Validation for Survival Data

STRATOS: David McLernon, . . . , Terry Therneau

July 12, 2022

Big picture

- Stop and think: what is the target?
- Survival data is not yes/no bionomial
  - But we can copy some ideas
- There are a *lot* of papers (not all good)
- High level
  - Match method to the target
  - Censoring drives technical issues

- Reference model + new situation
  - Is the model useful in this context?

- Reference model + new situation
  - Is the model useful in this context?
- Define "useful"
  - Stratify subjects for a clinical trial
  - Make treatment choices
  - Counsel a patient
  - Make global statements about the model itself
  - Understand the model better

- Reference model + new situation
  - Is the model useful in this context?
- Define "useful"
  - Stratify subjects for a clinical trial
  - Make treatment choices
  - Counsel a patient
  - Make global statements about the model itself
  - Understand the model better

"If you don't know where you are going, you might end up someplace else." Yogi Berra

In pratice, this often leads to a time horizon τ

- Utility is focused on an interval

In pratice, this often leads to a time horizon τ

- Utility is focused on an interval
- Limited data

In pratice, this often leads to a time horizon τ

- Utility is focused on an interval
- Limited data
- "Predicted τ year survival" is often a simple way to communicate

Stop and *think*

- D. Altman and P. Royston, What do we mean by validating a prognostic model? (2000)
- E. Korn and R. Simon, Measures of explained variation for survival data (1990)

Metrics

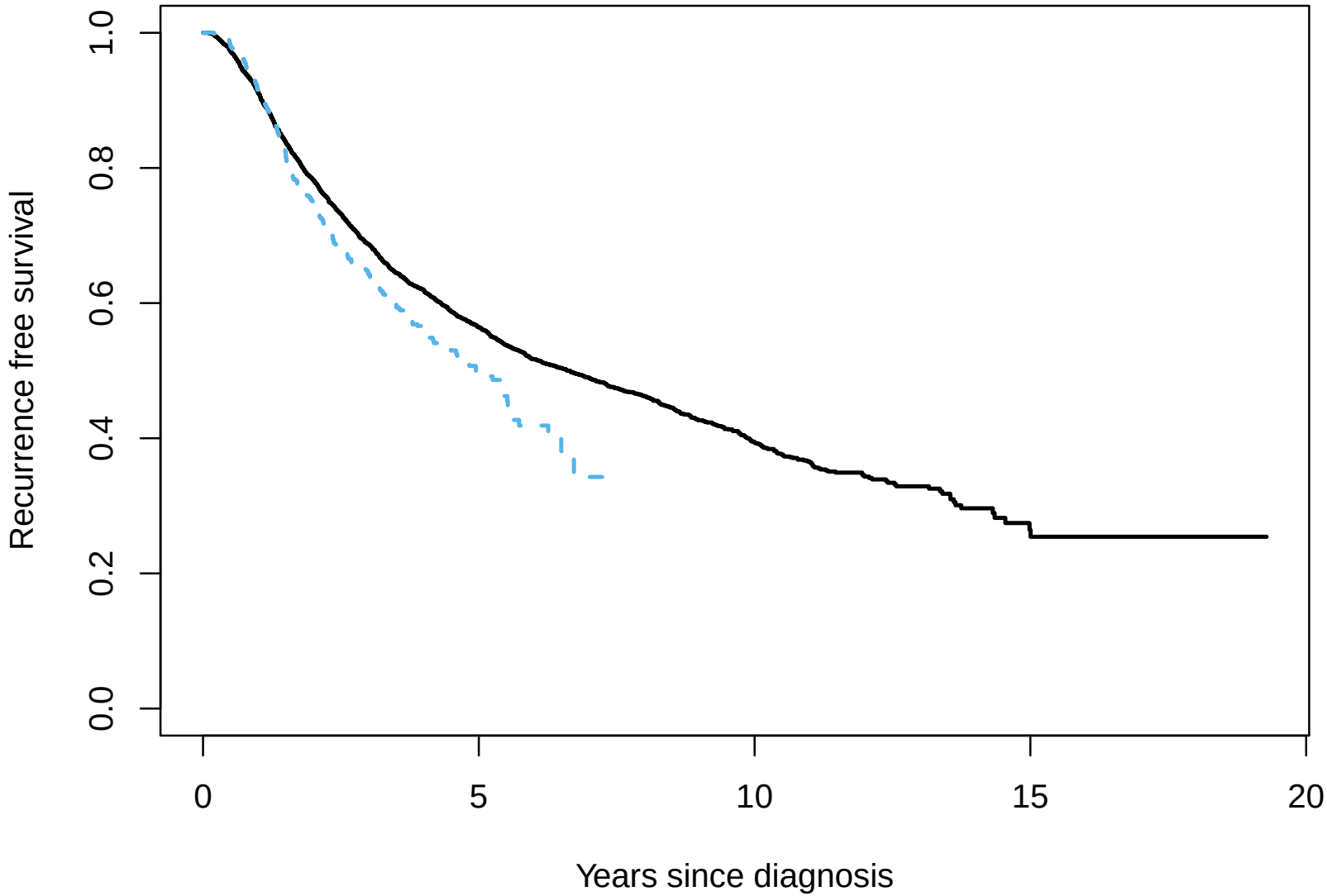
- Discrimination
  - are the predictions in the right order?
  - concordance, AUC

Metrics

- Discrimination
  - are the predictions in the right order?
  - concordance, AUC
- Calibration
  - absolute prediction: "the 5 year death rate is 27%"
  - t_i vs $\hat{t}_i$: difficult
  - absolute risk: usual
    - observed vs expected events by time τ
    - observed vs expected survival probability at τ

Rotterdam and GBSG data sets

- Reference model: 2892 breast cancer patients from the Rotterdam tumor bank
- Validation data: 686 patients from a German Breast Cancer Study Group trial
- (One of the very few publicly available data set pairs.)


Recurrence free survival
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Years since diagnosis

```
Call:
coxph(formula = Surv(ryear, rfs) ~ size + grade + pmin(node
    10), data = rott2)

                    coef exp(coef) se(coef)    z       p
size20-50         0.3051    1.3568   0.0547  5.6  2e-08
size>50           0.5242    1.6891   0.0824  6.4  2e-10
grade             0.3223    1.3803   0.0596  5.4  6e-08
pmin(nodes, 10) 0.1317    1.1407   0.0071 18.5 <2e-16

Likelihood ratio test=563  on 4 df, p=<2e-16
n= 2982, number of events= 1713
```

Discrimination

- $P(y_i > y_j \,|\, \hat{y}_i > \hat{y}_j)$
- Ties: Kendall's tau-a, Kendall's tau-b, Goodman's gamma, Somers' d = opinions
- -1 to 1 versus 0 to 1: $C = (d + 1)/2$

Discrimination

- $P(y_i > y_j \,|\, \hat{y}_i > \hat{y}_j)$
- Ties: Kendall's tau-a, Kendall's tau-b, Goodman's gamma, Somers' d = opinions
- -1 to 1 versus 0 to 1: $C = (d + 1)/2$
- If y is 0/1, C = AUROC
- For survival, ignore unrankable pairs ($i = 10+$, $j = 20$). Harrell's C
- y= survival, x= 0/1: Harrell's C = Gehan-Wilcoxon

Discrimination

- $P(y_i > y_j \,|\, \hat{y}_i > \hat{y}_j)$
- Ties: Kendall's tau-a, Kendall's tau-b, Goodman's gamma, Somers' d = opinions
- -1 to 1 versus 0 to 1: $C = (d + 1)/2$
- If y is 0/1, C = AUROC
- For survival, ignore unrankable pairs ($i = 10+$, $j = 20$). Harrell's C
- y= survival, x= 0/1: Harrell's C = Gehan-Wilcoxon
- y= survival, x= 0/1: Uno C = Peto-Wilcoxon
- Reprise of log-rank vs Gehan-Wilcoxon vs rho-gamma vs Schemper vs . . . debate
  - If all risk sets are > 20 it hardly matters
  - Uno C can get odd for small risk sets
- A storm in a teacup

Dichotomania

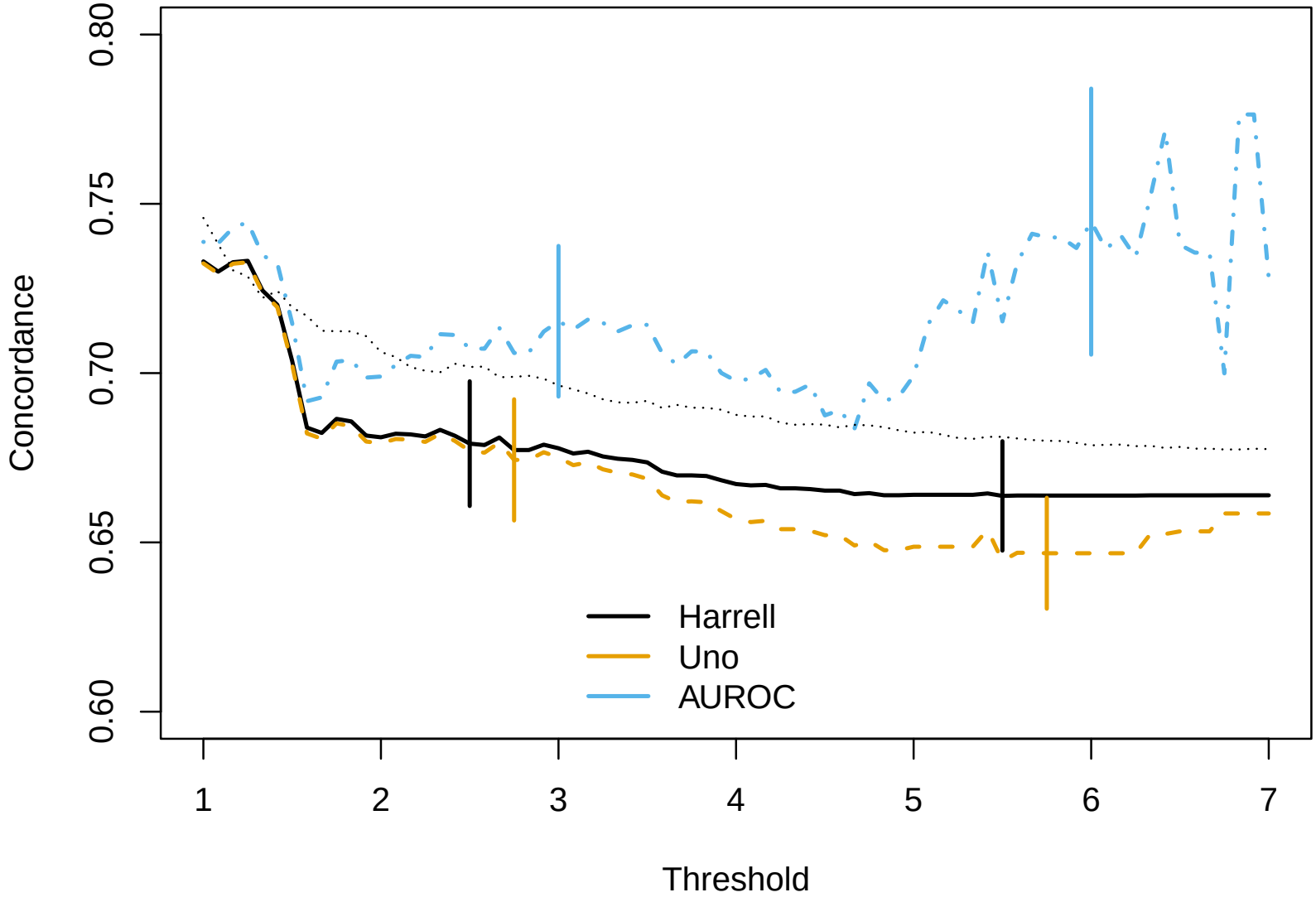
- Serious disease in clinical researchers

Dichotomania

- Serious disease in clinical researchers
- Time Dependent AUROC
  - Dichotomize time at some point τ
  - $P(I\{y_i > \tau\} > I\{y_j > \tau\} \,|\, \hat{y}_i > \hat{y}_j)$
  - Many papers
  - Reprise censoring weight arguments

Dichotomania

- Serious disease in clinical researchers
- Time Dependent AUROC
  - Dichotomize time at some point τ
  - $P(I\{y_i > \tau\} > I\{y_j > \tau\} \,|\, \hat{y}_i > \hat{y}_j)$
  - Many papers
  - Reprise censoring weight arguments
- `www.senns.uk/wprose.html#Dance`
- (My opinion)
- $P(\min(y_i, \tau) > \min(y_j, \tau) \,|\, \hat{y}_i > \hat{y}_j)$


Concordance
0.60
0.65
0.70
0.75
0.80
1
2
3
4
5
6
7
Threshold
Harrell
Uno
AUROC

Calibration

- Need $\hat{p}$; this requires the baseline hazard
- How to handle censoring in the validation data
  - say your target $\tau = 5$ years
  - $\hat{p}_i(5) =$ predicted is available for all
  - $y_i(5) = 0/1 =$ death is not known for someone censored at 3
    1. Eliminate, using IPC weights
    2. Impute (and smooth), using a survival model
    3. Counting process approach (SIR)
    4. Pretend they aren't there

Eliminate

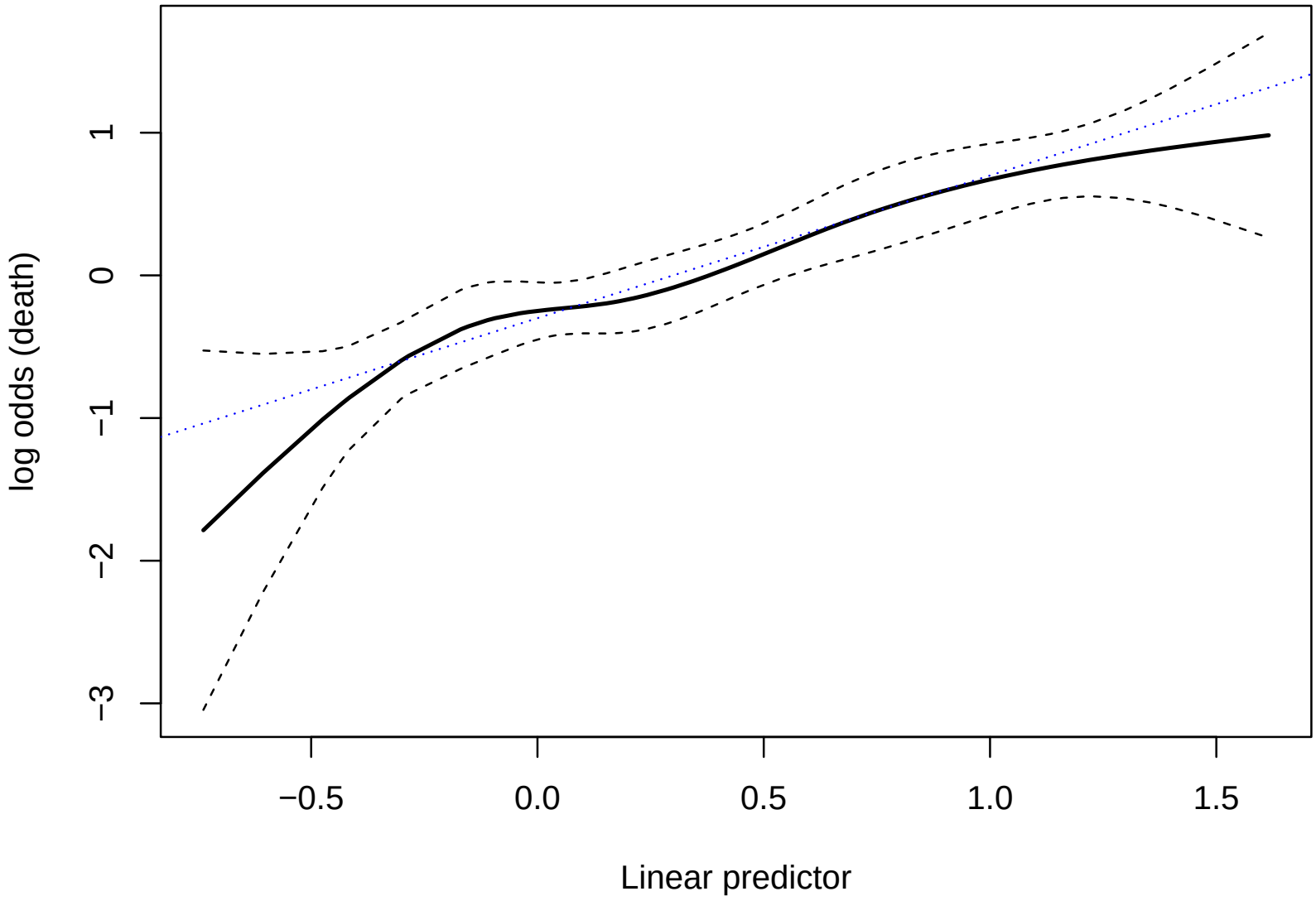
- Redistribute to the right: starting at the left, censored subjects give their case weight to those with more follow-up. Stop at τ.
- All the problem cases now have a weight of 0
- Use your favorite binomial tools: ROC curve, sensitivity, specificity, ... (with case weights)
- Logistic regression of new y vs spline(log(-log(phat)))
- Weighted R^2 (Brier score at τ)

Eliminate

- Redistribute to the right: starting at the left, censored subjects give their case weight to those with more follow-up. Stop at τ.
- All the problem cases now have a weight of 0
- Use your favorite binomial tools: ROC curve, sensitivity, specificity, ... (with case weights)
- Logistic regression of new y vs spline(log(-log(phat)))
- Weighted R^2 (Brier score at τ)
- old familiar metrics
- need a robust variance
- assumes censoring is independent

```
> eta <- predict(rfit, newdata=gbsg2)
> wt5 <- rttright(Surv(ryear, rfs) ~ 1, data = gbsg2,
                  times = 5)
> table(wt5 ==0)

FALSE  TRUE
  406   280
> fit5 <- glm(I(ryear < 5) ~ ns(eta,4), weights= wt5,
              family= quasibinomial(link= "cloglog"),
               data= gbsg2)
```

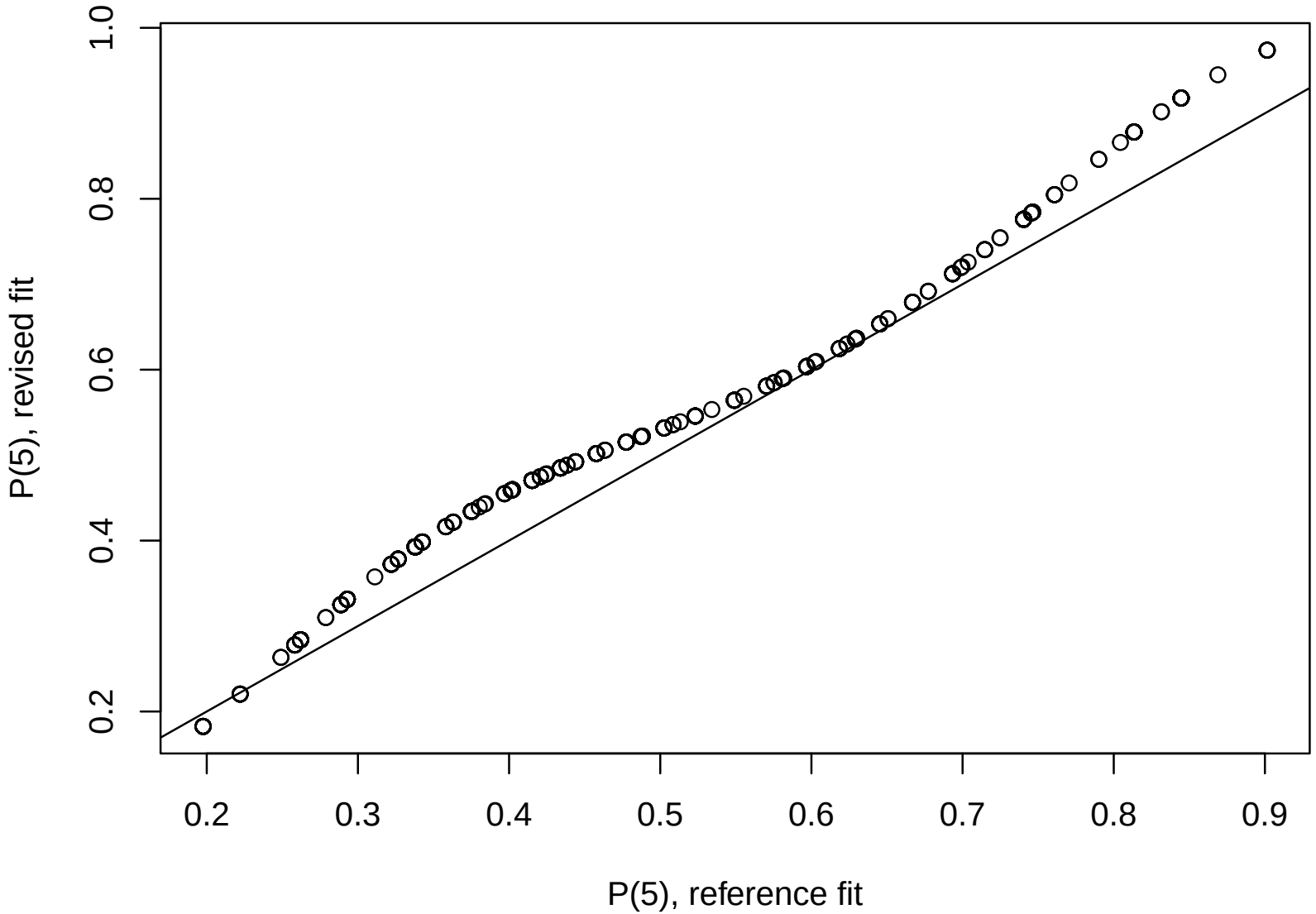

log odds (death)
1
0
−1
−2
−3
−0.5
0.0
0.5
1.0
1.5
Linear predictor

Impute and smooth

- Fit a survival model to the validation data, with $z = \log(-\log(1 - \hat{p}_i))$ as the predictor;
- The new model should be flexible, e.g., spline(z)
- Obtain a new prediction $\tilde{p}_i$
- Compare $\hat{p}_i$ to $\tilde{p}_i$

Impute and smooth

- Fit a survival model to the validation data, with $z = \log(-\log(1 - \hat{p}_i))$ as the predictor;
- The new model should be flexible, e.g., spline(z)
- Obtain a new prediction $\tilde{p}_i$
- Compare $\hat{p}_i$ to $\tilde{p}_i$
- Simple to do, simple graph
- A Cox model is quick but may be too inflexible.
- (Better models are less accessible.)
- Do not label the plot as "observed vs predicted"

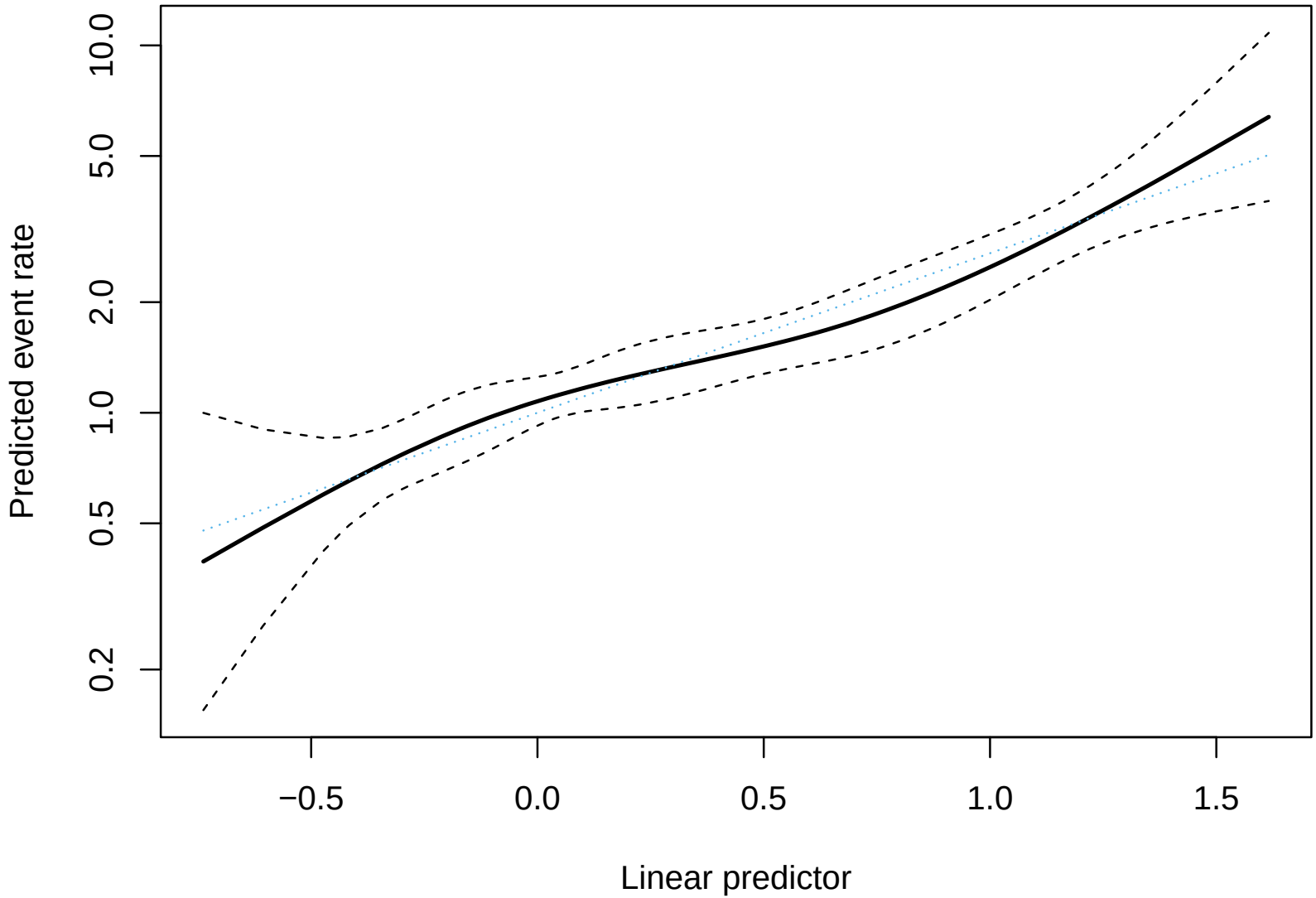

P(5), revised fit
1.0
0.8
0.6
0.4
0.2
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
P(5), reference fit

Observed vs Expected deaths

- For someone censored at 2.5 years, use $\hat{p}(2.5)$ rather than trying to impute $y(2.5)$
- Leads to "observed deaths in 5 yrs" vs "predicted deaths in 5"
- The standardized incidence ratio (SIR). Common in epidemiology; observed/expected
- Simple computational trick (poisson glm + offset)

Observed vs Expected deaths

- For someone censored at 2.5 years, use $\hat{p}(2.5)$ rather than trying to impute $y(2.5)$
- Leads to "observed deaths in 5 yrs" vs "predicted deaths in 5"
- The standardized incidence ratio (SIR). Common in epidemiology; observed/expected
- Simple computational trick (poisson glm + offset)
- Simple to do, simple graph
- Solid theory based on counting processes
- Unaffected by censoring issues
- Reliable confidence intervals and p-values
- Unfamiliar to many


Predicted event rate
0.2
0.5
1.0
2.0
5.0
10.0
−0.5
0.0
0.5
1.0
1.5
Linear predictor

Ignore censoring

- Shown to be a bad idea for ordinary survival curves in 1952 (Berkson), referred to by Kaplan and Meier (1958).
- Lives on
  - Censored before τ: treated as alive
  - Censored before τ: toss the observation

Summary

- It isn't hard
- Concordance, RTTR (IPCW), refit survival, and O/E are all easy
- O/E $>$ refit $>$ RTTR in terms of fewer assumptions, wider validity
- But all are pretty good

Summary

- It isn't hard
- Concordance, RTTR (IPCW), refit survival, and O/E are all easy
- O/E $>$ refit $>$ RTTR in terms of fewer assumptions, wider validity
- But all are pretty good
- Think