Understanding progression of ovarian cancer in US from SEER data: a semiparametric joint model for bivariate survival distribution with interval sampling

Hong Zhu, PhD, Candidate
Advisor: Prof. Mei-Cheng Wang
Department of Biostatistics, Johns Hopkins University

Background
- Ovarian cancer is the 5th leading cause of death from cancer in women in US.
- 1 out of 40-60 women in US has a lifetime chance of developing ovarian cancer.
- Diagnosis with a late stage and poor prognosis, due to non-specific symptoms.

Objective
- Identify the dynamic disease progression: Birth→Ovarian cancer onset→Death.
- Outcomes of interest: age of onset and survival after onset.
- Model bivariate survival distribution of outcomes: Estimate marginal distributions of age of onset, survival time after onset, and their joint distribution.

Bias from interval sampling
Data are collected under ‘Interval sampling’:
- Cancer-onset is identified within a calendar time period: double truncation.
- Birth time can be retrospectively confirmed.
- Death is observed subject to right censoring.

Conventional statistical methods (empirical distribution, Kaplan-Meier estimator) may fail and lead to biased results.

Statistical method: Notations and Assumption
- T: birth time with distribution \( G(t) \), Y: time from birth to cancer onset, Z: time from cancer onset to death.
- \( Y \) observed subject to double truncation, \( Z \) observed subject to right censoring, \( Y \) and \( Z \) are correlated.

Model assumption: Semimodular condition
Age of onset and survival after onset are independent of birth cohort.

Statistical method: Probability Structure
Joint density of observed uncensored \((T, Y, Z)\) under semimodular condition:
\[
 p(t, y, z) = \frac{G(t) f_T(t) f_Y(y) I(T < t < T_0-y-z) g(Y, Z; T)}{G_T(t) f_Y(y) I(T_0-y-z \leq Y) g(Y, Z; T)}
\]
where \( g, f \) denote population densities of \( T \) and \( (Y, Z) \).

Semiparametric joint model
Consider a joint model of \( T \) and \( (Y, Z) \):
- Focus on the estimation of parameter in \( g(\cdot | \theta) \) and semiparametric estimation of \( S(y, z) \).

Model based birth rate
Model-based birth density plots

Exploratory plots of ovarian cancer statistics

Estimated joint survival functions
Table: Joint estimates of \( S(y, z) \) from the SEER ovarian cancer data
(a) joint survival function estimates by conventional method, (b) Proposed estimates of joint survival function \( S(y, z) \), (c) Standard error estimates of \( S(y, z, \theta) \) (Bootstrap).

What the table tells?
- Conventional method generally leads to underestimation comparing to the proposed method.
- Proportions of patients who was diagnosed later than 62 years old and survived longer than 4.6 years by conventional method and proposed method are 16% (95% CI: 15% – 17%) vs 22% (95% CI: 19% – 25%).
- 27% (95% CI: 21% – 33%) of the ovarian cancer patients in the SEER program were diagnosed later than 70 years old and survived longer than 1.6 years.

Age of cancer onset distribution and Survival after onset
Plots of estimated distributions of age of onset and survival functions after cancer
(red lines are by proposed method, black lines are by conventional method.)

Naive estimators based on observed data

Exploratory analysis of age of onset impact on survival time after onset
Plots of estimated survival functions of time after cancer, by age of onset

Estimated median age of onset is 77.0 years (observed: 69.8). Estimated median survival time after onset is 2.62 years (observed: 1.58).

Later age of onset corresponds to lower survival probability.

Conclusions and ongoing research
- Method developed to study the ovarian cancer progression in SEER data: Potential bias from data collection mechanism is corrected.
- Assessment of treatment effect (Surgery, Chemotherapy...): bivariate regression model with interval sampling.
- Proposed statistical method extend to Rakai community-based cohort data to study HIV/AIDS epidemiology and prevention (PIs. Dr. Ronald Gray and Dr. Maria Wauer, JHSPPH)