Biases in the Identification of Risk Factor Thresholds and J-Curves

Ian C. Marschner¹², R. John Simes², and Anthony Keech²

¹ Asia Biometrics Center, Pfizer Australia, Sydney, Australia.
² National Health and Medical Research Council Clinical Trials Center, University of Sydney, Sydney, Australia.

Received for publication March 15, 2006; accepted for publication April 9, 2007.

For some diseases, there has been controversy about whether key risk factors are related linearly to the occurrence of disease events. This issue has important implications for strategies to modify risk factors, since nonlinear threshold or J-curve associations imply that risk factor modification is not beneficial beyond a certain level. This paper considers whether nonlinear risk factor associations can arise spuriously from selection mechanisms common in prospective cohort studies. Using theory, simulation, and cohort data, the authors show that selecting individuals based on their prior disease status leads to the primary risk factor being negatively confounded with other residual risk factors. If this confounding combines with effect modification between the primary and residual risk factors, as exists in cardiovascular disease, then the aggregate effect is nonlinear distortion of the risk factor relation. Such distortion can produce an apparent threshold or J-curve relation, even if the true underlying relation is linear. The authors conclude that nonlinear risk factor associations observed in primary or secondary prevention cohorts should be interpreted with caution because they may be consistent with an underlying linear lower-is-better relation. Randomized studies provide an important complement to prospective cohort studies when choosing between intensive and moderate risk factor modification strategies in high-risk populations.

Abbreviation: LIPID, Long-Term Intervention with Pravastatin in Ischemic Disease.

Understanding the relation between continuous risk factors and disease events is central to understanding how to improve health outcomes through risk factor modification. Of particular importance is whether the relation between a risk factor and the occurrence of disease events is linear, on an appropriate scale, so that a lower value of the risk factor is associated with a lower incidence of disease events. Linear risk factor associations rarely exist on the absolute incidence scale but are common on the log-incidence scale, where each unit change in the risk factor leads to a constant, proportionate change in incidence (1). Such a relation suggests a lower-is-better approach to risk factor modification (1–3).

An alternative to the lower-is-better relation occurs when the risk factor is related nonlinearly to the occurrence of disease events. An important special case is the threshold relation, where the risk factor and disease events are positively related above a threshold risk factor level but are unrelated below the threshold level. A further variation is the J-curve relation, where the occurrence of disease events changes from negative to positive. Thresholds and J-curves have important implications for risk factor modification, indicating that beyond a certain level, risk factor modification is beneficial and possibly harmful. It is therefore important to take great care when identifying threshold and J-curve associations and to be aware of biases that can influence identification of such associations.
Threshold and J-curve associations have been observed in a variety of risk factor contexts. Our research is motivated by cardiovascular risk factors, particularly cholesterol level and blood pressure as risk factors for coronary and vascular events. However, the results are relevant to any relation between a continuous risk factor and the occurrence of disease events. For cardiovascular risk, cholesterol and blood pressure thresholds and J-curves have been observed in populations with existing disease (4–6) and populations without existing disease (7, 8). Nonetheless, large-scale randomized studies comparing intensive cholesterol lowering with moderate lowering support a lower-is-better strategy (9–12), and meta-analyses of prospective blood pressure and cholesterol studies do likewise (1, 13–15). Thus, while the lower-is-better paradigm has prevailed, the accumulation of evidence leading to this conclusion was complicated by studies reporting apparent threshold and J-curve associations.

This paper presents a new explanation for conflicting evidence in the assessment of risk factor thresholds and J-curves. Using theory, simulation, and data, we argue that two influences can combine to distort risk factor associations such that a lower-is-better relation may appear to be a threshold or J-curve relation. The first influence is the restriction of cohorts to either individuals with no prior disease, called primary prevention cohorts, or individuals with a prior disease event, called secondary prevention cohorts, which leads to confounding of the primary risk factor of interest with other, or residual, risk factors. For example, an individual qualifying for a secondary prevention study with a low level of the primary risk factor is likely to have high levels of the residual risk factors for a prior event to have occurred. Likewise, individuals who have not experienced a prior event but have high levels of the primary risk factor are likely to have low levels of the residual risk factors so that a prior event did not occur. In both instances, this leads to a negative relation between the primary risk factor and the residual risk factors. The second influence is effect modification between the primary risk factor and the residual risk factors, which occurs, for example, in cardiovascular disease. In this paper, we demonstrate that the aggregate effect of these two influences may produce spurious threshold or J-curve associations, even when an underlying lower-is-better relation exists.

MATERIALS AND METHODS

Theoretical studies

A log-linear proportional hazards model was used to relate the incidence of a disease event with the primary risk factor of interest and other, or residual, risk factors. The term residual risk level is used to describe a measure of the aggregate risk level associated with the residual risk factors. For example, if the primary risk factor is cholesterol level, then the residual risk level might be measured by the number of noncholesterol risk factors. Letting P stand for the primary risk factor and R stand for the residual risk level, the log-linear model is

\[
\log \lambda_{adj}^\tau (P, R) = \log \lambda_{adj}^\tau + aP + bR, \tag{1}
\]

where \(\lambda_{adj}^\tau (P, R)\) is the disease incidence rate at time \(t\) adjusting for \(P\) and \(R\), while \(\lambda_{adj}^\tau\) is a baseline incidence rate that is independent of \(P\) and \(R\). This model reflects a lower-is-better relation between the primary risk factor and the disease incidence rate, after adjustment for the residual risk level.

The apparent relation between disease incidence and the primary risk factor was investigated when confounding and effect modification are present. Effect modification was incorporated into the model by allowing the rate ratio associated with the primary risk factor in equation 1 to be linearly dependent on the residual risk level:

\[
a = a_0 + a_1 R. \tag{2}
\]

Even with effect modification, equation 1 represents a lower-is-better relation because, for any given residual risk level, the relation between the primary risk factor and the (log) incidence rate is linear.

Confounding due to selection mechanisms was incorporated through a linear relation between the average residual risk level and the primary risk factor:

\[
\hat{R}(P) = E(R \mid P) = c + dP. \tag{3}
\]

As explained in the introductory paragraphs of this paper and confirmed in the Results section, to study confounding in primary and secondary prevention cohorts, it is appropriate to assume that \(d \leq 0\), reflecting the negative relation that selection mechanisms would produce.

In practice, continuous risk factors such as cholesterol level and blood pressure are subject to measurement error and fluctuation over time, leading to a phenomenon known historically as attenuation by errors (16) and more recently as regression dilution (17, 18). Regression dilution leads to underestimation of the apparent relation between a risk factor and disease events. The regression dilution ratio, \(0 < g \leq 1\), is the ratio of the steepness of the apparent relation to that of the true relation (17). Regression dilution makes the incidence rate appear to depend on the primary risk factor according to \(\lambda_{adj}^\tau (gP, R)\) rather than \(\lambda_{adj}^\tau (P, R)\).

The incidence rate as a function of \(P\) alone was approximated by replacing \(R\) by \(\hat{R}(P)\) in the model specified by equations 1 and 2. The resulting function, \(\lambda_{adj}^\tau (P) = \lambda_{adj}^\tau (gP, \hat{R}(P))\), provides information about the apparent dependence of the incidence rate on \(P\), not adjusting for \(R\). By varying the model parameters, it was investigated whether spurious threshold and J-curve associations could arise in the unadjusted incidence rate, despite the underlying linear relation. Because various methods exist to correct for bias due to regression dilution (19, 20), these theoretical calculations investigated the relation that would occur both before \((g < 1)\) and after \((g = 1)\) such correction.

Simulation studies

The theoretical studies mentioned above can investigate the effect of confounding and effect modification, if they exist. Evidence that selection mechanisms can actually lead to confounding between the primary and residual risk levels was provided by simulation of risk factor associations in primary and secondary prevention cohorts.
A population was simulated in which the prevalence $\theta$ of prior coronary heart disease was related to cholesterol level $P$ and the number of noncholesterol risk factors $R$ according to a complementary log-log relation equivalent (21) to the proportional hazards model in equation 1:

$$\log\{-\log(1-\theta)\} = \mu + aP + bR.$$  

A rate ratio of 2.0 was assumed for each mmol/liter higher cholesterol level ($a = \log(2)$), and a rate ratio of 1.5 was assumed for each additional noncholesterol risk factor ($b = \log(1.5)$). The parameter $\mu$ was calculated such that the prevalence of prior coronary heart disease in the population was 15 percent. Cholesterol level in the population was assumed to be distributed normally with mean 5.5 and variance 1.0, and it was assumed to be independent of noncholesterol risk level. Effect modification and regression dilution were absent in the simulation studies so that the manner in which selection mechanisms induce confounding could be studied in isolation. A maximum of eight noncholesterol risk factors was possible, which is the number of noncholesterol risk factors in a prior model derived from the Long-Term Intervention with Pravastatin in Ischemic Disease (LIPID) study (22). The average number of noncholesterol risk factors was two in the first simulation and four in the second simulation and had a beta-binomial distribution with variance 0.09375 and 0.125, respectively. This variance

---

**FIGURE 1.** Simulation studies of a population in which cholesterol level and noncholesterol risk level are unrelated. The confounded associations in the primary and secondary prevention subpopulations are displayed, assuming an average of either two (panel A) or four (panel B) for the number of noncholesterol risk factors in the full population.

**FIGURE 2.** Cohort studies depicting confounding and effect modification between cardiovascular risk factors and residual risk level. Panel A, average number of noncholesterol risk factors vs. cholesterol level at baseline in the Long-Term Intervention with Pravastatin in Ischemic Disease (LIPID) study (23), Australia and New Zealand, 1990–1997. Panel B, coronary event rate ratio (log scale) per mmol/liter higher cholesterol level for the placebo cohort of the LIPID study (23), Australia and New Zealand, 1990–1997, and for 10 cohort studies reported by Law et al. (14), United Kingdom, United States, Sweden, and Israel, 1950–1991; and stroke mortality rate ratio per 10-mmHg higher diastolic blood pressure for 61 cohort studies in the Prospective Studies Collaboration (13), Europe, United States, Canada, Japan, China, and Australia, 1949–1997. Residual risk level is based on either number of noncholesterol risk factors (LIPID study) or age (other studies).
is equivalent to each risk factor being present with an average probability of 0.25 in the first simulation and 0.5 in the second simulation but with the actual probabilities varying between individuals, such that noncholesterol risk factors have a 0.5 correlation with each other.

When these assumptions were used, a population of one million individuals was simulated and was separated into the primary and secondary prevention subpopulations, which consisted of those individuals with, and respectively without, prior coronary heart disease. For each of the two subpopulations, the relation between cholesterol level and number of noncholesterol risk factors was plotted and inspected for evidence of confounding, despite the full population being simulated so that no such confounding was present.

**Cohort studies**

To investigate the relevance of selection bias and effect modification to real data, evidence was examined from prospective cohorts. The LIPID study (23) was used to investigate the presence of confounding and effect modification in relation to cholesterol level and coronary event for a secondary prevention cohort of individuals. This study randomly assigned 9,014 individuals with existing coronary heart disease to receive either placebo or cholesterol-lowering pravastatin therapy. Confounding of cholesterol level and noncholesterol risk level was investigated by calculating the average number of noncholesterol risk factors in each decile of baseline cholesterol level. Eight noncholesterol risk factors were examined, as determined by a previously developed risk factor model for the LIPID study, and baseline cholesterol levels were adjusted for regression dilution as in previous analyses of the LIPID data (22, 24).

Effect modification in the LIPID study was investigated by fitting a proportional hazards model for the incidence of coronary event, with baseline cholesterol level and noncholesterol risk level as predictors. Noncholesterol risk level was classified into low (two or fewer risk factors), medium (three to five risk factors), or high (six or more risk factors), and the rate ratio per unit of baseline cholesterol level was estimated for each noncholesterol risk level. Only the 4,502 individuals from the placebo cohort were included to prevent the treatment-mediated improvements in coronary risk from complicating interpretation of analyses. Evidence of effect modification was also sought from previously published cohort data on the relation between cholesterol level and coronary events, as well as blood pressure and vascular events (13, 14, 25). As reported in the original publications, all cohort analyses discussed here were adjusted for regression dilution.

**RESULTS**

**Confounding in primary and secondary prevention cohorts**

Based on the simulation results, figure 1 shows the average number of noncholesterol risk factors for each 0.5-mmol/liter increment in cholesterol level. These results cover a range of populations, from low-risk primary prevention (figure 1A, lower curve) to high-risk secondary prevention (figure 1B, upper curve). Although there was no relation between cholesterol level and number of noncholesterol risk factors in the full population, a negative relation emerged in both the primary and secondary prevention subpopulations. When the noncholesterol risk level was lower in the full
population, the confounding was more pronounced in the secondary prevention subpopulation, whereas, when the noncholesterol risk level was greater, the confounding was more pronounced in the primary prevention subpopulation. In both cases, there was a substantive decline in the number of noncholesterol risk factors over the range of cholesterol levels.

Because the assumptions of the simulation studies were necessarily simplified compared with a real population, evidence was sought for a similar level of confounding in a real population. Figure 2A shows the relation between the average number of noncholesterol risk factors and cholesterol level for individuals enrolled in the LIPID study. The negative trend is consistent with observations from the simulation studies. These results support the expectation that a primary risk factor will be negatively confounded with residual risk level in subpopulations chosen on the basis of prior occurrence of disease events.

Effect modification

Evidence for effect modification between cholesterol level and noncholesterol risk level is presented in figure 2B for the LIPID study. The rate ratio for each higher unit of cholesterol level was lower among individuals with a higher residual risk level, as measured by the number of noncholesterol risk factors. This finding concords with previously published cohort data on the relation between cholesterol level and coronary events. In particular, also plotted in figure 2B is the relation between age and the rate ratio for each higher unit of cholesterol level in a combined analysis of 10 cohort studies provided by Law et al. (14).

The type of effect modification depicted in figure 2B suggests that cholesterol level is a stronger predictor of coronary events in patients with a lower noncholesterol risk level. This relation has also been observed in cohort data relating blood pressure to vascular events. Two large cohort study overviews of the relation between blood pressure and stroke, involving well over a million individuals, were provided by the Prospective Studies Collaboration (13) and Asia Pacific Cohort Studies Collaboration (25). In both of these overviews, there was pronounced effect modification in which blood pressure was a stronger predictor of stroke events for younger individuals. This is illustrated in figure 2B, which provides the stroke mortality rate ratio for each 10-mmHg higher diastolic blood pressure, as observed in 61 cohort studies in the Prospective Studies Collaboration. Such effect modification led to a pronounced lack of parallelism in the relation between the rate ratio for stroke and the blood pressure level (13, 25, 26).

Combined influence of confounding and effect modification

Theoretical studies considered the extent to which the combined effect of confounding and effect modification could explain nonlinear risk factor associations. On the basis of the proportional hazards model specified by equations 1 and 2, the form of the incidence rate as a function of cholesterol level and noncholesterol risk level is presented in figure 2B for the LIPID study. The rate ratio for each higher unit of cholesterol level was approximated on the log scale by

\[
\log \lambda(t)(P) = \log \lambda^{adj}_{rel}(gP, \bar{R}(P)) = \log \lambda^{*}_{rel} + \{g(a_0 + a_1(c + dP))\}P + b(c + dP) = \log \lambda^{*}_r + (ga_0 + ga_1(c + bd))P + ga_1dP^2, \tag{4}
\]

with \(\lambda^*_r = \exp(bc)\lambda^{*}_{rel}\). Given the proportional hazards structure in equation 4, the dependence of the unadjusted incidence rate on \(P\) is captured by the incidence rate ratio \(\Lambda(P)\), which is independent of time. When the unadjusted incidence rate specified in equation 4 was used, the incidence rate ratio associated with a primary risk factor value \(P\) relative to a referent value \(P_o\), is given by

\[
\log \Lambda(P) = k_2P^2 + k_1P + k_0, \tag{5}
\]

where \(k_2 = ga_1d, k_1 = ga_0 + ga_1c + bd,\) and \(k_0 = (ga_0 + ga_1c + bd)P_0 + ga_1dP_0^2\). Equation 5 shows that, although the underlying incidence model in equation 1 reflects a linear dependence on the primary risk factor after adjustment for
residual risk level, a quadratic relation arises when residual risk level is not adjusted for. This indicates that nonlinear distortions of the primary risk factor relation are possible in unadjusted analyses, which is true regardless of whether regression dilution is present, since the assumption \( g = 1 \) would not alter the quadratic dependence on \( P \). It is also true regardless of the selection mechanism that induced the confounding and so applies to both the primary and secondary prevention situations.

By examining the role of \( a_1 \) and \( d \) in equation 5, it follows that both confounding and effect modification are needed to produce nonlinear distortions of the primary risk factor relation. In particular, when there is an absence of either confounding (\( d = 0 \)) or effect modification (\( a_1 = 0 \)), then \( k_2 = ga_1d = 0 \), and the apparent dependence on \( P \) remains linear. However, if \( a_1 = 0 \), the true gradient specified by equation 1 is \( a_0 \) and the apparent gradient specified by equation 5 is \( g a_0 + bd \leq a_0 \), since \( d \leq 0 \) and \( g \leq 1 \). This indicates that the apparent linear relation is attenuated compared with the underlying relation, when there is confounding with no effect modification. This statement is true even when regression dilution is absent (\( g = 1 \)).

Figure 3 illustrates how confounding and effect modification combine to produce nonlinearity in equation 5. In each of the four panels, linear associations between the rate ratio and the primary risk factor are depicted by thin lines, with upper thin lines corresponding to higher residual risk level \( R \). Because of confounding, higher values of the primary risk factor are associated with lower values of the residual risk level. This means that the apparent rate ratio for a given value \( P \) of the primary risk factor comes from a lower thin line if \( P \) is high and from an upper thin line if \( P \) is low. This situation is depicted by marking points on figure 3 that jump from lower thin lines to upper thin lines as \( P \) moves lower. Although this is only an intuitive demonstration because the precise relation between \( P \) and \( R \) is not stated, it does illustrate how the unadjusted relation will behave under different scenarios, as depicted by the thick lines. This apparent relation ranges from linear attenuation when no effect modification is present (figure 3A), to mild nonlinearity when there is mild effect modification (figure 3B), to threshold and J-curve associations when the effect modification is substantial (figures 3C and 3D).

Figure 3 is a graphic aid to assist in understanding the nonlinearity in equation 5, but it does not depict the actual nonlinear relation represented in equation 5. This nonlinearity was investigated in figure 4 both before (panel A) and after (panel B) correction for regression dilution, using a range of parameter values in table 1 that are broadly relevant to the relation between cholesterol level and coronary events (combinations 5 and 6 were omitted from panel B to avoid clutter, and they were close to the solid line). For the basic parameter assumptions (solid lines), the apparent relation between cholesterol level and coronary events shows a nonlinear leveling off below 4.5–5.0 mmol/liter. When the assumptions were varied to include greater effect modification or confounding, the apparent nonlinearity included J-curve associations. When these assumptions were varied in the opposite direction, the apparent nonlinearity was less pronounced. A comparison of panels A and B in figure 4 shows that regression dilution accounts for some but not all of the nonlinearity. Even after adjustment for regression dilution, the relation between cholesterol and coronary events can appear as a threshold or J-curve.

### DISCUSSION

The observation of threshold or J-curve risk factor associations should be interpreted with caution. Analyses examining the relation between a primary risk factor and the occurrence of disease events should include careful adjustment for residual risk factors and allow for effect modification between the primary and residual risk factors. Such

**TABLE 1. Parameter assumptions for theoretical calculations of the unadjusted relation between cholesterol level and coronary event, as plotted in figure 4**

<table>
<thead>
<tr>
<th>Basic assumptions</th>
<th>Interpretation</th>
<th>Sensitivity analysis assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_0 = b = \log(2) )</td>
<td>Incidence rate is doubled for each unit of cholesterol and for each noncholesterol risk factor.</td>
<td>1. ( a_0 = b = \log(1.5) )</td>
</tr>
<tr>
<td>( a_1 = \log(0.9) )</td>
<td>Rate of increase in incidence for each unit of cholesterol decreases by 10% for each noncholesterol risk factor.</td>
<td>2. ( a_0 = b = \log(3) )</td>
</tr>
<tr>
<td>( c = 6.75, d = -0.5 )</td>
<td>Average number of noncholesterol risk factors decreases by 0.5 for each unit of cholesterol. The intercept of 6.75 in equation 3 is determined by the assumed value for ( d ), an assumed mean cholesterol level of 5.5 mmol/liter, and an assumed mean number of noncholesterol risk factors of 4, i.e., ( c = 4 - 5d ).</td>
<td>3. ( a_1 = \log(0.85) )</td>
</tr>
<tr>
<td>( g = 0.75 ) or 1.0</td>
<td>For ( g = 0.75 ), the relation is that which would occur with no correction for regression dilution, when measurement error and fluctuations account for one quarter of the variation in observed cholesterol levels. For ( g = 1.0 ), the relation is that which would occur after correction for regression dilution.</td>
<td>4. ( a_1 = \log(0.95) )</td>
</tr>
<tr>
<td>( g = 0.75 ) or 1.0</td>
<td>Not applicable</td>
<td>5. ( c = 8.125, d = -0.75 )</td>
</tr>
<tr>
<td>( g = 0.75 ) or 1.0</td>
<td></td>
<td>6. ( c = 5.375, d = -0.25 )</td>
</tr>
</tbody>
</table>

Am J Epidemiol 2007;166:824–831
adjustment may eliminate an apparent threshold or J-curve, leaving a linear lower-is-better relation. Confounding between the primary and residual risk factors can arise from selection mechanisms common in epidemiologic and clinical studies, where eligibility is restricted based on prior occurrence of disease. While it is well known that any epidemiologic investigation of risk factors should consider the effects of confounding, these results point to specific mechanisms by which confounding can arise from selection bias and can combine with effect modification to produce spurious threshold and J-curve associations.

In this paper, we have presented a new source of bias in the study of risk factor extremities that should be considered in conjunction with other, more well-known sources, particularly reverse causation. This situation arises when a disease event causes a reduction in the primary risk factor, making the event rate appear higher than it should be at lower risk factor levels. This alone could produce a threshold, or alternatively may exacerbate the distortions illustrated in figures 3 and 4, allowing a lower level of confounding and effect modification to produce a threshold.

Effect modification and confounding will not always lead to threshold or J-curve associations, since the apparent curvature can be minimal and appear as nonlinear attenuation (figure 3B). For example, the LIPID study displayed both confounding (figure 2A) and effect modification (figure 2B), yet prior analyses displayed a linear relation, albeit with wide confidence intervals not inconsistent with a threshold relation (24). Despite the lack of a pronounced threshold, these prior LIPID study analyses did demonstrate attenuation that was ameliorated by carrying out adjusted analyses (24), as expected based on the results presented here. For studies in which a pronounced threshold or J-curve has been observed, confounding and effect modification could provide an explanation consistent with an underlying linear relation.

Our results do not justify a conclusion that threshold associations can always be explained by selection bias in primary and secondary prevention studies. Rather, they show that observed threshold and J-curve associations may have other explanations that are difficult to rule out based on cohort data. For evaluating intensive and moderate risk factor modification in secondary prevention and high-risk primary prevention populations, randomized studies provide an important complement to prospective cohort studies. Such studies involve randomized comparison of more intensive intervention strategies appropriate for lower-is-better associations, with less intensive strategies appropriate for threshold or J-curve associations. Randomized studies are generally impractical for low-risk primary prevention populations, which are more suited to prospective cohort studies with careful adjustment for residual risk factors and effect modification. Some authors have argued that randomized studies are unnecessary for the adoption of intensive risk factor modification strategies when cohort data point strongly to a lower-is-better relation (1). This depends on an assumption that treatment or lifestyle-mediated modifications to risk factors convey comparable benefit to those imparted by natural differences, the plausibility of which would require assessment in each disease context. For cardiovascular risk factors, despite the existence of earlier threshold and J-curve associations in prospective cohort data, comprehensive overviews and large-scale randomized trials have now largely confirmed the lower-is-better paradigm (1, 9–13, 15, 26). Our research provides a new explanation for apparent inconsistencies observed during the accumulation of evidence leading to this conclusion, and it has relevance in other disease contexts.

ACKNOWLEDGMENTS

Dr. Marschner holds stock in Pfizer Inc. and is currently employed by this company.

REFERENCES

14. Law MR, Wald NJ, Thompson SG. By how much and how quickly does the reduction in serum cholesterol concentration...


