ARIC Manuscript Proposal #4076

PC Reviewed: 7/12/22 **Status: _____** **Priority: 2**
SC Reviewed: _____ **Status: _____** **Priority: ____**

1.a. Full Title: Inverse Probability Censoring Weighted Super Learner for Survival Prediction in Case-Cohort Studies

b. Abbreviated Title (Length 26 characters): **Case-Cohort Studies**

2. Writing Group:
Writing group members: Haolin Li, Haibo Zhou, David Couper, Jianwen Cai

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. _____ **[please confirm with your initials electronically or in writing]**

First author: Haolin Li
Address: Department of Biostatistics, CB#7420
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7420

Phone: (919)450-8771 Fax:
E-mail: haolin@live.unc.edu

ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).
Name: David Couper
Address: 123 W. Franklin Street, Suite 450, CB #8030
Chapel Hill, NC 27516
USA

Phone: (919) 962-3229 Fax:
E-mail: david_couper@unc.edu

3. Timeline: Work will begin as soon as approval is received.

4. Rationale:

In epidemiologic cohort studies and disease prevention trials, it is often of interest to investigate the association between various risk factors and time-to-event outcomes. However, when only

a very small fraction of the full cohort develops diseases in the given study time frame and the exposure measurement of interest is expensive to obtain, it can be very costly to obtain the exposure measurement for the entire cohort. The case-cohort sampling scheme (Prentice, 1986) is one of several designs that have been proposed to reduce the cost of such type of study. The key principle of this design is to obtain covariate information on only a random subset of the cohort and all other subjects in the cohort who have developed the event of interest. This leads to the case-cohort studies being considered an appealing alternative to full-cohort studies for large-scale epidemiological studies where the event is rare or the covariates are expensive to measure. Since the proposal of the case-cohort study design, most of the method development has been focused on the statistical inference using a particular model (Prentice, 1986; Self and Prentice, 1988; Borgan, Goldstein, and Langholz, 1995; Lin and Ying, 1993; and Barlow, 1994), but there is very limited literature focusing on survival prediction using the data from the case-cohort study.

The super learner approach (Breiman, 1996; LeBlanc and Tibshirani, 1996; and Van der Laan, Polley, and Hubbard, 2007) is a very flexible statistical learning method that finds the best weighted ensemble of a set of individual candidate prediction rules by minimizing the cross-validated risk, and it has been shown to have superior performances of prediction in various settings of regression and classification. There have also been some super learner algorithms that have been proposed in recent years that can be used for survival prediction in right-censored settings (Van der Laan and Rose, 2011; Wey, Connett, and Rudser, 2015; and Golmakani and Polley, 2020) and in competing risk setting (Sachs et. al., 2019; and Gonzalez Ginestet et. al., 2021). However, it is still unclear that, (1) how a super learner algorithm can be formulated when the training data come from a case-cohort design; and (2) whether the case-cohort study design is cost-effective in survival prediction when a super learner is used (i.e., does a super learner trained by data from case-cohort design have better prediction accuracy than the one trained by data from a cohort study of the same size).

In this statistical methodology paper, we propose a super learner algorithm based on Brier Scores (Brier, 1950) for case-cohort studies. It can be used for survival prediction with right-censored data as well as in competing risk settings. It can also be easily extended to stratified case-cohort studies (Hu, Cai, and Zeng, 2014) and generalized case-cohort studies (Chen, 2001; and Cai and Zeng, 2007). We proved that the proposed super learner algorithm has asymptotic model selection consistency and uniform consistency, and we also conducted extensive simulation studies to show that the super learner trained by data from case-cohort design has better prediction accuracy than the one trained by data from a simple random sample with the same sample size as the case-cohort sample. As part of this research, we will also apply our method to a real case-cohort data set to illustrate its use. This is the reason we are requesting data from the ARIC study. These data will be published only as part of a methodological paper with acknowledgement of the ARIC study.

Reference

1. Barlow, W. E. (1994). Robust variance estimation for the case-cohort design. *Biometrics*, 1064-1072.
2. Borgan, O., Goldstein, L., & Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *The Annals of Statistics*, 1749-1778.

3. Breiman, L. (1996). Stacked regressions. *Machine learning*, *24*(1), 49-64.
4. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1-3.
5. Cai, J., & Zeng, D. (2007). Power calculation for case–cohort studies with nonrare events. *Biometrics*, *63*(4), 1288-1295.
6. Chen, K. (2001). Generalized case–cohort sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(4), 791-809.
7. Golmakani, M. K., & Polley, E. C. (2020). Super learner for survival data prediction. *The International Journal of Biostatistics*, *16*(2).
8. Gonzalez Ginestet, P., Kotalik, A., Vock, D. M., Wolfson, J., & Gabriel, E. E. (2021). Stacked inverse probability of censoring weighted bagging: A case study in the InfCareHIV Register. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *70*(1), 51-65.
9. Hu, W., Cai, J., & Zeng, D. (2014). Sample size/power calculation for stratified case–cohort design. *Statistics in medicine*, *33*(23), 3973-3985.
10. LeBlanc, M., & Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association*, *91*(436), 1641-1650.
11. Lin, D. Y., & Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, *88*(424), 1341-1349.
12. Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, *73*(1), 1-11.
13. Sachs, M. C., Discacciati, A., Everhov, Å. H., Olén, O., & Gabriel, E. E. (2019). Ensemble prediction of time-to-event outcomes with competing risks: a case-study of surgical complications in Crohn's disease. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *68*(5), 1431-1446.
14. Self, S. G., & Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*, 64-81.
15. Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, *6*(1).
16. Van der Laan, M. J., & Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data* (Vol. 4). New York: Springer.
17. Wey, A., Connett, J., & Rudser, K. (2015). Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. *Biostatistics*, *16*(3), 537-549.

5. Main Hypothesis/Study Questions:

We would like to use the proposed super learner algorithm to construct a prediction rule for incident diabetes using the data collected under the case-cohort design.

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present).

We will apply the proposed super learner algorithm (to be more specific, to find an optimal ensemble of weighted Weibull proportional hazards model, weighted gamma accelerated failure time model, weighted Cox model, and weighted survival random forest) to the case-cohort data set that was created for the following two manuscripts to construct a survival prediction rule for incident diabetes:

- MS#1571: Luft, V. C., Schmidt, M. I., Pankow, J. S., Couper, D., Ballantyne, C. M., Young, J. H., & Duncan, B. B. (2013). Chronic inflammation role in the obesity-diabetes association: a case-cohort study. Diabetology & metabolic syndrome, 5(1), 1-8.
  MS#2330: Luft, V. C., Duncan, B. B., Schmidt, M. I., Chambless, L. E., Pankow, J. S., Hoogeveen, R. C., ... & Heiss, G. (2016). Carboxymethyl lysine, an advanced glycation end product, and incident diabetes: a case–cohort analysis of the ARIC Study. Diabetic Medicine, 33(10), 1392-1398.

7.a. Will the data be used for non-ARIC analysis or by a for-profit organization in this manuscript? ____ Yes __X__ No

b. If Yes, is the author aware that the current derived consent file ICTDER05 must be used to exclude persons with a value RES_OTH and/or RES_DNA = "ARIC only" and/or "Not for Profit" ? ____ Yes ____ No
(The file ICTDER has been distributed to ARIC PIs, and contains the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript? ____ Yes __X__ No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the current derived consent file ICTDER05 must be used to exclude those with value RES_DNA = "No use/storage DNA"? ____ Yes ____ No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status.
ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: http://www.cscc.unc.edu/aricproposals/dtSearch.html

___X___ Yes _____ No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?

MS1571 and MS2330

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? __X__ Yes ____ No

11.b. If yes, is the proposal

___ A. primarily the result of an ancillary study (list number* _AS195.09____)

___ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* _____ _____ _____)

*ancillary studies are listed by number https://sites.cscc.unc.edu/aric/approved-ancillary-studies

12a. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is **your responsibility to upload manuscripts to PubMed Central** whenever the journal does not and be in compliance with this policy. Four files about the public access policy from http://publicaccess.nih.gov/ are posted in http://www.cscc.unc.edu/aric/index.php, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to PubMed central.