Survival in Random Forests

John Ehrlinger

Department of Quantitative Health Sciences
Lerner Research Institute
Cleveland Clinic
john.ehrlinger@gmail.com

Random Forest

Mature statistical “machine learning” method for

- Regression (continuous outcomes)
- Classification (categorical outcomes)
- Survival (time to event outcomes)
- Others (competing risk, unsupervised, etc.)

Similar to C4.5

Breiman's Two Cultures

Machine Learning vs. Statistics

Machine Learning:

- Prediction, Prediction, Prediction
- Black box modeling

Statistics:

- Why?
- Information on underlying process

Random Forest:

- Why not both?
- Insight into the black box of prediction

Example

Primary Biliary Cirrhosis (PBC) of the liver data set

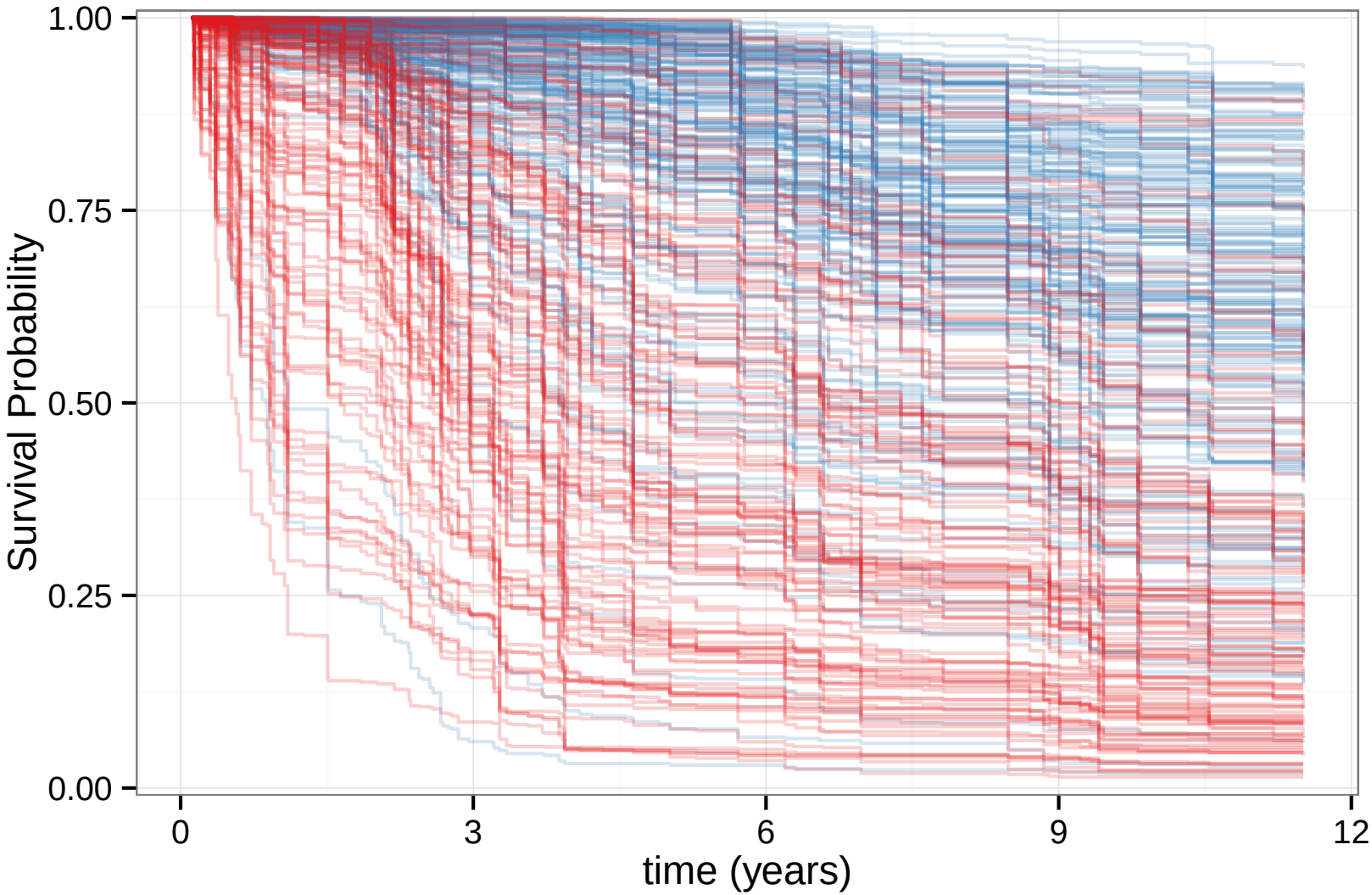
(Fleming and Harrington 1991)

Randomized trial of D-penicillamine (DPCA) at Mayo Clinic

312 patients from 1974 to 1984

- 125 deaths
- 17 variables

Random Survival Forest



Variable (Feature) Selection

Two independent ranking methods

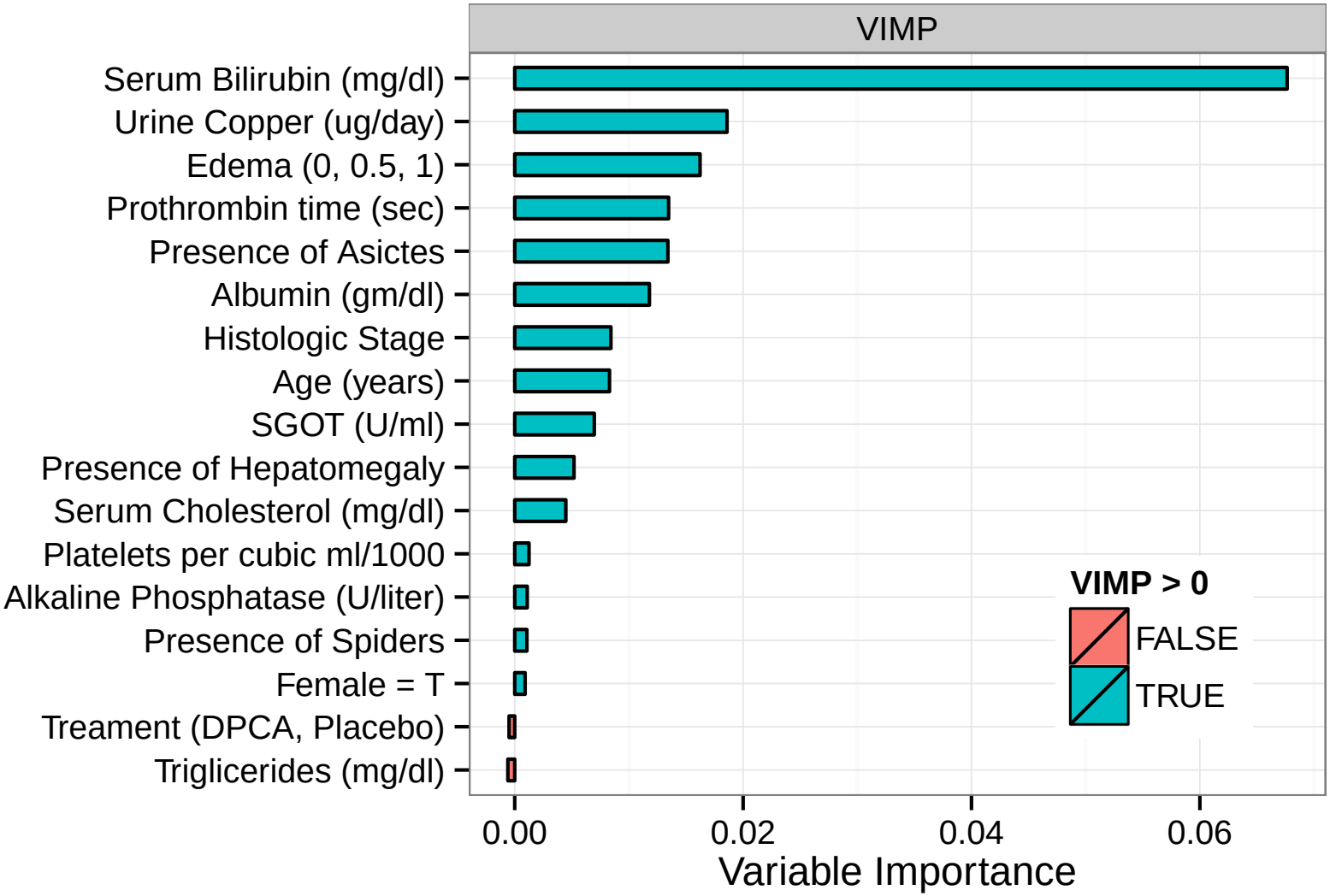
Variable IMPortance (VIMP)

- Based on RF Prediction Error
- Measures the impact of variable misspecification

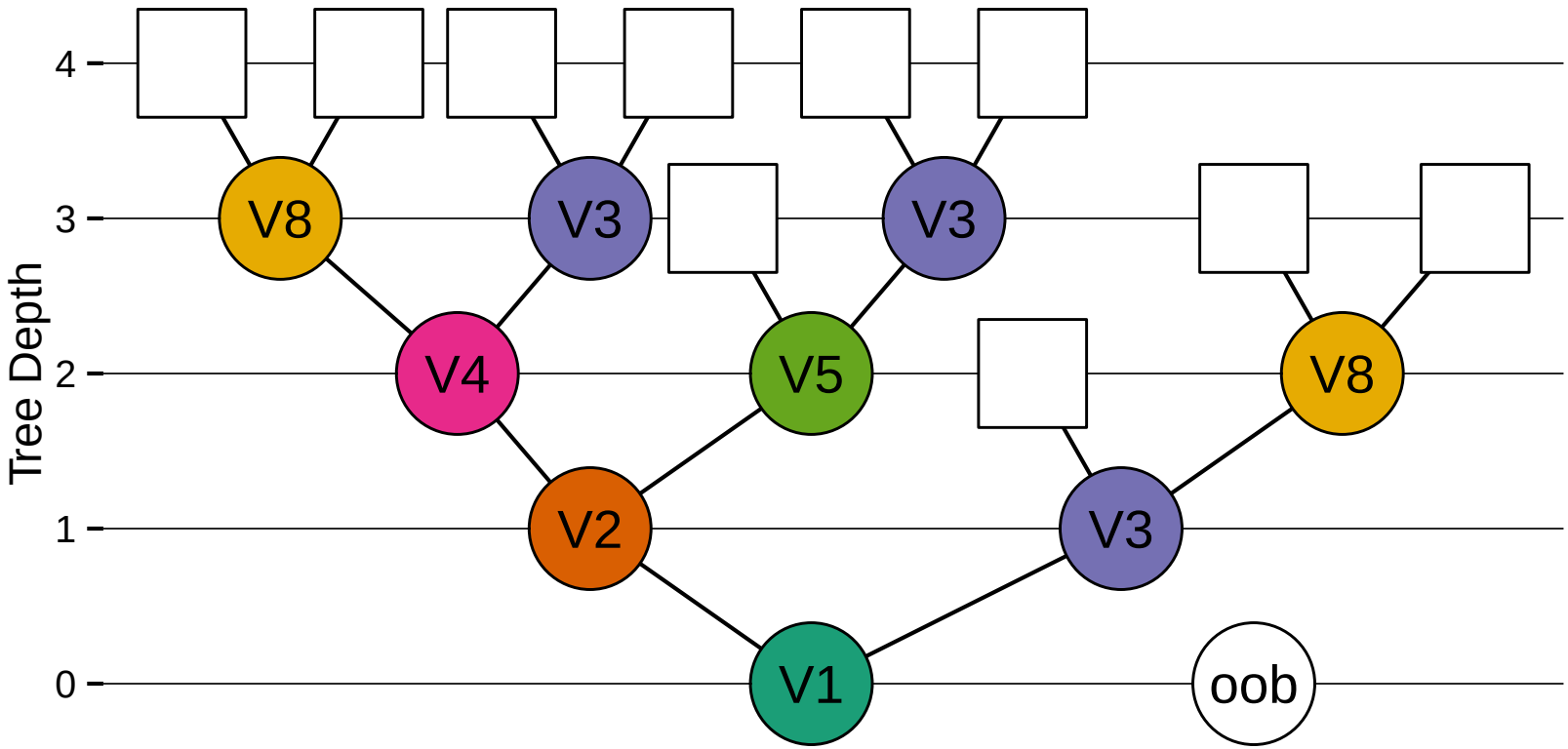
Minimal Depth

- Property of decision tree construction
- Measures how a variable segments nodes

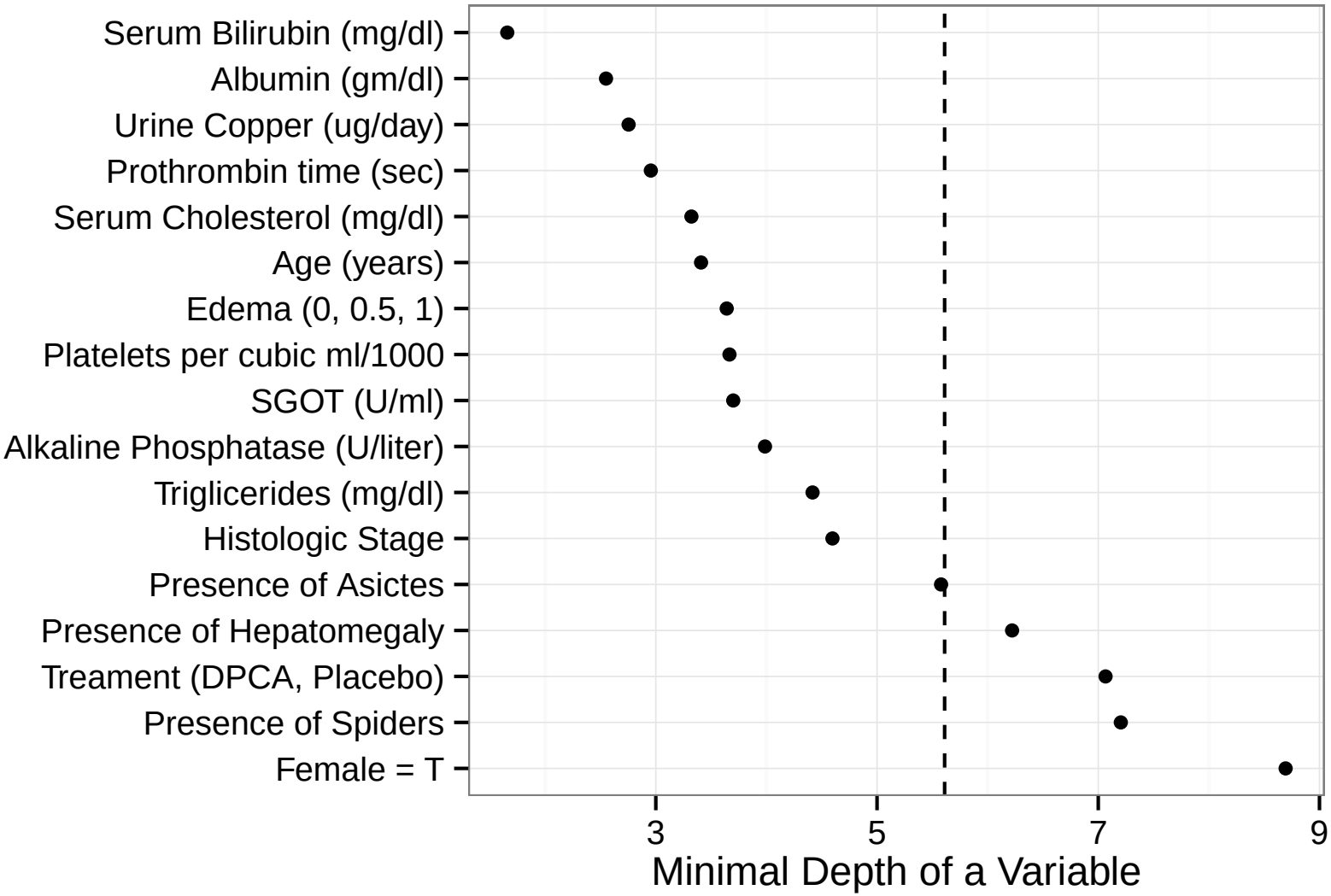
Variable Selection - VIMP



Variable Selection - Minimal Depth



Variable Selection - Minimal Depth



Random Forest

Which variables contribute to forest prediction?

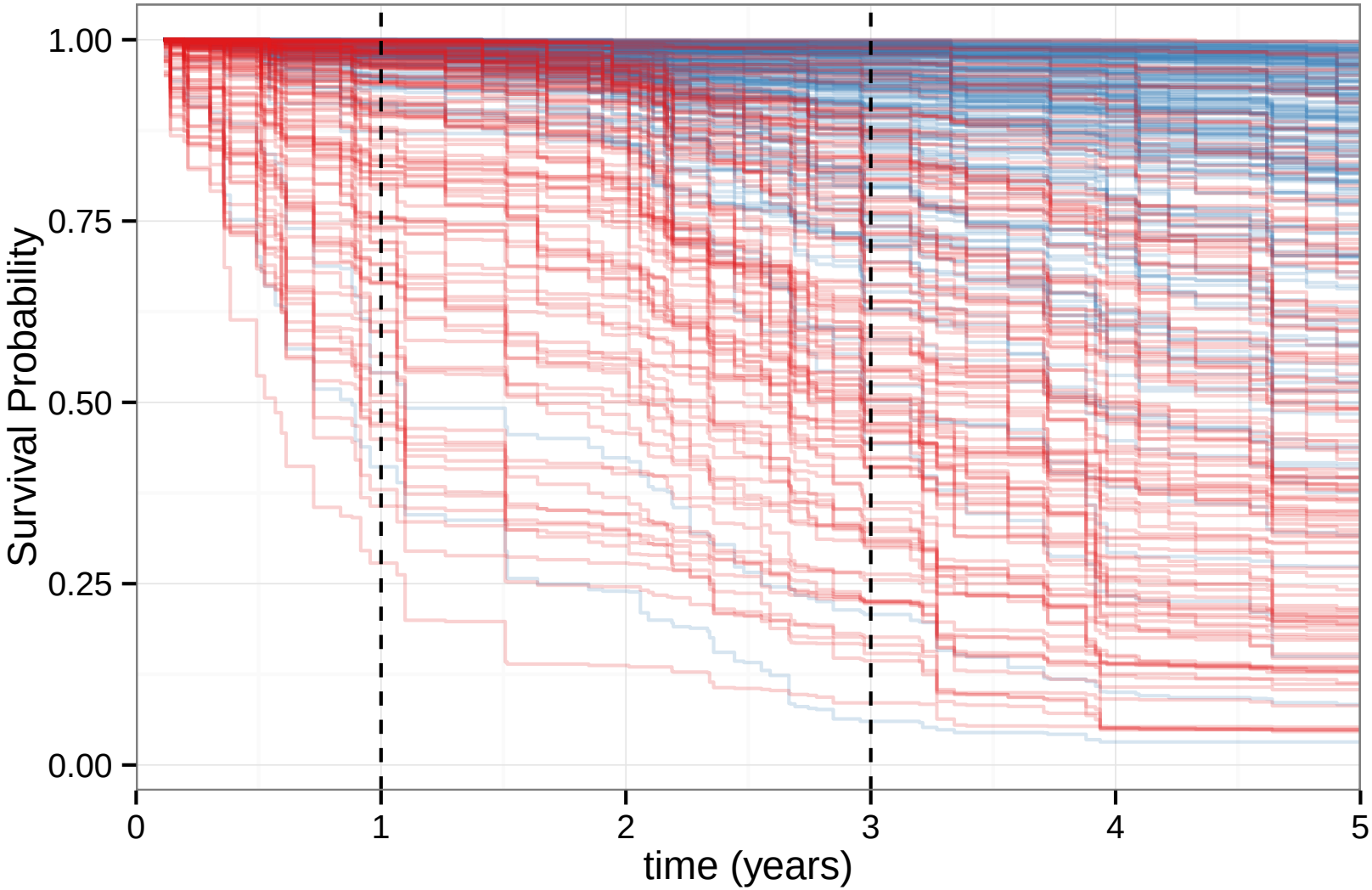
- “Stacking” VIMP and Minimal Depth

How does response depend on variables?

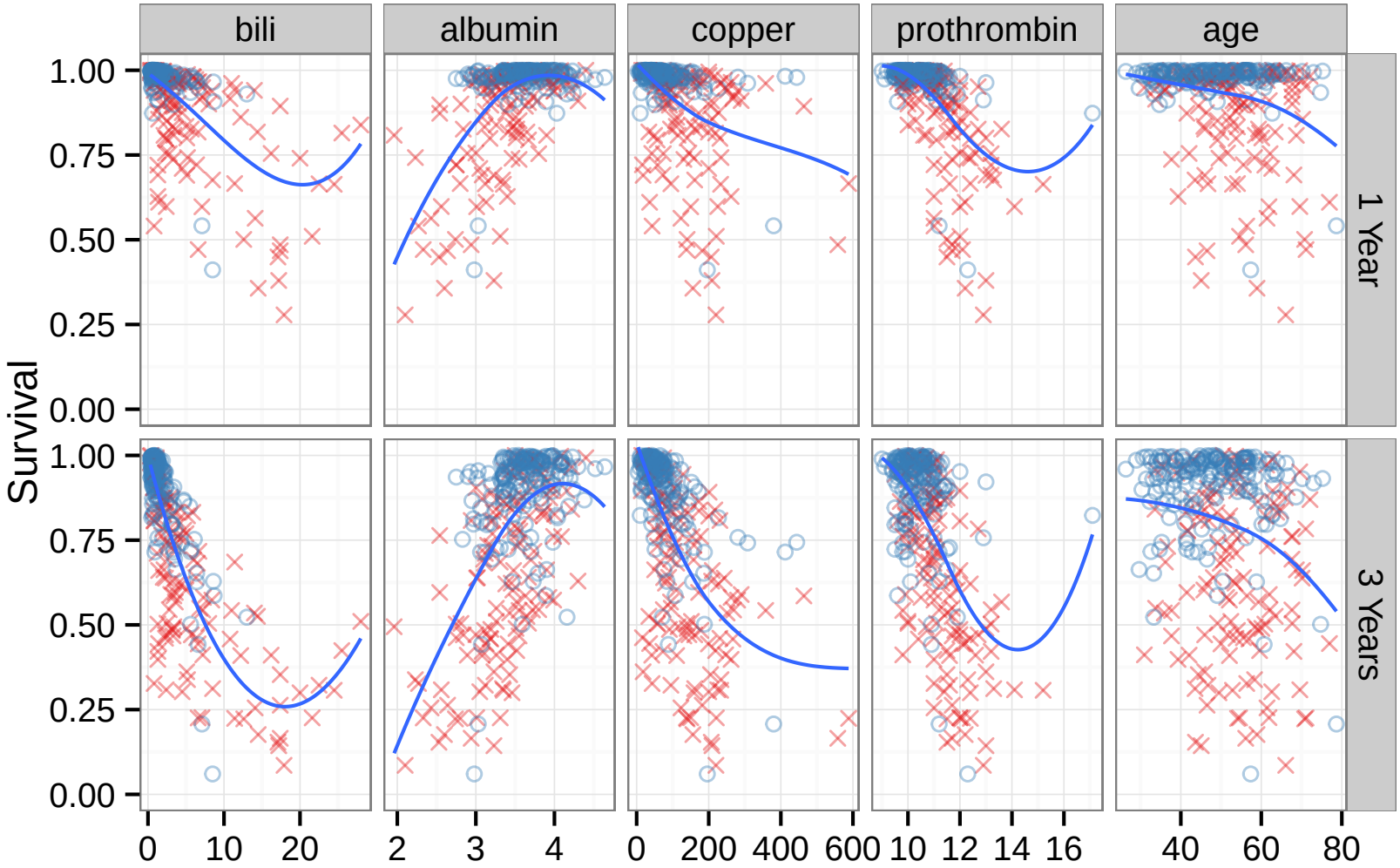
- Variable Dependence - Observation Based
- Partial Dependence - Population Based

Variable Dependence

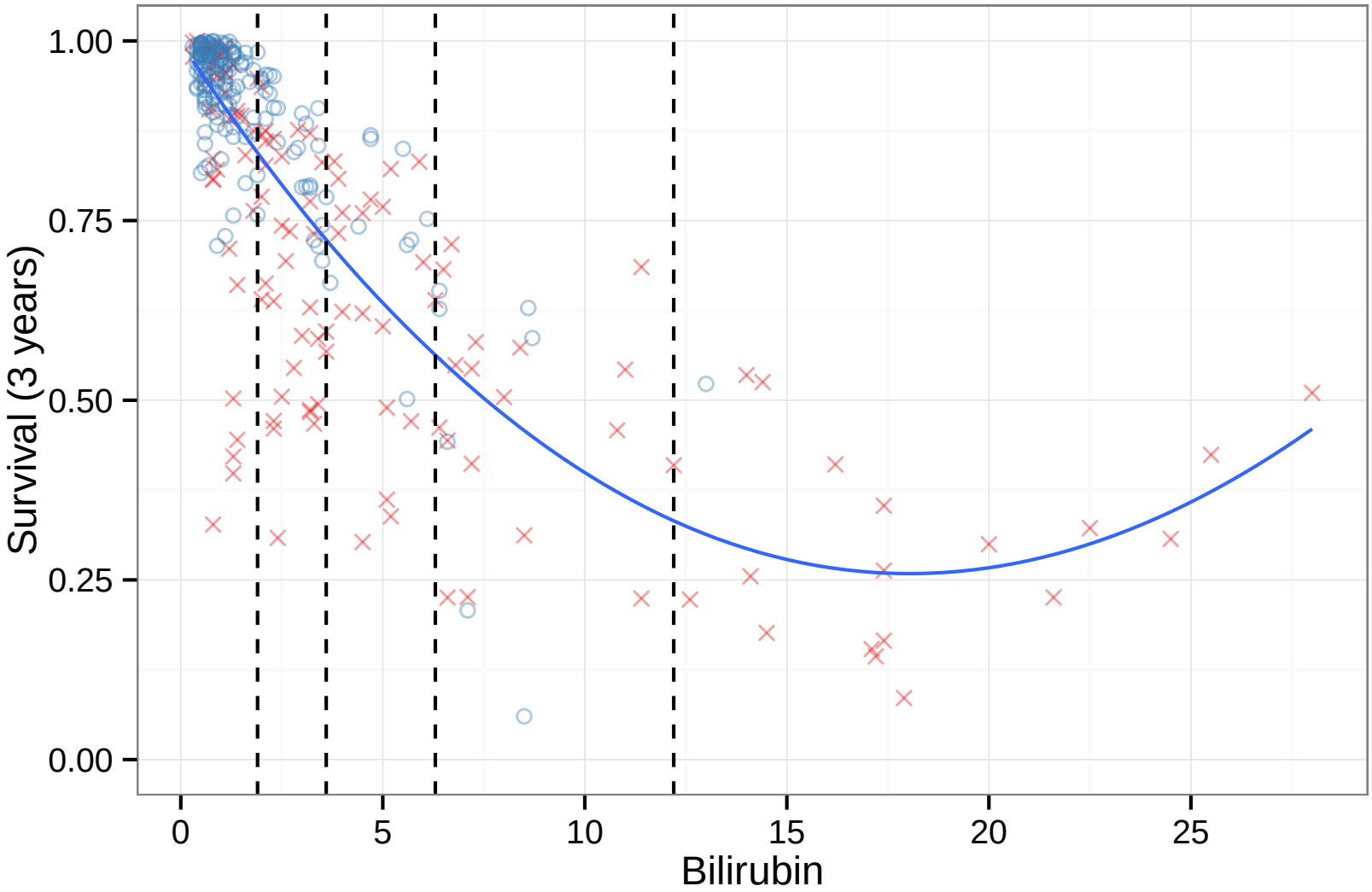
Observation based



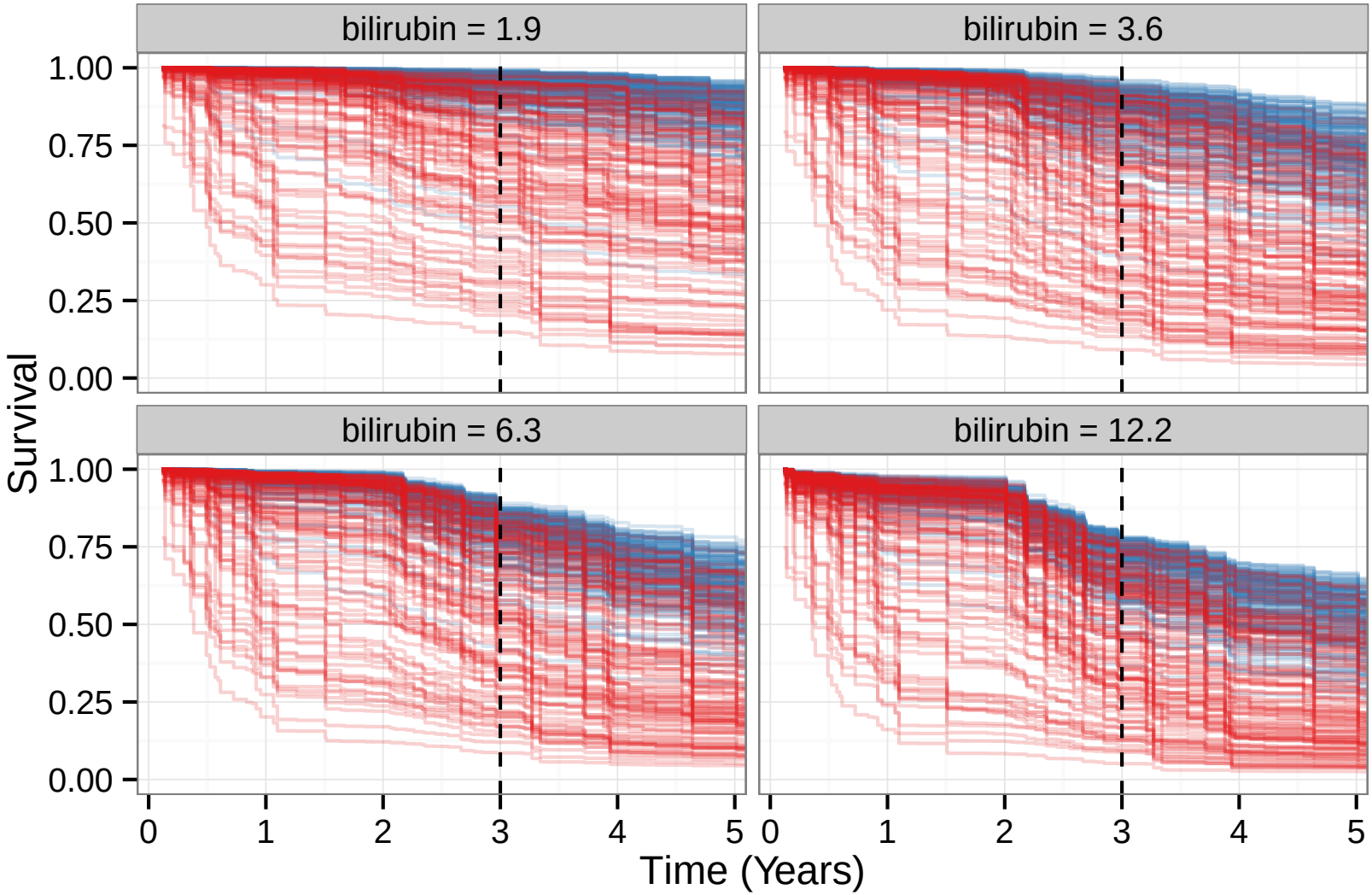
Variable Dependence



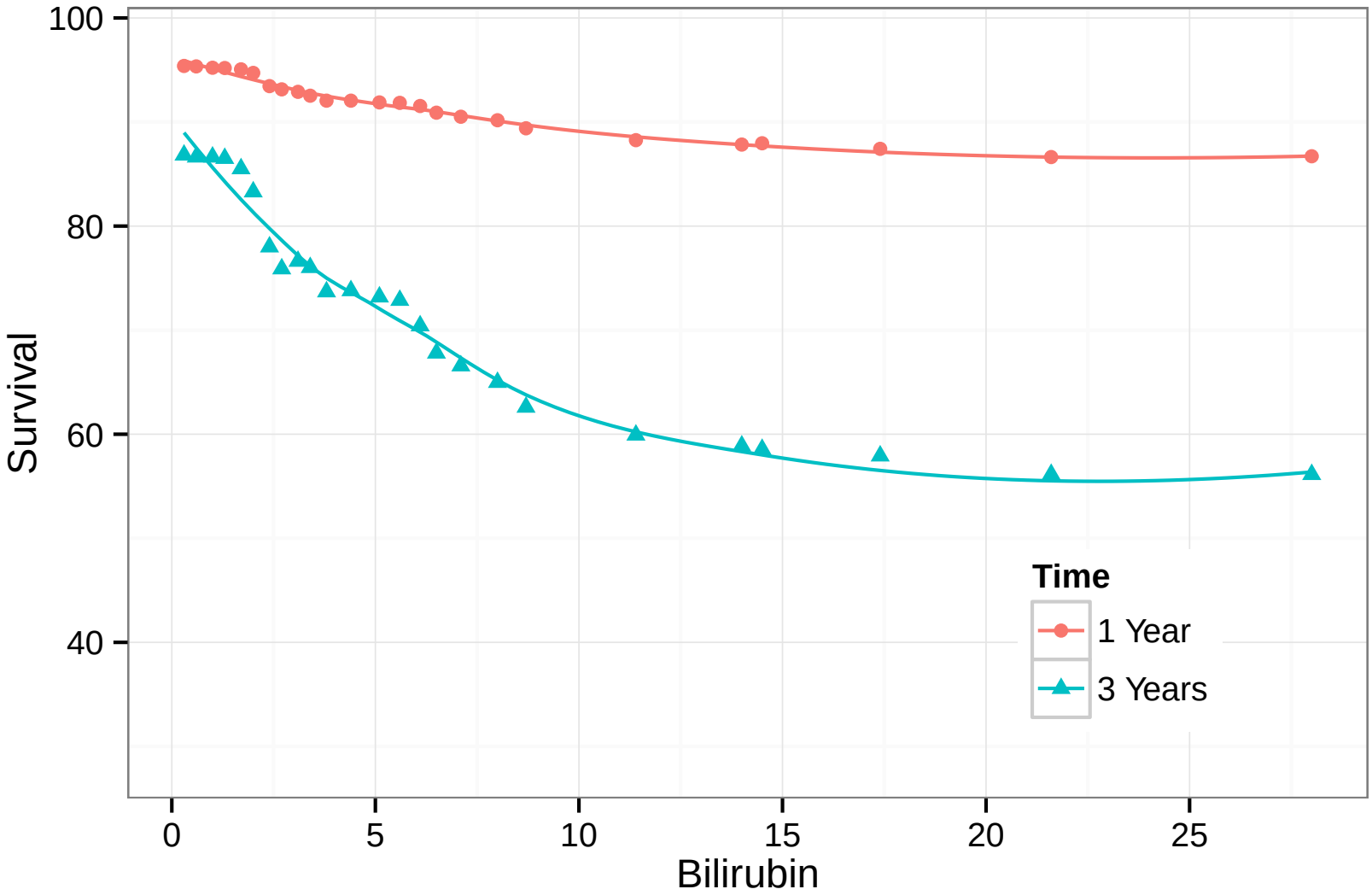
Partial Dependence



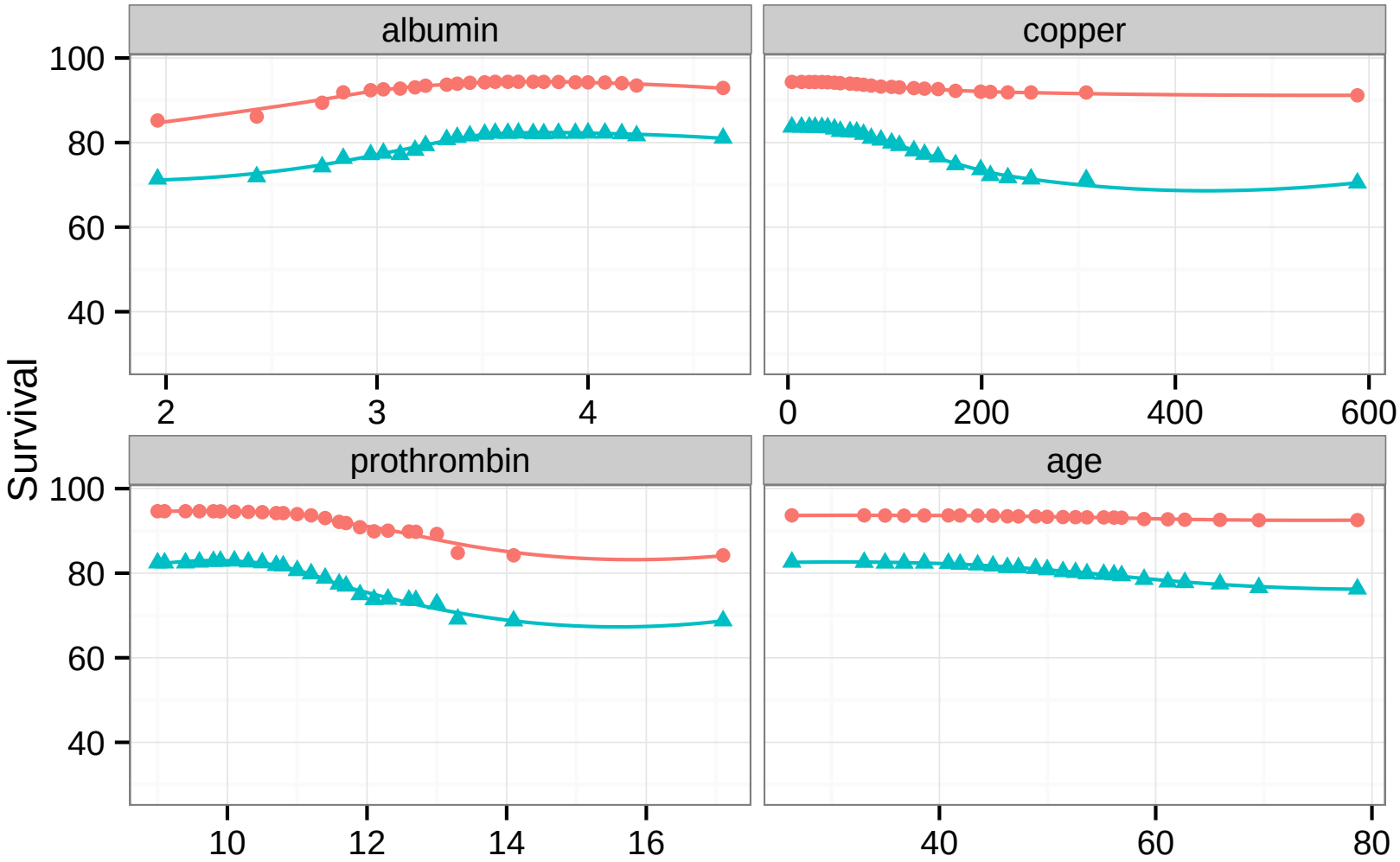
Partial Dependence



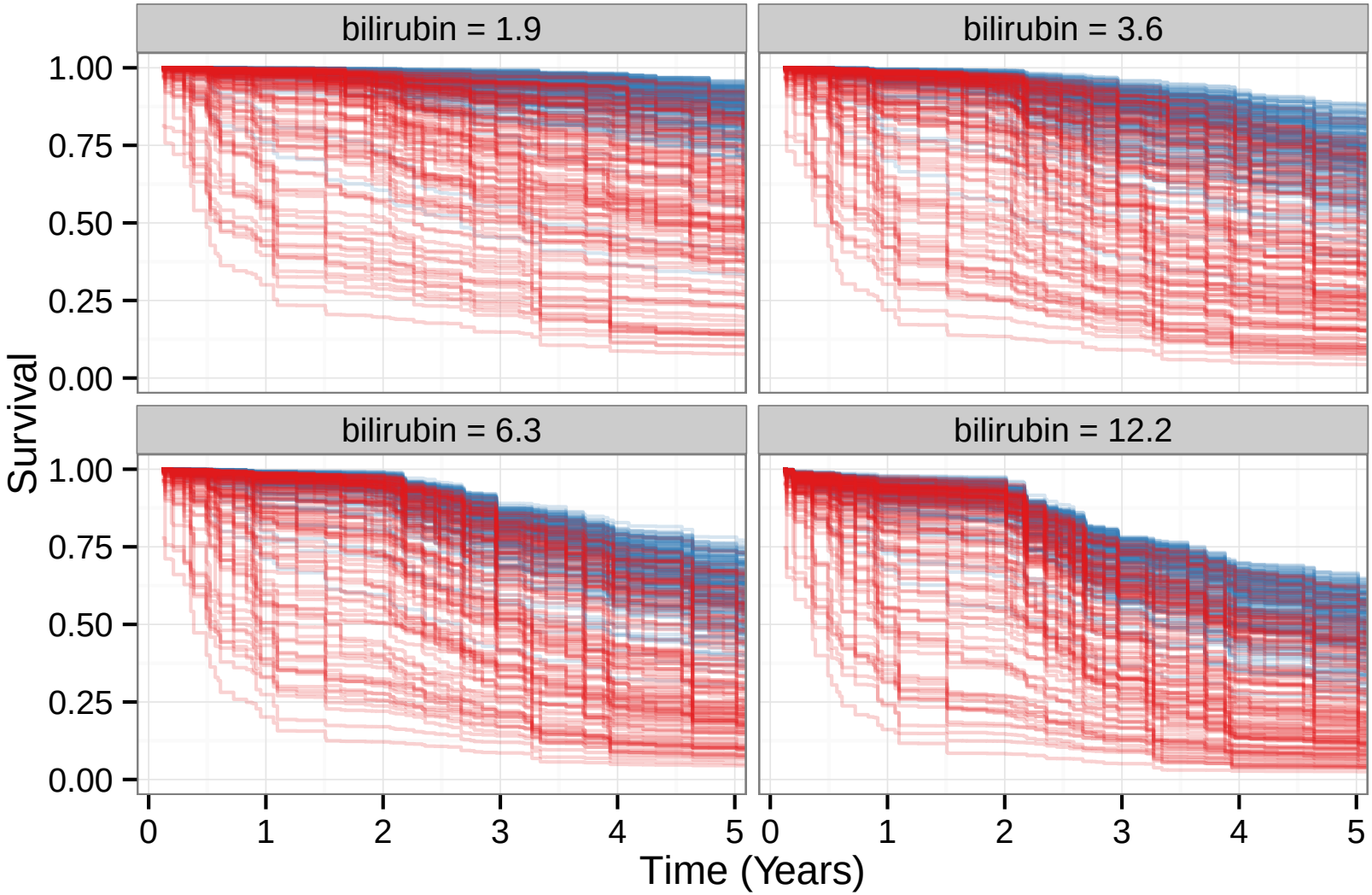
Partial Dependence



Partial Dependence



Partial Dependence



Partial Dependence