Statistica Sinica **36** (2026), 1-21
doi:https://doi.org/10.5705/ss.202023.0373

ROBUST AND EFFICIENT CASE-CONTROL STUDIES WITH CONTAMINATED CASE POOLS: A UNIFIED M-ESTIMATION FRAMEWORK

Guorong Dai* and Jinbo Chen

Fudan University and University of Pennsylvania

Abstract: We consider a general M-estimation problem based on contaminated case-control data, including the primary and secondary analyses of case-control studies as special examples. The case pool contains ineligible patients who should be excluded from the study if known, but the true status of an individual in the case pool is unclear except in a small subset. Through imputing the possibly unobserved status variable with a function of all available relevant predictors, followed by an appropriate debiasing procedure, we exploit the whole sample to develop a family of robust and efficient estimators, eliminating bias from the case contamination. With the help of cross-fitting, the imputation function can be constructed using any reasonable regression or machine learning approaches. Our estimators are always root-n-consistent and asymptotically normal regardless of the imputation function's limit. Further, we explore relaxation of requirements on the imputation function. We show even without any assumption on its convergence properties, our estimators are still root-n-consistent while asymptotic normality can be achieved by a sample-splitting variant. We also demonstrate results of this type, which are entirely free of convergence assumptions on the nuisance estimators, can be extended to other problems involving nuisance functions. The finite-sample superiority of our method is demonstrated by comprehensive simulation studies. We also apply our method to analyze sepsis-related death based on a real data set from electronic health records.

Key words and phrases: Contaminated case pool, estimating equation, nuisance estimation, primary and secondary analyses of case-control data, robustness and efficiency.

1. Introduction

In epidemiology and many other biomedical fields, case-control designs have been serving as flexible and cost-effective tools for investigating risk factors for conditions of interest, e.g., the occurrence of rare diseases and disease-related mortality. In stark contrast with prospective cohort designs, a case-control sample is assembled by combining two independent subsamples drawn separately from two groups: individuals with (cases) and without (controls) the condition of interest. A detailed overview of case-control methods can be found in Breslow (1996). In biomedical research, case-control data are popularly used for two

*Corresponding author. E-mail: guorongdai@fudan.edu.cn