Commentary on Chen et al. (2022): The need for continued methodological research on leveraging information in secondary endpoints for more efficient RCTs

Jack M. Wolf[a*], Joseph S. Koopmeiners[a], David M. Vock[a]

[a]Division of Biostatistics and Health Data Science, School of Public Health, University of Minnesota, 2221 University Ave SE Minneapolis, MN 55414

*Corresponding author. *E-mail address*: wolfx681@umn.edu.

Abstract: Chen et al. (2022) recently proposed a set of estimating equations that incorporate data from secondary endpoints to improve precision in parameter estimates related to a primary endpoint. We were motivated to translate their methodology to the context of randomized controlled trials to gain precision in treatment effect estimation using data from secondary endpoints. Our results suggest that this estimator cannot gain efficiency in this context because of random treatment assignment, especially when there is a treatment effect on secondary endpoints, and that further methodological work in this area is needed.

Keywords: secondary endpoints, randomized controlled trial, estimation precision

1. Introduction

Recently, Chen et al. developed a set of augmented estimating equations that directly incorporate data from auxiliary sources to gain efficiency in parameter estimates (1,2). While they were motivated by the analysis of observational data, it is natural to consider if their approach can be adapted to gain efficiency in RCTs.

Our interest is motivated by research on very low nicotine content cigarettes to inform policy-level nicotine standards that would impact all people who smoke in the United States. Although previous RCTs (3–7) have shown that these standards reduce tobacco product use behaviors on average in the general United States smoking population, additional investigations are necessary to ensure that these policies are effective in specific vulnerable subpopulations including but not limited to people who smoke with schizophrenia, Black people who smoke, and low socioeconomic status people who smoke. However, it is financially and practically infeasible to run fully powered RCTs to precisely estimate subpopulation treatment effects in all of these subpopulations. Thus, we wish to leverage additional information to gain efficiency to make such analyses feasible with smaller sample sizes–either in small individual trials or in subgroup analyses of completed trials. We hope to gain efficiency using information encoded in secondary endpoints regularly collected in these trials including biomarkers of nicotine and tobacco use and psychological measures of nicotine dependence.

2. Summary of Chen et al. (2022)

We begin by briefly summarizing the proposed estimator. Consider the following estimating equation for estimating the parameter of interest, β: $\sum_{i=1}^{n} g\left(Y_{i,1}; \beta\right) = 0$, where $Y_{i,1}$ is the

primary endpoint. They propose to fit a working model for a secondary endpoint, $Y_{i,2}$, to improve upon the efficiency for estimating β by solving the following weighted estimating equations based on empirical likelihood theory (8):

$$\sum_{i=1}^{n} \hat{p}_i \, g(Y_{i,1}; \beta) = 0. \tag{1}$$

Here, $\hat{p}_i$ maximize $\prod_{i=1}^{n} p_i$ under the following constraints:

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \, h(Y_{i,2}; \theta) = 0, \tag{2}$$

where $h(Y_{i,2}; \theta)$ is a set of estimating equations for $Y_{i,2}$ such that $\mathrm{E}\{h(Y_{i,2}; \theta_*)\} = 0$ for some θ_*. Importantly, h must be over-identified to gain efficiency. For example, instead of standard least squares linear regression on covariates X_i, which would imply $h(Y_{i,2}; \theta) = X_i(Y_{i,2} - X_i^T \theta)$, they suggest adding "redundant" covariates Z_i which are assumed to be orthogonal to the residuals: $h(Y_{i,2}; \theta) = (X_i^T, Z_i^T)^T (Y_{i,2} - X_i^T \theta)$. Simulation studies and theoretical arguments indicate that this over-identification and weighting lead to efficiency gains with $\mathrm{Var}(\hat{\beta}_{Y_1 \& Y_2}) \leq \mathrm{Var}(\hat{\beta}_{Y_1})$ where $\hat{\beta}_{Y_1}$ and $\hat{\beta}_{Y_1 \& Y_2}$ are the respective solutions to $\sum_{i=1}^{n} g\,(Y_{i,1}; \beta) = 0$ and $\sum_{i=1}^{n} \hat{p}_i \, g(Y_{i,1}; \beta) = 0$.

3. Translation to Randomized Controlled Trials

We now consider this estimator in the context of a RCT. Suppose we have data $(A_i, X_{i,1}, X_{i,2}, Y_{i,1}, Y_{i,2})_{i=1}^{n}$ where A_i is a binary treatment indicator, $(X_{i,1}, X_{i,2})$ are two baseline

characteristics, and $(Y_{i,1}, Y_{i,2})$ are measurements of the primary endpoint and secondary endpoint, respectively. Importantly, due to randomization, A_i is independent of $(X_{i,1}, X_{i,2})$. In this setting, the standard estimating equation for the average treatment effect, identified as $\beta_1 = \mathrm{E}(Y_i|A = 1) - \mathrm{E}(Y_i|A = 0)$, is given by

$$g(Y_{i,1}; \beta) = (1, A_i)^T\{Y_{i,1} - (\beta_0 + \beta_1 A_i)\}$$

and the difference in sample means is the solution for β_1. To gain efficiency on $\hat{\beta}$, we need to specify an over-identified set of estimating equations for the secondary endpoint: $h(Y_{i,2}; \theta)$. Following the authors' suggestion, we will consider estimating equations with additional constraints for over-identification. For illustrative purposes, we suppose that the second endpoint's first moment is linear in $(1, A_i, X_{i,1}, X_{i,2})$.

First, we consider estimating equations that place constraints on the treatment indicator:

$$h(Y_{i,2}; \theta) = (1, X_{i,1}, X_{i,2}, A_i)^T\{Y_{i,2} - (\theta_0 + \theta_1 X_{i,1} + \theta_2 X_{i,2})\}.$$

This can be thought of as a linear model for $Y_{i,2}$ as a function of $(1, X_{i,1}, X_{i,2})$ which is fit via empirical likelihood with the additional constraint that $\sum_{i=1}^{n} \hat{p}_i\, A_i\{Y_{i,2} - (\hat{\theta}_0 + \hat{\theta}_1 X_{i,1} + \hat{\theta}_2 X_{i,2})\} = 0$. Recall that these estimating equations must satisfy $\mathrm{E}\{h(Y_{i,2}; \theta_{1*})\} = 0$ for some θ_{1*}. For this equality to hold, it must be assumed that A_i is orthogonal to $Y_{i,2} - \mathrm{E}(Y_{i,2}|X_{i,1}, X_{i,2})$. However, this orthogonality cannot hold if there is a treatment effect on $Y_{i,2}$. Indeed, in our preliminary simulations, the iterative estimation algorithm for $\hat{p_i}$ and $\hat{\theta}$ was unable to converge upon a global maximizer for $\prod_{i=1}^{n} p_i$ with these constraints, further supporting the infeasibility of these estimating equations for the secondary endpoint. Thus, we cannot include A_i as a

"redundant" covariate to gain efficiency and must strategically add constraints based on other covariates, which may be conditionally independent of the secondary endpoint, as we do in the following simulation study.

4. Simulation Study and Results

We consider data resembling a moderate-sized RCT with 500 participants to detect a hypothesized standardized treatment effect of 0.25 on the primary endpoint with 80% power. There are two baseline covariates, $(X_{i,1}, X_{i,2})$, which are independent of treatment and follow a standard bivariate normal distribution with correlation 0.7. The primary endpoint has mean $\mathrm{E}(Y_{i,1}|A_i, X_{i,1}, X_{i,2}) = 0.25A_i + 0.5X_{i,1}$ and the secondary endpoint has mean $\mathrm{E}(Y_{i,2}|A_i, X_{i,1}, X_{i,2}) = 0.5A_i + X_{i,2}$, where, conditional on $(A_i, X_{i,1}, X_{i,2})$, the endpoints follow a bivariate normal distribution with variance 1 and correlation 0.7. Here, the two covariates can be viewed as baseline measurements of the two endpoints.

To gain efficiency from the secondary endpoint, we use the following correctly specified working model: $\mathrm{E}(Y_{i,2}|A_i, X_{i,1}, X_{i,2}) = \theta_0 + \theta_1 A_i + \theta_2 X_{i,2}$ and place an additional orthogonality constraint with respect to $X_{i,1}$, resulting in the estimating equations:

$$h(Y_{i,2}; \theta) = (1, A_i, X_{i,1}, X_{i,2})^T \{Y_{i,2} - (\theta_0 + \theta_1 A_i + \theta_2 X_{i,2})\}.$$

These estimating equations were then used in the constraints given in Equation 2 to find the optimal subject-specific weights $\hat{p}_i$ maximizing $\prod_{i=1}^{n} p_i$. We then considered two sets of estimating equations for the treatment effect on the primary endpoint. The first corresponds to the sample mean difference:

$$g_1(Y_{i,1};\beta) = (1, A_i)^T\{Y_{i,1} - (\beta_0 + \beta_1 A_i)\}.$$

And the second adjusts for the baseline covariates:

$$g_2(Y_{i,1};\beta) = (1, A_i, X_{i,1}, X_{i,2})^T\{Y_{i,1} - (\beta_0 + \beta_1 A_i + \beta_2 X_{i,1} + \beta_3 X_{i,2})\}.$$

In both cases, the treatment effect is given by the parameter β_1. We then obtained the unweighted and weighted solutions to both sets of estimating equations for four total estimates per simulation.

We performed 2000 Monte Carlo simulations. For each simulation, we recorded the parameter estimates for the treatment effect β_1 as well as other nuisance parameters. We summarized the empirical standard error for each parameter under each estimator (noting that all estimators had negligible bias). Our results indicate that this approach offers no efficiency gain for the treatment effect, β_1. However, in the covariate adjusted model, there are notable efficiency gains for the effects of $X_{i,1}$ and $X_{i,2}$. Efficiency was gained on the effect of $X_{i,1}$ through β_2, which was not included in the working model for the secondary endpoint but used to generate an additional constraint and on the effect of $X_{i,2}$ through β_3, which appeared in the working model for the secondary endpoint but is correlated with $X_{i,1}$ (Table 1). Similar findings were observed under simulations with a null treatment effect for the primary and secondary endpoint.

Table 1: *Empirical standard errors estimating model coefficients with and without using data from the secondary endpoint. The estimate for the treatment effect is given via* β_1.

Parameter	Unadjusted		Adjusted	
	Primary Endpoint Only	Secondary Endpoint Added	Primary Endpoint Only	Secondary Endpoint Added
β_0	0.069	0.069	0.062	0.062
β_1	0.098	0.099	0.088	0.088
β_2			0.061	0.044
β_3			0.062	0.055

5. Discussion

Most, if not all trials, collect data on secondary endpoints. However, there has been limited methodological research to leverage information from these endpoints to gain efficiency in RCTs. We attempted to translate an estimator that uses data from secondary endpoints proposed by Chen et al. (2022) to improve treatment effect estimation. Our results indicate that in order to gain efficiency for a specific parameter under the proposed estimator with one secondary endpoint, the covariate associated with that parameter must (a) be omitted from the outcome model for the secondary data and used to generate an additional constraint or (b) be associated with a covariate satisfying (a). When considering a treatment indicator in an RCT, the constraint required in (a) cannot be satisfied so long as there is a treatment effect on the secondary endpoint, and due to randomization, the treatment will be independent of all covariates, thus not satisfying (b). Although this novel estimator demonstrates strong potential when analyzing observational data, other innovative methods are required to borrow efficiency from secondary endpoints within RCTs.

Acknowledgements

This study was funded by the National Institute on Drug Abuse (Award Numbers R01DA046320, and U54DA031659) and the National Center for Advancing Translational Science (Award Number UM1TR004405). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and Food and Drug Administration Center for Tobacco Products.

References

1. Chen C, Han P, He F. Improving main analysis by borrowing information from auxiliary data. Stat Med. 2022;41(3):567–79.

2. Chen C, Wang M, Chen S. An efficient data integration scheme for synthesizing information from multiple secondary datasets for the parameter inference of the main analysis. Biometrics. 2023 Mar 24;79(4):2947–60.

3. Donny EC, Denlinger RL, Tidey JW, Koopmeiners JS, Benowitz NL, Vandrey RG, et al. Randomized trial of reduced-nicotine standards for cigarettes. N Engl J Med. 2015 Oct 1;373(14):1340–9.

4. Hatsukami DK, Luo X, Jensen JA, al'Absi M, Allen SS, Carmella SG, et al. Effect of immediate vs gradual reduction in nicotine content of cigarettes on biomarkers of smoke exposure: a randomized clinical trial. JAMA. 2018 Sep 4;320(9):880–91.

5. Smith TT, Koopmeiners JS, Tessier KM, Davis EM, Conklin CA, Denlinger-Apte RL, et al. Randomized trial of low-nicotine cigarettes and transdermal nicotine. Am J Prev Med. 2019 Oct;57(4):515–24.

6. White CM, Tessier KM, Koopmeiners JS, Denlinger-Apte RL, Cobb CO, Lane T, et al. Preliminary evidence on cigarette nicotine reduction with concurrent access to an e-cigarette: manipulating cigarette nicotine content, e-liquid nicotine content, and e-liquid flavor availability. Prev Med. 2022 Dec 1;165:107213.

7. Hatsukami DK, Jensen JA, Carroll DM, Luo X, Strayer LG, Cao Q, et al. Reduced nicotine in cigarettes in a marketplace with alternative nicotine systems: randomized clinical trial. Lancet Reg Health – Am. 2024 Jul 1;35:100796.

8. Owen AB. Empirical likelihood. Boca Raton, Fla: Chapman & Hall/CRC; 2001. (Monographs on statistics and applied probability; 92).