Super Learner for Survival Prediction in Case-Cohort and Generalized Case-Cohort Studies

Jianwen Cai, Haolin Li, Haibo Zhou, David Couper

Department of Biostatistics, University of North Carolina at Chapel Hill

ABSTRACT

The case-cohort study design is often used in modern epidemiological studies of rare diseases, as it can achieve similar efficiency as a much larger cohort study with a fraction of the cost. Previous work focused on parameter estimation for case-cohort studies based on a particular statistical model, but few discussed the survival prediction problem under such type of design. In this article, we propose a super learner algorithm for survival prediction in case-cohort studies. We further extend our proposed algorithm to generalized case-cohort studies. The proposed super learner algorithm is shown to have asymptotic model selection consistency as well as uniform consistency. We also demonstrate our algorithm has satisfactory finite sample performances. Simulation studies suggest that the proposed super learners trained by data from case-cohort and generalized case-cohort studies have better prediction accuracy than the ones trained by data from the simple random sampling design with the same sample sizes. Finally, we apply the proposed method to analyze a generalized case-cohort study conducted as part of the Atherosclerosis Risk in Communities (ARIC) Study.

Keywords: Cost-Efficient Design; Ensemble Learning; Epidemiological Studies; Survival Analysis

Efficient Case-Cohort Design Using Balanced Sampling

Kaeum Choi[1], **Sangwook Kang[2]**

[1]*Department of Biostatistics and Bioinformatics, Emory University, Georgia, USA*

[2]*Department of Applied Statistics, Yonsei University, Seoul, Republic of Korea*

ABSTRACT

The case-cohort design is a cost-efficient two-phase design for analyzing survival data when key risk factors are expensive to assess and the event rate is low. Traditionally, subcohorts are selected via simple random sampling, which might not fully utilize available information. In this talk, we introduce an efficient sampling design using balanced sampling for subcohort selection within the case-cohort design. A notable benefit of employing balanced sampling is the automatic calibration of auxiliary variables available for the entire cohort. Under a Cox model, it has been demonstrated that the calibration of sampling weights, utilizing auxiliary variables highly correlated with the main risk factor, significantly enhances the efficiency of regression coefficient estimators. Extensive simulation experiments show the reduced variabilities under the proposed approach in comparison to those under both simple random sampling. The proposed design and estimation procedure are further illustrated using the well-established National Wilms Tumor Study dataset.

Keywords: Calibration; cohort sampling; Cox model; sampling weights; survival analysis

Improving Efficiency of Risk Prediction with Subsampled Cohort Data

Yei Eun Shin[1]

[1]*Department of Statistics, Seoul National University*

ABSTRACT

In large cohort studies, subsampling designs such as case-cohort and nested case-control sampling are widely used to improve efficiency and reduce costs. However, these designs introduce methodological challenges for survival analysis, particularly when estimating absolute risk, considering competing events, or handling time-varying covariates. This talk introduces recent methods that aim to improve the performance of survival analysis under such designs. The first part presents approaches for improving the estimation and validation of risk prediction models including absolute risks, proportional and additive hazards models when risk factors are not fully available in a cohort. For competing risk analysis with event-specific subsamples, a proportional risk model provides a simple and statistically efficient way in a joint framework. Additional topics include estimating transition probabilities in multi-state models and applying landmarking methods when only limited subsampled data are available. Together, these methods illustrate how targeted use of subsampled data via influence functions can support efficient and flexible survival analysis, even when full cohort data are not available.

Keywords: influence function; risk prediction model; two-phase sampling designs; weight calibration

Semiparametric Regression Analysis of Case-Cohort Studies with Multiple Interval-Censored Disease Outcomes

Haibo Zhou[1], Qinging Zhou[2], Jianwen Cai[1]

[1]*University of North Carolina at Chapel Hill, USA*

[2]*University of North Carolina at Charlotte, USA*

ABSTRACT

In this work, we formulate the case-cohort design with multiple interval-censored disease outcomes and generalize it to nonrare diseases where only a portion of diseased subjects are sampled. We develop a marginal sieve weighted likelihood approach, which assumes that the failure times marginally follow the proportional hazards model. We consider two types of weights to account for the sampling bias, and adopt a sieve method with Bernstein polynomials to handle the unknown baseline functions. We employ a weighted bootstrap procedure to obtain a variance estimate that is robust to the dependence structure between failure times. The proposed method is examined via simulation studies and illustrated with a dataset on incident diabetes and hypertension from the Atherosclerosis Risk in Communities study.

Keywords: case-cohort design, robust inference, sieve estimation, survival analysis