AMERICAN JOURNAL OF INDUSTRIAL MEDICINE 49:683–689 (2006)

Use of Multiple Cause of Death Data in Cancer Mortality Analyses

David B. Richardson*

Background *In a cancer mortality study, the decision of whether to define a study outcome via underlying cause of death (UCD) or via multiple cause of death (MCD) information may impact relative risk (RR) estimates and associated confidence intervals.*
Methods *Simple equations are presented that relate RR estimates obtained in a cancer incidence study to the RR estimates obtained in mortality studies using UCD and MCD information. Data from the Surveillance, Epidemiology and End Results program were used to obtain information about the detection and confirmation rates of cancer diagnoses made via UCD. Data from US cause of death data tapes were used to obtain information on the ratio of UCD to MCD listings for cancer outcomes. Numerical examples illustrate the use of these equations.*
Results *In our examples, the RRs obtained via analyses of MCD were close to those obtained via analyses of UCD (but of greater precision), even when assuming that the confirmation rate of cancer diagnoses made via MCD listing was substantially lower than that of diagnoses made via UCD.*
Conclusions *These finding are supportive of the use of MCD information in cancer mortality studies.* Am. J. Ind. Med. 49:683–689, 2006. © *2006 Wiley-Liss, Inc.*

KEY WORDS: occupational mortality; multiple cause of death; death certificates; cancer; epidemiologic methods

INTRODUCTION

In cancer mortality studies, the outcomes under investigation are often defined in terms of the underlying cause of death (UCD) listed on the death certificate. Until 1968, cause of death information was routinely tabulated at the national level only for UCD; consequently, until recently an investigator necessarily had to define study outcomes in terms of UCD for US cohort studies that involved comparisons to national mortality rates (i.e., analyses of standardized mortality ratios).

However, in mortality studies that involve internal comparisons between groups with different exposures, the investigator has had the option of making use of all listed causes on the death certificate which may include contributory causes as well as other diseases not related to the underlying cause (listed as ''other conditions'' prevalent at death). The use of multiple cause of death (MCD) information may be particularly valuable for studies of chronic diseases, such as cancer, which tend to occur at older ages in patients with multiple morbid conditions prevalent at death [Israel et al., 1986]. Several previous authors have noted that entirely discounting available information on non-underlying causes of death may be a waste of information, and further argued that focus on a single underlying cause often provides an incomplete picture when there are multiple conditions present at death [Dorn and Moriyama, 1964; Markush and Seigel, 1968; Israel et al., 1986].

Department of Epidemiology, School of Public Health, University of North Carolina, Chapel Hill, North Carolina
Contract grant sponsor: National Institute for Occupational Safety and Health; Contract grant number: RO1 OH007871.
*Correspondence to: David B. Richardson, Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7435.
E-mail: david.richardson@unc.edu

Accepted 17 April 2006
DOI 10.1002/ajim.20343. Published online in Wiley InterScience (www.interscience.wiley.com)

In this article, the decision of whether to define cancer mortality outcomes in terms of UCD or MCD information is framed in terms of the impact on relative risk (RR) estimates and associated confidence intervals. I focus on the scenario of a cancer mortality study that contrasts incidence proportions between two subgroups of the study population. My working assumption is that the aim of this hypothetical cancer mortality study is to approximate the RR that would be obtained in a study of cancer incidence. Equations are developed that permit one to contrast the RRs obtained via analyses of UCD and MCD information to the RR that would be obtained via analyses of cancer incidence. These equations are formulated in terms of the sensitivity and positive predictive value of cancer diagnoses made via UCD and MCD information. These values are derived, where possible, from published data in which the ''gold standard'' was considered to be the ascertainment of cancer cases via a cancer registry [Percy et al., 1990]. Therefore, the numerical examples contrast the RRs derived via analyses of UCD and MCD information to the RR that would be obtained in an incidence study that ascertained cases via a cancer registry.

MATERIALS AND METHODS

Consider a hypothetical study comparing cancer risk in two groups, with cancer incidence ascertained in a closed cohort via a cancer registry with characteristics comparable to those of the Surveillance, Epidemiology, and End Results (SEER) program. Let's say that the cancer risk in the exposed subgroup is r_1, the cancer risk in the unexposed group is r_0, and the cancer risk overall is $r_\bullet$, where r_1, r_0, and $r_\bullet$ denote incidence proportions (see Appendix I).

Now consider the same population examined in the context of a mortality study of cancer outcomes in a closed cohort followed to extinction. If case ascertainment via the cancer registry is our standard, we can then refer to information about the sensitivity (sometimes referred to as detection rate) and confirmation rate (sometimes referred to as positive predictive value) for cancer classifications based upon UCD information. Table I reports sensitivity and confirmation rates for three cancer outcomes; these values, denoted Se_{UCD} and Cf_{UCD}, were originally reported by Percy et al. [1990] and were obtained by linkage of cancer registry records with UCD information.

Using these reported values for the sensitivity and confirmation rate for cancer outcomes we can derive Sp_{UCD} (see Appendix I), the specificity of cancer outcomes classified via UCD information, as

$$Sp_{UCD} = 1 - \frac{Se_{UCD}(r_\bullet)}{Cf_{UCD}(1-r_\bullet)} + \frac{Se_{UCD}(r_\bullet)}{(1-r_\bullet)}$$

Assuming that Se_{UCD} and Sp_{UCD} are equivalent in the exposure groups, an analysis in which cancer outcomes were defined via UCD would yield an estimate of cancer risk in the exposed subgroup,

$$r'_1 = Se_{UCD}(r_1) + (1 - Sp_{UCD})(1 - r_1)$$

an estimate of cancer risk among the unexposed,

$$r'_0 = Se_{UCD}(r_0) + (1 - Sp_{UCD})(1 - r_0)$$

and a risk ratio estimate of r'_1/r'_0.

TABLE I. Estimates of the Sensitivity, Se_{UCD}, and Confirmation Rate, Cf_{UCD}, for Classification of Three Cancer Outcomes Via Underlying Cause of Death Information Derived From the US Death Certificate

Cancer category (International Classification of Diseases, Ninth Revision)	Se_{UCD}[a]	Cf_{UCD}[a]	x[b]
Leukemia (204–208)	0.74	0.94	0.30
Pancreas (157)	0.92	0.90	0.07
Lung and bronchus (162.2–162.9)	0.95	0.94	0.09[c]

Also shown is the ratio of the number of cases listed as a non-underlying cause of death to the number of cases listed as the underlying cause.
[a]From Percy et al. [1990].
[b]From Steenland et al. [1992]. The ratio of cases listed anywhere on the death certificate to the number of cases listed as underlying cause minus one.
[c]Value reported by Steenland et al. [1992] is for lung, bronchus, and trachea.

Multiple Cause of Death Information

Use of MCD information classifies as diseased all of those people classified as diseased via UCD information plus adds other people to the diseased category. If UCD information leads to the identification of t people with a disease, then use of MCD information leads to identification of $(t + xt)$ cases, where x denotes the ratio of the number of deaths for which a cancer was listed anywhere on the death certificate to the number of deaths for which the cancer was listed as the UCD, minus one. In a cohort mortality study in which all cause of death information has been collected, x can be directly calculated. However, an estimate of x for a given cause of death can be obtained via analyses of US multiple cause of death data tapes (Table I).

The confirmation rate for the additional cancer cases listed as non-underlying causes may differ from the confirmation rate for cancer cases classified based upon UCD information. If the confirmation rate for the additional cancer cases listed as non-underlying causes is denoted Cf_{ac} then the sensitivity and specificity of outcome classification using MCD information can be expressed as,

$$Se_{MCD} = Se_{UCD} + xCf_{ac}\,\frac{Se_{UCD}r_\bullet + (1-Sp_{UCD})(1-r_\bullet)}{r_\bullet}$$

and

$$Sp_{MCD} = Sp_{UCD} - x(1 - Cf_{ac})\,\frac{Se_{UCD}r_\bullet + (1-Sp_{UCD})(1-r_\bullet)}{1-r_\bullet}$$

Assuming that Se_{MCD} and Sp_{MCD} are equivalent in the exposure groups, an analysis in which cancer outcomes were defined via MCD would yield an estimate of cancer risk in the exposed subgroup,

$$r''_1 = Se_{MCD}(r_1) + (1 - Sp_{MCD})(1 - r_1)$$

an estimate of cancer risk among the unexposed,

$$r''_0 = Se_{MCD}(r_0) + (1 - Sp_{MCD})(1 - r_0)$$

and a risk ratio estimate of r''_1/r''_0.

Numerical Examples

These relationships are illustrated for several scenarios. Numerical examples are presented for analyses of leukemia, pancreatic cancer, and lung cancer. I specify that r_0 is 0.010 for leukemia, 0.014 for pancreatic cancer, and 0.080 for lung cancer.

Results are presented for three values for RR (1.5, 2.0, 4.0), where $RR = r_1/r_0$. I applied the sensitivity and confirmation rate for outcome classification via UCD using the values in Table I; and, I specified the ratio of the number of deaths for which a cancer was listed as a non-underlying cause to the number of deaths for which the cancer was listed as the UCD using the values in Table I.

Results are shown for the scenario in which the confirmation rate for the cases ascertained via non-underlying cause of death information (Cf_{ac}) is equal to the confirmation rate for cases ascertained via UCD, Cf_{UCD}. Results are also shown for scenarios in which Cf_{ac} is lower than Cf_{UCD}; specifically, I specified that $Cf_{ac} = 0.90(Cf_{UCD})$ and $Cf_{ac} = 0.75(Cf_{UCD})$.

In order to illustrate the precision of risk estimates derived via analyses of UCD information relative to the precision of risk estimates derived via analyses of MCD information, I calculated 95% confidence intervals for the relative risk estimates for the scenario of a cohort study with 1,000 exposed and 1,000 unexposed subjects. The standard deviation for the log relative risk was estimated as $SD = \left(\frac{1}{1000r'_1} - \frac{1}{1000} + \frac{1}{1000r'_0} - \frac{1}{1000}\right)^{\frac{1}{2}}$ and Wald 95% confidence limits were obtained as $\exp(\log(r'_1/r'_0) \pm 1.96\ SD)$ [Greenland and Rothman, 1998].

RESULTS

Table II presents numerical examples for three hypothetical cohort studies comparing leukemia risk in two exposure groups. In the first example the risk of leukemia incidence among the exposed is specified to be 1.5 times the risk of leukemia in the unexposed; in the second and third examples the relative risk is 2.0 and 4.0, respectively. In each example, the sensitivity of leukemia classification via UCD information is 0.74 and the confirmation rate of leukemia diagnoses in the full cohort is 0.94. In each of the three examples the risk ratios obtained via analyses of UCD are biased towards the null when compared to the RR specified for leukemia incidence.

TABLE II. Relative Risks for Three Cancer Causes in Analyses Using UCD (RR_{UCD}) and MCD (RR_{MCD}) Information*

Outcome	RR	RR_{UCD}[a]	95% CI*	RR_{MCD}[b]	95% CI*
Leukemia	1.5	1.46	(0.60, 3.58)	1.46	(0.67, 3.20)
	2.0	1.91	(0.82, 4.45)	1.91	(0.91, 4.00)
	4.0	3.58	(1.69, 7.58)	3.58	(1.85, 6.90)
Pancreatic cancer	1.5	1.44	(0.74, 2.78)	1.44	(0.76, 2.72)
	2.0	1.85	(1.00, 3.45)	1.85	(1.02, 3.38)
	4.0	3.33	(1.93, 5.72)	3.33	(1.97, 5.62)
Lung cancer	1.5	1.46	(1.12, 1.90)	1.46	(1.13, 1.88)
	2.0	1.90	(1.48, 2.44)	1.90	(1.50, 2.41)
	4.0	3.49	(2.81, 4.34)	3.49	(2.84, 4.30)

Confirmation rate for cases listed as non-underlying cause of death (Cf_{ac}) is assumed to be the same as the confirmation rate for cases listed as underlying cause of death.
[a]Given values for Cf_{UCD} and Se_{UCD} listed in Table I.
[b]Given values for x listed in Table I and assuming $Cf_{ac} = Cf_{UCD}$.
*95% confidence interval calculated for the scenario in which there are 1,000 exposed cohort members and 1,000 unexposed cohort members.

Table II also reports the relative risk obtained via analyses of MCD, under the assumption that the confirmation rate for cases ascertained via non-underlying cause of death information was the same as for cases ascertained via UCD. The relative risks derived in analyses of cases identified via MCD information are identical to those obtained in analyses of cases identified via UCD; however, use of MCD information results in slightly tighter confidence intervals than obtained via analyses of UCD.

The results shown for numerical examples of analyses of pancreatic cancer and lung cancer are similar to those shown for leukemia (Table II). Risk ratios based upon analyses of UCD are biased towards the null when compared to the RR specified for incidence in each example. The relative risks derived via MCD information are the same as that obtained via UCD although use of MCD information results in slightly tighter confidence intervals than obtained via analyses of UCD.

In Tables III and IV, the confirmation rate for cases ascertained via non-underlying cause of death information was less than the confirmation rate for cases ascertained via UCD. Specifically, for the numerical examples in Table III the confirmation rate for cases ascertained via non-underlying cause of death information was 90% of the confirmation rate for cases ascertained via UCD; and, for the numerical examples in Table IV the confirmation rate for cases ascertained via non-underlying cause of death information was 75% of the confirmation rate for cases ascertained via

TABLE III. Relative Risks for Three Cancer Causes in Analyses Using UCD (RR_{UCD}) and MCD (RR_{MCD}) Information*

Outcome	RR	RR_{UCD}[a]	95% CI*	RR_{MCD}[b]	95% CI*
Leukemia	1.5	1.46	(0.60, 3.58)	1.45	(0.66, 3.17)
	2.0	1.91	(0.82, 4.45)	1.88	(0.90, 3.93)
	4.0	3.58	(1.69, 7.58)	3.44	(1.80, 6.59)
Pancreatic cancer	1.5	1.44	(0.74, 2.78)	1.43	(0.76, 2.71)
	2.0	1.85	(1.00, 3.45)	1.85	(1.01, 3.36)
	4.0	3.33	(1.93, 5.72)	3.29	(1.95, 5.55)
Lung cancer	1.5	1.46	(1.12, 1.90)	1.45	(1.13, 1.87)
	2.0	1.90	(1.48, 2.44)	1.89	(1.49, 2.40)
	4.0	3.49	(2.81, 4.34)	3.44	(2.80, 4.22)

Confirmation rate for cases listed as non-underlying cause of death (Cf_{ac}) is assumed to be 90% of the confirmation rate for cases listed as underlying cause of death.
[a]Given values for Cf_{UCD} and Se_{UCD} listed in Table I.
[b]Given values for x listed in Table I and assuming $Cf_{ac} = 0.90(Cf_{UCD})$.
*95% confidence interval calculated for the scenario in which there are 1,000 exposed cohort members and 1,000 unexposed cohort members.

UCD. The relative risks obtained in analyses of MCD are slightly more attenuated than the relative risk estimates obtained in analyses of cases ascertained via UCD. This occurs because the confirmation rate for the additional cancer cases ascertained via non-underlying cause of death information is lower than the confirmation rate for cases ascertained via UCD. However, even under these scenarios the relative risks obtained in analyses of MCD are similar in magnitude to the relative risks obtained in analyses of UCD (differing by less than 10%); and, use of MCD information results in tighter confidence intervals than those obtained via analyses of UCD.

TABLE IV. Relative Risks for Three Cancer Causes in Analyses Using UCD (RR_{UCD}) and MCD (RR_{MCD}) Information*

Outcome	RR	RR_{UCD}[a]	95% CI*	RR_{MCD}[b]	95% CI*
Leukemia	1.5	1.46	(0.60, 3.58)	1.43	(0.65, 3.13)
	2.0	1.91	(0.82, 4.45)	1.84	(0.88, 3.83)
	4.0	3.58	(1.69, 7.58)	3.25	(1.71, 6.17)
Pancreatic cancer	1.5	1.44	(0.74, 2.78)	1.43	(0.76, 2.70)
	2.0	1.85	(1.00, 3.45)	1.83	(1.01, 3.34)
	4.0	3.33	(1.93, 5.72)	3.24	(1.93, 5.46)
Lung cancer	1.5	1.46	(1.12, 1.90)	1.45	(1.12, 1.86)
	2.0	1.90	(1.48, 2.44)	1.87	(1.48, 2.37)
	4.0	3.49	(2.81, 4.34)	3.35	(2.73, 4.11)

Confirmation rate for cases listed as non-underlying cause of death (Cf_{ac}) is assumed to be 75% of the confirmation rate for cases listed as underlying cause of death.
[a]Given values for Cf_{UCD} and Se_{UCD} listed in Table I.
[b]Given values for x listed in Table I and assuming $Cf_{ac} = 0.75(Cf_{UCD})$.
*95% confidence interval calculated for the scenario in which there are 1,000 exposed cohort members and 1,000 unexposed cohort members.

DISCUSSION

The numerical examples presented here are generally supportive of the use of MCD data for cancer mortality analyses. The equations and numerical examples suggest that use of MCD information should result in RR estimates that are similar to those obtained via use of UCD information but of slightly greater statistical precision. Ideally, a decision about whether to define a study outcome via UCD or via MCD information would be informed by empirical data about the sensitivity and specificity of disease diagnoses using UCD and MCD information. Unfortunately, there is minimal information available in the published literature on the reliability of MCD information for classification of cancer outcomes. Therefore, numerical examples have been used to illustrate the likely impact on relative risk estimates of use of UCD and MCD information. It is shown that, even in scenarios in which the confirmation rate of the cases ascertained via non-underlying cause information is 75% of the confirmation rate for cases ascertained via UCD information, the relative risk estimate obtained in analyses of MCD typically is only modestly attenuated relative to that obtained in analyses of UCD. It should be noted that there is little empirical evidence to suggest that cancer diagnoses noted as non-underlying conditions are markedly less reliable than those noted as the UCD.

For simplicity, examples focused on the scenario of a closed cohort followed to extinction. Often, of course, in a cohort mortality study a large proportion of the cohort survives to the end of follow-up. Among those alive at the end of follow-up, the sensitivity of the death certificate for ascertaining cancer incidence is, by definition, 0; and, the specificity of the death certificate for ascertaining cancer incidence is, by definition, 1 (since false positive cases can only be ascertained among the members of the cohort who are deceased). Appendix II presents equations for calculating the sensitivity and specificity classifications of cancer cases for the scenario in which a portion of the cohort remains alive at the end of follow-up.

The equations and numerical examples presented here are premised on the assumption that outcome misclassification is non-differential with respect to exposure (i.e., sensitivity and specificity of case classification does not differ by exposure status). There are scenarios, of course, in which differential misclassification may occur. For example, deaths that occur at a hospital are more likely to have multiple causes listed than deaths that occur in other settings [Wall et al., 2005]; therefore, Se_{MCD} and Sp_{MCD} may vary with place of death. If place of death was related to exposure status then this would be a scenario in which differential outcome misclassification would be more likely to arise in analyses that define outcomes via MCD information than in analyses that defined the study outcome via UCD. Of course concerns about differential outcome

misclassification are not unique to analyses of MCD data; if place of death is related to exposure then differential case misclassification may occur in analyses based upon UCD as well, since UCD tends to be more accurately recorded for deaths that occur at a hospital (i.e., Se_{UCD} and Sp_{UCD} may also vary with place of death). There are also scenarios in which differential misclassification of case status may be *less* likely in analyses that utilize MCD information than in analyses that utilize UCD information. For example, often in cohort mortality studies cause of death information is coded to the ICD revision in effect at time of death (this is the case, for example, in studies that obtain cause of death information from the US National Death Index). The rules for selection of a single UCD from the listed causes of death have changed between revisions of the ICD. Therefore, for a given cause of death Se_{UCD} and Sp_{UCD} may vary over calendar time; if calendar year of death were related to exposure then this could be a source of differential misclassification in analyses based upon UCD information; in contrast, this is not an issue if study outcomes are defined in terms of MCD information.

This discussion has been framed in terms of analyses of mortality risk ratios. However, analyses of cohort data often involve estimation of rate ratios. Previous authors have noted that for non-differential outcome misclassification, the relationship between sensitivity/specificity and attenuation bias in relative risk and relative rate estimates is similar as long as one assumes that false positive cases do not result in much improper truncation of follow-up time for truly non-diseased subjects [Rothman and Greenland, 1998]. For analyses in which case classification is based upon death certificate information, this assumption is held since follow-up time is truncated for deceased subjects regardless of their case status (therefore, false positive cases never result in improper truncation of follow-up time). Consequently, by reference to sensitivity and positive predictive value of case classification one might describe the apparent rate ratio in a similar manner to that illustrated in this paper for analyses of risk ratios.

In these numerical examples, the use of MCD information typically resulted in a modest increase in precision of relative risk estimates over that achieved in analyses that used UCD information. In order to achieve a comparable increase in precision in a cohort study that utilized UCD information to that achieved via analyses of MCD information, however, an investigator would need to increase the study size in proportion to the ratio of cases listed as MCD to the number listed as UCD (Table I). For the example of leukemia mortality shown in the first row of results in Table II, the analysis using UCD information resulted in a $RR_{UCD} = 1.46$ (95% CI 0.60, 3.58) and was based upon a hypothetical cohort of 2,000 subjects. In order to obtain a RR_{UCD} estimate with confidence intervals as narrow as those obtained via analyses of MCD information (0.68, 3.20), an investigator would require a cohort study of approximately 2,600 subjects. The gain in efficiency obtained via use of MCD information will be greater for outcomes that have a high ratio of multiple cause listings to underlying cause listing [Steenland et al., 1992]. In general these include diseases that tend to have a long morbidity period and low case fatality rate, yet are serious enough to be noted by the certifier on the death certificate.

It is assumed that the aim of an investigator conducting a hypothetical cancer mortality study is to approximate the RR that would be obtained in a study of cancer incidence. Of course, a study that relies upon death certificate data assesses an association that is different from the association assessed in a cancer incidence study. Mortality studies are open to potential confounding by treatment quality, for example, and may underestimate the role of risk factors that influence the incidence of less severe or non-fatal disease. From this perspective, as well, the use of MCD information in cancer studies might be expected to minimize these limitations and increase the comparability of RR estimates obtained in cancer mortality and incidence studies. Analyses based upon MCD information will typically encompass some of the cases of disease that were prevalent at death but not selected as the UCD. Of course, an investigator has the option of considering results obtained via analyses of outcomes defined via MCD and via UCD information. Further empirical evaluation of the detection and confirmation rates for cancer classifications based upon MCD information would be a useful addition to the literature and provide further information about the appropriateness of using MCD information to define outcomes in cancer mortality studies.

ACKNOWLEDGMENTS

The author thank Steve Wing, Beverly Rockhill, and Kyle Steenland for their comments.

REFERENCES

Dorn HF, Moriyama IM. 1964. Uses and significance of multiple cause tabulations for mortality statistics. Am J Public Health Nations Health 54:400–406.

Greenland S, Rothman K. 1998. Introduction to categorical statistics. In: Rothman K, Greenland S, editors. Modern epidemiology, 2nd ed. Philadelphia: Lippincott, Williams, & Wilkins.

Israel RA, Rosenberg HM, Curtin LR. 1986. Analytical potential for multiple cause of death data. Am J Epidemiol 124:161–179.

Markush RE, Seigel DG. 1968. Prevalence at death. I. A new method for deriving death rates for specific diseases. Am J Public Health Nations Health 58:544–557.

Percy CL, Miller BA, Gloeckler Ries LA. 1990. Effect of changes in cancer classification and the accuracy of cancer death certificates on trends in cancer mortality. Ann NY Acad Sci 609:87–97.

Rothman KJ, Greenland S. 1998. Modern epidemiology. 2nd ed. Philadelphia, PA: Lippincott-Raven. xiii, 737 p.

Steenland K, Nowlin S, Ryan B, Adams S. 1992. Use of multiple-cause mortality data in epidemiologic analyses: Us rate and proportion files developed by the national institute for occupational safety and health and the national cancer institute. Am J Epidemiol 136:855–862.

Wall MM, Huang J, Oswald J, McCullen D. 2005. Factors associated with reporting multiple causes of death. BMC Med Res Methodol 5:4.

APPENDIX I

Consider a hypothetical study comparing cancer risk in two groups, where the case classification based upon cancer registry data conforms to Table AI. The risk in the exposed subgroup is denoted $r_1 = A/n_1$; the cancer risk in the unexposed group is denoted $r_0 = B/n_0$; and, the overall risk is denoted $r_\bullet = m_1/N$.

A mortality study that uses underlying cause of death (UCD) information to classify subjects with respect to cancer case status has sensitivity, Se_{UCD}, and specificity, Sp_{UCD} would result in the classification shown in Table AII.

The overall classification of subjects in the study with respect to disease status would conform to Table AIII.

TABLE AI. Distribution of Subjects by Disease Status and Exposure Status

	Exposed	Unexposed	
Diseased	A	B	m_1
Non-diseased	C	D	m_0
	n_1	n_0	N

The confirmation rate (i.e., positive predictive value) would be:

$$Cf_{UCD} = \frac{Se_{UCD}(A+B)}{Se_{UCD}(A+B)+(1-Sp_{UCD})(C+D)}$$

Dividing the numerator and denominator by *N*, the confirmation rate can be expressed in terms of the average risk in the study population,

$$Cf_{UCD} = \frac{Se_{UCD}(r_\bullet)}{Se_{UCD}(r_\bullet)+(1-Sp_{UCD})(1-r_\bullet)}$$

So, given Se_{UCD} and Cf_{UCD}, the specificity can then be expressed as,

$$Sp_{UCD} = 1 - \frac{Se_{UCD}(r_\bullet)}{Cf_{UCD}(1-r_\bullet)} + \frac{Se_{UCD}(r_\bullet)}{(1-r_\bullet)}$$

APPENDIX II

Consider a study of cancer mortality in a closed cohort in which a proportion, S, of the cohort is alive at the end of study. We wish to contrast the risk ratio for cancer incidence to the risk ratio obtained via analyses in which cancer cases were ascertained via UCD and via MCD.

Let's say that P is the ratio of deceased cases to incident cases ascertained over the follow-up period. Then, the sensitivity of case classification based upon UCD information will be Se_{UCD}*P. Similarly, the sensitivity of case classification based upon MCD information will be Se_{MCD}*P. If most incident cases are followed to extinction (e.g., the disease under study is rapidly fatal) then $P \sim 1$, Se_{UCD}*P $\sim Se_{UCD}$, and Se_{MCD}*P $\sim Se_{MCD}$.

The specificity of case classification based upon UCD information can be expressed as $Sp_{UCD,S} = [S+(1 - S - r_\bullet)\ Sp_{UCD}]/(1 - r_\bullet)$, where $r_\bullet$ is the average risk in the study population. When the disease risk is low, $Sp_{UCD,S}$ is well approximated by $Sp_{UCD,S} = S+(1 - S)\ Sp_{UCD}$. Similarly, the specificity of case classification based upon MCD

TABLE AII. Expected Distribution of Subjects By Exposure Status, Disease Status Defined by Cancer Registry Information, and Disease Status Defined by Underlying Cause of Death Information

		Exposed		Unexposed	
		Classification based upon registry information		Classification based upon registry information	
		Diseased	Non-diseased	Diseased	Non-diseased
Classification based upon UCD	Diseased	Se_{UCD} A	$(1 - Sp_{UCD})$C	Se_{UCD} B	$(1 - Sp_{UCD})$D
	Non-diseased	$(1 - Se_{UCD})$A	Sp_{UCD} C	$(1 - Se_{UCD})$B	Sp_{UCD} D

TABLE AIII. Expected Distribution of Subjects by Disease Status Defined by Cancer Registry Information and Disease Status Defined by Underlying Cause of Death Information

		Classification based upon registry information	
		Diseased	**Non-diseased**
Classification based upon UCD	Diseased	$Se_{UCD}(A + B)$	$(1 - Sp_{UCD})(C + D)$
	Non-diseased	$(1 - Se_{UCD})(A + B)$	$Sp_{UCD}(C + D)$

information can be expressed as $Sp_{MCD,S} = [S+(1 - S - r_{\bullet}) Sp_{MCD}]/(1 - r_{\bullet})$. Consequently, as the proportion of the cohort alive at the end of follow-up increases $Sp_{UCD,S}$ and $Sp_{MCD,S}$ approach unity.

Similar to the numerical examples shown in Tables II–IV, with the additional specification of values P and S, one can derive RR_{UCD} and RR_{MCD} for a study of mortality in a closed cohort that is not followed to extinction.