Date: February 15, 2005

Bobby R. Scott, Ph.D.
Lovelace Respiratory Research Institute
2425 Ridgecrest Drive SE
Albuquerque, NM, 87108 USA
bscott@LRRI.org

For: Notes on the Relative Hazard in the Context of the Cox Regression, Radiation Exposure, Cancer Occurrence, and the Linear Non-threshold Risk Model

Reference on Statistical Methods: Nicholas P. Jewell. *Statistics for Epidemiology*. Chapman & Hall, Boca Raton, London, New York, Washington, DC, 2004.

If the cumulative risk for cancer is small over the follow-up time interval [0,*T*] and the proportional hazard (hazard function at time *T* for exposed persons divided by the hazard function at time *T* for unexposed persons) is constant over the interval (see Figure 1), then the odds ratio (*OR*) at the end of the interval [i.e., *OR(T)*] is approximately equal to the constant relative hazard (*RH*) evaluated at *T* [i.e., *RH(T)*].

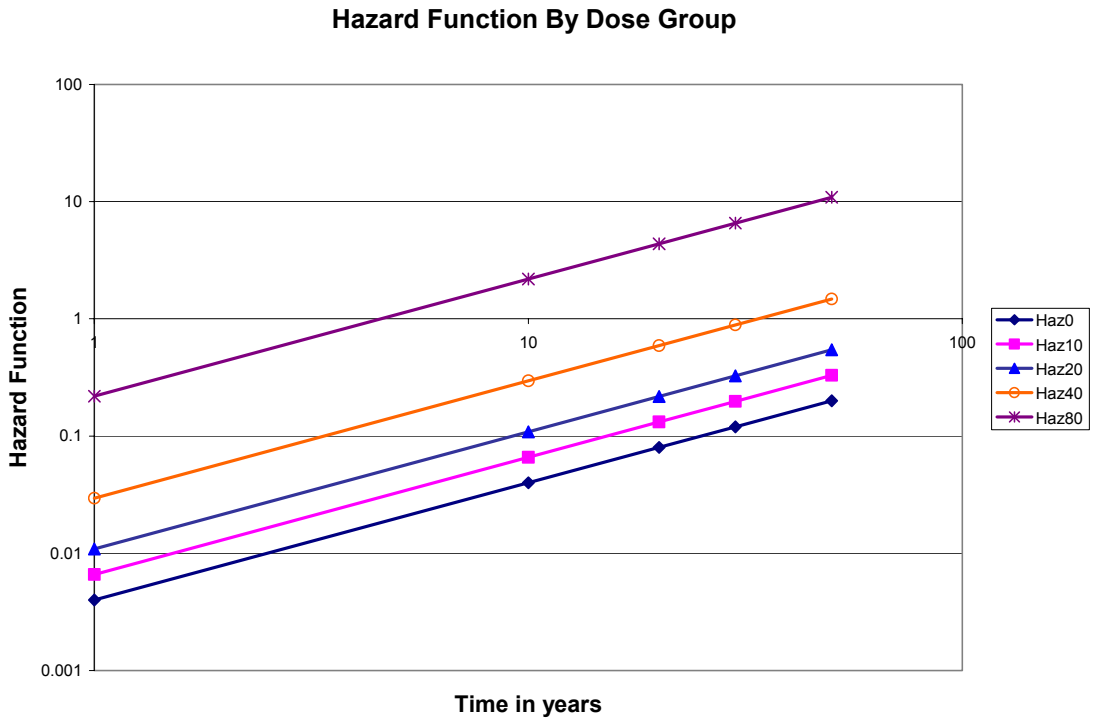

Figure 1: Hypothetical family of hazard functions, $h_x(t)$, for cancer induction for follow-up times from 1 to 50 years post exposure. Haz0, Haz10, etc. refer to doses of zero, 10, etc., in arbitrary units. Note that a logarithmic scale is used on both axes. Thus, the ratio of corresponding hazard function values on any two curves evaluated at the same time point is a constant (this is how the proportional hazard model works). The linear model in Equation 3 was used to generate these curves. The area under each curve evaluated from time zero gives the cumulative hazard (*CH*) whose log is also a linear function of dose x.

With logistic regression you focus on estimating *OR*.

With Cox regression you focus on estimating *RH* (also called hazard ratio).

With logistic regression you use the following model in the case of what is called linearity by some:

$$log[p_x/(1-p_x)] = a + bx, \qquad (1)$$

where x is dose; a and b are model parameters; p_x is the frequency (probability) of the specific disease of interest in the population of interest. The $log(OR)$ evaluated at dose x_1 relative to dose x_0 is just $b(x_1-x_0)$. Here log is evaluated relative to base e. Thus, the *OR* equals $exp[b(x_1-x_0)]$ and is therefore a highly nonlinear function of (x_1-x_0) for $b(x_1-x_0) >> 0$. When $x_0=0$, the *OR* is therefore a nonlinear function of x_1 under the condition of Equation 1. Note that with Equation 1, it is the log of the odds [i.e., *log* of $p_x/(1-p_x)$] that is being modeled as a linear, non-threshold function of dose x. It is especially important to recognize that this is not the same as modeling cancer risk p_x as a linear non-threshold function of dose. Thus, Equation 1 does not represent the linear non-threshold model used in cancer risk assessment, which when excess cancer risk is plotted vs. dose gives a straight line through the origin!

With Cox regression and when dealing with rare diseases you can also estimate *log*(*RH)* at dose x_1 relative to dose x_0 and use the results as an estimate of *log(OR)*, i.e.:

$$log[RH(T)] = c(x_1 - x_0) \approx log(OR). \qquad (2)$$

Note that the ratio of hazard functions at doses x_1 and x_0 are being used here. Note also that it is the *log[RH]* that is being modeled as being a linear, non-threshold function of dose increment (x_1-x_0). When $x_0=0$, then

$$log[RH(T)] = cx_1. \qquad (3)$$

Thus, with Cox regression, when one says there is a linear dose-response curve, this appears to refer to the *log[RH(T)]* being linear rather than to risk p_x being a linear function of dose!

This means that the Cox model when, based on Equation 2, cannot be used to estimate cancer risk per unit dose (which is the slope of the linear cancer risk, p_x, vs. dose relationship). While the above slope parameter c indeed has units of dose^{-1}, the slope c gives the increase in *log* of the *RH* when evaluated at *dose=1*. The relative hazard [an estimate of *OR* and therefore an estimate of relative risk (RR) for rare diseases] is given by $exp(cx_1)$. Thus, the estimate of excess relative risk per unit dose, which is given by $[exp(cx_1) -1]/x_1$ is not constant but changes as the dose x_1 changes since *RH* is a nonlinear function of dose x_1 under the circumstances being considered. This of course means that

an estimate of *RR* based on *RH* will also be a nonlinear function of dose under the circumstances being considered.

Because the cumulative hazard function *CH* (evaluated over the time interval of interest) is related to p_x by the equation

$$p_x = 1 - exp(-CH), \qquad (4)$$

p_x cannot be a linear, non-threshold function of x so long as either Equation 1 or 2 holds! Thus, "neither" logistic regression or Cox regression under conditions of Equations 1 or 2 can be used to obtain the slope of the linear risk (p_x) vs. dose response relationship. Similarly, neither logistic regression or Cox regression when based on either Equation 1 or 2 can be used to obtain excess relative risk per unit dose under the condition where p_x is a linear non-threshold function of dose (i.e., linear, non-threshold risk model).

For the proportional hazard assumption to be valid, the *RH* for the exposed vs. unexposed groups has to remain at least approximately constant at all times in the interval $[0,T]$ being considered (see Figure 1). Use of Cox regression is generally done when one has time to tumor occurrence (or time to cancer diagnosis) information which is used in applying the Cox model. Time-dependent dose (including discounted dose such as dose to 10 years before cancer diagnosis) can also be accommodated. When dummy time points are used, this would be expected to introduce systematic error. Actual time to diagnosis (or estimated time to tumor occurrence) from first exposure should be used.

Note that for the linear, non-threshold risk model where $p_x = a+bx$, then, Equation 1 should be replaced with:

$$log[p_x/(1-p_x)] = log[(a+bx)/(1-a- bx)]. \qquad (5)$$

log(OR) evaluated at dose x_1 relative to dose x_0 should then be given by:

$$log[OR] = log\,[(a+bx_1)/(1-a-bx_1)]-log[(a+bx_0)/(1-a-bx_0)]. \qquad (6)$$

When $x_0=0$, then Equation 6 reduces to:

$$log[OR] = log\,[(a+bx_1)/(1-a-bx_1)] - log\,[a/(1-a)]. \qquad (7)$$

Thus, relative risk is approximately given by:

$$RR \approx OR = exp\{log[(a+bx_1)/(1-a-bx_1)]-log[a/(1-a)]\}, \qquad (8)$$

which is a nonlinear function of x_1! Please note that the *OR* per unit dose (i.e., *OR* divided by dose) changes with dose and the excess *OR* per unit dose at $x_1=1$ is given by:

$$Excess\ OR\ per\ unit\ dose = exp\{log[(a+b)/(1-a-b)] - log\,[a/(1-a)]\} - 1. \qquad (9)$$

Under a true linear, non-threshold risk model (for p_x vs. dose), the exponential relationship in Equation 9 estimates the excess relative risk per unit dose, under the condition of logistic regression according to Equation 6.

From Equation 8, it can be shown that when $a+bx_1 << 1$, then the right-hand side of the equation is approximately given by $exp\{log[(a+bx_1)/a]\}$ which is just $exp(log[RR])$. It therefore follows that $log\{OR\} \approx log\{RR\}$. This explains the circumstances in which you can use *OR* as and estimate of *RR*. *OR* can be used as an estimate of *RR* only when the disease under study is rare. Note that the disease must remain rare even after a dose x_1. In the current example, this occurs when $a+bx_1 << 1$ (i.e., risk of disease occurrence is very small compared to a risk of 1).

Regarding Cox regression, a linear p_x vs dose curve implies a nonlinear curve for the cumulative hazard CH_x vs dose x, i.e.:

$$CH_x = -log(1 - p_x). \qquad (10)$$

This implies that the hazard rate function $h_x(t)$ (derivative of CH_x with respect to time t) used in the Cox model will be a nonlinear function of dose x of the form:

$$h_x(t) = -log(1 - f - gx), \qquad (11)$$

where f and g are fixed parameters and $h_x(t)$ has units of time^{-1}. Thus, the relative hazard, *RH*, under conditions of a linear non-threshold risk model for cancer induction is given by the following relationship when evaluated at dose x_1 relative to dose x_0:

$$RH = -log(1-f-gx_1)/[-log(1-f-gx_0)] = log(1-f-gx_1)/log(1-f-gx_0). \qquad (12)$$

For $x_0=0$ and rare diseases, Equation 12 reduces to:

$$RH = log(1-f-gx_1)/log(1-f) \approx OR. \qquad (13)$$

Thus, the *RH* (evaluated relative to the baseline hazard) is a non-linear function of dose x_1 in the case of the linear non-threshold risk (p_x) vs. dose model. Under conditions where the linear non-threshold risk model applies for cancer risk (p_x) vs. dose, excess relative risk per unit dose $[(p_x-p_0)/(xp_0)]$, abbreviated *ERRPD* and based on spontaneous risk, $p_0=a$, can be estimated based on Equation 13 (setting $x_1=1$), using:

$$ERRPD \approx [log(1-f-g)/log(1-f)]-1. \qquad (14)$$

Note that the dose $x_1 = 1$ in the denominator of Equation 14 has been omitted since division is by 1. Equation 14 has units of dose^{-1}.

The views expressed in these notes are those of the author and do not reflect in any way any formal position of the Lovelace Respiratory Research Institute or the U. S. DOE Office of Science (who supported this work) with respect to the application of the

statistical methods discussed. Notes presented were formulated as part of work carried out under DOE Grants DE-FG02-03ER63657 and DE-FG02-03ER63671.