UAW-GM Cohort Study

Predicting survival to 1985

July 30, 2020

Population

- Restricted to those:
  - Still alive in 1941
  - Hired in or after 1938, but no later than 1982
  - Missing no more than half of their work record
- Individuals contributed person-time from three years after hire or 1941 (whichever came first) to death or loss to follow-up
- Individuals were considered lost to follow-up upon reaching the oldest observed age at death (106.56 years)
- N = 36 549, (698 394 person-years)
- Deaths due to natural causes by end of 1984: 4 405 (11.4%)

ICD codes for natural causes of death

- ICD-9: all codes codes in [001, 799]
  - Excludes the categories labeled as "Injury and poisoning" and "external causes of injury and supplemental classification."
- ICD-10: all codes, except those with prefix S, T, V, W, X, or Y.

Estimating survival in each year using Super Learner

- We use the Super Leaner `tlverse/sl3` to estimate the probability dying due to natural causes by the end of each year of follow-up, conditional on covariates and survival prior to that year
- The Super Learner library included:
  - Covariate combination-specific mean `mean`
  - Pooled logistic regression `glm`
  - PLR with reglarization by lasso or elasticnet `glmnet`
  - Generalized additive model `gam`
  - Random forests `ranger`
  - Extreme gradient boosting: additive decision tree ensembles `xgboost`

Covariates

- Duration of employment
- Calendar year
- Age
- Race
- Plant
- Sex
- Cumulative time spent off
- Year of hire
- Cumulative exposure to straight, soluble, and synthetic MWFs
- Employment status

Calculating cumulative survival

1. Extract the fitted probabilities $\hat{p}_{ti}$ from the learner(s)
2. For each individual, with obserations ordered by time t, take the cumulative product $\prod_t (1 - \hat{p}_{ti})$
3. The cumulative product in each row represents the probability of survival (for natural cause mortality) to the end of that row's year

Average survival probabilities by age group

Age	sl	mean	glm	glmnet	gam	ranger	xgboost
[16.24, 55]	0.97	0.92	0.98	0.98	0.98	0.98	0.97
(55, 70]	0.82	0.84	0.83	0.83	0.83	0.83	0.81
(70, 106.6]	0.59	0.82	0.59	0.59	0.59	0.62	0.58

Average survival probabilities by age group (deciles)

Age	sl	mean	glm	glmnet	gam	ranger	xgboost
[16.24, 44.71]	0.99	0.94	0.99	0.99	0.99	0.99	0.98
(44.71, 50.71]	0.96	0.89	0.97	0.97	0.97	0.97	0.94
(50.71, 54.64]	0.93	0.87	0.94	0.94	0.94	0.95	0.90
(54.64, 57.76]	0.89	0.85	0.90	0.90	0.90	0.92	0.86
(57.76, 60.97]	0.86	0.84	0.87	0.86	0.87	0.90	0.83
(60.97, 64.04]	0.82	0.83	0.83	0.83	0.83	0.88	0.80
(64.04, 67.27]	0.79	0.83	0.79	0.79	0.79	0.85	0.77
(67.27, 70.73]	0.73	0.82	0.73	0.73	0.73	0.81	0.73
(70.73, 76.08]	0.65	0.82	0.65	0.65	0.65	0.76	0.66
(76.08, 106.6]	0.43	0.81	0.42	0.42	0.42	0.62	0.46

Average survival probabilities by other covariates

Race	sl	mean	glm	glmnet	gam	ranger	xgboost
White	0.91	0.89	0.91	0.91	0.91	0.92	0.90
Black	0.92	0.91	0.93	0.93	0.93	0.93	0.92

Sex	sl	mean	glm	glmnet	gam	ranger	xgboost
Male	0.90	0.89	0.91	0.91	0.91	0.91	0.90
Female	0.95	0.92	0.96	0.96	0.96	0.96	0.94

Year of hire	sl	mean	glm	glmnet	gam	ranger	xgboost
[1938, 1945]	0.68	0.78	0.69	0.69	0.69	0.71	0.67
(1945, 1960]	0.80	0.83	0.81	0.81	0.81	0.82	0.79
(1960, 1982]	0.98	0.93	0.98	0.98	0.98	0.98	0.97

Employment status	sl	mean	glm	glmnet	gam	ranger	xgboost
At work	0.98	0.93	0.98	0.98	0.98	0.98	0.97
Left work	0.82	0.85	0.82	0.82	0.82	0.83	0.80

Average survival probabilities by other covariates (deciles)

Year of hire	sl	mean	glm	glmnet	gam	ranger	xgboost
[1938, 1942]	0.69	0.78	0.70	0.69	0.70	0.80	0.67
(1942, 1945]	0.71	0.79	0.71	0.71	0.71	0.80	0.69
(1945, 1946]	0.69	0.80	0.70	0.69	0.70	0.78	0.68
(1946, 1948]	0.73	0.81	0.73	0.73	0.73	0.81	0.71
(1948, 1950]	0.76	0.82	0.77	0.77	0.77	0.83	0.74
(1950, 1952]	0.79	0.82	0.79	0.79	0.79	0.85	0.77
(1952, 1953]	0.83	0.83	0.84	0.84	0.84	0.88	0.80
(1953, 1954]	0.84	0.83	0.85	0.85	0.85	0.89	0.81
(1954, 1960]	0.88	0.85	0.89	0.89	0.89	0.92	0.86
(1960, 1982]	0.98	0.93	0.98	0.98	0.98	0.99	0.97

Average survival probabilities by MWF exposure

Cumulative exposure	sl	mean	glm	glmnet	gam	ranger	xgboost
Straight							
0	0.92	0.90	0.93	0.93	0.93	0.93	0.91
(0, 0.7836]	0.92	0.91	0.93	0.93	0.93	0.93	0.92
(0.7836, 1.337]	0.89	0.88	0.89	0.90	0.89	0.90	0.88
(1.337, 293.4]	0.88	0.87	0.89	0.89	0.89	0.89	0.87
Soluble							
0	0.94	0.92	0.95	0.95	0.95	0.95	0.93
(0, 14.6]	0.93	0.91	0.93	0.93	0.93	0.93	0.92
(14.6, 19.02]	0.83	0.85	0.84	0.84	0.84	0.85	0.82
(19.02, 240.8]	0.82	0.83	0.83	0.83	0.83	0.83	0.81
Synthetic							
0	0.91	0.89	0.92	0.92	0.92	0.92	0.90
(0, 0.1692]	0.94	0.93	0.95	0.95	0.95	0.95	0.94
(0.1692, 105]	0.90	0.89	0.91	0.91	0.91	0.91	0.89

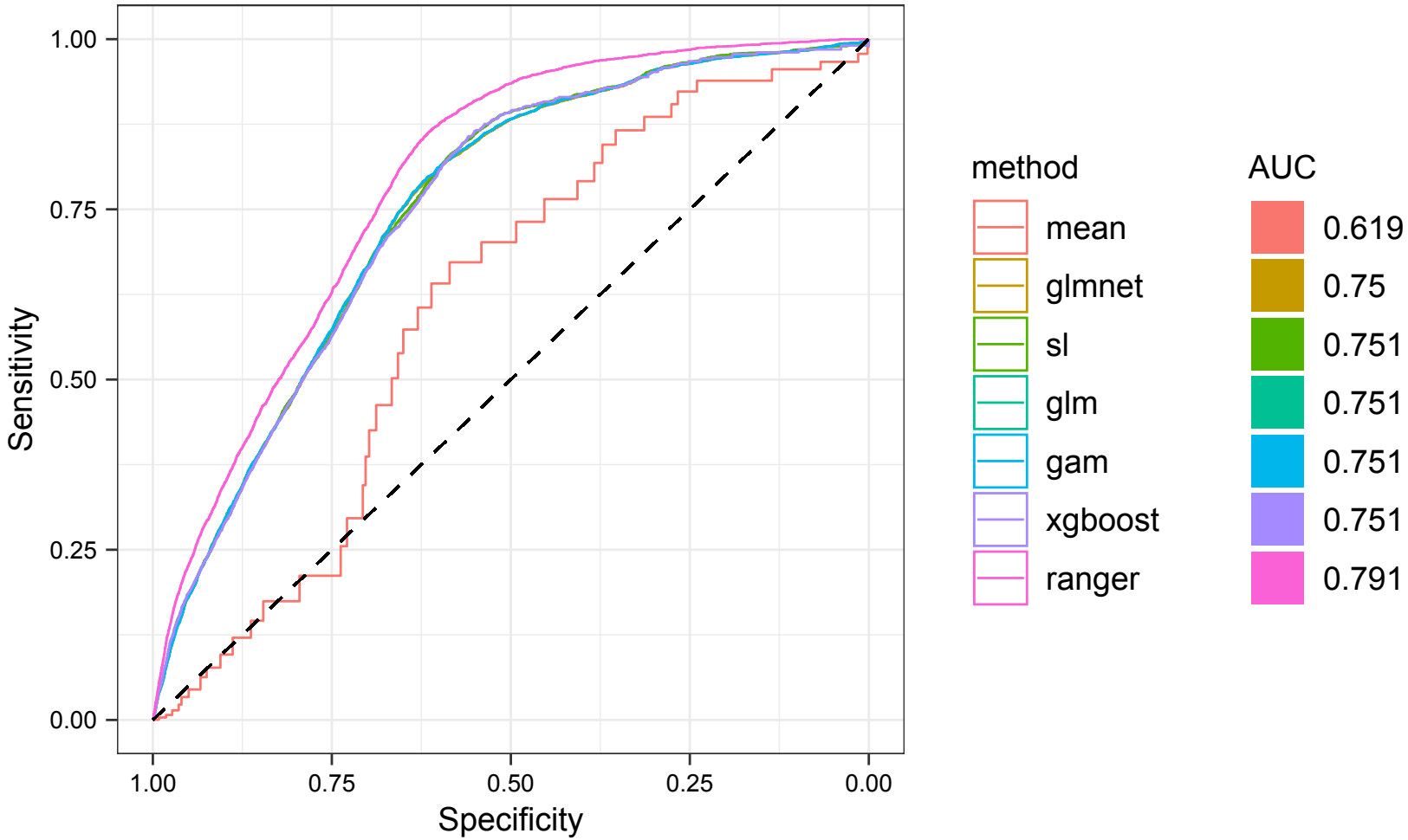
Average survival probabilities by MWF exposure (deciles, as is possible)

Cumulative exposure	sl	mean	glm	glmnet	gam	ranger	xgboost
Straight							
0	0.93	0.90	0.93	0.93	0.93	0.95	0.91
(0, 0.1412]	0.93	0.91	0.93	0.93	0.93	0.95	0.91
(0.1412, 0.432]	0.94	0.91	0.94	0.94	0.94	0.96	0.92
(0.432, 0.968]	0.92	0.90	0.92	0.92	0.92	0.94	0.90
(0.968, 1.844]	0.89	0.88	0.89	0.89	0.89	0.92	0.87
(1.844, 5.676]	0.90	0.87	0.90	0.90	0.90	0.93	0.88
(5.676, 293.4]	0.88	0.85	0.89	0.89	0.89	0.91	0.86
Soluble							
0	0.95	0.92	0.95	0.95	0.95	0.97	0.93
(0, 0.1795]	0.96	0.93	0.97	0.97	0.97	0.97	0.95
(0.1795, 2.047]	0.96	0.93	0.97	0.97	0.97	0.98	0.95
(2.047, 3.872]	0.94	0.91	0.95	0.94	0.95	0.96	0.92
(3.872, 6.208]	0.91	0.88	0.92	0.91	0.92	0.94	0.88
(6.208, 8.7]	0.89	0.87	0.90	0.89	0.90	0.93	0.87
(8.7, 11.79]	0.87	0.87	0.87	0.87	0.87	0.91	0.86
(11.79, 16.3]	0.86	0.86	0.86	0.86	0.86	0.90	0.85
(16.3, 22.35]	0.84	0.85	0.85	0.85	0.85	0.89	0.83
(22.35, 34.74]	0.84	0.84	0.84	0.84	0.84	0.88	0.83
(34.74, 240.8]	0.82	0.83	0.83	0.83	0.83	0.86	0.81

Average survival probabilities by MWF exposure (deciles, as is possible)

Cumulative exposure	sl	mean	glm	glmnet	gam	ranger	xgboost
Synthetic							
0	0.92	0.89	0.92	0.92	0.92	0.94	0.90
(0, 0.01152]	0.92	0.92	0.92	0.92	0.92	0.95	0.91
(0.01152, 0.5001]	0.95	0.93	0.96	0.95	0.96	0.97	0.94
(0.5001, 1.774]	0.90	0.88	0.90	0.90	0.90	0.93	0.88
(1.774, 105]	0.87	0.86	0.88	0.88	0.88	0.91	0.85

ROC Curve



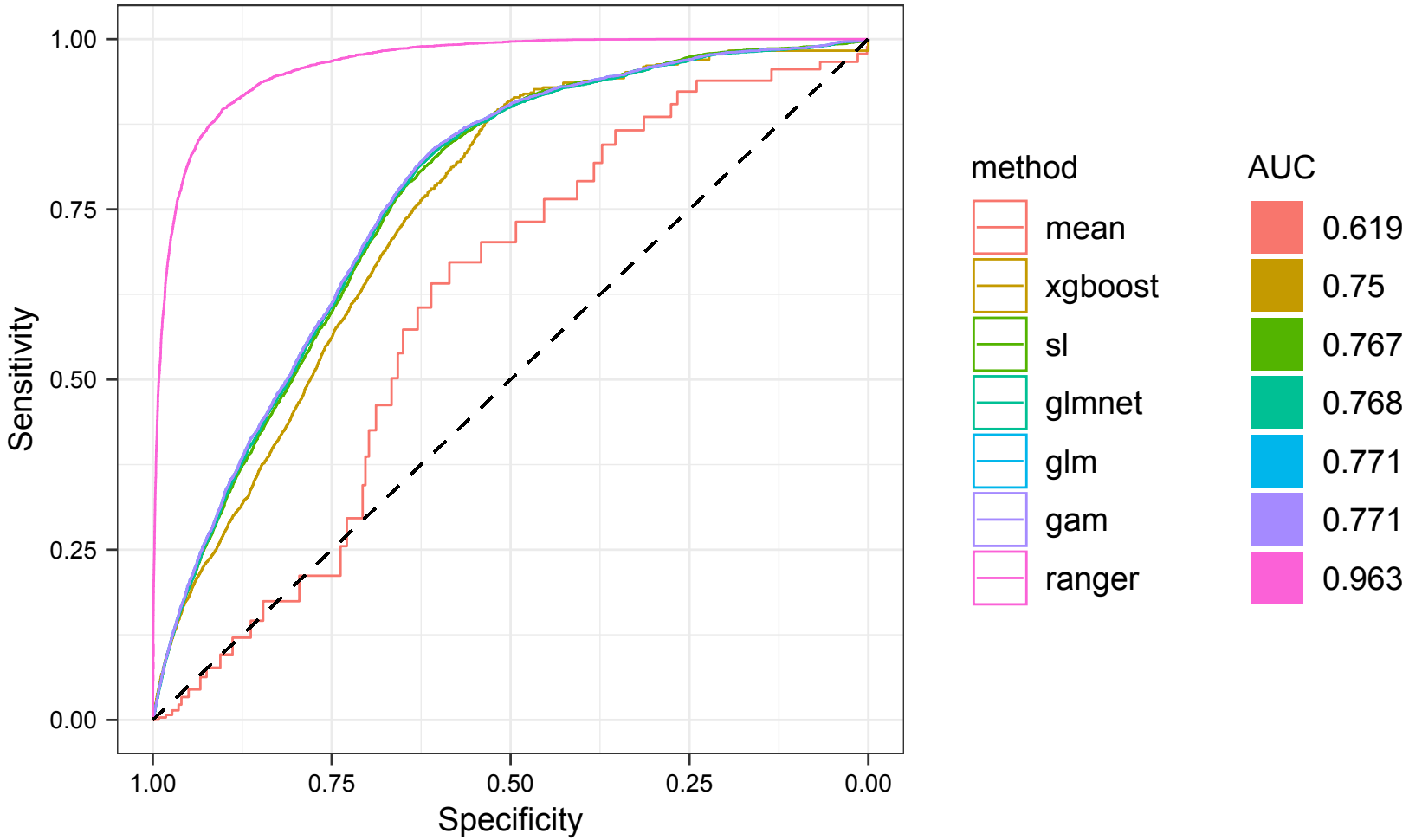
Outcome: Natural cause mortality status in 1984

An individual's probability of not dying due to natural causes was calculated as: $\prod_t (1 - \hat{p}_t)$ where $\hat{p}_t$ is the predicted probability of death due to natural causes for the t^{th} year of follow-up.

ROC Thresholds

method	min	Q2	med	Q3	max
sl	0.0010	0.0203	0.0617	0.1852	0.8537
mean	0.0094	0.0731	0.1327	0.1884	0.2406
glm	0.0003	0.0144	0.0465	0.1717	0.8677
glmnet	0.0003	0.0148	0.0471	0.1725	0.8640
gam	0.0003	0.0144	0.0465	0.1717	0.8677
ranger	0.0003	0.0103	0.0535	0.1837	0.8292
xgboost	0.0015	0.0783	0.1584	0.2771	0.8551

ROC Curve (deciles)



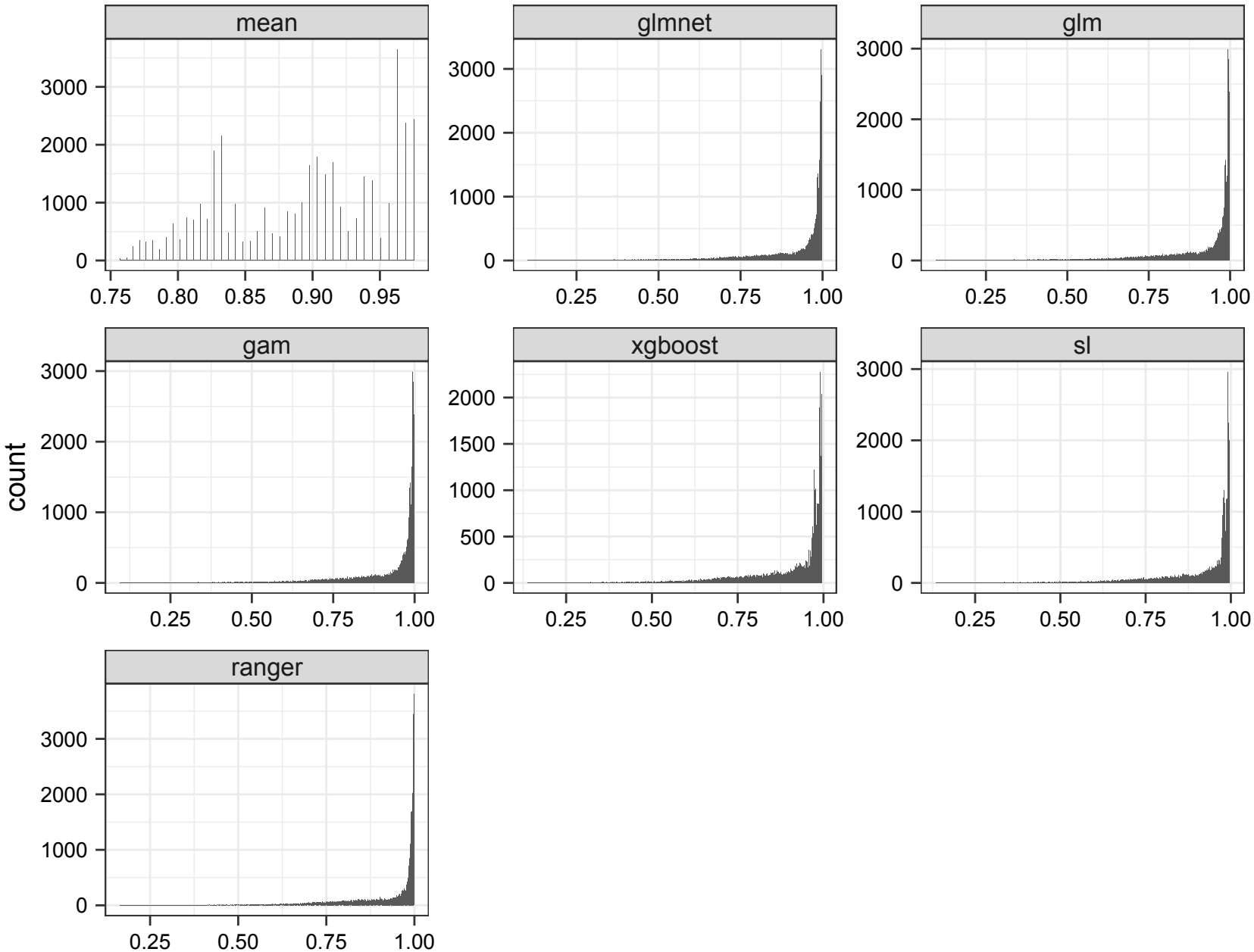
Outcome: Natural cause mortality status in 1984

An individual's probability of not dying due to natural causes was calculated as: $\prod_t (1 - \hat{p}_t)$ where $\hat{p}_t$ is the predicted probability of death due to natural causes for the t^{th} year of follow-up.

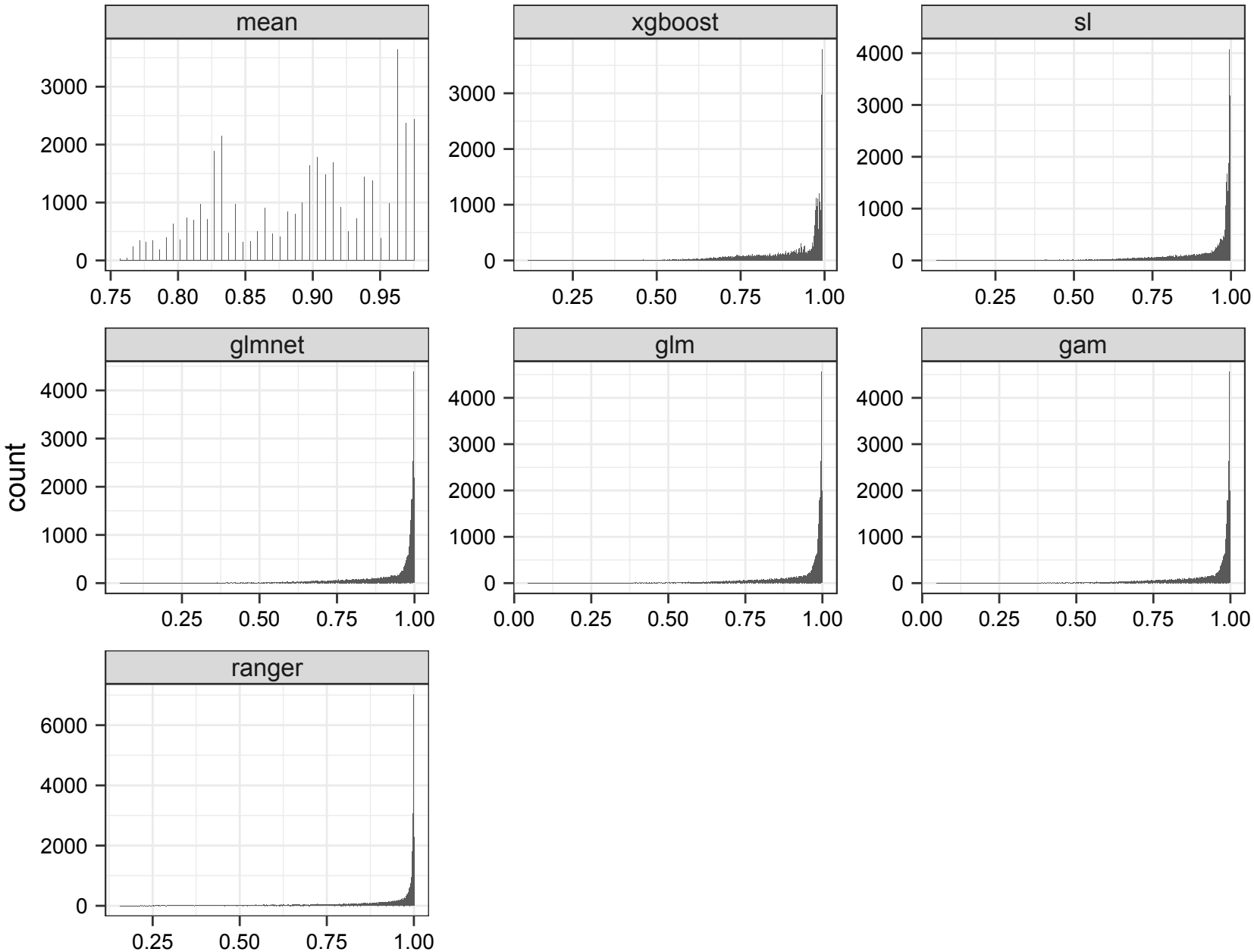
ROC Thresholds (deciles)

method	min	Q2	med	Q3	max
sl	0.0004	0.0092	0.0308	0.1426	0.9283
mean	0.0094	0.0731	0.1327	0.1884	0.2406
glm	0.0002	0.0071	0.0247	0.1344	0.9409
glmnet	0.0002	0.0075	0.0256	0.1354	0.9368
gam	0.0002	0.0071	0.0247	0.1344	0.9409
ranger	0.0000	0.0025	0.0200	0.1334	0.8416
xgboost	0.0020	0.1006	0.1850	0.2851	0.8784

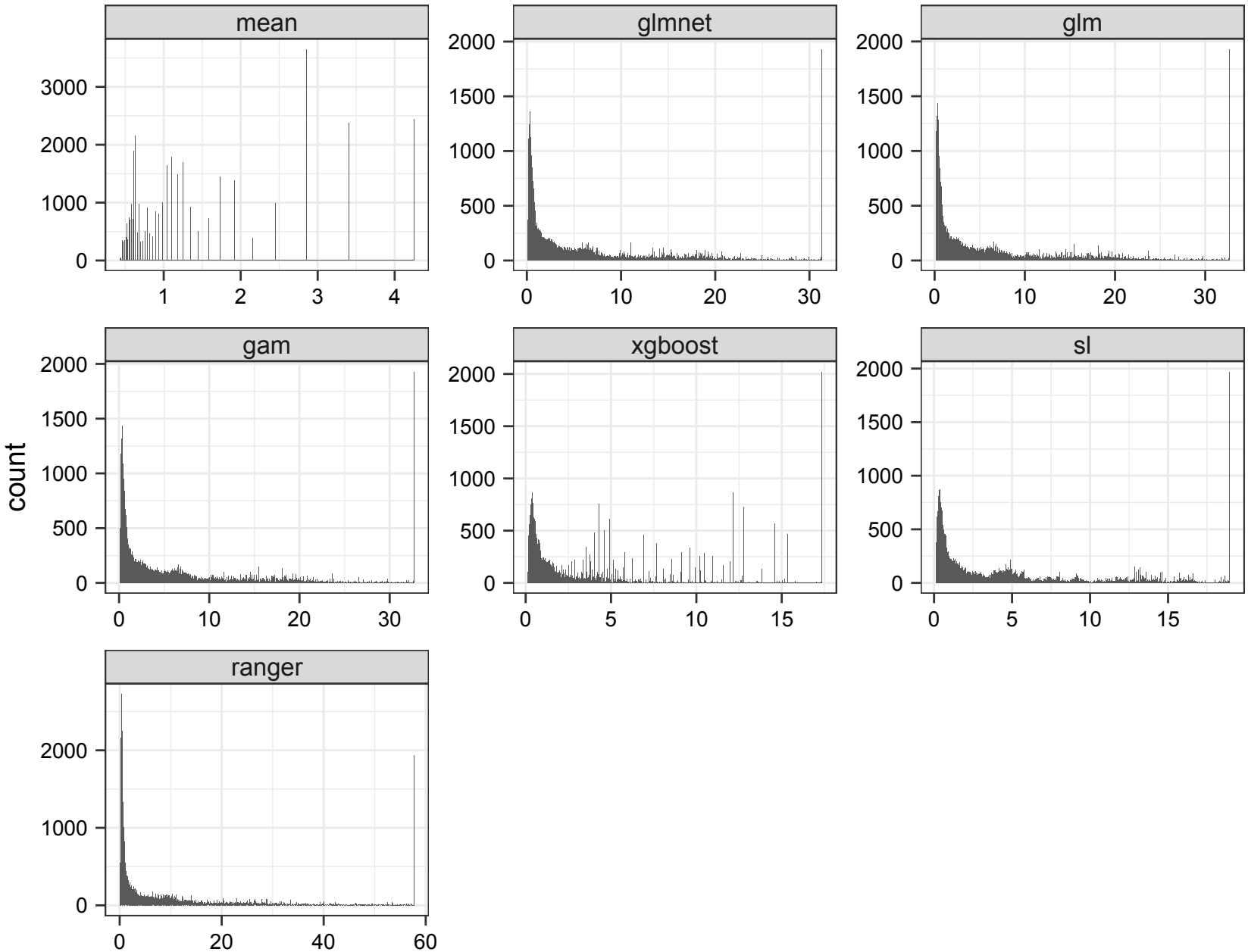
Distribution of survival probabilities, truncated at the 95th percentile



Distribution of survival probabilities, truncated at the 95th percentile (deciles)



Distribution of stabilized weights, truncated at the 95th percentile



Distribution of stabilized weights, truncated at the 95th percentile (deciles)