AD________________

Award Number: DAMD17-99-1-9390

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

PRINCIPAL INVESTIGATOR: George Wong, Ph.D.

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

REPORT DATE: July 2001

TYPE OF REPORT: Annual

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

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

# REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

| | | |
|---|---|---|
| **1. AGENCY USE ONLY (Leave blank)** | **2. REPORT DATE**<br>July 2001 | **3. REPORT TYPE AND DATES COVERED**<br>Annual (1 July 2000 - 30 June 2001) |

**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 Wong, Ph.D.

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Strang Cancer Prevention Center
New York, New York 10021-4601
email: 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 $d$ ($\geq 2$) 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 second year of research, we have established asymptotic distribution of the GMLE and the weighted Kaplan-Meier test statistics. Additionally, we have resolved statistical problems arising from an unexpected finding that $\hat{S}$ may be non-unique in multivariate interval-censored data. 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-Cencored Data,
Genrealized Maximum Likelihood Consistency

**15. NUMBER OF PAGES**
11

**16. PRICE CODE**

| | | | |
|---|---|---|---|
| **17. SECURITY CLASSIFICATION OF REPORT**<br>Unclassified | **18. SECURITY CLASSIFICATION OF THIS PAGE**<br>Unclassified | **19. SECURITY CLASSIFICATION OF ABSTRACT**<br>Unclassified | **20. LIMITATION OF ABSTRACT**<br>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, 6/28/01
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 will be investigated and computer software package for maximum likelihood inference and Kaplan-Meier type survival plots will be 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

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

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

$$L_i = R_i = X_i, \qquad (C.2)$$

then (C.1) and (C.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 $\Lambda_n$, 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

$$s_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{ij} s_j}{\sum_{k=1}^{m} \delta_{ik} s_k}, \qquad (C.3)$$

$j = 1$, ..., $m$, 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.

We have implemented an algorithm to identify MI's corresponding to a set of rectangle $I_1$, ..., $I_n$, and a computer program to calculate the GMLE iteratively. The programs are installed in the internet site math.binghamton.edu/ftp/pub/qyu. This completes Task 1. We have established uniform consistency of $\hat{S}$ by proving

$$\Pr\{\lim_{n\to\infty} \sup_{\underline{x} \text{ is observable}} |\hat{S}(\underline{x}) - S(\underline{x})| = 0\} = 1$$

under condition C1 (Task 2a) and under condition C2 (Task 2b):

C1. The censoring vectors $(U_1, V_1)$ and $(U_2, V_2)$ take on countably many values.

C2. The censoring distribution $G$ of $(U, V)$ is continuous, and some regularity assumptions are imposed on either $S$ or $G$.

The above consistency results that we have accomplished in our first two years of research are published in a peer-reviewed statistical journal ([2]) and reported in the Ph.D thesis [9] of Dr. Shaohua Yu under the supervision of Professor Qiqing Yu.

Asymptotic normality results are fundamentally important for confidence statements and hypothesis testing in data analysis. We have proved that $\sqrt{n}(\hat{S}(\underline{x}) - S(\underline{x}))/\hat{\sigma} \xrightarrow{\mathcal{D}} N(0,1)$, where $\hat{\sigma}^2$ is the inverse of the observed Fisher information number, or equivalently, $\hat{S}$ is both asymptotically normal and asymptotically efficient under condition D1 (Tasks 3a,b):

D1. $(U_1, V_1)$ and $(U_2, V_2)$ take on finitely many values, say $\underline{a_1}$, ..., $\underline{a_N}$, and $S(\underline{a_k}) > S(\underline{a_l})$, if $a_{k1} \le a_{l1}$ and $a_{k2} \le a_{l2}$ with at least one strict inequality, $\underline{a_k} = (a_{k1}, a_{k2})$ and $\underline{a_l} = (a_{l1}, a_{l2})$.

The above asymptotic normality results that we have accomplished in our first two years of research are published in a peer-reviewed statistical journal ([2]).

Our research effort for the second year is focused on the asymptotic normality of the GMLE under conditions D2 and D3 (Tasks 3c,d,e,f):

D2. $S$ is arbitrary, $(U_1, V_1)$ and $(U_2, V_2)$ takes on countably many values, and the strict monotonicity condition in D1 holds.

D3. $S$ is arbitrary, $G$ is continuous, and either $S$ or $G$ meets some reasonable smooth regularity conditions.

The following are results we have established in the second year of our research:

1. If there are no exact observations in the data, then the asymptotic normality for the GMLE does not hold under assumption D3. If the GMLE has an asymptotic normal distribution, then its marginal asymptotic distribution function must also be a univariate normal distribution. However, Groeneboom and Wellner [3] have shown that if the underlying distribution functions are continuous, then the GMLE will not have an asymptotic normal distribution. Thus the GMLE of $S$ with multivariate interval-censored data cannot be asymptotic normal under conditions D3.

2. If there exist exact observations in the data, then under assumption D2 or D3, the GMLE of $S$ has asymptotic normality and efficiency under the mixed interval censorship

model. A manuscript that summarizes the asymptotic normality under assumptions D2 or D3 for such a model is being prepared.

We have also worked on Task 4. In particular, we have studied the large-sample properties of the weighted Kaplan-Meier test statistics given by

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

where $W(\cdot)$ is a given weight function, and $A$ and $B$ refer to two comparison conditions. Under condition D1, we have established consistency and asymptotic normality of the statistic $D$ (Task 4a).

When $W(\underline{x}) = 1$, and $A$ and $B$ represent two independent samples, then $D = \int_{\underline{x}\geq 0} \hat{S}_A(\underline{x})d\underline{x} - \int_{\underline{x}\geq 0} \hat{S}_B(\underline{x})d\underline{x}$, $Var(D) = Var(\int_{\underline{x}\geq 0} \hat{S}_A(\underline{x})d\underline{x}) + Var(\int_{\underline{x}\geq 0} \hat{S}_B(\underline{x})d\underline{x})$. A consistent estimator of $Var(D)$ can easily be derived, and the P-value of $D$ can be computed.

We have also studied other weight functions $W(\cdot)$ and the case that $A$ and $B$ are not independent. The derivation is not as simple and will not be discussed here. A manuscript that derives the asymptotic distribution of $D$ when $W(x) \not\equiv 1$ or the sets $A$ and $B$ are not independent is under preparation (Tasks 4b,c).

Furthermore, we have encountered a non-uniqueness problem in the GML inferences that we had not expected when we submitted our proposal, namely, the solution of the GMLE of $S$ for some multivariate interval-censored data is not unique. We point out that the GMLE solution for univariate interval-censored data is always unique. As a consequence, the sample information matrix $J_{\hat{S}} = -\left(\frac{\partial^2 \log \Lambda_n}{\partial s_i \partial s_j}\right)_{(m-1)\times(m-1)}\Big|_{\mathbf{S}=\hat{\mathbf{S}}}$, where $\Lambda_n$ is the generalized likelihood function, is singular and its inverse does not exist. Since the variance estimation of the GMLE is based on the inverse of $J_{\hat{S}}$, it is therefore important to resolve the non-uniqueness problem.

The program for deriving a GMLE estimator of $S$ that we have accomplished in the first year of our project is still applicable, even if there are multiple solutions. However, if there are multiple solutions, the program cannot provide an estimator of the variance of the GMLE and thus cannot provide confidence intervals. We present an artificial bivariate data set that gives rise to non-uniqueness of the GMLE solution.

**Example 1** Suppose that a sample of size 4 consists of observations $(L_{i1}, R_{i1}, L_{i2}, R_{i2})$, $i = 1, ..., 4$, which equal $(1,6,1,3)$, $(1,6,4,6)$, $(1,3,1,6)$ and $(4,6,1,6)$, respectively. Then the MI's are $A_1 = (1,3]\times(1,3]$, $A_2 = (1,3]\times(4,6]$, $A_3 = (4,6]\times(1,3]$ and $A_4 = (4,6]\times(4,6]$. $\hat{\mathbf{S}}_q = q(1/2,0,0,1/2) + (1-q)(0,1/2,1/2,0)$, $q \in (0,1)$, are all GMLEs of $\mathbf{S}$. The sample information matrix $J_{\hat{\mathbf{S}}}$ is

$$\begin{pmatrix} a_{12}+a_{13}+a_{24}+a_{34} & a_{12}+a_{34} & a_{13}+a_{24} \\ a_{12}+a_{34} & a_{12}+a_{34} & 0 \\ a_{13}+a_{24} & 0 & a_{13}+a_{24} \end{pmatrix}$$

where $a_{ij} = (s_i + s_j)^{-2}$. Note that the first column of the matrix is the sum of the next two columns. Consequently, $J_{\hat{\mathbf{S}}}$ is singular, and the inverse matrix of $J_{\hat{\mathbf{S}}}$ does not exist.

We have proposed a method to resolve the non-uniqueness problem. By proving a result in linear algebra, we have proposed to estimate $S$ by a special GMLE, and to estimate its covariance matrix by a certain procedure. There are four main steps:

1. Using the self-consistent algorithm, we first find a GMLE of $F$, denoted by $\hat{F}$.
2. We established in [8] that each solution to the system of equations

$$\sum_{j=1}^{m} \delta_{ij} s_j = \mu_{\hat{F}}(I_i), \ i = 1, ..., n, s_j \geq 0, \ \sum_{j=1}^{m} s_j = 1, \tag{C.4}$$

is a GMLE of $\underline{s}$, where $\mu_{\hat{F}}$ is the probability measure induced by $\hat{F}$. The information in (C.4) can be formulated in a matrix form

$$A\underline{s} = \underline{b}. \tag{C.5}$$

Before we discuss the last two steps, we pointed out that in general, given an $(n+1) \times m$ dimensional matrix $A$ with rank $r < m-1$, an $m \times 1$ dimensional vector $\underline{s}$ and an $(n+1) \times 1$ dimensional vector $\underline{b}$, if the linear equation $A\underline{s} = \underline{b}$ has a non-zero solution, then the solution is not unique and the solutions can be written as the form

$$(s_{i_{r+1}}, ..., s_{i_m})' = B(s_{i_1}, ..., s_{i_r})', \ s_{i_1}, ..., s_{i_r} \in \mathcal{R},$$

where $B$ is a $(m-r) \times r$ dimensional matrix and $(i_1, ..., i_m)$ is a permutation of $(1, ..., m)$. However, there is no guarantee that

$$\text{there is a solution that satisfies } s_i \geq 0, \ i = 1, \ ..., \ m \text{ and } s_{i_{r+2}} = \cdots = s_{i_m} = 0. \tag{C.6}$$

3. We established in [8] that (C.6) holds for equation (C.4) or (C.5) and proposed a procedure to identify the indexes $i_{r+2}$, ..., $i_m$.
4. Then the likelihood function of $\underline{s}$ with $s_{i_{r+2}} = \cdots = s_{i_m} = 0$ and $\sum_{j=1}^{r+1} s_{i_j} = 1$ will have a non-singular Fisher information matrix. We propose to find a GMLE of $\underline{s}$ with $s_{i_{r+2}} = \cdots = s_{i_m} = 0$.

For ease in understanding, we illustrate our idea with Example 1 above. Note that a GMLE assigns weight 1/4 to each of the 4 MI's (Step 1). Each solution $\underline{s}$ $(= (s_1, s_2, s_3, s_4))$ to the set of linear equations

$$s_1 + s_2 = 1/2, \ s_1 + s_3 = 1/2, \ s_2 + s_4 = 1/2, \ s_3 + s_4 = 1/2, \ \sum_{i=1}^{4} s_i = 1,$$

is a GMLE of $\underline{s}$ (Step 2). The equations can be written in the matrix form $A\underline{s} = \underline{b}$:

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 1 \end{pmatrix},$$

where the rank of $A$ is $r = 2$.

For these equations, there exists a solution such that $s_4 = 0$ and $s_i \geq 0$. In fact, the solution $\underline{\hat{s}} = (0, 0.5, 0.5, 0)$ is another GMLE (Step 3).

The likelihood function is given by

$$\Lambda_n = (s_1 + s_2)(s_1 + s_3)s_2 s_3 = (1 - s_3)(1 - s_2)s_2 s_3.$$

For this choice of $\underline{\hat{s}}$, the covariance matrix of $(\hat{s}_2, \hat{s}_3)$ is estimated by the inverse of the matrix

$$-\frac{\partial^2 \log \Lambda_n}{\partial s_2 \partial s_3}\bigg|_{(s_2,s_3)=(\hat{s}_2,\hat{s}_3)} = \begin{pmatrix} \frac{1}{s_2^2} + \frac{1}{(1-s_2)^2} & 0 \\ 0 & \frac{1}{s_3^2} + \frac{1}{(1-s_3)^2} \end{pmatrix}\bigg|_{(s_2,s_3)=(\hat{s}_2,\hat{s}_3)} \quad \text{(Step 4).}$$

We have established consistency and asymptotic normality of this procedure under certain regularity conditions. Our research here extends the requirements of Tasks 2 and 3 in the original proposal. For more details we refer to [8].

## D. KEY RESEARCH ACCOMPLISHMENTS IN THE SECOND YEAR

- We have completed most of Task 3.

  The GMLE of the distribution function is studied and its consistency and asymptotic normality are established under various assumptions (C!, C2, D1, D2) on the censoring random vectors. Part of our results are published in a peer-reviewed journal (see [2]). The rest will be organized in two manuscripts under preparation.

- We have completed most of Task 4.

  We have established consistency and asymptotic normality of the statistic $D$. The results are summarized in a manuscript which is under preparation.

- We have resolved the non-uniqueness problem in the GML inferences. Consistency and asymptotic normality (Extensions of Tasks 1, 2 and 3) of the procedure proposed have been established. The result is published in a peer-reviewed journal (see [8]).

- We have developed computer software packages for implementing the GML inferences when the GMLE with multivariate interval-censored data is not unique. It is an extension of Task 1.

## E. REPORTABLE OUTCOMES

- 6 published articles in journals cited in the science citation index: [2], [4], [5], [6], [7], [8].

- Computer programs installed in math.binghamton.edu.ftp/pub/qyu.

## F. CONCLUSIONS

In the first two year of our DOD grant, we have successfully accomplished our research objectives stated in Tasks 1, 2, 3 and 4. Under the multivariate interval censorship models, we have established consistency, asymptotic normality and asymptotic efficiency of the GMLE under various assumptions. We have solved a non-uniqueness problem that occurs in multivariate interval censoring, but not in univariate interval censoring. Moreover, 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] Groeneboom, P. and Wellner, J.A. (1992). Information bounds and nonparametric maximum likelihood estimation. *Birkhäuser Verlag, Basel.*

[4] Yu, Q. Q., Li, L. X. and Wong, G. Y. C. (2000). On consistency of the self-consistent estimator of survival functions with interval censored data. *Scan. J. of Statist.*, 27, 35-44.

[5] Schick, A. and Yu, Q. Q. (2000). Consistency of the GMLE with mixed case interval-censored data. *Scan. J. of Statist.*, 27, 45-55.

[6] Yu, Q. Q. and Li, L.X. (2001). Asymptotic properties of the GMLE of self-consistent estimators with doubly-censored data. *Acta Math. Sinica*, (In press).

[7] Yu, Q. Q., Wong, G. Y. C. and Li, L. X. (2001). Asymptotic properties of self-consistent estimators with mixed interval-censored data. *Ann. Inst. Statist. Math.*, (In press).

[8] 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.

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