derived using a composite of injury occurrence and exposure time. Conceptually, this assumes that any uncertainty about the calculation (an unobserved quantity) is solely due to random sampling variability, rather than a lack of knowledge about the underlying process which resulted in the 12 observed injuries. As a result, while it may seem (and is often misunderstood) that the confidence interval noted above (5.2-18.8) represents a range of values that this rate may take, this is not the case. Instead, the appropriate interpretation of this interval would indicate that if data collection and analysis were repeated similarly across 100 similar camps, 95 intervals constructed in the same manner would be expected to contain the true training injury rate for this population. We contend that a formulaic, fixed summary measure and its accompanying interval offer limited practical value to the stakeholders we commonly work with, such as coaches and administrative staff.[8]

Communicating injury data to coaches and administrators often involves addressing practical questions about what they might expect during certain scenarios. For instance, after the training camp, it would be reasonable for coaching or administrative staff to ask, "We are planning another training session with this group next month - how many injuries might we expect in a typical session lasting roughly 90 minutes?" A common reaction to this question might begin by extrapolating the observed injury rate per 1,000 minutes to this shorter time frame. That is, if 12

Bayesian injury rates in sport injury epidemiology

injuries were observed over 1,000 minutes, it would not be unreasonable to calculate the expected number of injuries per 90 minutes as: $12/1000 \times 90$, yielding ~1 injury per session. This provides a straightforward, re-standardized rate per 90 minutes, but the rate remains anchored on the previously observed injury rate of 12 per 1,000 minutes. This does not answer the question of how many injuries might occur in a future 90-minute session. The coaches are not simply interested in what was observed in the past; they are interested in what might occur during a future session. To match the nuance of this question, any response should acknowledge the uncertainties involved in attempting to make such predictions. Yet, by treating the observed rate of 12 per 1,000 minutes as a fixed, true value, the fact that injury rates are unobservable, variable, and that our understanding of them is incomplete, is overlooked. For example, in any 90-minute period, there may be no injuries recorded at all. However, when using the fixed-point calculation of 12 per 1,000 minutes, this possibility is never fully acknowledged. Instead, the fixed rate implies certainty in the expected outcome, which does not align with real-world variability.

Reconfiguring the Injury Rate and the Practical Implications for Commonly Reported Outputs

An alternative approach to calculating an injury rate would be to start with an intuition about what the injury rate might be and refine those expectations on the basis of observed data under the premise that the injury rate is inherently variable. Approaching calculations in this manner reflects the basic premise of Bayesian inference. The difference between Bayesian inference and the classical (frequentist) approach is that in frequentist analysis, values such as an injury rate would be considered fixed whereas in Bayesian analysis, it may be considered to be variable, and calculations may be re-positioned according to the combined insights from both intuition and observed data.[9,10] Importantly, Bayesian thinking does not deny the existence of a true value, rather it acknowledges that there is uncertainty in what we know about it in our

Bayesian injury rates in sport injury epidemiology

calculations. By viewing the injury rate as inherently unobserved and variable, it becomes possible to calculate and communicate this measure differently.[11] This approach allows for a natural characterization of variability and helps answer questions such as the expected number of injuries over a training session, with a plausible range of injuries that might occur during a 90-minute period. Returning to the hypothetical training camp data from above, rather than reporting a single expected value of ~ 1 injury per session, this reconceptualization could yield results similar to those presented in Figure 1, where the range of plausible outcomes are clearly defined.

If the injury rate is viewed as an variable measure, a Bayesian approach to calculating it by incorporating both prior beliefs about it as well as observed data allows for a more natural characterization of its precision as well.[11] Sparing the mathematical details, it is straightforward to see how intuitions about the injury rate can place reasonable constraints on it. For example, injury rates cannot be lower than 0 and may reasonably span a wide range of positive values. There may also be some ideas of what the population injury rate would be, based on existing research. In this approach, the injury rate is considered to be an underlying driver of what is in fact observed — injuries across a period of time within a set sample or population. By taking this Bayesian approach, it is also possible to obtain a distribution of potential injury rate values (Figure 2). In addition to reconceptualizing the injury rate as a driver of observed injury counts, this approach allows for direct interpretation of intervals in terms of the rate itself. For instance, the distribution of injury rate values shown in Figure 2 represents the actual range of possible values for the injury rate according to the Bayesian analysis. Moreover, because of how the Bayesian approach is implemented, it can accommodate more nuanced and probabilistic questions about the injury rate — which can be useful to both researchers and clinicians.

Further Possibilities with a Re-configured Injury rate

Bayesian injury rates in sport injury epidemiology

As described above, the Bayesian perspective allows us to obtain distributions of parameters such as the injury rate, enabling both summary measures (such as the mean), and predictive capabilities.[8] The distributions illustrated in Figures 1, and 2 demonstrate the range of possible outcomes for injury counts, and for the injury rate, according to Bayesian analysis. This highlights why interval estimation within the Bayesian framework not only represents a natural way of thinking but also aligns more closely with how clinicians, for instance, balance event probabilities in their practice. These distributions depict the range of values that injury counts, or the true rate can take. This can be configured to produce credible intervals at a desired level of certainty (e.g., 90%, 95%), analogous to classical confidence intervals. Moreover, such distributions also offer answers to more nuanced questions, such as the probability that the true rate exceeds a given value. For instance, Bayesian analysis allows us to determine the probability that a parameter of interest is less than or greater than a particular value.

To illustrate this concept, let's return to the hypothetical training camp example. Concerned about the number of injuries observed in the last camp, the administrators now approach you for predictions regarding the likelihood of observing more than a specific number of injuries — say, 15 — during a similar 1,000-minute camp (again, for simplicity, this may be considered to be team-level). Bayesian inference offers a way to address this question using the same framework and model used to produce the results described above. Figure 3 demonstrates how this probability can be estimated directly, yielding a Bayesian-derived probability of ~0.37. As a result, we can inform the administrators that there is approximately a 37% probability of observing more than 15 injuries during a camp of similar duration. This analytical framework can be extended to evaluate various scenarios across the spectrum of possible outcomes, as illustrated in Figure 3. For instance, our analyses indicate that the probability of observing more than 5 injuries during a similar period is quite high (approximately 0.93), while the probability of observing more than 30 injuries is quite low (approximately 0.04). These estimated probabilities

Bayesian injury rates in sport injury epidemiology

align with both the available data as well as basic intuitions regarding injury rates in this context. This example highlights how Bayesian thinking expands inferential possibilities and offers practical value, particularly when consumers seek to assess injury likelihoods in real-world scenarios.

Conclusions and Bottom-line

The purpose of this article is not to promote the use of one analytical approach over another. Instead, we aim to highlight that the same dataset can yield different inferences depending on the analytical approach taken. For instance, a recent paper describing injury incidence during the 2022 FIFA World Cup in Qatar reported an overall injury event rate of 5.6 per 1,000 hours of total exposure (81 time loss injuries recorded over 14,362 exposure hours).[6] Now imagine the value of being able to make probabilistic statements on the basis of such data. For example, using the analytical framework described above, we may estimate that the probability of observing over 100 injuries during a comparable tournament would be ~ 0.32 , while the probability of observing over 120 injuries would be ~ 0.16 . We acknowledge that these represent basic estimations obtained using simple models which may be refined by incorporating additional contextual factors. However, this example illustrates how a more nuanced approach to framing questions and conducting analyses can yield meaningful and practical insights.

Ultimately, our intention with this work is to emphasize the importance of considering how data-generation processes are conceptualized in sports injury epidemiology. While Bayesian inference offers advantages as we discuss here, it has its limitations and requires careful consideration during implementation. For instance, subjective decisions are often involved in the analytical process, and Bayesian analysis can be computationally intensive.[12] Nonetheless, we sought to contrast the traditional approach with the Bayesian perspective in conducting epidemiological analyses and deriving known parameters. As the field of sports medicine

Bayesian injury rates in sport injury epidemiology

continues to expand, it is crucial to maximize the utility of collected data through nuanced techniques that allow clinicians and other consumers to extract actionable insights for practice.

Acknowledgements: Authors acknowledge the contributions of physiotherapists, athletic trainers, and other sports medicine staff in the successful execution of epidemiological efforts such as injury surveillance. Authors also acknowledge the use of OpenAI's GPT-4/GPT-4o for code troubleshooting and language editing during the preparation of this manuscript.

Ethics Declarations

Funding: No funding was received for the preparation or publication of this article.

Conflicts of Interest Statement: Author AC acknowledges funding from the National Collegiate Athletic Association as the Managing Director of the NCAA Injury Surveillance Program. The Datalys Center is an independent nonprofit organization that manages the operations of the NCAA ISP. The content of this report is solely the responsibility of the authors and does not necessarily represent the official views of any funding organization. The remaining authors do not have relevant conflicts of interest to declare.

Availability of Data and Material: This article does not utilize real-world raw data and relies on simulations and synthetic data for illustrative purposes. As such, there are no datasets available for sharing.

Ethics Approval: Not applicable.

Consent to Participate: Not applicable.

Consent for Publication: Not applicable.

Code Availability: Not applicable.

Author Contributions: Authors AC and BC contributed to article conception. AC led writing efforts and all simulations; SWM collaborated with AC on simulations and analyses. Authors BC, IV, KK, SWM, and RB reviewed the manuscript and technically contributed to the quality of the manuscript. All authors read and approved the final version of the manuscript.

Figure Legends

Figure 1. Distribution of predicted injuries per 90 minutes using Bayesian inference. The histogram reflects a range of possible injury counts, with a dashed vertical line indicating the observed injuries per 90 minutes calculated from the data (scaled from total injuries and exposure). The dot beneath the plot represents the posterior mean expected injury count from the Bayesian analysis, while the horizontal bar depicts a plausible range (95% credible interval) of expected injury counts over a 90-minute period.

Figure 2. Plot illustrates the distribution of estimated injury rates as a density plot, considering the injury rate as being variable. The observed injury rate is marked by a vertical dashed line within this distribution. This density plot highlights the range and uncertainty of injury rate estimates, integrating prior intuitions about the rate with observed data to provide a probabilistic view of injury incidence.

Figure 3. Plots represent the predictive distribution of injury counts based on Bayesian modeling. The vertical dashed line mark various thresholds (of 5, 10, 15, 30), and the shaded regions to the right of the thresholds indicate the respective probabilities of exceeding these values. The predictive approach provides probability estimations, integrating observed data and prior information, offering a nuanced perspective on the likelihood of injury counts in similar scenarios.

References

1. Ekstrand J, Waldén M, Häggglund M. Hamstring injuries have increased by 4% annually in men's professional football, since 2001: A 13-year longitudinal analysis of the UEFA Elite Club injury study. *Br J Sports Med*. 2016;50:731–7.
2. Chandran A, Boltz AJ, Morris SN, Robison HJ, Nedimyer AK, Collins CL, et al. Epidemiology of Concussions in National Collegiate Athletic Association (NCAA) Sports: 2014/15-2018/19. *Am J Sports Med*. 2022;50:526–36.
3. Kerr ZY, Putukian M, Chang CJ, DiStefano LJ, Currie DW, Pierpoint LA, et al. The First Decade of Web-Based Sports Injury Surveillance: Descriptive Epidemiology of Injuries in US High School Boys' Soccer (2005–2006 Through 2013–2014) and National Collegiate Athletic Association Men's Soccer (2004–2005 Through 2013–2014). *J Athl Train*. 2018;53:893–905.
4. Zuckerman SL, Kerr ZY, Yengo-Kahn A, Wasserman E, Covassin T, Solomon GS. Epidemiology of sports-related concussion in NCAA athletes from 2009-2010 to 2013-2014. *Am J Sports Med*. 2015;43:2654–62.
5. Ekstrand J, Häggglund M, Waldén M. Injury incidence and injury patterns in professional football: The UEFA injury study. *Br J Sports Med*. 2011;45:553–8.
6. Serner A, Chamari K, Hassanmirzaei B, Moreira F, Bahr R, Massey A, et al. Time-loss injuries and illnesses at the FIFA world cup Qatar 2022. *Sci Med Footb*. 0:1–8.
7. Chandran A, Morris S, Wasserman E, Boltz A, Collins C. Methods of the National Collegiate Athletic Association Injury Surveillance Program (NCAA ISP), 2014/15 through 2018/19. *J Athl Train*.
8. Hespanhol L, Vallio CS, Costa LM, Saragiotto BT. Understanding and interpreting confidence and credible intervals around effect estimates. *Braz J Phys Ther*. 2019;23:290–301.
9. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. 3rd ed. New York: Chapman and Hall/CRC; 2015.
10. Lambert B. *A Student's Guide to Bayesian Statistics*. 1st edition. Los Angeles: SAGE Publications Ltd; 2018.
11. Chandran A, Lambert B. Bayesian methods for estimating injury rates in sport injury epidemiology. *Inj Epidemiol*. 2025;12:31.
12. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A Probabilistic Programming Language. *J Stat Softw*. 2017;76:1.