Statistics and Its Interface Volume 2 (2009) 123–131

The use of the risk percentile curve in the analysis of epidemiologic data

Nilanjan Chatterjee, Barry I. Graubard and Joseph L. Gastwirth

Economists and social scientists have used percentile-based curves, e.g., the Lorenz curve, to summarize data from positive random variables, especially skewed data such as income. Measures of interest, e.g., the Gini index of relative inequality, correspond to areas defined by the curves. In this paper we explore the usefulness of risk-percentile and related curves in epidemiology, especially when the exposure data is skewed. These curves are defined and risk measures, e.g. the population attributable risk are related to areas under them for data from either a cohort or a case-control study. Regression spline methods of estimating these curves are used as they do not require a pre-specified risk model. The concepts are illustrated by analyzing data from a cohort study of dietary red meat consumption and all-cause mortality and a case-control study of serum homocysteine level and colorectal cancer. These examples show that the risk percentile curves often are more useful than presenting the risk as a function of the raw exposure data as the later graph is often dominated by the tails when the data is skewed. Furthermore, the risk percentile curve is more informative than the commonly used method of presenting the average risk in categories defined by several fixed percentiles such as quartiles or quintiles. Indeed, the risk averages for these categories can be obtained from the risk-percentile curve.

Keywords and phrases: Relative risk, Absolute risk, Population attributable risk, Logistic regression, Cox proportional hazard regression, Case-control study, Cohort study, Attributable risk reduction curve, Expectancy curve, Survey data.

1. INTRODUCTION

Due to the skewed distribution of many exposure variables, the absolute or relative risks obtained in epidemiologic studies are often reported by quartiles. It is known, however, that such grouping typically entails a loss of statistical information (Zhao and Kolonel, 1992). This paper demonstrates the usefulness of the full Risk Percentile Curve, RPC, which can be used to present either absolute or relative risks as a function of the percentile of the exposure distribution. When the exposure variable has long tails, as often occurs in nutritional and occupational epidemiological data, it is difficult to graph the risk profile over the original scale of exposure because most of the x-axis will be devoted to the few percent of extreme exposures. This concentrates the bulk of the population in a small interval, making it difficult to see the important features of the risk profile in the region where the bulk of the population lies. Plotting the risk profile against the corresponding percentiles takes care of this problem. An additional advantage of these curves is that the population attributable risk (the proportion of diseased cases attributable to the exposure of interest) is interpretable as a ratio of the excess risk, the area over which the relative risk exceeds 1.0, to the area under the entire curve, RPC. If the exposure is both protective and harmful then the area in which the relative risks are less than 1.0 are subtracted from the area of the curve that exceeds 1.0 to obtain the net excess risk. This excess risk is divided by the area under the entire curve to obtain the attributable risk.

Percentile-based curves have been used in economics and social science. The Lorenz curve, which is defined in terms of percentiles, and the area between it and a diagonal line or the Gini index, are standard tools used in the analysis of income data (Gastwirth, 1972; Cowell, 1995). In related work Mahalanobis (1960) introduced fractile analysis to summarize the relationship between two variables and the expectancy curve (Gastwirth et al., 2003) is used to assess the utility of pre-employment tests.

In medical research applications, Eide and Heuch (2001) identified how the ratio of certain areas under the risk percentile curve is the attributable risk. The RPC was used, at the suggestion of the authors of this paper to display odds ratios, which approximate relative risks, of prostate cancer by levels of serum selenium in a case-control study (Vogt et al., 2003). Recently, Pepe et al. (2008) used risk percentile curves (called predictiveness curves) to assess how well biomarkers predict risk of disease.

In this paper we describe how the RPC can be estimated from data obtained in a cohort or case-control study. When the proportional hazard model is applicable to data from a cohort study, the RPC is based on the relative hazards instead of relative risks. Similarly, this ratio of the two areas with respect to the curve that is described at the end of the first paragraph of this paper is shown to correspond to an attributable risk, which is independent of time. For case-control studies, a relative risk percentile curve is based on

the odds ratio and again the population attributable risk is a ratio of the area corresponding to the "excess" risk to the area under the curve.

When the biologic relationship between the level of exposure and disease is the same in two populations, it is often useful to generalize the results of a study from one population to the other population. Typically studies that estimate risks or relative risks are not based on a true random sample of the population, so that the exposure levels in the study group may not reflect that of the overall population. However, if a national sample of the exposure level of the population is available, a standardized, RPC, for the entire population can then be constructed using the estimated risks from the study and the exposure distribution of the population. By taking exposures in one population as the standard, one can compare the risk profiles of two populations. Plotting both RPCs on the same scale provides a visual comparison.

The basic concepts are described in section two. A simple example assuming a logistic risk function and an exponential exposure distribution illustrates the ideas. Section 3 is devoted to the statistical methodology used in the analysis. The use of spline regression to obtain the estimated risk profile and creating the corresponding risk-percentile curve is described. The methodology is illustrated by reanalyzing two epidemiologic studies. In the first study, we study the relationship between dietary red meat consumption and all-cause mortality using data from the National Institutes of Health (NIH)-AARP (formerly known as the American Association for Retired Persons) Diet and Health Study, a cohort study. In previous analyses of this data, Sinha et al. (2009) reported a statistically significant positive association between total red meat consumption and all-cause mortality in men. The second study is a community population-based case-control study of the relationship between homocysteine levels and invasive cervical cancer from five study sites in the U.S. (Weinstein et al., 2001). Here we estimate the odds ratio of cervical cancer as a function of age-adjusted serum homocysteine levels. The method of obtaining a standardized RPC for national data using estimated relative risks from a case-control study is illustrated in section four as national levels of homocysteine were obtained in the Third National Health and Nutrition Examination Survey (NHANES III) (National Center for Health Statistics, 1994).

2. THE RISK-PERCENTILE CURVE

Assume that an exposure, X, has a continuous distribution $F(x)$ in the study population and that the probability of contracting the disease given $X = x$ is $P(x)$. The relative risk of the disease at $X = x$ in reference to a fixed level of exposure x_0 (e.g. unexposed) is given by $R(x) = P(x)/P(x_0)$. Defining the q-th percentile of the exposure distribution $x_q = F^{-1}(q), 0 \le q \le 1$, "the (absolute) risk percentile curve" is $RPC(q) = P(x_q)$ and "the relative risk percentile curve" is defined as $RRPC(q) = R(x_q)$. The risk percentile curve displays the dose-exposure relationship reflecting the inherent variation in the exposure distribution in the natural scale of the exposure that is with respect to its distribution. The following example illustrates this point. Suppose $F(x)$ is an exponential distribution with parameter λ and $P(x)$ is logistic with intercept α and slope β. Since $F(x) = 1-e^{-\lambda x} = q$, $x_q = \lambda^{-1}\ln(1/1-q)$. Letting $\bar{q} = 1-q$, we have

$$RPC(q) = P(x_q) = \frac{e^{\alpha+\tau\ln(1/\bar{q})}}{1+e^{\alpha+\tau\ln(1/\bar{q})}},$$

where $\tau = \beta/\lambda$. Thus the value of the RPC at the q$^{\text{th}}$ fractile in this example is a logistic function of $\ln(1/\bar{q})$ with $\tau = \beta/\lambda$ as the slope. Since the slope β is interpreted as the increase in log-odds of disease for each unit increase in the exposure x, and λ is the standard deviation of the exponential distribution, then $\tau = \beta/\lambda$ is interpreted as the increase in the log-odds of the disease for each standard deviation unit increase in exposure. For example, in Figure 1 the risk percentile curves for $\tau = 0.25$, 0.50 and 0.75 are plotted. For each value of τ, the value of α was chosen to ensure that the marginal probability of the disease in the population is 0.05. The area under the curve $RPC(q)$ equals

$$\int_0^1 RPC(q)dq = \int_0^\infty P(x)dF(x),$$

which is the average risk, μ_p, in the population due to this exposure. The population attributable risk (Benichou, 2001) or fraction of cases that could be eliminated if the population would no longer be exposed ($x = 0$) to the risk factor is given by

$$\frac{\mu_p - P(0)}{\mu_p}.$$

Note that this is the ratio of the area of the portion of the curve above $y = P(0)$ to the entire area. We will denote the area between the curve and the line $y = P(0)$ by H and the ratio by A. Here it is implicitly assumed that the risk of the disease is at its minimum at $x = 0$ and monotonically increases with x. In practice, however, this may not always be the case. For example, the relationship between an exposure and the risk of some diseases may be a J or U-shape where the minimum risk occurs at an intermediate or middle value of the exposure and higher risks occur for individuals with very low or higher levels of exposure. For example, this type of risk relationship occurs for consumption of alcohol and cardiovascular mortality where drinking one to four drinks per day has the lowest risk (Gunzerath et al., 2004). In such cases, the attributable risk could be defined relative to the exposure level corresponding to the minimum risk. When the exposure is beneficial and the risk of disease increases montonically in x, e.g., a nutrient reduces the risk (e.g., see

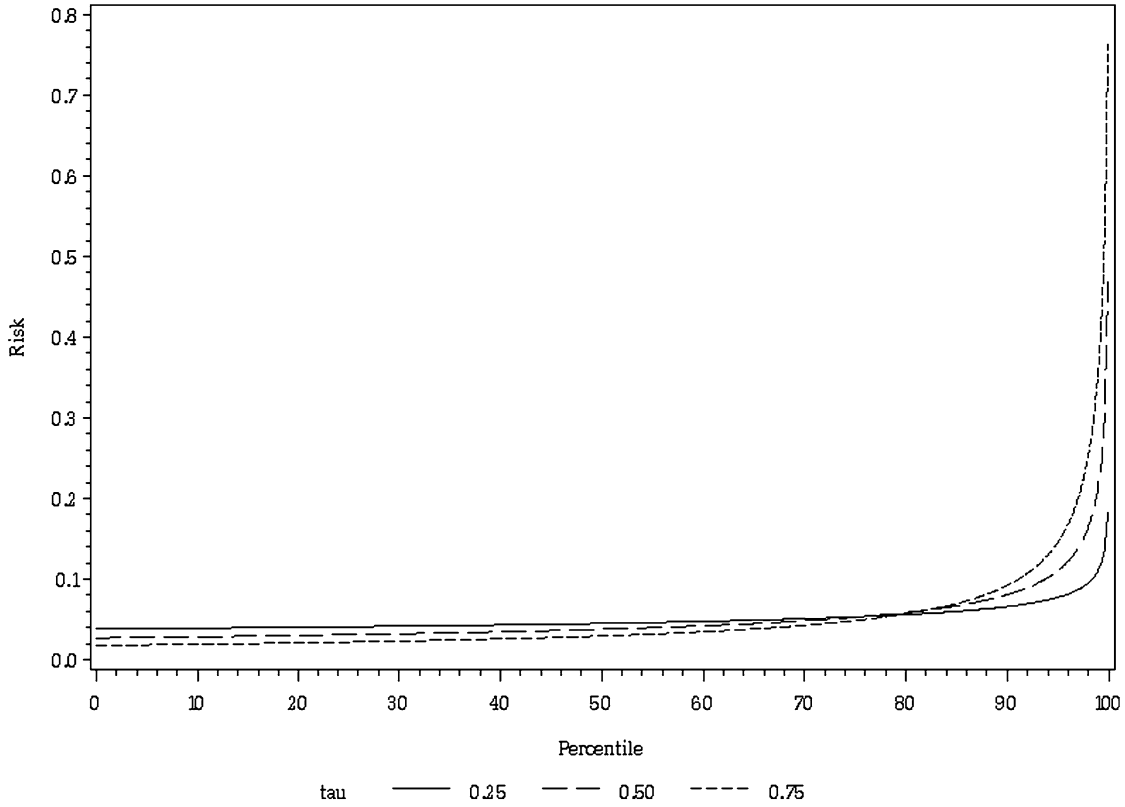

Figure 1. Risk percentile curves for logistic regression model with exponentially distributed exposure.

our example for red meat and all-cause mortality, Figure 2a), the area between the curve $RPC(q)$ and the line $y = P(0)$ is called the prevented fraction (Benichou, 2001). If everyone in the population had the maximum level of this exposure $x_{\max}$, their risk would be $RPC(1)$ and the maximum preventable fraction would be $[RPC(1) - RPC(0)]/RPC(1)$.

Since case-control studies enable one to estimate the odds-ratio, which for rare diseases is a good approximation to the relative risk (Cornfield, 1951), the above measures can be estimated because both the population attributable risk and the preventable fraction are functions of the relative risk. Indeed, if one multiplies the RPC by $1/P(0)$ one obtains a similar curve for the relative risk (RRPC). The population attributable risk is given by the ratio (A) of the area between the curve and $y = 1$ to the total area under the curve.

For cohort data that fits the proportional hazard regression model, there are analogs of the RRPC in which the corresponding ratio of areas A can be interpreted as an attributable risk. Explicitly, the proportional hazard model for time to event data (T) can be written as

$$P(T = t|T \geq t, X = x) = \lambda_0(t)R(x),$$

where $\lambda_0(t)$ is the baseline hazard function, i.e., the hazard function for the unexposed individuals ($x = 0$) and $R(x)$ is the relative hazard or relative risk function. If we draw the RRPC corresponding to the relative risk function $R(x)$, the ratio of the areas between the RRPC and 1 and the total area under the RRPC is given by

$$\frac{\int R(x)dF(x) - 1}{\int R(x)dF(x)}, \tag{1}$$

where the distribution of the risk factor, X, in the population is given by $F(x)$. When the incidence of the disease is small so that $F(x|T \geq t)$ is approximately $F(x)$, then multiplying the numerator and denominator of (1) by $\lambda_0(t)$ we obtain the following interpretation of the formula for the area ratio, A:

$$\frac{\int [P(T = t|T \geq t, X = x) - P(T = t|T \geq t, X = 0)]\, dF(x)}{\int P(T = t|T \geq t, X = x)dF(x)}. \tag{2}$$

The denominator is the average instantaneous probability of an event at time t, given survival until t, averaged over the exposure distribution. The integrand of the numerator is the difference between this instantaneous probability and the corresponding instantaneous probability for unexposed individuals. The numerator of (2) is this integrand averaged over the exposure distribution. Thus, the ratio, (2), of the areas represents the attributable risk of an event at time t given survival until t, i.e., the fraction of cases occurring at time t, are attributable to the risk factor X. The propor-

tional hazards model implies that $R(x)$, the risk function, is constant with respect to time of follow-up. i.e., remains the same over time so that the attributable risk in (1) does not depend on follow-up time.

In practice, often it is not feasible to completely eliminate an exposure from a population, and it is useful to assess the benefit of reducing the exposure to an attainable level that will result in a public health benefit. The potential reduction in the population attributable risk if the exposure of individuals in the top $100 \times (1-q)$ percentiles of the population is reduced to that of the $100 \times q$-th percentile can be defined as the area, $A(q)$, under the RPC (or RRPC) over the interval q to 1 that is formed by the region between the curve and the line $y = RPC(q)$ (or $y = RRPC(q)$). This measures the reduction in risk that would result from a proposed regulation that reduces the maximum exposure to x_q and may be more relevant for public health purposes than the usual attributable risk since it is often unrealistic to completely eliminate an exposure. This leads to the consideration of the Population Attributable Risk Reduction Curve, $PARRC(s)$, which express the potential reduction in the population attributable risk if the exposure of individuals in the top 100s% of the population is reduced to that of the $(1-s)$th percentile. For example if $s = .1$ then all individuals in the upper 10^{th} percent would have exposure reduced to the 90^{th} percentile. Formally, let $s = 1-q$ and $PARRC(s) = A_q/\mu_p$, where $\mu_p$is the total area under the RPC. For purposes of health policy the proportional reduction in the excess risk due to the exposure is of interest. This Attributable Risk Reduction Curve (ARRC) is simply obtained by replacing μ_p by $\mu_p - 1$ in the PARRC. Another advantage of the ARRC or PARRC is that their curvature indicates where the greatest benefit from reducing exposure levels is obtained. The ARRC will be illustrated Section 4.2 when we discuss the homocysteine and cervical cancer example.

3. ESTIMATING THE RISK PERCENTILE CURVE

In order to estimate the RPC, one first fits the risk model in the original scale of the exposure. In this step one can use standard methods, parametric or non-parametric, to obtain an estimate $\hat{P}(x)$ of the risk curve $P(x)$ at x. The RPC at the q-th percentile now can be estimated as $\hat{RPC}(q) = \hat{P}(\hat{x}_q)$, where $\hat{x}_q$ is the sample q-th percentile of the exposure of interest. The percentiles of the exposure can be estimated either from internal or external data. For cohort studies, one can use the empirical distribution of the exposures among the study subjects to internally estimate the percentiles. When the rare disease assumption is valid, the percentiles of exposure in the population are obtained from the empirical distribution based in the controls. The risk percentile curve and the associated attributable risk for the population under study can also be estimated if external data on the exposure distribution are available for the same population, e.g., from a census or a large scale survey. In this situation, the estimate of the risk curve from the current study and the estimates of the percentiles from the external data can be combined to estimate the risk percentile curve in the underlying population. External standardization can also be used to obtain an estimate of the RPC for a different population than the one being studied. While the exposure distribution in a different population can be quite different from that of the study population it often is biologically plausible that the relationship between the exposure and the disease risk is the same in the two populations. In this case, one can project the risk curve from the population being studied to another population by standardizing the estimated risk curve of the former with the exposure percentiles of the later. However, one needs to exercise care when extrapolating if the exposure patterns in the two populations differ so substantially that there is little overlap in the exposure distributions of the two populations.

In this paper to estimate the RPC, we utilize restricted cubic regression splines with fixed knots (Durrleman and Simon, 1989). Use of spline regression for epidemiologic data has been advocated by various researchers to study risk of disease as a function of a continuous exposure of interest (Greenland, 1995). This method has several advantages over two most commonly used methods for the analysis of continuous exposure data in practice. These are (1) grouping continuous exposure variables into a few categories and assume the disease risk is constant in each of these categories; (2) assume that the effect of the exposure is linear on the disease risk in an appropriate scale. While the first approach loses power by ignoring the variation of risk within the defined categories, the linearity assumed in the second approach may not be sufficiently flexible to capture all the important aspects of the disease exposure relationship. Furthermore, even if the assumed linearity is essentially correct for the main body of exposure levels, departure from linearity or the assumed parametric form in extreme ranges may unduly influence the estimate of the regression parameters. This occurs because the extreme observations are overweighted in the estimation of the regression parameter, so either non-linearity in the extreme region or measurement error may lead to an erroneous estimate. This problem occurs frequently in nutritional studies as some extremely high or low intake values are often recorded. Whether they are real or due to measurement error, it is preferable to reduce their influence on the shape of the risk curve in the region containing the bulk of the data.

4. EXAMPLES OF RISK PERCENTILE CURVES

4.1 Diet and mortality example

This section applies the spline method to estimate the risk percentile curves and related measures to two data sets. The first example is a prospective cohort study of the relationship of dietary red meat consumption to all-cause mortality. The National Cancer Institute followed a cohort of 617, 119 men and women in the NIH-ARRP Diet and Health

Study from October 25, 1995 until December 31, 2005. These were the subjects who returned the baseline questionnaire, which provided data about diet from food frequency questions and about other risk factors such as smoking. Sinha et al. (2009) investigated the association of all-cause mortality and dietary red meat intake separately among men and women. A significant increased risk of mortality with increasing levels of red meat intake was found separately in men and women. To illustrate the risk percentile curve, we examine the subgroup of 85,781 white men who never smoked to control for race and confounding due to smoking status, which was the major confounder in the Sinha et al. (2009) analysis. Age is adjusted in our analysis by using it as the time metric in Cox proportional regression analysis to estimate the relative hazards (Korn, Graubard and Midthune, 1997). During follow-up of this subgroup 7,954 men died.

Since the effect of red meat consumption may also be affected by the total calorie intake, we follow Sinha et al. (2009) and consider the risk factor to be the ratio of daily dietary red meat intake in grams divided by daily total calorie intake per 1000 kilo calories (DM). This method is a standard approach of pre-adjusting for total calorie intake used in investigating the effect of nutrients on disease risk and is called the nutrient density energy adjustment method (Willett, Howe and Kushi, 1997; Brown et al., 1994). In Figure 2a the relative risk curve in the original scale and in Figure 2b the risk percentile curve are given. The curves are obtained by first fitting a proportional hazards model, where the effect of DM is modeled by spline methods. A cubic spline, constrained to be linear in the two extreme intervals, was chosen with knots at the following percentiles, 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} (Durrleman and Simon, 1989).

Examining the relative risk curve as a function of DM, the RRPC suggests that the risk of mortality increased slowly with DM until DM reaches about the 95%. After 95%, the risk of mortality seems to increase rapidly with DM up to 100%. As mentioned earlier, a problem with plotting the relative risk curve in the original scale is that the long tail of an exposure distribution can have a disproportionate influence on the plot. In Table 1, for example, one can see that the highest 1 percent of DM values (101.2–220.6) correspond to about 50% of the total range of DM (0.0–220.6). This problem does not arise in the RRPC as this region represents only 1% of the total data. The risk percentile curve indicates that most of the increased risk is seen only in the upper 5 percent of the data. Before that the risk remains fairly flat, only rising from 1.0 to 1.6 for most of the distribution. Thus the RRPC tells us that the relative risk goes from 1.6 to 3.1 between the 95^{th} and 100^{th} percentile. The estimated attributable risk, the fraction of the total area under the curve that corresponds to a RR > 1, is 0.10. This implies that if red meat consumption among nonsmoking white men can be reduced to that of the non-red meat eaters then approximately 10 percent of mortality at each age would be eliminated. Another insight obtained from the RRPC is that in Western populations, where the overwhelming majority (upper 95 percentiles) of nonsmoking white men have red meat/energy consumption levels that put them in the "flat" region where the risk is less than 1.6, it will be difficult for the typical epidemiologic study, which is much smaller than the sample of nearly 100,000 in the NIH-AARP Diet and Health Study, to detect an association between mortality and red meat consumption, if indeed one exists.

Table 1. Percentiles of daily dietary red meat intake from white men in the NIH-ARRP Diet and Health Study and age-adjusted serum homocysteine level from women in the Third National Health and Nutrition Examination Survey

Percentiles (%)	Daily Dietary Red Meat Intake (grams/1000 kcal)	Age-Adjusted Serum Homocysteine Level[1]
100 Maximum	220.618	4.114
99	101.208	2.514
95	75.882	1.859
90	64.400	1.600
75	48.441	1.239
50	33.459	1.000
25	20.610	0.821
10	10.811	0.671
5	6.082	0.590
1	1.125	0.474
0 Minimum	0.000	0.278

[1]The 13 highest values of age-adjusted serum homocysteine were deleted because of extreme values.

4.2 Serum homocysteine and cervical cancer example

The second example uses data from a multi-center population-based case-control study in which Weinstein et al. (2001) investigated the relationship of serum homocysteine levels to invasive cervical cancer. The cases consisted of all subjects with histologically confirmed, primary invasive cervical cancer incident cases occurring between April 1982 and January 1984 in five US study sites that reported to the Comprehensive Cancer Patient Data System. Random digit dialing was used to select two controls, matched on ethnicity and telephone exchange. The same sample as Weinstein et al. (2001), which consisted of 183 cases and 540 controls, will be re-analyzed here. For purposes of illustration, we will ignore the matching in the analysis. Weinstein et al. (2001) found a statistically significant association of increasing risk of invasive cervical cancer with increasing levels of serum homocysteine levels among their sample of women aged 19 to 74 years.

Here, we illustrate the standardized RRPC and ARRC. The serum homocysteine data collected from a nationally representative sample of women in the Third National Health and Nutrition Examination Survey (NHANES III) will provide the data for the standardization. NHANES III has a medical examination component that included drawing blood and measuring serum homocysteine levels for

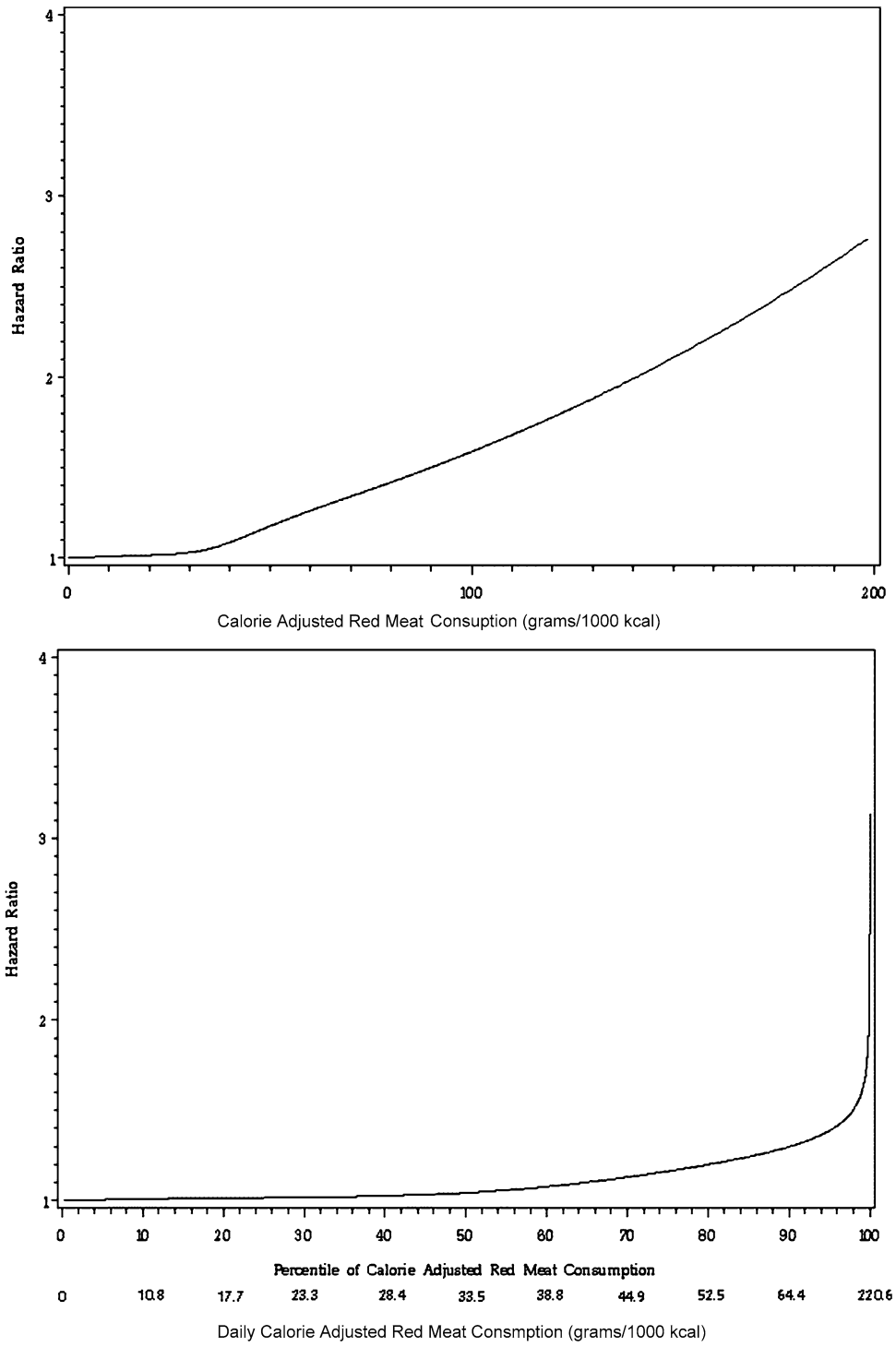

Figure 2. Relative risk of all-cause mortality associated with red meat (daily dietary red meat intake in grams divided by daily total calorie intake per 1000 kilo calories) shown in original scale (upper panel, Figure 2a) and percentile scale (lower panel, Figure 2b) with the original scale values displayed below the percentiles.

a random subsample of its participants. The case-control study used the same laboratory and procedures to obtain the homocysteine measurements as was used in NHANES III. Since NHANES III is a weighted sample, all the analyses that use NHANES III will be weighted by its sample weights.

Because the effect of serum homocysteine may be confounded by age, we pre-adjust homocysteine for age as fol-

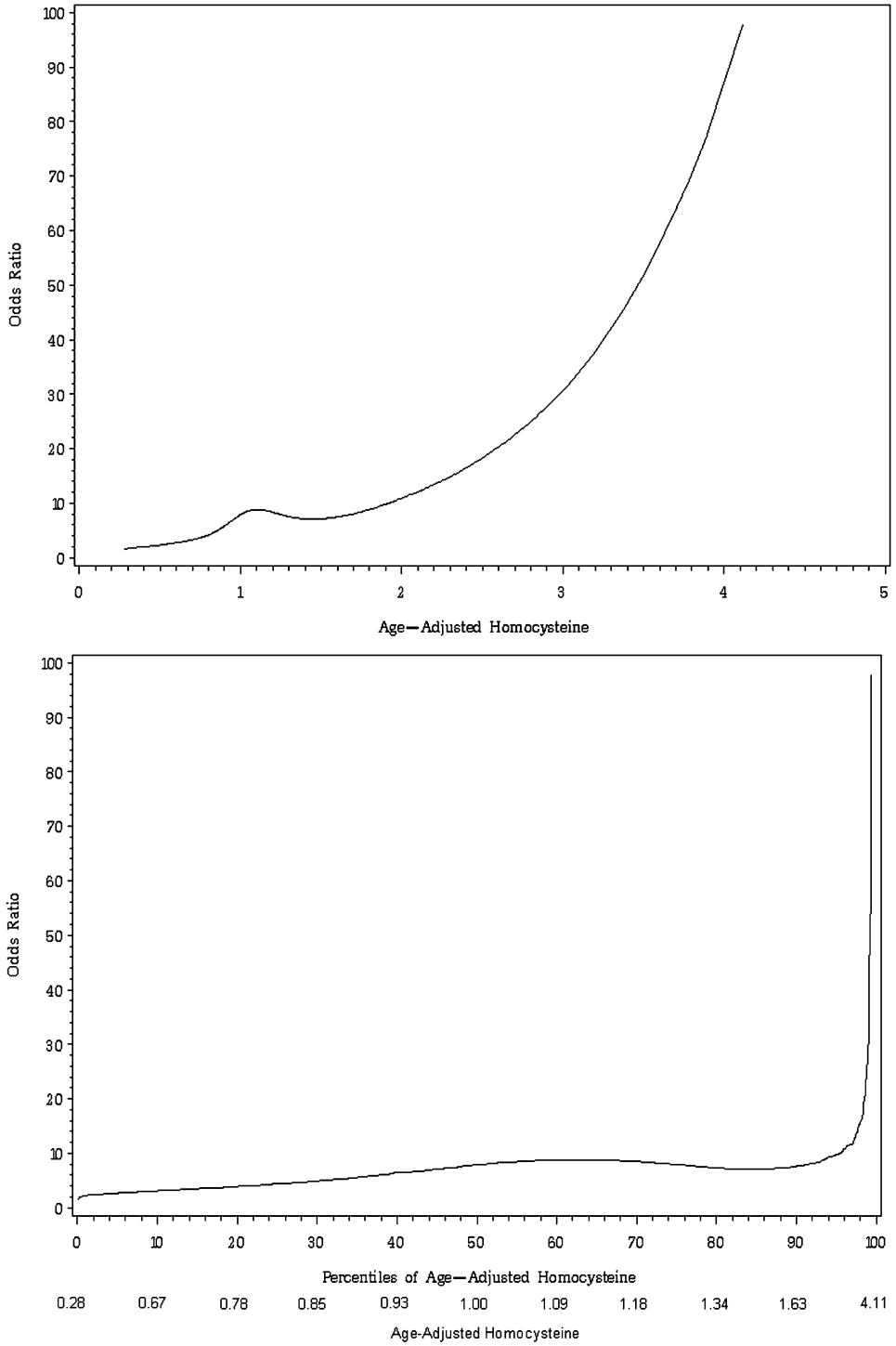

Figure 3. Relative risk of cervical cancer associated with serum homocystiene (age-adjusted) shown in original scale (upper panel, Figure 3a) and percentile scale (lower panel, Figure 3b) with the original scale values displayed below the percentiles.

lows. Using only the NHANES III data we estimate the conditional median serum homocysteine given age by applying a kernel smoothing method for percentiles of weighted survey data (Korn and Graubard, 1999, pp. 86–89). The age-adjusted homocysteine (AAH) values for the case–control study are computed by dividing each homocyteine observation from the case-control study by the conditional median value of homocysteine from the NHANES III, i.e., using the median level of homosysteine at the same age as the observation.

Figure 3a presents the relative risk curve (approximated by odds ratios) in the original scale, and Figure 3b presents

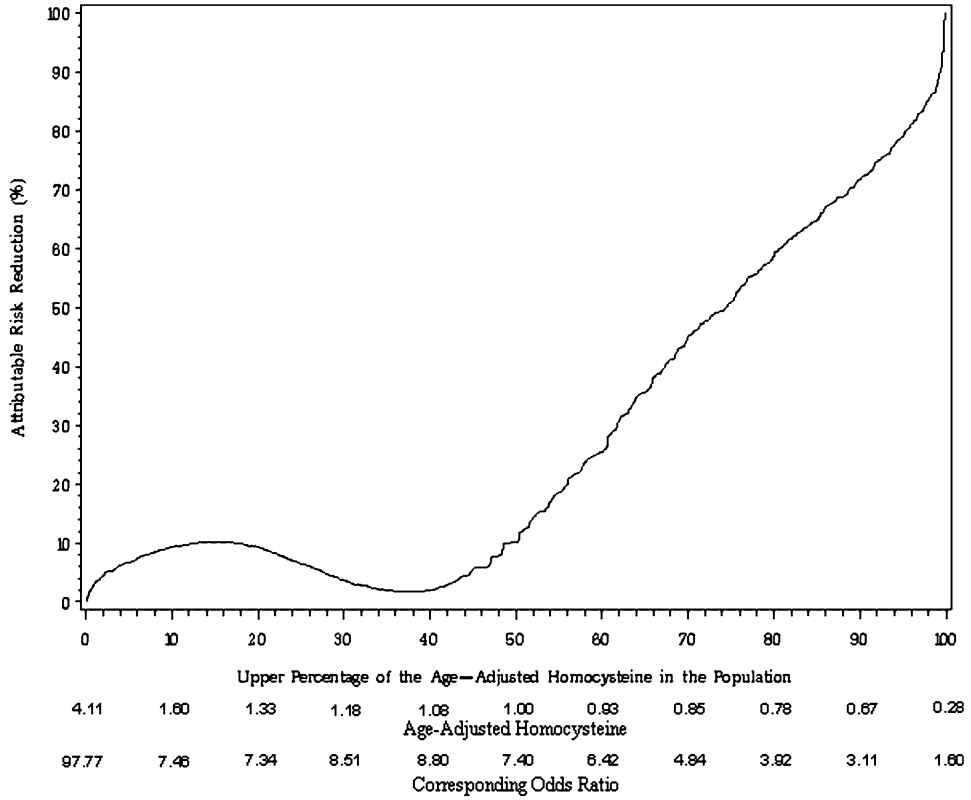

Figure 4. Attributable risk reduction curve (ARRC) of cervical cancer with upper percentage of serum homocysteine (age-adjusted) in the population and with the original scale values displayed below the percentages and the corresponding odds ratios below the original scale values.

the standardized RRPC for the age-adjusted values where the percentiles are taken from the distribution of AAH from NHANES III. Theses curves are obtained by first fitting a logistic regression model, where the effect of AAH is modeled by a 5-knot spline as in the diet and all-cause mortality example. Using this risk model the relative risk (odds ratio) is plotted against the values of AAH (Figure 3a) and then against the percentiles of AAH from the NHANES III sample (the percentiles from the NHANES III are weighted to estimate the percentiles in the US population). Although not shown here, the RRPC, using percentiles of AAH derived from only the controls, is very similar to Figure 3b.

The RRPC shows a monotonic increase in the relative risk of cervical cancer with AAH until AAH reaches the 60^{th} percentile, where the relative risk levels off and then slowly decreases until about the 85^{th} percentile; after that the relative risk increases sharply. As was the case for the meat and dietary example, the long tail in the homocysteine exposure distribution has a disproportionate influence on the relative risk curve (Figure 3a). In Table 1, for example, one can see that the highest 1 percent of AAH values (2.51–4.11) correspond to about 42% of the total range of AAH (0.28–4.11). Again this problem does not arise in the RRPC as this region represents only 1% of the total data. The risk percentile curve indicates that most of the increased risk is seen only in the upper 4 percent of the data. Before that the risk gradually rises from about 1.6 to about 10.0 for most of the distribution. Thus the RRPC tells us that the relative risk goes from 10.0 to 97.8 between the 96^{th} and 100^{th} percentile. The RRPC in Figure 3a better reflects the population impact of the levels of the AAH on cervical cancer whereas the relative risk curve distorts this relationship.

The corresponding ARRC is displayed in Figure 4, which is based on percentiles of AAH from the NHANES III. The ARRC shows where there is a large potential reduction in the population attributable risk by lowering the maximum level of homocysteine to a particular percentile. The ARRC rises at the very highest percentiles of AAH but then falls at about the upper 15 percent of AAH until about the upper 38^{th} percent point, which reflects the decreasing odds ratio in the RRPC (Figure 3b). The ARRC increases after the upper 38^{th} percent of AAH. The ARRC tells us that individuals in the upper 80 percent of the AAH would have to reduce their AAH to the 20^{th} percentile in order to reduce the excess risk by 50%.

5. DISCUSSION

The use of the full percentile curve provides more information about the relationship of exposure to disease risk than the common method of grouping the data by specific percentiles, e.g. quartiles. Compared to a plot of the risk as a function of the original exposure levels, the RPC has the advantage of not being dominated by a small fraction of data in the tails of the exposure distribution. As epidemiological data is often right skewed, the RPC enables one to focus attention on the exposure data pertaining to the large majority of the population. Moreover, areas under this curve correspond to established public health measures, e.g., attributable risk or the related new attributable risk reduction curve defined here.

Both the RPC and ARRC can be calculated separately for well-defined sub-groups of the overall population. These curves may differ noticeably for a variety of reasons: the exposure distributions can be different or the underlying relationship between risk and exposure might be different in the subgroups, perhaps due to different behavioral factors, environmental exposures or genetic characteristics. Further exploration of the reasons underlying such observed difference may lead to the adoption of better strategies for risk reduction in important sub-groups of the population.

In this paper we used pre-adjustment and stratification methods to account for some major covariates that might confound the effect of the exposure of interest. Further research concerning alternative methods of adjustment is needed when several important covariates should be considered.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Stephanie Weinstein and Rashmi Sinha and the NIH-ARRP Diet and Health Study for making their data available for use in our examples and a referee for helpful comments.

Received 4 December 2008

REFERENCES

Benichou, J. (2001). A review of adjusted estimators of attributable risk. *Statistical Methods and Medical Research* **10**(3) 195–216.

Brown, C. C., Kipnis, V., Freedman, L. S., Hartman, A. M., Schatzkin, A., Wacholder, S. (1994). Energy adjustment methods for nutritional epidemiology: the effect of categorization. *American Journal of Epidemiology* **139**(3) 323–338.

Cowell, F. A. (1995). *Measuring Inequality*. London: Prentice Hall/ Harvestor.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Application to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute* **11** 1269–1275.

Durrleman, S., Simon, R. (1989). Flexible regression models with cubic splines. *Statistics in Medicine* **8**(5) 551–561.

Eide, G. E., Heuch, I. (2001). Attributable fractions: fundamental concepts and their visualization. *Statistical Methods in Medical Research* **10**(3) 159–193.

Gastwirth, J. L. (1972). The estimation of the Lorenz curve and Gini index. *Review of Economics and Statistics* **54** 306–316. MR0314429

Gastwirth, J. L., Miao, W. W., Zheng, G. (2003). Statistical issues arising in disparate impact cases and the use of the expectancy curve in assessing the validity of pre-employment tests. *International Statistical Review* **71**(3) 565–580.

Greenland, S. (1995). Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology* **6**(4) 356–365.

Gunzerath, L., Faden, V., Zakhari, S., Warren, K. (2004). National Institute on Alcohol Abuse and Alcoholism Report on moderate drinking. *Alcoholism Clinical and Experimental Research* **28**(6) 829–847.

Korn, E. L., Graubard, B. I. (1999). *Analysis of Health Surveys*. New York, Wiley.

Korn, E. L., Graubard, B. I., Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up for a survey: choice of the time scale. *American Journal of Epidemiology* **145** 72–80.

Mahalanobis, P. C. (1960). A method of fractile graphic analysis. *Econometrika* **28** 325–351. MR0179865

National Center for Health Statistics (1994). Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988–94. *Vital and Health Statistics* **1** (32).

Pepe, M. S., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I. M., Zheng, Y. (2008). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology* **167**(3) 362–368.

Sinha, R., Cross, A. J., Graubard, B. I., Leitzmann, M. F., Schatzkin, A. (2009, in press). Meat intake and mortality: a prospective study of over half a million people. *Archives of Internal Medicine*.

Vogt, T. M., Ziegler, R. G., Graubard, B. I., Swanson, C. A., Greenberg, R. S., Schoenberg, J. B., Swanson, G. M., Hayes, R. B., Mayne, S. T. (2003). Serum selenium and risk of prostate cancer in U.S. Blacks and Whites. *International Journal of Cancer* **103** 664–670.

Weinstein, S. J., Ziegler, R. G., Selhub, J., Fears, T. R., Strickler, H. D., Brinton, L. A., Hamman, R. F., Levine, R. S., Mallin, K., Stolley, P. D. (2001). Elevated serum homocysteine levels and increased risk of invasive cervical cancer in U.S. women. *Cancer Causes and Control* **12**(4) 317–324.

Willett, W. C., Howe, G. R., Kushi, L. H. (1997). Adjustment for total energy intake in epidemiologic studies. *American Journal of Clinical Nutrition* **65**(4 Suppl) 1220S–1228S.

Zhao, L. P., Kolonel, L. N. (1992). Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. *American Journal of Epidemiology* **136** 464–474.

Nilanjan Chatterjee
Biostatistics Branch
Division of Cancer Epidemiology and Genetics
National Cancer Institute, Bethesda, MD

Barry I. Graubard
Biostatistics Branch
Division of Cancer Epidemiology and Genetics
National Cancer Institute, Bethesda, MD
E-mail address: graubarb@mail.nih.gov

Joseph L. Gastwirth
Department of Statistics
George Washington University
Washington, DC