Fundamentals of Sports Analytics

Erin B. Wasserman, PhD[a,*], Mackenzie M. Herzog, MPH[b,c], Christy L. Collins, PhD[a], Sarah N. Morris, PhD[a], Stephen W. Marshall, PhD[d]

KEYWORDS

• Epidemiology • Study design • Analytics • Sports performance • Injury occurrence

KEY POINTS

- There are a variety of research study designs that can be used to identify risk factors for injury, including cohort, case control, and case series.
- Each has advantages and disadvantages and their use should be determined by the data available and the research question at hand.
- Sports injury surveillance systems are useful for collecting injury incidence data but have unique limitations.
- Common analytical measures used in sports injury research include injury rate, injury risk and odds, incidence rate ratio, and risk difference and can be calculated based on the study design and research question.
- There has been increased emphasis on using technology to measure athletic performance, but much is still unknown about best practices with these data.

INTRODUCTION

Analytics, in some form or another, have always been a part of sports. Basic statistics, such as the score of the game, or the number of receptions or hits, provide the basis for athletic competition. Recently, however, the importance of statistics and analytics in sports has increased, with emphasis on measures that improve the likelihood of winning or may provide an "edge" over the competition who has not yet discovered the value of these measures. These analytics include measures of sport aptitude

Disclosure Statement: No conflicts of interest to disclose.
[a] Datalys Center for Sports Injury Research and Prevention, Inc, 401 West Michigan Street, Suite 500, Indianapolis, IN 46202, USA; [b] Department of Epidemiology, The University of North Carolina at Chapel Hill, Suite 500, Bank of America Building 7505, Chapel Hill, NC 27599, USA; [c] Real-World Evidence, IQVIA, Research Triangle Park, 4820 Emperor Boulevard, Durham, NC 27703, USA; [d] Department of Epidemiology, The University of North Carolina at Chapel Hill, Injury Prevention Research Center, Suite 500, Bank of America Building 7505, Chapel Hill, NC 27599, USA
* Corresponding author.
E-mail address: ewasserman@datalyscenter.org

Clin Sports Med 37 (2018) 387–400
https://doi.org/10.1016/j.csm.2018.03.007
sportsmed.theclinics.com

(eg, strikeouts in baseball, free-throw percentage in basketball), physical location (eg, pitch location in baseball, distance run in basketball), economic value, and, most relevant to medicine, injury incidence and metrics of physical performance.

Just as measures of sport aptitude have been used in the sports setting to increase win probability, there is increasing recognition that understanding injury occurrence and identifying factors that can prevent injury can provide a team with an advantage on the field or court via implementation of data-driven injury prevention strategies. Additionally, recent advances in technology have led to an improved understanding of physical ability, functional movement, training load, and fatigue. Understanding an athlete's workload can lead to improved training that maximizes athletic performance and minimizes fatigue and injury.

The emphasis of this review is to describe measures of sports injury and fundamentals of sports injury research with a brief overview of some of the emerging measures of sports performance. First, we describe research study designs that can be used to identify risk factors for injury. Second, we discuss an important source of injury incidence data: surveillance programs. Third, we describe common measures of injury risk and association. Finally, we describe measures of physical performance and training and considerations for using these measures. This review provides sports medicine clinicians with an understanding of current research measures and considerations for designing sports injury research studies.

SPORTS INJURY RESEARCH STUDY DESIGNS

The goal of research is to answer a question about a broader population using a sample of data. Here, the population is the entire group that we want to study (eg, all adolescent boys' basketball players in the United States), and the study sample is a subset (eg, a sample of United States high school boys' basketball teams) that we actually examine to make an inference about the population of interest. Our ability to answer research questions depends on the strength of the study design, including the overall framework used, the validity of the data collected, and the rigor of the analysis and interpretation.

Evidence can be generated from both experimental and observational study designs. Experimental studies, such as randomized controlled trials, are largely considered the gold standard for evidence generation owing to the validity afforded by randomization, that is, the ability to balance participant characteristics that may affect the outcome of interest ("confounders") between groups, control treatment, or exposure delivery, and isolate treatment effects among specific populations of interest.[1–3] However, randomized controlled trials and other experimental study designs also have limitations, including high costs and time required for study completion, potential lack of real-world applicability and generalizability, a need for equipoise, and an exposure that can be ethically randomized.[1–3] Owing to these caveats, many research questions cannot be answered using experimental study designs.

Observational studies can also generate high-quality evidence when rigorous research methods are used. These studies typically fall within 3 broad study designs: cohort, case control, and case series studies (**Table 1**).[1,3] Here we provide a high-level overview of the fundamental concepts underpinning these study designs. These study designs can be used with data stemming from a variety of sources, including surveillance data, electronic medical records, administrative claims or billing information, or data specifically collected for research purposes, and the source of the data and study design are not inherently linked together, but rather separate concepts.