# Novel graphical diagnostics for assessing the fit of logistic regression models

Venkata Pavan Kumar Vajjah
Stephen Duffull

University of Otago, Dunedin, New Zealand

# Outline

- Introduction
- Motivating example
- Aim
- Novel graphical diagnostics
- Simulation study
- Discussion
- Conclusions

# Introduction

- Models for binary data
  - Logistic regression (LR) models are used to understand the relationship between the probability ($\pi$) of binary response variable (event or no event) and exploratory variable (dose ($D$))

$$\ln\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 \times D$$

# Introduction

- Model evaluation
  - "*Do the model's deficiencies have a noticeable effect on the substantive inferences?*" (Gelman, c 1995)

- Model diagnostics
  - Techniques used to examine the adequacy of a fitted model (Collett 1999)

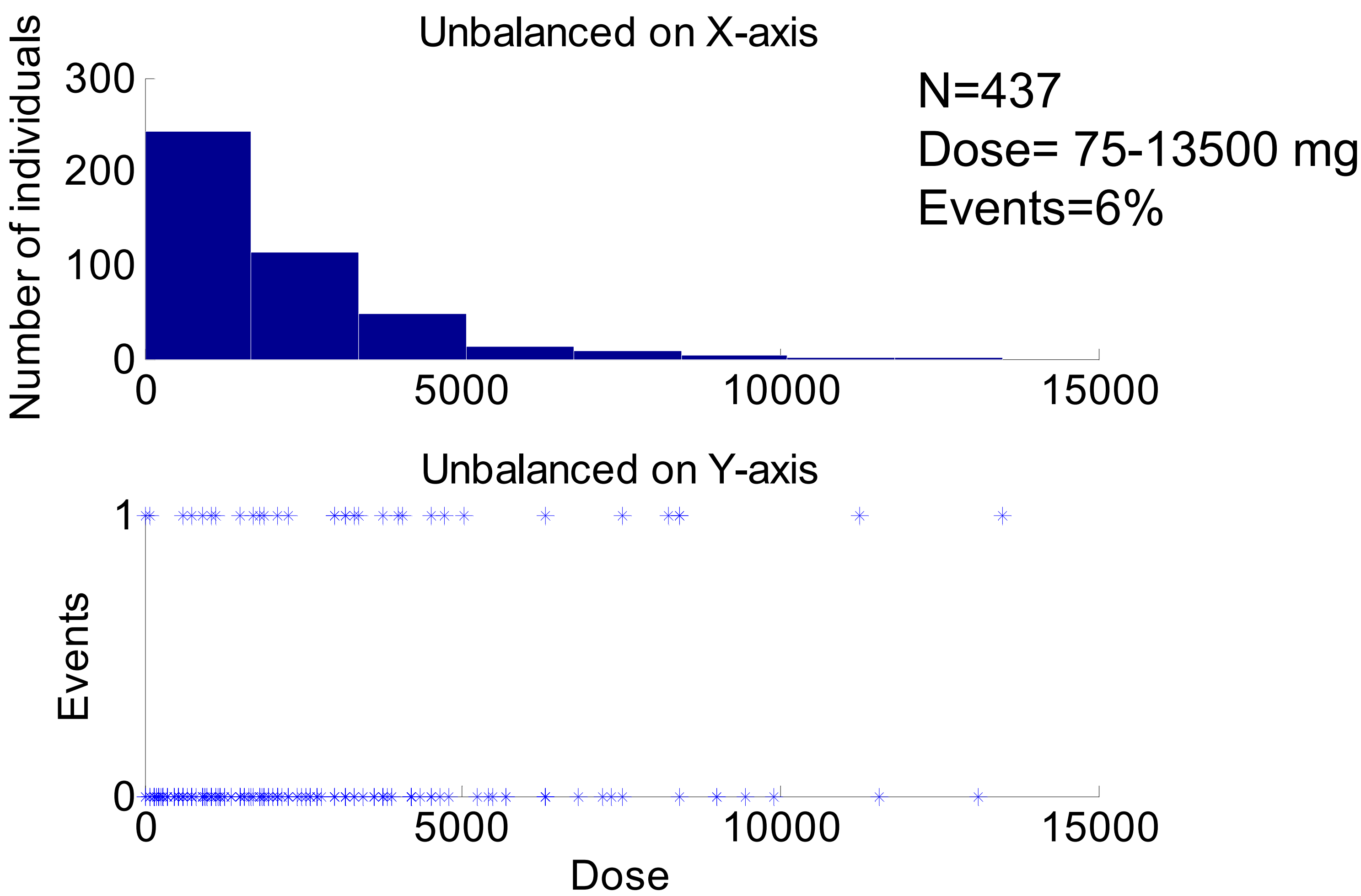
# Graphical diagnostics used in LR

- Simple binning
  - Grouping of measured data into data classes
    - Based on dose
    - Based on individuals
  - Estimate empirical probability and compare it with model predictions $(\hat{\pi})$

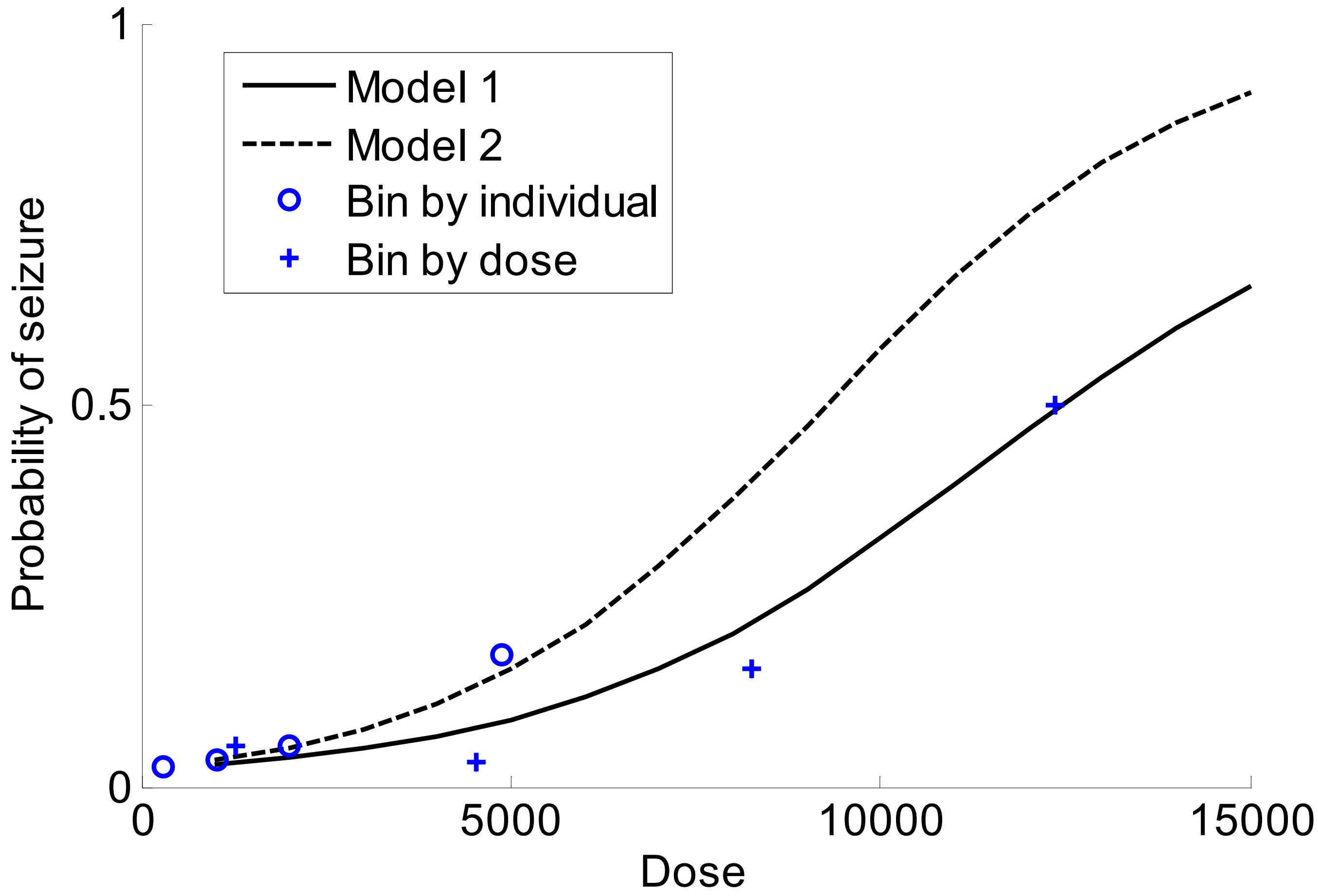
$$\tilde{\pi}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, where, i = 1,2,3....L; j = 1,2,3...n_i; y_{ij} \in \{0,1\}$$

$L$=Number of bins; $n_i$=number of individuals in $i^{th}$ *bin*

# Motivating example venlafaxine



# Simple binning as a "diagnostic"



# Aim

- To develop graphical diagnostics that are informative about fit of logistic regression model

# Model diagnostics

- Random binning
- Simplified Bayes Marginal Model Plots (SBMMP)

(Pardoe I, The American Statistician.2002, 56(4): 263-272)

# Random binning

- Generate a distribution of empirical probabilities of events $f\left(\tilde{\pi} | E, binning\right)$

- Compare empirical probability with model predictions $\left(\hat{\pi}\right)$

  Plot $f\left(\tilde{\pi} | y, binning\right)$ versus $D$

  overlay

  Plot $\hat{\pi}$ versus $D$

# Simplified Bayes Marginal Model Plots

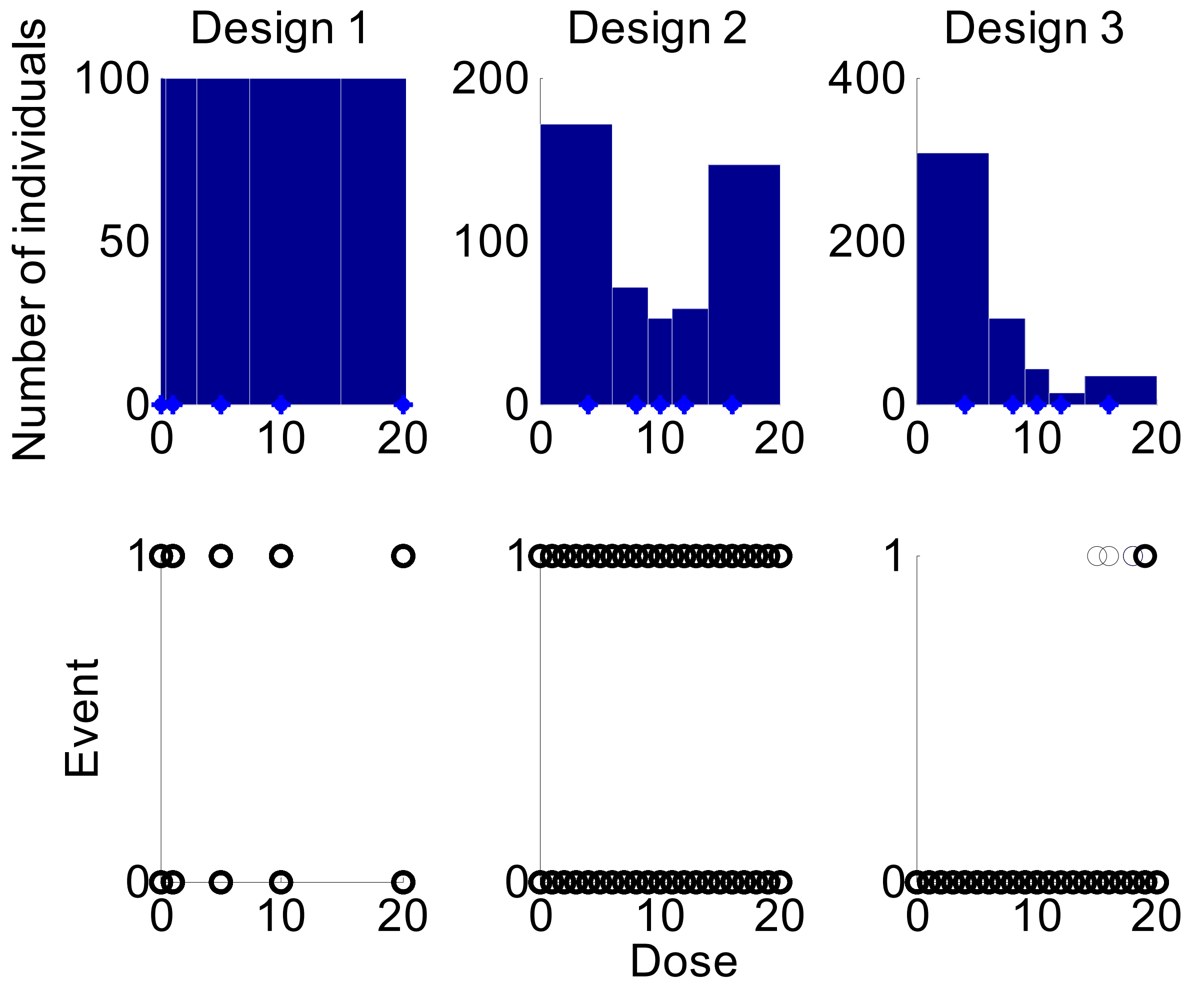
- Hypothesis: If the model describes data, then if we simulate ‘n’ observations, from posterior distribution of MODEL, SPLINE should be one of those observations

- Splines are believed to be the best possible empirical fit to the data

# Simulation study

- Simulation
  - Emax model (MATLAB)
- Estimation
  - Emax model (WinBUGS)
  - Linear model (WinBUGS)
- Evaluation (MATLAB)

# Simulation- Study design



# Simulation parameters

| | Design 1 | Design 2 | Design 3 |
|---|---|---|---|
| No.of simulations | 30 | 30 | 30 |
| No.of individuals | 500 | 500 | 500 |
| No.of dose levels | 5 | Random | Random |
| Doses | 0,1, 5, 10, 20 | Random between 0 & 20 | Random between 0 & 20 |
| No.of individuals/dose level | 100 | Random | Random |
| No.of events | 50%(approx) | 50%(approx) | 10%(approx) |
| $ED_{50}$ | 5 | 5 | 5 |
| $Pr(E_0)$ | 0.2 | 0.2 | 0.05 |
| $Pr(E_{max})$ | 0.9 | 0.9 | 0.825 |
| Variance$E_0$ | 0.025 | 0.025 | 0.025 |
| Variance$E_{max}$ | 0.025 | 0.025 | 0.025 |
| Variance$ED_{50}$ | 0.025 | 0.025 | 0.025 |

# Evaluation

1. Simple binning
   - Based on dose
   - Based on individuals
2. Random binning
   - Based on dose
   - Based on individuals
3. Simplified Bayes Marginal Model Plots (SBMMP)

# Random binning

- Number of bins = 5
- Minimum number of observations/bin=5

Step 1: Sort data by dose

Step 2: Generate 4 bin boundaries randomly based on dose or individuals

Step 3: Group data based on bin boundaries generated above

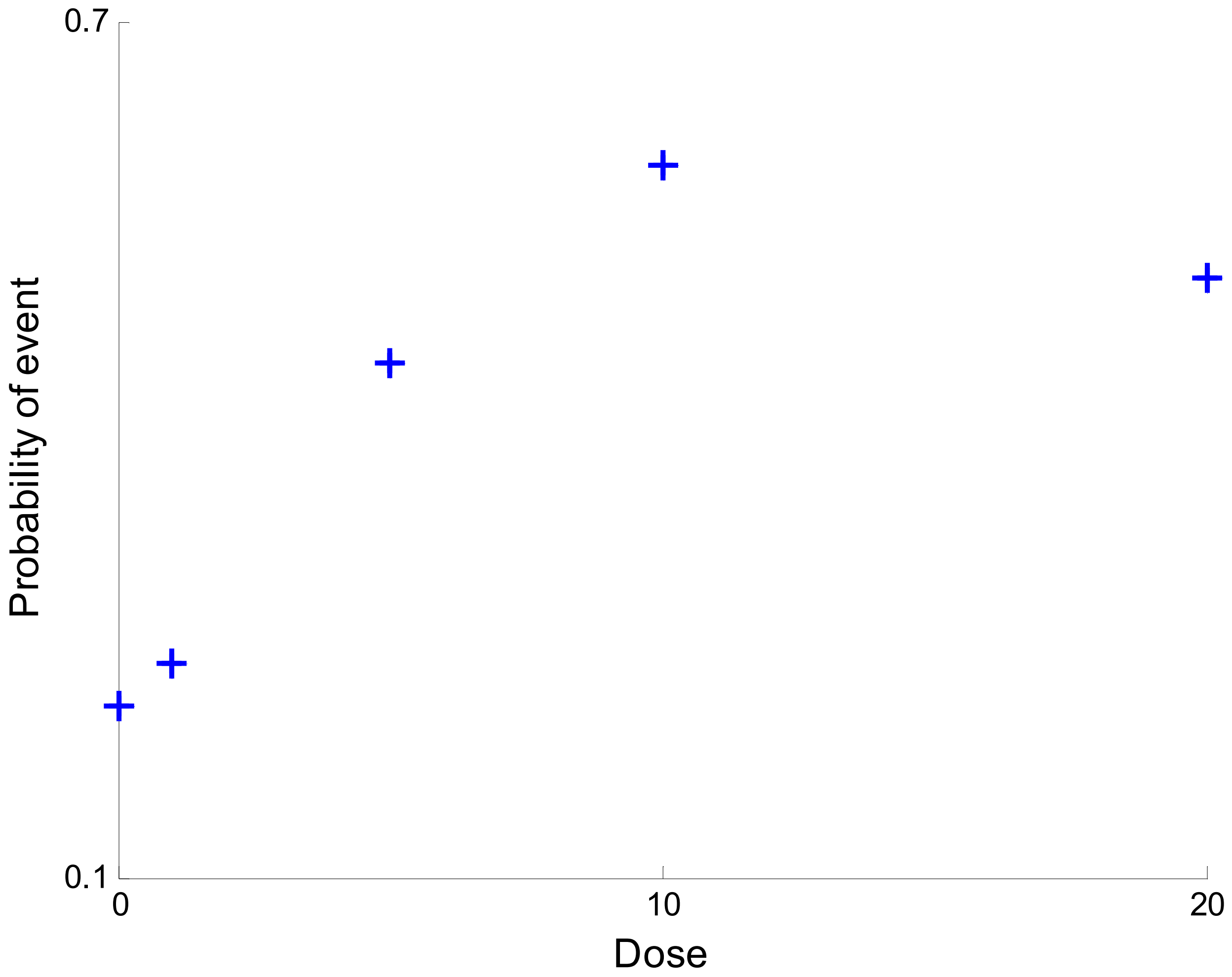
Step 4: Estimate $\tilde{\pi}$

Step 5: Repeat steps 2 - 4 1000 times

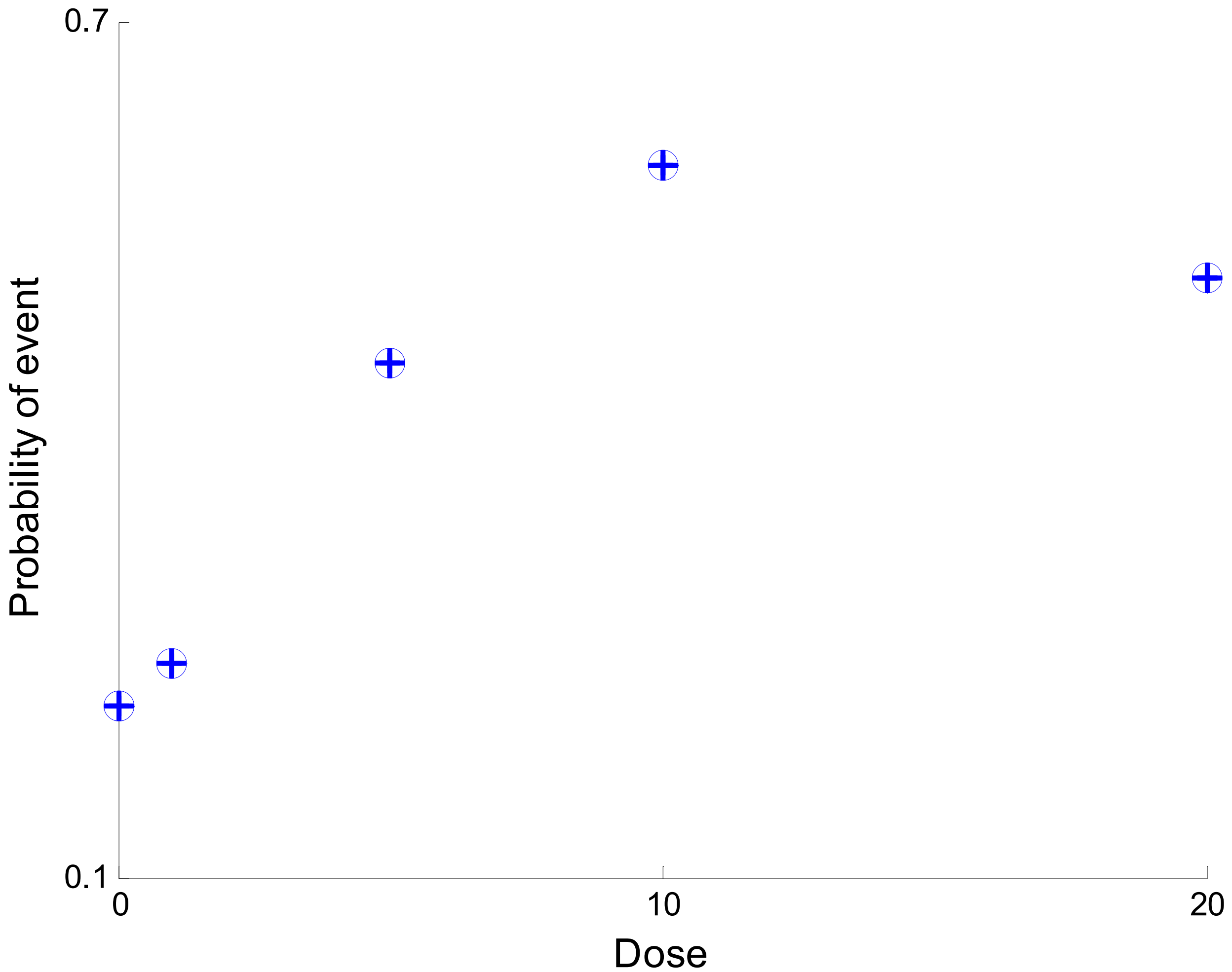
# SBMMP

- A linear spline was fitted to data with a maximum of 2 knots

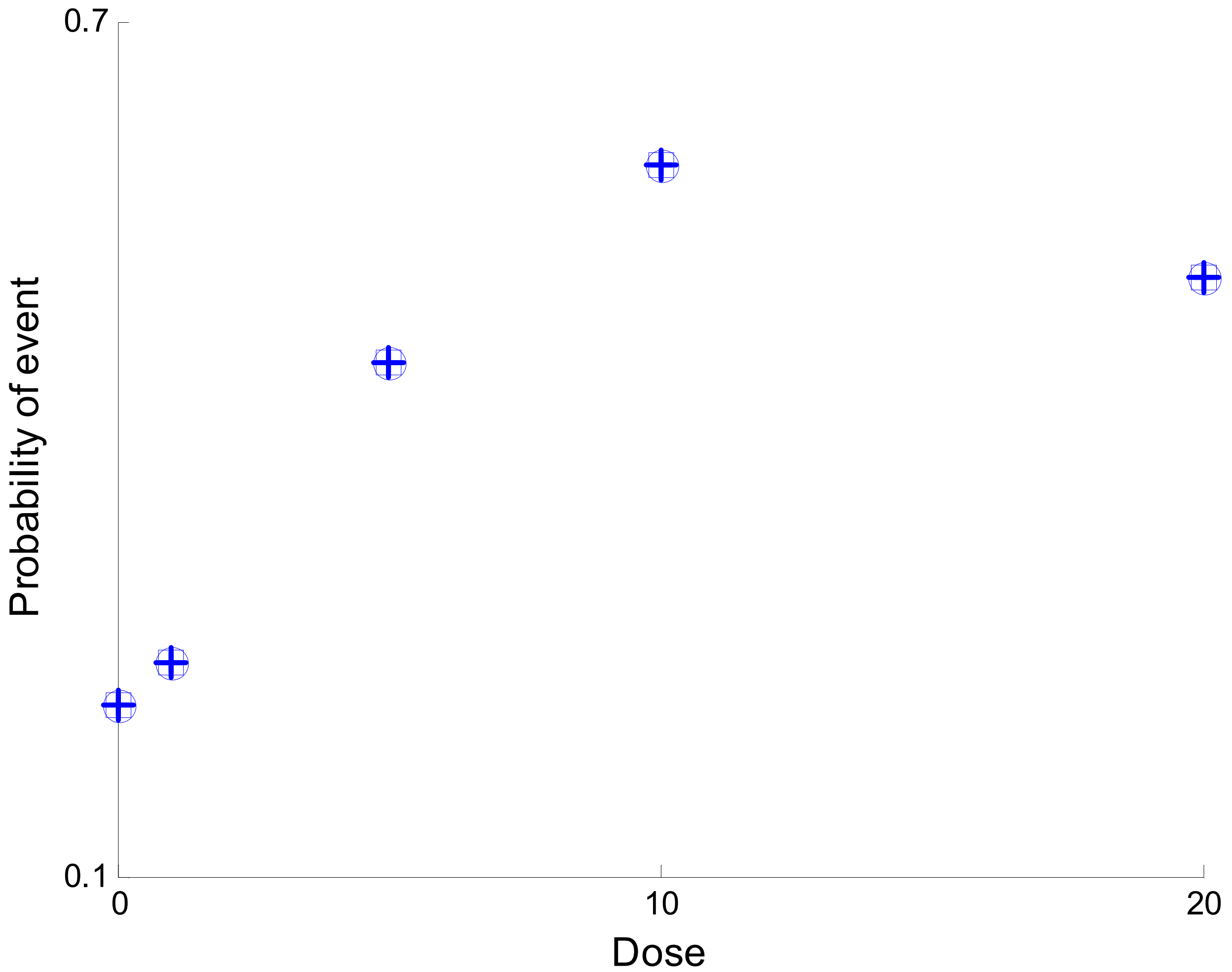
# Design 1
# Simple binning dose



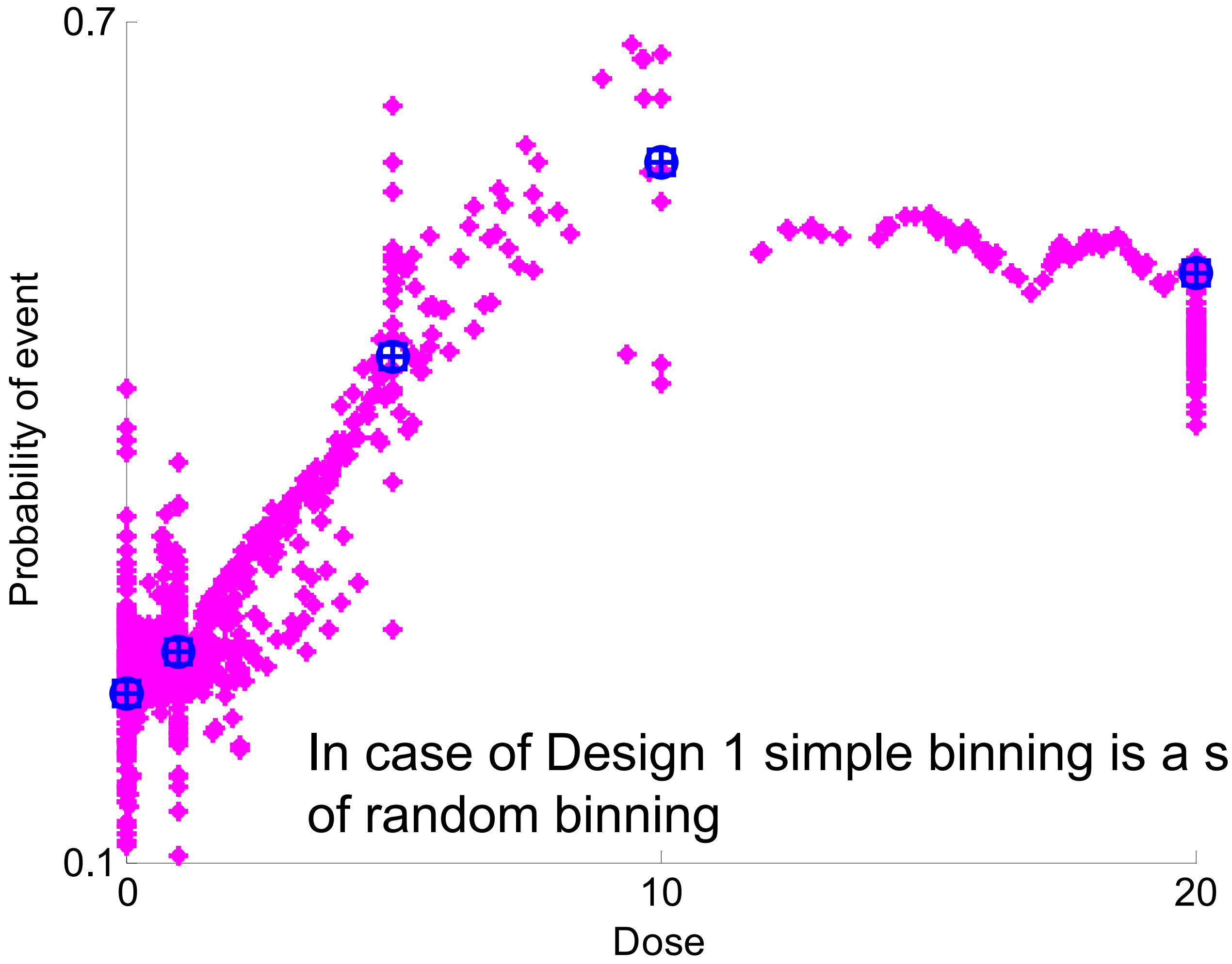
# Design 1
# Simple binning individuals



# Design 1
# Random binning dose

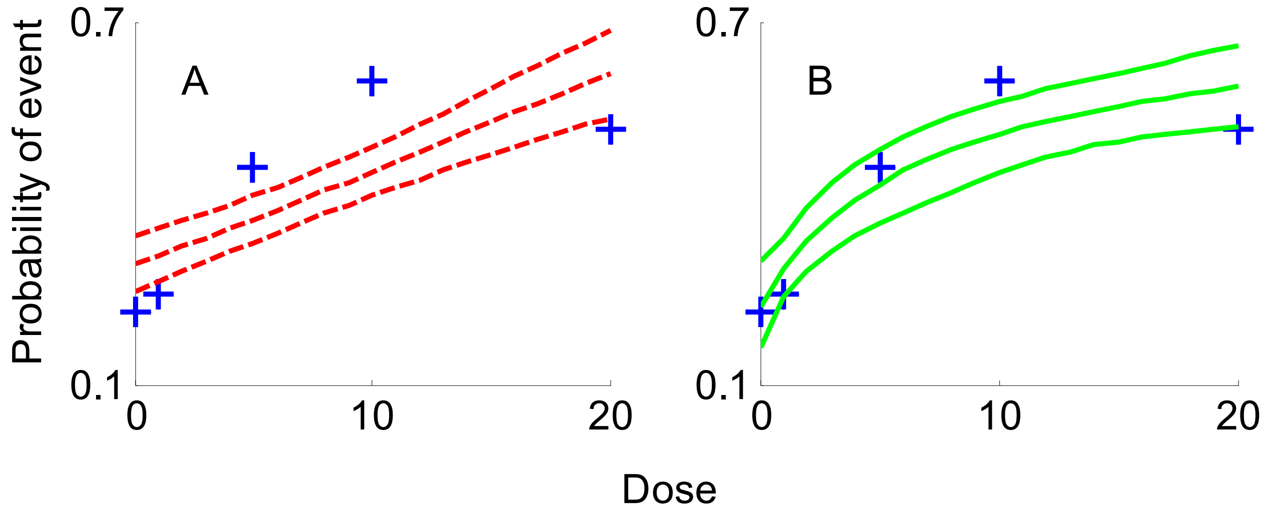


# Design 1
# Random binning individuals

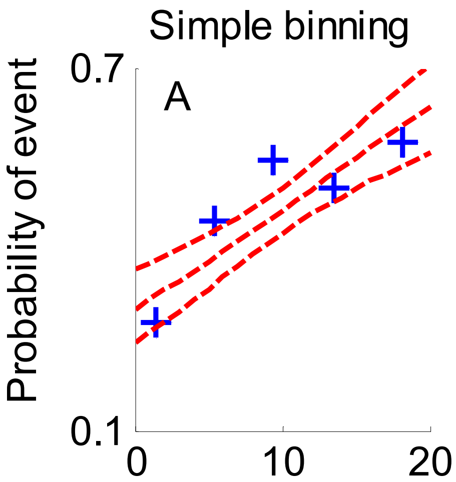



In case of Design 1 simple binning is a special case of random binning

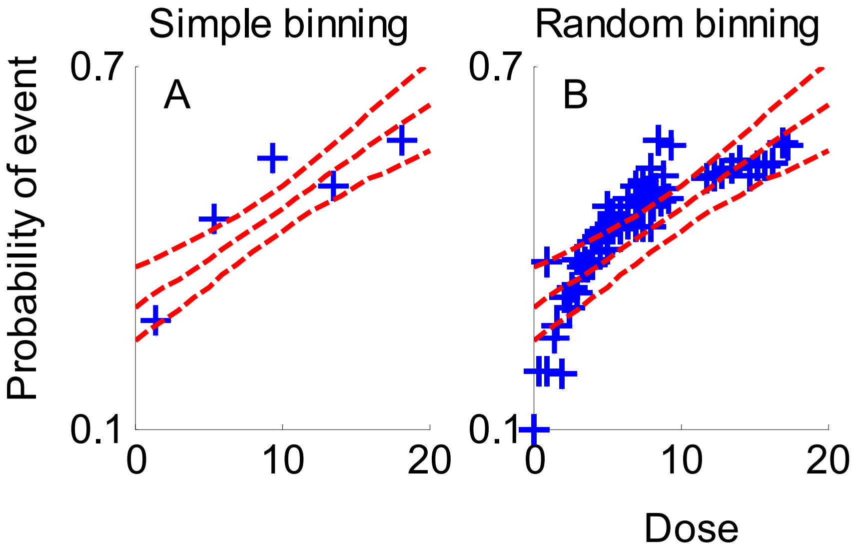
# Design 1
# Model evaluation



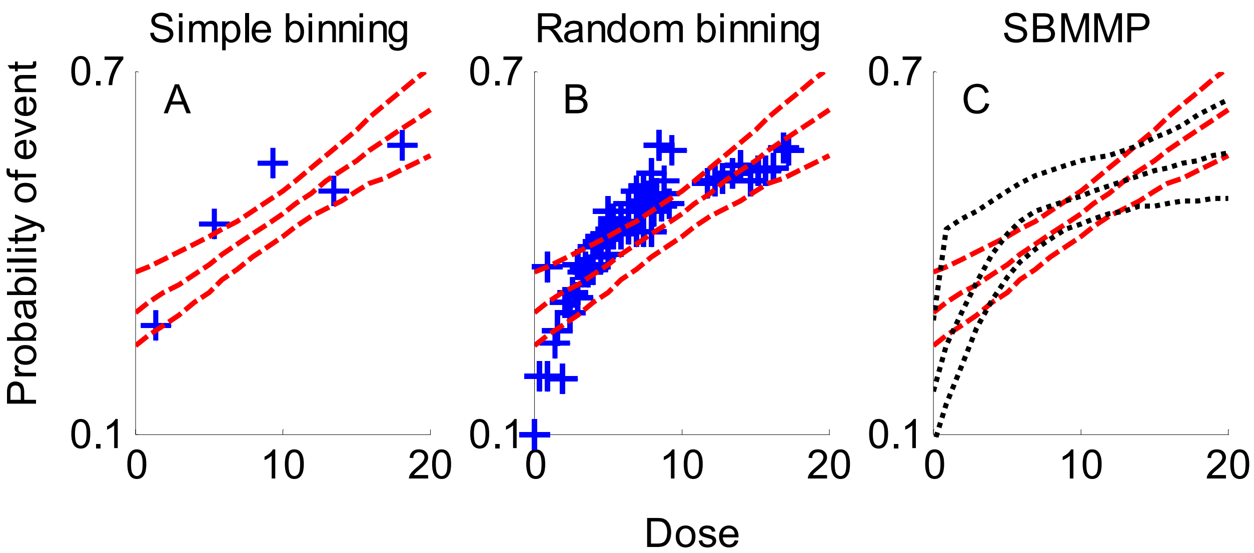
# Design 2
# May be correct model or may be not!!!



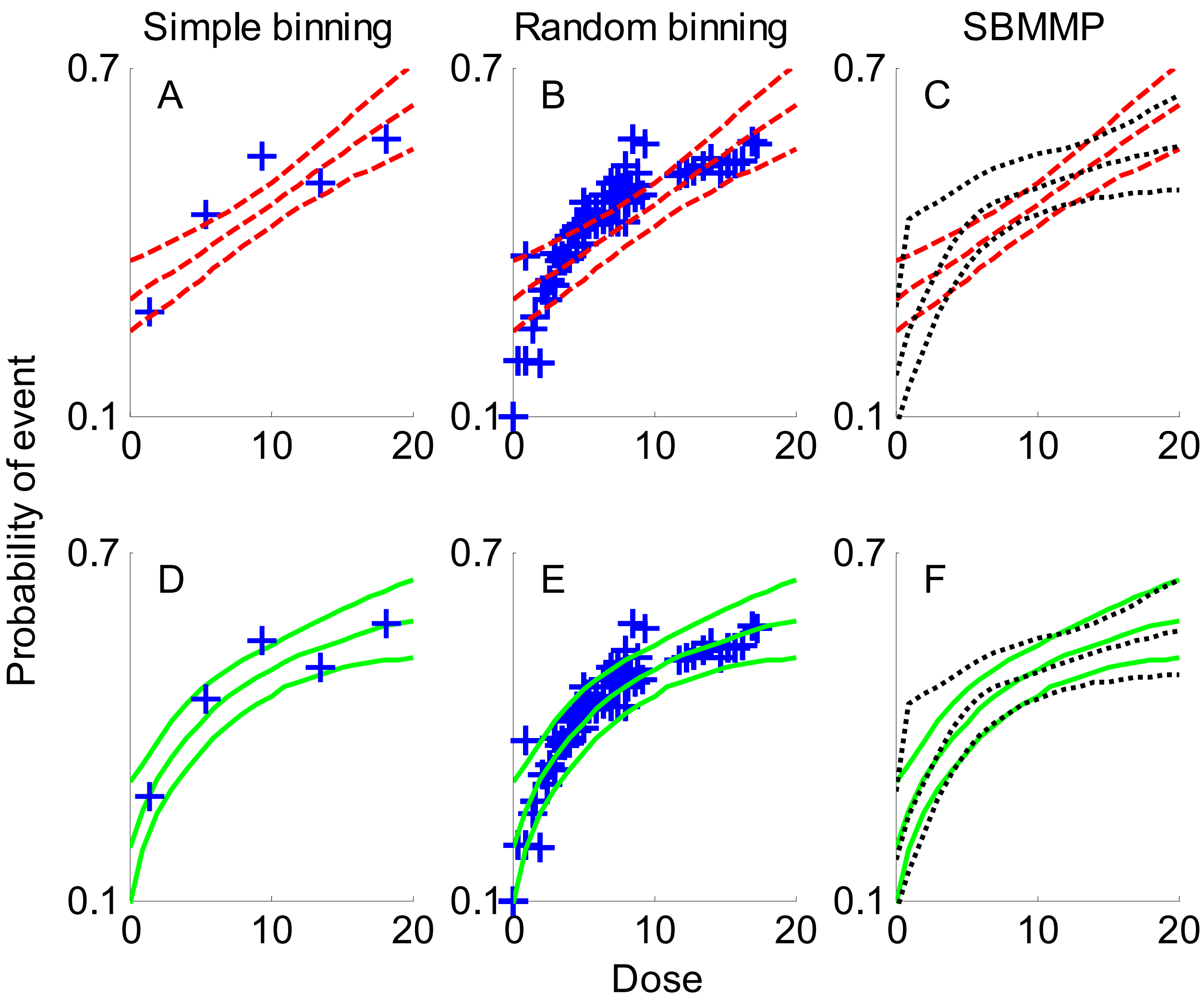
# Design 2
# May be not!!!



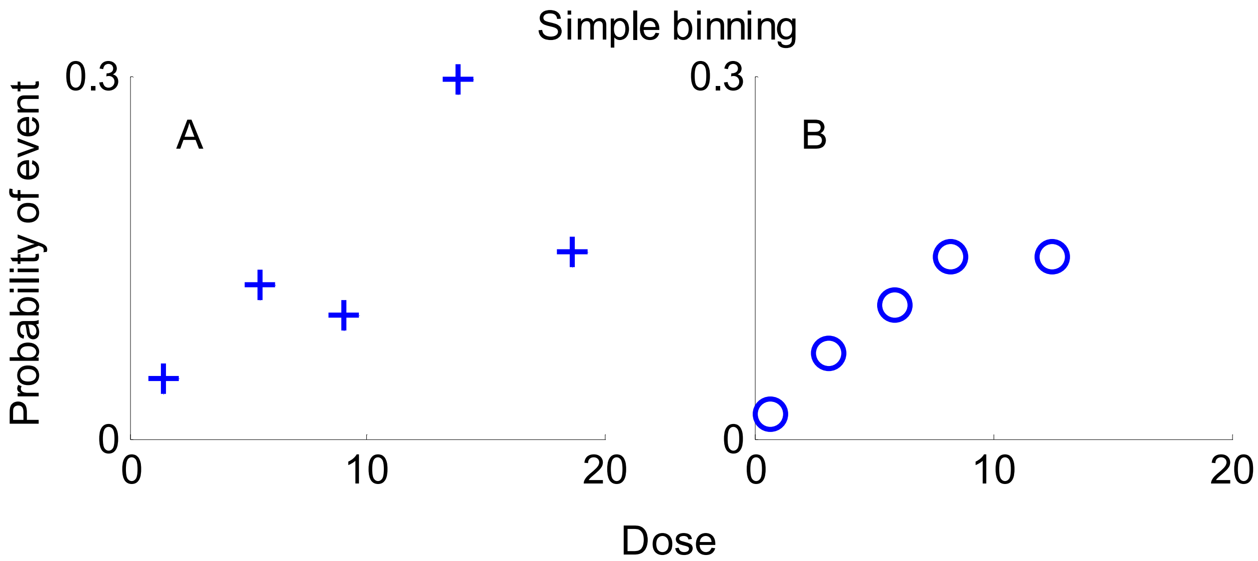
# Design 2
# This is wrong model!!!



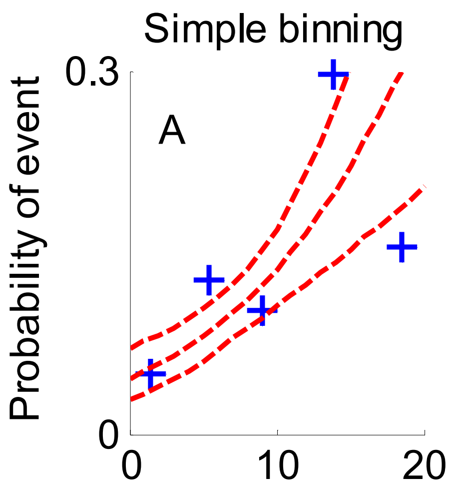
# Design 2
# How about the correct model?



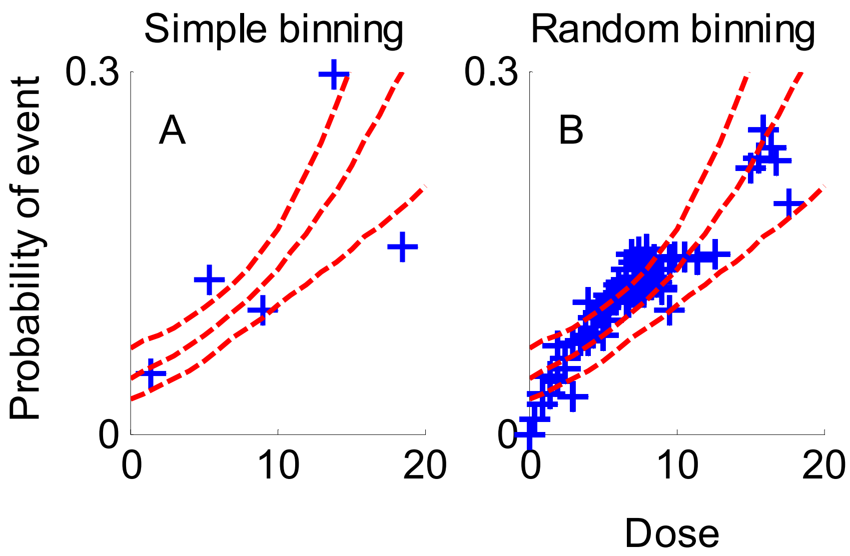
# Design 3
# Which binning method should I use??



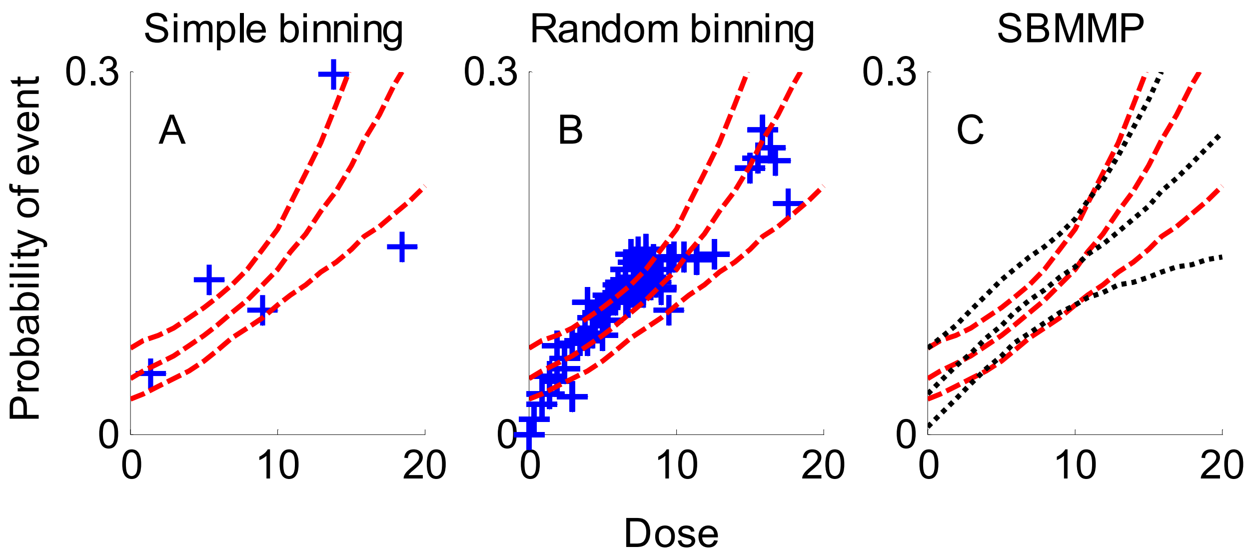
# Design 3
# Model describes data well!!!



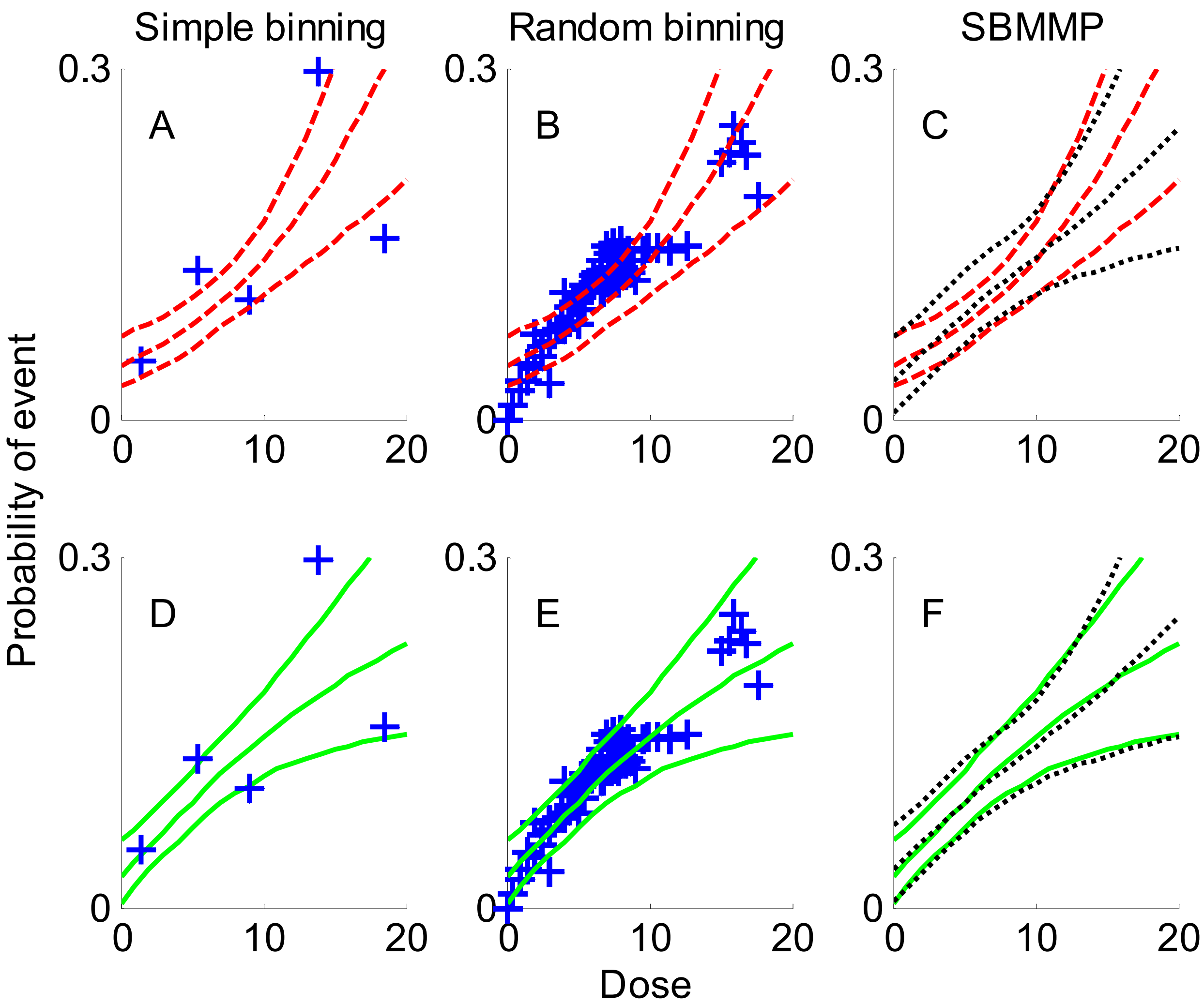
# Design 3
# May be not!!!!



# Design 3
# This is wrong model!!!



# Design 3
# How about correct model?



# Discussion

- Simple binning
  - Easy to do
  - Single realisation of a set of possible empirical probabilities, hence biased
  - Data is discrete
- Random binning
  - Random binning on average is unbiased, but adds noise
  - Data is discrete
- SBMMP
  - Additional model to be fitted to the data
  - The spline which represents the data is continuous
- Both the Random binning and SBMMP are computationally intensive

# Conclusions

- Simple binning is a useful diagnostic for completely balanced designs
- In case of unbalanced designs random binning acts as much better diagnostic than simple binning
- SBMMP is the best diagnostics studied here

# Acknowledgements


- Prof. Stephen Duffull
- Dr. Geoff Isbister
- Friends from M&S lab, University of Otago, New Zealand
- School of Pharmacy, University of Otago, New Zealand
- University of Otago postgraduate scholarship
- PAGE-Pharsight sponsorship

Thank you

# Why 30 simulations?

- Assumption
- Number of simulations required for 90% chance of getting the best and worst plot