AUTHOR PAGE

Category: Literature of EOH

Real World Occupational Epidemiology, Part 1: Odds Ratios, Relative Risk and Asbestosis

Eric J. Beh, BMath (Hons), PhD.

School of Mathematical & Physical Sciences, Faculty of Science and Technology, University of Newcastle, Callaghan, Australia

Derek R. Smith, PhD, DrMedSc, MPH.

School of Health Sciences, Faculty of Health, University of Newcastle, Ourimbah, Australia

Corresponding author:

Associate Professor Eric J. Beh

School of Mathematical and Physical Sciences,

University Drive, Callaghan, New South Wales 2308, Australia

Phone: +61 2 4921 5113

Fax: +61 2 4921 6898

Email: eric.beh@newcastle.edu.au

Real World Occupational Epidemiology, Part 1: Odds Ratios, Relative Risk and Asbestosis

Background

Asbestos represents one of the most heavily studied occupational and environmental hazards of human history.[1] From an epidemiological perspective it also represents one of the most important and historically significant case studies in *Environmental and Occupational Health* (EOH).[2] The sub discipline of EOH which examines and evaluates workplace risks, occupational epidemiology, has itself been evolving as new statistical techniques emerge and the complexities of workplace exposures increase.[3] Nonetheless, three fundamental goals of statistical techniques in EOH are, and have always been, clear: (1) to establish which exposures are related to which disease, (2) to determine in what magnitude they may be related, and (3) to differentiate (or eliminate) whether the results identified are simply due to chance. The term 'statistical significance' is commonly used in occupational epidemiology to denote the rejection of a null hypothesis, that is, to establish whether the differences between two groups (such as exposed and non-exposed individuals) is as great or greater than could have occurred due to chance alone. Although most studies tend to cite p-values as being less than or equal to 0.05, the actual choice of 0.05 (referred to as the *level of significance*) is entirely arbitrary – an investigator may choose any

value that seems reasonable to them.[4] For whatever reasons, be they pragmatic, traditional or otherwise, science and scientists have gradually accepted 0.05 (or having only a 1 in 20 chance of being wrong) as the de facto standard for scientific research. So too has occupational epidemiology.

Most contemporary statistical techniques are based on early studies of games of chance. Although seminal events include the chi squared test for goodness of fit by Pearson[5] and Fisher's exact test,[6] it is the tests of significance in contingency tables that represents what is perhaps one of the most useful techniques in occupational epidemiology. Odds Ratios (OR) can be derived from contingency tables and provide a ratio of the probability that an event occurs compared to the probability that it does not.[7] The concept of 'risk' with respect to a particular exposure has always been fundamental in EOH, and as a result, it is the relative risk estimated by OR that has become a de facto standard for 'hazard' in modern EOH.[8] OR are often interpreted as Relative Risk (RR), and given certain conditions, may give a reasonable approximation of it.[9] OR have become popular as they provide an estimate for the relationship between two variables, they enable examination of the effects of other variables on the relationship and they offer a convenient interpretation of case-control studies.[10]

The study of industrial exposure to asbestos and its relationship with subsequent disease represents an excellent example in this regard. Not only because the topic regularly appears in the literature of EOH,[11] but also considering the fact that workers are still being exposed to it.[7] In the current article we explore the use of the OR and RR, as well as reviewing many of the statistical and practical aspects of its

implementation using data from Irving Selikoff's classic asbestos research conducted during the 1960s and 70s. Although the main study began in the early 1960s, some interesting and demonstrative raw data was later published in a 1981 issue of the *Bulletin of the New York Academy of Medicine* (used with permission).[12] Parts of this data are analysed in our current article for demonstrative purposes.

Odds Ratios in Occupational Epidemiology

To begin our discussion of OR usage in the 'real' world of occupational epidemiology, we shall consider Table 1 which is based on Selikoff's paper from 1981.[12] This table summarises data collected in 1963 from 1117 insulation workers in metropolitan New York, including duration of asbestos exposure and diagnosis of asbestosis. Interestingly during clinical x-ray examinations, Selikoff noticed that most chest films were normal among workers with less than 20 years exposure, while among those with 20 years exposure or more, most roentgenograms displayed abnormal findings. He termed this the '20-year rule' and thus, classified the period of exposure as either 0-19 years or 20+ years.

Table 1 Contingency Table Based on Selikoff's Original Asbestosis Data*

	Asbestosis		
Onset of Exposure	**No**	**Yes**	**Total**
0-19 years	522	203	725
20+ years	53	339	392
Total	575	542	1117

*Adapted from Selikoff (1981)[12] published in the *Bulletin of the New York Academy of Medicine* (used with permission)

Using data from Table 1, the prevalence of asbestosis can be measured in a variety of different ways. For example, for those with an asbestos exposure of less than 20 years the odds of contracting asbestosis can be calculated as: 203 / 522 = 0.39. Since this result is less than 1 it suggest that individuals exposed to asbestos for less than 20 years are not likely to be diagnosed with asbestosis. On the other hand, for individuals exposed to asbestos for 20 years or more the odds that they are then diagnosed with asbestosis is: 339 / 53 = 6.396. That is, workers exposed for a relatively long period of time are around 6.4 times more likely to be diagnosed with asbestosis when compared to their colleagues with shorter exposures. Therefore a person has been exposed to asbestos for 20 years or more is (339 / 53) / (203 / 522) = 6.396 / 0.39 = 16.4 times more likely to contract asbestosis than those who have been exposed to asbestos for less than 20 years. This simple calculation certainly provides compelling evidence to support Selikoff's '20-year rule'.[12] The ratio of 16.4 calculated above can be referred to as the OR and succinctly helps quantify the association between asbestos exposure and being diagnosed with asbestosis. Another method for calculating the odds ratio is by looking at the ratio of the product of values diagonally on Table 1. That is: (522 x 339) / (53 x 203) = 16.4.

It is important to note that the OR cannot be negative since this would imply that there were a 'negative' number of individuals sharing certain characteristics - a result which is simply not possible. As such, the odds ratio can range from 0 to practically any large number, including up to infinity (although the latter result would invoke some degree of suspicion in occupational epidemiology for obvious reasons!). An OR greater than 1 indicates a 'positive' association between the rows and columns of the 2x2 contingency table. When considering the data from Table 1 in a practical

manner, our derived OR of 16.4 suggests a positive association – the longer a worker is exposed to asbestos, the more likely (or the higher the odds) that they will be diagnosed with asbestosis. On the other hand, had the OR fallen between 0 and 1, this would suggest a ‘negative’ association between the rows and columns - a finding that is often interpreted in EOH as being a ‘protective’ effect. Where there is no more of a preference for one response over another (so that there is no association), the OR will be 1.

Zero and Undefined Odds Ratios

As convenient as OR can be for epidemiologists, difficulties in interpretation arise when calculated values lie at the extremes of its permissible range. Indeed, there are many practical situations where this value equals zero, or simply cannot be calculated. Consider, for example, Table 2 which summarises the onset of exposure to asbestos versus the number of workers diagnosed with either of the two extremes of asbestosis: Grade 1 (the least severe) and Grade 3 (the most severe). By following the same process as described above, the OR calculation in Table 2 is: (194 x 50) / (172 x 0), which gives an undefined result. Note that dividing by zero does not lead to a quantity of infinity. Rather, if we divide a number by a value that approaches zero, then that ratio will approach infinity, but will not in itself, be infinite.

Table 2 Contingency Table Based on Selikoff's Original Asbestosis Data*

	Asbestosis Grade**		
Onset of Exposure	**1**	**3**	**Total**
0-19 years	194	0	194
20+ years	172	50	222
Total	366	50	416

*Adapted from Selikoff (1981)[12] published in the *Bulletin of the New York Academy of Medicine* (used with permission), **Asbestosis graded from least severe (Grade 1) to most severe (Grade 3)

While this may seem obvious, dividing, or multiplying, by zero represents a serious practical problem since it does not elucidate any potential association between rows and columns. In both cases, one way to overcome the presence of a zero is to add a small quantity (typically 0.5) to each cell in the table, yielding Haldane's odds ratio.[13] The choice to use 0.5 has become popular as it is thought to best reduce the possibility of bias in small samples.[14] When applied to the current example from Table 2, this gives a revised odds ratio of: (194.5 x 50.5) / (172.5 x 0.5) = 113.9. Interpreting these numbers suggests that an individual exposed to asbestos for at least 20 years is nearly 114 times more likely to be diagnosed with Grade 3 rather than Grade 1 asbestosis. Care must be taken, however, to avoid too literal an interpretation of the meaning OR in such cases. This is because any other value close to zero could be added to each of the counts in the contingency table to overcome the zero count. For example, 0.05 could be added to each of the four cells (since it closer to 0 than 0.5), resulting in an odds ratio of: (194.05 x 50.05) / (172.05 x 0.05) = 1129. Alternatively, since the sample size is relatively large (at 416), an additional unit can be added to each cell so that the OR becomes: (194 x 51) / (173 x 1) = 57.5. Similarly, dealing with a zero count by adding a constant value to each cell

in the table leads to a questionable interpretation of the magnitude of that OR. Despite these caveats and regardless of the value added, the preliminary findings from our basic analysis of Selikoff's data is clear – individuals exposed to asbestos for at least 20 years are far more likely to be diagnosed with Grade 3 asbestosis than Grade 1 asbestosis.

Advantages and Disadvantages

An advantage of the OR in occupational epidemiology is that, regardless of how big or small the sample actually is, it continues to preserve the underlying association of what is being analysed. For example, if Selikoff's study had recruited 10 times the number of workers, the calculation (based on the data in Table 1) would have been: (5220 x 3390) / (530 x 2030) = 16.45, in other words, exactly the same as the original result. It is important to note that this does not mean that the sample size can be discounted, as it plays a key role in the determination of confidence intervals, and will be discussed later. Small cell counts on the other hand, lead to a more biased OR and increase the variance. It is also important to remember when examining categorical variables that *association does not necessarily mean causation*. Although our previously described analysis clearly demonstrates that exposure to asbestos (for a certain time period) and subsequent diagnosis of clinical asbestosis are associated, from a theoretical perspective this does not mean that one always causes the other. Although medical research has now shown this to be true, it was far from certain when the original studies of asbestosis were conducted over 50 years ago. Until the link could be conclusively proven and the sceptics silenced, Selikoff and his team were working with this caveat in mind. Indeed, all researchers who use OR in their results should be doing the same.

Log-Odds Ratios and Confidence Intervals

In the previous sections we have described how OR range between 0 and infinity and the range of possible values permissible for a positive association is far greater than the range of possible values when a negative association exists. This apparent bias can lead to some intuitive problems in providing a meaningful interpretation of the derived values. There is also great difficulty in using the OR to formally test whether a statistically significant association exists between the rows and columns of a 2x2 contingency table. To overcome these problems, one may consider the (natural) logarithm of an OR, thereby calculating the log-OR of a 2x2 table. One key advantage of this method is that the log-OR is approximately standard normally distributed, thereby allowing for formal tests of the association to be conducted.[15]

From an intuitive perspective, the logarithm of any OR value that lies between 0 and 1 is negative, the logarithm of any value that is greater than 1 is positive while the logarithm of 1 is zero. Therefore, when a negative association exists (when the OR is between 0 and 1), the log-odds ratio is negative. When there exists a positive association between the rows and columns (where the odds ratio is greater than 1), the log-odds ratio is positive. When there is no association (when the OR is exactly 1) the log-OR is also zero. Consider Serikoff's data summarised in Table 1 from which we earlier derived an OR of 16.4. Taking the logarithm of this quantity gives a log-OR of 2.8, thus confirming the existence of a positive association between the onset of asbestos and being diagnosed with asbestosis. Similarly, with an OR of 8.2, the log-OR of Table 2 is 2.1.

Another important advantage of considering the log-OR is that it allows for the estimation of the likely variation of the measure of association in the population under study. One key aim of Selikoff's study was to understand the relationship between exposure to asbestos and the odds of being diagnosed with asbestosis among a sample of 1117 insulation workers from New York. The OR, and therefore the log-OR, only summarises this link based on the sample collected. If another sample of 1117 was randomly selected, the log-OR would be different, thereby implying variation in this quantity. To estimate the log-OR in the population of insulation workers in New York, we first need to calculate the standard error to quantify the variation. Note that according to Selikoff's study, the population of insulation workers in the city is quoted as 1250. However, we shall assume in our calculation here that it is actually unknown. This quantity can be measured by the Standard Error (SE):

$$\text{SE (log(OR))} = \sqrt{\frac{1}{522}+\frac{1}{203}+\frac{1}{53}+\frac{1}{339}} = 0.17$$

Using this SE, the 95% confidence interval for the log-OR in the population of all insulation workers is calculated by log(16.4) ± 1.96 x 0.17 and therefore lies between 2.47 and 3.13. As such, this can be interpreted to mean that we are 95% confident (95 times out of 100) that the log-OR in the population lies between 2.47 and 3.13. Again, this highlights the possibility of there exists a very strong positive association between exposure to asbestos and the likelihood of being diagnosed with asbestosis. Although we have determined the 95% Confidence Interval (CI) of the log-OR for a population of all insulation workers, the decision to use 95% is arbitrary and can be changed to reflect the genuine level of confidence that needs to be considered. As previously described, the level of significance can be any value that one feels

comfortable with. In fact, in many practical situations 0.05 may instil sufficient confidence since it implies that only 5% of the time (1 in 20 times), the population log-odds ratio will be poorly estimated.

To be more certain of estimating a true log-OR, one may wish to consider instead a 99% CI, or even a 99.5% CI. In doing so, these intervals are (based on the sample OR of 16.4 generated from Table 1) 2.36 – 3.23 (calculated by considering log(16.4) ± 2.576 x 0.17) and 2.32 – 3.27 (calculated by considering log(16.4) ± 2.81 x 0.17) respectively. Note that as we increase the level of confidence, the interval widens. If a tighter confidence interval is more desirable (one which more precisely estimates the population log-OR), this can be achieved in one of two ways. One could either reduce the level of confidence, for example to 50%, giving a CI of 2.68 – 2.91 (generated by considering log(16.4) ± 0.67 x 0.17). This first option would clearly be unacceptable as it affords only a 50-50 chance that the log-OR for all insulation workers has been accurately estimated. Alternatively, increasing the same size will reduce the width of the confidence interval. For example if the original sample size obtained by Selikoff was actually multiplied by a factor of 10, the OR remains 16.4, but the SE becomes:

$$\text{SE } (\log(OR)) = \sqrt{\frac{1}{5220}+\frac{1}{2030}+\frac{1}{530}+\frac{1}{3390}} = 0.054$$

thereby giving a 95% CI of 2.69 – 2.90 (calculated by considering log(16.4) ± 1.96 x 0.054). The CI of the OR can be simply derived by taking the antilogs of each of the limits of the intervals from the log-OR. If so, the 95% CI for the OR of Table 1 is 11.82 to 22.87. By interpretation, this suggests we are 95% confident that, for the

population, those exposed to asbestos for at least 20 years are between 11.82 and 22.87 times more likely to be diagnosed with asbestosis than those who are exposed to asbestos for less than 20 years - a calculation that again validates the aforementioned 20-year rule proposed by Selikoff almost half a century ago.

Relative Risk and Odds Ratios

Earlier in this article we highlighted how OR calculations are often interpreted as 'risk' in the field of occupational epidemiology. When considering Selikoff's data from Table 1, suppose we consider an insulation worker with asbestos exposure in excess of 20 years. The probability of being diagnosed with asbestosis is: $P_1 = 339 / 392 = 0.865$. Similarly, for a worker exposed for less than 20 years, the probability of being diagnosed with asbestosis is: $P_2 = 203 / 725 = 0.28$. As such, a worker exposed to asbestos for at least 20 years appears to be at far greater risk of being diagnosed with asbestosis than his or her counterparts exposed for less than 20 years. This risk, referred to as the Relative Risk (RR) can be quantified as: $RR = P_1 / P_2 = 0.865 / 0.28 = 3.09$. The derived result suggests that an asbestos worker is more than 3 times more likely to be diagnosed with asbestosis than a worker selected at random. There is a clear difference in the magnitude of the two measures however - the OR being 16.4 and the RR being 3.09. Indeed, the OR always overestimates the RR, with discrepancies being particularly marked when the incidence of the outcome of interest is high and when the OR itself is high.[16]

To establish the source of discrepancy between an OR and an RR, one must first assume (or reject), that the column totals are fixed. When investigating the odds of being diagnosed with asbestosis in the current example, the RR assumes that the

column totals (number of workers who have been exposed to asbestos for less than 20 years, or more than 20 years), is fixed. On the other hand, the OR does not make such an assumption. Rather, the latter calculation considers the magnitude of each cell in the table. Fixed row and / or column totals are a common feature in many statistical techniques designed to measure association, the most popular of which is probably the chi-squared test of independence. There is clearly a link between the odds ratio and the relative risk, however.[16] It can be defined as: $\mathrm{RR} = OR / \left[\left(1 - \mathrm{P}_2\right) + OR \times \mathrm{P}_2 \right]$ and is well understood in the statistics discipline, especially when examining the association present in 2x2 contingency tables. Its use and limitations in EOH, occupational epidemiology and clinical medicine has also been explored and debated in depth. Various methods have been proposed, such as that by Zhang,[9] which helps correct the OR during cohort studies with common outcomes. Zhang states that an algebraic relationship that exists between the OR and the RR in terms of P_2. In fact a similar relationship may also be derived in terms of P_1; $\mathrm{RR} = OR + \mathrm{P}_1\left(1 - OR\right)$. Note that when there is no association (such that the OR = 1), the RR = 1 which implies that no one outcome is more at risk of occurring than any other outcome.

Conclusion

In the current article we have explored the use of the OR and RR, as well as reviewing many of the statistical and practical aspects of their implementation. We have used Irving Selikoff's classic asbestos research from the 1960s as it serves not only as an excellent example for highlighting issues with the calculation and interpretation of data in occupational epidemiology, but also as an important case study for understanding a 'real world' hazard in EOH. In 2011 it seems almost certain

that asbestos will continue to remain an important issue for epidemiologists to tackle in the future. The incidence of asbestos-related disease continues to rise in countries such as Australia,[17] and a universal ban for asbestos has still not been achieved.

References

1. Guidotti TL. Why study asbestos? *Arch Environ Occup Health.* 2008;63:99-100.
2. Smith DR, Beh EJ. Occupational epidemiology in the real world: Irving Selikoff, odds ratios and asbestosis. *Arch Environ Health.* 2011;66:(in press).
3. Guidotti TL. Occupational epidemiology. *Occup Med (Lond).* 2000;50:141-5.
4. Hammond EC. Statistical significance. *Am J Ind Med.* 1983;4:397-8.
5. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5.* 1900;50:157-175.
6. Fisher RA. On the mathematical foundations of theoretical statistics. *Philos Trans R Soc Lond A.* 1922;222:309-368.
7. Ramazzini C. Asbestos is still with us: repeat call for a universal ban. *Arch Environ Occup Health.* 2010;65:121-6.
8. Smith DR, Attia J, McEvoy M. Exploring new frontiers in occupational epidemiology: the Hunter Community Study (HCS) from Australia. *Ind Health.* 2010;48:244-8.
9. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA.* 1998;280:1690-1.
10. Bland JM, Altman DG. Statistics notes. The odds ratio. *BMJ.* 2000;320:1468.

11. Smith DR. Highly cited articles in environmental and occupational health, 1919-1960. *Arch Environ Occup Health.* 2009;64 (Suppl.):32-42.

12. Selikoff IJ. Household risks with inorganic fibers. *Bull N Y Acad Med.* 1981;57:947-61.

13. Haldane JB. The estimation and significance of the logarithm of a ratio of frequencies. *Ann Hum Genet.* 1955;20:309-311.

14. Breslow N. Odds ratio estimators when the data are sparse. *Biometrika.* 1981;68:73-84.

15. Agresti A, *An Introduction to Categorical Data Analysis (2nd Ed)*. 2007, New Jersey: Wiley-Interscience. p.31.

16. Schmidt C, Kohlmann T. When to use the odds ratio or the relative risk? *Int J Public Health.* 2008;53:165-167.

17. Smith DR, Leggat PA. 24 years of pneumoconiosis mortality surveillance in Australia. *J Occup Health.* 2006;48:309-13.