International initiative: STRengthening Analytical Thinking for Observational Studies (STRATOS)

UNRESOLVED ISSUES IN MODELING OF FUNCTIONAL FORMS FOR CONTINUOUS VARIABLES IN MULTIVARIABLE ANALYSES

Michal Abrahamowicz*

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada.

*** for the STRATOS Topic Group TG2**

(Selection of variables and functional forms in multivariable analysis):

Heiko Becher, Harald Binder, Frank E. Harrell, Georg Heinze,

Matthias Schmid, Aris Perperoglou, Patrick Royston, & Willi Sauerbrei

Overview

- Goals of STRATOS Topic Group 2 (TG2)
- Importance of Flexible Modeling of Non-Linear (NL) associations of Continuous Variables with the Outcome:
  - Drawbacks of Categorization
  - (selected) *'Smoothers'* to estimate NL associations
  - *Empirical Examples* of important, Clinically Plausible NL associations
  - Reasons for, and Implications of *Bias due to imposing Linearity A Priori*

- Comparisons between different Smoothers:
  - Summary of (very *few)* published *Simulation* studies
  - Selected empirical comparisons in *Real-life analyses*

- Survival analysis: NL vs Time-Dependent (non-PH) effects

- Impact of Modeling of Continuous Variables on Variable Selection
- (selected) Variable Selection approaches
- Summary of recent Simulations re: Variable Selection methods
- CONCLUSIONS:

  Un-resolved issues that need to be addressed by TG2

MAIN ISSUES ADDRESSED BY TG2

- TG2 focuses on **2 inter-related questions,** common to all multivariable *explanatory** models:
  - Selection of 'relevant' Variables (*independently*** related to the outcome)
  - Choice of the Functional Forms for Continuous Variables. (modeling of the independent (adjusted) associations of each Continuous Variable with the Outcome)

** In *explanatory* models (as opposed to prediction), we:

- try to determine 'which variables have true' associations, and
- Not to include 'spurious' variables

(even if they may improve prediction of the outcome)

DRAWBACKS OF CATEGORIZATION OF CONTINUOUS PREDICTORS

- Review of Recent Literature indicates that **CATEGORIZATION of Continuous Variables is still Very Common in Both Clinical & Epidemiological research**

- Yet, **Several Drawbacks of Categorization were demonstrated [1]**:
  - Implausibility of the Step-Function effect & 'Local Bias' [2]
  - Arbitrary cut-offs for categories often vary wildly across studies of the same predictor-outcome association [3], inducing spurious differences
  - 'Bad' *A Priori* selection of cut-offs results in worse fit to data and increased Type II error
  - If cut-offs selected *A Posteriori:* standard Inference is Not valid, increased risk of Type I error and overfit bias [4]

1] Royston et al. *Stat Med* 2006, 25: 127-141.

[2] Sauerbrei et al. *Br J Cancer* 1999; 79: 1752-1760.

[3] Malats et al. *Lancet Oncology* 2005, 6:678-686.

[4] Schulgen et al. *AJE* 1994, 140(2): 172-184.

FUNCTIONAL FORM FOR CONTINUOUS INDEPENDENT VARIABLE

- To understand the role of Continuous Predictor (X) in an Explanatory Model (for a given outcome), **we need to estimate the etiologically correct' Dose-Response function *g(x)*** (a continuous, *smooth* transformation of X)

- **Conventional models usually *A Priori* assume that: (a) either g(x) is Linear** & include Un-transformed X: **g(x) = βx, or (b) g(x) is a specific conventional simple parametric transformations [e.g. ln(x) or exp(x)]**

- <u>Linearity assumption is convenient</u> (effect of X summarized by a single ß, parsimony = improved power), <u>and often adequate</u>

- **Yet, Linearity should not be imposed *a priori*:**

  there are numerous examples of systematically Non-Linear or even Non-Monotone associations, e.g.:

  - **BMI → all-causes mortality**: both Obese and 'Skinny' subjects have Increased Risks)
  - **Age at diagnosis → mortality in different cancers**:

    Young age at Diagnosis -> more aggressive disease,

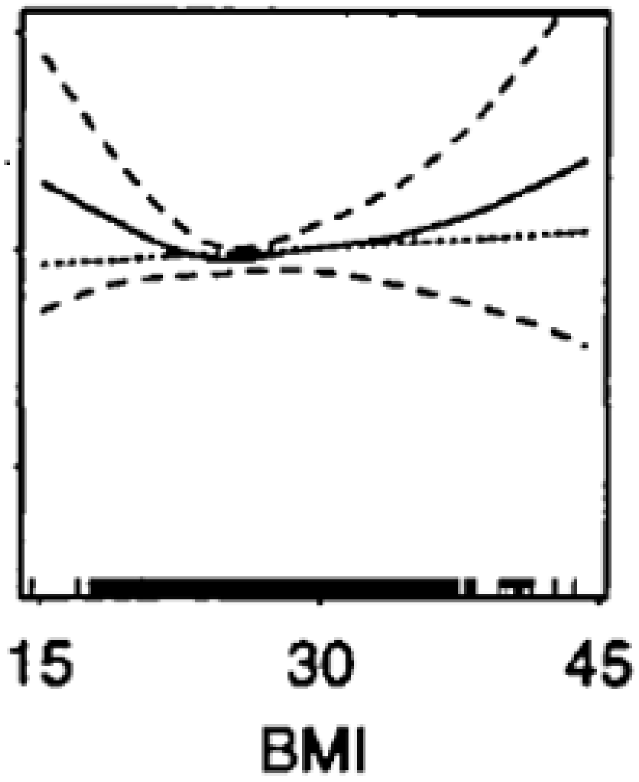
    Old age -> increased risk of all-cause mortality, worse tolerance of treatment

OPTIONS FOR FLEXIBLE MODELLING OF THE FUNCTIONAL FORM FOR CONTINUOUS VARIABLES

- **Flexible Modeling techniques, proposed to estimate Non-linear (NL) effects of Continuous Variables, with different Smoothers, include e.g.:**
  - Fractional Polynomials (FP) [Royston & Sauerbrei2008; Royston & Altman 1994]
  - (un-penalized) Regression Splines [Ramsay 1988; Abrahamowicz & MacKenzie 2007]
  - Restricted Cubic Splines [Harrell (2001; 2015)]
  - Penalized Smoothing Splines [Gray JASA 1992]
  - Generalized Additive Models (GAM) [Hastie & Tibshirani, 1990]
  - ... + several other types of Splines (I- , P- ... – splines)

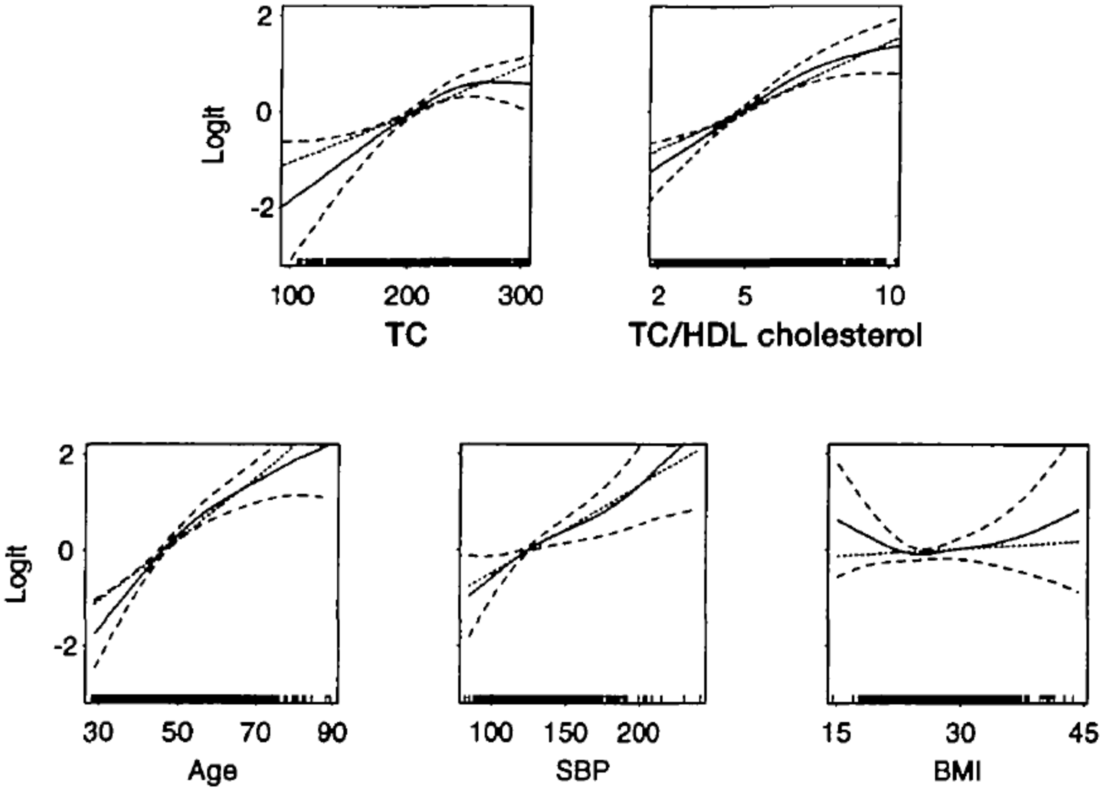
Smoothing spline estimate of non-linear effect on BMI on the logit of the probability of Coronary Heart Disease (CHD) death

[Abrahamowicz *et al*, *Am J Epidemiology (AJE)* 1997]



Smoothing splines (GAM) estimates of non-linear effects of conventional CHD risk factors on the logit of Coronary Heart Disease (CHD) mortality

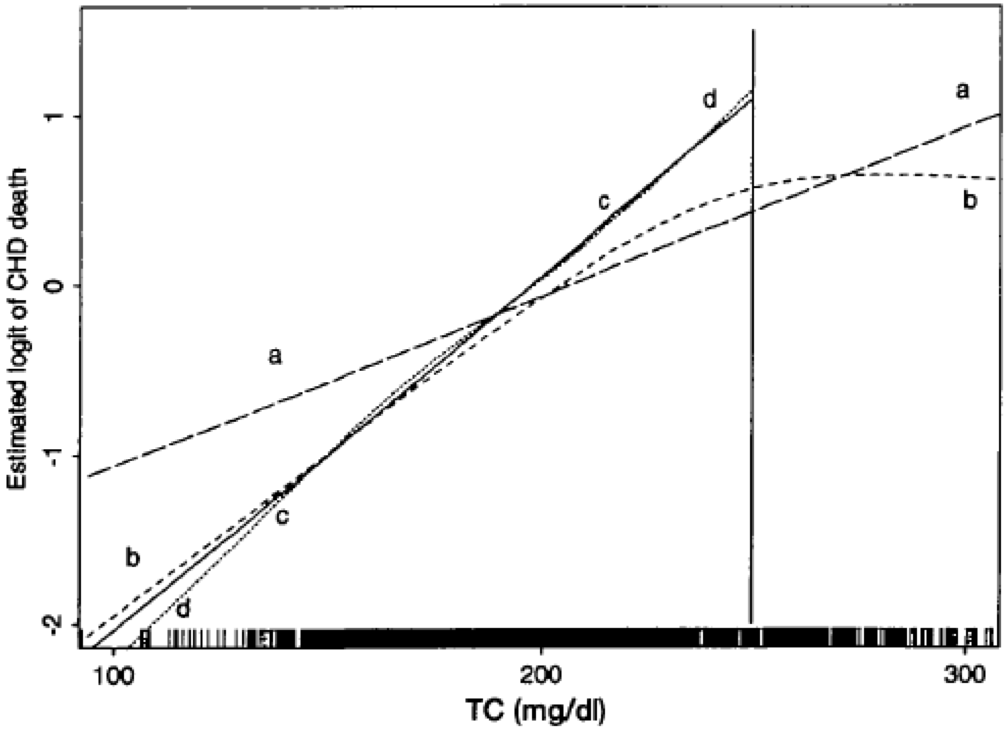
[Abrahamowicz *et al*, *Am J Epidemiology (AJE)* 1997]



Flexible modelling of continuous variable avoids "local biases" of a linear function

Cholesterol (X) vs. Logit of prob. of CHD Death (Y)

[Abrahamowicz *et al*, *AJE* 1997]



- (a) & (b): full range of X; (c) & (d) X<250; (a) & (c) linear (ßx);
- (b) & (d) Smoothing Spline (GAM)

Estimated reduction in CHD death risk (due to cholesterol lowering interventions) Depends strongly on Linear (L) vs NL/GAM (G) estimation

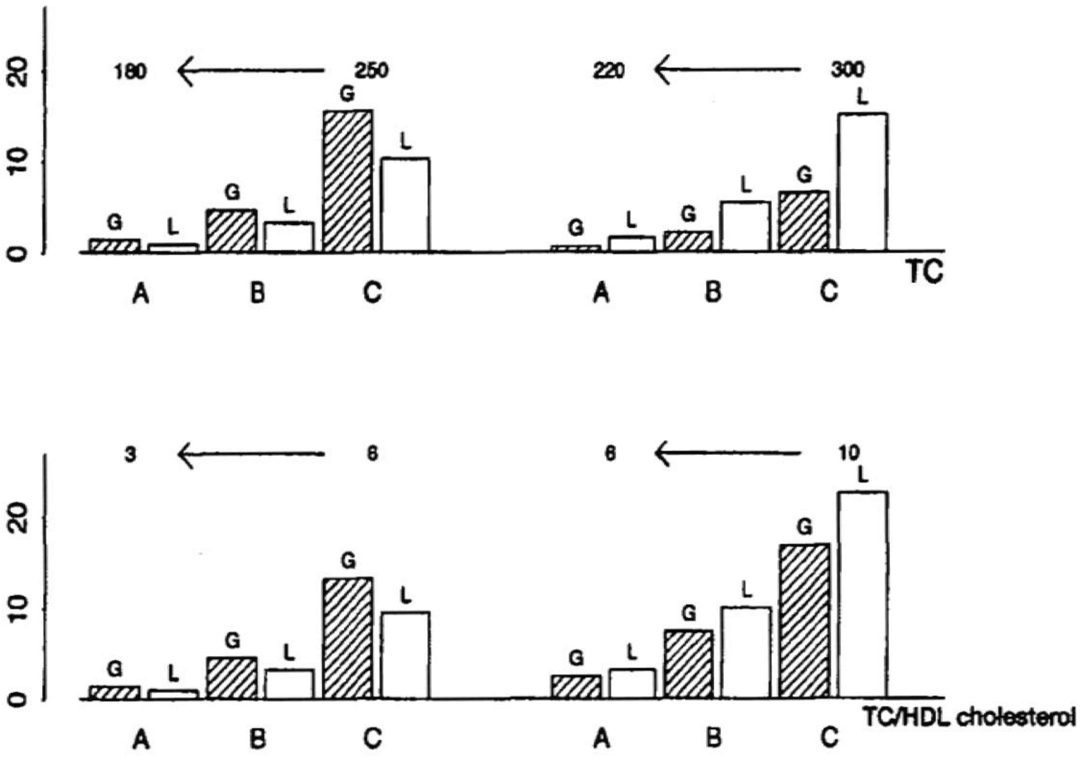
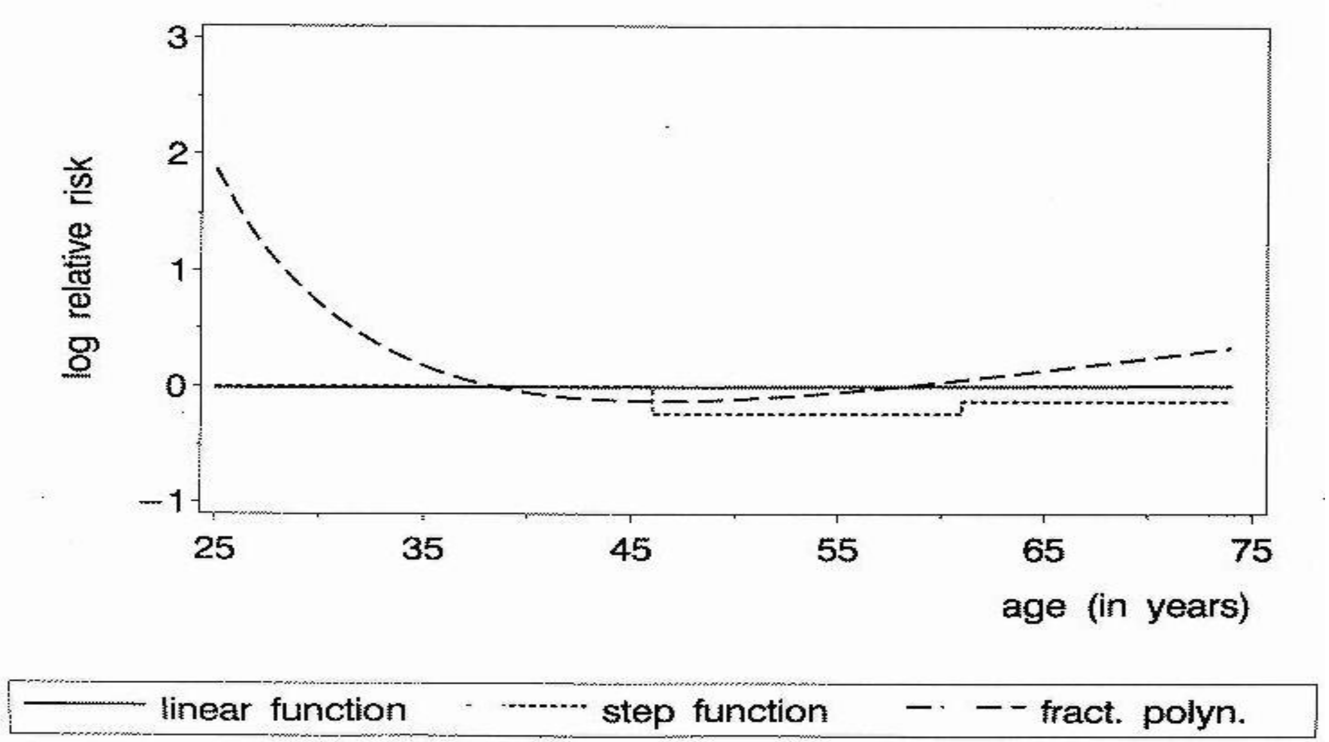


FIGURE 6. Comparison of the estimated effects of selected risk factor modifications for hypothetical risk profiles. The three risk profiles are: *A*, healthy nonsmoker with mean values of continuous risk factors; *B*, smoker with elevated values of systolic pressure and BMI (corresponding to the 75th percentile of the respective distributions in the random sample); and *C*, nonsmoking diabetic with a history of previous CHD and mean values of continuous risk factors. The top panel shows the estimated decrease in the probability of a CHD death during a 12-year period because of lowering total serum cholesterol (TC) from 250 to 180 mg/dl (left top panel) as well as from 300 to 220 mg/dl (right top panel). The bottom panel shows the corresponding effects of decreasing the ratio of TC to high density lipoprotein cholesterol (TC/HDL cholesterol) from 6 to 3 (left bottom panel) and from 10 to 6 (right bottom panel). The GAM estimates are denoted by G (above barred boxes), and the logistic model estimates are denoted by *L* (above white boxes). Results are based on males in the random sample of LRC Prevalence and Follow-up Studies, 1972–1987.

Different conclusions re: stat. Significance

(depending on how a continuous variable is modeled)

AGE as predictor of Death or Recurrence in Breast Cancer (adjusted) [Sauerbrei et al, *Br J Cancer* 1999]



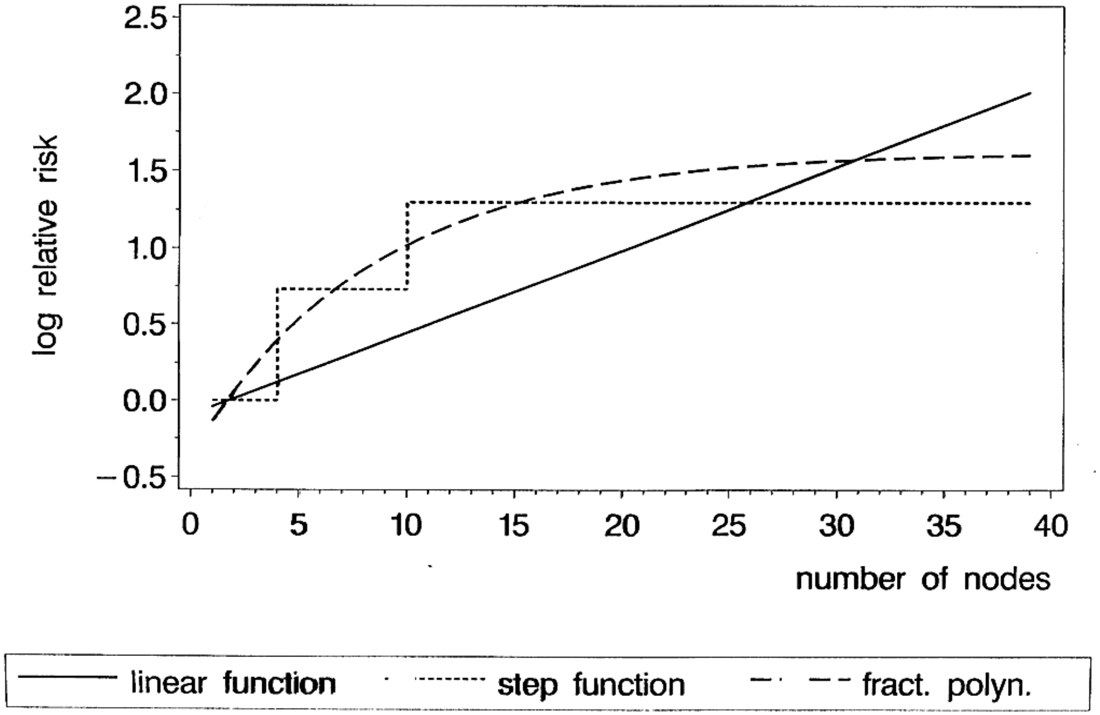
P-value 0.9 0.2 0.001

Similar P-values but DIFFERENT ESTIMATES

(depending on how a continuous variable is modeled)

# + NODES as predictor of Death or Recurrence in Breast Cancer:

[Sauerbrei et al, *Br J Cancer* 1999]



P-value 0.001 0.001 0.001

UN-RESOLVED ISSUES IN FLEXIBLE MODELING OF FUNCTIONAL FORMS

- GENERAL: **Which Smoother ?**

  Quasi-parametric Smoothers such as FP's or Regression Splines facilitate Implementation & Statistical Inference

  [Wegman & Wright, *JASA* 1982]

- Specific for SPLINES:
- **Type of Splines?**
- **Number of Knots** (=> **DF** ?) ***

  Criteria for Choice: AIC, BIC, A Priori, Other ?

  [e.g. Ruppert, *JCGS* 2002]
- **Knot Location ?** ***

  (only minor impact on the estimates)

*** **for Inference: Impact of Data-dependent choices on the Variance** must be accounted for (e.g. by Bootstrap)

[Abrahamowicz et al, *JASA* 1996]

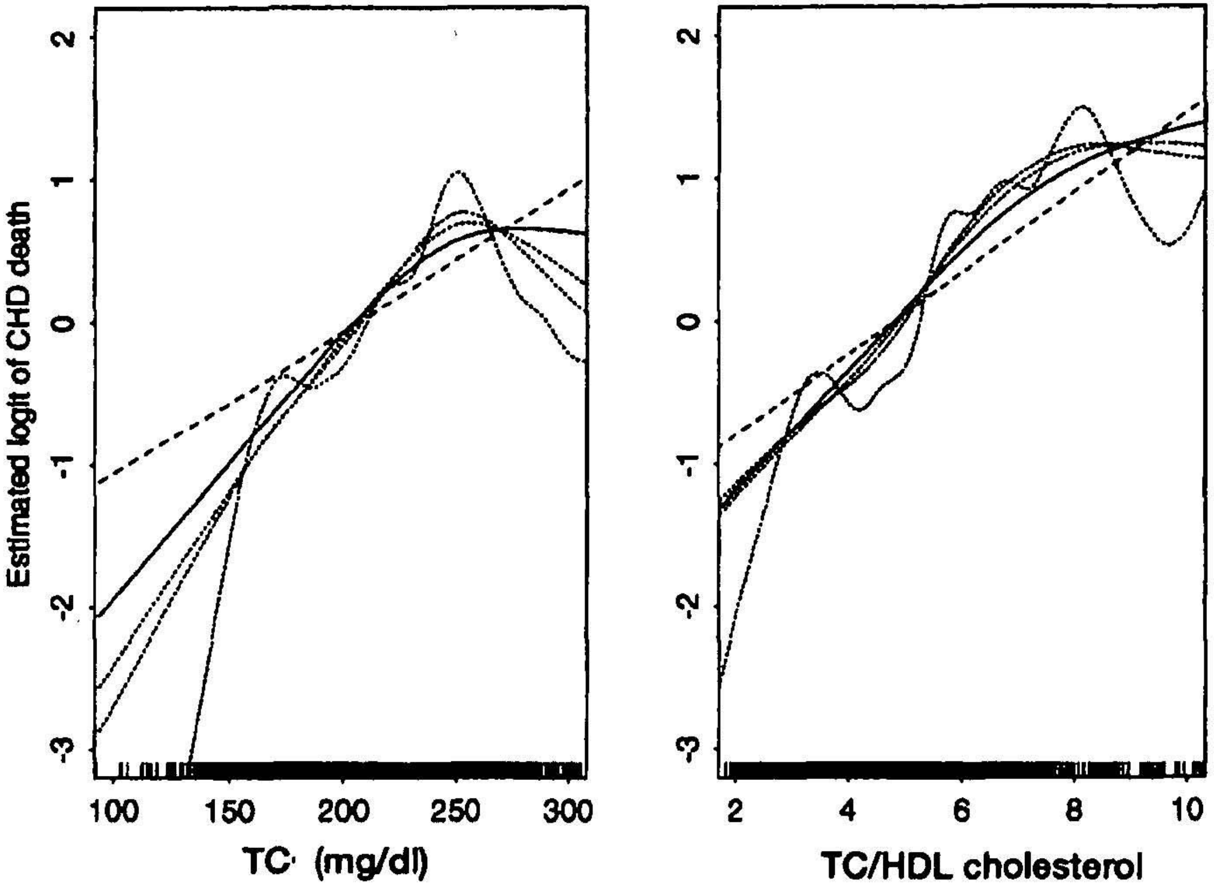
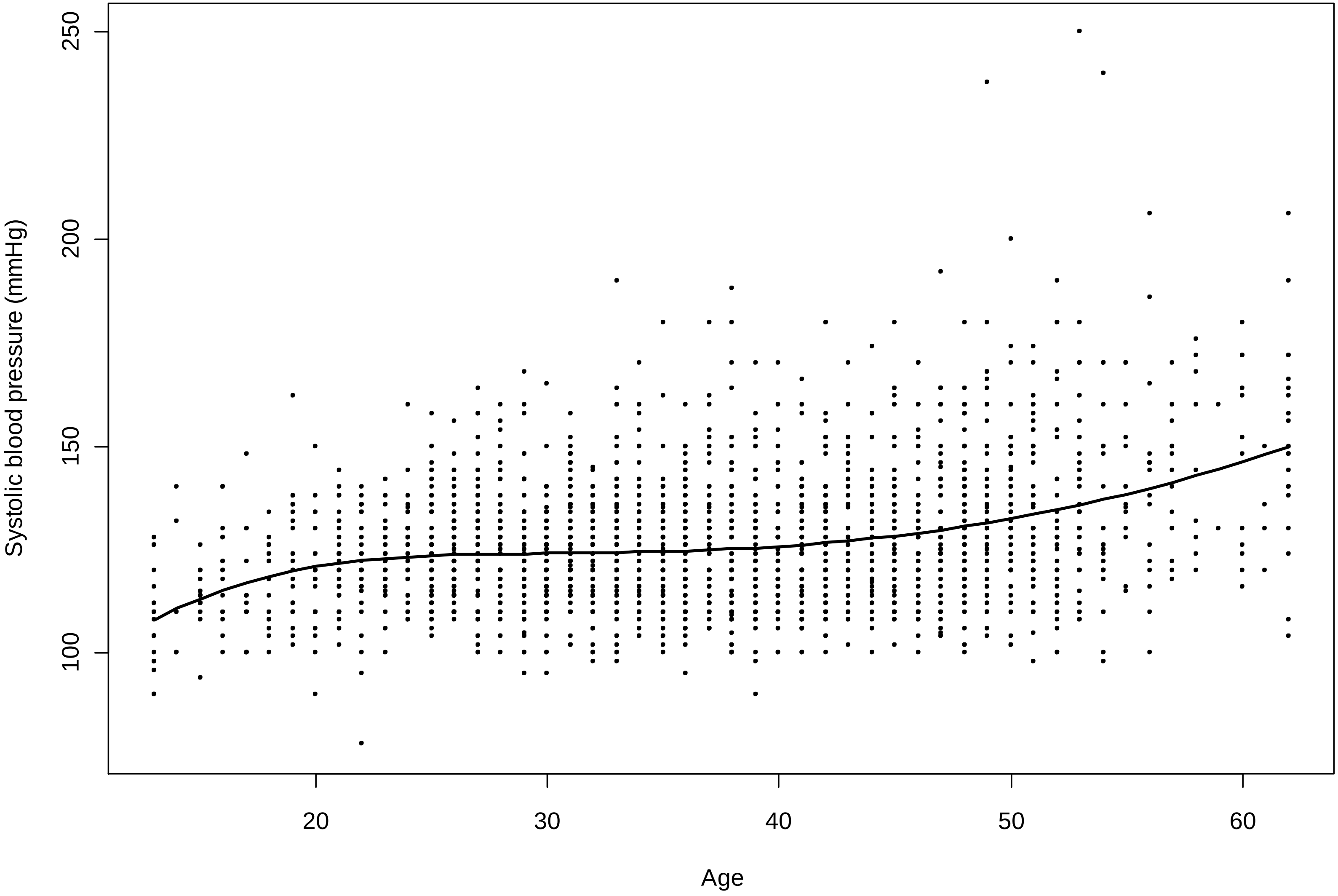


FIGURE 1. Comparison of parametric and nonparametric estimates of the effects of total serum cholesterol (TC) (left panel) and the ratio of TC to high density lipoprotein cholesterol (TC/HDL cholesterol) (right panel) for males in the random sample of the LRC Prevalence and Follow-up Studies, 1972–1987. The tick marks on the abscissa describe the sample distribution of the respective risk factor. The logistic regression estimate is denoted by a dashed line. GAM estimates are displayed for 2 (solid curve) as well as 3, 4, and 10 (dotted curves) df. All effects are adjusted for risk factors listed in table 1. The graphs correspond to an arbitrary vector of nonlipid risk factor values, so that the logit scale on the vertical axis has a valid unit but an arbitrary zero. Accordingly, the relative risks between different risk factor values are accurately represented, but absolute risk levels should not be interpreted. CHD, coronary heart disease.

Real-life example where **Cubic regression B-splines are robust (1, 2, or 3 knots yield identical estimates) :** AGE-SBP RELATIONSHIP IN FRAMINGHAM STUDY (~ **2,500** men)



Summary of the Results of (Very few) simulation studies that systematically compared selected smoothers

- [**Binder et al. *SIM* 2013**]: FP's vs Splines in a complex multivariable linear regression:

  FP's fit better than Splines if 'true' *g(X)* simple, but Splines better for complex g*(X)* (non-monotone; sharp peaks …)

- [**Govindaraju et al. *Int J Biost* 2009**]: 3 types of Splines (Penalized P; Restricted Cubic RC; Natural) & FP's in Cox:

  FP's best (min MSE) if true g(X) linear, P-splines best for Non-linear or Non-monotone *g(X)* but inflated Type I error for testing NL effects; RC splines showed under-fit Bias; AIC too 'liberal'

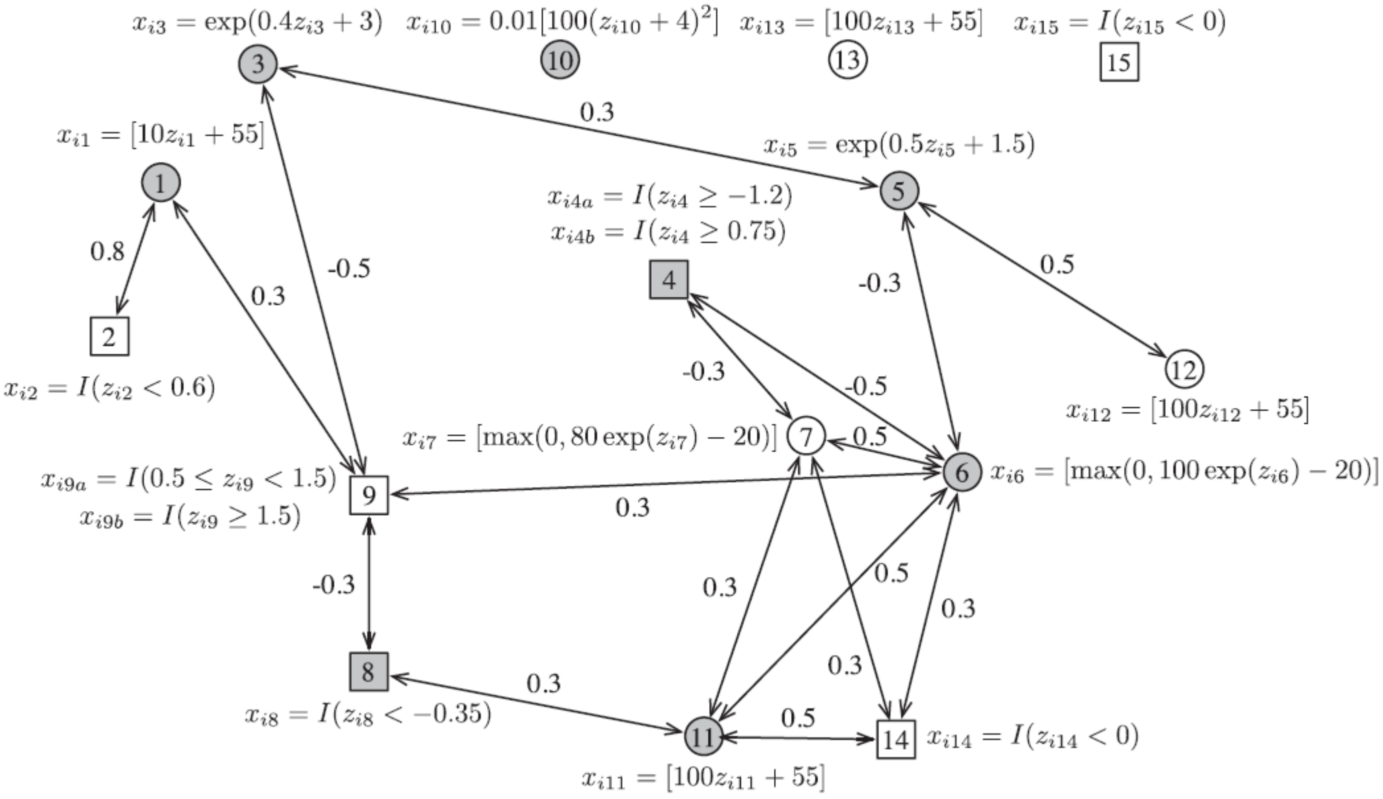
- [**Hastie & Tibshirani, GAM monograph 1990**]: Smoothing splines vs Regression splines vs Loess in GAM: similar point estimates for all smoothers (for equivalent df's)

- [**Hollander & Schumacher, *CSDA* 2004**]: FP's vs RCS in Survival Analysis: FP's had lower Type I error & lower MSE

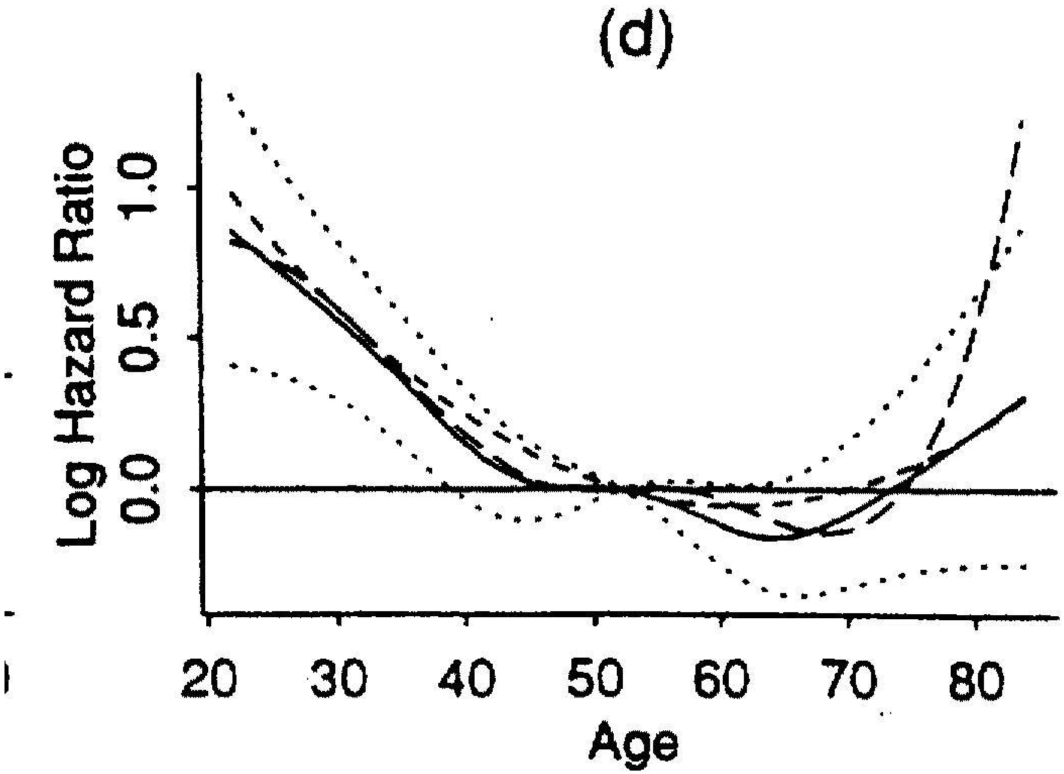
- [**Wand, *Comp Stat* 2000**]: different Spline Smoothing procedures (with shrinkage): Natural Splines reduce the in-stability of conventional Polynomial Splines (e.g. B-splines) in the Tails

Example of a Clinically-Motivated Complex Simulation Design

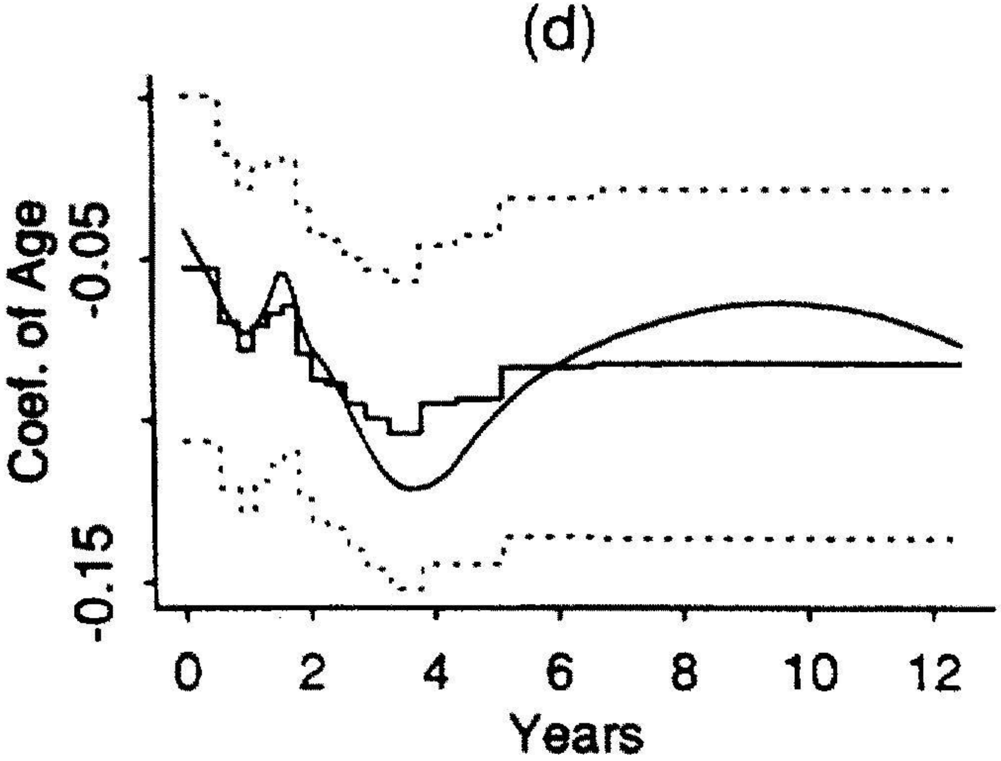
Binder H, Sauerbrei W, Royston P, *Statistics in Medicine*, 2013


$x_{i3} = \exp(0.4z_{i3} + 3)$
$x_{i10} = 0.01[100(z_{i10} + 4)^2]$
$x_{i13} = [100z_{i13} + 55]$
$x_{i15} = I(z_{i15} < 0)$
3
10
13
15
0.3
$x_{i1} = [10z_{i1} + 55]$
$x_{i5} = \exp(0.5z_{i5} + 1.5)$
1
5
$x_{i4a} = I(z_{i4} \geq -1.2)$
$x_{i4b} = I(z_{i4} \geq 0.75)$
0.8
-0.5
0.3
4
-0.3
0.5
2
$x_{i2} = I(z_{i2} < 0.6)$
-0.3
-0.5
12
$x_{i12} = [100z_{i12} + 55]$
$x_{i7} = [\max(0, 80\exp(z_{i7}) - 20)]$
7
0.5
6
$x_{i6} = [\max(0, 100\exp(z_{i6}) - 20)]$
$x_{i9a} = I(0.5 \leq z_{i9} < 1.5)$
$x_{i9b} = I(z_{i9} \geq 1.5)$
9
0.3
-0.3
0.3
0.5
0.3
0.3
8
0.3
$x_{i8} = I(z_{i8} < -0.35)$
11
0.5
14
$x_{i14} = I(z_{i14} < 0)$
$x_{i11} = [100z_{i11} + 55]$

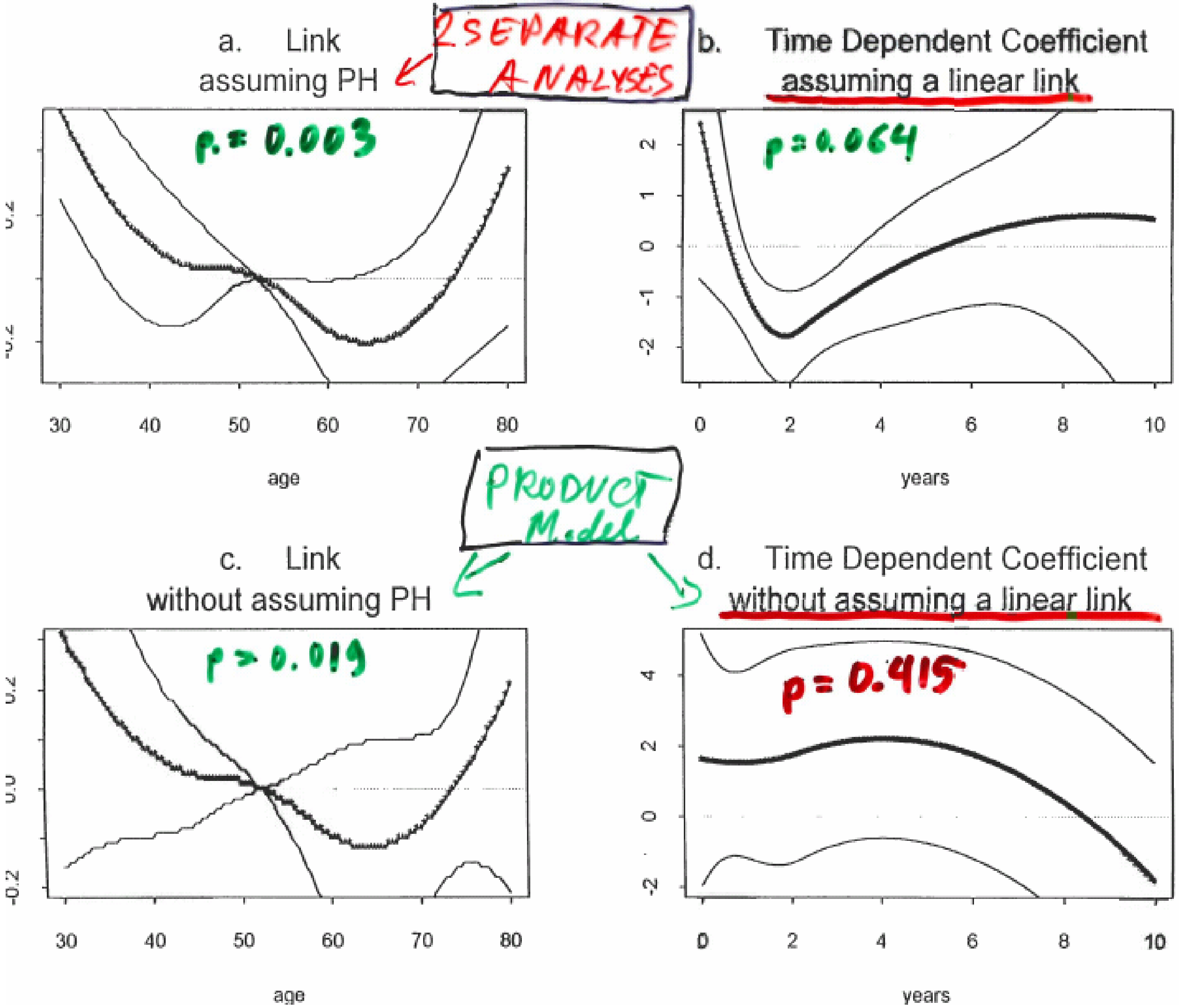
REAL-LIFE RESULTS: [GRAY, *JASA* 1992] **SMOOTHING SPLINE** ESTIMATE OF **NL EFFECT OF AGE** IN BREAST CANCER RECURRENCE (**ASSUMING PH**)



SMOOTHING SPLINE ESTIMATE OF TIME-DEPENDENT (TD) EFFECT OF AGE (ASSUMING LINEARITY IN LOG HAZARD) [GRAY, *JASA* 1992]



Quadratic Regression Spline estimates of effects of AGE: NL (Top-Left) & TD (TOP-Right) [Abrahamowicz & MacKenzie, *SIM* 2007] in data from R. Gray (JASA 1992)



2 SEPARATE ANALYSES

a. Link assuming PH

p.= 0.003

b. Time Dependent Coefficient assuming a linear link

p=0.064

PRODUCT Model

c. Link without assuming PH

p>0.019

d. Time Dependent Coefficient without assuming a linear link

p=0.415

Comments on 3 previous slides

- **Penalized Smoothing Spline *versus* (un-penalized) Quadratic Regression B-splines yielded Very Similar Estimates** of both NL & TD (non-PH) effects of Age at Diagnosis
- WARNING:

  Both types of Estimates may be Inaccurate if the *A Priori* imposed assumptions of (a) PH (for NL estimation) and/or (b) Linearity (for TD estimate) are Invalid ?

INTER-DEPENDENCE OF NL VS TD ESTIMATES FOR CONTINUOUS VARIABLES IN SURVIVAL ANALYSIS [ABRAHAMOWICZ & MACKENZIE, *SIM* 2007]

- **Next Slide illustrates** how the

  TD Estimates for Age, and their 'Statistical Significance' differ depending on whether:

- (a) **Linearity of Age effect is imposed**

  (TD estimate in top-right panel, **p=0.064**)

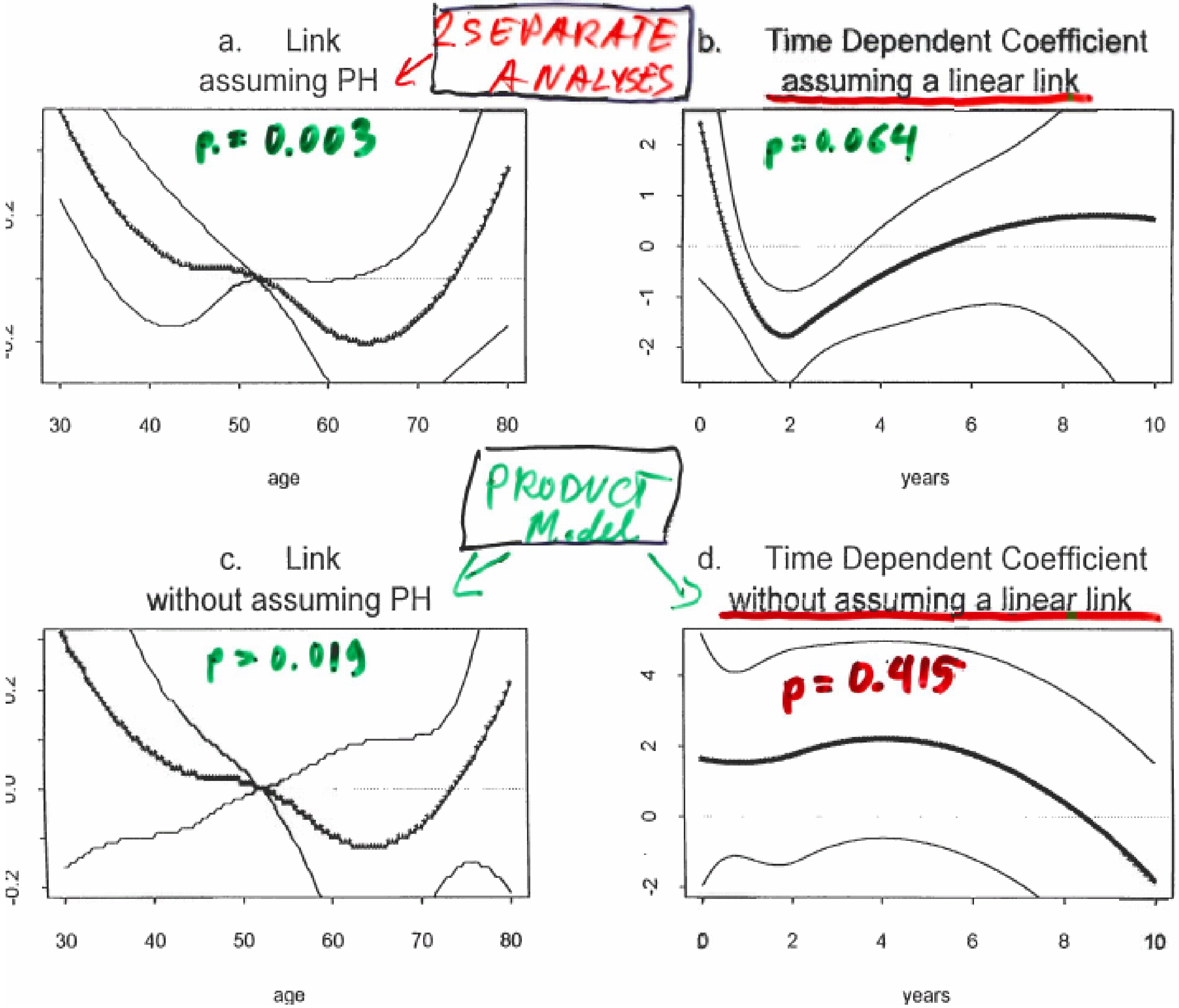
OR

- (b) **NL effect of Age is accounted for**

  (TD estimate in bottom-right panel, **p=0.415**)

[Abrahamowicz & MacKenzie, *SIM* 2007]

NL (Left) & TD (Right) effects of AGE: Estimates (P-values): separate analyses (Top graphs) vs Mutually adjusted for (Bottom graphs)



a. Link assuming PH

2 SEPARATE ANALYSES

b. Time Dependent Coefficient assuming a linear link

p = 0.003

p = 0.064

PRODUCT Model

c. Link without assuming PH

d. Time Dependent Coefficient without assuming a linear link

p > 0.019

p = 0.415

Impact of imposing incorrect PH assumption on inflated type I error rate for testing NL effect (Simulations in [Abrahamowicz & MacKenzie, *SIM* 2007])

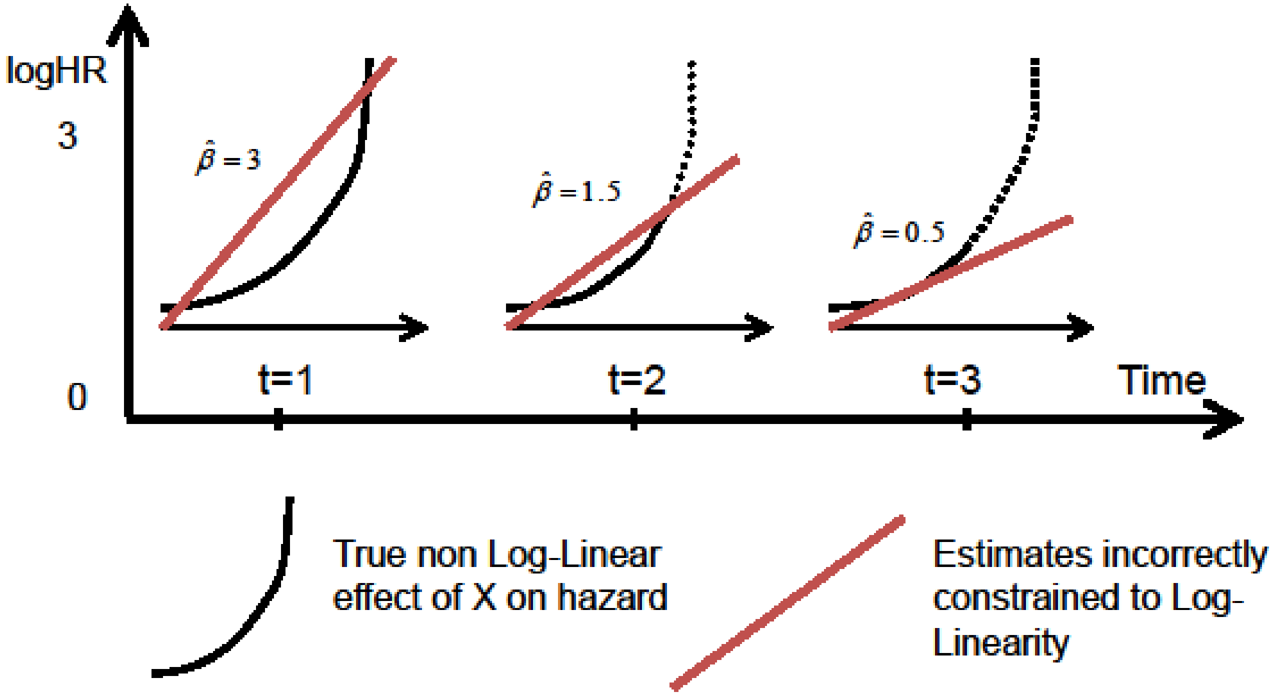
			Empirical Rejection Rates (α=0.05)					
		H1	**PH *lin***	**TD *lin***	**PH *flex***	**TD *flex***	**TD *flex***	**TD *flex***
TDF	**DRF**	vs.						
True TDC	**True link**	H0	**Null**	**PH *lin***	**PH *lin***	**TD *lin***	**PH *flex***	**PH *lin***
Null	**-**		0.06	0.05	0.04	0.15	0.15	0.10
Constant	**Linear**		0.90	0.05	0.04	0.09	0.09	0.06
Constant	**LLH**		0.69	0.23	0.65	0.57	0.13	0.59
Constant	**LHL**		0.37	0.07	0.98	0.96	0.07	0.96
Constant	**LHH**		0.59	0.10	0.40	0.41	0.10	0.35
Decay	**Linear**		1.00	0.97	0.75	0.04	0.69	0.92
Decay	**LLH**		0.98	1.00	1.00	0.80	0.58	1.00
Decay	**LHL**		0.41	0.12	1.00	1.00	0.95	1.00
Decay	**LHH**		0.56	0.09	0.14	0.40	0.31	0.33
Invert	**Linear**		0.75	0.98	0.57	0.05	0.88	0.94
Invert	**LLH**		0.82	0.97	1.00	0.76	0.50	1.00
Invert	**LHL**		0.47	0.20	0.58	0.92	0.92	0.94
Invert	**LHH**		0.14	0.81	0.09	0.53	0.90	0.88

"SELF-CONFOUNDING"

- (Traditional) CONFOUNDING of

  "Exposure Effect" occurs if another risk factor, correlated with "Exposure", is Omitted or Mis-modeled (Residual Confounding)
- **"SELF-CONFOUNDING" of a given effect of "Exposure"** (e.g. TD) **occurs if a Different Effect (e.g. NL) of the SAME "Exposure" is Ignored or Mis-modeled**
- Such 'Self-Confounding' may explain results for: (i) Real-life analyses of Age in Breast Cancer Recurrence and (ii) Simulations (shown, respectively, on the 2 Previous Slides)

Why self-confounding occurs?


logHR
3
0
$\hat{\beta} = 3$
$\hat{\beta} = 1.5$
$\hat{\beta} = 0.5$
t=1
t=2
t=3
Time
True non Log-Linear effect of X on hazard
Estimates incorrectly constrained to Log-Linearity

Failure to account for nonlinearity of the effect of X resulted in a spurious evidence of time-dependence. The linear slope (in red) appears to be decreasing over time

IMPACT OF MODELING OF FUNCTIONAL FORM (FOR CONTINUOUS VARIABLES) ON RESULTS OF SIGNIFICANCE TESTS & VARIABLE SELECTION

- e.g. in Survival analysis:

Failure to account for NL and/or TD effects of a Continuous Variable may result in a serious Type II error for testing its association with the hazard and, thus, **Incorrect Elimination** from the final multivariable model.

- **Simulation results for a binary covariate with a TD effect** ('Crossing Hazards'):

- **% rejection of H0** of No Association (at α=0.05) **->**

% Selection into 'Final' multivariable model:

6 % in PH model **vs** 94% in Flexible TD/NL model

- **Mean LRT statistic for H0** of No Association:

0.88 in PH model **vs** 27.43 in Flexible TD/NL model

[Wynant & Abrahamowicz, *Stat Med* 2014]

Real-life example of link modeling<->variable selection: Albumin is 'Significant' Predictor for Mortality in non-small cell lung cancer Only (Significant) IF NL & TD effects are accounted for (p=0.49 in cox PH vs p<0.001 in flexible NL/TD model)

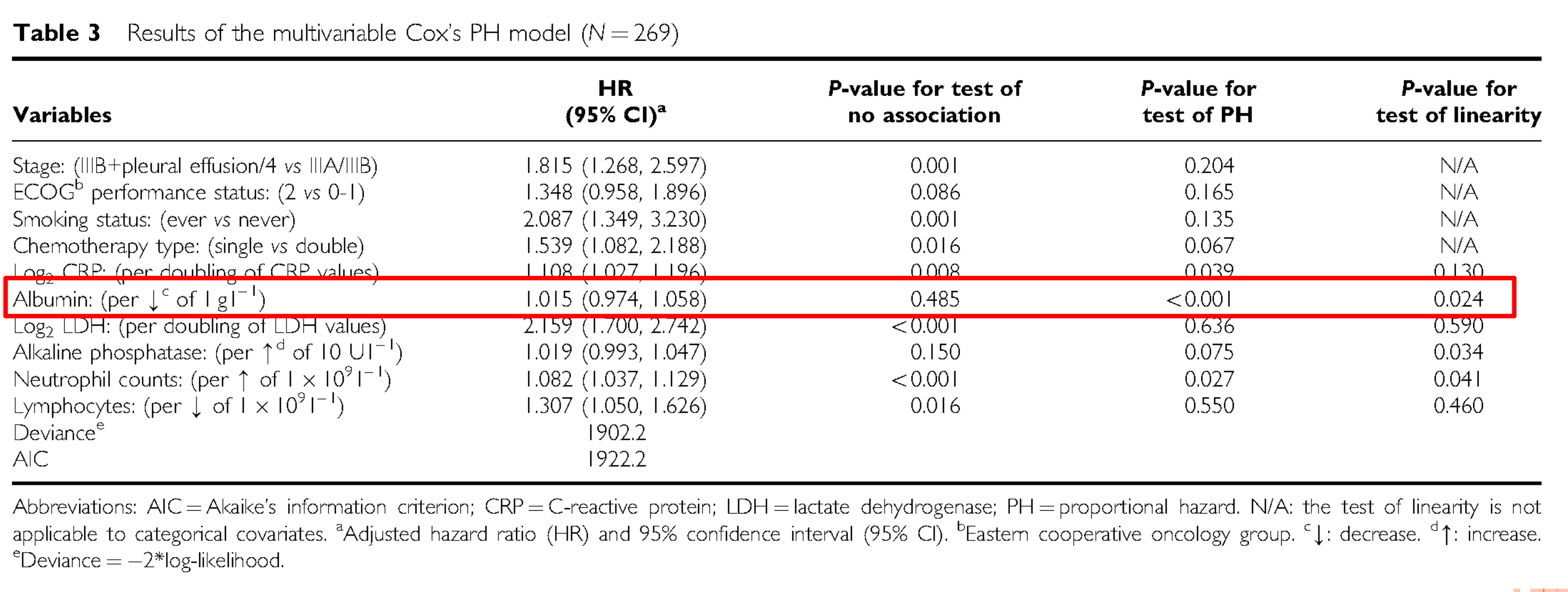


Table 3 Results of the multivariable Cox's PH model (*N* = 269)

Variables	HR (95% CI)[a]	*P*-value for test of no association	*P*-value for test of PH	*P*-value for test of linearity
Stage: (IIIB+pleural effusion/4 *vs* IIIA/IIIB)	1.815 (1.268, 2.597)	0.001	0.204	N/A
ECOG[b] performance status: (2 *vs* 0-1)	1.348 (0.958, 1.896)	0.086	0.165	N/A
Smoking status: (ever *vs* never)	2.087 (1.349, 3.230)	0.001	0.135	N/A
Chemotherapy type: (single *vs* double)	1.539 (1.082, 2.188)	0.016	0.067	N/A
Log_2 CRP: (per doubling of CRP values)	1.108 (1.027, 1.196)	0.008	0.039	0.130
Albumin: (per ↓[c] of 1 g l^{-1})	1.015 (0.974, 1.058)	0.485	<0.001	0.024
Log_2 LDH: (per doubling of LDH values)	2.159 (1.700, 2.742)	<0.001	0.636	0.590
Alkaline phosphatase: (per ↑[d] of 10 U l^{-1})	1.019 (0.993, 1.047)	0.150	0.075	0.034
Neutrophil counts: (per ↑ of $1 \times 10^9\,\text{l}^{-1}$)	1.082 (1.037, 1.129)	<0.001	0.027	0.041
Lymphocytes: (per ↓ of $1 \times 10^9\,\text{l}^{-1}$)	1.307 (1.050, 1.626)	0.016	0.550	0.460
Deviance[e]	1902.2			
AIC	1922.2			

Abbreviations: AIC = Akaike's information criterion; CRP = C-reactive protein; LDH = lactate dehydrogenase; PH = proportional hazard. N/A: the test of linearity is not applicable to categorical covariates. [a]Adjusted hazard ratio (HR) and 95% confidence interval (95% CI). [b]Eastern cooperative oncology group. [c]↓: decrease. [d]↑: increase. [e]Deviance = −2*log-likelihood.

British Journal of Cancer (2010) **102**(7), 1113–1122

NL effects of albumin differ across follow-up times (TD effect)

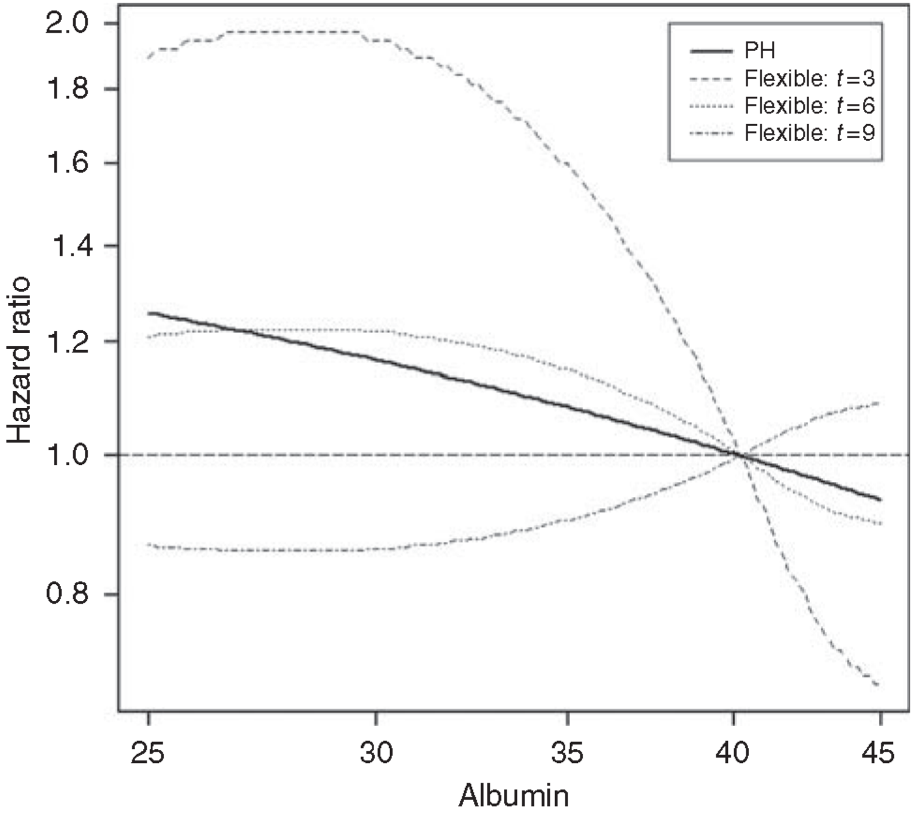


Figure 3 Results of the Cox's PH and flexible spline-based multivariable modeling of the effect of albumin on survival. The bold line represents the linear estimate from the Cox's PH model. The curves correspond to the flexible spline estimates at different times from 3 months ($t = 3$) to

British Journal of Cancer (2010) **102**(7), 1113–1122

IMPACT OF MODELING OF CONTINUOUS VARIABLES ON SIGNIFICANCE TESTING & VARIABLE SELECTION: **RESIDUAL CONFOUNDING**

Simulations demonstrate that:

Failure to account for NL (and/or TD) effects of a Continuous Confounder (correlated with the "Exposure' of main interest) may result in:

- a serious Inflation of either Type I or Type II error for testing the effect of exposure

  [Benedetti & Abrahamowicz *Stat Med* 2004]

- Incorrect Elimination of Truly Prognostic Variables or their NL/TD effects from the final multivariable model, OR Incorrect Selection of 'Spurious' Variables and/or Effects

  [Wynant & Abrahamowicz *Stat Med* 2014]

- Conclusion:

the 2 Challenges on which STRATOS TG2 focuses are Inherently Related and need to be addressed Simultaneously

MULTIVARIABLE MODELS – METHODS FOR VARIABLE SELECTION

Full model

- variance inflation in the case of multicollinearity

Stepwise procedures ⇒ prespecified (α_{in}, α_{out}) and actual significance level?

- forward selection (FS)
- stepwise selection (StS)
- backward elimination (BE)

All subset selection ⇒ which criteria?

AIC	Akaike Information Criterion	= n ln (SSE / n)	+ p 2
BIC	Bayes Information Criterion	= n ln (SSE / n)	+ p ln(n)
		fit	penalty

Central issue: MORE OR LESS COMPLEX MODELS?

Stepwise procedures

- Central Issue: Cut-off for Significance level (popular 0.157 (=AIC), 0.05, 0.01)

Criticism

- FS and StS start with ‚bad‘ univariate models (under-fitting, No Control for Confounding)
- BE starts with too many variables (over-fitting, over-adjustment) -> usually less critical than problems with FS
- INCORRECT INFERENCE: due to Multiple Testing
- BIAS: Estimates biased (over-estimation of the strength of the association for variables that were selected into the final model)

Many more strategies…

- Boosting

- Triggered by high dimensional data:

  Many approaches based on regularization techniques

  Combining selection with shrinkage, e.g. LASSO

  [Tibshirani, *JRSS B* 1996)

- Techniques based on resampling

Diverging opinions re: Variable Selection approaches for multivariable explanatory models

- 1st Question:

  Do we Need Any Variable selection ?

- e.g. **F. Harrell** [2001; 2015] **argues that NOT** :
- Penalized MLE can ‘take care’ of spurious variables by shrinking their estimated effects to (or close to…) 0
- However, Assessing of the Performance of Penalized MLE in the context of Multivariable Models with several correlated Continuous Variables (possibly with Non-linear effects) requires More Empirical (Simulation-based) Evidence

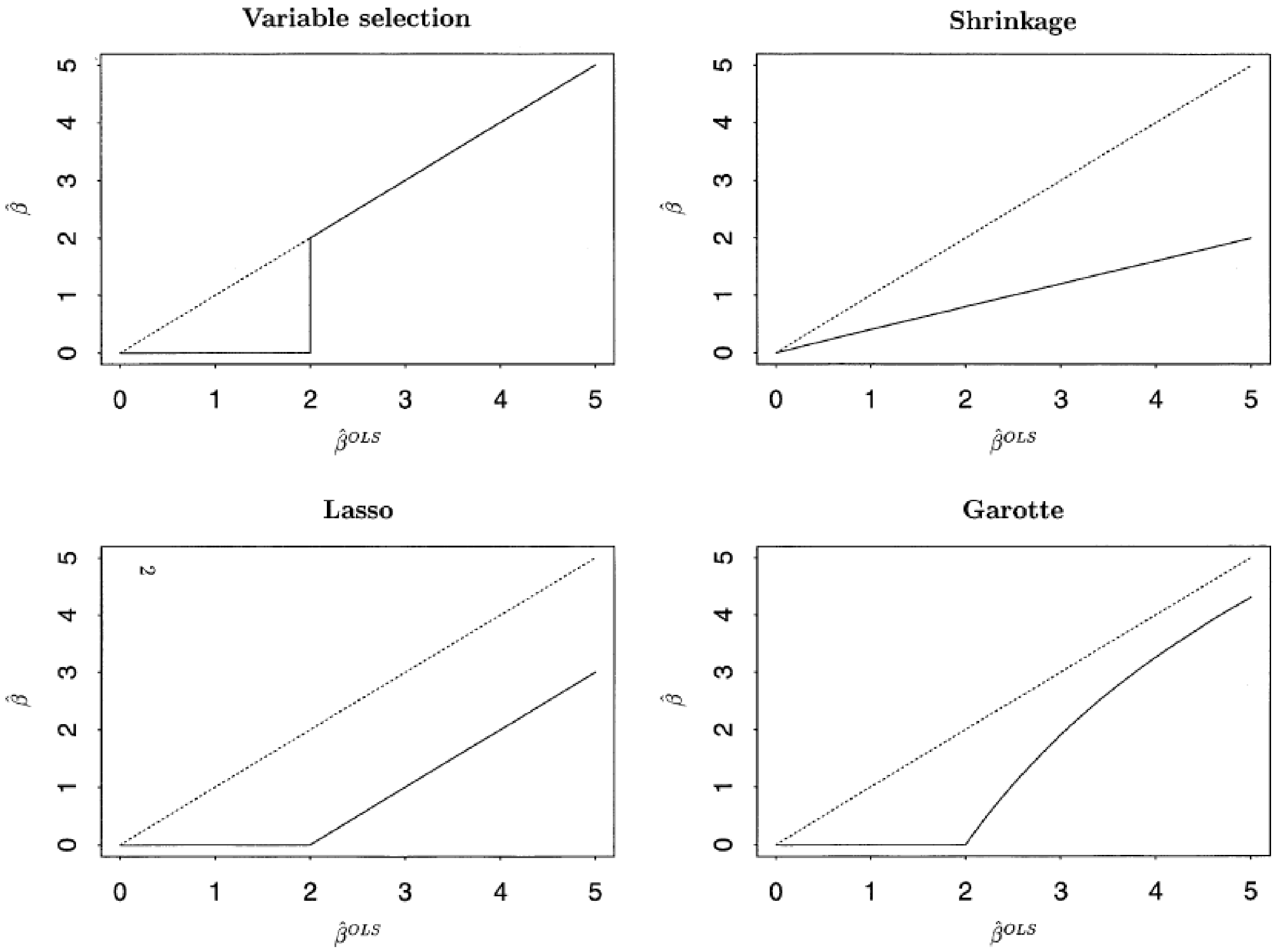
SHRINKAGE VS VARIABLE SELECTION

- Regularized regression
  - Add penalty to log-likelihood
  - L_1 penalty: $\sum|\beta_j|$, L_2 penalty: $\sum\beta_j^2$
- Estimates of β_j are shrunken towards zero
- Selection = extreme shrinkage

  “If it’s close to 0, set it to 0.”
- Consequences of shrinkage:
  - Controlling variance, not bias

    (bias-variance trade-off)
  - Confidence intervals?

Variable selection and shrinkage



EVOLVING 'PREFERENCES' FOR METHODS OF VARIABLE SELECTION IN EPIDEMIOLOGY

- In Epidemiological studies, multivariable models are typically built to **Assess the Unbiased ("causal") effect of a Single Exposure**/Risk Factor/Treatment. Thus, **other variables are selected mostly to reduce risk of Confounding Bias.**
- This led to **Popularity of the "Change-in-the-estimate"** [Mickey & Greenland, *AJE* 1993] in Epi research [Madonado & Greenland, *AJE* 1993], **and 'rejection' of Stepwise methods**
- Yet, **recent Simulations results suggest Combining Backward Elimination with Change-in-Estimate**

  [Dunkler et al, *PLoS One*2014]
- Interestingly, **Causal Inference experts** Recently argued in favor of **Adjusting for All Variables considered 'Causes' of (a) Either (a) the Exposure, (b) Or the Outcome** (i.e. May Not act as traditional 'confounders').
- [Vanderwelde & Shpitser, B*iometrics* 2011]

(MOSTLY RECENT) SIMULATION-BASED FINDINGS FOR VARIABLE SELECTION METHODS

- Dunkler et al [*PLoS One* **2014**]:

  Augmented Backward Elimination (BE): **combining** BE with Standardized Change-in-Estimate criterion Reduces :

  (i) Bias & (ii) Risk of Elimination of predictive variables

- Van Houwelingen & Sauerbrei [*Open J Statistics* **2013**]:

  Shrinkage should be **combined** with Variable Selection

- Tibshirani [*JRSS B* 1996]:

  Advantages of LASSO: **combining** interpretability of results of traditional variable selection & stability of ridge regression

- Vanderwelde & Shpitser [B*iometrics* **2011**]:

  combining traditional Propensity Score-based approach of Exposure determinants with selection of Variables related to Outcome

COMBINING VARIABLE SELECTION & MODELING OF NON-LINEAR ASSOCIATIONS

- **Raheem et al [Comp Stat & Data Analysis 2012]:**
- **semi-parametric Positive-part Shrinkage estimator** that, for a 'partially linear model', **combines (i) shrinkage of the parametric (linear) coefficients with (ii) B-spline modeling of NL effects**
- in Simulations: the semi-parametric estimator (a) agrees with BIC selection for the 'parametric' sub-model, (b) reduces MSE relative to kernel estimators of the NL effects, (c) outperformed the absolute penalty estimators when the 'true model' was large
- Preliminary results of Comparisons with Lasso & (computationally expensive) Adaptive Lasso in-conclusive: Need for Further Simulations ?

(SELECTED) UN-RESOLVED ISSUES & CHALLENGES FOR STRATOS TG2:

- (1) Systematic **Comparisons of Alternative Smoothers and Modeling Strategies for Splines** Fitting
- (2) Optimization of **procedures for selecting NL and/or TD effects of Continuous Variables in Survival** (link to **TG8**)
- (3) Assessment of Performance of Alternative (traditional & computer-intense) **Algorithms for Variable Selection and/or Shrinkage** in Complex Multivariable Analyses
- (4) Novel Comprehensive Model Building Strategies for **Combining Flexible Modeling of Continuous Variables & complementary approaches for Variable Selection/Shrinkage** ?
- (5) Developing & Validating **Inference Methods to Account for Impact of Data-Dependent choices in (2)-(4) on the Variance**

*** (6) All issues (1)-(5) **need Innovative, Complex, Clinically Realistic Simulation Designs for Multivariable modeling with Correlated variables and NL/TD effects!**

THANK YOU

- **Michal.Abrahamowicz@McGill.ca**

References (part 1)

- Abrahamowicz M, du Berger R, Grover SA. Flexible modeling of the effects of serum cholesterol on coronary heart disease mortality. *Am J Epidemiol* 1997; 145:714-729.
- Abrahamowicz M, MacKenzie T, Esdaile JM. Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis. *JASA* 1996;91(436):1432-1439.
- Abrahamowicz M, MacKenzie T. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Stat Med* 2007; 26:392-408.
- Benedetti A, Abrahamowicz M. Using generalized additive models to reduce residual confounding. *Stat Med* 2004; 23:3781-3801.
- Binder H, Sauerbrei W, Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Statistics in Medicine* 2013; **32**: 2262-2277.
- Dunkler D, Plischke M, Leffondre K, Heinze G. Augmented backward elimination. PLoS ONE 2014; 9(11): e113677.
- Gagnon B, Abrahamowicz M, Xiao Y et al. Flexible modeling improves assessment of prognostic value of C-reactive protein in advanced non-small cell lung cancer. *Br J Cancer* 2010; 102:1113-1122.
- Govindarajulu U, Malloy E, Ganguli B, Spiegelman D, Eisen E. The comparison of alternative smoothing methods for fitting non-linear exposure-response relationships with Cox model in simulation. Int J Biostat 2009. Vol 5: article 2.
- Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *JASA* 1992; 87:942-951.
- Harrell FE. Regression *Modeling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY, 2001 (1st edition), 2015 (2nd edition).
- Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Chapman & Hall/CRC: New York, 1990.
- Hollander N, Schumacher M. Estimating the functional form of a continuous covariate's effect on survival time. *Comp Stat Data Analysis (CSDA)* 2004; 50:1131-1151.
- Maldonado G, Greenland S. Simulation study of confounder selection criteria strategies. *AJE* 1993; 138:923-936.

References (Part 2)

- Malats N, Bustos A, Nascimento C et al. Molecular Biomarkers in Bladder Cancer: Novel Potential Indicators of Prognosis and Treatment Outcomes. *Lancet Oncology* 2005, 6:678-686.
- Mickey RM, Greenland S. The impact of confounder selection criterion on effect estimation. *AJE* 1993; 129:125-137.
- Raheem SME, Ahmed SE, Doksum KA. Absolute penalty and shrinkage estimation in partially linear models. *CSDA* 2012; 56: 874-891.
- Ramsay JO. Monotone regression splines in action (with discussion). *Statistical Sciences* 1988; 3:425-461.
- Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006; 25:127-141.
- Royston P, Sauerbrei W. Multivariable model building: a pragmatic approach to regression analyses based on fractional polynomials for modelling continuous variables. Wiley. Chichester, UK. 2008.
- Ruppert D. Selecting the number of knots for penalized splines. *J Comp Graph Stat (JCGS)* 2002; 11: 735-737.
- Sauerbrei W, Royston P, Bojar H, Schmoor C, Schumacher M. Modelling the effects of standard prognostic factors in node-positive breast cancer. *Br J Cancer* 1999; 79:1752-1760.
- Schulgen G, Lausen B, Olsen JH et al. Outcome-oriented cut-points in analysis of quantitative exposure. *Am J Epidemiol* 1994; 140:172-184.
- Tibshirani R. Regression shrinkage and selection via the Lasso. *J Royal Statistical Society B* 1996; 58: 267–268.
- Van Houwelingen, HC, Sauerbrei W. Cross-validation, shrinkage and variable selection in linear regression revisited. *Open Journal of Statistics* 2013; 3: 79–102.
- Vanderwelde TJ, Shpitser I. A new criterion for confounder selection. B*iometrics* 2011: 67: 1406-1413.
- Wand MP. A comparison of regression spline smoothing procedures. *Computational Stat* 2000; 15:443-462.
- Wegman EJ, Wright IW. Splines in statistics. *Journal of the American Statistical Association* 1983; **78**: 351-365
- Wynant W, Abrahamowicz M. Impact of the model building strategy on the inference about time-dependent and non-linear covariate effects in survival analysis. *Stat Med* 2014 ;33::3318-37.