AD_____

Award Number: DAMD17-99-1-9390

TITLE: Statistical Analysis of Multivariate Interval Censored Data in Breast Cancer Follow-Up Studies

PRINCIPAL INVESTIGATOR: George Y. C. Wong, Ph.D.

CONTRACTING ORGANIZATION: Strang Cancer Prevention Center
New York, New York 10021-4601

REPORT DATE: July 2002

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 2002	Annual (1 Jul 01 - 30 Jun 02)

4. TITLE AND SUBTITLE
Statistical Analysis of Multivariate Interval Censored Data in Breast Cancer Follow-Up Studies

5. FUNDING NUMBERS
DAMD17-99-1-9390

6. AUTHOR(S)
George Y. C. Wong, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Strang Cancer Prevention Center
New York, New York 10021-4601

E-MAIL:
gwong@strang.org

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

The overall objective of our research proposal is nonparametric inference of the joint survival function $S(x_1,...,x_d) = \Pr(X_1 > x_1,...,X_d > x_d)$ of d (≥ 2) time-to-event variables X_1, ..., X_d, each of which is subject to interval censoring. The standard estimator of S is the generalized maximum likelihood estimator (GMLE) $\hat{S}$. However, $\hat{S}$ cannot be expressed in a closed-form expression and its statistical properties have not been studied in the multivariate case. The technical objectives of this pioneer methodological research proposal are to develop asymptotic generalized maximal likelihood (GML) inference of S and to derive efficient computational algorithms for the GML procedure. In our third year of research, we have investigated the asymptotic behavior of the GMLE $\hat{\rho}$ of the correlation coefficient ρ between a pair of time-to-event variables. We have established consistency and asymptotic normality for $\hat{\rho}$ under the assumption that the censoring distribution is discrete and finite. The results will be useful to breast cancer researchers pursuing chemoprevention intervention trials involving multiple surrogate endpoint biomarkers, and genetic epidemiologists conducting studies on familial aggregation of breast cancer and related cancers.

14. SUBJECT TERMS
breast cancer, multivariate interval censored data, generalized maximum likelihood, consistency, asymptotic normality and efficiency

15. NUMBER OF PAGES
10

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

___ Where copyrighted material is quoted, permission has been obtained to use such material.

___ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

___ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

N/A For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

George Wong

7/29/02

principal Investigator

A. TABLE OF CONTENTS

B. INTRODUCTION

Interval-censored (IC) data are encountered in three areas of breast cancer research. The most common application is in clinical relapse follow-up studies in which the study endpoint is disease-free survival. When a patient relapses, it is usually known that the relapse takes place between two follow-up visits, and the exact time to relapse is unknown. In statistics, we say relapse time is interval censored. Interval censoring is also encountered in breast cancer registry studies in which information on family history of cancer is updated periodically. The Strang Breast Surveillance Program for women at increased risk for breast cancer, for instance, has enlisted over 800 women with complete pedigree information which is verified and updated continuously. Family history data such as age at diagnosis of a specific cancer, or a benign but risk-conferring condition, are obtained from each registrant at each update. Time to a cancer event, and definitely time to first detection of a benign condition, are at best known to fall in the time interval between the last update and age at diagnosis. A third but increasingly important area of application of interval censoring is in breast cancer chemoprevention experiments or prevention trials, which involve the observation of one or more surrogate endpoint biomarkers (SEB) over time. The scientific question of interest here is the estimation of time for the SEB to reach a target value, and time from cessation of intake of a chemopreventive agent to the loss of its protective effect. Unfortunately, the exact values of both these time variables are known only to lie in between two successive assay inspection times.

Let X denote a time-to-event variable with distribution $F(x) = Pr(X \leq x)$, or equivalently, survival function $S(x) = 1 - F(x)$. In interval censoring, X is not observed and is known only to lie in an observable interval (L, R). In our previous DOD funded grant, we have made fundamental contributions to both the theory of the generalized maximum likelihood (GML) estimation of S, and the computation in connection with the inference of GML estimator (GMLE) $\hat{S}$ of S. These contributions are restricted to the case of univariate interval-censored data.

Multivariate interval censoring involves $d \geq 2$ correlated X variables, each of which is subject to interval censoring. The main statistical concern here is the GML estimation of the joint survival function $S(x_1, ..., x_d) = Pr(X_1 > x_1, ..., X_d > x_d)$, and the correlations among the variables. Our interest in multivariate IC data is driven by needs arising from two related areas of breast cancer research at Strang. First, our investigators in the Strang Cancer Genetics Program want to study various patterns of familial aggregation of breast, ovarian and other forms of cancer using family history data from the Strang Breast

Surveillance Program. Studies of familial early onset of breast cancer, breast-ovarian and breast-prostate associations will lead to multivariate IC data of high dimensions; therefore, a proper statistical procedure together with a feasible software to deal with such data are very much needed. Second, we are conducting a one-year chemoprevention trial of indole-3-carbinol (I3C) for breast cancer prevention. In this prevention trial we are monitoring the levels of two SEB's, a urinary estrogen metabolite ratio and a blood counterpart, both of which are subject to interval censoring. An earlier dose-ranging study of I3C conducted by Wong *et al* [1] has been published.

Statistical analysis of multivariate IC data has never been attempted. In the multivariate situation, modeling of the intercorrelated time-to-event variables and their dependency structure will require a great deal of innovative thinking; moreover, GML computation in realistic sample sizes can be prohibitively difficult.

The overall aim of this research proposal is to develop statistical inference for multivariate interval-censored data that are encountered in breast cancer chemoprevention trials employing multiple surrogate endpoint biomarkers, and in breast cancer registry follow-up studies of familial aggregation of breast and other forms of cancer. Asymptotic generalized maximum likelihood theory has been investigated and computer software package for maximum likelihood inference and Kaplan-Meier type survival plots has been implemented.

C. BODY

Consider nonparametric estimation of the joint survival function $S(x_1, ..., x_d) = \Pr(X_1 > x_1, ..., X_d > x_d)$ of $d \geq 2$ intercorrelated time-to-event variables X_1, ..., X_d, each of which is subject to interval censoring. For ease of presentation and without any loss of generality, we shall restrict our discussion to the bivariate case $\underline{X} = (X_1, X_2)$.

Let (U_i, V_i) denote two consecutive follow-up times corresponding to X_i, and (L_i, R_i) denote the observable interval-censored (IC) data for X_i defined as

$$(L_i, R_i) = \begin{cases} (0, U_i) & \text{if } X_i \leq U_i, \\ (U_i, V_i) & \text{if } U_i < X_i \leq V_i, \\ (V_i, +\infty) & \text{if } X_i > V_i, \end{cases} \tag{1}$$

for $i = 1, 2$. Under this two-dimensional interval censorship model, data are always interval censored, i.e., $L_i < R_i$ with probability one. If we allow the possibility of having exact observations in the data, so that

$$L_i = R_i = X_i, \tag{2}$$

then (1) and (2) together define a two-dimensional mixed interval censorship model.

Let B_i denote any one of $[0, U_i]$, $(U_i, V_i]$ and $(V_i, +\infty)$. Therefore, a bivariate IC data point is a rectangular region in $\mathcal{R}^2$ taking one of the nine forms in $\mathcal{B} = \{B_k \times B_l : k, l = 1, 2, 3\}$. Given a sample of size n, the observations $(L_{i1}, R_{i1}, L_{i2}, R_{i2})$ can be represented by rectangle subsets $I_i \in \mathcal{B}$, for $i = 1, ..., n$. Define a maximal intersection (MI) A of the observable rectangles $I_1, ..., I_n$, to be a nonempty finite intersection of the I_i's such that $A \cap I_i = \emptyset$ or A, for each i. Let $A_1, ..., A_m$, denote the distinct maximal intersections with respect to $I_1, ..., I_n$.

The generalized likelihood function of S is given by $\Lambda_n = \mu_S(I_1) \times \cdots \times \mu_S(I_n)$, where $\mu_S(\cdot)$ is the probability measure induced by S. Wong and Yu [2] show that the GMLE $\hat{S}$, which maximizes Λ_n, must assign all the probability masses $s_1, ..., s_m$ to $A_1, ..., A_m$. In general, $\hat{S}$ has to be obtained iteratively. Since $\hat{S}$ is also a self-consistent estimate (SCE), we can implement the SCE algorithm by solving for $\hat{s}_1, ..., \hat{s}_m$ in

$$s_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{ij} s_j}{\sum_{k=1}^{m} \delta_{ik} s_k},$$

$j = 1, ..., m$, where $\delta_{ij} = \mathbf{1}[A_j \subset I_i]$, $\mathbf{1}[\cdot]$ denoting the indicator function, and obtain an SCE of $S(\underline{x})$

$$\tilde{S}(\underline{x}) = \sum_{A_j \subset (x_1, +\infty) \times \cdots \times (x_d, +\infty)} \hat{s}_j.$$

With starting values $s_j^{(0)} = 1/m$ for all j, $\tilde{S}(\underline{x})$ is the GMLE at convergence.

In the first and second years of our research, we have established consistency and asymptotic normality of the GMLE $\hat{S}(\underline{x})$ under both discrete and continuous assumptions. Additionally, we have derived asymptotic properties of the weighted Kaplan-Meier test statistics given by

$$D = \int_{\underline{x} \geq 0} W(\underline{x})(\hat{S}_A(\underline{x}) - \hat{S}_B(\underline{x}))d\underline{x},$$

where $W(\cdot)$ is a given weight function, and A and B refer to two comparison conditions.

A key feature of multivariate IC data and a parameter of substantive importance is the correlation coefficient ρ between pair of the X variables, say X_1 and X_2. The GMLE of $\rho(X_1, X_2)$ is

$\hat{\rho}(x_1, x_2)$

$$= \frac{\int\int x_1 x_2 d\hat{F}(x_1, x_2) - \int\int x_1 d\hat{F}(x_1, x_2) \int\int x_2 d\hat{F}(x_1, x_2)}{\{[\int\int x_1^2 d\hat{F}(x_1, x_2) - (\int\int x_1 d\hat{F}(x_1, x_2))^2][\int\int x_2^2 d\hat{F}(x_1, x_2)(\int\int x_2 d\hat{F}(x_1, x_2))^2]\}^{1/2}}.$$

In a follow-up study involving interval censoring, it is often the case that not all events will take place by the end of the study. In this situation, $\hat{\rho}$ will not provided a consistent estimate of ρ. Let τ denote the largest follow-up time. A more appropriate correlation coefficient to consider is

$$\rho_\tau(x_1, x_2) = \frac{Cov(X_1, X_2 | X_1, X_2 \leq \tau)}{\sqrt{Var(X_1|X_1 \leq \tau) Var(X_2|X_2 \leq \tau)}}.$$

$\hat{F}$, the GMLE of F_o, is a discrete cdf with discontinuity points at the upper-right vertexes of the maximum intersections. Without loss of generality, let $a_1 < \cdots < a_m$ be the set of partition points of the real line such that the set $\{(a_i, a_j) : \ i, j \in \{0, 1, ..., m, m+1\}\}$ contains all the discontinuity points of $\hat{F}$, where $a_0 = -\infty$ and $a_{m+1} = \infty$. Let $\hat{s}_{ij}$ denote the GMLE of the bivariate probability weight assigned to (a_i, a_j) by $\hat{F}$. The GMLE of ρ_τ is given by

$$\hat{\rho}_\tau = \frac{E_{00}E_{12} - E_{10}E_{02}}{\sqrt{[E_{00}E_{11} - (E_{10})^2][E_{00}E_{22} - (E_{02})^2]}},$$

where $E_{12} = \sum_{a_i, a_j < \infty} a_i a_j \hat{s}_{ij}$, $E_{00} = \sum_{a_i, a_j < \infty} \hat{s}_{ij}$, $E_{10} = \sum_{a_i, a_j < \infty} a_i \hat{s}_{ij}$, $E_{02} = \sum_{a_i, a_j < \infty} a_j \hat{s}_{ij}$, $E_{11} = \sum_{a_i, a_j < \infty} a_i^2 \hat{s}_{ij}$, and $E_{22} = \sum_{a_i, a_j < \infty} a_j^2 \hat{s}_{ij}$.

From the consistency results of Wong and Yu [2], and Yu [3] we can show that $\hat{\rho}_\tau$ is consistent under the assumption that the union of the support sets of censoring variables is dense. Moreover, if the range of the censoring vector is finite, $\hat{\rho}_\tau$ can be shown to be asymptotically normally distributed. The asymptotic variance of $\hat{\rho}_\tau$ can be estimated by

$$\hat{\sigma}^2 = B\mathcal{I}^{-1}B',$$

where $B = \frac{\partial \rho_\tau}{\partial \mathbf{s}}$, $\mathbf{s} = \{s_{ij} : \ (i,j) \neq (m,m)\}'$, and $\mathcal{I}$ is the information matrix, that is

$$\mathcal{I} = -\frac{\partial^2 \mathrm{lnL}}{\partial \mathbf{s}' \partial \mathbf{s}}.$$

We are preparing a manuscript to report these findings.

When the finite distribution assumption regarding the censoring vector is not met, we shall have to resort to the proposed bootstrap method (Task 5) to investigate the asymptotic behavior of $\hat{\rho}_\tau$. We shall devote our effort to this research topic in the fourth year of no-cost extension of our DOD grant.

D. KEY RESEARCH ACCOMPLISHEMENTS

- We have expanded the scope of Task 5 to include a theoretical consideration for the GMLE $\hat{\rho}$ of the correlation coefficient ρ between a pair of correlated time-to-event variables.
- We have established consistency and asymptotic normality of $\hat{\rho}$ under a finite distribution assumption.

E. REPORTABLE OUTCOMES

- 2 published articles in journals : [2], [4].
- Computer programs for comprehensive GML inferences installed in http://www.math.binghamton.edu/qyu/index/html.

F. CONCLUSIONS

In the past three year of our DOD grant, we have successfully accomplished our research objectives stated in Tasks 1 - 4 and part of Task 5. Under the multivariate interval censorship model, we have established consistency, asymptotic normality and asymptotic efficiency of the GMLE under various assumptions. We have encountered and conquered a methodological problem arising from the unexpected outcome that $\hat{S}$ may not be unique in multivariate interval censoring. Also, we have derived asymptotic results for the GMLE of the correlation coefficient between a pair of correlated time-to-event variables under finite distribution assumption. Finally, we have implemented computer programs for carrying out the asymptotic GML procedure.

The results which we have established will be useful to breast cancer researchers pursuing chemoprevention intervention trials involving multiple surrogate endpoints biomarkers, and genetic epidemiologists conducting studies on familial aggregation of breast cancer and related cancers.

G. REFERENCES

[1] Wong, G. Y. C., Bradlow, H. L., Sepkovic, D., Mehl, S., Mailman, J. and Osborne, M. P. (1997). A dose-ranging study of indole-3-carbinol for breast cancer prevention. *Journal of Cellular Biochemistry Supplements* 28/29, 111-116.

[2] Wong, G. Y. C. and Yu, Q. Q. (1999). Generalized MLE of a joint distribution function with multivariate interval-censored data. *J. of Multi. Anal.* 69, 155-166.

[3] Yu, Shaohua. (2000). Consistency of the generalized MLE with multivariate mixed case interval-censored data. Ph.D thesis, Binghamton University.

[4] Yu, Q.Q, Wong, G.Y.C. and He, Q.M. (2000). Estimation of a joint distribution function with multivariate interval-censored data when the nonparametric MLE is not unique. *Biometrical Journal*, 42, 747-763.