STATISTICAL THEORY AND RELATED FIELDS
2018, VOL. 2, NO. 1, 21–22
https://doi.org/10.1080/24754269.2018.1493630

Taylor & Francis
Taylor & Francis Group

Discussion of the paper by R. L. Prentice and Y. Huang: Optimal designs and efficient inference for biomarker studies

D. Y. Lin

Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA

ARTICLE HISTORY Received 28 May 2018; Accepted 23 June 2018

It gives me great pleasure to congratulate Drs. Prentice and Huang on an excellent, thought-provoking article on statistical issues and opportunities in nutritional epidemiology research. The authors addressed several important challenges in obtaining reliable information on dietary intake, such as random and systematic biases in self-reported dietary data and high cost of biomarker measurements. They also suggested ways to develop additional biomarkers and to improve statistical strategies. In this brief commentary, I will focus on a couple of questions posed in Section 3 of the article, namely, how to make efficient statistical inference when expensive biomarkers are measured only on a subset of cohort members and how to optimally select cohort members for biomarker measurements.

Let T denote the failure time, $\mathbf{X}$ denote the set of expensive biomarkers, $\mathbf{Z}$ denote the set of inexpensive covariates that is potentially correlated with $\mathbf{X}$ and $\mathbf{W}$ denote the set of inexpensive covariates that is known to be independent of $\mathbf{X}$. We specify that the hazard function of T conditional on $\mathbf{X} = \mathbf{x}$, $\mathbf{Z} = \mathbf{z}$ and $\mathbf{W} = \mathbf{w}$ satisfies the proportional hazards model (Cox, 1972)

$$\lambda(t|\mathbf{x},\mathbf{z},\mathbf{w}) = \lambda_0(t)\, e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}+\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}+\boldsymbol{\eta}^{\mathrm{T}}\mathbf{w}}, \tag{1}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function, and $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ are unknown regression parameters.

The failure time T is subject to right censoring by C, such that we observe $\widetilde{T}$ and Δ instead of T, where $\widetilde{T} = \min(T, C)$, $\Delta = I(T \le C)$, and $I(\cdot)$ is the indicator function. Let $\mathcal{S}$ denote the set of cohort members who are selected for measurements of $\mathbf{X}$, and $\overline{\mathcal{S}}$ denote the complement of $\mathcal{S}$. The selection can depend on $(\widetilde{T}, \Delta, \mathbf{Z}, \mathbf{W})$ in any manner.

Write $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\eta}^{\mathrm{T}})^{\mathrm{T}}$ and $\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds$, and let $P(\cdot|\cdot)$ denote a conditional density function. For a subject in $\mathcal{S}$, the likelihood contribution is the density of $(\widetilde{T}, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W})$; for a subject in $\overline{\mathcal{S}}$, the likelihood contribution is the density of $(\widetilde{T}, \Delta, \mathbf{Z}, \mathbf{W})$. Thus, the log-likelihood function concerning $\boldsymbol{\theta}$, Λ_0 and $P(\mathbf{x}|\mathbf{z})$ takes the form

$$\sum_{i\in\mathcal{S}} \{\log P(\widetilde{T}_i, \Delta_i|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i) + \log P(\mathbf{X}_i|\mathbf{Z}_i)\} + \sum_{i\in\overline{\mathcal{S}}} \log \int P(\widetilde{T}_i, \Delta_i|\mathbf{x}, \mathbf{Z}_i, \mathbf{W}_i) P(\mathbf{x}|\mathbf{Z}_i)\, d\mathbf{x}.$$

Under the assumption that C_i is independent of T_i conditional on $(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i)$,

$$P(\widetilde{T}_i, \Delta_i|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i) \propto \left[\lambda_0(\widetilde{T}_i)\, e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i+\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i+\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_i}\right]^{\Delta_i} \exp\left[-\int_0^{\widetilde{T}_i} e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i+\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i+\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_i}\lambda_0(t)\, dt\right]$$

(Kalbfleisch & Prentice, 2002, p. 54). If C_i is independent of T_i and $\mathbf{X}_i$ conditional on $(\mathbf{Z}_i, \mathbf{W}_i)$, then $P(\widetilde{T}_i, \Delta_i|\mathbf{X} = \mathbf{x}, \mathbf{Z}_i, \mathbf{W}_i)$ is the product of

$$\left[\lambda_0(\widetilde{T}_i)\, e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}+\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i+\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_i}\right]^{\Delta_i} \exp\left[-\int_0^{\widetilde{T}_i} e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}+\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i+\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_i}\lambda_0(t)\, dt\right]$$

and a function that does not involve $\mathbf{x}$ or $(\boldsymbol{\theta}, \Lambda_0)$ and thus can be factored out of the integral in the second term of the log-likelihood function.

We adopt nonparametric maximum likelihood estimation, under which both Λ_0 and $P(\mathbf{x}|\mathbf{z})$ are nonparametric. The estimation can be carried out through EM algorithms. The resulting estimators of $\boldsymbol{\theta}$ and Λ_0 are consistent and asymptotically normal. In addition, the estimator of $\boldsymbol{\theta}$ achieves the semiparametric efficiency bound. The interested readers are referred to Zeng and Lin (2014,1) for details.

Drs. Prentice and Huang described case–cohort and nested case–control designs, which assume that all cases are selected. In large cohorts with relatively common diseases, it may not be economically feasible to measure biomarkers on all cases. Indeed, it is unclear whether or not cases should take

CONTACT D. Y. Lin lin@bios.unc.edu Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-7420, USA

precedence over controls or which cases and controls are the most informative. Lawless (2018) suggested to select the cases with the smallest failure times and the controls with the largest censoring times. In addition, Borgan, Langholz, Samuelsen, Goldstein and Pagoda (2000) stratified the selection of the subcohort in the case–cohort design on inexpensive covariates, and Langholz and Borgan (1995) used inexpensive covariates to select 'counter-matched' controls at the failure time of each case.

In recent unpublished work, my colleagues Drs. Ran Tao and Donglin Zeng and I investigated the efficiency of such sampling designs. The design efficiency pertains to the semiparametric efficiency bound for estimating the regression coefficients of expensive covariates. We developed optimal designs that are the most efficient among all possible sampling designs. We found that the design suggested by Lawless (2018) is optimal if there are no inexpensive covariates. In the presence of inexpensive covariates, a design that selects an equal number of subjects at the two extreme tails of martingale residuals in each stratum of inexpensive covariates is optimal and can be substantially more efficient than the existing sampling designs.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by National Institutes of Health [R01GM047845, R01HG009974, and P01CA142538].

Notes on contributor

D. Y. Lin is the Dennis Gillings Distinguished Professor of Biostatistics at the University of North Carolina at Chapel Hill. He is an internationally recognized expert on survival analysis, with many influential publications. One of this current research interests is efficient designs and analysis of two-phase studies. He is a fellow of IMS and ASA and an Associate Editor for Biometrika and JASA.

References

Borgan, Ø., Langholz, B., Samuelsen, S. O., Goldstein, L., & Pagoda, J. (2000). Exposure stratified case–cohort designs. *Lifetime Data Analysis*, *6*, 39–58.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, *34*, 187–220.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). Hoboken: Wiley.

Langholz, B., & Borgan, Ø. (1995). Counter-matching: A stratified nested case–control sampling method. *Biometrika*, *82*, 69–79.

Lawless, J. F. (2018). Two-phase outcome-dependent studies for failure times and testing for effects of expensive covariates. *Liftetime Data Analysis*, *24*, 28–44.

Zeng, D., & Lin, D. Y. (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *Journal of the American Statistical Association*, *109*, 371–383.

Zeng, D., & Lin, D. Y. (2018). Maximum likelihood estimation for case–cohort and nested case–control studies. In Ø. Borgan, N. Breslow, N. Chatterjee, M. Gail, & A. Scott (Eds.), *Handbook of statistical methods for case–control studies* (Chapter 24). New York: Chapman and Hall Press.