The Multivariable Fractional Polynomial approach, with thoughts about opportunities and challenges in Big Data

Willi Sauerbrei[1] and Patrick Royston[2]

[1] Institute for Medical Biometry and Statistics, University of Freiburg, Germany
[2] Medical Research Council, Clinical Trials Unit, London, UK

UNIVERSITÄTS KLINIKUM FREIBURG

Content

- Big data – Aims? How collected?

- Health sciences
  To answer many questions good data (from well designed studies) and suitable statistical approaches are required

- Fractional polynomials (FPs)
- Variable selection
- Multivariable fractional polynomials (MFP) approach

- Extensions (Interactions)

- FPs and Big Data

Big Data

- Term ‘Big Data’ used for many different situations. Extremely confusing
- Hand (JRSSA, 2016) distinguishes **two types**
  First type: primarily data manipulation
  sorting, searching, matching, …
  Examples include online route finders, apps for updated status of bus traffic
  → Mostly addressed by computer scientists & mathematicians
  Second type: uses data to derive models for prediction or understanding of the mechanisms and processes that have generated the collected data
  Achieving these goals will rely primarily on state-of-the-art statistical and machine learning methods

- Aim, design and type of data are key issues
- Data from well designed experiments, systematically collected (eg. registries) or ‘found‘ data?

Comparing two treatments - Typical questions

- **Q1: Using survival time as outcome**
  Is one treatment better?

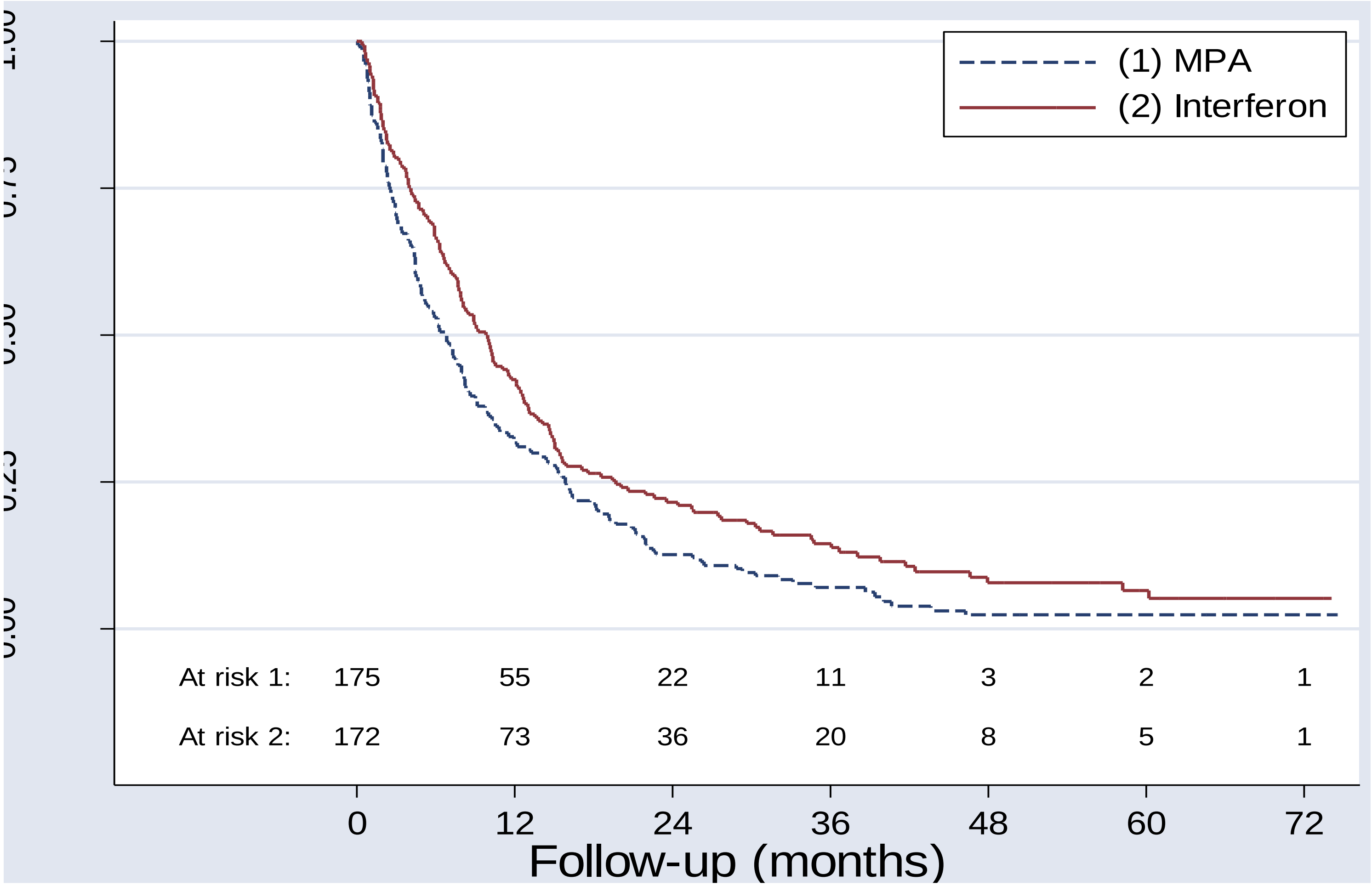
  Well designed experiment required!

RCT in UK (MRC trial) to compare
interferon-α with MPA in renal cancer patients

N = 347, 322 Death

14 potential prognostic factors

RCT to compare MPA and interferon in renal cancer patients



Comparing two treatments
- Typical questions

- **Q1: Using survival time as outcome**
  Is one treatment better?

Main analysis:

Interferon improves survival time

HR: 0.75 (0.60 - 0.93), p = 0.009

Could that have been investigated by using ‚BIG' observational data?

→ NO! Risk of bias is severe!

Comparing two treatments - Typical questions

- **Q2: Is the treatment effect similar in all patients?**

Sensible question?

→ Yes, at least for hypothesis generation

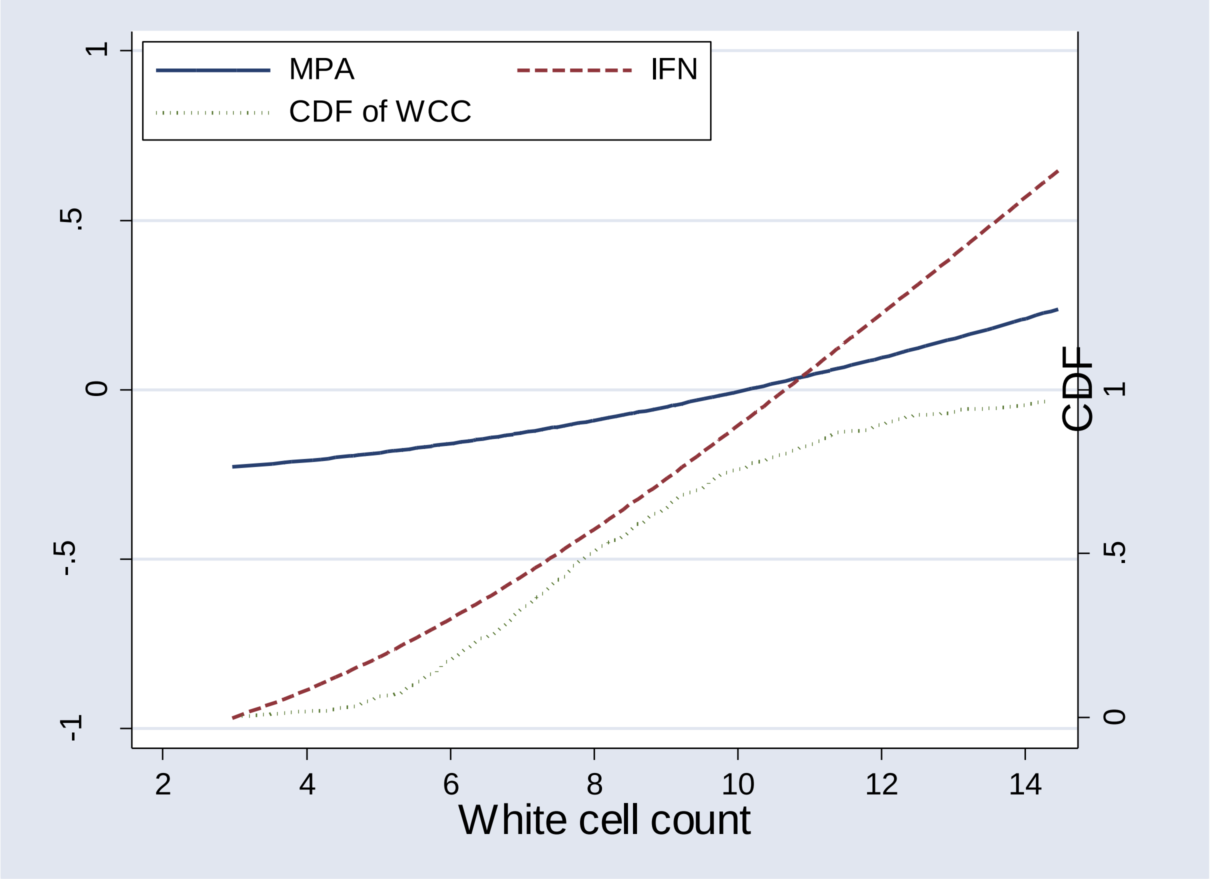
Ten continuous covariates available for the investigation of treatment-covariate interactions –

Using the multivariable fractional polynomial interaction (MFPI) approach one (White Cell Count - WCC) is significant

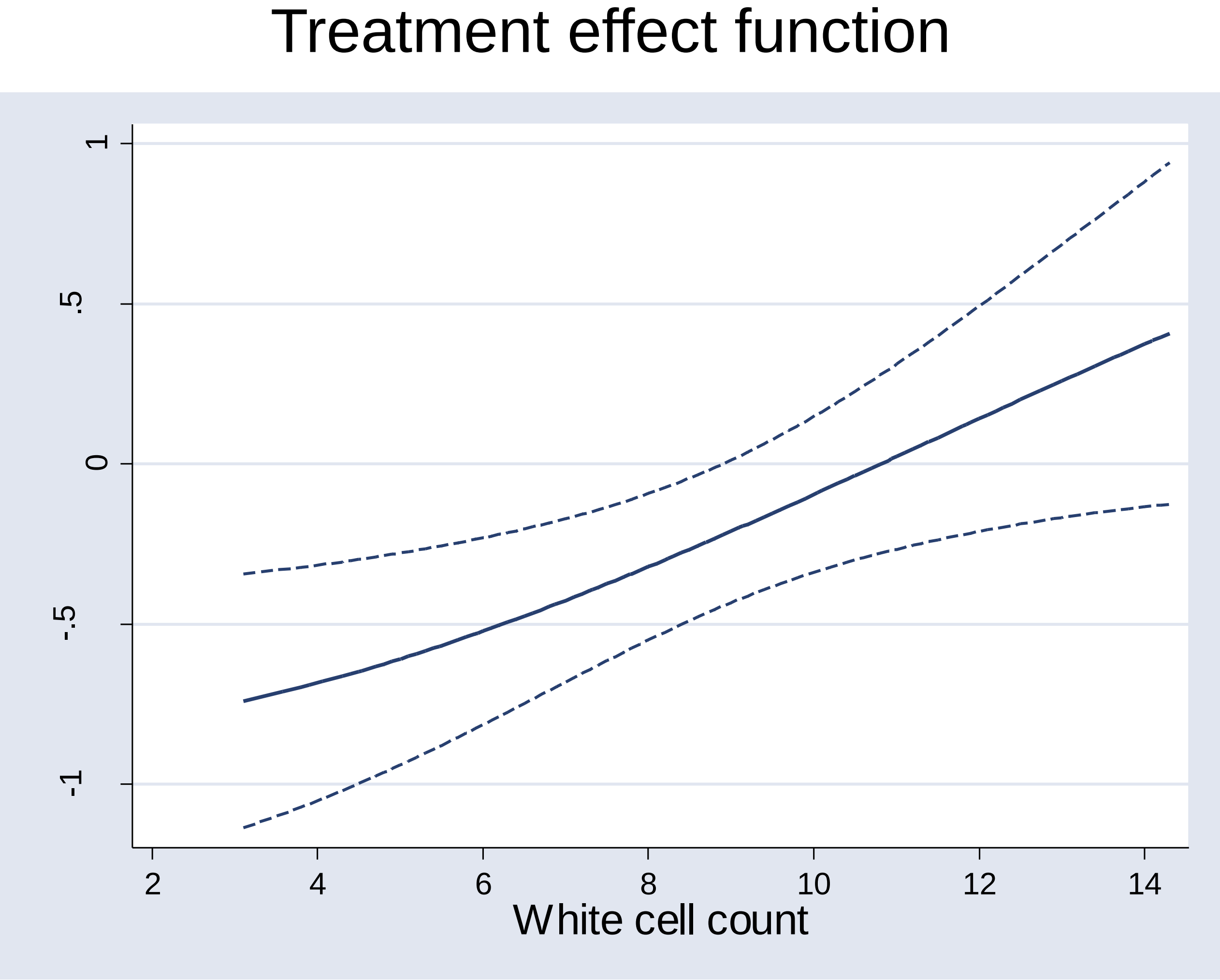
Hypothesis generation:
does the treatment effect depend on any factor?

Functions derived with the fractional polynomial approach

Effect of WCC is best modelled with an FP2 (2, 3).



Treatment effect seems to depend on WCC

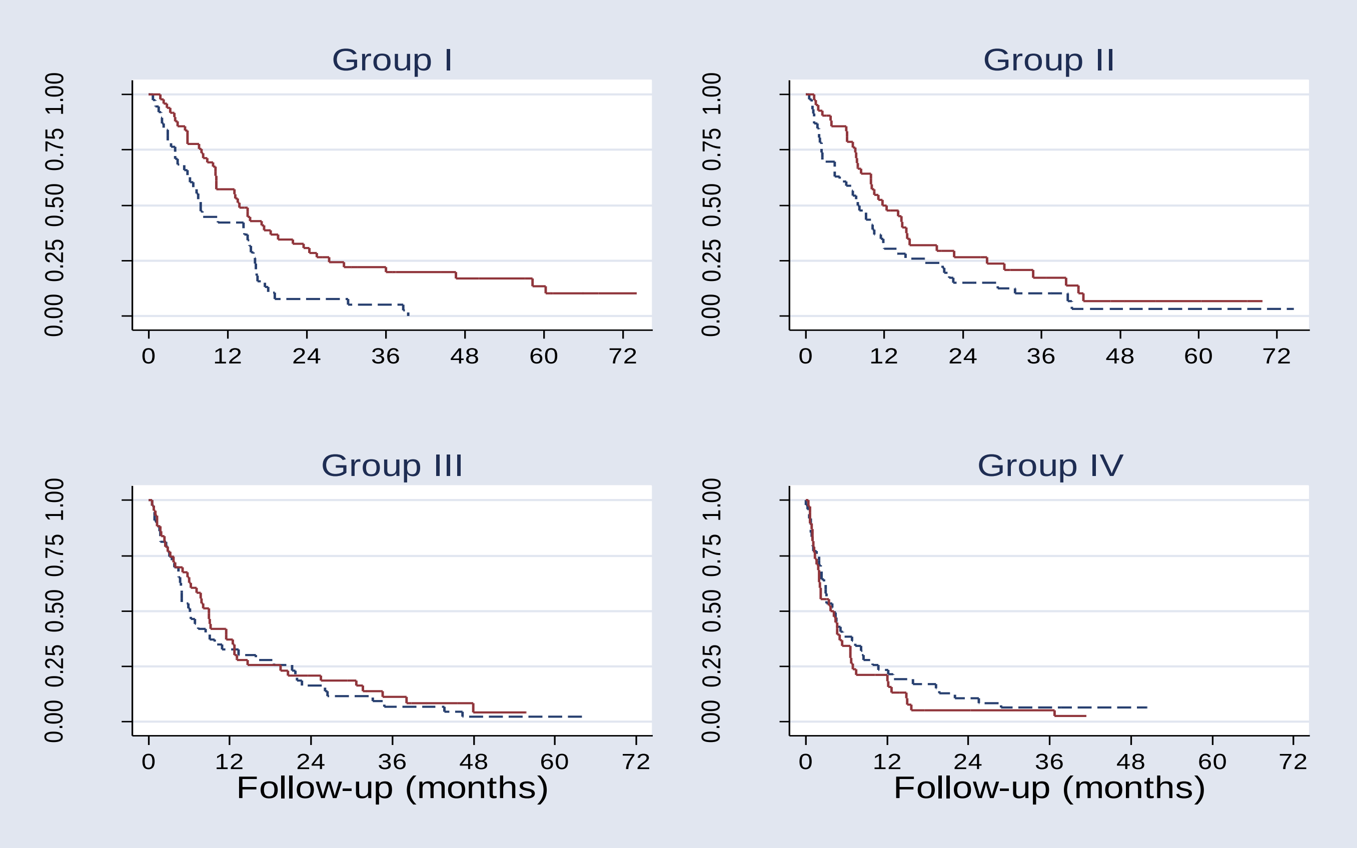


About 25% of patients with WCC > 10 seem not to benefit from interferon

Does model agree with data?

Check proposed trend

Treatment effect in subgroups defined by WCC



HR (Interferon to MPA; adjusted values similar)

overall: 0.75 (0.60 – 0.93)

I : 0.53 (0.34 – 0.83) **II : 0.69 (0.44 – 1.07)**

III : 0.89 (0.57 – 1.37) **IV : 1.32 (0.85 –2.05)**

Comparing two treatments
- Typical questions

- **Q2: Is the treatment effect similar in all patients?**

→ Suitable statistical modelling and larger RCTs are needed to answer such questions reliably.

Observational Studies

Typical situation

Several variables, mix of continuous and (ordered) categorical variables

Different aims of model building

- Prediction
- Explanation
  Shmueli, G.(2010) To explain or to predict?
- Adjust for relevant confounders

Aim has a severe influence on the suitability of modelling approaches

Explanation is of main interest here:

- Identify variables with (strong) influence on the outcome
- Determine functional form (roughly) for continuous variables

The issues are very similar in different types of regression models (linear regression model, GLM, survival models ...)

Use subject-matter knowledge for modelling ...
... but for some variables, data-driven choice inevitable

Regression models

$X=(X_1, \ldots, X_k)$ covariate, prognostic factors

g(x) $= ß_1 X_1 + ß_2 X_2 + \ldots + ß_k X_k$ (assuming effects are linear)

normal errors (linear) regression model

Y normally distributed

$E(Y|X) = ß_0 + g(X)$

$\mathrm{Var}(Y|X) = \sigma^2 I$

logistic regression model

Y binary

$$\text{Logit } P(Y|X) = \mathbf{ln} \frac{P(Y=1|X)}{P(Y=0|X)} = \beta_0 + \mathbf{g(X)}$$

survival times

T survival time (partly censored)

Incorporation of covariates

$$\lambda(t|X) = \lambda_0(t)\exp\,(\mathbf{g(X)})$$

Implicit assumptions

- Subject matter knowledge (if available) determines (parts) of the model
- About 5 to 30 candidate variables
- No ‚small sample size‛ situation
- No missing data problem

Central issues

To select or not to select (full model)?

Which variables to include?

How to model continuous variables?

Continuous variables – The problem

“Quantifying epidemiologic risk factors using non-parametric regression: model selection remains the greatest challenge”

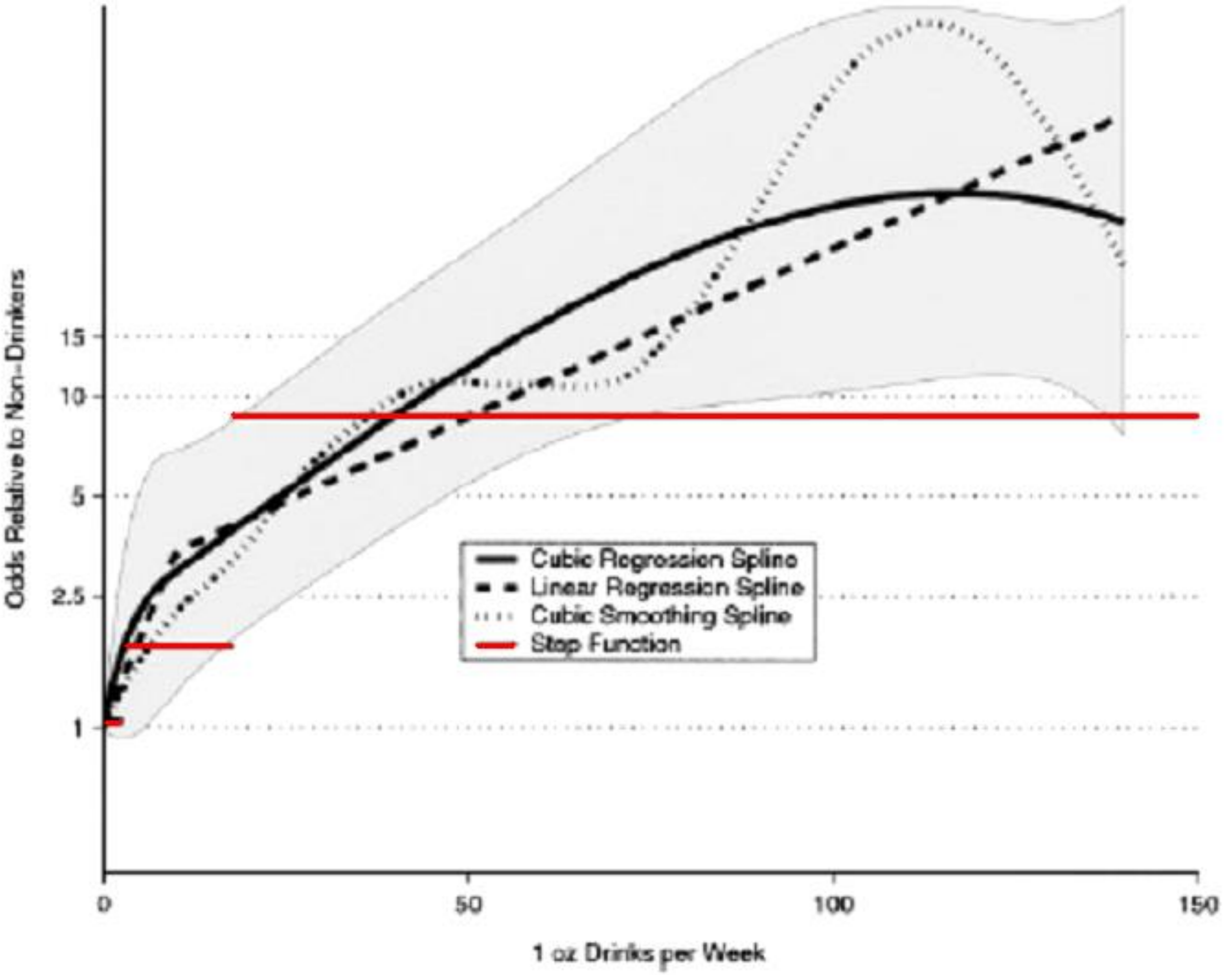
Rosenberg P. et al., StatMed 2003

Discussion of issues in (univariate) modelling with splines

Trivial nowadays to *fit* almost any model

To *choose* a good model is much harder

Alcohol consumption as risk factor for oral cancer



Rosenberg et al, StatMed 2003

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?

• C_p	Mallows	=	$(SSE / \hat{\sigma}^2)$	$- n + p\ 2$
• AIC	Akaike Information Criterion	=	$n \ln (SSE / n)$	$+ p\ 2$
• BIC	Bayes Information Criterion	=	$n \ln (SSE / n)$	$+ p \ln(n)$
			fit	penalty

Combining selection with Shrinkage

Bayes variable selection

Recommendations???

Central issue: MORE OR LESS COMPLEX MODELS?

Stepwise procedures

- Central Issue: significance level

Criticism

- FS and StS start with ‚bad‛ univariate models (underfitting)
- BE starts with the full model (overfitting),

  less critical
- Multiple testing, P-values incorrect

Variable selection

All procedures have severe problems!
⇒ Full model? No!

Illustration of problems
Too often with small studies
(sample size versus no. variables)

Arguments for the full model
Often by using published data
Heavy pre-selection!

What is the full model?

Big data – much bigger problems

Type I error of selection procedures
Actual significance level
(linear regression model, uncorrelated variables)

- BE: For moderate sample size only slightly higher than α_{in}

- For all-subset methods in good agreement with asymptotic results for one additional variable (Teräsvirta & Mellin, 1986)

All-AIC	~	15.7 %
All-BIC	~	P (χ_1^2 > ln (n))
		0.032 N = 100
		0.014 N = 400

→ BE is the only of these approaches where the level can be chosen flexibly depending on the modelling needs!

Backward elimination is a sensible approach

- Significance level can be chosen
- Reduces overfitting

Of course required

- Checks
- Sensitivity analysis
- Stability analysis

Continuous variables – what functional form?

Traditional approaches

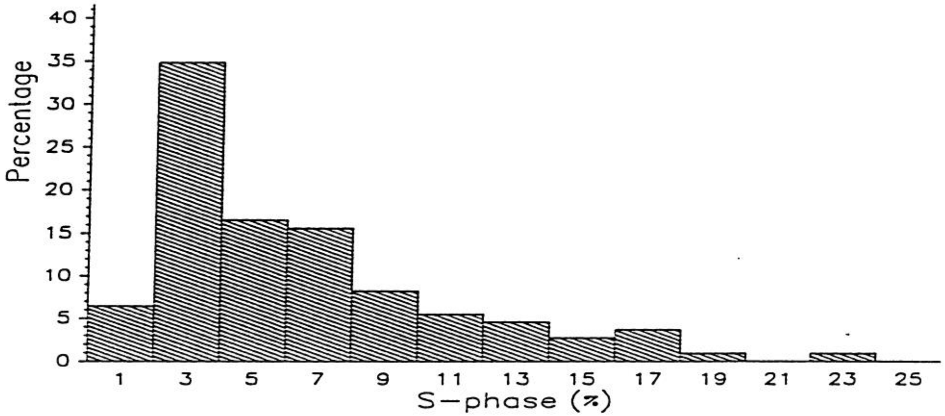
a) Linear function
   - may be inadequate functional form
   - misspecification of functional form may lead to wrong conclusions

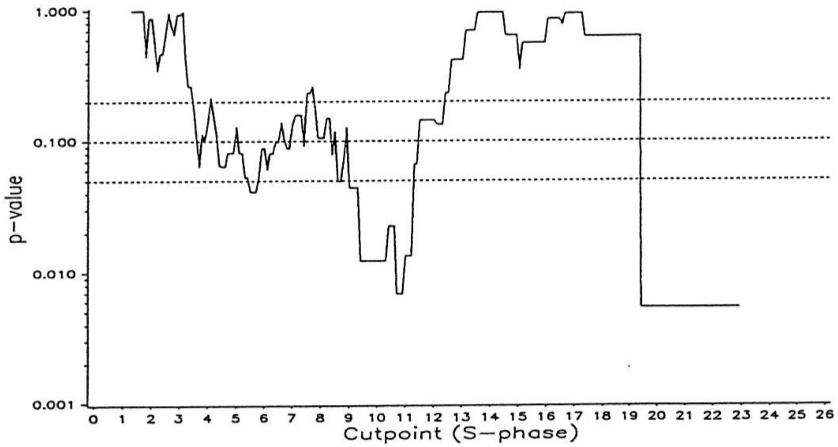
b) ‛best‛ ‛standard‛ transformation

c) Step function (categorial data)
   - Loss of information
   - How many cutpoints?
   - Which cutpoints?
   - Bias introduced by outcome-dependent choice

Searching for optimal cutpoint minimal p-value approach

Prognostic effect of SPF in breast cancer patients (Altman et al 1994)





Problem

multiple testing => inflated type I error

Example: Prognostic factors

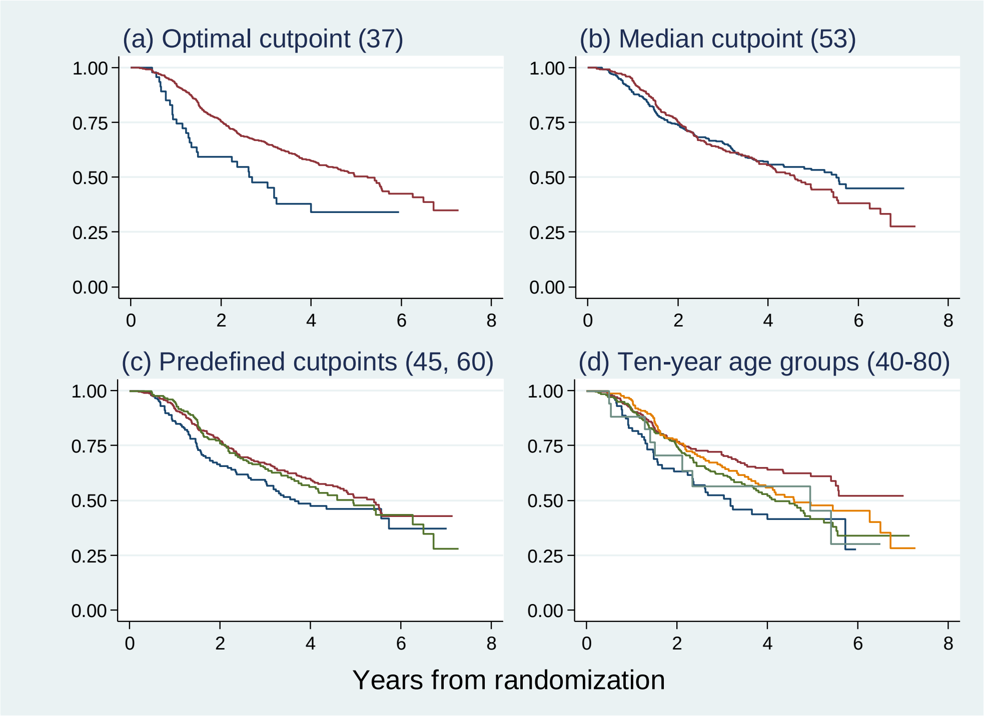
GBSG-study in node-positive breast cancer

299 events for recurrence-free survival time (RFS) in
686 patients with complete data

7 prognostic factors, of which **5** are continuous

Tamoxifen yes/no

Age as prognostic factor – cutpoint analyses



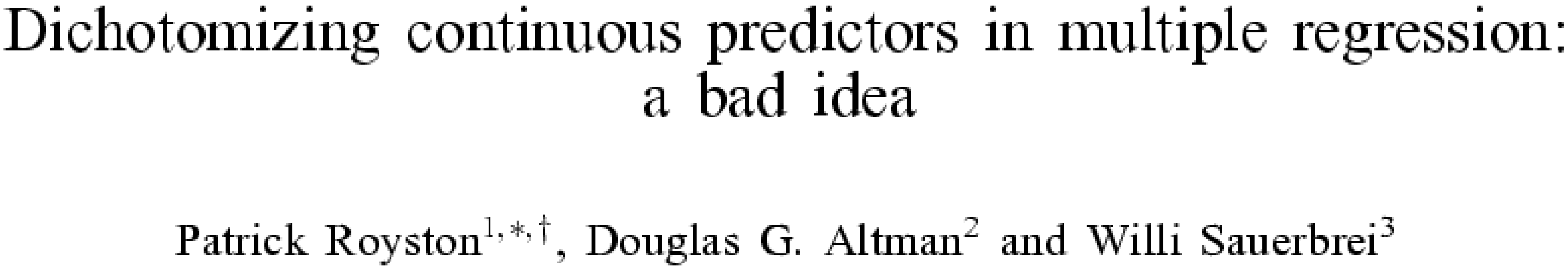
The youngest group is always in blue.

(a)‘Optimal’ (37 years); HR (older vs younger) 0.54, p= 0.004

(b) median (53 years); HR (older vs younger) 1.1, p= 0.4

(c) predefined from earlier analyses (45, 60years);

(d) popular (10-year groups)



Dichotomizing continuous predictors in multiple regression: a bad idea

Patrick Royston[1,*,†], Douglas G. Altman[2] and Willi Sauerbrei[3]

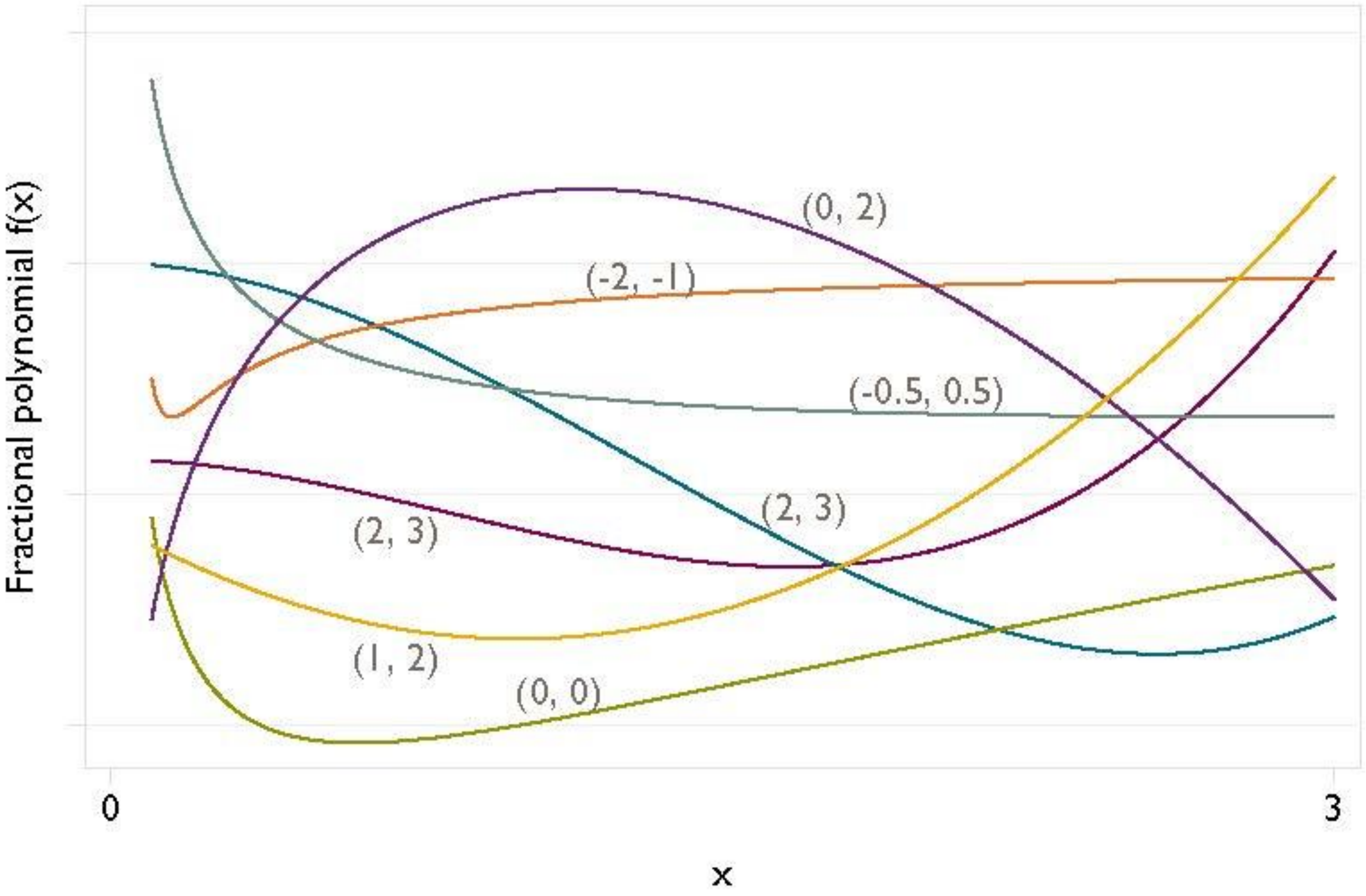
Statistics in Medicine, 2006, 25:127-141

Fractional polynomial models

- Describe for one covariate, *X*
- Fractional polynomial of degree *m* for *X* with powers $p_1, \dots, p_m$ is given by
$$\mathrm{FP}m(X) = \beta_1 X^{p1} + \dots + \beta_m X^{pm}$$
- Powers $p_1,\dots, p_m$ are taken from a special set
$$\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$$
- Usually $m = 1$ or $m = 2$ is sufficient for a good fit
- Repeated powers ($p_1=p_2$)
$$\beta_1 X^{p1} + \beta_2 X^{p1} \log X$$
- 8 FP1, 36 FP2 models

FP2 family

- simple but varying powers allow very different shapes



Our philosophy of function selection

- Prefer simple (linear) model
- Use more complex (non-linear) FP1 or FP2 model if indicated by the data

- Contrasts to more local regression modelling (eg splines)
  - Already starts with a complex model

- These issues influence our way of defining a **function selection procedure (FSP)**

FP analysis for the effect of age

Degree 1		Degree 2								
Power	Model	Powers		Model	Powers		Model	Powers		Model
	chi-square			chi-square			chi-square			chi-square
-2	6.41	-2	-2	17.09	-1	1	15.56	0	2	11.45
-1	3.39	-2	-1	17.57	-1	2	13.99	0	3	9.61
-0.5	2.32	-2	-0.5	17.61	-1	3	12.37	0.5	0.5	13.37
0	1.53	-2	0	17.52	-0.5	-0.5	16.82	0.5	1	12.29
0.5	0.97	-2	0.5	17.30	-0.5	0	16.18	0.5	2	10.19
1	0.58	-2	1	16.97	-0.5	0.5	15.41	0.5	3	8.32
2	0.17	-2	2	16.04	-0.5	1	14.55	1	1	11.14
3	0.03	-2	3	14.91	-0.5	2	12.74	1	2	8.99
		-1	-1	17.58	-0.5	3	10.98	1	3	7.15
		-1	-0.5	17.30	0	0	15.36	2	2	6.87
		-1	0	16.85	0	0.5	14.43	2	3	5.17
		-1	0.5	16.25	0	1	13.44	3	3	3.67

Function selection procedure (FSP)

Effect of age at 5% level?

	χ^2	df	p-value
Any effect?			
Best FP2 versus null	**17.61**	**4**	**0.0015**
Linear function suitable?			
Best FP2 versus linear	**17.03**	**3**	**0.0007**
FP1 sufficient?			
Best FP2 vs. best FP1	**11.20**	**2**	**0.0037**

Best FP2 (-2, -0.5) function selected

Many predictors – MFP

With many continuous predictors selection of best FP for each becomes more difficult → MFP algorithm as a standardized way to variable and function selection

(usually binary and categorical variables are also available)

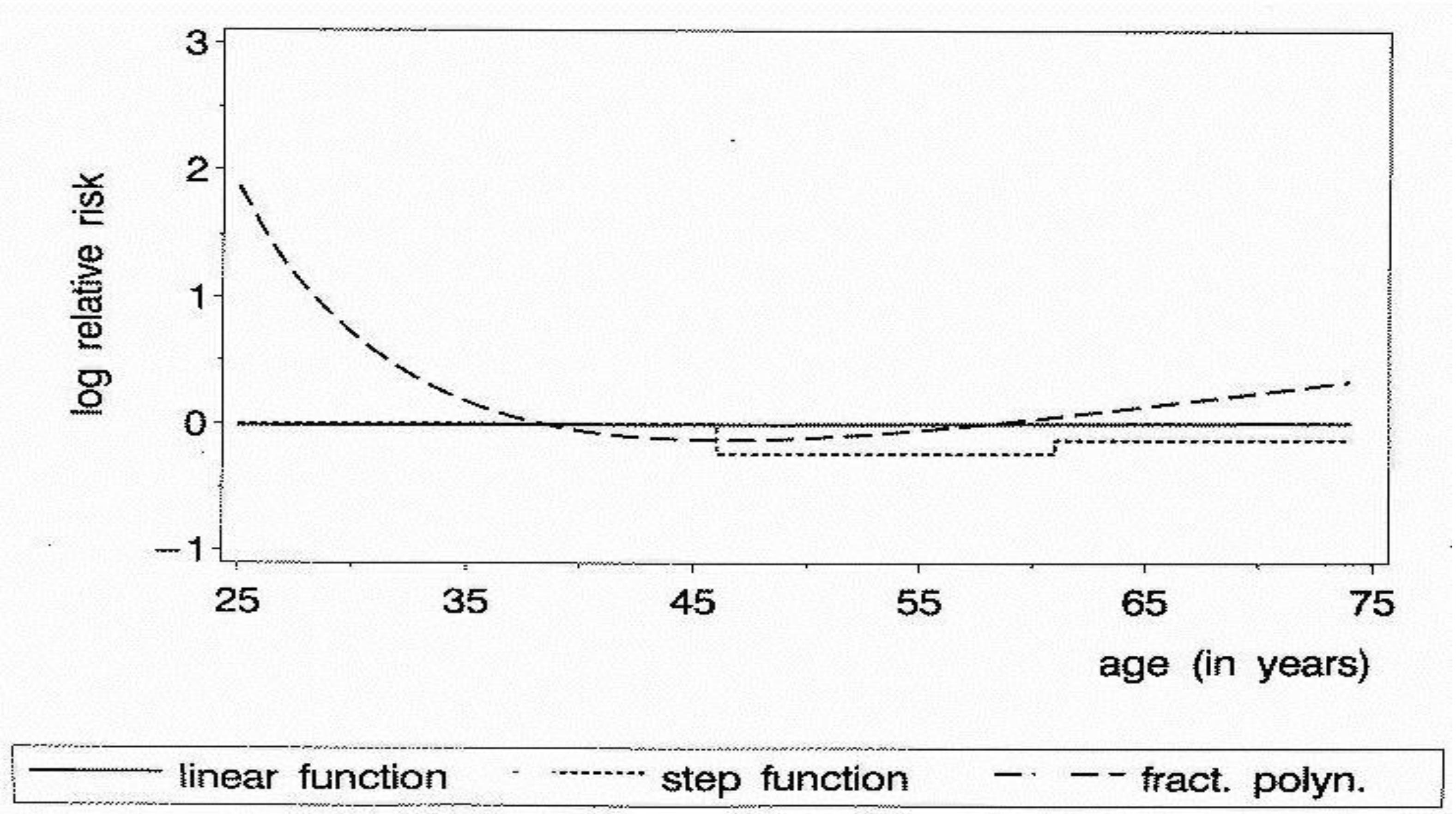
MFP algorithm combines

backward elimination with
FP function selection procedures

Continuous factors

Different results with different analyses

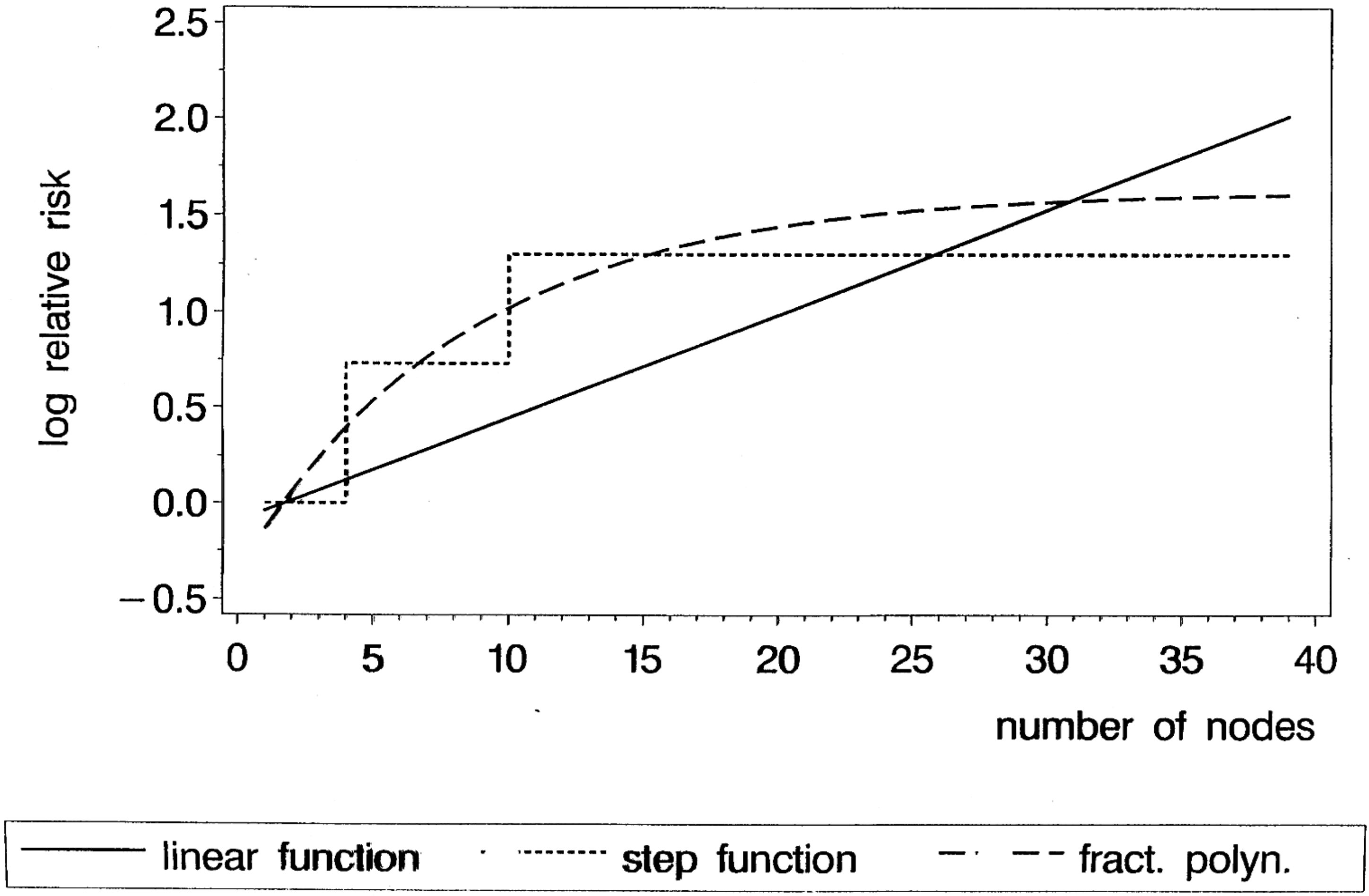
Age as prognostic factor in breast cancer (adjusted)



P-value 0.9 0.2 0.001

Results similar?

Nodes as prognostic factor in breast cancer (adjusted)



P-value 0.001 0.001 0.001

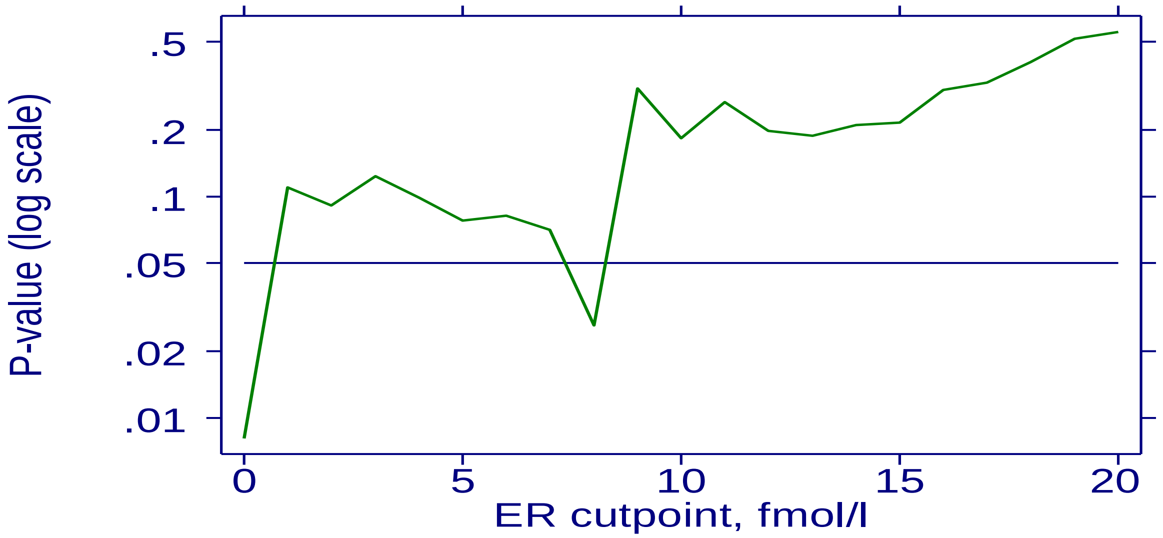
Back to interaction
- treatment and continuous covariates

- GBSG-study in breast cancer
- Hormonal treatment tamoxifen (TAM): yes/no
- Known from overviews that TAM interacts with oestrogen receptor status (ER) of primary tumour
- **But the research community needed many years to realize and to accept it**

- For illustration: investigate ER × TAM interaction

Interaction with treatment
- Standard approach for continuous covariates

- Based on binary predictor
- Need cut-point for continuous predictor
- Illustration - problem with cut-point approach



Interactions – MFPI method

- Have continuous *X* of interest, binary treatment variable *T* and other covariates *Z*
- Select ‘adjustment’ model *Z** on *Z* using MFP
- Find best FP2 function of *X* (in all patients) adjusting for *Z** and *T*
- Test FP2(*X*) × *T* interaction (2 d.f.)
  - Estimate β’s separately in 2 treatment groups
  - Standard test for equality of β’s
- May also consider simpler FP1 and linear functions

Royston and Sauerbrei (2004)

Interactions
- treatment effect function

- Have estimated two FP2 functions – one per treatment group
- Plot difference between functions against *X* to show the interaction
    - i.e. the treatment effect at different *X*
- Pointwise 95% CI shows how strongly the interaction is supported at different values of *X*
    - i.e. variation in the treatment effect

  (see WCC in renal cancer)

Slight variations were later proposed.

A large simulation study showed that MFPI has advantages to several other approaches (Royston and Sauerbrei 2013, 2014).

Interactions

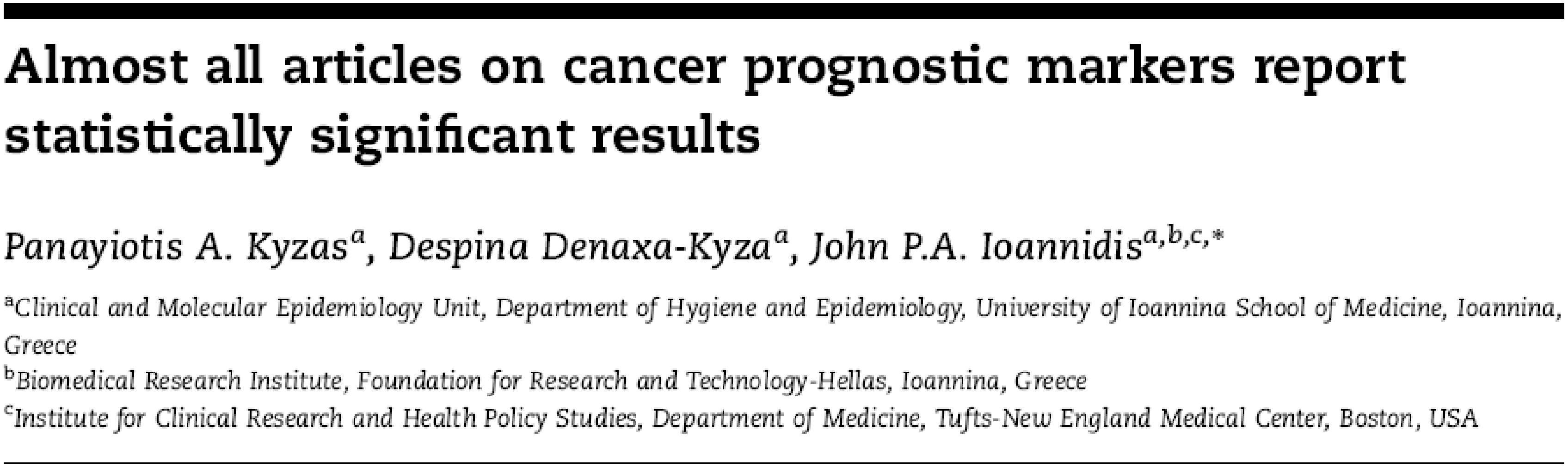
- Interactions are often ignored by analysts

- Continuous × categorical has been studied in FP context because clinically very important (MFPI)
- Continuous × continuous is more complex (MFPIgen)
- Interaction with time important for long-term FU survival data.
  Time-varying effects in survival data (MFPT)

FP methodology – further extensions

- Spike at zero approach
- Preliminary transformation to improve robustness
- Extend FP class to include sigmoid curves
- Add ‘local’ features

In medicine the literature is heavily criticized



Almost all articles on cancer prognostic markers report statistically significant results

Panayiotis A. Kyzas[a], Despina Denaxa-Kyza[a], John P.A. Ioannidis[a,b,c,*]

[a]Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece
[b]Biomedical Research Institute, Foundation for Research and Technology-Hellas, Ioannina, Greece
[c]Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Boston, USA

EJC 2007, 43:2559-79

Database 1: 340 articles included in meta-analysis
Database 2: 1575 articles published in 2005

In medicine the literature is heavily criticized

Reporting guidelines are one part to improve this situation

REporting recommendations for tumor MARKer prognostic studies (REMARK)

Lisa M McShane*, Douglas G Altman, Willi Sauerbrei, Sheila E Taube, Massimo Gion and Gary M Clark
for the Statistics Subcommittee of the NCI–EORTC Working Group on Cancer Diagnostics

published simultaneously in 5 journals, August 2005
Extended paper April 2012

The EQUATOR network acts as an umbrella
http://www.equator-network.org/

Reporting of MFP and MFPI analysis

The following variables were considered as candidates $x_1, \ldots x_k$

MFP(α_1, α_2) ; FP2 allowed

MFPI (α), adjusted for MFP(α_1, α_2) model

Candidates $x_1, \ldots x_k$

all continuous variables truncated (1%, 99%)

Important for **transparency** and **reproducible research**

Software sources MFP

- Most comprehensive implementation is in Stata
  - Command **mfp** is part since Stata 8 (now Stata 14)
- Versions for SAS and R are available
  - SAS
  - R version available on CRAN archive
    - **mfp** package
- Extensions to investigate interactions
    - So far only in Stata

SAS macro available on website

http://mfp.imbi.uni-freiburg.de/software

Concluding comments – MFP

- FPs use full information from continuous data – in contrast to a priori categorisation
- FPs search within flexible class of functions (FP1 and FP2 - 44 models)
- MFP is a well-defined multivariate model-building strategy – combines search for transformations with BE
- Important that model reflects medical knowledge, e.g. monotonic / asymptotic functional forms

Towards recommendations for model-building by selection of variables and functional forms for continuous predictors under several assumptions

Issue	Recommendation
Variable selection procedure	Backward elimination; significance level as key tuning parameter, choice depends on the aim of the study
Functional form for continuous covariates	Linear function as the 'default', check improvement in model fit by fractional polynomials. Check derived function for undetected local features
Extreme values or influential points	Check at least univariately for outliers and influential points in continuous variables. A preliminary transformation may improve the model selected. For a proposal see R & S 2007
Sensitivity analysis	Important assumptions should be checked by a sensitivity analysis. Highly context dependent
Check of model stability	The bootstrap is a suitable approach to check for model stability
Complexity of a predictor	A predictor should be 'as parsimonious as possible'

Sauerbrei et al. SiM 2007, Royston & Sauerbrei 2008

FPs and Big Data opportunities and challenges

- **large(r) n**
  - Test-based FP function selection (FSP)
    FSP needs to be adapted, for example replace p-value by improvement of area between curves (Govindarajulu et al., 2007)
    adaptation for categorical covariates needed
  - only monotonic functions – restrict to FP1 class
    non-monotonic functions – best FP2
  - required to investigate for interactions (MFPT, MFPIgen)
  - chances to validate a (MFP) model
- **large p, relatively small n (- omics)**
  - restrict to best FP1 transformation. Much better (Govindarajulu measure?) than linearity?
  - check for influential points (Boulesteix and Sauerbrei, 2011)

Current situations in the health sciences

Many claims…
often related to Big Data
(eg. personalized medicine)

… reality

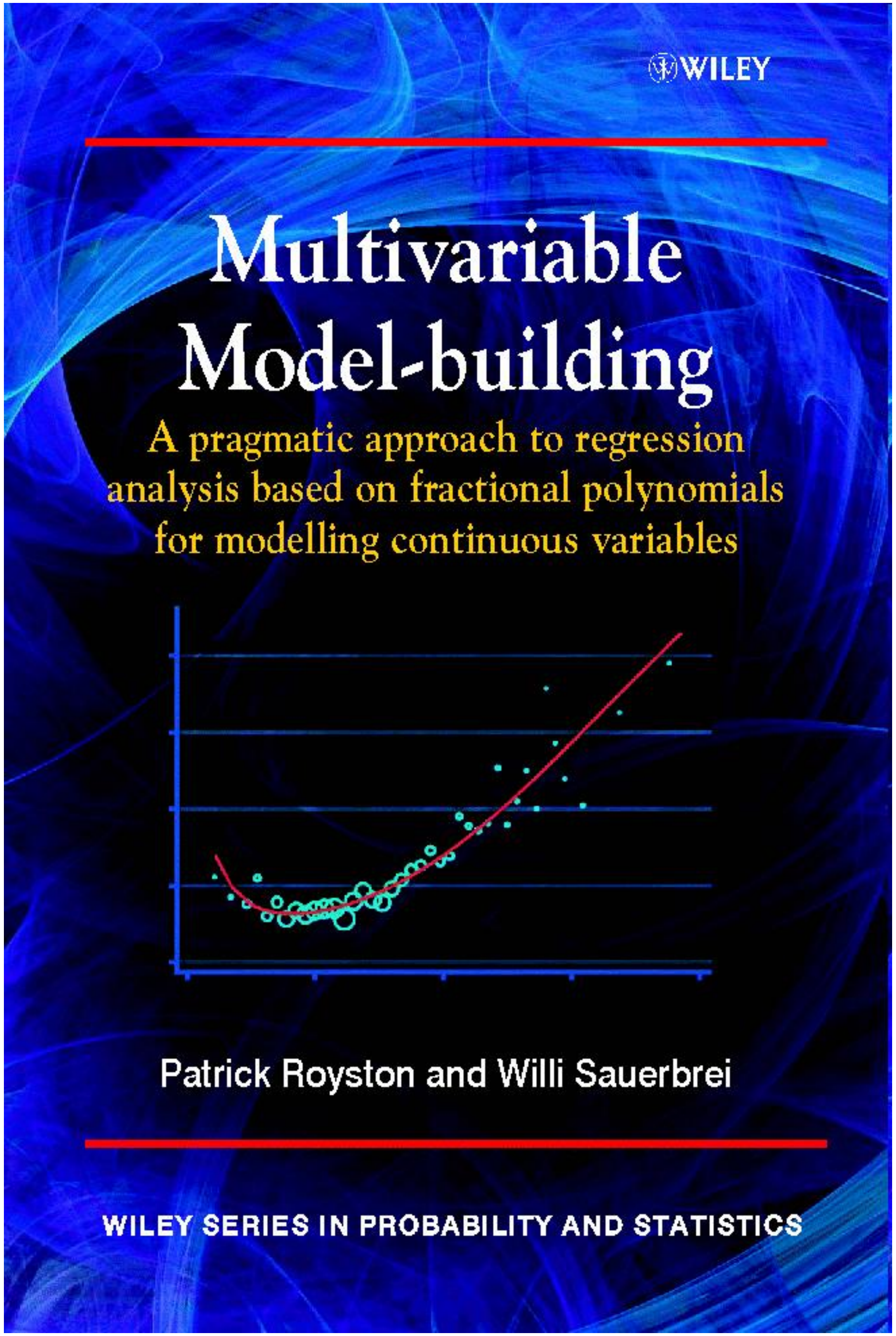
Lancet series 2014
Increasing value, reducing waste

Summary

- Usefulness of Big Data
  - what is the aim and how was the data collected?

- Statistical modeling
  - Getting the big picture right is more important than optimising aspects and ignoring others
    - strong predictors
    - strong non-linearity
    - strong interactions
    - strong non-PH in survival model

Guidance would be most helpful (is required!)



WILEY

Multivariable Model-building

A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables

Patrick Royston and Willi Sauerbrei

WILEY SERIES IN PROBABILITY AND STATISTICS

References

Hand DJ (2016): Editorial: ‘Big data’ and data sharing, Journal of the Royal Statistical Society, Series A 179, 3: 629–631.

Harford T (2014): Big data: are we making a big mistake? Financial Times, March 28 2014.

Rosenberg PS, Katki H, Swanson CA, Brown LM, Wacholder S and Hoover RN (2003): Quantifying epidemiologic risk factors using nonparametric regression: model selection remains the greatest challenge, Statistics in Medicine 22: 3369–3381.

Royston P, Altman DG (1994): Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). Applied Statistics, 43:429-467.

Royston P, Sauerbrei W (2004): A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Statistics in Medicine, 23:2509-2525.

Royston P, Sauerbrei W (2008): Multivariable Model-Building - A pragmatic approach to regression analysis based on fractional polynomials for continuous variables. Wiley.

Sauerbrei W, Royston P (1999): Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. Journal of the Royal Statistical Society A, 162:71-94.

Sauerbrei W, Royston P, Binder H (2007): Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Statistics in Medicine, 26:5512-28.

Sauerbrei W, Royston P, Zapien K (2007): Detecting an interaction between treatment and a continuous covariate: a comparison of two approaches. Computational Statistics and Data Analysis, 51:4054-4063.

Shmueli G (2010): To explain or to predict?, Statistical Science 3:289–310.

Claims

from many opinion articles, editorials, comments

- beginning of a golden age for patients and healthy people

- patients
  diagnosed much earlier/more correctly and
  treated accurately, efficiently, free of side effects by
  personalized medicine

- healthy people
  protected from becoming sick by perfect
  preventative healthcare

Antes, Gerd (2015) - A new Science(ability)?

Claims

achieved with systems medicine

Biological mechanisms of a pathogenesis (disease development) may be better understood by using methods from

- omics research
- systems biology
- computer science
- network theory

Antes, Gerd (2015) - A new Science(ability)?

Big Data

- Big Data is omnipresent as a universal tool

- Obstacles & barriers almost non-existent

- required are …
  - unlimited computer performance
  - unrestricted data acquisition (+ storage in limitless clouds)
  - free access to data and resulting journal articles
    → Open Access

- obtaining necessary resources seems to be easy

Claims…

and reality

… this Land of Cockaigne contrasts with dark, conservative world of science …

… where distortions, aberrations, failure and waste are commonplace (rather the rule than the exception)…

Antes, Gerd (2015) - A new Science(ability)?

The Lancet Research: Increasing Value, Reducing Waste Series

How should medical science change?

In 2009, we published a Viewpoint by Iain Chalmers and Paul Glasziou called "Avoidable waste in the production and reporting of research evidence", which made the extraordinary claim that as much as 85% of research investment was wasted.

Our belief is that research funders, scientific societies, school and university teachers, professional medical associations, and scientific publishers (and their editors) can use this Series as an opportunity to examine more forensically why they are doing what they do—the purpose of science and science communication—and whether they are getting the most value for the time and money invested in science.

Kleinert and Horton 2014

Of 1575 reports about cancer prognostic markers published in 2005, 1509 (96%) detailed at least one significant prognostic variable. However, few identified biomarkers have been confirmed by subsequent research and few have entered routine clinical practice. This Pattern — initially promising findings not leading to improvements in health care — has been recorded across biomedical research. So why is research that might transform health care and reduce health problems not being successfully produced?

Global biomedical and public health research involves billions of dollars and millions of people. In 2010, expenditure on life sciences (mostly biomedical) research was US$240 billion. The USA is the largest funder, with about $70 billion in commercial and $40 billion in governmental and non-profit funding annually, representing slightly more than 5% of US health-care expenditure. Although this vast enterprise has led to substantial health improvements, many more gains are possible if the waste and inefficiency in the ways that biomedical research is chosen, designed, done, analysed, regulated, managed, disseminated, and reported can be addressed.

Macleod et al. 2014

Increasing value – reducing waste

Five papers on

- priorities
- design, conduct and analysis
- biomedical research regulation and management
- addressing inaccessible research
- incomplete or unusable reports

Publication system…

- publication system is not complete in the presentation of results

  → collectively “turning a blind eye”

- globally consistent publication rate of ~50 % (of studies in recent decades)

in addition to 20,000 randomized clinical trials (annually recorded in Medline) there are a further 20,000 that are **never published**

‘vanished’ trials not a random selection

Overestimation of treatment effects

… can therefore **not be used as knowledge base** upon which to make decisions

Antes, Gerd (2015) - A new Science(ability)?

More data – will it solve every problem?

Does it come from a well planned study – or is it ‚found data‘?

- methodological statements only followed in sheer astonishment
- belief in correlation as the sole carrier of information
- age of causality is over → now in new era of correlation
- search for explanations unnecessary & waste of resources
  → trust in the power of data and its correlations
- correlations can indeed show connections, but in terms of causality can also be extremely misleading (only mentioned in passing and only regarded as problem if there is not enough data available)

→ Hence, "more data" solves every problem

Antes, Gerd (2015) - A new Science(ability)?