Statistica Sinica **35** (2025), 1-22
doi:https://doi.org/10.5705/ss.202023.0006

IDENTIFICATION AND ESTIMATION OF TREATMENT EFFECTS ON LONG-TERM OUTCOMES IN CLINICAL TRIALS WITH EXTERNAL OBSERVATIONAL DATA

Wenjie Hu[1], Xiao-Hua Zhou[1,2] and Peng Wu*[3]

[1]*Peking University,* [2]*Pazhou Lab*
and [3]*Beijing Technology and Business University*

Abstract: In biomedical studies, estimating drug effects on chronic diseases requires a long follow-up period, which is difficult to meet in randomized clinical trials (RCTs). The use of a short-term surrogate to replace the long-term outcome for assessing the drug effect relies on stringent assumptions that empirical studies often fail to satisfy. Motivated by a kidney disease study, we investigate the drug effects on long-term outcomes by combining an RCT without observation of long-term outcomes and an observational study in which the long-term outcome is observed but unmeasured confounding may exist. Under a mean exchangeability assumption weaker than the previous literature, we identify the average treatment effects in the RCT and derive the associated efficient influence function and semiparametric efficiency bound. Furthermore, we propose a locally efficient doubly robust estimator and an inverse probability weighted (IPW) estimator. The former attains the semiparametric efficiency bound if all the working models are correctly specified, which may be hard to achieve due to the intertwined working models, while the latter has a simpler form and requires much fewer model specifications. The IPW estimator using estimated propensity scores is more efficient than the one using true propensity scores and achieves the semiparametric efficient bound in the case of discrete covariates and surrogates with finite support. Both estimators are shown to be consistent and asymptotically normally distributed. Extensive simulations are conducted to evaluate the finite-sample performance of the proposed estimators. We apply the proposed methods to estimate the efficacy of oral hydroxychloroquine on renal failure in a real-world data analysis.

Key words and phrases: Data fusion, long-term treatment effects, semiparametric efficiency, surrogate.

1. Introduction

In biomedical research, randomized clinical trials (RCTs) are the gold standard for drug or therapy evaluation (Cartwright, 2010). However, the high cost of labour and material resources restricts the sample size and the duration of RCTs. Especially for chronic diseases, the important long-term outcomes are difficult to observe during the period of RCTs. As a motivating

*Corresponding author. E-mail: pengwu@btbu.edu.cn