Open Access Medical Statistics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

METHODOLOGY

Assessing the variability of the attributable causes of death

This article was published in the following Dove Press journal:
Open Access Medical Statistics
22 September 2011
Number of times this article has been viewed

Wenjiang J Fu[1]
Tianshuang Wu[2]
Yu Wang[3]
Haiying Meng[4]
Jianshi Huang[3]

[1]Department of Epidemiology, [2]Department of Mathematics, Michigan State University, East Lansing, MI, USA; [3]Department of Epidemiology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences; School of Basic Medicine, Peking Union Medical College, Beijing, People's Republic of China; [4]Center for Disease Control and Prevention of Chaoyang District, Beijing, People's Republic of China

Abstract: The study of attributable causes of death (ACD) provides a new venue to quantify the external (nongenetic) causes of mortality, and may guide policymaking to address emerging issues in public health by focusing on the largely preventable risk factors. Given such importance, systematic methods to assess the variability of the attributable number of deaths (AND), including the standard errors and confidence intervals, need to be developed. In this article, we develop two statistical methods of the estimation of the standard errors and confidence intervals for the ANDs, one using multinomial distribution and the other using bootstrap sampling, and study the effect of the size of the mortality through simulations. Both methods are easy to implement and provide valid and efficient estimation of the standard errors and confidence intervals. While AND estimates and their standard errors increase with the size of the mortality, the ratio of the standard error to the AND estimate decreases. We demonstrate the methods with two data sets, the US national mortality data during the year 2006 and the mortality data of Chaoyang district of Beijing, China during the year 2007. We conclude that assessment of the variability is needed for small size mortality as the uncertainty is relatively large, but not for large size mortality.

Keywords: attributable causes, bootstrap, confidence interval, mortality, population attributable fraction

Introduction

Premature death is one of the primary foci of public health concerns. Epidemiological studies of mortality data can help to guide policymaking to address emerging public health issues, by monitoring the dynamic trend of mortality of a given population over time and designing corresponding public health programs to reduce the risks that are largely preventable. So far, many intervention programs have proved to be successful, such as smoking cessation programs to prevent lung cancer, etc. However, it is known that diseases are often multifactorial, and that one single risk factor may contribute simultaneously to multiple diseases. Therefore, at the population level, intervention studies targeting a single disease may seem to be less effective than they actually are since other diseases not considered in the study programs may also benefit but are not accounted. For example, smoking cessation may prevent not only lung cancer, but also cardiovascular diseases.[1] Thus studies focusing on the primary prevention of lung cancer through smoking cessation programs may account for less benefit than its actual effect since the benefit on cardiovascular diseases is not accounted. Better understanding of the mechanism of how the risk factors lead to the diseases and death, particularly answers to the questions: "How many risk factors contribute to each disease?", "How do they interact?" and "How many diseases does each risk

Correspondence: Wenjiang J Fu
Department of Epidemiology, Michigan State University, West Fee Hall, Suite B601, East Lansing, MI 48824 USA
Tel +1 517 353 8623 ext 113
Fax +1 517 432 1130
Email fuw@msu.edu

factor contribute to?" will surely increase the likelihood of the success of the intervention programs. To achieve this, one needs valid estimation of the risk factors to each specific disease and also needs to assess the weight or proportion that is attributable to the risk factors.

In two pioneering articles, the attributable causes of death (ACDs) were studied for the US national mortality data.[2,3] In these two articles, the authors named the external (nongenetic) risk factors of death or the ACDs, the actual causes of death. A recent ACD study incorporated further dietary factors into consideration.[4] These studies identified and quantified the major risk factors leading to a large portion of the US national mortality, including high blood pressure, tobacco use, alcohol consumption, poor diet and physical inactivity, etc. It was also noticed that these ACDs are largely preventable through intervention programs.

To identify and quantify the risk factors, one needs to study the mortality data and the population attributable fractions (PAFs) of the multiple risk factors for each given disease leading to death.[2–4] For a given population, the attributable number of deaths (AND) of a specific risk factor can be computed with the PAFs estimated with relative risks or the odds ratio of the diseases[3,4] together with the population mortality data by disease, and a list of leading ACDs can then be generated.[2–4] The PAFs have been studied extensively for various types of study designs, including retrospective studies adjusting for confounding factors,[5] studies with stratified data,[6] studies based on logistic regressions,[7] unmatched case–control studies using logistic regression,[8] studies of fitted incidence ratios and exposure survey data,[9] and case–control studies with matched pairs,[10] etc. Based on the above methods, the World Bank conducted a risk factor study for many diseases leading to death among the worldwide population by gender, age group, and geographic region.[11]

Although the mortality of large geographic regions, such as a metropolitan, a state or a country, is very often of interest to investigators, small geographic regions or regions with sparse population may also be of interest.[6] While the mortality in large populations and the estimated ANDs are in general stable and have relatively small variation, those in moderate size populations or smaller may present large variation. Thus it is of importance to study how the ANDs vary with the population size. Although the variability of the PAFs has been studied,[6–10] that of the ANDs has not been studied systematically. Specifically, the variability of the ANDs was not assessed in the two studies of US mortality[2,3] although large variability was not expected for such large mortality data. While confidence intervals were reported through a simulation approach accounting for the variability from the PAFs based on several assumptions of the distributions of the uncertainties, including normal, Chi-squares, binomial, and log-normal distributions,[4] the variability from the mortality data was not accounted for. In theory, the variability of the ANDs may be attributed to either the PAFs estimation, or the mortality, or both. In practice, it is even more important to assess the variability from the mortality data because the mortality data may vary largely from state to state, or from city to city, while the PAFs remain stable for the same population within a country or a geographic region. In particular, it is extremely important to gain a priori knowledge about when an assessment of the uncertainty or variability is needed and how the variability varies with the population size. It is noted that since the mortality data are archived data, their distribution may vary with the population of interest. Thus, it may be less apparent to assess the variability of the mortality and of the ANDs.

In this article, we develop two methods to assess the variability of the ANDs, a parametric method using multinomial distribution of the mortality data and a nonparametric method using bootstrap sampling method. This work is of particular importance in determining if the AND estimates require variability assessment and for what population size it becomes necessary. Through simulation studies, we examine the effect of the mortality size on the AND estimates and the variability. We illustrate the method with two mortality data sets and provide the estimate, standard error, and confidence interval for the leading ACDs of the mortality.

Methods

Mortality data

The national mortality data of the US during 2006 were obtained from the National Vital Statistics Reports on deaths for the year 2006 published by the Centers for Disease Control and Prevention.[12] Among all the diseases, we selected 16 leading causes of death (shown in Table 1) that have the PAFs in the World Bank report.[11] We used the mortality data of these 16 diseases only in this study.

The mortality data of Chaoyang District of Beijing, China during 2007 were analyzed and 19 diseases as the leading causes of death were reported in Wang et al (2009) for the study of actual causes of death in Chaoyang district of Beijing.[13] These diseases were selected to match the PAFs reported in the World Bank Report. Among them were malignant neoplasms, diseases of the heart, cerebrovascular

Table 1 Sixteen leading causes of death in the US during year 2006[a]

Disease	Number of deaths
Ischemic heart disease	425425
Malignant neoplasms of trachea, bronchus, and lung	158664
Cerebrovascular diseases	137119
Diabetes mellitus	72449
Malignant neoplasm of colon, rectum and anus	53549
Malignant neoplasm of breast	41210
Malignant neoplasm of pancreas	33454
Hypertensive heart disease	29788
Leukemia	21944
Malignant neoplasm of liver and intrahepatic bile ducts	16525
Other chronic liver disease and cirrhosis	14505
Malignant neoplasm of bladder	13474
HIV/AIDS	12113
Malignant neoplasm of stomach	11345
Malignant neoplasm of corpus uteri and uterus, part unspecified	7384
Malignant neoplasm of cervix uteri	3976

Note: [a]These 16 diseases were chosen from the National Vital Statistics Report to match their PAFs in the World Bank Report.

disease, diabetes mellitus, chronic obstructive pulmonary diseases, accidents, pneumonia and influenza, chronic liver disease and cirrhosis, infectious diseases, etc.

Statistical analysis

The AND (N_f) of each risk factor *f* in a given population was calculated by

$$N_f = \sum_i n_i a_i, \quad (1)$$

ie, the number of deaths n_i of each disease *i*, which is likely to be causal to the deaths, multiplied by its PAF (a_i).[3,4,13] The a_i is the PAF to the given risk factor by disease *i* under consideration. This multiplication yielded the AND of the risk factor under consideration for disease *i*. The summation was over all possible diseases that were partially attributed to the risk factors of consideration, and yielded the total number of deaths in a given population attributable to the risk factor of consideration, namely, the AND.

To calculate the standard error and the confidence interval attributable to the mortality, we noticed that the counts of deaths from *K* different diseases (n_1, …, n_K) follow a multinomial distribution for a given fixed total number of deaths $N = n_1 + \cdots + n_K$ with probability (p_1, …, p_K) estimated by (n_1/N, …, n_K/N). The variance of n_i is $\text{var}(n_i) = Np_i(1-p_i)$ and the covariance between n_i and n_j is $\text{cov}(n_i, n_j) = -Np_ip_j$ for two different diseases *i* and *j*. Therefore, for a given risk factor of interest, let A = (a_1, …, a_K) denote the PAFs for the *K* diseases. The AND $N_f = \sum_i n_i a_i$ has a variance of a quadratic form of a_i and p_i for given fixed PAFs A:

$$\sum_{i=1}^{K} a_i^2 Np_i(1-p_i) + \sum_{\substack{i\neq j \\ i,j=1}}^{K} a_i a_j(-Np_ip_j) = N\left[\sum_{i=1}^{K} a_i^2 p_i(1-p_i) + \sum_{\substack{i\neq j \\ i,j=1}}^{K} a_i a_j(-p_ip_j)\right] \quad (2)$$

which is increasing with N. This provides an easy-to-implement parametric method to calculate the variance of the AND of each risk factor, and the 95% confidence interval can be calculated as (N_f – 1.96 se, N_f + 1.96 se), where se is the standard error of the N_f, and is calculated as the square-root of var(N_f). It is noted that equation (2) only accounts for the variation in the mortality data, but not in the PAFs.

The standard error and the confidence interval of the AND can also be calculated through the nonparametric bootstrap method with 200 bootstrap samples.[14] Each bootstrap sample was obtained through random sampling from the original data. Let S = {Y_1, Y_2, …, Y_K} be a collection of counts of deaths from K diseases, with Y_k = {k, …, k} being indices of n_k deaths from disease k, k = 1, …, K. The total number of deaths in the given population was $N = n_1 + n_2 + \cdots + n_K$. Let S_b be a bootstrap sample of S with b = 1, …, B. S_b was obtained by the bootstrap sampling: a random sample {1, …, 1, 2, …, 2, …, K, …, K} of S with replacement. The following procedure was carried out to compute the bootstrap standard error and the confidence interval. For each bootstrap sample S_b, the AND to the risk factor of consideration N_{fb}, was calculated. The standard error of the AND (N_f) was computed to be the standard deviation of {N_{fb}, b = 1, …, B} with B = 200. That is, $SE(N_f) = SD(N_{f1}, \ldots, N_{fB})$. The lower and upper bounds of the 95% confidence interval were computed to be the 0.025 and 0.975 quantiles of {N_{fb}, b = 1, …, B} with B = 200, respectively. We noticed that in general, 200 bootstrap samples is enough to yield accurate estimation for the variance and confidence interval.[14]

Simulation study

Ten diseases and ten risk factors were assumed in the simulation study. The number of deaths from each disease was specified to be random integers (10000, 14210, 23430, 253210, 34550, 634630, 34530, 25220, 23420, 84540). The PAFs were randomly generated fractions between 0 and 1 with the sum of the ten PAFs for each disease less than or equal to 1, and were kept fixed for the assessment of the variability attributable to the mortality. In the simulations, the sum of PAFs of a disease less than 1 indicates that some risk factors, that the disease is attributable to, may not be known or considered in the study.

To examine the effect of the population size, we computed the AND estimates and their standard errors in a simulation study by varying the mortality size with a varying multiplier C valued at 0.0001, 0.001, 0.01, 0.1 and 1 so that the structure of the mortality data remains the same as real mortality data. For each value of C, we computed $SE(N_f)$ for a given risk factor using the above methods with B = 200 for the bootstrap sampling, and the standard error $SE(N_f)$ was computed for all ten risk factors. We report the AND estimate N_f, its standard error $SE(N_f)$ and the ratio $SE(N_f)/N_f$ by the varying multiplier C. We also compare the two methods of computing the standard errors for the simulated data with C = 1 using the scatter plot. The software package R (R Development Core Team; Auckland, New Zealand) was used in this study for the statistical computation.

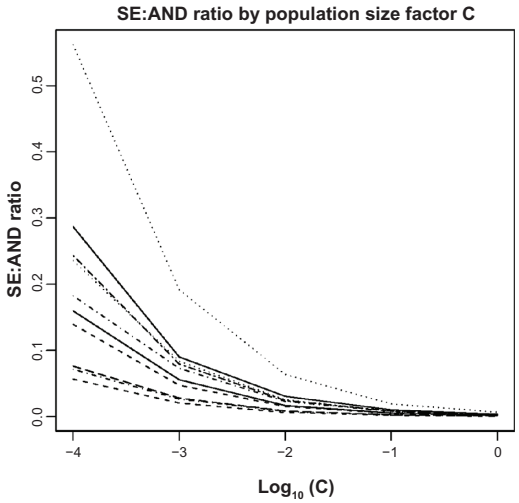

Figure 2 Plot of SE:AND ratio of ten risk factors in the simulation study.

Results

Effect of the mortality size on AND and the variability

The AND estimate N_f increased proportionally with the constant C through the size N of the mortality as in the definition, and the standard error $SE(N_f)$ also increased as in equation (2). Figure 1 illustrates the AND estimates of the ten risk factors (Figure 1A) and their standard errors (Figure 1B) by the multiplier C from 0.0001 to 1 in log scale. It is noticed with equation (2) that the standard error increases proportionally with the square-root of N. While the increase of the standard error can be explained by the fact that the larger the mortality data, the larger the variability, this variability does not reflect the significance of the AND estimates since the AND also increased. In order to assess the significance of the estimates, one needs to examine the relative variability, eg, the ratio of the standard error to the estimate, which is inversely proportional to the square root of mortality size. We thus examined the ratio $SE(N_f)/N_f$ with the multiplier C. Figure 2 illustrates the ratio $SE(N_f)/N_f$ for these ten risk factors. It shows that unlike the estimates and their standard errors, the ratio $SE(N_f)/N_f$ decreased with the scaling parameter C. This clearly indicates that the relative variability of the ACD estimates decreased quickly with the mortality size, and reflects the fact that the larger the sample size, the smaller the (relative) variability. Figure 3 illustrates the comparison of the standard errors $SE(N_f)$ between the bootstrap method and the multinomial distribution method. It is clear that the standard errors barely differ between these two methods.

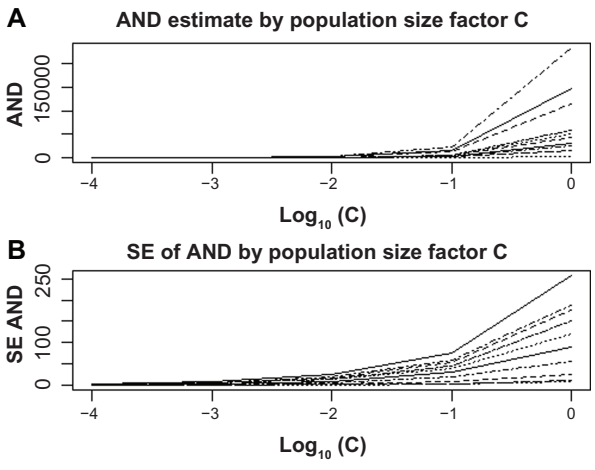

Figure 1 Plot of ANDs of ten risk factors and their standard errors in simulation study. (**A**) Plot of ten ANDs by the population multiplier C at logarithmic scale. (**B**) Plot of the SE of the ten ANDs by the population multiplier C at logarithmic scale.
Abbreviations: AND, attributable number of deaths; SE, standard error.

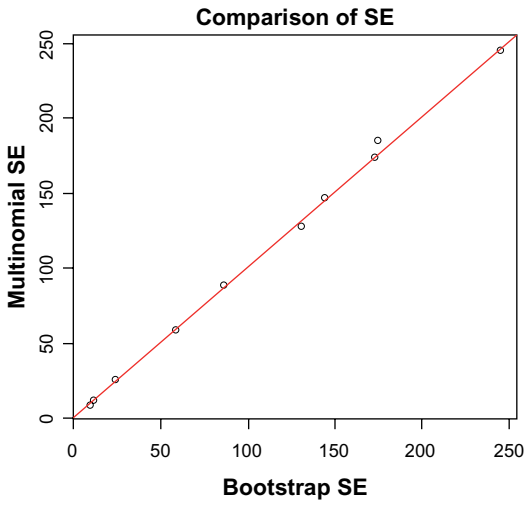

Figure 3 Scatter plot comparing the standard errors $SE(N_f)$ between the bootstrap method and the multinomial distribution method for the simulated data with C = 1. The red line is the diagonal as a reference for the comparison.

Table 2 Leading attributable causes of death of the US mortality during year 2006

Attributable cause of death (ACD)	Attributable number of deaths (AND)		
	Estimate	SE	95% CI
High blood pressure (HBP)	298146.5	342.1	(297542.8, 298629.3)
High cholesterol (HCL)	230429.8	266.5	(229903.3, 230825.3)
Overweight and obesity (ONO)	195734.6	240.8	(195390.0, 196062.9)
Smoking	178081.0	270.5	(177600.1, 178546.1)
Low fruit and vegetable consumption (LVC)	164144.8	168.9	(163908.3, 164369.1)
Physical inactivity (PhInA)	118472.3	123.8	(118296.8, 118631.7)
Alcohol use	64339.8	93.7	(64191.7, 64481.3)
Unprotected sex	14727.9	121.8	(14491.4, 14970.2)
Urban air pollution	6347.4	14.9	(6317.9, 6376.3)
Illicit drug use	1088.3	10.9	(1067.7, 1110.9)
Contaminated injections in healthcare settings (CIHS)	596.2	3.1	(590.8, 601.7)

Abbreviations: CI, confidence intervals; SE, standard error.

Leading ACDs of the mortality data

National mortality of the US during 2006

The leading ACDs of the national mortality of the US in 2006 included high blood pressure, high cholesterol, low fruit and vegetable consumption, smoking, physical inactivity, and overweight and obesity. Table 2 displays the AND estimates, the bootstrap standard errors, and 95% confidence intervals. Table 3 further displays these separately for males and females. The standard errors are very small relative to the estimates and thus not needed in the analysis of the national mortality data of the US for the leading ACDs.

Mortality of Chaoyang district of Beijing, China during 2007

The overall leading ACDs of Chaoyang district in 2007 included high blood pressure, smoking, alcohol use, low fruit and vegetable consumption, high cholesterol, physical inactivity, and overweight and obesity. Table 4 displays by age and sex the leading ACDs and the AND estimates, their standard errors and the 95% confidence intervals. It is noted that in most age groups, the leading risk factors remained the same between male and female. It is also noted that alcohol use was more severe in the young population and became less severe in the senior population. It is further noted that although unprotected sex was among the leading ACDs for female groups of age 15–29 and 30–44 years old, the estimates were not statistically significant at $P = 0.05$ level, and thus was not alarming. This observation is consistent with the report by Wang et al[13] in 2009 and illustrates the importance of the standard error and the confidence interval for the AND estimates.

Discussion

Many leading causes of death, such as lung cancer and cardiovascular diseases, are multifactorial, which makes it difficult to address all related issues of the same disease at once in a conventional disease-based study. Besides, intervention studies aiming at preventing or lowering one risk factor for a specific disease may also alter the risk factors of other diseases not under consideration. Therefore, an intervention study (eg, smoking cessation program) focusing on one primary disease (eg, lung cancer) may underestimate the actual overall benefit (including effect on lung cancer, cardiovascular diseases, etc)

Table 3 Leading attributable causes of death of the US mortality by sex during year 2006[a]

Male	Est (SE)	95% CI	Female	Est (SE)	95% CI
HBP	143032.6 (219.6)	(142646.8, 143344.3)	HBP	156277.9 (240.8)	(155863.0, 156710.4)
HCL	118770.0 (186.8)	(118420.2, 119123.5)	HCL	113677.0 (169.7)	(113356.6, 114012.1)
Smoking	114648.5 (235.5)	(114279.5, 115043.9)	ONO	104961.9 (186.9)	(104707.8, 105216.7)
ONO	88605.5 (154.3)	(88371.3, 88855.4)	LVC	76524.5 (104.4)	(76354.8, 76681.2)
LVC	88080.6 (122.7)	(87912.8, 88248.6)	PhInA	63550.8 (86.7)	(63423.1, 63676.0)
PhInA	56905.0 (82.9)	(56769.1, 57042.9)	Smoking	55086.5 (127.3)	(54867.0, 55299.3)
Alcohol use	47887.0 (85.4)	(47744.7, 48023.3)	Alcohol use	15594.3 (46.9)	(15514.0, 15676.8)
Unprotected sex	7791.3 (77.4)	(7634.9, 7957.5)	Unprotected sex	6970.1 (85.5)	(6794.4, 7129.5)
Urban air pollution	3571.8 (11.3)	(3550.3, 3593.2)	Urban air pollution	2774.3 (10.5)	(2752.2, 2792.8)
Illicit drug use	700.3 (7.0)	(686.3, 715.3)	Illicit drug use	302.2 (5.3)	(292.2, 3212.3)
CIIHS	470.7 (3.0)	(465.5, 477.6)	CIIHS	209.7 (1.8)	(206.2, 213.0)

Abbreviations: CI, confidence intervals; CIHS, contaminated injections in healthcare settings; HBP, high blood pressure; HCL, high cholesterol level; IndrAir, indoor air pollution from household use of solid fuels; LVC, low fruit and vegetable consumption; ONO, overweight and obesity; PhInA, physical inactivity; SE, standard error.

Table 4 Leading attributable causes of death: AND (SE) and 95% CI by age and gender in Chaoyang District of Beijing, China in 2007[a,b]

Age group	Male	Est (SE)	95% CI	Female	Est (SE)	95% CI
15–29	Alcohol use	3.0 (0.49)	(2.07, 3.95)	Alcohol use	1.5 (0.38)	(2.07, 3.95)
	HBP	1.8 (0.94)	(0.41, 3.82)	Smoking	1.5 (0.75)	(0.54, 3.01)
	LVC	1.2 (0.47)	(0.32, 2.02)	Unprotected sex	1.0 (1.04)	(0, 3.03)[b]
	Smoking	1.1 (0.47)	(0.44, 2.08)	HBP	0.9 (0.60)	(0, 2.32)[b]
30–44	HBP	31.2 (3.39)	(24.79, 37.55)	HBP	8.1(1.68)	(4.96, 11.77)
	Alcohol use	20.4 (1.73)	(17.37, 23.66)	Smoking	7.0 (1.22)	(5.07, 9.54)
	Smoking	14.2 (1.62)	(10.99, 17.41)	LVC	5.1 (0.77)	(3.62, 6.47)
	LVC	12.7 (1.27)	(10.37, 14.80)	Alcohol use	4.5 (0.70)	(3.16, 5.85)
	HCL	11.4 (1.53)	(8.64, 14.4)	ONO	3.2 (0.81)	(1.85, 4.93)
	CIHS	11.2 (1.44)	(8.60, 14.28)	PhInA	3.1 (0.61)	(2.01, 4.34)
	PhInA	7.7 (0.95)	(6.04, 9.37)	HCL	3.0 (0.81)	(1.52, 4.72)
	ONO	6.4 (0.82)	(4.80, 7.93)	Unprotected sex	3.0 (1.67)	(0, 6.00)[b]
45–59	HBP	151.9 (7.11)	(138.8, 166.0)	HBP	45.6 (4.16)	(38.56, 54.91)
	Smoking	82.9 (3.76)	(75.88, 89.93)	Smoking	29.0 (2.50)	(24.65, 34.04)
	LVC	71.7 (2.86)	(66.58, 77.34)	LVC	22.8 (1.69)	(19.94, 26.53)
	Alcohol use	65.0 (3.17)	(59.03, 71.68)	PhInA	19.6 (1.29)	(17.12, 21.98)
	HCL	57.6 (3.44)	(51.59, 63.60)	ONO	19.4 (1.65)	(16.10, 22.28)
	PhInA	40.5 (2.22)	(36.27, 45.09)	Alcohol use	18.0 (1.62)	(15.20, 21.33)
	ONO	36.7 (2.38)	(32.57, 41.54)	HCL	16.7 (1.98)	(12.80, 20.90)
	CIHS	36.5 (2.59)	(30.79, 42.04)	CIHS	8.8 (1.39)	(5.85, 11.22)
60–69	HBP	173.2 (6.55)	(159.5, 185.8)	HBP	112.7 (5.44)	(100.2, 121.3)
	Smoking	101.3 (4.25)	(93.51, 109.5)	Smoking	67.3 (3.77)	(60.65, 74.42)
	LVC	85.0 (2.89)	(79.56, 90.60)	LVC	57.7 (2.27)	(53.29, 61.99)
	HCL	69.8 (3.63)	(62.71, 76.02)	HCL	47.7 (2.77)	(42.00, 52.72)
	PhInA	50.9 (2.25)	(46.80, 55.46)	ONO	42.8 (2.63)	(37.66, 47.77)
	ONO	47.4 (2.96)	(41.93, 52.99)	PhInA	41.3 (1.78)	(37.97, 44.28)
	Alcohol use	41.9 (2.28)	(37.83, 46.68)	Alcohol use	24.7 (1.77)	(21.13, 28.11)
70–79	HBP	454.3 (11.93)	(431.3, 474.9)	HBP	337.5 (9.64)	(319.3, 355.2)
	Smoking	248.8 (6.20)	(237.6, 260.6)	LVC	153.3 (4.16)	(145.6, 161.1)
	LVC	213.0 (4.46)	(204.9, 221.3)	Smoking	148.8 (5.55)	(138.7, 159.8)
	HCL	179.8 (5.85)	(168.0, 191.5)	HCL	147.8 (5.58)	(138.6, 158.1)
	PhInA	125.3 (3.36)	(118.7, 131.5)	PhInA	108.4 (3.40)	(101.5, 114.7)
	ONO	110.0 (3.65)	(103.9, 117.7)	ONO	102.5 (3.90)	(95.05, 110.8)
	Alcohol use	86.7 (3.02)	(80.49, 92.45)	Alcohol use	54.3 (3.03)	(48.74, 60.80)
80+	HBP	411.8 (9.99)	(393.7, 431.1)	HBP	430.3 (9.97)	(410.2, 448.2)
	LVC	176.4 (4.20)	(167.7, 183.9)	HCL	185.1 (5.74)	(171.4, 193.7)
	HCL	171.4 (5.35)	(161.0, 181.0)	LVC	168.5 (4.45)	(158.6, 175.2)
	Smoking	165.1 (5.70)	(154.5, 175.7)	PhInA	122.4 (3.26)	(114.6, 127.7)
	PhInA	108.5 (3.22)	(102.5, 114.5)	Smoking	116.6 (4.40)	(108.1, 126.3)
	ONO	83.1 (2.90)	(78.48, 89.52)	ONO	103.9 (3.07)	(98.35, 109.7)
	IndrAir	56.4 (3.49)	(49.46, 62.05)	Alcohol use	52.0 (2.10)	(48.10, 56.49)
	Alcohol use	55.9 (2.43)	(51.30, 60.88)	IndrAir	45.7 (3.21)	(38.90, 51.88)

Notes: [a]PAF of pooled population, rather than of age-gender specific group, was used to compute the ANDs since group-specific PAFs are not available for some diseases (risk factors). Thus the ANDs may differ from the one reported in Wang et al 2009.[13] [b]Not statistically significant at $P = 0.05$ level.

Abbreviations: ANDs, attributable number of deaths; CI, confidence intervals; CIHS, contaminated injections in healthcare settings; HBP, high blood pressure; HCL, high cholesterol level; IndrAir, indoor air pollution from household use of solid fuels; LVC, low fruit and vegetable consumption; ONO, overweight and obesity; PhInA, physical inactivity; SE, standard error.

of the intervention program. It is thus of crucial importance to make public health policies based on study outcomes that consider multiple risk factors and multiple diseases simultaneously. The study of ACDs provides such a venue to study the nongenetic risk factors of death, where many risk factors are often found largely preventable. Despite the above advantages of the ACD studies, many quantitative methods remain to be developed, in particular, methods for systematic estimation of the standard errors and the confidence intervals of the ANDs.

In this study, we developed two methods for the variance estimation of the AND, the first assumes the mortality data follow a multinomial distribution and the second takes a nonparametric bootstrap approach. Simulation studies show that both lead to valid estimation of the variance of the AND. Both methods are easy to implement through a standard

statistical software package, including R, SAS, MATLAB, STATA, etc.

Using the bootstrap method, we studied the effect of the size of the mortality data on the AND estimates and the standard errors through simulations. We found that unlike other statistical estimates that usually achieve small variability with large samples, the AND standard error increases with the size of mortality. However, it is more interesting to note that the SE:AND ratio decreases with the size of mortality, which reflects the fact that the larger the mortality sample, the more accurate the estimation. This seemingly confusing outcome can be explained as follows. In general, the variability of estimates decreases with sample size while the estimates themselves converge to true values as sample size increases, ie, the estimates become stable around the true values. Therefore, the ratio of standard error to estimate effectively decreases with the sample size. Similarly in the AND estimation, the ratio decreases although both the standard error and the estimate increase with the size of the mortality.

We further illustrated with two mortality data sets of different sizes, one large data set of the national mortality of the US during 2006 and one moderate size set of mortality data in Chaoyang District of Beijing, China in 2007. We found that for the large national mortality data of the US, the SE:AND ratios are very small and thus standard errors are not needed, while for the moderate size Chaoyang District mortality data, standard errors are needed for the AND estimates. In particular, the point estimate of the AND for the unprotected sex was ranked high in the female groups of age 15–30 years and 30–44 years, had a large standard error and was not statistically significant. Therefore, the ranking of unprotected sex as a leading ACD may require further investigation. This demonstrates the needs of the variability assessment of the AND estimates in studying small-to-moderate size mortality data.

In comparing the results of the two mortality data sets, we found that although China and the US have very distinct culture and food consumption, and are located geographically far apart, the patterns of the leading ACDs are surprisingly similar. The top ACDs include high blood pressure, smoking, low fruit and vegetable consumption, alcohol use, high cholesterol level, overweight and obesity, and physical inactivity, among others. In particular, high blood pressure was ranked number one in both studies.

In summary, the multinomial distribution-based estimation method and the bootstrap method for the standard errors and confidence intervals for the AND estimates are useful tools in the study of attributable causes of death, and may play an important role in epidemiological studies of mortality and risk factors.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Ockene JK, Kuller LH, Svendsen KH, Meilahn E. The relationship of smoking cessation tocoronary heart disease and lung cancer in the multiple risk factor intervention trial (MRFIT). *Amer J Publ Heal*. 1990;80:954–958.
2. McGinnis JM, Foege WH. Actual causes of death in the United States. *J Amer Med Assoc*. 1993;270:2207–2212.
3. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. *J Amer Med Assoc*. 2004;291:1238–1245.
4. Danaei G, Ding EL, Mozaffarian D, et al. The preventable causes of death in the United States: comparative risk assessment of dietary, lifestyle, and metabolic risk factors. *PLoS Med*. 2009;6(4):e1000058.
5. Whittemore AS. Statistical methods for estimating attributable risk from retrospective data. *Statist Med*. 1982;1:229–243.
6. Greenland S. Variance estimators for attributable fraction estimates consistent in both large strata and sparse data. *Statist Med*. 1987;6: 701–708.
7. Benichou J, Gail MH. Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models. *Biometrics*. 1990;46:991–1003.
8. Kooperberg C, Petitti DB. Using logistic regression to estimate the adjusted attributable risk of low birthweight in an unmatched case-control study. *Epidemiology*. 1991;2:363–366.
9. Greenland S. Estimation of population attributable fractions from fitted incidence ratios and exposure survey data, with an application to electromagnetic fields and childhood leukemia. *Biometrics*. 2001;57: 182–188.
10. Liu K-J. Interval estimation of the attributable risk in case-control studies with matched pairs. *J Epid Comm Heal*. 2001;55:885–890.
11. Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJL. *Global burden of disease and risk factors*. Oxford, UK: Oxford University Press and The World Bank; 2006.
12. Heron MP, Hoyert DL, Murphy SL, Xu JQ, Kochanek, KD, Tejada-Vera B. *Deaths*: *final data for 2006*. *National vital statistics report*; *vol 57 no 14*. Hyattsville, MD: National Center for Health Statistics; 2009. Available from: http://www.cdc.gov/nchs/data/nvsr/nvsr57/nvsr57_14.pdf. Accessed July 7, 2011.
13. Wang Y, Zhang L, Huang J, Fu WJ, Li X, Meng H. Actual causes of death in Chaoyang District of Beijing, China, 2007. *Postgrad Med J*. 2009;87(1023):4–11.
14. Efron B, Tibshirani R. *An introduction to the bootstrap*. New York: Chapman and Hall/CRC; 1993.