Human and Ecological Risk Assessment

ISSN: 1080-7039 (Print) 1549-7860 (Online) Journal homepage: https://www.tandfonline.com/loi/bher20

Probability Models of Occupational Injury

James T. Wassell

To cite this article: James T. Wassell (1998) Probability Models of Occupational Injury, Human and Ecological Risk Assessment, 4:6, 1275-1283, DOI: 10.1080/10807039891284659

To link to this article: https://doi.org/10.1080/10807039891284659

Published online: 03 Jun 2010.

Human and Ecological Risk Assessment: Vol. 4, No. 6, pp. 1275-1283 (1998)

Probability Models of Occupational Injury

James T. Wassell*
National Institute for Occupational Safety and Health, Morgantown, WV

Key Words: injury risk assessment, occupational injury, probability

INTRODUCTION

Probability models are an attempt to describe the random mechanism underlying events for which the outcome is uncertain. In injury risk assessment, these models provide the basis for developing and evaluating summary measures of the incidence of occupational injuries for comparison of occupations and industries. Probability models are used to represent the mechanism by which injuries occur to some individuals, while other individuals are not injured, under similar circumstances. Methods for statistical analysis of occupational injury data require assumptions based on a reasonable mechanism describing the random nature of the occurrence of events, after accounting for all observable injury determinants.

Probability models are an important consideration whether the data are obtained as a sample of a larger population, or if the data describe all the events in a well-defined population (*i.e.*, a census). In either case, events, such as occupational injuries, require a probability model in the absence of complete information to perfectly predict injury events. After accounting for individual characteristics and explanatory factors that may affect injury risk, the fact remains that some individuals are injured and some, under similar circumstances, are not injured. Thus a probability model is needed to describe the nondeterministic nature of injury incidence.

Probability models must be distinguished from statistical models which are usually formulated with the primary consideration of residual error or sampling variability (Choi and Bang, 1998). In the application of regression or

* National Institute for Occupational Safety and Health, Division of Safety Research, 1095 Willowdale Road, Morgantown, WV 26505; Tel: (304) 285-5946; Fax: (304) 285-6046 or (304) 285-6047

linear models, emphasis is often placed on the necessary assumption that residual error terms have a gaussian (normal) distribution with zero mean and constant variance. Researchers are frequently willing to accept this assumption, almost without validation, because it permits inference and testing of the parameter estimates. Assessing the validity of the underlying random process for a specific application requires going beyond the routinely recognized textbook language of model assumptions. It is important to recognize the limitations of probability models that we find useful, through a complete exploration of the underlying random mechanism they attempt to describe.

Identifying, with certainty, those occupations and industries that are at high injury risk can be a difficult analytical problem that may be resolved through the consideration of probability models. Explicit definition of the random variables and their associated probability distributions permits a greater understanding of the strengths and limitations of comparisons based on the summary measures. The *working lifetime risk*, which is the probability that a work related fatal injury occurs during a 45-year working lifetime, has been proposed as a measure that allows comparison of occupational illness and occupational injury (Fosbroke, Kisner, and Myers, 1997; Bailer *et al.*, 1998). A threshold value of one fatality in 1000 workers in a working lifetime has been used for setting occupational health standards and provides motivation for the use of this metric in setting occupational safety standards (Martonic, Grossman, and Gordon, 1998). The working lifetime risk can be shown to be the result of several different probability models, which describe different types of random variables, based on different mechanisms for the random occurrence of injury.

This paper explores the applications of some elementary probability distributions for occupational injury data; examines the concept of working lifetime risk using several different probability models; considers the working lifetime risk and alternative metrics for comparison of occupation and industry groups; introduces the limit form of the working lifetime risk; demonstrates that time-until-event methods can be used to estimate the working lifetime risk; and mentions additional unique characteristics of occupational injury data. The limit form of the working lifetime risk demonstrates that the application of survival analysis methods to occupational injury data provides a generalization of the working lifetime risk.

PROBABILITY MEASURES OF OCCUPATIONAL INJURY EVENTS

Binomial Probability Models

The most common way in which probability is defined is as a *relative frequency* measure of the probability of an event. Simply stated, a researcher first identifies a situation involving repetition of identical trials and counts the number of times a particular outcome occurs. A ratio of the count of the number of some specific outcome, to the total number of trials observed, provides an estimate of the probability that the outcome will be observed on future attempts.

To estimate the probability of occupational fatal injury, researchers observe the number of workers experiencing a fatal injury in a particular occupation, over some specific time period (*e.g.*, 1 year), and divide by the number of persons employed in that occupation. For this discussion, this ratio will be denoted as $\hat{p}$. This number describes a situation that has already occurred, based on past data, and provides some estimate of the probability that a worker in a particular occupation will experience a fatal injury in the future, if all other factors that might affect fatal injury incidence remain the same. The number is useful even without the need to consider future events; this number can be used to characterize the past, if the goal is to compare past performance of different industries or occupations.

In occupational injury risk assessment, interest is usually centered on the probability of a fatal injury occurring to an individual worker during some fixed, usually pre-specified time interval. Considering a group of employees in an occupation or industry, more than one fatal injury may occur, to a different individual, of course, without much change in the estimated probability of injury, $\hat{p}$, when the total number of workers is large. The numerator of the relative frequency estimate counts each fatality only once and represents a subset of the total number of workers at risk in the denominator. The discussion of repeated non-fatal, recurrent injuries poses difficult analytical problems related to correlated random variables requiring complex probability models that form the basis of advanced analytical methods which will be discussed later in the context of survival analysis techniques.

Depending on the intended use of the relative frequency estimate $\hat{p}$, it will be interpreted differently. An individual who works in a particular industry or occupation may consider this number as an estimate of the probability of a fatal injury on the job and it might direct individual career goals or financial planning. The number provides an answer for the question: "What is the chance I will be injured in this job during the next year?" Researchers, for example, in the insurance industry, may interpret this relative frequency estimate as the probability that one or more claims will be received during some period of time. The proportion might be used to determine the expected number of fatalities by multiplying $\hat{p}$ times the best estimate of the number of employed that are "at risk" for injury. In this setting, the number is considered to represent a probability for the members of a large group of employees rather than for an individual employee. This is more consistent with the process routinely used to obtain the data in occupational injury studies and extends the interpretation of $\hat{p}$ to characterize a group of workers in a particular occupation or industry.

The number of observed injuries may be represented as a binomial random variable, $X = 0, 1, 2, \ldots, n$, with the following formula representing the probability that X injuries are observed: $P(X = x) = {}_nC_x p^x(1 - p)^{n-x}$ where the notation ${}_nC_x$ is the combination or number of ways of selecting X items out of a group of N. In order to use this probability model, the probability of injury for each individual must be constant, and not be affected by whether or not another worker is injured (independent events). The relative frequency estimate, $\hat{p}$, serves as an estimate for p, in this formula.

Consider a modification of the previous situation, where we observe a group of workers over a long period of time, say 45 years, from age 20 to age 65. The binomial probability model can be used to address the question: "What is the probability that there is *at least one* year in which there is at least one work injury among a large group of employees during the working lifetime of 45 years?" The probability estimate needed should indicate the relative frequency of occupational injury occurring in any one year, *i.e.*, the number of years in which injuries occur divided by the number of years for which data are available. Ideally, the estimate would be obtained by counting the number of fatal injuries that occur in a large group of employees over a 45-year observation period. The estimate, $\hat{p}$, based on injuries per workers, is used as an estimate of the probability of injuries per year, in the absence of the ideal estimate.

Age-group or other group specific probabilities can be used to describe the injury counts stratified by age or other groups, if available, call them $\hat{p}_i$, where i is an index that identifies the age-specific probabilities. Another option is to make the simplifying assumption that the probability of injury is constant for all ages, $p_i = p$ for all values of i.

Note that there are 45 different ways that exactly one injury can occur in 45 years and there will be ${}_{45}C_2 = 990$ ways that two injuries can occur in 45 years, ${}_{45}C_2 = 14{,}190$ ways that three injuries can occur in 45 years, etc. We calculate the sum of all the probabilities associated with the different combinations or, to simplify the problem, consider the *complement* of the event to obtain an answer. $P(X \geq 1) = 1 - P(X = 0)$. Because the probability of zero injuries in a 45-year lifetime is estimated as $(1 - \hat{p})^{45}$ or $\Pi_i(1 - \hat{p}_i)$ for age-specific probabilities, then the probability of *at least one* year in which there is at least one work injury during the working lifetime is $1 - (1 - \hat{p})^{45}$ (Fosbroke *et al.*, 1997) or $1 - \Pi_i(1 - \hat{p}_i)$ (See and Bailer, 1998) which are two representations of the *working lifetime risk* for occupational injury. (Note that, for convenience, the working lifetime risk is usually defined as the value obtained above multiplied by 1000; because the constant multiplier has no effect on the discussion of probability models, it will be ignored in this discussion.)

This model is based on a prospective viewpoint, anticipating all possible outcomes for the entire 45-year period. The working lifetime risk uses no information from historical data about the proportion of years in which injuries have been known to occur, although it is rare to find a well defined occupational group with a reasonable number of employees who were injury-free for an entire year. Note that the use of $\hat{p}_i = \hat{p}$, for all i, results in an estimate of the working lifetime risk that is equivalent to an order invariant transformation of the $\hat{p}$ (working lifetime risk is a nearly linear function of $\hat{p}$ over the range of typical values) when comparing different occupations or industries (Chen and Fosbroke, 1998).

The Geometric Distribution

An alternative formulation is more specific to fatal occupational injury risk assessment from the viewpoint of the probabilities perceived by an individual. Consider the following setup: over a period of 45 years an individual worker may have an estimated probability, $\hat{p}$, of an occupation-related fatal injury each year. The number of years involved in the calculation of the lifetime probability varies from individual to individual and ends on the year of the worker's fatal injury. If the number of years survived, $Y = 0, 1, 2, \ldots$, is counted for each individual, this number is a random variable with the geometric distribution, often referred to as the number of trials preceding an event. Unlike the binomial random variable, the maximum of a geometric random variable is not restricted to the number of trials, although the underlying probabilistic structure is identical, requiring independent trials with a constant probability of a fatal injury. This is a retrospective approach, starting with the year in which the fatal injury might occur and multiplying by the probabilities that no fatal event occurs in the years of life prior to the year of the fatal injury, while the worker is "at risk" of a fatal injury. In the case of working lifetime risk, the $P(Y < 45)$ is the sum of the probability for the events where an individual dies from a fatal injury in year one, added to the probability of the event that an individual dies in year two, etc.

$P(Y = 0) + P(Y = 1) + P(Y = 2) + \cdots + P(Y = 44)$, which is: $\hat{p} + (1 - \hat{p})*\hat{p} + (1 - \hat{p})^2*\hat{p} + \cdots + (1 - \hat{p})$. In general, this is the sum of the geometric series: $P(Y < k) = \Sigma_{y=0}^{k-1} (1 - p)^y p = 1 - (1 - p)^k$. Again, this quantity, which has developed out of a different probability model, is known as the *working lifetime risk*.

A few important points are also necessary for the application of these probability models. In order to multiply (or exponentiate) probabilities, the probabilities must represent independent events. In this setting, this is only appropriate if the occurrence of occupational injuries in one year does not influence the probability of injury in subsequent years, so that the probabilities are constant over years. This is a fairly restrictive assumption; however, the model is often used even when time trends are present or other departures from the independence requirement are not too severe.

Poisson Probability Model

Another interpretation of the relative frequency measure, $\hat{p}$, is to consider it as an estimate of the *rate* at which injuries occur, taking the number of injuries, $D = 0, 1, 2, \ldots$, observed in a fixed time period and for a fixed number of workers "at risk", w, as having a Poisson distribution (Bailer, Reed, and Stayner, 1997). This requires that the assumptions of a Poisson process apply to injury incidence, requiring that (the probability of) the number of injuries in any time period is independent of (the probability of) the injury counts in other time periods (Hogg and Craig, 1995). As a consequence, both the mean and variance of D are estimated by $\hat{p}$ multiplied by the number of workers at risk, $\mu = \hat{p}w$ and $\sigma^2 = \hat{p}w$. The probability associated with the different injury counts are: $P(D = d) = ((\hat{p}w)^d \exp(-\hat{p}w)) / (d!)$. Notice that, for $w = 1$ in any single year, $P(D = 0) = \exp(-\hat{p})$. Over a working lifetime, $P(D \geq 1) = 1 - P(D = 0)^{45} = 1 - \exp(-45\hat{p})$, which is a limit representation of the *working lifetime risk* provided by the Poisson probability model. The Poisson distribution is well known to be the limiting distribution for a binomial random variable with large n and small values for the binomial parameter, p (Hogg and Craig, 1995).

Given that injuries occur according to this probability model, interest is usually directed toward comparing different rate estimates based on different occupations, industries or other groups, $\hat{p}_i$, where the subscript i may be used to indicate different industries, occupations or other groups. The expected value for the number of injuries over a 45 year period, $\hat{p}_i \times 45$ years, may be shown to be an approximation of the *working lifetime risk* based on a Taylor series expansion of the $P(D = 0)^{45} = \exp(-45\hat{p})$.

Fortunately, comparison of group-specific rates does not require accurate characterization of the distribution. An approach, based on the empirical Bayes method, has recently been used for identification of elevated risks for different types of cancer in various occupations (Carpenter *et al.*, 1997). Another quantitative approach is to estimate the variance and derive confidence intervals for the working lifetime risk as a function of the individual group-specific rates (See and Bailer, 1998). Other methods may work well for identifying groups of industries and occupations with unusually high rates, *e.g.*, identifying as "high risk" those groups with estimated lifetime risk values greater than 1 death in 1000 working lifetimes and with a minimum of 5 deaths (Myers, Kisner, and Fosbroke, 1998).

Application of Life Table and Survival Analysis Techniques

Life table and survival analysis techniques (Klein and Moeschberger, 1997) have a long history of application in evaluating differences between groups using the survival function, $S(t) = P(T > t)$. For occupational injury risk assessment, the random variable T may represent a discrete, nonnegative random variable, representing the number of years worked until an event (fatal or nonfatal injury) occurs. Because there is no interest in events that occur prior to age 20 or beyond age 65, this is, more specifically, a left-truncated, right-censored survival function. If we consider the probability of death during the working lifetime of 45 years, $t_i = 1, \ldots, 45$, for $i = 1, \ldots, 45$, this is the complement of the event of surviving 45 working years, $P(45 \leq T) = F(t_i = 45) = 1 - S(t_i = 45)$. If $\hat{p}_i$ represents the probability of death in year t_i, then $1 - \hat{p}_i$ can be used to obtain an estimate of the survival function at year t_i. The nonparametric actuarial method for estimating the survival function is determined by the product of survival probabilities for all preceding years, $S(t_i) = \Pi_i(1 - \hat{p}_i)$, for $i = 1, \ldots, 45$. Now, $F(t_i) = 1 - S(t_i) = 1 - \Pi_i(1 - \hat{p}_i)$, which again is the *working lifetime risk* estimate.

Occupational injury studies are typically based on data obtained in accordance with *current life tables* rather than the long term *cohort* studies of a selected group more typical of occupational illness studies. If data were available on the exact ages at which fatal injuries occur, so that T is a continuous random variable, and a more specific determination of the numbers of workers at risk for injury at any given time, a more accurate estimate of $S(t)$ could be obtained using the product-limit or Kaplan-Meier estimate.

As an alternative to the nonparametric approach, a particular parametric form for the distribution of the continuous random variable, T, could be specified. The hazard function is useful in this regard: it is the ratio of the probability density to the survival function and represents the instantaneous probability of a fatal injury conditional on prior survival. The relative frequency estimate, $\hat{p}_i$, may be considered an estimate of the hazard, if the time until injury, T, is taken to have an exponential distribution, $P(t \le T) = 1 - \exp(-\hat{p}t)$, with constant hazard equal to $\hat{p}_i$. Note that $P(45 \le T) = 1 - \exp(-45\hat{p})$, which is equivalent to the limit representation of the *working lifetime risk* provided by the Poisson probability model. This result demonstrates that comparisons of the working lifetime risk are equivalent to comparing the 45-year survival probability for different occupations and industries using the exponential model.

The assumption of a constant hazard is not suitable for all survival data; therefore other parametric survival probability models allow more general hazard functions. Based on some knowledge of the change in hazard for injury with age, parametric models with a decreasing hazard, such as the Weibull distribution with a shape parameter < 1, could be useful in modeling situations where increasing experience on the job decreases a worker's risk over time. Occupations where hazards increase with loss of facility, as a result of aging, could be modeled by an increasing hazard, such as the Weibull distribution with a shape parameter > 1. This suggests another generalization of the working lifetime risk, $P(45 \le T) = 1 - \exp(-\hat{p} \times 45^{\eta})$ with the Weibull shape parameter η.

Survival analysis comprises a collection of methods that are powerful for the evaluation of subject-specific or group-specific covariate information with application to the problem of distinguishing the most hazardous occupations and industries. The location or scale of the Weibull distribution can be expressed as a function, $\hat{p}_i$, = $\exp(\beta Z_i)$, of a vector of individual worker-specific covariates, Z_i, and regression coefficients, β, as a further generalization of the working lifetime risk for comparison of different occupations and industries. The Cox (1972) proportional hazard regression method provides for the most powerful evaluation of covariate effects with an arbitrary nonspecified baseline hazard.

More sophisticated techniques, based on the use of *counting process* methods, permit the consideration of recurrent events especially important for the analysis of nonfatal injuries (Andersen *et al.*, 1993). Both marginal and conditional estimation methods, with new methods for variance estimation of regression coefficients, have been developed. These techniques permit different types of interpretation for the estimated covariate effects, which are useful in clinical studies, but the application of these methods for analysis of occupational injury data remains unexplored (Wei and Glidden, 1979). Recent developments (Wassell and Kulczycki, 1995; Wassell, Wojciechowski, and Landen,

1999) permit the modeling of repeated non-fatal injuries as recurrent events, to account for unobserved random effects in addition to observed covariates. These more advanced methods have greater data requirements and rely on cohort studies rather than more routinely collected census or survey data, which limit their application for occupational injury risk assessment.

SUMMARY

Probability models provide a useful and unifying framework for understanding the wide range of techniques and different interpretations of quantitative methods used for analysis of occupational injury data. Probability models specify a mechanism for the random occurrence of injuries after accounting for the deterministic effects of all measurable risk factors. The relative frequency estimate of the probability of injury has a different interpretation depending on the viewpoint of an individual worker, insurance claims estimator, or researcher interested in the probability applied to a large group of workers.

One metric, the *working lifetime risk*, is described using different probability structures which demonstrates that these definitions are complementary. The estimate of the working lifetime risk can be derived by considering the binomial random variable of the number of injuries, the geometric random variable indicating the number of years until a fatal injury occurs and, the nonparametric estimate of the survival function of the time until a fatal injury occurs. A limit form of the working lifetime risk is provided by considering a probability distribution of the number of events which occur from a Poisson process and the parametric exponential distribution of survival times. This limit form can be generalized by permitting greater flexibility in the hazard function, not restricting the hazard function to a constant. The inclusion of covariates, as a regression model representing the working lifetime risk, may overcome the restriction of the simpler methods which are an order invariant transformation of the fatality rates. More advanced survival analysis methods permit generalizing the concept of the working lifetime risk to permit recurrent (nonfatal) injuries.

Comparison of occupations and industries or other groups, for the purpose of identifying those at high risk, to establish research and intervention priorities, can be effectively accomplished using straightforward techniques. There are many advantages to using well known and studied statistical techniques for estimation and inference for the working lifetime risk. A better understanding of the probability mechanisms describing injury incidence, through examination of probability models, leads to an appreciation of the utility of standard statistical methods for analysis of occupational injury data.

ACKNOWLEDGMENT

The author is grateful to A. John Bailer for valuable review comments and suggestions. Discussions with John Myers, David Fosbroke and William Miller

influenced the content of this paper and are greatly appreciated. Furthermore, the idea for this paper evolved out of having the opportunity to review the contributions of the authors whose papers are included in this special issue on Occupational Injury Risk Assessment.

REFERENCES

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. 1993. *Statistical Models Based on Counting Processes*. New York, Springer-Verlag.

Bailer, A. J., Reed, L. D., and Stayner, L. T. 1997. Modeling fatal injury rates using poisson regression: a case study of workers in agriculture, forestry and fishing. *Journal of Safety Research* **28**, 177–186.

Bailer, A. J., Stayner, L. T., Halperin, W., Reed, L. D., and Seymour, T. 1998. Comparing injury and illness risk assessments for occupational hazards. *Human and Ecological Risk Assessment*. **4**, this issue.

Carpenter, L. M, Maconochie, N. E. S., Roman, E., and Cox, D. R. 1997. Examining associations between occupation and health by using routinely collected data. *J. R. Statist. Soc. A*. **160**, 507–521.

Chen, G-X. and Fosbroke, D. 1998. Work-related fatal injury risk of construction workers by occupation and cause of death. *Human and Ecological Risk Assessment*. **4**, 6, 1371–1390.

Choi, J. W. and Bang, K. M. 1998. Relative risk of involuntary injuries among currently employed adults in the U.S., 1994. *Human and Ecological Risk Assessment*. **4,** 6, 1353–1370.

Cox, D. R. 1972. Regression models and life tables (with discussion). *J. R. Statist. Soc. B*. ***34***, 187–220.

Fosbroke, D. E., Kisner, S. M., and Myers, J. R. 1997. Working lifetime risk of occupational fatal injury. *Am. J. Ind. Med*. **31**, 459–467.

Hogg, R.V. and Craig, A. T. 1995. *Introduction to Mathematical Statistics, Fifth Edition*. New Jersey, Prentice Hall.

Klein, J. P. and Moeschberger, M. L. 1997. *Survival Analysis: Techniques for Censored and Truncated Data*. New York, Springer-Verlag.

Martonic, J. F., Grossman, E. A., and Gordon, C. P. 1998. Injury risk assessment for occupational safety standards. *Human and Ecological Risk Assessment*. **4**, 6, 1259–1264.

Myers, J. R., Kisner, S. M., and Fosbroke, D. E. 1998. Lifetime risk of fatal occupational injuries within industries, by occupation, gender and race. *Human and Ecological Risk Assessment*. **4**, 6, 1291–1307.

See, K. and Bailer, A. J. 1998. Estimates of lifetime risk of occupational fatal injury from age-specific rates. *Human and Ecological Risk Assessment*. **4**, 6, 1309–1319.

Wassell J. T. and Kulczycki, G. W. 1995. Frailty analysis of repeated injuries. *ASA 1995 Proceedings of the Section on Epidemiology*. 130–134.

Wassell, J. T., Wojciechowski, W. C., and Landen, D. D. 1999 (anticipated). Recurrent injury event-time analysis. Submitted to *Statistics in Medicine*.

Wei, L. J. and Glidden, D. V. 1997. An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine*. **16**, 833–839.