NOVA

University of Newcastle Research Online

nova.newcastle.edu.au

Tran, Duy; Beh, Eric J.; Smith, Derek R. "Real-world occupational epidemiology, Part 3: an aggregate data analysis of Selikoff's '20-year rule'" Archives of Environmental & Occupational Health Vol. 67, Issue 4, p. 243-248 (2012)

Available from: <http://dx.doi.org/10.1080/10937404.2012.678766>

This is an Accepted Manuscript of an article published in Archives of Environmental & Occupational Health on 17/10/2012, available online:
<http://www.tandfonline.com/10.1080/10937404.2012.678766>

Accessed from: <http://hdl.handle.net/1959.13/1053709>

Real World Occupational Epidemiology, Part 3: An Aggregate Data Analysis of Selikoff's '20 Year Rule'

Duy Tran, BMathIT (Hons).

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

Eric J. Beh, PhD.

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

Derek R. Smith, PhD.

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

Keywords: asbestosis, asbestos, aggregate association index, Irving Selikoff

Background

A number of different statistical techniques can be used to analyse potential associations in occupational epidemiology, such as the calculation of the odds ratios, correlations or chi-squared statistics. Most of these calculations however, rely on having a complete dataset, even though the collection of data is rarely straightforward in real life. In many practical situations when researching in the field of *Environmental and Occupational Health* (EOH) one may have knowledge of the row and column totals for example, but have little or no information on the value of the cells themselves. This may be because such data were not recorded at the time the study was taken, the data that was collected may have been unreliable or incomplete; or simply because the data one needs to undertake a thorough analysis were not made public for reasons of confidentiality.¹ In such situations, an analysis of *aggregate data* can be very useful for deriving a meaningful result from previously 'incomplete' datasets.

Asbestos is one of the most heavily studied EOH hazards of human history,² partly because it represents an archetype example of the expose-response relationship, and partly because the problem has simply not gone away. Due to its long latency period, in certain parts of the world where asbestos use is now banned, mortality rates continue to rise.³ One of the most influential asbestos researchers of the 20th Century was Irving Selikoff (1915-1992), who was one of the first to notice an exposure level (20 years) after which disease tends to occur. His aforementioned '20 year rule' was based on a summary of data from his pioneering study of New York asbestos workers (see Table 1 below), which was originally published in the *Bulletin of the New York Academy of Medicine* in 1981 (used with permission).⁴

We have previously explored and reanalysed Selikoff's pioneering asbestosis data in two earlier articles: Part 1⁵ which focussed on its relationship with odds ratios and relative risk, and Part 2,⁶ which described a visual interpretation of statistical significance. The objective of the current article, Part 3, is to apply a new statistical approach called the *Aggregate Association Index* (AAI) to Selikoff's published research data, in order to elucidate some additional aspects of his original findings. The current article will focus on analysis based only on the availability of the row and column totals.

The issue of how useful row and column totals of a 2x2 table are for understanding an association between categorical variables has long been debated - indeed, from a statistical perspective it has been a topic of discussion for almost a century.⁷⁻⁹ More recent work has demonstrated that marginal information can be used to detect the association structure that exists within cells of a 2x2 table,^{10, 11} thereby making it a potentially valuable tool in occupational epidemiology - a field where certain aspects of one's dataset are often incomplete or missing.

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

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

*Adapted from Selikoff (1981)⁴ published in the *Bulletin of the New York Academy of Medicine* (used with permission), **As clinically diagnosed, ***Years of occupational exposure to asbestos

Since we are interested in whether individuals exposed to asbestos for at least 20 years are more likely to contract asbestosis than those who were exposed for less than 20 years, one must define P_1 as the proportion of those workers who contracted asbestosis given that they were exposed to the fibre for at least 20 years. In particular, when considering Selikoff's data from Table 1, the value of P_1 can be calculated as: $339/392=0.86$. Similarly, we may be interested in the likelihood of being diagnosed with asbestosis in a worker who was exposed to asbestos for less than 20 years. By doing so, we can define P_2 to be the proportion of those people who were diagnosed with asbestosis (at any level of severity) given that they were exposed for less than 20 years. For example, from Table 1, we can calculate that P_2 is $203/725=0.28$. Therefore, based on the data collected by Selikoff, a worker who was exposed to asbestos for at least 20 years is $0.86/0.28=3.07$ times more likely to be diagnosed with asbestosis than a worker who was exposed for less than 20 years. Such figures provide compelling evidence for Selikoff's '20 year rule', although one must remain cognisant of the fact that such a conclusion is based on his single sample of 1117 individuals.

Understanding the potential relationship between onset of exposure and whether a worker is subsequently diagnosed with disease represents one of the cornerstones of occupational epidemiology.¹² Despite this fact, however, the collection of data is rarely straightforward and in many practical situations, one may have knowledge of the row and column totals of a simple table such as Table 1, but have little or no information on the value of the cells. This may be because the data were not recorded at the time the study was taken, or because the data were not made public for reasons of confidentiality. Therefore, we shall investigate how helpful the row and

column totals of Table 1 are for providing us with an indication of the relationship between onset of asbestos and whether a worker is diagnosed with asbestosis.

In order to investigate this issue, we shall focus our attention on the proportion P_1 . By doing so, we centre our discussion on the proportion of workers who contract asbestosis given they have been exposed to asbestos fibres for at least 20 years. Before proceeding to the issue of analysing the association given only the row and column totals of Table 1, we shall explore some common measures of association on Selikoff's data.

Measuring Association

There are a number of different statistical techniques that can be used to analyse the association between the rows and columns of Table 1. Some of the simplest techniques include calculation of the odds ratio, correlation and chi-squared statistic.¹³ In a previous article, for example⁵ we described an analysis of Table 1 from the point of view of the odds ratio, thereby providing insight into the calculation and the interpretation of simple statistical measures. Alternatively, graphical techniques such as correspondence analysis may also be used to analyse Selikoff's '20 year rule'.⁶ Irrespective of the statistical analysis used to analyse the association, there are a few key questions to consider:

1. Is there sufficient evidence of a statistically significant association between the onset of exposure to asbestos and subsequent diagnosis of asbestosis?

2. If a statistically significant association does exist between the two, how is the association best quantified?
3. Does a positive association exist, such that being diagnosed with asbestosis is related to *long-term* (>20 years) exposure to asbestos?*
4. If a statistically significant association does exist, can we visualise the nature of this association?

(*Perhaps a negative association exists, suggesting that asbestosis is more likely to be diagnosed among workers with less than 20 years exposure, although such a situation appears unlikely given the brief analysis conducted above)

The first question can be answered by considering a number of different techniques, such as calculating the odds ratio of Table 1 and its chi-squared statistic. In a previous article, we described how the odds ratio of Table 1 can be calculated as: $(339 \times 522) / (203 \times 53) = 16.45$.⁵ This suggests that a person who has been exposed to asbestos for 20 years or more is 16.45 times more likely to contract asbestosis than those who have been exposed for less than 20 years. Taking the logarithm of this quantity gives a log-odds ratio of 2.8, thus confirming the existence of a positive association between exposure of asbestos and being diagnosed with asbestosis.

However, the odds ratio we have calculated does not allow us to infer the odds ratio at the population level, only what is observed in our sample. If we chose another sample, the odds ratio is very likely to be different and the variation in the odds ratio needs to be taken into consideration. This may be done by determining, for example, the odds ratio's 95% confidence interval. To do so, we must first consider the log-odds ratio which is calculated by finding the (natural) logarithm of the odds ratio. It is

from this quantity that the 95% confidence interval of the odds ratio can be obtained, which in this case is: 11.82 – 22.87. Therefore, when the true odds ratio is unknown, we can be 95% confident that it will lie somewhere between 11.82 and 22.87. Since such a confidence interval does not contain 1 (which would indicate a zero association), there is still very strong evidence to suggest that the association is statistically significant. Furthermore, since the interval contains values greater than 1, this provides evidence of a statistically significant *positive* association. That is, we can conclude that, based on the data collected by Selikoff, workers who have been exposed to asbestos for at least 20 years are more likely to be diagnosed with asbestosis when compared to those who have been exposed for less than 20 years. Thereby providing more evidence supporting Selikoff's aforementioned '20 year rule'.

Alternatively, one may examine the statistical significance of the variables of interest by calculating the chi-squared statistic. Doing so gives a statistic of 348.35, with a p-value of <0.001. Therefore, at any level of significance less than 0.1%, there is sufficient evidence to suggest a statistically significant association between the number of years that a worker is exposed to asbestos, and whether or not they are likely to be diagnosed with asbestosis. While the chi-squared statistic provides a formal means of quantifying the statistical significance of the association, unlike the odds ratio, it does not elucidate whether the association is positive or negative. Whether a positive or negative association exists may be established by considering the correlation between the variables. A correlation value between -1 and 0 implies that there is a negative association. Similarly, a correlation value between (but not including) 0 and 1 indicates that a positive association exists. Note that when the correlation is 0, no association exists and the rows and columns of Table 1 are

statistically independent. Analysis of the data in Table 1 gives a correlation value of 0.56 indicating the existence of a positive association between the onset of exposure to asbestos and a subsequent diagnosis of asbestosis.

Analysing the Aggregate Data

The analysis of aggregate data can be very useful for occupational epidemiologists in situations where data is incomplete, such as when information regarding workplace exposure is missing, unreliable or incomplete. Suppose that for the sample of 1117 workers, the only data available was that 392 individuals had been exposed to asbestos for at least 20 years, and 542 had been diagnosed with asbestosis. Given that we know the total sample size, it can be deduced that 725 workers were exposed to asbestos for less than 20 years and that 575 workers did not contract asbestosis. Thus, we can now consider a situation where the data is missing with regard to how many workers were exposed to asbestos for at least 20 years *and* had been diagnosed with asbestosis. Table 2 provides a summary of such a situation.

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

Onset of Exposure	Asbestosis**		Total
	Yes	No	
20+ years***	?	?	392
0-19 years***	?	?	725
Total	542	575	1117

*Adapted from Selikoff (1981)⁴ published in the *Bulletin of the New York Academy of Medicine* (used with permission), **As clinically diagnosed, ***Years of occupational exposure to asbestos

When only the row and column totals of Table 1 (or Table 2) are known, statistical calculations described in the previous section can no longer be undertaken as they rely on knowing the cell counts: 339, 53, 203 and 522. In this case, we no longer have cell, or individual, level data to analyse - all that is available is the *aggregate data* of Table 1. Such a case is likely to arise in many practical situations in EOH. For example, many agencies that collect data do not release information at the individual level, in order to preserve the confidentiality of those who participated in the study. Instead, data is often provided at an aggregate level, for example in a local government area (such as a post/zip code) or an institutional group (by workplace, school), and so on.

Aggregate level data may also only be available because, at the time the study was conducted, a limited amount of information was collected. This may be due to budget or other practical constraints, or simply because such information was not considered important at the time. Not collecting such data may also be the result of an oversight when designing the data collection process utilised in the study, or simply that some of the data that was collected, was later shown to be incomplete or unreliable. As a result, when only aggregate data is available, identifying the nature of the association between the row and column categories becomes increasingly important.

When considering Selikoff's original data from New York study, the challenge is then to implement statistical techniques to help elucidate the association between onset of exposure to asbestos and subsequent asbestosis – if only aggregate data was

available. For a single table such as we are considering in this paper, the issue is longstanding and often unresolved. However, for stratified data of the form provided in Table 1 (for example, if Selikoff had collected the same information throughout New York, or even nationwide), a suite of mathematically sophisticated techniques are available. Such tools lie within the realm of ecological inference and have been previously studied with mathematical rigour.¹⁴⁻¹⁷

In the case of Table 2, where only the row and column totals are known, the proportion P_1 is unknown. However, since P_1 is a proportion, we know it must lie between 0 and 1. Thus, for each possible value of P_1 that lies between 0 and 1, the chi-squared statistic, the odds ratio and the log-odds ratio can all be calculated. Since, in this case, the magnitude of P_1 is not known, Figures 1-3 provide a graphical indication of changes in the p-value of the chi-squared statistic, odds ratio and log-odds ratio, respectively, as P_1 ranges from 0 to 1. Figure 1 indicates that for values of P_1 that lie between 0.446 and 0.526, the p-value of the chi-squared statistic is large, and certainly larger than the nominal 0.05 value that is often used. Outside this interval, the p-value appears to be very small, thereby providing an indication of when it is likely that an association exists between the variables in Table 2.

Figure 1 Graphical representation of p-values across the range of P_1

Figure 2 provides an indication of the change in the odds ratio for P_1 ranging from 0 to 1, indicating that the odds ratio increases as P_1 increases. That is, for workers exposed to asbestos for at least 20 years, the more individuals who contract asbestosis, the stronger the positive association between the variables. Note that a P_1 value of 0.86 (such as that in Table 1) coincides with an odds ratio of approximately 16.5.

Figure 2 Graphical representation of odds ratios across the range of P_1

When compared with Figure 2, Figure 3 more clearly depicts the association that exists between the rows and columns of Table 1. This is because Figure 2 graphically demonstrates that the odds ratio can vary from 0 to infinity; a negative association exists when the odds ratio is between 0 and 1, while a positive association exists when the odds ratio is greater than 1. By considering Figure 3, the log-odds ratio provides a more intuitive view of the association; a negative log-odds ratio means that there exists a negative association between the rows and columns while a positive association arises when the log-odds ratio is positive.

Figure 3 Graphical representation of log-odds ratios across the range of P_1

Although in the current article, we have considered a P_1 ranging from 0 to 1, It is important to remain cognisant of the boundary proposed by Duncan and Davis¹⁸ which, for many practical situations, provides an interval of P_1 that is narrower than 0 to 1. However, for the current analysis of Selikoff's data in Table 2, the Duncan and Davis¹⁸ range for P_1 is 0 to 1.

The Aggregate Association Index (AAI)

Recall that the significance of the association between the two variables of Table 1 can be assessed by considering the chi-squared statistic (348.35 in the case of Table 1) and a chosen level of significance with which to determine the relative magnitude of the p-value. Typically a 5% level of significance (<0.05) is used, although larger or smaller values may be considered based on practical issues concerning the analysis under consideration.¹⁹

However, when cell values are unknown (such as in Table 2), the chi-squared statistic changes as the value of P_1 changes. In fact, the minimum possible chi-squared statistic is 0 (indicating with no association) and there is a quadratic relationship between the chi-squared statistic and P_1 . It can be demonstrated that a chi-squared value of 0 will be obtained when P_1 is $542/1117=0.485$. Similarly, it can be demonstrated and confirmed^{10, 17} that any value of P_1 smaller than 0.485 reflects a negative association between onset of exposure to asbestos and subsequent diagnosis of asbestosis. On the other hand, a value of P_1 that is greater than 0.485 indicates a positive association between the variables.

As a result, depending on how big or small the proportion of P_1 is, this will impact on the magnitude of the chi-squared statistic, and therefore, the significance and direction of the association within Table 2. This property may be exploited by considering a new measure of association derived when only aggregate data is available. This new measure, originally proposed in 2008¹⁰ and expanded in 2010,¹¹ has been termed the AAI.

When considering Selikoff's data, the AAI takes into consideration the strength of the association (as measured using the chi-squared statistic at the 5% level of significance) as P_1 ranges from 0 to 1. An AAI of 0 indicates that there is no information in the row and column totals of Table 1 with which to conclude that an association exists at a particular level of significance. Alternatively, an AAI of close to 100 indicates considerable information within the row and column totals from which one may conclude that an association exists at a chosen level of significance.

In the current analysis, if we analyse the row and column totals of Selikoff's data (in Table 2) at the 5% level of significance, the AAI can be calculated to be 98.20. Thereby providing strong evidence to suggest an association between exposure to asbestos and whether a worker subsequently contracted asbestosis (when testing the association at a 5% level of significance). When testing the association between the row and column categories of Table 1, any chi-squared value that exceeds the critical value at the 5% level of significance (in our case, this critical value is 3.84) indicates a statistically significant association in the data. However, as previously described, a change in the proportion P_1 will also change the magnitude of the chi-square statistic - indeed, some statistics will be greater than 3.84 and some will be smaller. The AAI reflects the area that lies under the curve representing the relationship between the chi-squared statistic and P_1 , but above the critical value of 3.84. Thus, it can be demonstrated graphically that a statistically significant association between onset of exposure to asbestos and subsequent diagnosis of asbestosis will exist for just about all values of P_1 (98.5% of them, in fact). Refer to Figure 4.

Figure 4 A visual representation of the *Aggregate Association Index* (AAI)

Considering that a positive association between onset of exposure of asbestos and the diagnostic status of asbestosis arises when P_1 is greater than 0.485 (while a negative association arises when P_1 is less than 0.485), Figure 4 suggests that it is (slightly) more likely that there is a positive association. In fact, the AAI can be partitioned into a ‘positive association component’ and a ‘negative association’ component. For the aggregate data from Selikoff’s study, the positive association component of the index is 53.5, while the negative association component is 44.7 (note that $53.5 + 44.7 = 98.2$, the AAI of Table 2). This indicates therefore, that if there was in fact a statistically significant association, it is more likely to be positive than negative. Thus, confirming Selikoff’s original ‘20-year rule’.

Conclusion

In the current paper we have demonstrated the application of a new statistical measure, the AAI, which can help elucidate associations between variables when only aggregate data is available. To do so we reanalyse some data from a classic exposure versus risk study originally published by Irving Selikoff in the 1980s. One of

Selikoff's original key findings was that a worker exposed to asbestos for at least 20 years is more likely to be diagnosed with asbestosis when compared to a worker exposed for less than 20 years - what became known as the '20 year rule'. In our current analysis, we have demonstrated that, when the cell values are known, the association between onset of exposure to asbestos and the diagnostic status of asbestosis is positive, thereby statistically confirming the significance of Selikoff's original postulate using modern techniques. Similarly, and somewhat pioneeringly, we have also demonstrated that even if these cell values were *not* known, the 20 year rule still holds true.

Acknowledgements

The authors are grateful to Glenn Courtenay from the University of Newcastle in Australia, for his assistance in editing an early draft of this article.

References

1. Smith DR, Beh EJ. Asbestos kills - no matter how you cut the data. *Arch Environ Occup Health*. 2012;67:(in press).
2. Guidotti TL. Why study asbestos? *Arch Environ Occup Health*. 2008;63:99-100.
3. Smith DR, Leggat PA. 24 years of pneumoconiosis mortality surveillance in Australia. *J Occup Health*. 2006;48:309-13.
4. Selikoff IJ. Household risks with inorganic fibers. *Bull N Y Acad Med*. 1981;57:947-61.

5. Beh EJ, Smith DR. Real world occupational epidemiology, part 1: odds ratios, relative risk, and asbestosis. *Archives of Environmental & Occupational Health*. 2011;66:119-123.
6. Beh EJ, Smith DR. Real world occupational epidemiology, part 2: a visual interpretation of statistical significance. *Arch Environ Occup Health*. 2011;66:245-8.
7. Fisher RA. The logic of inductive inference. *Journal of the Royal Statistical Society*. 1935;98:39-82.
8. Plackett RL. The marginal totals of a 2x2 table. *Biometrika*. 1977;64:37-42.
9. Yates F. Test of significance for 2x2 contingency tables. *J R Stat Soc Ser A*. 1984;147:426-463.
10. Beh EJ. Correspondence analysis of aggregate data: The 2x2 table. *Journal of Statistical Planning and Inference*. 2008;138:2941-2952.
11. Beh EJ. The aggregate association index. *Computational Statistics & Data Analysis*. 2010;54:1570-1580.
12. 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.
13. Smith DR, Beh EJ. Hirayama, passive smoking and lung cancer: 30 years on and the numbers still don't lie. *Public Health*. 2011;125:179-81.
14. Hudson IL, Moore L, Beh EJ, Steel DG. Ecological inference techniques: an empirical evaluation using data describing gender and voter turnout at New Zealand elections, 1893–1919. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2010;173:185-213.
15. King G, *A Solution to the Ecological Inference Problem*. 1997, Princeton: Princeton University Press. 342.

16. King G, Rosen O, Tanner MA, *Ecological Inference: New Methodological Strategies*. 2004, Cambridge: Cambridge University Press. 421.
17. Wakefield J. Ecological inference for 2x2 tables. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2004;167:385-445.
18. Duncan OD, Davis B. An alternative to ecological correlation. *American Sociological Review*. 1953;18:665-666.
19. Smith DR, Beh EJ. Occupational epidemiology in the real world: Irving Selikoff, odds ratios, and asbestosis. *Archives of Environmental & Occupational Health*. 2011;66:63-64.