Estimating and Presenting Nonlinear and Interaction Effects with Restricted Cubic Splines

Andrea Bellavia, PhD

TIMI Study Group, Brigham and Women's Hospital, Harvard Medical School

Department of Environmental Health, Harvard T.H. Chan School of Public Health

abellavia@bwh.harvard.edu

Harvard Catalyst, Biostatistics Journal Club, January 29 2025

1) Introduction: assumptions in regression modeling

   Linearity

   Additivity

2) Relaxing linearity: intro to restricted cubic splines (RCS)

3) Relaxing additivity

4) RCS framework to integrate non-linear and non-additive effects

5) Software material

1) Introduction: assumptions in regression modeling

Linearity

- Regression modeling is ubiquitous in clinical and epidemiological research to connect one or more covariates (ind. variables) and a certain health outcome (dep.)
- Basic idea: identify a functional form between dependent and independent variables
- Commonly, these functional forms involve linear relationships

Common examples

- Linear regression [linear on the expected value of a continuous outcome]

$$E[Y] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

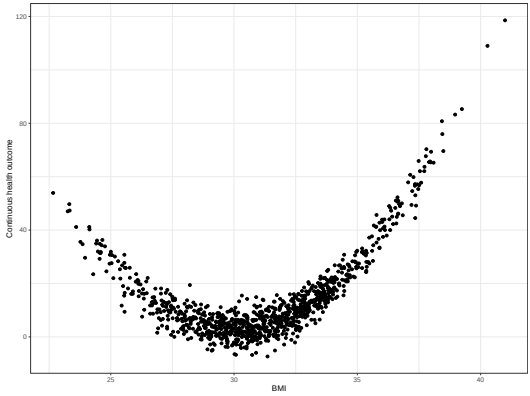
- Logistic regression [linear on the logarithm of the odds (logit) of a binary outcome]

$$logit(p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Cox regression [linear on the logarithm of hazard ratio for a time-to-event outcome]

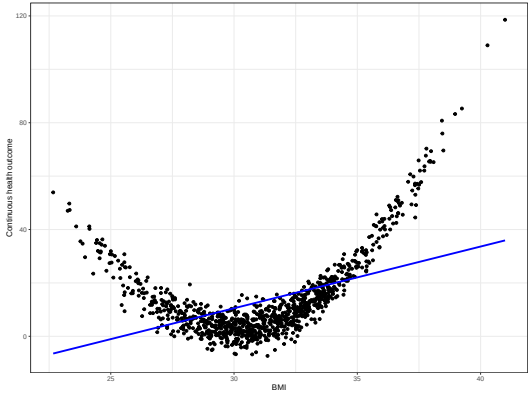
$$log(HR) = \beta_1 X_1 + \cdots + \beta_p X_p$$

- The use of linear functions implies that specific assumptions are made for continuous predictors
- Example: continuous predictor (e.g. BMI) and continuous outcome Y



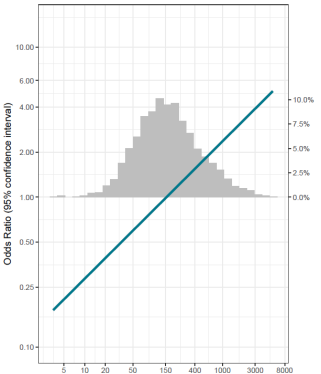
Linearity assumption in practice

- Fit a linear regression: $E[Y|\text{BMI}] = \beta_0 + \beta_1 \cdot \text{BMI}$. Result: $\hat{\beta}_1 = 2.3$
- Interpretation: difference in $E[Y]$ **for each** unit increase in BMI [blue line]
- The effect is the same when we compare BMI of 21 vs 20, 11 vs 10, 56 vs 55 etc.



Log-linear models (e.g. logistic, Cox)

- Linearity assumptions have slightly different interpretations in logistic and Cox model, which define linear assumptions on the logarithmic scale (log-linear models)
- Plotting figures *on the log-scale* is required to visualize linearity (and, later, potential departures)



Additivity

- Regression models make several additional assumptions (e.g. residuals normality, homoscedasticity, proportionality of the hazards ...)
- Additivity is another silent assumption with relevant implications on results' interpretation and translation
- Additivity assumptions are made for any combination of covariates included in a regression model

Additivity in practice

- AB pushes the car at 1 mph
- GF pushes the car at 2 mph
- How fast do they go when they push together?
- 3 mph: perfect additivity (assumption of a linear regression model)
- $>$3 mph: additive interaction
- $<$3 mph: negative interaction

Additivity: implications

The assumption of additivity between two covariates implies:

- Their joint effect equals the sum of the two main effects: absence of interaction
- The effects of each covariate are constant over levels of the other covariate: absence of effect modification

Note: in log-linear models, additivity assumption translates into a multiplicative assumption on the OR and HR scale [1]

[1] For more details see: VanderWeele TJ, Knol MJ. A tutorial on interaction. Epidemiologic methods. 2014 Dec 1;3(1):33-72.

2) Relaxing linearity: intro to restricted cubic splines (RCS)

2) Relaxing linearity: intro to restricted cubic splines (RCS)

A common approach to relax the linearity assumption is by creating a categorical version of the continuous covariate, included in regression models using dummy variables

Example: create 4 groups using quartiles of the distribution

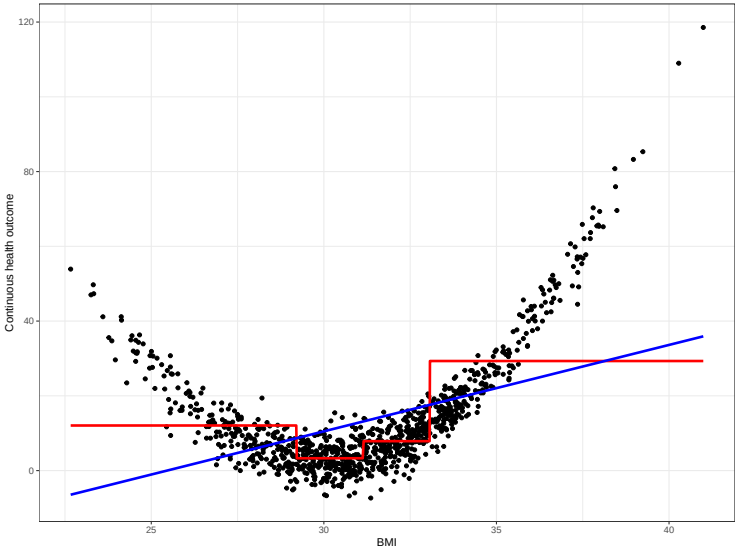
- Old (linear) model: $E[Y] = \beta_0 + \beta_1 X_1$
- New (categorical) model: $E[Y] = \beta_0 + \beta_1 X_{25th-50th} + \beta_2 X_{50th-75th} + \beta_3 X_{75th-100th}$

with $X_{25th-50th}, X_{50th-75th}$, and $X_{75th-100th} = (0,1)$

Table 1: Continuous and categorical version

x1	x1cat
2.7565865	3
3.1283981	4
0.6720901	2
-0.3604915	2
2.8176624	3
-4.4538679	1

Categorization in practice



We are replacing an assumption (linearity) with another assumption (step function)

- We now assume that the predicted response will be exactly the same for all individuals in the same subgroups
- We are also assuming that the change in the outcome will occur at specified (a priori and often subjectively) jumps

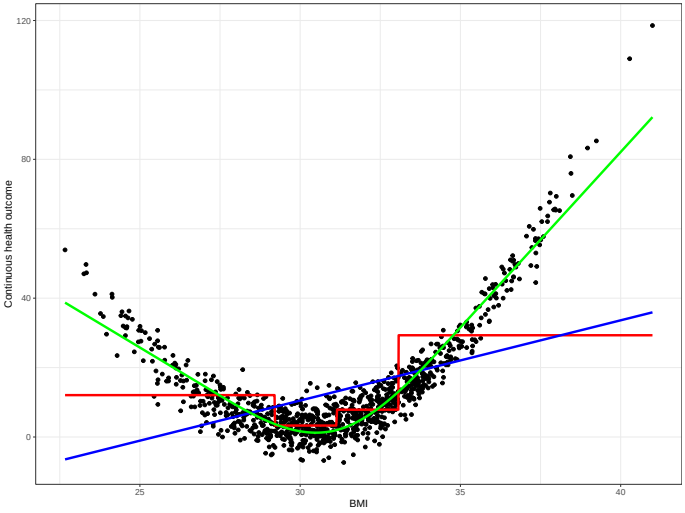
This assumption is often unrealistic.

Issues with categorization have long been recognized

- Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology 1995
- Greenland S. Problems in the Average-Risk Interpretation of Categorical Dose-Response Analyses. Epidemiology 1995.
- Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006

Splines

A more flexible solution: splines modeling



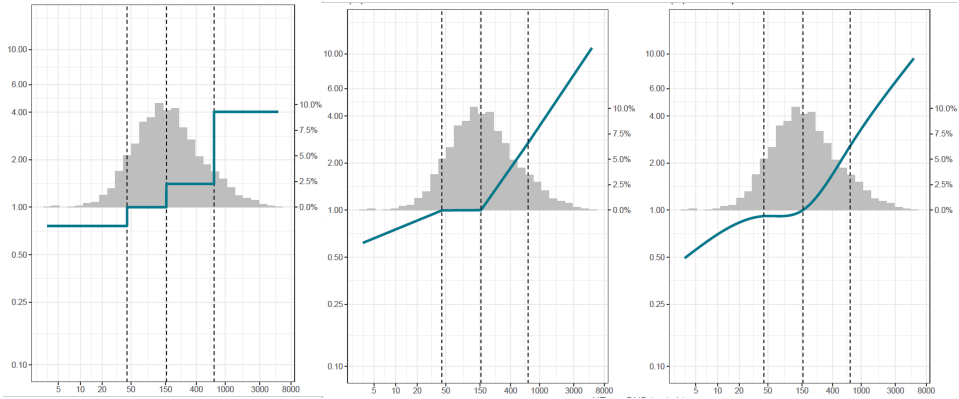
Splines transformations involve 2 steps:

- Select how many knots and where to place them
  - Conventionally at distribution percentiles. In general, 3 or 4 might suffice. Explore more with skewed distributions or if there is specific interest at the tails
- Select how to model in between knots

This interactive website provides a great tool to understand more the different assumptions and impact of knots numbers and locations: link

How to model in between knots?

- Categorical analysis is actually a particular case of splines modeling (degree 0 splines) where we assume constant outcome levels between knots
- Alternatively, we could use linear function that changes slope at each knot (degree 1 splines, aka piecewise modeling)
- Degree 3 (cubic) splines: between knots, the curve is a cubic polynomial, a smooth function of the form $\beta_0 + \beta_1 x + \beta_2 x + \beta_3 x$



Degree 0 (left, categorical approach), degree 1 (center, piecewise linear model, and degree 3 (right, cubic) splines for modeling the association between a continuous predictor and a binary outcome (OR scale)

Restricted Cubic Splines

- Restricted cubic splines add a constraint of linearity before the first and after the last knot
- Practically, the continuous covariate is replaced by $k-1$ new variables, where k is the number of knots. A key feature of RCS is that the first of these new variables coincide with the original covariate:

$$s_1 = x$$

$$s_i = \frac{(x - t_{i-1})_+^3 - (x - t_{n-1})_+^3 \frac{(t_n - t_{i-1})}{(t_n - t_{n-1})} + (x - t_n)_+^3 \frac{(t_n - t_{n-1})}{(t_n - t_{n-1})}}{(t_n - t_1)^2}$$

Table 2: RCS transformation with 3 knots. The original covariate (first column) is replaced by 3-1 new variables, of which the first once coincides with the original covariate

Original predictor	1st splines transform.	2nd splines transform.
2.7565865	2.7565865	2.7126776
3.1283981	3.1283981	3.2072301
0.6720901	0.6720901	0.6792160
-0.3604915	-0.3604915	0.2281211
2.8176624	2.8176624	2.7919021
-4.4538679	-4.4538679	0.0000000

Example of R implementation and output interpretation

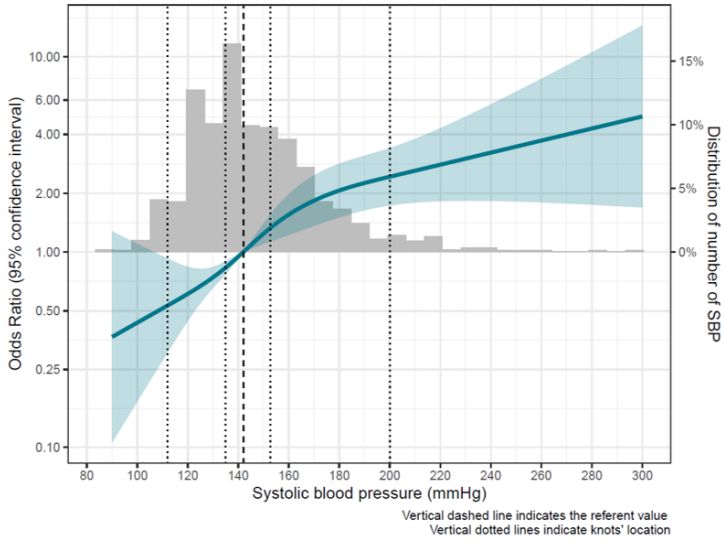
```
rcs<-glm(y~rcs(x1,3),data=cov)

round(summary(rcs)$coefficients,3)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.929      0.222  -8.694        0
## rcs(x1, 3)x1    -5.541      0.109 -50.912        0
## rcs(x1, 3)x1'   10.240      0.126  81.303        0
```

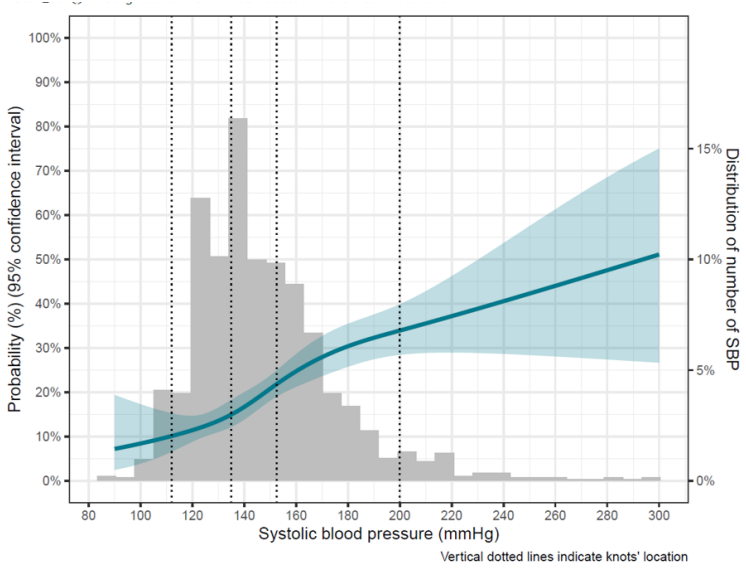
- Because of the interpretation of the first term (corresponding to the original X), we can interpret the second term as the *"non-linear term"*
  - A coefficient for this term close to 0 implies negligible departure from linearity
  - The p value can be used (with caution) as a statistical test for linearity

- The actual parameters, however, do not carry a clear interpretation.
- We need a graphical display to describe the non-linear association
- Results from the previous slide can be used to compute measures of interest and CIs across the continuous predictor. See section 5 for software material to reproduce these figures
- Example with logistic regression:



OR of CHD as a function of SBP, modeled with RCS (4 knots)

- With models such as logistic or Cox, we can either plot the OR (or HR) as a function of X, or model predictions such as the predicted probability (or absolute risk)
  - When presenting ORs, it is important to select a meaningful comparison point (e.g. the median.)
  - Note that, CIs for predictions should not cross, while CIs for ORs should cross at the reference value (OR=1)



Probability of CHD as a function of SBP, modeled with RCS (4 knots)

3) Relaxing additivity

3) Relaxing additivity

This is more straightforward: inclusion of a product term (aka interaction term) relaxes assumption of additivity and can be used to assess interaction or effect modification

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2$$

- β_3 can be used to test for interaction and/or effect modification
- Interpretation of the results should take into account the scale (i.e. additive interaction vs multiplicative interaction)

A note on overfitting

Do we always need complex (e.g. non-linear and non-additive) models?

- Key question: do we want our statistical model to describe as best as possible our data, or to be generalizable (i.e. to predict better)?
- Overfitting: the model perfectly fits our data but has limited generalizability
- Indexes like AIC/BIC are a better way to account for model complexity when comparing models. They might be preferred to conventional Wald tests in this context
- Pragmatically speaking, consider results' interpretation
  - For example, if the non-linearity occurs at a distribution tail with limited individuals, we are likely overfitting -> Always recommended to include a histogram or a rug plot under the splines plots

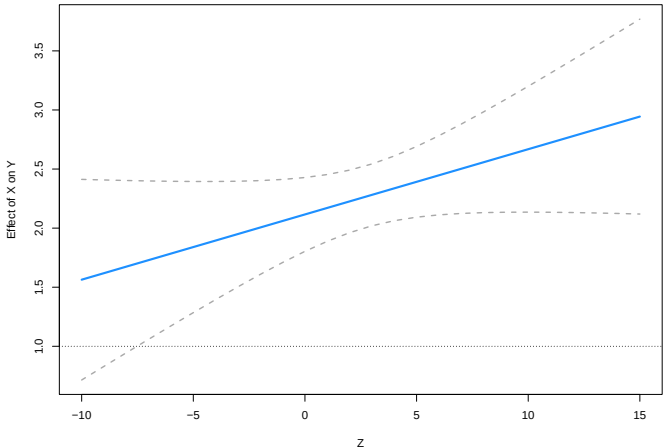
4) RCS framework to integrate non-linear and non-additive effects

4) RCS framework to integrate non-linear and non-additive effects

- Suppose we are conducting a study on a binary exposure X (e.g. TRT) and we want to study how the effects of X on Y vary over levels of a continuous variable Z (e.g. age)
- We include an interaction (product) term between X and Z

$$\log(HR) = \beta_1 x + \beta_2 z + \beta_3 z \cdot x$$

- The linearity assumption made for Z is extended to the interaction term. With the product term we are relaxing the assumption that the effect of X is constant over levels of Z, but we are now assuming that it changes (log-)linearly



Assumption of a model with interaction term involving a continuous predictor. The effect of X on Y changes (log-)linearly over levels of Z

RCS in an interaction model

- Let's address the problem with RCS. We transform Z using the RCS transformation from slide 22. With 3 knots:

z -> $sp_1(z) + sp_2(z)$, where

$sp_1(z) = z$, and $sp_2(z) = [\frac{(z-t_1)_+^3 - (z-t_2)_+^3 \cdot \frac{t_3-t_1}{t_3-t_2} + (z-t_3)_+^3 \cdot \frac{t_2-t_1}{t_3-t_2}}{(t_3-t_1)^2}]$

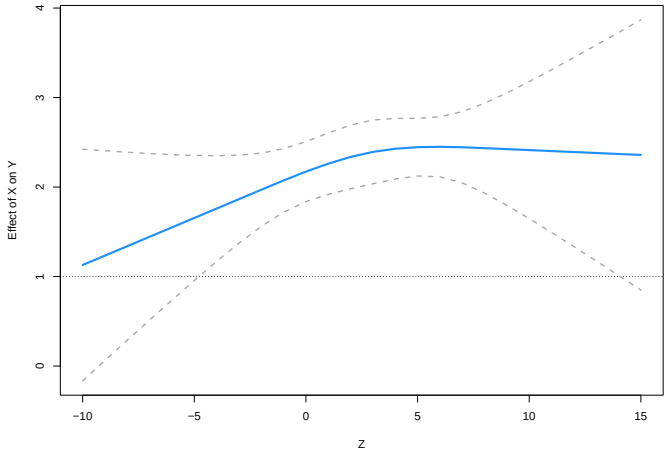
and we fit the model

$$log(HR) = \beta_1 x + \beta_2 z + \beta_3 sp_2(z) + \beta_4 x \cdot z + \beta_5 x \cdot sp_2(z)$$

- terms for non-linear Z
- terms for non-linear $X \cdot Z$ interaction

- From this model, we can predict the OR for X over levels of Z [2]
- And derive the SE by using either the delta method or bootstrap
- R package `interactionRCS` to derive graphical displays of non-linear interaction from the models
  - Incorporates linear, logistic, and Cox
  - Online vignette (link) also includes formulas for more than 3 knots

[2] details in 2024 AJE paper

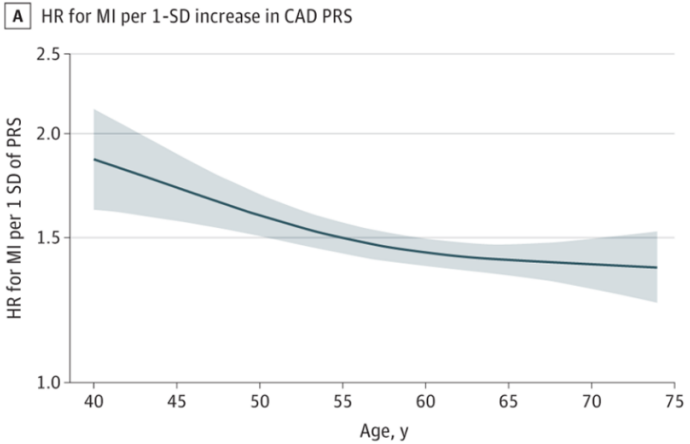


Example of figure from interactionRCS

Example from real data

From Marston et al, JAMA cardiology 2022

Figure 1. Hazard Ratio (HR) for Myocardial Infarction (MI) as a Function of Age at Baseline



Model predictions

- As we can predict event probabilities and absolute risks from a splines model, so we can predict non-linear interactions on these additional scales
- These will provide indication of interaction on the (additive) absolute risk/probability scale, to complement those on the (multiplicative) OR/HR scale.

Example: smoking effect of the risk of MACE over levels of hemoglobin (2024 AJE paper)

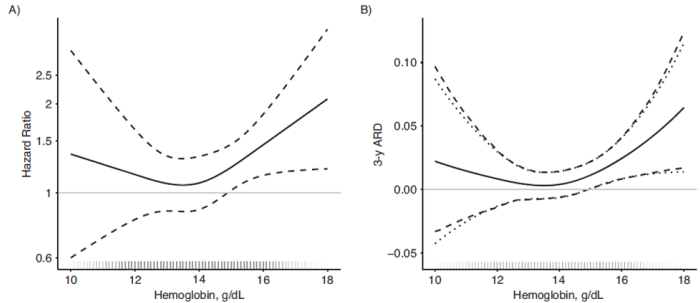


Figure 1. Hazard ratio for major adverse cardiovascular events (A) and 3-year absolute risk difference (ARD) (B) between current smokers and never smokers according to hemoglobin level, flexibly modeling the interaction with restricted cubic splines. Data were obtained from a randomized, double-blind, multinational, placebo-controlled phase 3 trial of patients with type 2 diabetes and established atherosclerotic cardiovascular disease or multiple risk factors for atherosclerotic cardiovascular disease. This analysis focused on a subsample of 16 694 individuals with complete data on covariates of interest. Tick marks inside the x-axis indicate the distribution of hemoglobin levels in the population, with lower opacity at a lower frequency. Ninety-five percent CIs for ARDs were calculated with both an SE-based approach (dotted lines) and a bootstrap (dashed lines).

Figure 1 from Bellavia et al., AJE 2024

1) Point estimates derived through individual predictions from the splines-interaction model (consider a grid based on covariates values to improve computational speed)
2) Estimate confidence intervals for predictions

- Bootstrap
- Based on SE estimated from the Cox model

See R code (link later in the slides) for implementation of these steps

5) Software material

5) Software material

We [3] have developed code and tutorials for all steps discussed (https://github.com/andreabellavia/RCSplines)

The page includes:

- 1) Code and tutorials in SAS, Stata, and R, for splines modeling and graphical presentations for linear, logistic, and Cox (including model predictions)
- 2) R material for non-linear interactions
  - R package `interactionRCS` documentation
  - Code and tutorial for absolute risk predictions

[3] Credits to: Andrea Discacciati, Giorgio Melloni, Michael Palazzolo, Jeong-Gun Park, Hong Xiong

Thanks for your attention!

Contact:

abellavia@bwh.harvard.edu

andreabellavia.github.io

timi.org/biostatistics/

Acknowledgments:

Andrea Discacciati, Sabina Murphy, Giorgio Melloni, Michael Palazzolo, Jeong-Gun Park, Hong Xiong

Disclaimer: AB is a member of the TIMI Study Group which has received institutional research grant support through Brigham and Women's Hospital from: Abbott, Abiomed, Inc., Amgen, Anthos Therapeutics, ARCA Biopharma, Inc., AstraZeneca, Boehringer Ingelheim, Daiichi-Sankyo, Ionis Pharmaceuticals, Inc., Janssen Research and Development, LLC, MedImmune, Merck, Novartis, Pfizer, Regeneron Pharmaceuticals, Inc., Roche, Saghmos Therapeutics, Inc., Siemens Healthcare Diagnostics, Inc., Softcell Medical Limited, The Medicines Company, Verve Therapeutics, Inc., Zora Biosciences