

How many events do we need to fit a risk model?

Gareth Ambler and Rumana Omar

Statistical Science, University College London

Joint UCLH/UCL Biomedical Research Unit

Background: risk models

- Risk prediction models are commonly used to make predictions regarding a patient's health
- I will be discussing risk models for binary outcomes
 - e.g. predicting the probability of mortality within 30 days following heart valve surgery
 - typically developed using logistic regression
- Such models are useful for:
 - facilitating case-mix adjusted comparisons between institutions
 - assisting patients to make informed decisions regarding their treatment

Background: EPV

- Ideally we should have enough data to develop a useful, reliable risk model with good calibration and discrimination
- A common problem is over-fitting
 - the model works well with development data but performs poorly in new validation data
- To avoid this it has been suggested that the EPV should be at least 10 when developing risk models (Peduzzi *et al*, 1996; Harrell, 2001)
 - EPV = Events Per Variable
 - e.g. 100 deaths, 10 predictors \Rightarrow EPV = 10
- This sort of calculation is used routinely in sample size calculations

Background: Peduzzi / Concato paper

- Case study: based on a cardiac trial of 673 patients
 - 252 deaths
 - 7 predictors of mortality
- Simulation is used to assess:
 - bias in the (logistic) regression coefficients
 - coverage of confidence intervals
 - power (significance testing)
- Conclude that problems more likely to occur for EPVs less than 10
- Similar paper for survival analysis

Objective

- The main objective is to assess the rule of 10 guideline in the context of risk models
- Assess the calibration and discrimination of models developed with different EPV in different scenarios:
 - calibration: accuracy of predictions
 - discrimination: separation of low- and high-risk patients
- Various factors were investigated including:
 - strength of the risk model (amount of prognostic information)
 - addition of noise predictors
 - correlated predictors
 - binary / survival outcome
 - prevalence of outcome / censoring

Calibration and Discrimination

- Discrimination assessed using the ROC area / c-index (Harrell, 2001)
 - consider all patient pairs where subjects had discordant outcomes
 - ROC area is proportion of pairs whose predictions and outcomes are concordant
 - predicted probability is higher for the subject who had an event
- Calibration was assessed using calibration slopes (Miller *et al.*, 1991)
 - the predicted log-odds ($\eta = X\beta$) fitted as a single term to the validation data
 - regression slope $\neq 1$ suggests over-fitting in the original data

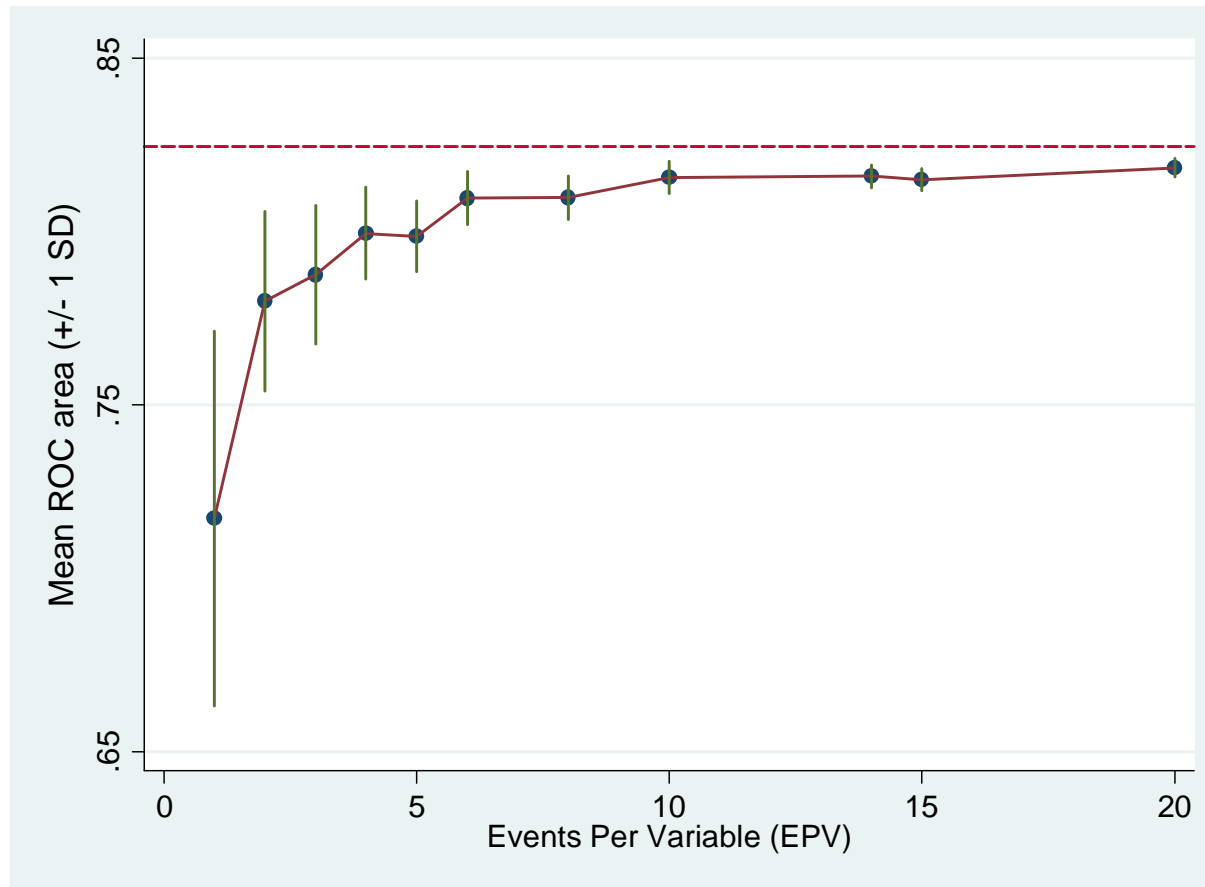
Methods: example

- Simulation details:
 - 20,000 observations generated: (Y, X_1, \dots, X_{10})
 - 10 correlated predictors:
 - $X_1, \dots, X_{10} \sim N(0,1)$
 - pairwise correlation = 0.30
 - $Y \sim \text{Bernoulli}(p)$ where:
 - $\log(p/1-p) = \beta_0 + \beta_1 X_1 + \dots + \beta_{10} X_{10}$
 - $\beta_1 = \dots = \beta_{10} = 0.25$
 - β_0 chosen to give an outcome prevalence of 20%

Methods: example

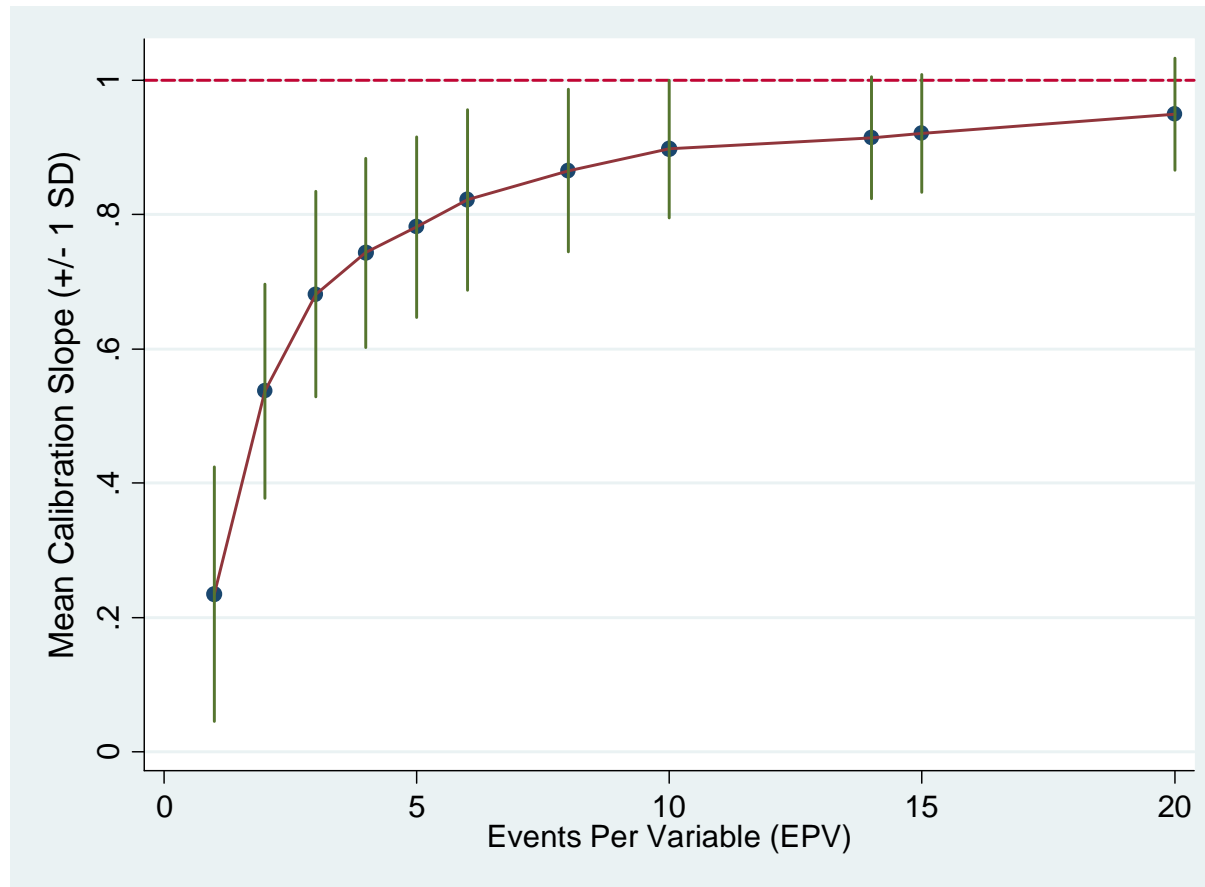
- Model development
 - development data sampled
 - different sample sizes to give different average EPV
 - logistic regression model fitted with all 10 predictors
- Model validation
 - models validated on remaining data
- 200 datasets sampled for each sample size / EPV

Results: ROC area



- Clearly, more data results in a better discrimination
 - reasonable discrimination for $EPV \geq 6$
 - a lot of variability for the lower EPV scenarios

Results: calibration slope

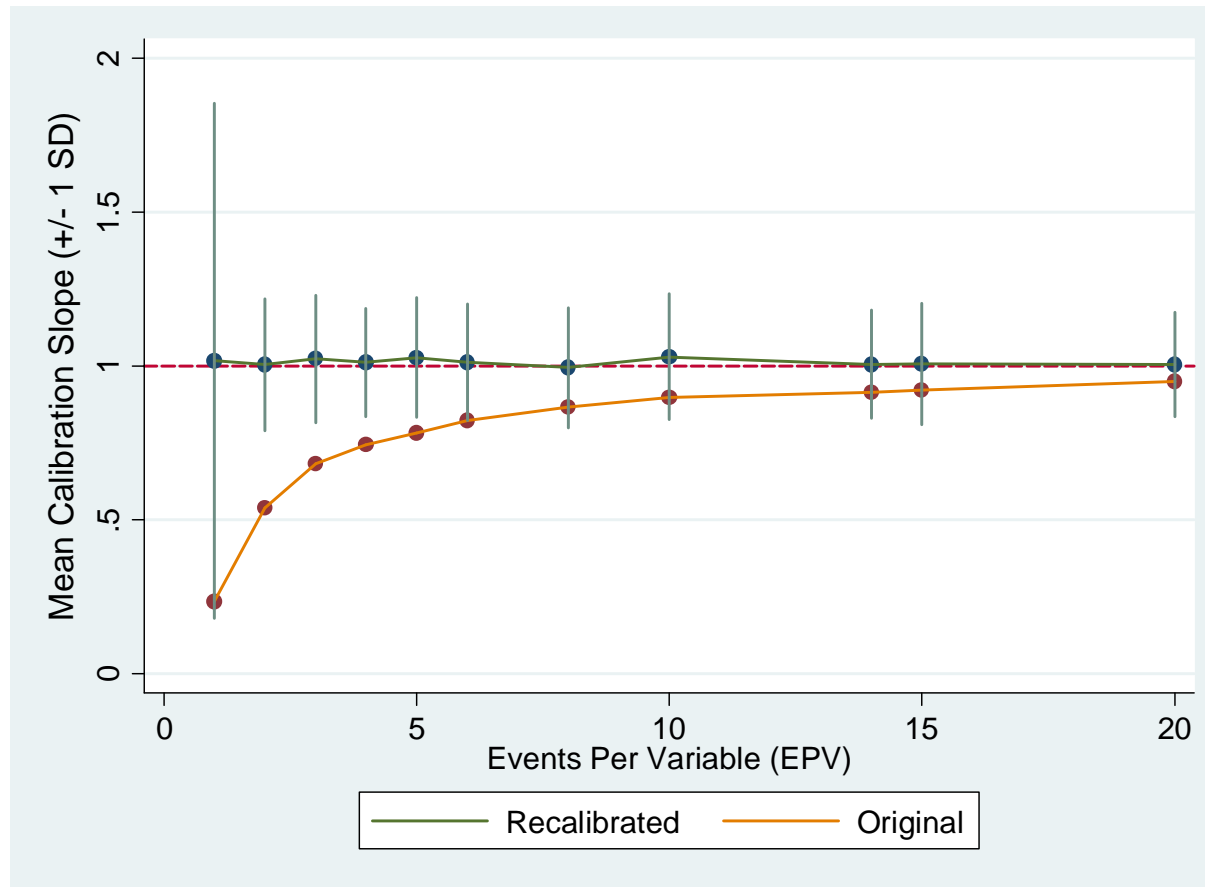


- The calibration is extremely poor for low EPV (< 6)
 - severe overfitting - range of predictions too extreme

Re-calibration

- A simple re-calibration approach was investigated to correct for over-fitting
 - extra ‘re-calibration’ datasets were also sampled
 - 200 observations (approx. 40 events)
 - the predicted log-odds was fitted as a single term to the re-calibration data to give the corrected log-odds (predictions)
 - $\log(p/1-p) = \alpha_0 + \alpha_1\eta$
- ROC area unaffected (same rank order)
- Other methods exist:
 - bootstrap, cross-validation, shrinkage (Steyerberg et al, 2001), penalisation (Ambler et al., 2002)

Results: re-calibration – calibration slope

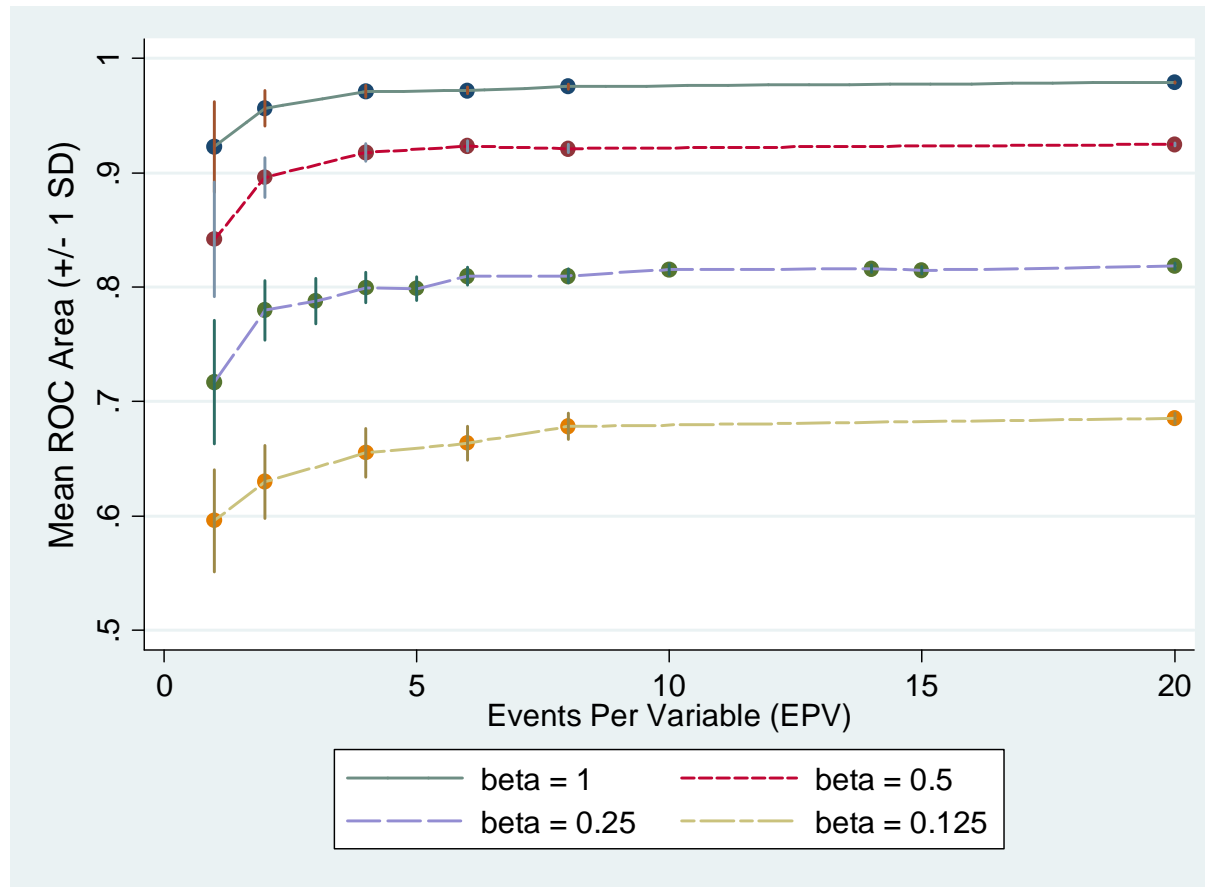


- This simple approach works extremely well, even for very low EPV
 - more variability: could increase size of re-calibration dataset

Methods: effect of model strength

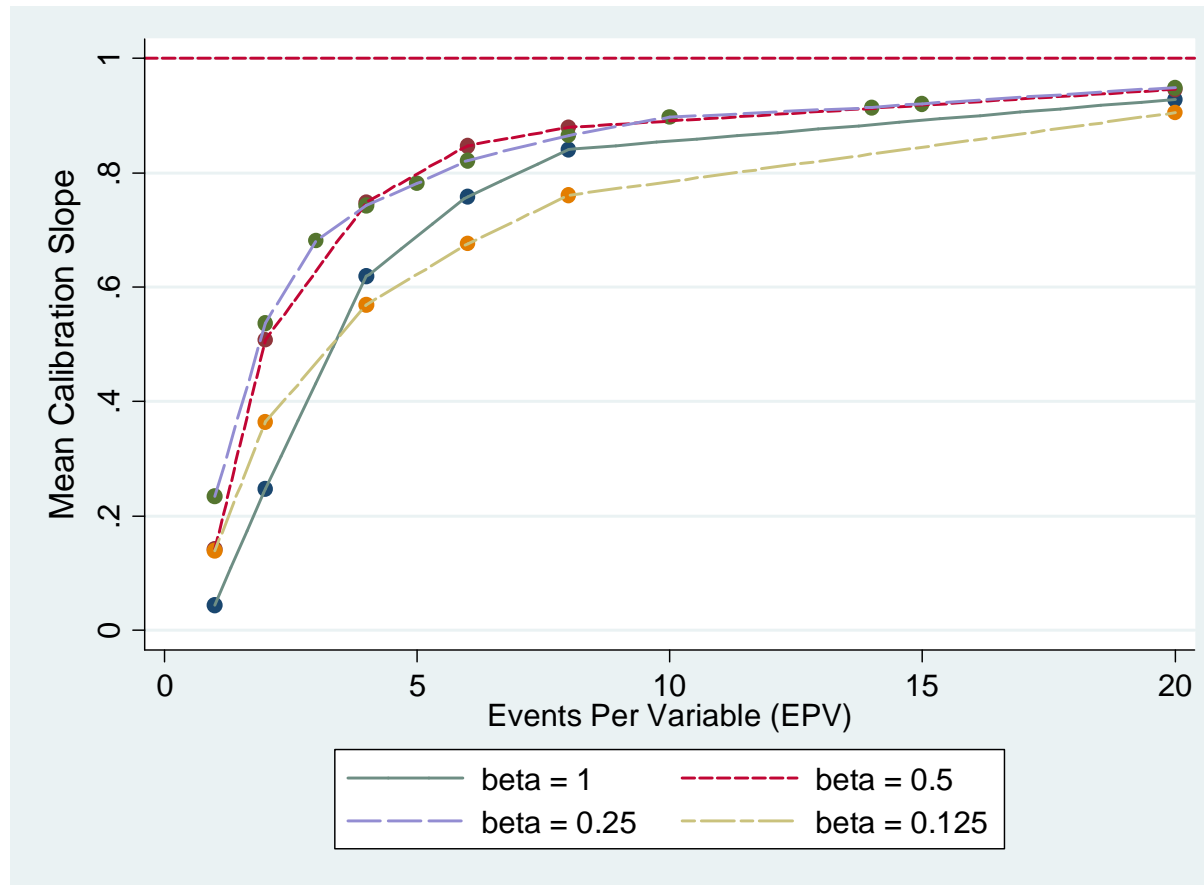
- Similar scenario to before:
 - true model: $\log(p/1-p) = \beta_0 + \beta_1 X_1 + \dots + \beta_{10} X_{10}$
 - logistic regression models fitted with all 10 predictors
 - β_j ($j > 1$) changed to give different ranges of predictions
 - different model strengths

Effect of model strength



- Different ROC areas indicative of different model strengths
 - ROC areas deteriorate at different EPV values

Effect of model strength

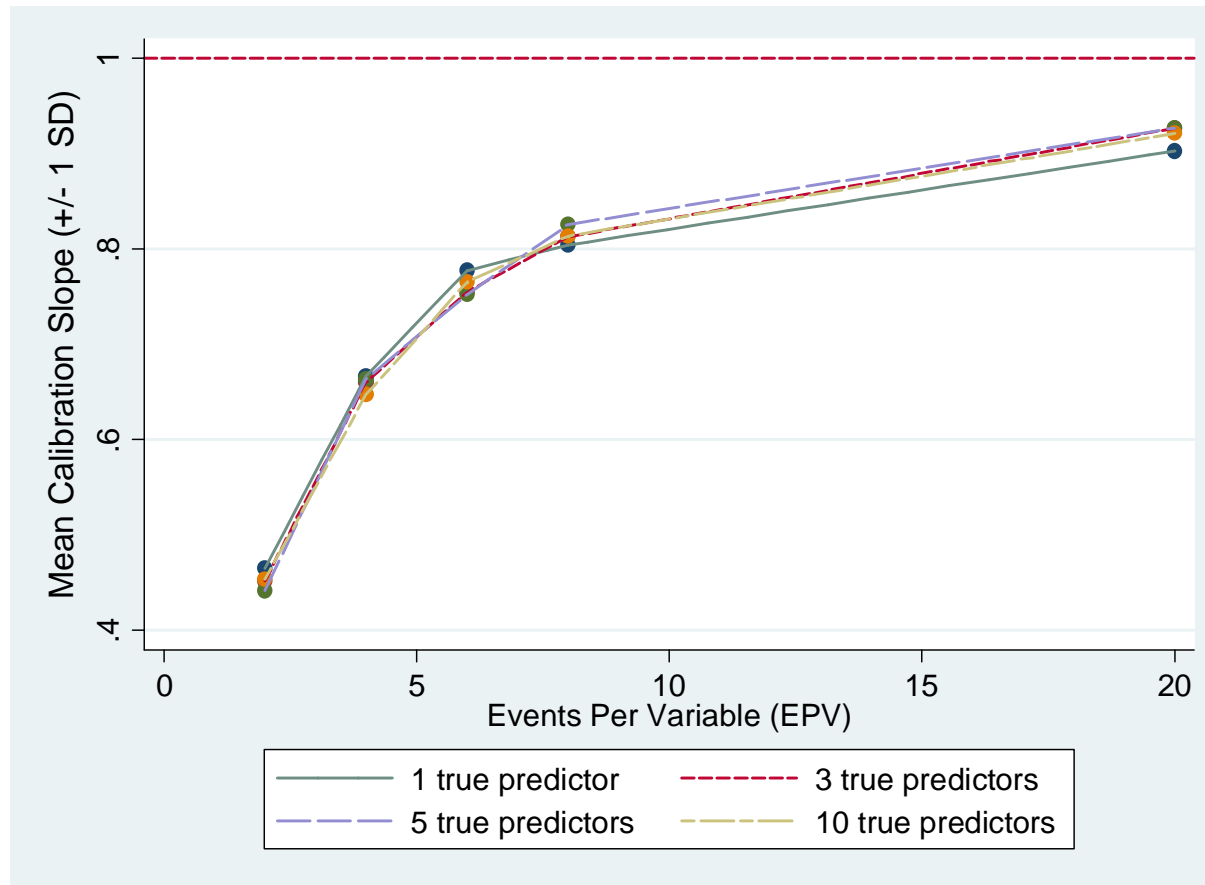


- Similar patterns observed
 - re-calibration was poor for weakest model (not shown)

Methods: effect of noise predictors

- Similar scenario to before:
 - true model: $\log(p/1-p) = \beta_0 + \beta_1 X_1 + \dots + \beta_q X_q$
 - $q = 1, 3, 5$ or 10
 - logistic regression models fitted with all 10 predictors
 - β_j s ($j > 1$) depend on q
 - set to give same range of predictions

Results: noise predictors - calibration slope



- Same pattern is observed regardless of number of true predictors
 - again, re-calibration works well (not shown)

Summary

- EPV is the major factor affecting risk model performance
- Useful risk models can be developed using datasets with low EPV values
 - re-calibration is essential
- Model strength may also be important
 - weakest models (less prognostic information) performed less well in low EPV scenarios
 - re-calibration problems
- Work in progress
 - other scenarios to consider