1

What is the Minimum Risk that can be Estimated from an Epidemiology Study?

Mark J. Nicolich[1] and John F. Gamble[2]
[1]*Cogimet, Lambertville, NJ*
[2]*Somerset, NJ*
USA

1. Introduction

Risk is commonly thought of as related to the chance of suffering a loss. In the following discussion we formally define risk and show that epidemiology and risk assessment are related to the probability that a specified event occurs. The risk will be in the form of a *relative risk* where the probability of an event occurring under specified conditions is expressed as a multiplier for the probability of the event occurring at some defined 'background level.'

The goal of the paper is to determine if there is a smallest relative risk that can be determined. That is, is there a risk level at which we can say 'An estimated risk less than X cannot be considered different from background level no matter what estimation methods are used or what the estimated statistical significance level is.' We are assuming the estimated risk level is based on data from an epidemiological study such as the health effects of ambient air pollution or arsenic in the public drinking water. We assume that there is a True Risk Level that applies to a defined population and we wish to estimate it based on the data from an epidemiology study. We assume the data are statistically analyzed either by simple ratios of the observed data or by model-based Bayesian or Classical statistical methods. There are specific epidemiological and statistical considerations that are needed to develop the formal estimate of the true risk as diagramed in Figure 1.

Figure 1 indicates the flow is from a true, but unknown, risk to a final estimate of the risk. It could have been shown as a flow from the initial hypothesis, through the design of the experiment, the analysis of the data, and the final estimate of the true risk. Either structure can be used, but we have chosen the method shown in Figure 1.

In practice the statistical and epidemiological considerations are not separate and distinct steps in the estimation process, but are shown and discussed separately for convenience. The following sections will cover:

- Basic Conclusions,
- Meaning of relative risk (RR),
- Epidemiological considerations in estimating a RR,
- Statistical considerations in estimating a RR,
- Discussion of how to answer the question "what is the smallest relative risk that can be determined"?
- Conclusion and summary,
- An appendix of quotes from knowledgeable researchers in the field

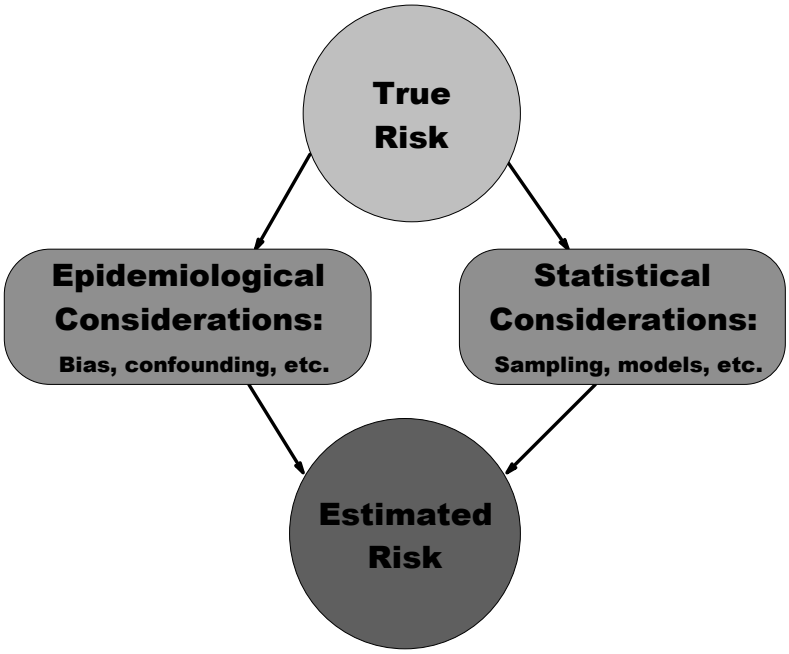


Fig. 1. Considerations needed to develop the point estimate and the significance level of an estimated risk.

2. Basic conclusions

There is no minimum estimable risk in the sense that there is no minimum measure of length. To accurately and precisely measure risk, or length, one needs a tool that is accurate and precise enough for the measurement being made. For example, a meterstick is inadequate for measuring length to a precision of 0.1 mm. When measuring risk in epidemiology, precision and accuracy depend on many factors including the degree to which the experimental data and analysis are free of the errors associated with experimental design flaws (i.e. confounding, bias, reliability, measurement error and misclassification, potential alternate risk factors), and statistical analysis errors (i.e. sampling error, violations of model assumptions, model shopping, multiple subgroup analyses, reliance on significance levels or *p*-values). Indeed, it is theoretically possible to correctly estimate significant risks near 1.000001 if there were no confounding, no biases, no measurement error, exact model specifications, etc. But it is unlikely, if not impossible, these conditions ever exist for an epidemiology study.

While there is no 'lower bound' to an estimable risk, and while many experienced researchers have made opinions about approximate lower bounds, a practical lower limit does exist. To believe an estimated risk the researchers must demonstrate that they have carefully considered the array of possible errors that can arise from the tools and techniques used. As important as these demonstrations are, the difficulty of actually performing these demonstrations leaves many questions about the 'true risk,'and at some point the demonstration may become

infeasible or meaningless and leave the realm of conventional science. As exposures approach background levels and apparent risks approach 1.0, the estimates enter the realm of 'trans-science.'When a problem is in the area of trans-science, risk assessment tools of standard science are, by definition, not useful. The most useful outcome possible may be the understanding that the problem belongs to the area of trans-science.

Epidemiology has provided society with many benefits by providing clues to the causes of infectious diseases and identifying factors that contribute to the start and development of non-communicable diseases. Causal knowledge allows quick and effective preventive measures to lessen or halt the diseases without a full understanding of the underlying mechanism. As the RR to be estimated becomes smaller the need for the analyses to be free of errors becomes greater and at some point it becomes impossible to reduce the error further.

3. Meaning of Relative Risk

We will consider epidemiologically based risk as measured by the relative risk (RR). By definition, RR is the ratio of disease frequency in exposed individuals divided by the disease frequency in unexposed individuals (or controls). RR can be thought of as the multiplier of the probability of an event occurring in a special situation relative to the probability of the event occurring in a baseline situation. For example, if the probability of developing lung cancer among smokers was 15% and among non-smokers was 3%, then the relative risk of cancer associated with smoking would be 5. Smokers would be five times as likely as non-smokers to develop lung cancer

Related to the concept of a RR is the idea of Association. Association is measured by absolute differences in rates of disease and by ratios of disease frequency.

- The difference measure of association is the absolute difference in disease frequency between exposed and unexposed. It is an estimate, in the absence of bias, of the absolute amount of disease caused by or attributable to exposure. This absolute difference is proportional to the prevalence of exposure, and so is not especially useful in assessing causality
- The ratio measure is the ratio of disease frequency between exposed and unexposed. It is an estimate of the proportional increase or excess of disease in exposed relative to unexposed and is unaffected by prevalence of exposure. The ratio measure of an association is used when determining whether there is a plausible evidence for a causal association between exposure and disease, or the etiology of disease. The evidence for a causal association is strengthened when there is a strong association, or when the relative risk (RR) is high. The larger the RR the greater the likelihood that an association is causal. Associations may be misleading, because other factors can cause apparent, but spurious, associations that are not true associations. The larger the RR the more unlikely it is that unrelated factors (e.g., bias, confounding) could overcome a true association. However, a RR cannot be used as the sole proof of causality.

The RR is derived from study designs where the study population is selected on the basis of exposure (e.g. cohort study), or disease (e.g. case-control design) and may have different names depending on the context:

- In a cohort mortality study the RR is called a standardized mortality ratio (SMR).
- In a cohort incidence study the RR is called a standardized incidence ratio (SIR).
- In a morbidity cohort or cross-sectional study the RR may be called a prevalence ratio.
- In a case-control study the RR is called an odds ratio (OR).

In a case-control study design the study population is selected on the basis of disease, and the strength of association is based on the odds of exposure (odds ratio or OR), which is the ratio of the proportion of exposed among cases divided by the proportion of exposed among non-cases. The OR is based on the assumption that the disease is rare. [There are methods for estimating risk ratios and confidence intervals from adjusted odds ratios (Greenland 2004, and King and Zeng, 2010).] In everyday applications the OR is used in a way similar to a RR for assessing causality. Differences include the fact that the risk is for exposure, not disease, as in the cohort-based RR. And there are usually differences in some of the confounders. For example, recall bias is likely to be a problem in a case-control study because cases are more likely than non-cases to remember more or exaggerate exposures related to their disease. This is not a likely confounder in a cohort study since the exposure data should be collected blind to disease status.

A weak association means the proportional excess of disease in the exposed relative to unexposed (RR) is small in magnitude, or less than about 2-fold. As associations becomes increasingly weak the importance of spurious factors (e.g., confounding, bias) increases until a point is reached where they overcome (become larger) than the effect of interest. Or in other words, as the power of the signal decreases it will eventually be unintelligible when it is overwhelmed by the noise level in the system.

In epidemiology practice 'weak' and 'strong' relate only to ratio associations and not to the difference measure of association (Rothman and Poole 1988). The strength of association for a particular agent varies over time because the estimated effect depends on the "time-specific distribution of its causal complements in the population" (Rothman and Greenland 2005), which varies over time and between populations.

Strength of association is a major guideline in determining causality (Hill 1965). But there is no agreed upon value for an association to be considered weak or strong. A less than two-fold increased risk has been generally accepted as a 'weak' association. For example, in litigation the court accepts a causal association if the connection between exposure and disease is 'more likely than not.' In epidemiology an *etiological fraction* or *attributable risk* (AR) can be calculated to estimate what proportion of a disease is attributable to exposure to a risk factor when that risk factor is believed to be a cause of the disease. The formula for this calculation expressed as a percent (x 100) is:

$$AR = 100 \bullet \frac{RR - 1}{RR} \tag{1}$$

If the RR for a risk factor of interest = 2.0, then 50% of the disease is attributed to that risk factor. Therefore RR must be greater than 2.0 for the exposure to 'have more likely than not' caused the disease in a particular individual.

4. Epidimiological considerations

In this section we consider areas dealing with study design, confounding, biases, measurement error, misclassification, reliability, and the effects of multiple risk factors.

Epidemiology is an observational, not an experimental science. Because humans are the study subjects it is not possible to manipulate or control exposure. Exposure often consists of the exposures experienced at work, as in occupational epidemiology. The investigators (and the study subjects) lack the capability of producing a specific intensity of exposure. The

duration and intensity of exposure cannot be controlled, and only the population can be defined. So there is a distribution of exposures of different intensity and duration among the study population. This is in contrast to the experimental design of animal studies where a desired exposure intensity and duration are produced for a selected number of animals selected at random from the same population.

In epidemiology, estimates of uncertainty are limited by real world limitations. The investigator in an industrial exposure study can increase the size of the study population by including more work sites (or cases with disease), but the total size of the study population is often limited.The investigator has no control regarding intensity and duration of exposures in the workplace, and exposures are generally not to a pure substance but to a mixture of different chemicals that varies over time. Also the investigator has no control on the presence of potential lifestyle confounders (e.g., smoking, obesity, alcohol, age, sex). It is possible to statistically account for confounders, but adjustments will be incomplete because a) the risk associated with each confounder is an imprecise estimate; and b) because there are some unknown confounders for which no statistical adjustment is possible. Note that we use the term 'account for confounders' or 'adjust for confounders' instead of the more popular 'correct for confounders' because an additional term in a statistical model can not *correct* for the presence of confounders, but can only apply a term that might account for the confounder if the adjustment term is adequate.

All estimates of associations in epidemiology are uncertain because of "unknown levels of systematic error from measurement, uncorrected confounding, selection bias, and other biases" that "may even dwarf the random sampling error" (Phillips 2003). The significance of random sampling error is estimated via *p*-values or confidence intervals. But statistical uncertainty tells us nothing about hidden uncertainties with regard to bias and confounding.

Measurement error is a major problem in evaluating causality in epidemiology. Rothman and Greenland (2005) argue that measurement error includes all errors, which range from statistical error to uncertainty problems relating to study design (such as subject selection and retention) through uncontrolled confounding and all other forms of bias. The "true risk"is unknown and is always an estimate because every study has some type of error. "The real issue is to quantify the errors. As there is no precise cutoff with respect to how much error can be tolerated before a study must be considered invalid, there is no alternative to the quantification of study errors to the extent possible." Whether or not one is assessing causality via a set of guidelines such as those of Hill, (Hill, 1965) one must estimate the "total error that afflicts the study" to obtain a reliable estimate of risk. This part of the project further defines and explains this goal of quantifying study errors but excludes consideration of statistical errors.

What are the guidelines for estimating reliability? A minimally reliable RR must be the same order of magnitude or greater than the maximum risks from unknown confounding and bias. The magnitude of the unknown confounding should be less than the minimum known confounding, assuming the investigator is rigorous in the design of the study.

Taking statistical significance into account, guidelines for reliability of a RR are outlined in Table 1.

4.1 Minimal reliable RR

There are several points that define a minimal reliable RR (MRRR).

	Confounding	Observational Study Results	Experimental (clinical) Study Results
Statistically significant	NOT controlled	Unreliable	Unreliable
	Controlled	Possible Real Effect	More likely Real Effect
Not statistically Significant; RR > Minimal RR from hidden uncertainties	NOT Controlled	More Unreliable	More Unreliable
	Controlled	Possible Real Effect; Not significant because of small n?	Unreliable
Not Statistically Significant; RR < minimal RR from hidden uncertainties	NOT controlled	Very Unreliable- likely no relation	Very likely no association
	Controlled	Likely No association	Very likely no association

Table 1. Guidelines for Reliability of a RR

4.1.1 Point 1: There is no single number for a minimal reliable risk that pertains to all studies

The strength of association and magnitude of the uncertainty (and therefore the MRRR) can vary from study to study because of the influence from several sources such as: study population, the causal components of the disease, the frequency and magnitude of the biases, measurement error, etc. MRRR is not dependent on sample size, and is therefore independent of statistical uncertainty or the *p*-value.

As an example of this point, Maldonado et al. (2003) conducted sensitivity analyses on four studies investigating relationships between occupational exposures to glycol ether and congenital malformations. They specifically examined methodological errors that produced biased results that could have incorrectly increased RRs from 24% to 300%. Unfortunately, "without information that can be used to judge the plausibility of different error-producing scenarios, we cannot know the true impact of error on study results….currently available evidence does not refute the [alternative hypothesis] conjecture that elevated OR estimates are due to error." (Maldonado, et al. 2003)

Study Example 1: (Cordier, Bergeret et al. 1997)

This is a case-control study of glycol ether and congenital malformations with multiple positive associations that were substantively modified by adjustments for potential biases. They reported

- Overall OR = 1.4 (1.1-1.9) for all malformations combined.
- Elevated ORs ranging from 1.3 to 2.0 for all eight major categories of malformations
- Elevated ORs ranging from 1.2 to 2.6 for 17/18 subcategories of malformations
- Most likely causal association was for cleft lip with an OR of 2.51 and monotonic E-R trend with ORs of 1.0, 1.5, 2.2, and 2.9 with increasing exposures from no to certain exposure.

These results were consistent with previous studies showing similar associations. Are these risks greater than a MRRR? Are the risks strong enough to support a causal association and greater than hidden uncertainties such as bias? Sensitivity analyses address these questions.

- Potential selection bias from unreported abortion rates differing between exposed and controls could produce errors estimated to range from 0.85 to 1.17. A 3% difference was calculated to produce a spurious OR of about 1.1.
- Potential selection bias from case non-response may have produced error factors ranging from 0.68 to 1.61. If 50% of identified cases were not included in the study the error factor was calculated to be 1.03.
- Potential selection bias may occur if controls are selected differently than cases and are from different populations, as was the case in this study. They suggest the error factor might range from 0.50 to 2.0, and the original authors needed to assess effects of this bias.
- Potential information bias due to exposure measurement error was inadequately presented so the magnitude of the error factor could not be reliably estimated. Maldonado et al. (2003) suggested information bias might produce error factors ranging from 1.4 to 2.3 for exposure misclassification effects related to cleft lip.
- Potential biases in combination are estimated by multiplication of the error factors, which in this case was an error factor of 2.47 (= 1.1 x 1.03 x1.09 x 2). With crude ORs of 2.51 (for cleft lip) and an error factor of 2.47, the corrected OR is 2.51/2.47, or 1.02. Cordier et al. (1997) adjusted for other biases and estimated an error factor of 1.24 (which changed the crude OR of 2.51 to 2.03). Combining the error factors of Maldonado et al. (2003) and Cordier et al. (1997) (2.47 x 1.24 = 3.06) changes the original crude OR of 2.51 to an adjusted OR of 0.82.

Similar sensitivity analyses of three reproductive studies are summarized in Table 2.

Types of Bias	(Cordier, Bergeret et al. 1997) Error Factor	(Shaw, Velie et al. 1999) Error Factor	(Cordier, Szabova et al. 2001) Error Factor
Due to losses	1.1	0.91	1.03
Case non-response	1.03	0.89	0.96
Non-random Control selection	2.0	0.50	2.0
Exposure measurement error	1.09	0.63	1.83
Subtotal of scenarios	2.46	0.26	3.62
Confounding	Largely unknown	Largely unknown	Largely unknown
Author adjustments	1.24	None	None
EF: Bias in combination	2.46 x 1.24 =3.06	0.26	3.62
Adjusted ORs	2.50 / 3.06 → 0.82	0.93 / 0.26 →3.58	3.4 / 3.62 → 0.94

Table 2. Summary of potential error factors (EF) due to bias and confounding in three studies of occupational exposure to glycol ethers and congenital malformations. Based on plausible scenarios (Maldonado, Delzell et al. 2003)

These combined biases are quite large and indicate a MRRR in these studies is far from 1.0. Reproductive studies of this type tend to have large error factors, and the minimal interpretable RRs for these 3 studies should be at least >3, <0.2 and > 3.6 before assessing associations of glycol ether and congenital malformations in these studies. Maldonado, et al. (2003) indicate that because there is a lack of investigation and information available to estimate the extent of the biases, results of these studies are basically uninterpretable.

4.1.2 Point 2: If the estimated RR is smaller than the estimated bias of one of the confounders, then the estimated RR is not reliable and there is a low probability that a causal association exists

Or in more general terms, if the estimated uncertainty or noise is greater than the estimated RR a causal association is unlikely and undetectable.

This is the situation for the results outlined in Table 2, which are based on scenarios whose plausibility is basically unknown. Until the extent of the biases is measured with more reliability, determination of a causal association will remain speculative because of the large size of the potential biases in both positive and negative directions.

There are several examples of studies where the strength of the associations is quite low and the question has arisen as to whether it is possible to measure such associations. These include dozens of studies examining associations between ambient concentrations of air pollutants (especially PM) and acute mortality and morbidity. These are mostly time-series studies where daily concentrations of pollutants in a city are correlated with the daily number of some health effect such as deaths, hospitalizations or asthma cases. Associations in these studies are quite weak with relative risks often around the null value of 1.0. For example, associations with mortality range from negative (<1.0) to a high of about 1.05. That is, the statistical model suggests that as PM concentrations increase by some amount (e.g.10 $\mu g/m^3$) there is 5% increase in mortality or morbidity. For a RR of 0.95 the model suggests a decrease in disease of 5% for every unit increase in exposure.

The primary known confounders in these studies are weather and co-pollutants. Can adequate adjustments be made of these confounders? Are the effects of these confounders greater than the effect of the pollutant of interest? If the answers are no and yes to each question respectively, it is probable the results are spurious and the PM effect cannot be distinguished from the confounding effect.

Lumley and Sheppard (Lumley and Sheppard, 2000) assessed the magnitude of biases in a time-series study of asthma admissions and ambient PM concentrations in Seattle, WA. In the original study the estimated RR was 1.03 per 10 $\mu g/m^3$ increase in PM. They then changed the exposure data (PM concentrations) to different time periods so no association would be expected (RR = 1.0). Deficits or increases in RRs under these conditions would indicate the magnitude of uncontrolled bias. Exposure data from 1 and 2 years in the future produced negligible bias. Exposure data 3 and 4 years in the future produced RRs larger than effect estimates from the real data. The authors concluded "the bias is small in absolute terms but of the same order as the estimated health impacts." Results from these types of studies where RRs are on the order of 1.03 may be measurements of bias not of effects. If so, the observed RR is spurious and can produce an incorrect and misleading interpretation.

The HEI Health Effects Committee (2003) reviewed a selected number of time-series studies after it was discovered that there were errors in the calculation of risk ratios and confidence intervals in these air pollution studies. They indicated there is *no* satisfactory way to adjust for weather and other time-varying confounders such as co-pollutants (HEI 2003).They

concluded "no strictly data-based (i.e., statistical) method can exist to insure that the amount of residual confounding due to that variable is small. This means that no matter what statistical method one uses…it is always logically possible that even if the true effect of pollution is null, the estimated effect is far from null due to confounding bias." Because the associations are so weak in these studies the effects that are being attributed to air pollution may "actually be due to…other factors, including weather (typically temperature and relative humidity), as well as unmeasured factors that also vary with time."

The Lumley and Sheppard paper suggests a minimal RR for a causal association in time-series studies must be greater than about 1.05 or the RR said to be associated with air pollution may, in actuality, be produced by unmeasured factors. A RR <1.05 should be considered too weak an association to interpret in the context of time-series air pollution studies, no matter how significant the *p*-value. And the RR may need to be even greater than that since there is no objective method to adjust for confounding from weather or time-related factors. Since weather may have a larger effect on health than PM we may be in an untenable situation with risk factors (e.g., weather, co-pollutants) having potentially greater effects than the exposure of interest.

Exposure misclassification is another unresolved problem in time-series studies. All cases in a time-series study are considered exposed to the same concentration of pollution in the ambient air. Ambient ozone does not permeate into homes, so cases may have 1% or less actual exposure than measured ozone exposure, and minimal estimated risk may be much higher than 1.05. The effect of this exposure misclassification error in this type of study should be studied and adjusted for.

4.1.3 Point 3: There are two kinds of uncertainty associated with relative risks and confounding: Known risk called RR (kc) and unknown risk called RR (uc)

Epidemiology is an observational science so there is natural variation or background uncertainty inherent in every study. Sometimes some of the effects of this uncertainty can be estimated. For example, a heavy smoker is at about a 20-fold greater risk of lung cancer than a non-smoker.

Issues of hidden uncertainty (e.g., bias, confounding) may be more important than frequentist statistical significance (Sterne and Smith 2001; Phillips 2003) and are important concerns that must be considered when evaluating possible etiological agents.

Statistical inferences (e.g., *p*-values) are based on the assumption that the statistical model is correct. However, statistical models can *never* account for all factors. Incompletely accounted for factors are subsumed under the general rubric of bias and confounding. If these factors are not accounted for in comparisons between exposed and non-exposed populations, results may be rendered invalid or unreliable.

Confounding occurs when effects of two agents are not separated and they are counted as one effect. As a result, a difference in disease rates is not due to the exposure of interest alone, but is due to effects of exposure plus other factors such as smoking, diet, etc. By definition a confounder must (1) cause the disease (or be associated with a causal risk factor); (2) be correlated with exposure, either positively or negatively; and (3) not be affected by the exposure.

Confounding arises due to a lack of comparability between exposed and unexposed groups. If both exposed and non-exposed are all non-smokers, smoking cannot be a confounder

because smoking is not associated with exposure and exposed/unexposed groups are comparable with respect to the prevalence of smoking. Confounding is an issue that must always be addressed in assessing issues of causality (McNamee 2003).

Strictly speaking there will always be differences between exposed and unexposed groups, but sometimes the differences are small and unimportant. What is important is the magnitude and direction of confounding effects on the estimated relative risk. The degree of confounding is measured as the ratio of the measured confounded RR divided by the true unconfounded RR. For a single dichotomous confounder, the degree of confounding depends on 1) the strength of the association between confounder and disease (RR); and 2) the percentage (p) of subjects in the exposed (p_1) and unexposed (p_0) groups. It is calculated from the formula:

$$\frac{confounded\,RR}{true\,RR} = \frac{(100 - p_1) + RR \bullet p_1}{(100 - p_0) + RR \bullet p_0} \tag{2}$$

For example, a confounded RR will be 1.4 times the true RR if the strength of association for the confounder is 5 and the prevalence of the confounder in exposed and unexposed populations are 50% and 30% respectively. The statistical significance between p_1 and p_0 should not be used to assess the potential importance of confounding. If there were 40 subjects in each group in this example, the difference between 50 and 30% is not significant at the 5% level. But a 1.4-fold difference between confounded and true RR is important. Table 3 illustrates the degree of confounding that can occur for a strength of association = 5 (McNamee 2003).

Unexposed	Exposed Group: prevalence of confounder (p_1)								
p_0	10%	20%	30%	40%	50%	60%	70%	80%	90%
10%		1.3	1.6	1.9	2.1	2.4	2.7	3.0	3.3
20%	0.8		1.2	1.4	1.7	1.9	2.1	2.3	2.6
30%	0.6	0.8		1.2	1.4	1.5	1.7	1.9	2.1
40%	0.5	0.7	0.8		1.2	1.3	1.5	1.6	1.8
50%	0.5	0.6	0.7	0.9		1.1	1.3	1.4	1.5
60%	0.4	0.5	0.6	0.8	0.9		1.1	1.2	1.4
70%	0.4	0.5	0.6	0.7	0.8	0.9		1.1	1.2
80%	0.3	0.4	0.5	0.6	0.7	0.8	0.9		1.1
90%	0.3	0.4	0.5	0.6	0.7	0.7	0.8	0.9	

Table 3. Degree of confounding due to differences in prevalence between exposed and unexposed when the true risk=5

The degree of confounding is affected more by differences in prevalence of a confounding factor than by strength of the associations, especially above RRs of 5 or so. For example, at prevalences of 60 vs. 40% in exposed and unexposed, the degree of confounding at RR of 2, 5, 10, 15, and 20 are 1.14, 1.31, 1.39, 1.42, and 1.44 respectively.

4.1.4 Point 4: Reliance on *p*-values is an ill-founded strategy. The veracity of a research finding is not dependent on a *p*-value and vice-versa

Too commonly results are considered conclusive on the basis of a *p*-value less than 0.05. This is an 'ill-founded strategy' because of the high rate of non-replication (lack of confirmation) from a "single study assessed by formal statistical significance' (Ioannidis 2005).Others have made the same point (Sterne and Davey Smith 2001; Wacholder, Chanock, et al. 2004). The probability that a finding is actually true when based simply on achieving statistical significance is called the positive predictive value, PPV (Ioannidis 2005). The formula for calculating PPV is

$$\text{PPV (Positive Predicted Value)} = (1-\beta)R \,/\, ((1-\beta)R + \alpha) \tag{3}$$

where
β = Type II error rate, usually 0.20
α = Type I error rate, usually 0.05
R = ratio of the number of 'true relationships' to 'no relationships' among tested hypotheses.
That is, the higher the PPV the greater the chance the estimated RR is an accurate estimate of the true risk. If was assume β = 0.20, α = 0.05, and R=1.0 (half the hypotheses that were tested were found to be statistically significant), the probability of a statistically significant RR being true is estimated to be 0.89. If R = 0.06 (as when 3 of 53 of the hypotheses being tested was statistically significant), then PPV < 0.5 and chances are better than even that the results are not true.
But bias and repeated independent testing by different investigators may further reduce the probability of a true (replicable) result. One definition of bias is 'the combination of various design, data, analysis and presentation factors that tend to produce research findings when they should not be produced' (Ioannidis 2005). This definition is analogous to publication bias *in situ (PBIS)* (Phillips 2004). It can involve manipulation of the analysis and cherry picking of the reported results. In the presence of this kind of bias the probability of a true finding is further reduced and can be estimated from the formula

$$\text{PPV} = ([1-\beta]\,R + u\beta R) \,/\, (([1-\beta]\,R + [1-\alpha]u + \alpha + u\beta R) \tag{4}$$

where u = the proportion of probed analyses that would not have been "research findings" but nevertheless end up presented and reported as such.'
For example, in air pollution studies of PM thousands of models may be investigated, only a small percentage of which are reported. Assuming both R and u are 0.5 (not unusual in studies with many test models), there is less than a 50-50 chance that a finding is actually true (PPV = 0.46).
On the basis of the above considerations, Ioannidis (2005) proposes several corollaries about the probability that a research finding is true or not.
Corollary 1: The smaller the study, the less likely the findings are true.
Sample size is related to power, and PPV 'for a true research finding decreases as power decreases towards 1 – β = 0.05.'
Corollary 2: The smaller the RR (the weaker the association) the less likely the findings are to be true.
Findings are more likely to be true for strong risk factors such as effects of smoking on cancer or cardiovascular disease (RR 3-20) than for small RRs such as genetic risk factors

for multi-genetic diseases (RR 1.1-1.5). Modern epidemiology increasingly targets smaller effect sizes as most risk factors with large RRs have been studied. As a result the proportion of true findings is expected to decrease. Similarly, if the majority of RRs are very small (e.g., RR < 1.05 as in genetic, nutritional and air pollution epidemiology), 'this field is likely to be plagued by almost ubiquitous false positive claims…[and are] largely utopian endeavors.'

Corollary 3: 'The greater the number and the lesser the selection of tested relationships' the more likely the findings will be spurious.

The higher the probability that the pre-study hypothesis is true, the more likely results of a study testing that hypothesis will be true and vice-versa. For example, meta-analyses and studies attempting to confirm hypotheses are more likely to produce true findings than hypothesis-generating studies.

Corollary 4: Increased flexibility in design, definitions, outcomes, and analytical modes increases the likelihood of spurious findings.

Adherence to common standards and protocol is likely to decrease the probability of false findings. Flexibility (e.g., multifarious outcomes such as scales for schizophrenia; experimental analytical methods such as artificial intelligence methods; reporting only 'best' results) 'increases the potential for transforming what would be "negative" results into "positive" results, i.e., bias, μ.' But cherry-picking the data and manipulating outcomes and analyses reported remains a 'common problem' even within the most stringent research designs such as randomized trials (Chan, Hrobjartsson, et al. 2004).

Corollary 5: 'The greater the financial and other interests and prejudices…[of the investigator] the less likely the research findings are to be true.'

These factors increase bias and may be common and widespread. Examples cited included prejudice because of belief in a particular theory or commitment to their own previous findings; university-based studies conducted for no other reason than to increase qualifications for promotion or tenure; peer reviews written to suppress findings that refute the reviewer's research findings. And there is evidence that expert opinion is 'extremely unreliable' (Antman, Lau et al. 1992).

Corollary 6: 'The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true.'

For example in molecular genetics there is competition between laboratories to be the first to publish positive associations (e.g., race to describe DNA). This can lead to each lab pursuing and publishing their most impressive 'positive' results, or 'negative' results to contradict positive results from another lab. This is called the 'Proteus phenomenon,' where there are rapidly alternating findings followed by extremely opposite refutations (Ioannidis and Trikalinos 2005).

Based on this reasoning, it is difficult to exceed a 50% probability that results are likely to be true. Considering effects of power, ratio of true to non-true relationships, and bias in various settings and study designs, Ioannidis (Ioannidis 2005) conducted simulations to estimate PPV. These are summarized in the Table 4.

This analysis indicates the majority of modern biomedical research operates where the pre- and post-probability of true findings is very low. In a 'null field' where true positive associations are unlikely to occur (or where RRs are very low), one 'would ideally expect all observed effect sizes to vary by chance around the null in the absence of bias. The extent that observed findings deviate from what is expected by chance alone would be simply a

pure measure of the prevailing bias.' For example, suppose 60 nutrients had been examined for causing a specific cancer and RRs were in the range 1.2-1.4 comparing high to low exposed groups. 'Then the claimed effect sizes are simply measuring nothing else but the net bias that has been involved in the generation of this scientific literature.'

1-β	R	μ	Practical Example	PPV
0.80	1:1	0.10	Adequately powered randomized control trial with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good quality randomized control trials	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II randomized control trials	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II randomized control trials	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

Table 4. Positive predictive Value (PPV) of research findings for various combinations of power (1-β), ratio of true to not-true relationships (R), and bias (μ) (Ioannidis 2005)

4.1.5 Point 5: Multiple exposures are a special case of confounding. Research studies investigating a single exposure in observational studies can lead to biased estimates of risk because there is no accounting for exposure to other risk factors that are always present

Exposure measurement errors and confounding are problem areas when studying effects of exposure to multiple chemical agents, and become more difficult when exposure to physical, biological, and psychological stressors are considered. In many epidemiology studies, exposure is limited to one, or at most two, agents in any analysis even though in most situations humans are exposed to chemical mixtures instead of a single chemical (Feron et al., 2002) as well as a wide variety of other stressors. Biological considerations have shown that exposure to some mixtures of chemical compounds can result in lower toxicity (antagonism) or higher toxicity (synergism) than expected based on knowledge of the potency of the individual components (Calabrese, 1991). There are many examples of these antagonistic and synergistic interactions in the literature (Chang et al., 2005).

A classic description of a set of confounders occurs when there are several risk factors, each of which engenders the same response as the agent under analysis and whose magnitudes co-vary. Confounding effects can be accounted for if it is known that these risk factors are present in the environment and their interaction patterns are known. Since it is rare that the presence and interaction patterns are well known, observed risk may be an under- or over-

estimate of the 'true risk." This problem is clearly exacerbated when the underlying risk is near unity.

4.2 Summary of the epidemiologically based points

It is difficult to account for all the pitfalls associated with designing an epidemiology study where potential risk is very low. The following quote (Ioannidis 2005) suggests the magnitude of the difficulties inherent in the air pollution field:

'For fields with very low PPV, the few true relationships would not distort this overall picture much. Even if a few relationships are true, the shape of the distribution of the observed effects would still yield a clear measure of the biases involved in the field. This concept totally reverses the way we view scientific results. Traditionally, investigators have viewed large and highly significant effects with excitement, as signs of important discoveries. Too large and too highly significant effects may actually be more likely to be signs of large bias in most fields of modern research. They could lead investigators to careful critical thinking about what might have gone wrong with their data, analyses, and results.'

Ioannidis (2005) suggest several steps be taken to improve post-study probability of reporting true findings and avoiding spurious interpretations:

i. Look at the totality of the evidence.
ii. Enhanced research standards and curtailment of prejudices may help. Part of this process is through 'developing and adhering to a protocol' in a fashion similar to that followed in randomized clinical trials.
iii. Improve your 'understanding of the range of R values-- the pre-study odds—where research efforts operate….[I]nvestigators should consider what they believe the chances are that they are testing a true rather than a non-true relationship.'
iv. Don't rely on significance as the primary basis for interpreting results. '[S]tatistical significance testing in the report of a single study gives only a partial picture, without knowing how much testing has been done outside the report and in the relevant field at large…[U]sually it is impossible to decipher how much data dredging…has preceded a reported research finding.'

5. Statistical considerations

Statistical considerations are not independent or exclusive from epidemiological considerations. The distinction provides a possible separation between the strictly epidemiological steps of study planning and sample selection and the statistical steps of data analysis. Both areas come into play in the critical interpretation and reporting steps of the study. In this section we will concentrate mainly on the ideas of violation of model assumptions, model shopping and multiple subgroup analyses, and reliance on significance levels or *p*-values.

5.1 Violation of model assumptions

Relative risks from epidemiological studies often are derived from complex statistical models based on statistical assumptions that are not always clearly understood by non-statisticians but which need to be met to reduce potential errors. The choice of a final model

in the model-building process is important because when the RR is small one cannot always verify whether the estimated risks are consistent with the graphic or tabular data. If statistical assumptions are not met the validity of the results is uncertain. We would like to be sure that an appropriate statistical model has been chosen and important assumptions have been met for applying the model to the data.

There is a wide array of concerns relating to the final model choice. Among them is the form of the model including questions such as:

- Is the risk assumed to be linear with exposure?
- Does the model admit to or allow for a threshold or a ceiling?
- Are confounders accounted for?

Other important concerns include

- Have the distributional assumptions of the model been met? For example, do counts follow a Poisson distribution, or are they better described by a negative binomial, or are the residuals normally distributed?
- Are observations independent of each other? Are the residuals free of any auto-correlation?
- Has the appropriate model estimation technique been applied, such as least squares, GAM, GLM?

5.2 Model shopping

The widespread availability of statistical and computing technology is an important factor contributing to the potential for estimating RR with unrealistic precision. It is now easy to routinely engage in sophisticated optimizations across a large number of models and/or variables to identify associations of potential scientific interest. Even with a single risk factor and a single response, it has become standard practice to consider a potentially large number of models in an effort to adjust for differences among the exposed and the unexposed (Peng, et al., 2006). This phenomenon is often called 'model-shopping.' It is an underlying assumption that the significance level of an estimate is developed from a model that was specified *before* statistical analyses were performed. In practice, models often are modified in ways that violate this basic assumption of a completely pre-specified model in order to maximize model efficacy (or maximize the ability to produce a desired result). These violations include such acts as choosing different forms of background effects, selecting smoothing parameters, or choosing different lags for explanatory variables. Such 'model shopping' produces a spuriously inflated significance level, or narrowed confidence interval, that often overstates the significance of the predictors unless there is some adjustment. Hodges (1987) pointed out that reporting only the 'best' model result and essentially ignoring the uncertainties associated with model assumptions may lead to overconfident predictions and policy decisions that are riskier and more uncertain than one thinks they are.

The degree of overstating is related to the number of models tested. Chatfield commented "It is indeed strange that we [statisticians] often admit model uncertainty by searching for the best model but then ignore this uncertainty by making inferences and predictions as if certain that the best fitting model is actually true" (Chatfield 1995).

Bayesian model averaging (BMA) is one method that can help eliminate the concern of multiple models and provide more realistic estimates of uncertainty of relative risks (Clyde

2000). BMA works on the principal that it is possible to calculate the Bayes probability that a model is the correct model for a given data set. After considering all possible models one can estimate a common parameter of interest and its standard deviation and take into account multiple testing. The parameter of interest in many of our examples that follow is the risk associated with an increase in PM concentration.

A cruder method to deal with problems of multiple model testing is to change the criteria for significance, as for example from $p < 0.05$ to $p < 0.005$. This method was suggested by the HEI Health Review Committee (2003) for revised analyses of ACS and Six Cities cohort studies (Krewski, Burnett et al. 2000).

Neither of these methods (BMA or changing the level at which significance is declared) has been universally applied, so the concern remains about minimizing the reported error in a RR associated with multiple testing. Note Hill's advice when interpreting for causality: don't over-emphasize statistical significance tests, as systematic error is often greater than random error. He questions the usefulness of statistical significance in situations where differences are negligible. What could be worse than when the "glitter of the *t* table diverts attention from the inadequacies of the fare" (Hill 1965).

5.3 Multiple subgroup analyses

Closely related to the phenomenon of model shopping is the analysis of subgroups of the original sample. Subgroups are often formed by 'interesting observations' seen during the initial analyses and can be useful and lead to important discoveries. "Nearly everything that has been learned in epidemiology has been derived from the analysis of subgroups. This is an incisive, effective technique to which we owe our sustenance." (Stallones, 1987). When subgroups are not selected *a priori* and are not part of the initial sample size determination, results from sub-group analyses must be looked on with caution and replications in a different experiment or sample are needed before any conclusions can be made.

5.4 Over-reliance on significance levels

It is a common misconception that the precision of a small RR depends on the significance level or *p*-value of an estimated RR. It is not true that the *p*-value or confidence interval "can provide a number that by itself reflects a probability of reaching erroneous conclusions." (Goodman 1999a)

The *p*-value is an informal index for measuring the difference between results of a study and the null hypothesis of no effect. A *p*-value of 0.05 does *not* mean that the null hypothesis has a probability of 5% of 'being true' (or that there is a 95% chance of no effect). This interpretation is incorrect because the study data alone cannot provide the probability that the null hypothesis is true. Tests based on probability (such as *p*-value and 95% confidence intervals) cannot provide "any valuable evidence of the truth or falsehood of a hypothesis" (Neyman and Pearson 1933; Goodman 1999a). The *p*-value *only* tells whether the results are statistically significant (Goodman 1999a).

The strength of evidence for an effect estimate with a *p*-value of 0.05 is actually much weaker than the number 0.05 suggests. Goodman (Goodman 1999b) suggests using the Bayes Factor in lieu of the *p*-value. The Bayes Factor, related to the likelihood ratio, is an index for evaluating evidence outside of individual study results that can be used for assessing the likelihood that a relationship is true. The Bayes Factor compares the predictive value of the null and alternative hypotheses based on the weight of evidence. It can help

keep chance (statistical evidence) distinct from conclusions, while being part of a calculus that formally links them (Goodman 1999b).

"We must expect that most truly new research findings in observational research will be unexpected and an unexpected result may well appear aberrant. To distinguish between those exceptional subgroups that have epidemiological significance and those that do not may be difficult. Statistical treatment is of little use in addressing the problem. A probability statement is of value in assessing the magnitude of an association in relation to the number of persons in the study; that is, it may help us restrain our enthusiasm over large difference in small groups. However, a *p*-value of 0.01 in an observational study does not mean that a difference as great or greater than the one observed will occur no more often than 1 in 100 trials. By and large, statistical significance is a meaningless term in observational research. This is true for many reasons, but mainly because a *p*-value does not know any epidemiology; it was born into a world of tossing pennies and urns full of black and white balls, and it never read a book about disease. It can't tell the difference between truth and bias, and it worships sample size more than truth" (Stallones, 1987).

6. Discussion

What should be done when an estimated RR is small, and consideration of epidemiological and statistical concepts do not provide a clear answer to the question of whether the estimated RR is too small to be accurately estimated? Two possible approaches are discussed in this section. The first is a technical solution that involves gathering more data and testing the statistical model; the second is a more philosophical approach. They are not mutually exclusive ways to approach this difficulty, and they are not always practical or satisfactory.

6.1 A Direct method

A direct method to demonstrate that a RR has been accurately estimated is to demonstrate that the same risk is seen in a similar, but disjoint, population using the same model used to develop the initial estimate. A statistical model developed to estimate an RR can be thought of as a descriptive model because it describes the data at hand. It was crafted to be 'the best' at estimating the RR, and by extension for describing the data set. However, when a descriptive model is going to be used to predict future effects that will accrue because of a change in exposure, thinking and application of the model must also change. Chatfield (1995) noted, "…model builders should adopt a more pragmatic approach in which they search not for a true model, but rather a *parsimonious* model giving an adequate approximation to the data at hand and then concentrate on determining the model's *accuracy* and *usefulness*, rather than testing it [for significance]."

For example, a descriptive time-series regression model in a mortality study of air-pollution can be used to describe relationships in a specific city for the previous 10 years say. For prediction of pollution effects, use the same model with the same coefficients to predict one year (or more) into the future. Then determine the accuracy and precision of the predicted measurements. A kind of sensitivity analysis could be done by developing two time-series models on the past data: one with an air pollution term in the model and one without the air pollution term. Then see if either can accurately predict the future data, and if the model with the air pollution term has a better fit to the "new data."

This type of validation would be a major demonstration that the estimated risk was useful both as a descriptor and a predictor.
However, in many situations this method is not feasible. In a cohort study all the 'cases' have usually been identified and there are no more cases to use in the second, confirmatory, study. Or the study population may be too small to divide it for both descriptive and predictive purposes. A similar problem occurs with case-control studies. In air pollution time-series studies there are technical difficulties that preclude using the exact coefficients from past data in a new data set, although there is some hope that this problem can be overcome. In the time-series example, it is not sufficient to develop a model in a different location and show the form of the model is the same. This procedure only demonstrates that the model is descriptive for another data set; the goal is to show the model is predictive because it may be used to set policy for future exposures.

6.2 Trans-science

Weinberg (1972) coined the term *trans-science* to describe the realm of scientific questions of facts that are properly formulated, but cannot be answered by science. The answers transcend science, usually because of the impracticality of expending the required time and money. Such questions include the shape of an exposure response curve at very low levels of exposure, or probabilities of extremely improbable events such as failures in complex systems like nuclear reactors and dams. His point is that we will never know, with a reasonable expenditure of resources, what the precise risks are at low (often environmental) levels. He discusses the Science Republic (the realm of scientists), and the Political Republic (the realm of politicians).These two are joined in the Republic of Trans-Science, the character of which depends on place and time. A major difficulty is to draw the line between matters of science and matters of trans-science, and is a crucially important role of the scientist. This is a difficult task, but often it is beneficial to draw the line and declare a question to be in the realm of trans-science. Then it can be dealt with on moral or aesthetic grounds, rather than having a scientific debate that has no hope of resolution. He concludes with the upbeat idea that the most science can do is to inject some intellectual discipline into the Republic of Trans-science and hope that politics in an open society will keep it democratic.
In a paper that deals with a topic similar to the current one about estimating minimum risk, Weinberg (1985) proposes a method for dealing with trans-science questions that are concerned with setting limits on an exposure response curve. He calls the trans-science problem of setting an environmental standard based on incomplete information 'The Regulator's Dilemma' and suggests initiating a branch of science called Regulatory Science. In this new branch of science the norms of scientific proof would be less demanding than in standard science. Instead of asking science for answers to questions that are unanswerable, the regulators should be content with less far-reaching and seemingly less precise answers. He suggests that where ranges of uncertainty can be well established, then regulations can be based on the uncertainty. Where the ranges of uncertainty are so wide as to be meaningless, then ask the question so that the regulation does not depend on answers to an unanswerable question.
Weinberg (1985) provides an idea on how to provide some assurance of public safety despite uncertainty in the estimation of the risk. The idea, called *de minimis*, is to set the

permissible exposure level at some fraction above the natural background level irrespective of the possible risk. The idea is that for almost all environmental hazards, or putative hazards, there is a natural background level, such as for cosmic radiation or exposure to the suite of pollutants regulated by the National Ambient Air Quality Standards. A *de minimis* standard would be at some level above background level, for example 1.0 or 1.5 standard deviations above the mean. The idea is that man has lived with the natural exposure, and if it was harmful man would not have survived. So, an increase that is small relative to natural background should be acceptable.

7. Conclusion

Theoretically, there is no relative risk that is too small to be estimated. The relative risk is a construct or a concept, not a physical reality. Since it is a mathematically defined concept it can be mathematically estimated to any degree of precision. However, we have shown in this paper that (1) there are many assumptions that must be met to make certain that the RR estimate is accurate and precise; and (2) the significance level or uncertainty associated with the RR estimate has its own set of assumptions that must be met. So, while there may be no theoretical minimum RR that can be estimated, in practice there is a minimum risk and varies depending on uncertainties present in the context of each study.

An analogy in the physical world of estimating a RR is to measure the length of an object. A meterstick is precise enough to determine the width of a table to see if it will fit through a doorway, but a meterstick is not precise enough to measure the diameter of a shaft in an automobile engine with a tolerance of ±1.0 mm. To measure the shaft diameter one would use a micrometer. The micrometer while sufficiently precise to measure the shaft is not adequate to determine the size of a dust mite, usually in the range of 200 to 300 μm. The analogy can be carried through to the size of molecules, to the wavelength of visible light, and to the diameter of an electron. The conclusion is that while all the tasks involve measuring length and there is no practical 'minimum length', different tools and considerations are needed depending on the object to be measured and the precision required.

The analogy carries over to the estimation of a RR. There is less concern about detailed assumptions when estimating a RR of lung cancer among pack-a-day smokers with estimated RRs ranging from 10 to 25. But there is greater concern about meeting detailed assumptions when estimating the RR of mortality from a 10 $\mu g/m^3$ increase in PM air pollution where RRs may be as low as 1.004 (or 3 to 4 orders of magnitude smaller than that of lung cancer and smoking), and concluding the association is causal.

What is to be done when an estimated RR is small and in the controversial range? Two suggestions were provided: (1) verify the estimated risk with new data; and (2) regulate on exposure not on the response as done in the regulatory arena when the consequences of exposure are controversial. In many cases neither of these solutions will be considered acceptable, putting us back into the realm of uncertainty.

Epidemiology has provided society with many benefits by providing clues to the causes of infectious diseases and identifying factors that contribute to the start and development of non-communicable diseases (Wynder, 1987). Causal knowledge allows quick and effective

preventive measures to lessen or halt diseases without a full understanding of the underlying mechanism. Some examples of interventions fruitfully undertaken without a full understanding of mechanisms include tobacco-related diseases, radiation and leukemia, and UV light and skin cancer. In these cases RRs were quite high, with magnitudes exceeding 10 -fold increases at reasonably low exposures.

However, when it comes to weaker associations, errors in the estimation of the risk can produce a false positive association when in fact no association exists, and vice-versa (Wynder, 1987). Epidemiological data may have the required quality to address a research hypothesis and estimate risk if thought, planning and care are taken in the design of the study with consideration of how cases and controls are selected or the cohort is defined, when possible biases and confounders have been avoided or properly considered, when problems of subgroup analysis are clearly understood, and when the protocol is carefully and accurately carried out. It is likely a correct interpretation may be possible if all available evidence is subjected to a careful review of the biological plausibility of the initial hypothesis, if the criteria discussed above are implemented, and if the data analyses are correctly carried out without pre-conceived bias. There are times, however, that such care is not given to epidemiological studies and their interpretations. There is a great concern that in the "rush to publish," often as a preliminary report, false associations are reported which do not do justice to the factor being incriminated as harmful, nor to public safety if a risk in fact does not exist, nor to the science of epidemiology.

Most of the problems in epidemiology, as they relate to reliable and interpretable estimations of small RRs, can be avoided if researchers pay attention and carefully consider issues relating to study design, confounding, the many forms of bias, reliability, measurement error and misclassification, multiple agents, sampling error, violation of model assumptions, model shopping, multiple subgroup analyses, and over-reliance on significance levels or *p*-values.

We agree with Wynder (1987) that epidemiology is able to correctly interpret relatively small relative risks, but only if the best epidemiological methodology is applied and only if the data are fully evaluated by examining all judgment criteria, especially those of biological plausibility. As RRs become smaller, the need for close adherence to these basic principles becomes greater. If these ideas are applied, a conclusion of no risk should reassure society. And when a risk is reported as positive, appropriate preventive measures to reduce avoidable illness can be used to successfully reach the ultimate goal of epidemiology and preventive medicine.

8. Appendix

8.1 Quotes from knowledgeable researchers in the field

There are many references from experienced health scientists as to what a reasonable minimum relative risk should be. Gary Taubes (1995) in his 1995 Science article had collected a number of quotes, not all of which could be independently verified. The following is a collection of some of Taubes' quotes and other comments from various sources, presented in almost alphabetical order.

"As a general rule of thumb, we are looking for a relative risk of 3 or more before accepting a paper for publication."

Marcia Angell, editor of the New England Journal of Medicine, reported in Taubes 1995.
"If it's a 1.5 relative risk and it's only one study and even a very good one, you scratch your chin and say maybe."
John Bailar, reported in Taubes 1995
"Relative risks of less than 2.0 may readily reflect some unperceived bias or confounding factor, those over 5.0 are unlikely to do so.**"**
Breslow and Day, 1980, Statistical methods in cancer research, Vol. 1, The analysis of case control studies. Published by the World Health Organization, International Agency for Research on Cancer, Sci. Pub. No. 32, Lyon, p. 36.
"Epidemiological studies, in general are probably not able, realistically, to identify with any confidence any relative risks lower than 1.3 (that is a 30% increase in risk) in that context, the 1.5 [reported relative risk of developing breast cancer after abortion] is a modest elevation compared to some other risk factors that we know cause disease."
Dr. Eugenia Calle, Director of Analytic Epidemiology, American Cancer Society, Washington Post - Oct 27,1994
" ... when relative risk lies between 1 and 2 ... problems of interpretation may become acute, and it may be extremely difficult to disentangle the various contributions of biased information, confounding of two or more factors, and cause and effect."
Richard Doll, F.R.S. and Richard Peto "The Causes of Cancer," Oxford-New York, Oxford University Press, 1981, p. 1219.
"An association is generally considered weak if the odds ratio [relative risk] is under 3.0 and particularly when it is under 2.0, as is the case in the relationship of ETS and lung cancer."
Dr. Geoffrey Kabat, Senior Epidemiologist, Albert Einstein College of Medicine, from E.L. Wynder & G.C. Kabat, Environmental Tobacco Smoke and Lung Cancer: A Critical Assessment, I.SAB.7.1 at 6 (JA 7,216),
http://www.forces.org/evidence/epafraud/files/osteen.htm, accessed 26 Dec 2007
"In epidemiologic research, [increases in risk of less than 100 percent] are considered small and are usually difficult to interpret. Such increases may be due to chance, statistical bias, or the effects of confounding factors that are sometimes not evident." "It is not size of the RR alone (but we have to agree at some point low is too low say 1.03 relative risk) but the results of other studies addressing the same issue and concerns about biological plausibility have to be factored in. Even though the size of the RR or OR is not necessarily determinative it is easy to cite a number of experts in the field who favor the notion that RR less than 2 should be- if not dismissed- at least looked at with a very skeptical eye."
National Cancer Institute, "Abortion and Possible Risk for Breast Cancer: Analysis and Inconsistencies," October 26, 1994 press release
"Differences in risk of 50% (Relative risk of 1.5) are small in epidemiological terms and severely challenge our ability to distinguish whether it reflects cause and effects or whether it simply reflects bias."
Lynn Rosenberg, Boston University School of Medicine quoted in Press Release U.S. National Cancer Institute Oct 26,1994
"Any scientist worth his qualifications knows that a RR of less than two or even three is unreliable and too shaky to place much reliance upon."
John K. Sutherland quoted from "The Week That Was", April 22, 2006,
"My basic rule is if the relative risk isn't at least 3 or 4, forget it."

Robert Temple, director of drug evaluation at the Food and Drug Administration, reported in Taubes 1995

"With epidemiology you can tell a little thing from a big thing. What's very hard to do is to tell a little thing from nothing at all."

Michael Thun, VP of Epidemiology and Surveillance Research at the American Cancer society reported in Taubes 1995

9. References

Antman, E. M., J. Lau, et al. (1992). "A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments of myocardial infarction." JAMA 268: 240-248.

Calabrese, 1991 E.J. Calabrese, Multiple Chemical Interactions, Lewis Publishers, Chelsea, MI, USA (1991).

Chan, A. W., A. Hrobjartsson, et al. (2004). "Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles." JAMA 291: 2457-2465.

Chang et al., 2005, "The effects of simultaneous exposure to methyl ehyl ketone and toluene on urinary biomarkers of occupational N,N-dimethylformamide exposure.", Tox Letters, Vol 155(3), 15 March 2005, Pages 385-395

Chatfield, C. (1995). "Model uncertainty, data mining and statistical inference." J Roy Stat Soc (Ser A) 158: 419-466.

Clyde, M. (2000). "Model uncertainty and health effect studies for particulate matter." Environmetrics 11: 745-763.

Cordier, S., A. Bergeret, et al. (1997). "Congenital malformations and maternal occupational exposure to glycol ethers." Epidemiology 8: 355-363.

Cordier, S., E. Szabova, et al. (2001). "Congenital malformations and maternal exposure to glycol ethers in the Slovak Republic." Epidemiology 12: 592-593.

Feron, VJ,. Cassee, J.P. Groten, P.W. van Vliet and J.A. van Zorge, (2002) "International issues on human health effects of exposure to chemical mixtures", Environ. Health Persp. 110 (2002), pp. 893–899.

Goodman, S. (1999a). "Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy." Ann Intern Med 130: 995-1004.

Goodman, S. (1999b). "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor." Ann Intern Med 130: 1005-1013.

Greenland, S. (2004). "Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies." Am J Epid 160: 301-305.

HEI (2003). Revised analyses of time-series studies of air pollution and health. Special Report: Revised analyses of the National Morbidity, Mortality, and Air Pollution Study. Boston, MA, Health Effects Institute.

Hill, A. (1965). "The environment and disease: association or causation?" Proc R Soc Med 58: 295-300.

Hodges, J. (1987). "Uncertainty Policy Analysis and Statistics." Stat Sci 2: 259-291.

Ioannidis, J. P. A. (2005). "Why most published research findings are false." PLoS Med 2(8): e124.

Ioannidis, J. P. A. and T. A. Trikalinos (2005). "Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials." J Clin Epidemiol 58: 543-549.

King, Gary, and Langche Zeng. "Inference in Case Control Studies." In Encyclopedia of Biopharmaceutical Statistics, edited by Shein-Chung Chow. 3rd ed. New York: Marcel Dekker, 2010.

Krewski, D., R. T. Burnett, et al. (2000). Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality. Special Report. Cambridge, MA, Health Effects Institute.

Lumley, T. and L. Sheppard (2000). "Assessing seasonal confounding and model selection bias in air pollution epidemiology using positive and negative control analyses." Environmetrics 11(6): 705-717.

Maldonado, G., E. Delzell, et al. (2003). "Occupational exposure to glycol ethers and human congenital malformations." Int Arch Occup Environ Health 76: 405-423.

McNamee, R. (2003). "Confounding and confounders." Occup Environ Med 60: 227-234.

Neyman, J. and E. Pearson (1933). "On the Problem of the most efficient tests of statistical hypotheses." Philosophical Transactions of the Royal Society. Series A 231: 289-337.

Peng, RD, Dominici, F and Zeger SL, (2006) "Reproducible Epidemiologic Research," Am J Epid 163: 783-789.

Phillips, C. (2003). "Quantifying and reporting uncertainty from systematic errors." Epidemiology 14: 459-466.

Phillips, C. (2004). "Publication bias *in situ*." BMC Medical research Methodology; online at http://www.biomedcentral.com/1471-22884/20. 4: 20.

Rothman, K. J. and C. Poole (1988). "A strengthening programme for weak associations." Int J Epidemiol 17((Suppl 4)): 955-959.

Rothman, K. J. and S. Greenland (2005). "Causation and causal inference in epidemiology." AJPH 95((Supplement 1)): S144-S150.

Shaw, G. M., E. M. Velie, et al. (1999). "Maternal occupational and hobby chemical exposures as risk factors for neural tube defects " Epidemiology 10:124-129 (erratum appears in Epidemiology 10:777).

Stalones, RA. (1987). "The use and abuse of subgroup analysis in epidemiological research." Preventive Medicine 16:183-194.

Sterne, J. A. and G. Davey Smith (2001). "Sifting the evidence-what's wrong with significance tests." BMJ 322: 226-231.

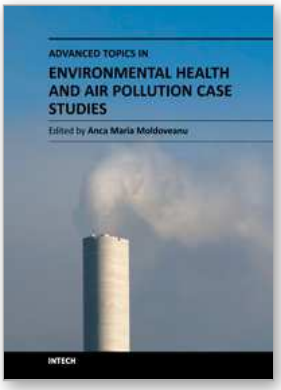
Taubes, G. (1995) "Epidemiology faces its limits." Science 269(5221):164-9.

Wacholder, S., S. Chanock, et al. (2004). "Assessing the probability that a positive report is false: an approach for molecular epidemiology studies." J Natl Cancer Inst 96: 434-442.

Weinberg, A. (1972) "Science and Trans-Science," Minerva 10:209-222.

Weinberg, A. (1985) "The regulator's dilemma." Issues in Science and Technology 2:59-72.

Wynder, E., "Workshop on Guidelines to the Epidemiology of Weak Associations." Preventive Medicine 16:139-212; 1987



Advanced Topics in Environmental Health and Air Pollution Case Studies

Edited by Prof. Anca Moldoveanu

ISBN 978-953-307-525-9
Hard cover, 470 pages
Publisher InTech
Published online 29, August, 2011
Published in print edition August, 2011

The book describes the effects of air pollutants, from the indoor and outdoor spaces, on the human physiology. Air pollutants can influence inflammation biomarkers, can influence the pathogenesis of chronic cough, can influence reactive oxygen species (ROS) and can induce autonomic nervous system interactions that modulate cardiac oxidative stress and cardiac electrophysiological changes, can participate in the onset and exacerbation of upper respiratory and cardio-vascular diseases, can lead to the exacerbation of asthma and allergic diseases. The book also presents how the urban environment can influence and modify the impact of various pollutants on human health.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Mark J. Nicolich and John F. Gamble (2011). What is the Minimum Risk that can be Estimated from an Epidemiology Study?, Advanced Topics in Environmental Health and Air Pollution Case Studies, Prof. Anca Moldoveanu (Ed.), ISBN: 978-953-307-525-9, InTech, Available from: http://www.intechopen.com/books/advanced-topics-in-environmental-health-and-air-pollution-case-studies/what-is-the-minimum-risk-that-can-be-estimated-from-an-epidemiology-study-

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821