Appendix for 'BoXHED: Boosted eXact Hazard Estimator with Dynamic covariates'

1. Cohort Selection

The data for our study is obtained from pooling together longitudinal records from two prospective cohorts: The Framingham Heart Study original cohort (FHS) (Dawber et al., 1951) and the Framingham Heart Study Offspring Cohort (FHS-OS) (Dawber et al., 1951). FHS-OS consists of the offspring of the FHS cohort. Time-varying risk factors are obtained from the physical exam results performed during (irregular) follow-up visits. The FHS cohort had physical exams approximately every two years, and for the FHS-OS cohort it was approximately every seven years. The event of interest is the first occurrence of any cardiovascular disease (CVD) event or diagnosis (fatal or non-fatal).

A total of 9,697 participants with 73,340 physical exam records are included in the analyses along with eight risk factors that were consistently collected across all exams and are commonly used in medical models: Age, gender, systolic blood pressure (SBP), diastolic blood pressure (DBP), smoking status, diabetes, total cholesterol (TC), and body mass index (BMI). Medical records are included if they are measured before the first occurrence of CVD, are not missing information on the risk factors, and do not have biologically implausible risk factor values. Implausible values include total cholesterol < 1.75 mmol/L or > 20 mmol/L, SBP < 70 mmHg or > 270 mmHg, and BMI > 80 kg/m^2 (Hajifathalian et al., 2015). Participants are included if they do not have a history of coronary heart disease or stroke at the time of study enrollment, and had at least one valid physical exam record before the first occurrence of CVD. A flowchart of cohort selection is shown in Figure 1.

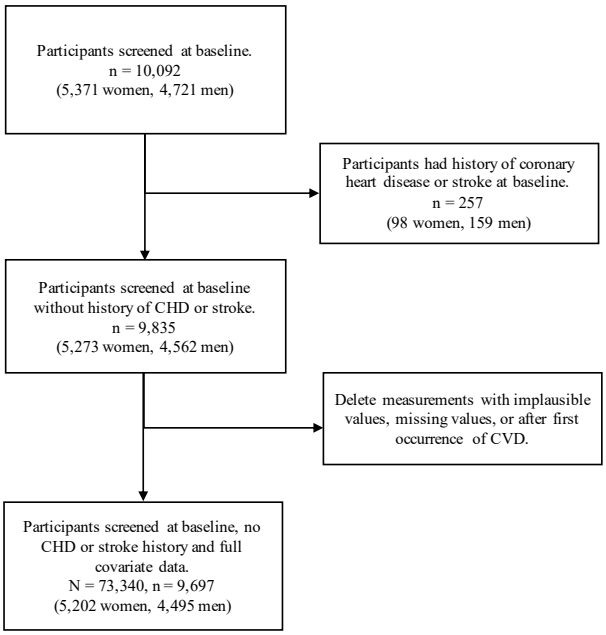


Figure 1. Flowchart for study participant selection in the Framingham study.

2. K-means Clustering

We vary the number of clusters K from 1 to 10, and compute the total within-sum of squares for each K. We identify the optimal number of clusters by the kink in the curve (Hastie et al., 2009). As seen in Figure 2, the optimal K is 4.

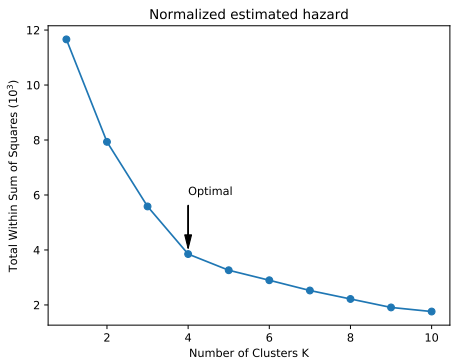


Figure 2. Total within sum of squares by number of clusters.

3. Baseline Comparisons

Details on the baseline comparison techniques and how they were tuned are presented here.

Kernel smoothing estimators. These are nonparametric and can handle time-dependent covariates (Nielsen and Linton, 1995). Tuning the kernel bandwidths for each covariate is

Table 1. The baseline characteristics of the Framingham data by cohort. Means and standard deviations for continuous risk factors, and percentages for categorical risk factors are reported.

	Overall	FHS	FHS-OS
# samples	9697 (100%)	4849 (50%)	4848 (50%)
Age, yr	39.9 (10.3)	43.9 (8.5)	36 (10.3)
SBP ($mmHg$)	128.2 (19.9)	135.1 (21.2)	121.2 (15.8)
Current smoker	6276 (65%)	3087 (64%)	3189 (66%)
Women	5202 (54%)	2683 (55%)	2519 (52%)
Diabetes	128 (1.3%)	46 (0.9%)	82 (1.7%)
DBP ($mmHg$)	81.3 (11.5)	84.3 (11.7)	78.3 (10.4)
Total Cholesterol ($mmol/L$)	5.3 (1.1)	5.7 (1.2)	5.1 (1.0)
BMI (kg/m^2)	25.4 (4.2)	25.5 (4.2)	25.3 (4.3)

computationally taxing, therefore we normalize the variances of the covariates in order to use the same bandwidth for all of them.[1] A grid search over $\{0.1, 0.3, \ldots, 2.0\}$ is performed *directly on the test data* to identify the optimal bandwidth that attains the best out-of-sample performance.

Parametric hazard estimators for time-dependent covariates. The `flexsurv` package in R allows us to fit survival models from eight parametric families: generalized Gamma, generalized F, Weibull, Gamma, exponential, log-logistic, log-normal, and Gompertz. The family that yields the best out-of-sample performance is chosen.

Boosted parametric hazard estimators for time-static covariates. The `blackboost` function in R performs tree-boosted survival estimation for the Weibull, log-normal, and log-logistic parametric families. This method only allows for time-static covariates, so for this we fix each study participant's risk factors at their first recorded values. The default maximum number of tree splits (two) is used, which coincides with the one chosen for `BoXHED` during cross-validation. The combination of parametric family and number of trees that yield the best out-of-sample performance is chosen.

4. AUC_t for the simulated data

See Figures 3, 4, 5, and 6.

5. Baseline characteristics of the Framingham data

See Table 1.

[1] All categorical variables are binary (smoking status, gender, history of diabetes) and are normalized in the same way as the continuous variables.

6. SBP×Gender Odds Ratio Analysis

Table 2. ORs with 95% confidence intervals for exploring the impact of the SBP×GENDER interaction on CVD hazard.

(a) Male

SBP \ DBP	<70	70-79	80-84	85-89	>90
<115	1.0	0.9 ± 0.2			
115-124	1.3 ± 0.3	1.2 ± 0.3	1.0 ± 0.2		
125-139	1.6 ± 0.3	1.3 ± 0.2	1.2 ± 0.3	1.7 ± 0.4	1.6 ± 0.4
140-149	1.7 ± 0.5	1.7 ± 0.4	1.5 ± 0.4	1.7 ± 0.4	2.0 ± 0.4
>150	2.5 ± 0.7	2.2 ± 0.5	2.2 ± 0.5	1.9 ± 0.4	2.4 ± 0.4

(b) Female

SBP \ DBP	<70	70-79	80-84	85-89	>90
<115	0.6 ± 0.1	0.5 ± 0.1			
115-124	0.9 ± 0.2	0.7 ± 0.1	0.7 ± 0.2		
125-139	1.1 ± 0.2	0.9 ± 0.2	0.8 ± 0.2	1.1 ± 0.2	1.1 ± 0.3
140-149	1.2 ± 0.4	1.2 ± 0.3	1.0 ± 0.2	1.2 ± 0.3	1.4 ± 0.3
>150	1.7 ± 0.4	1.6 ± 0.4	1.5 ± 0.4	1.4 ± 0.3	1.7 ± 0.4

Table 2 shows the odds ratios (ORs) stratified by gender. Cells containing < 10 events are left blank since we do not have enough data points to infer an odds ratio. For the table for females (Table 2b), reading down a given column reveals an increasing relationship between OR and SBP for a given level of DBP. The same relationship holds in the table for males as well (Table 2a). This suggests that SBP×GENDER is not responsible for the differences in the qualitative relationship between blood pressure and CVD risk among different patient cohorts.

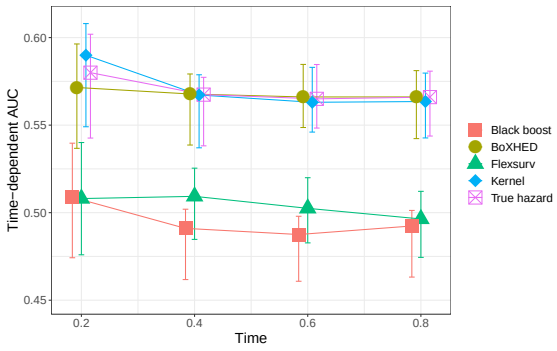
References

T.R. Dawber, G.F. Meadors, and F.E. Moore Jr. Epidemiological approaches to heart disease: the framingham study. *American Journal of Public Health and the Nations Health*, 41(3):279–286, 1951.

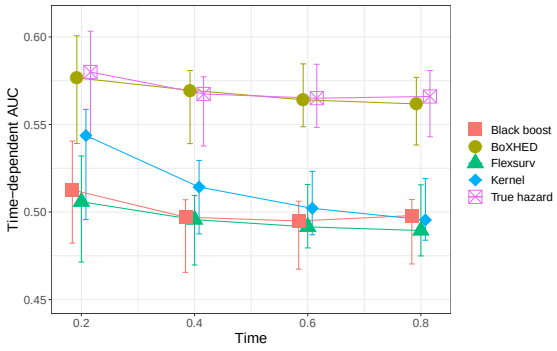
K. Hajifathalian, P. Ueda, Y. Lu, et al. A novel risk score to predict cardiovascular disease risk in national populations (globorisk): a pooled analysis of prospective cohorts and health examination surveys. *The Lancet Diabetes & Endocrinology*, 3(5):339–355, 2015.

T.J. Hastie, R. Tibshirani, and J.H. Friedman. *The elements of statistical learning (2nd edition)*. Springer NY, 2009.

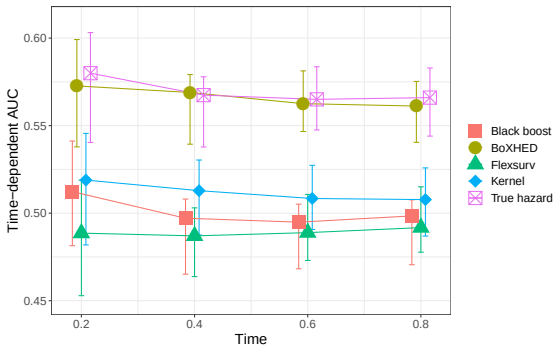
J.P. Nielsen and O.B. Linton. Kernel estimation in a nonparametric marker dependent hazard model. *Annals of Statistics*, 23(5):1735–1748, 1995.



(a)

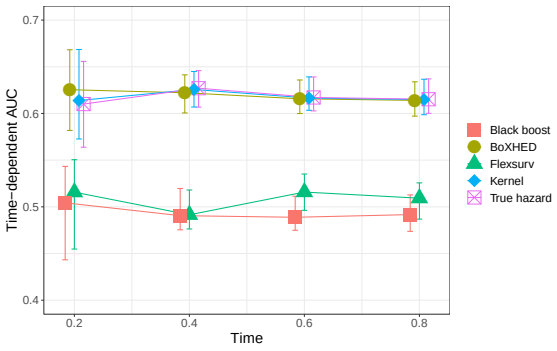


(b)

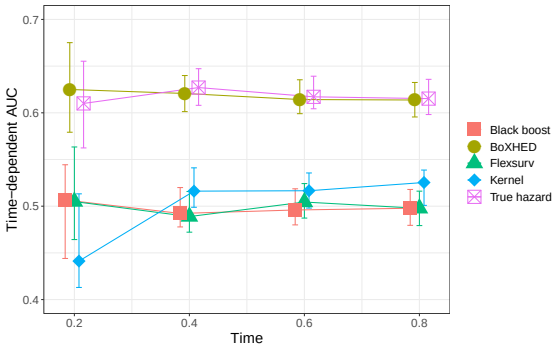


(c)

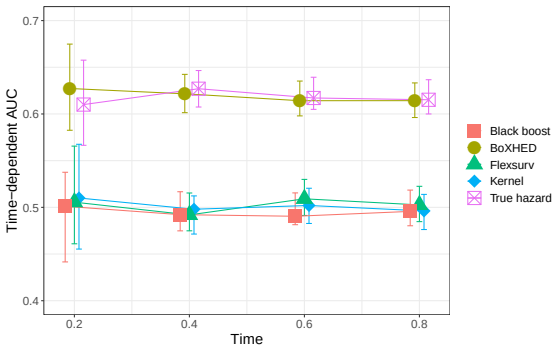
Figure 3. AUC_t versus time t for the estimators when applied to data simulated from λ_1. Larger AUC_t values are better. (a) No irrelevant covariates; (b) 20 irrelevant covariates; (c) 40 irrelevant covariates



(a)

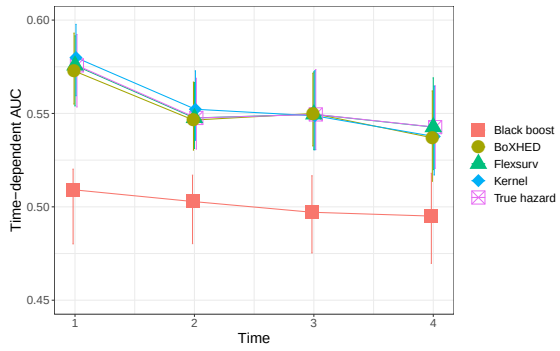


(b)

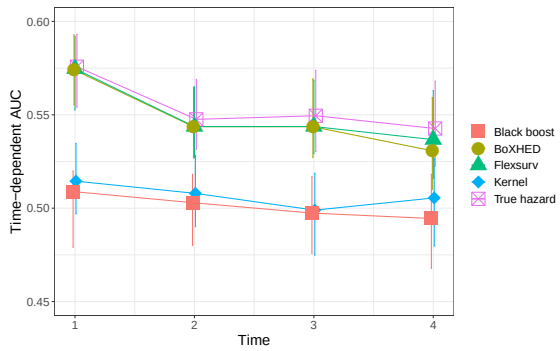


(c)

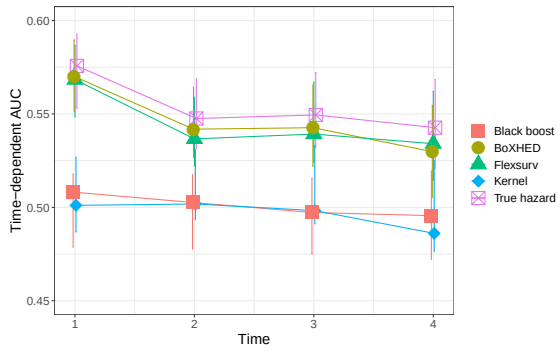
Figure 4. AUC_t versus time t for the estimators when applied to data simulated from λ_2. Larger AUC_t values are better. (a) No irrelevant covariates; (b) 20 irrelevant covariates; (c) 40 irrelevant covariates



(a)

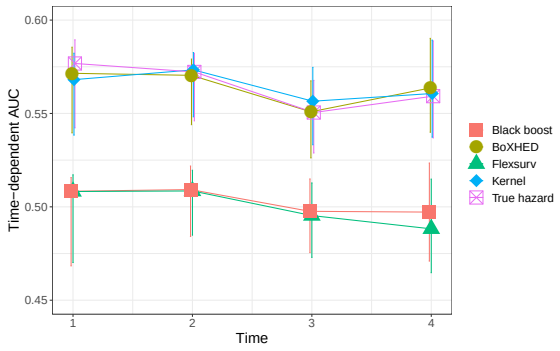


(b)

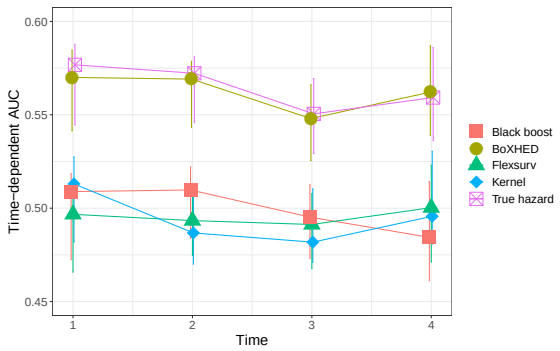


(c)

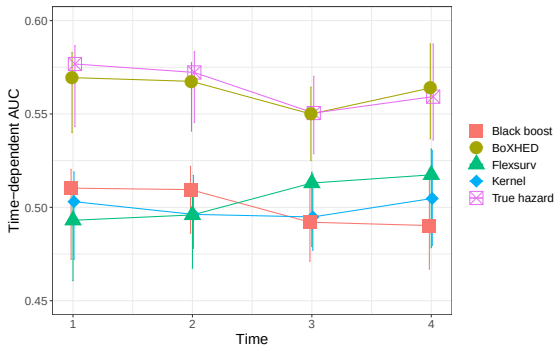
Figure 5. AUC_t versus time t for the estimators when applied to data simulated from λ_3. Larger AUC_t values are better. (a) No irrelevant covariates; (b) 20 irrelevant covariates; (c) 40 irrelevant covariates



(a)



(b)



(c)

Figure 6. AUC_t versus time t for the estimators when applied to data simulated from λ_4. Larger AUC_t values are better. (a) No irrelevant covariates; (b) 20 irrelevant covariates; (c) 40 irrelevant covariates