MRC Technical Summary Report # 1899

TESTING HYPOTHESES FOR EFFECTS ON SURVIVAL BY THE ANALYSIS OF A MATCHED RETROSPECTIVE DESIGN

Bernard Harris and Anastasios A. Tsiatis

LEVEL

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

November 1978

Received September 8, 1978

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

National Cancer Institute
9000 Rockville Pike
Maryland 20014

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

TESTING HYPOTHESES FOR EFFECTS ON SURVIVAL BY THE ANALYSIS OF A MATCHED RETROSPECTIVE DESIGN

Bernard Harris[†] and Anastasios A. Tsiatis[‡]

Technical Summary Report #1899
November 1978

ABSTRACT

In testing for the relation of risk factors to a particular cause of death, such as a rare disease, a longitudinal study requires the observation of many individuals for long periods of time before enough information has accrued to permit reliable statistical analysis. In the present paper, this difficulty is circumvented through the use of a matched retrospective design. In particular, tests of the hypothesis of no effect are obtained for the constant proportionality model and for a second model in which the risk factors are quantified. The asymptotic distributions of the test statistics are also derived.

AMS(MOS) Subject Classification: 62F05, 62P10

Key Words: Matched retrospective design; hazard function; constant proportionality model.

Work Unit Number 4 - Probability, Statistics and Combinatorics

[†] Professor, Mathematics Research Center and Statistics Department, University of Wisconsin, Madison.

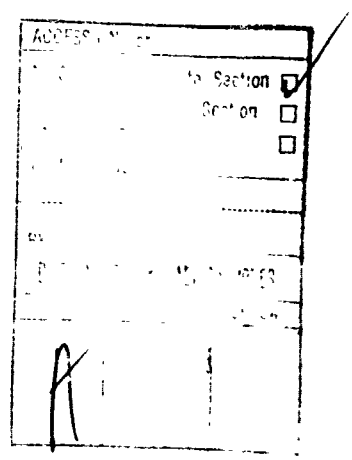
[‡] Assistant Professor, Departments of Statistics and Preventive Medicine, University of Wisconsin, Madison.

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024 and by the National Cancer Institute under Grant No. IROI CA 18332.

Significance and Explanation

This paper is concerned with testing hypotheses that certain presumed risk factors significantly affect survival. For example, it is desired to test whether certain specific risk factors affect the mortality rate from a particular disease, such as the relationship of exposure to polyurethane vapors and death due to leukemia. If the disease is rare, traditional methods of investigation involve the observation of many individuals over very long periods of time before enough mortalities have accrued to make statistical analysis feasible.

To reduce the time needed to acquire enough data for reliable statistical analysis a matched retrospective experimental design is suggested, in which each individual who died at time t from the disease under investigation is matched with an individual chosen at random from those alive at time t.



The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.

TESTING HYPOTHESES FOR EFFECTS ON SURVIVAL BY THE ANALYSIS OF A MATCHED RETROSPECTIVE DESIGN

Bernard Harris and Anastasios A. Tsiatis

1. INTRODUCTION AND SUMMARY

In this paper we construct tests of hypotheses for the existence of effects on survival due to the presence of risk factors; such as may be caused by unfavorable environmental situations. Such problems arise naturally in the comparison of the relationship between various environmental situations in employment and the possible effect that these may have on occupational health and safety.

Traditionally, longitudinal studies have been employed for this purpose. In such a study, risk factors are identified in advance and individuals exposed to these risk factors are observed for a predesignated length of time. Frequently, such studies have been utilized for the purpose of identifying risk factors as causes of death from a particular disease, such as the relationship of exposure to polyurethane vapors and death due to leukemia. However, if the disease under investigation is rare, then many individuals have to be observed for very long periods of time before enough mortalities have accrued to make statistical analysis feasible.

To circumvent this difficulty, a matched retrospective design is proposed. That is, each individual who died at age t from the disease under investigation is matched with an individual chosen at random from those alive at age t. We refer to the individual who died as the case and his matched counterpart as the control. For each such pair, we determine the risk factors to which they have been exposed. Let $\lambda_j(t)$, $j = 0,1,\ldots,k$, be the hazard function for the disease of interest for each individual exposed to risk factor j. Further let $\mu_j(t)$, $j = 0,1,\ldots,k$, be the hazard function for other causes of death for each individual exposed to risk factor j. We assume that for every pair i,j, $0 \leq i, j \leq k$, $\lambda_i(t)/\lambda_j(t) = \gamma_{ij} > 0$ a constant (independent of age). In the statistical literature, this is referred to as the constant proportionality hazards model (see Cox (1972)).

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024 and by the National Cancer Institute under Grant No. IROI CA 18332.

In section 2, we employ this model to obtain a test of the hypothesis $\gamma_{ij} = 1$, $0 \leq i, j \leq k$ and derive some of its large sample properties. Section 3 is devoted to the specific case in which the hazard functions satisfy a relationship of the form $\gamma_{ij} = \exp[\beta(v_i - v_j)]$ where the v_i, $0 \leq i \leq k$ are known constants. Such an assumption may be appropriate when the risk factors can be quantitatively measured, for example, when individuals have been exposed to specific levels of toxicity. This specific model has been proposed by Cox (1972).

2. TESTS OF HYPOTHESES FOR DIFFERENCES IN MORTALITY DUE TO RISK FACTORS IN A MATCHED RETROSPECTIVE EXPERIMENT

We divide a population Π into $k+1$ strata, $\pi_0, \ldots, \pi_k$. An element of the population will be placed in stratum π_j if it has been exposed to risk factor j. A particular cause of death, such as a specific disease, will be designated as the cause of death of interest. Data is to be collected as follows. If an individual dies from the cause of death of interest at age t, then a second individual alive at age t will be selected at random from the population and the stratum for each will be recorded. We denote the hazard function for the disease of interest by

$$\lambda_j(t) = \lambda(t)\exp\xi_j, \quad j = 0, \ldots, k, \tag{2.1}$$

and for the other causes of death by $\mu_j(t)$. With no loss of generality, we can set $\xi_0 = 0$. Using the above data, we will construct a test of the hypothesis $H_0: \xi_1 = \ldots = \xi_k = 0$, or equivalently, that the hazard rates do not depend on the risk factors.

Let T denote the age of death of any individual in the study. Assuming that the survival time, T_1, for the disease of interest and the survival time, T_2, for other causes of death are stochastically independent within each stratum, we get

$$P(T > t|\pi_j) = P(\min(T_1, T_2) > t|\pi_j) = \exp\{-\int_0^t [\lambda(x)e^{\xi_j} + \mu_j(x)]dx\}. \tag{2.2}$$

Let

$$v_j(t) = \lim_{h \to 0} P\{t < T_1 \leq t+h, T_2 > t|\pi_j\}/h . \tag{2.3}$$

It will be convenient to refer to $v_j(t)$ as the competitive hazard rate for the disease of interest. Since T_1, T_2 are stochastically independent within each stratum, it follows readily that

$$v_j(t) = \lambda(t)\exp\xi_j \exp\{-\int_0^t [\lambda(x)e^{\xi_j} + \mu_j(x)]dx\}. \tag{2.4}$$

Applying Bayes' Theorem, we get

$$P(\pi_j | T_1 = t, T_2 > t) = v_j(t)P(\pi_j)/\sum_{i=0}^{k} v_i(t)P(\pi_i) = p_{1j}(t), \tag{2.5}$$

and

$$P(\pi_j | T_1 > t, T_2 > t) = P(T > t|\pi_j)P(\pi_j)/\sum_{i=0}^{k} P(T > t|\pi_i)P(\pi_i) = p_{2j}(t). \tag{2.6}$$

Employing (2.2), (2.4), (2.5), (2.6) we obtain

$$\log(p_{1j}(t)/p_{10}(t)) - \log(p_{2j}(t)/p_{20}(t)) = \xi_j \ , \quad j = 1,\ldots,k, \tag{2.7}$$

independent of t.

Let $(Z_{i0\ell},\ldots,Z_{ik\ell})$, $Z_{ij\ell} \geq 0$, $j = 0,\ldots,k$, $\sum_{j=0}^{k} Z_{ij\ell} = 1$, $i = 1,2$; $\ell = 1,\ldots,n$, be independent multinomial random vectors with $P(Z_{ij\ell} = 1) = p_{ij\ell} > 0$ and $\sum_{j=0}^{k} p_{ij\ell} = 1$. Then

$$P(Z_{ij\ell} = z_{ij\ell},\ i = 1,2;\ j = 0,\ldots,k;\ \ell = 1,\ldots,n) = \prod_{i=1}^{2}\prod_{j=0}^{k}\prod_{\ell=1}^{n} p_{ij\ell}^{z_{ij\ell}}. \tag{2.8}$$

Now let

$$Z_{1j\ell} = \begin{cases} 1, & \text{if the } \ell\text{th case is in } \pi_j, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$Z_{2j\ell} = \begin{cases} 1, & \text{if } \ell\text{th control is in } \pi_j, \\ 0, & \text{otherwise,} \end{cases}$$

$\ell = 1,2,\ldots,n$.

Then if the random vectors $(Z_{i0\ell},\ldots,Z_{ik\ell})$ are conditionally independent given $T_1^{(1)} = t_1,\ldots,T_1^{(n)} = t_n$, where t_ℓ is the age at death of the ℓth case, the corresponding conditional likelihood is given by (2.8) upon setting $p_{ij}(t_\ell)$ equal to $p_{ij\ell}$.

We denote this conditional likelihood by

$L(\tilde{p};\tilde{Z},\tilde{t})$ where $\tilde{p} = (p_{ij\ell},\ i = 1,2;\ j = 1,\ldots,k;\ \ell=1,\ldots,n)$,

$\tilde{Z} = (Z_{ij\ell},\ i = 1,2;\ j = 1,\ldots,k;\ \ell = 1,\ldots,n)$ and $\tilde{t} = (t_1,\ldots,t_n)$.

Then from (2.8)

$$L(\tilde{p};\tilde{Z},\tilde{t}) = \left[\prod_{i=1}^{2}\prod_{\ell=1}^{n} p_{i0\ell}\right]\left[\prod_{i=1}^{2}\prod_{j=1}^{k}\prod_{\ell=1}^{n} (p_{ij\ell}/p_{i0\ell})^{Z_{ij\ell}}\right] =$$

$$= c(\tilde{p},\tilde{t}) \prod_{i=1}^{2}\prod_{j=1}^{k}\prod_{\ell=1}^{n} (p_{ij\ell}/p_{i0\ell})^{Z_{ij\ell}} \ . \tag{2.9}$$

From (2.7), it follows that

$$L(\tilde{p};\tilde{Z},\tilde{t}) = c(\tilde{p},\tilde{t})\exp\left\{\sum_{j=1}^{k}\xi_j\sum_{\ell=1}^{n}Z_{1j\ell} + \sum_{j=1}^{k}\sum_{\ell=1}^{n}(Z_{1j\ell}+Z_{2j\ell})\log(p_{2j\ell}/p_{20\ell})\right\}. \qquad (2.10)$$

From well known results on the properties of distributions in an exponential family (see E. L. Lehmann (1959)), the joint distribution of $\sum_{\ell=1}^{n} Z_{1j\ell}$ given $(Z_{1j\ell}, Z_{2j\ell})$, $j = 1,\ldots,k; \ell=1,\ldots,n$ is independent of $(p_{2j\ell}/p_{20\ell})$. This conditional distribution is given in the following theorem.

Theorem 1: Let $\{p_{\alpha\nu}\}$; $\alpha = 1,\ldots,n$; $\nu = 0,\ldots,k$ denote a family of multinomial distributions (i.e. with sample size unity), that is $p_{\alpha\nu} \geq 0$, $\sum_{\nu=0}^{k} p_{\alpha\nu} = 1$, for $\alpha = 1,\ldots,n$. Let $m_j(r)$ be the number of case-control pairs for which $W_{j\ell} = Z_{1j\ell} + Z_{2j\ell} = r$, $r = 0,1,2$ and let $m_{jj'}$ be the number of pairs for which $W_{j\ell} = 1$ and $W_{j'\ell} = 1$. Then the distribution of $\{(\sum_{\ell=1}^{n} Z_{10\ell},\ldots,\sum_{\ell=1}^{n} Z_{1k\ell}) \mid (W_{0\ell},\ldots,W_{k\ell}), \ell=1,\ldots,n\}$ is the distribution of the sum of n independent multinomial random vectors with $m_j(2)$ of them satisfying $p_{\alpha j} = 1$, $p_{\alpha j'} = 0$, $j \neq j'$, $j = 0,\ldots,k$ and $m_{jj'}$ of them satisfying $p_{\alpha j} = (1+\exp(\xi_j - \xi_{j'}))^{-1}$, $p_{\alpha j'} = (1+\exp(\xi_j - \xi_{j'}))^{-1}$, $p_{\alpha j''} = 0$, $j'' \neq j,j'$, $0 \leq j < j' \leq k$. Clearly

$$\sum_{j=0}^{k} m_j(2) + \sum_{0 \leq j < j' \leq k} m_{jj'} = n .$$

Proof: For fixed ℓ,

$$P\{Z_{10\ell} = z_{10\ell},\ldots,Z_{1k\ell} = z_{1k\ell} \mid W_{0m} = w_{0m},\ldots,W_{km} = w_{km},\ m = 1,\ldots,n\} =$$

$$P\{Z_{10\ell} = z_{10\ell},\ldots,Z_{1k\ell} = z_{ik\ell} \mid W_{0\ell} = w_{0\ell},\ldots,W_{k\ell} = w_{k\ell}\}$$ since the random vectors $(Z_{i0\ell},\ldots,z_{ik\ell})$, $\ell = 1,\ldots,n$; $i = 1,2$ are mutually independent.

All events of positive probability satisfy either

(a) $W_{j\ell} = 2$, $W_{j'\ell} = 0$, $j' \neq j$ or

(b) $W_{j\ell} = 1$, $W_{j'\ell} = 1$, $W_{j''\ell} = 0$, $j'' \neq j,j'$,

since by definition

$$\sum_{j=0}^{k} W_{j\ell} = 2.$$

We denote the events indicated by (a) by $E_{j\ell}(2)$ and the events indicated by (b) by $E_{jj'\ell}(1,1)$. Then

$$P\{Z_{10\ell}=z_{10\ell},\ldots,Z_{1k\ell}=z_{1k\ell}|E_{j\ell}(2)\}=\begin{cases}1, & \text{if } z_{1j\ell}=1,\ z_{1j'\ell}=0,\ j'\neq j\\ 0, & \text{otherwise,}\end{cases} \tag{2.11}$$

and

$$P(Z_{10\ell}=z_{10\ell},\ldots,Z_{1k\ell}=z_{1k\ell}|E_{jj'\ell}(1,1))=\begin{cases}\dfrac{p_{1j\ell}p_{2j'\ell}}{p_{1j\ell}p_{2j'\ell}+p_{1j'\ell}p_{2j\ell}}=(1+e^{\xi_{j'}-\xi_j})^{-1} & \text{if}\\ & z_{1j\ell}=1, z_{1h\ell}=0, h\neq j,\\ \dfrac{p_{1j'\ell}p_{2j\ell}}{p_{1j\ell}p_{2j'\ell}+p_{1j'\ell}p_{2j\ell}}=(1+e^{\xi_j-\xi_{j'}})^{-1} & \text{if}\\ & z_{1j'\ell}=1, z_{1h\ell}=0,\ h\neq j'\\ 0, \text{ otherwise.}\end{cases} \tag{2.12}$$

Corollary 1. Letting $T_j=\sum_{\ell=1}^{n} Z_{1j\ell}$, the conditional means, variances and covariances of T_j are given by;

$$\mu_j^{(n)}=E(T_j|W_{0\ell},\ldots,W_{k\ell};\ell=1,\ldots,n)=m_j^{(n)}(2)+\sum_{j'\neq j}m_{jj'}^{(n)}(1+e^{\xi_{j'}^{(n)}-\xi_j^{(n)}})^{-1},\ j=0,\ldots,k,$$

$$\sigma_{jj}^{(n)}=\sigma^2(T_j|W_{0\ell},\ldots,W_{k\ell};\ell=1,\ldots,n)=\sum_{j'\neq j}m_{jj'}^{(n)}(1+e^{\xi_{j'}^{(n)}-\xi_j^{(n)}})^{-1}(1+e^{\xi_j^{(n)}-\xi_{j'}^{(n)}})^{-1}, \tag{2.13}$$
$$j=0,\ldots,k,$$

$$\sigma_{jj'}^{(n)}=\mathrm{Cov}(T_j,T_{j'}|W_{0\ell},\ldots,W_{k\ell};\ell=1,\ldots,n)=-m_{jj'}^{(n)}(1+e^{\xi_{j'}^{(n)}-\xi_j^{(n)}})^{-1}(1+e^{\xi_j^{(n)}-\xi_{j'}^{(n)}})^{-1},$$

$$0\leq j<j'\leq k.$$

Theorem 2: If $m_{0j}^{(n)}(1+e^{\xi_{j'}^{(n)}-\xi_j^{(n)}})^{-1}(1+e^{\xi_j^{(n)}-\xi_{j'}^{(n)}})^{-1}\to\infty$ as $n\to\infty$ for all $j=1,\ldots,k$, then the distribution of

$$\sum_{j=1}^{k}a_j(T_j-\mu_j^{(n)})/\Big(\sum_{j=1}^{k}\sum_{j'=1}^{k}a_ja_{j'}\sigma_{jj'}^{(n)}\Big)^{\frac{1}{2}} \text{ given } (W_{0\ell}=w_{0\ell},\ldots W_{k\ell}=w_{k\ell},\ell=1,\ldots,n)$$

is asymptotically normal with mean zero and unit variance, whenever $a_1,\ldots,a_k$ are not all zero.

Proof: Defining $a_0=0$, we can write

$$\sum_{j=1}^{k}a_j(T_j-\mu_j^{(n)})=\sum_{\ell=1}^{n}\sum_{j=0}^{k}a_j(Z_{1j\ell}-p_{\ell j}^{(n)}),$$

where

$$p_{\ell j}^{(n)} = \begin{cases} (1+e^{\xi_{j'}^{(n)}-\xi_j^{(n)}})^{-1} & \text{if} \quad w_{j\ell}=1,\ w_{j',\ell}=1, \text{ for } j'\neq j, \\ 1 & w_{j\ell}=2, \\ 0 & \text{otherwise.} \end{cases}$$

Let $X_\ell = \sum_{j=0}^{k} a_j (Z_{1j\ell} - p_{\ell j}^{(n)})$, $\ell = 1,2,\ldots,n$. Then the conclusion will follow from Liapunov's theorem, upon establishing that for some $\delta > 0$,

$$\frac{\sum_{\ell=1}^{n} E(|X_\ell|^{2+\delta} | W_{0\ell},\ldots,W_{k\ell}, \ell=1,\ldots,n)}{\left[\sum_{\ell=1}^{n} E(X_\ell^2 | W_{0\ell},\ldots,W_{k\ell}, \ell=1,\ldots,n)\right]^{1+\delta/2}} \to 0 \tag{2.14}$$

as $n \to \infty$.

Since $Z_{1j\ell} - p_{\ell j}^{(n)} = 0$ with probability one given $W_{j\ell}=2$ for some j and $W_{j',\ell}=0$, $j'\neq j$, $0 \leq j, j' \leq k$. X_ℓ is non zero if and only if $W_{j\ell}=1$ and $W_{j',\ell}=1$, $j\neq j'$ and $W_{j''\ell}=0$, $j''\neq j,j'$. In this case, for $\delta \geq 0$

$$E\{|X_\ell|^{2+\delta} | W_{j\ell}=1, W_{j',\ell}=1, W_{j''\ell}=0, j''\neq j,j'\} =$$

$$|a_j - a_{j'}|^{2+\delta} \omega_{jj'}^{(n)}(1-\omega_{jj'}^{(n)})((\omega_{jj'}^{(n)})^{1+\delta} + (1-\omega_{jj'}^{(n)})^{1+\delta}),$$

where $\omega_{jj'}^{(n)} = (1+e^{\xi_{j'}^{(n)}-\xi_j^{(n)}})^{-1}$. Consequently, we can write (2.14) as

$$\frac{\sum_{0 \leq j < j' \leq k} m_{jj'}^{(n)} |a_j - a_{j'}|^{2+\delta} \omega_{jj'}^{(n)}(1-\omega_{jj'}^{(n)})((\omega_{jj'}^{(n)})^{1+\delta} + (1-\omega_{jj'}^{(n)})^{1+\delta})}{\left[\sum_{0 \leq j < j' \leq k} m_{jj'}^{(n)} |a_j - a_{j'}|^2 \omega_{jj'}^{(n)}(1-\omega_{jj'}^{(n)})\right]^{1+\delta/2}}$$

$$\leq \max_{j,j'} |a_j - a_{j'}|^{\delta} \Big/ \left[\sum_{0 \leq j < j' \leq k} m_{jj'}^{(n)} |a_j - a_{j'}|^2 \omega_{jj'}^{(n)}(1-\omega_{jj'}^{(n)})\right]^{\delta/2},$$

which tends to zero as $n \to \infty$.

<u>Corollary 2:</u> If $m_{jj'}^{(n)}/n \overset{a.s.}{\to} c_{jj'}$, $0 \leq j < j' \leq k$, $c_{0j} > 0$ for all $j = 1,\ldots,k$, and $\xi_j^{(n)} \to \xi_j$ as $n \to \infty$, $j=1,\ldots,k$, then the distribution of the random vector

$$n^{-\frac{1}{2}} \{(T_1 - \mu_1^{(n)}),\ldots,(T_k - \mu_k^{(n)})\}, \text{ given } (W_{0\ell},\ldots,W_{k\ell}; \ell=1,\ldots,n),$$

is almost surely asymptotically normal with mean zero and covariance matrix

$\Sigma = (\sigma_{jj'}; j,j'=1,\ldots,k)$, where

$$\sigma_{jj'} = \begin{cases} \sum_{j''\neq j} c_{jj''}\omega_{jj''}(1-\omega_{jj''}) & \text{if } j=j', \\ -c_{jj'}\omega_{jj'}(1-\omega_{jj'}) & \text{if } j\neq j', \end{cases}$$

and

$$\omega_{jj'} = (1+e^{\xi_{j'}-\xi_j})^{-1}.$$

Proof: As a consequence of the multivariate central limit theorem (See Rao (1973)), it suffices to show that for all $\underset{\sim}{a} = (a_1,\ldots,a_k) \neq \underset{\sim}{0}$, the distribution of

$$\sum_{j=1}^{k} a_j(T_j-\mu_j^{(n)}) \Big/ \Big[n\sum_{j,j'=1}^{k} a_j a_{j'}\sigma_{jj'}\Big]^{\frac{1}{2}} \quad \text{given } (W_{0\ell},\ldots,W_{k\ell};\ell=1,\ldots,n)$$

is asymptotically $N(0,1)$.

We can write

$$\frac{\sum_{j=1}^{k} a_j(T_j-\mu_j^{(n)})}{\left[n\sum_{j,j'=1}^{k} a_j a_{j'}\sigma_{jj'}\right]^{1/2}} = \frac{\sum_{j=1}^{k} a_j(T_j-\mu_j^{(n)})}{\left[\sum_{j,j'=1}^{k} a_j a_{j'}\sigma_{jj'}^{(n)}\right]^{1/2}} \cdot \left\{\frac{\sum\sum a_j a_{j'}\sigma_{jj'}^{(n)}/n}{\sum\sum a_j a_{j'}\sigma_{jj'}}\right\}^{1/2}.$$

By assumption $\sigma_{jj'}^{(n)}/n \xrightarrow{a.s.} \sigma_{jj'}$, therefore

$$\frac{\sum\sum a_j a_{j'}\sigma_{jj'}^{(n)}/n}{\sum\sum a_j a_{j'}\sigma_{jj'}} \xrightarrow{a.s.} 1, \quad \text{and the proof follows from Theorem 2.}$$

Corollary 3: Let $\mu_{j0}^{(n)} = E(T_j \mid W_{0\ell},\ldots,W_{k\ell},\ell=1,\ldots,n;\xi_j=0,\ j=1,\ldots,k)$. Then

$$\mu_{j0}^{(n)} = m_j^{(n)}(2) + \sum_{j'\neq j} m_{jj'}^{(n)}/2.$$

If $m_{jj'}^{(n)}/n \xrightarrow{a.s.} c_{jj'}$, $0 \leq j \leq j' \leq k$, $c_{0j} > 0$, $j=1,\ldots,k$, and $\xi_j^{(n)}\sqrt{n} \to \tau_j$ as $n \to \infty$, then the distribution of the random vector

$$n^{-\frac{1}{2}}\{(T_1-\mu_{10}^{(n)}),\ldots,(T_k-\mu_{k0}^{(n)})\} \quad \text{given } (W_{0\ell},\ldots,W_{k\ell};\ell=1,\ldots,n)$$

is almost surely asymptotically normal with mean $X\underset{\sim}{\tau}$, where $\underset{\sim}{\tau}' = (\tau_1,\ldots,\tau_k)$ and covariance matrix $X = (x_{jj'})$, where

$$x_{jj'} = \begin{cases} \sum_{j''\neq j} c_{jj''}/4, & j = j' \\ -c_{jj'}/4, & j \neq j', \end{cases}$$

Proof: We can write $n^{-\frac{1}{2}}(T_j-\mu_{j0}^{(n)}) = Q_j^{(n)} + R_j^{(n)}$, $j = 1,\ldots,k$ where $Q_j^{(n)} = n^{-\frac{1}{2}}(T_j-\mu_j^{(n)})$ and $R_j^{(n)} = n^{-\frac{1}{2}}(\mu_j^{(n)}-\mu_{j0}^{(n)})$. By Corollary 2, the random vector $(Q_j^{(n)}, j=1,\ldots,k)$ given $(W_{0\ell},\ldots,W_{k\ell}; \ell=1,\ldots,n)$ is asymptotically $N(0,X)$. Since

$$R_j^{(n)} = \sum_{j'\neq j} \left(\frac{m_{jj'}^{(n)}}{n}\right) n^{\frac{1}{2}} \left[\left(1+e^{\xi_{j'}^{(n)}-\xi_j^{(n)}}\right)^{-1} - \frac{1}{2}\right]$$

and $n^{\frac{1}{2}}\left[\left(1+e^{\xi_{j'}^{(n)}-\xi_j^{(n)}}\right)^{-1}-\frac{1}{2}\right] \to (\tau_j-\tau_{j'})/4$ as $n \to \infty$, then

$$R_j^{(n)} \xrightarrow{a.s.} \sum_{j'\neq j} c_{jj'}(\tau_j-\tau_{j'})/4 = \tau_j \sum_{j'\neq j} c_{jj'}/4 - \sum_{j'\neq j} \tau_{j'}c_{jj'}/4 .$$

Hence, the vector $(R_j^{(n)}, j=1,\ldots,k) \xrightarrow{a.s.} X\underset{\sim}{\tau}$ and the conclusion follows.

Remark 1: The proposed design matches an individual, who died of the disease under investigation at age t with a random individual alive at age t. Let the death times t_ℓ, $\ell=1,\ldots,n$ be independently distributed with density function $f(t)$ and let $c_{jj'}$ be the unconditional probability that the case is in stratum π_j and the control is in stratum $\pi_{j'}$. Then from (2.5) and (2.6) we get

$$c_{jj'} = \int_0^\infty [p_{1j}(t)p_{2j'}(t) + p_{1j'}(t)p_{2j}(t)]f(t)dt .$$

The marginal distribution of $m_{jj'}^{(n)}$ is then the multinomial distribution with sample size n and cell probabilities $c_{jj'}$, $0 \leq j < j' \leq k$, which is the notation for the hypotheses of Corollaries 2 and 3.

By Corollary 3, $Y = n^{-\frac{1}{2}}\{(T_1-\mu_{10}^{(n)}),\ldots,(T_k-\mu_{k0}^{(n)})\}$ given $W_{0\ell},\ldots,W_{k\ell}; \ell=1,\ldots,n$ is asymptotically $N(X\underset{\sim}{\tau}, \Sigma) = N(X\underset{\sim}{\tau}, X)$. From the theory of general linear models an efficient test statistic for testing $\underset{\sim}{\tau} = \underset{\sim}{0}$ is given by

$$\hat{\underset{\sim}{\tau}}'(\Sigma(\hat{\tau}))^{-1}\hat{\underset{\sim}{\tau}} \tag{2.15}$$

where $\hat{\underset{\sim}{\tau}}$, the weighted least squares estimate, is

$$\hat{\tau} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y = X^{-1}Y$$

and

$$\Sigma(\hat{\tau}) = (X'\Sigma^{-1}X)^{-1} = X^{-1} .$$

Consequently, the statistic (2.15) reduces to $Y'X^{-1}Y$. Under local alterantives this is distributed by the non-central χ^2 distribution with k degrees of freedom and non-centrality parameter $\underset{\sim}{\tau}'X\underset{\sim}{\tau}$. In practice, X has to be replaced by its consistent

estimate $\hat{X}$, where

$$\hat{X} = (\hat{x}_{jj'}; j,j'=1,\ldots,k),\ \hat{x}_{jj'} = n^{-1} \sum_{j''\neq j} m_{jj''}^{(n)}/4, \quad \text{if} \quad j=j' ,$$

$$= -n^{-1} m_{jj'}^{(n)}/4, \quad \text{if} \quad j\neq j'.$$

Therefore, the test of size α rejects H_0: $\xi_j=0, j=1,\ldots,k$, whenever

$$\underset{\sim}{T}^{0\prime} L^{-1} \underset{\sim}{T}^{0} > \chi^2_{\alpha;k},$$

where

$$\underset{\sim}{T}^{0} = (T_j^0, j=1,\ldots,k),\ T_j^0 = \sum_{\ell=1} Z_{1j\ell} - m_j^{(n)}(2) - \sum_{j'\neq j} m_{jj'}^{(n)}/2 ,$$

$$L = (\ell_{jj'}; j,j'=1,\ldots,k),\ \ell_{jj'} = \begin{cases} \sum_{j''\neq j} m_{jj''}^{(n)}/4 & \text{if } j=j' , \\ -m_{jj'}^{(n)}/4 & \text{if } j\neq j' , \end{cases}$$

and $\chi^2_{\alpha,k}$ is the $(1-\alpha)$th percentile of the chi-square distribution with k degrees of freedom.

Remark 2: For the case of two risk categories the problem has been studied by Miettinen (1968) who obtained a test previously given by McNemar (1947).

We also note that the test derived above is identical to the test for homogeneity of marginal distributions in a two way classification given by Stuart (1955).

3. QUANTITATIVELY ORDERED CATEGORIES

In some applications it may be possible to associate a quantitative measure to each stratum. For example, these measures may be the amounts of exposure to an environmental agent under investigation. Let $v_0, v_1, \ldots, v_k$ be the values assigned to each of the strata. Assume that the hazard functions for the disease of interest and for the other causes of death for individuals in stratum j are $\lambda(t)\exp(\beta v_j)$ and $\mu_j(t)$ respectively. This model has been proposed by D. R. Cox (1972).

The null hypothesis is $\beta=0$ and suggests no association between the strata and death due to the disease of interest. With these assumptions, analogously to (2.5) and (2.6), we get

$$\log(p_{1j}(t)/p_{10}(t)) - \log(p_{2j}(t)/p_{20}(t)) = \beta(v_j - v_0),$$

independent of t. Hence, analogous to (2.10), we have

$$L(\tilde{p};\tilde{z},\tilde{t}) = c(\tilde{p},\tilde{t})\exp\{\beta \sum_{j=1}^{k} (v_j - v_0) \sum_{\ell=1}^{n} z_{1j\ell} +$$

$$\sum_{j=1}^{k} \sum_{\ell=1}^{n} (z_{1j\ell} + z_{2j\ell}) \log(p_{2j\ell}/p_{20\ell})\}.$$

Using well known results on distributions in an exponential family, a UMP unbiased test for $H: \beta = 0$ vs $K: \beta > 0$, rejects H_0 for large values of $\sum_{j=1}^{k} (v_j - v_0) \sum_{\ell=1}^{n} z_{ij\ell}$, conditional on $((z_{1j\ell} + z_{2j\ell}), j=1,\ldots,k; \ell=1,\ldots,n)$. To obtain a large sample approximation to the distribution of $\sum_{j=1}^{k} (v_j - v_0) T_j$ given $(w_{j\ell}, j=0,\ldots,k; \ell=1,\ldots,n)$, we proceed as in Section 2, noting that under H_0

$n^{-\frac{1}{2}} \{(T_1 - \mu_{10}^{(n)}),\ldots,(T_k - \mu_{k0}^{(n)})\}$ given $(W_{j\ell}; j=0,\ldots,k; \ell=1,\ldots,n)$ is asymptotically normally distributed with mean 0 and covariance matrix X. Consequently,

$$n^{-\frac{1}{2}} \{\sum_{j=1}^{k} v_j' T_j - \sum_{j=1}^{k} v_j' \mu_{j0}^{(n)}\} \text{ given } W_{j\ell}; j=0,\ldots,k; \ell=1,\ldots,n$$

is asymptotically normal with mean 0 and variance $\chi' X \chi$, where $\chi' = (v_1 - v_0,\ldots,v_k - v_0)$.

We can estimate the variance by $\chi' \hat{X} \chi$ or

$$n^{-1} \sum_{j=1}^{k} \sum_{j' \neq 1}^{k} (v_j - v_0)(v_{j'} - v_0)\hat{x}_{jj'} =$$

$$= \frac{n^{-1}}{4} [-\sum_{j=0}^{k} \sum_{j' \neq j} (v_j - v_0)(v_{j'} - v_0) m_{jj'} + \sum_{j=0}^{k} (v_j - v_0)^2 \sum_{j' \neq j} m_{ij'}]$$

$$= \frac{n^{-1}}{4} [-\sum_{j=0}^{k-1} \sum_{j'=j+1}^{k} 2(v_j - v_0)(v_{j'} - v_0) m_{jj'} + \sum_{j=0}^{k-1} \sum_{j'=j+1}^{k} \{(v_j - v_0)^2 + (v_{j'} - v_0)^2\} m_{jj'}$$

$$= \frac{n^{-1}}{4} [\sum_{j=0}^{k} \sum_{j'=j+1}^{k} (v_j - v_{j'})^2 m_{jj'}] .$$

Under H_0

$$\frac{2 \sum_{j=1}^{k} (v_j - v_0)(T_j - \mu_{j0}^{(n)})}{[\sum_{j=0}^{k-1} \sum_{j'=j+1}^{k} (v_j - v_{j'})^2 m_{jj'}]^{1/2}} , \tag{3.1}$$

given $(w_{j\ell};j=0,\ldots,k;\ell=1,\ldots,n)$ is asymptotically distributed as a standard normal. Therefore the UMP unbiased level α test for $H_0:\beta=0$ vs $K:\beta>0$ consists of rejecting H_0 when the test statistic (3.1) is greater than z_α, where z_α is the $(1-\alpha)$th percentile of the standard normal.

REFERENCES

Cox, David R. (1972), "Regression Models and Life Tables (with Discussion)," Journal of the Royal Statistical Society, Ser. B, 34, 187-220.

Lehmann, E. L. (1959), Testing Statistical Hypotheses, New York: John Wiley & Sons.

McNemar, Q. (1947), "Note on the Sampling Error of the Differences Between Correlated Proportions or Percentages," Psychometrika, 12, 153-157.

Miettinen, O. S. (1968), "The Matched Pairs Design in the Case of All or None Responses," Biometrics, 24, 339-352

Rao, C. R. (1973), Linear Statistical Inferences and Its Applications, Second Edition, New York: John Wiley & Sons.

Stuart, A. (1955), "A Test for Homogeneity of the Marginal Distributions in a Two-Way Classification," Biometrika, 42, 412-416.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

MRC-TSR- REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 1899	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): TESTING HYPOTHESES FOR EFFECTS ON SURVIVAL BY THE ANALYSIS OF A MATCHED RETROSPECTIVE DESIGN,	5. TYPE OF REPORT & PERIOD COVERED: Summary Report - no specific reporting period
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Bernard Harris and Anastasios A. Tsiatis	8. CONTRACT OR GRANT NUMBER(s): DAAG29-75-C-0024, PHS - 1R01-CA-18332
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 4- Probability, Statistics and Combinatorics
11. CONTROLLING OFFICE NAME AND ADDRESS: See Item 18 below	12. REPORT DATE: November 1978
	13. NUMBER OF PAGES: 11
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U. S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

National Cancer Institute
9000 Rockville Pike
Maryland 20014

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Matched retrospective design;
hazard function
constant proportionality model

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In testing for the relation of risk factors to a particular cause of death, such as a rare disease, a longitudinal study requires the observation of many individuals for long periods of time before enough information is accrued to permit reliable statistical analysis. In the present paper, this difficulty is circumvented through the use of a matched retrospective design. In particular, tests of the hypothesis of no effect are obtained for the constant proportionality model and for a second model in which the risk factors are quantified. The asymptotic distribution of the test statistics are also derived.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)