

General aims of the STRATOS initiative illustrated by issues in variable selection and function selection

Willi Sauerbrei

for the STRATOS initiative

Institute for Medical Biometry and Medical Informatics,
Medical Center - University of Freiburg, Germany

Overview

- Background of the STRATOS initiative
- TG 2 – Variable and function selection
 - Issues in variable selection
 - Which functional form for continuous variables?
 - Requirements for evidence supported guidance

Statistical methodology – Current situation

- Statistical methodology has seen some substantial development
- Computer facilities can be viewed as the cornerstone
- Possible to assess properties and compare complex model building strategies using simulation studies
- Resampling and Bayesian methods allow investigations that were impossible two decades ago
- Wealth of new statistical software packages allow a rapid implementation and verification of new statistical ideas

Unfortunately, many sensible improvements are ignored in practical statistical analyses

Reasons why improved strategies are ignored

- Overwhelming concern with **theoretical aspects**
- Very **limited guidance** on key issues that are **vital in practice**, discourages analysts from utilizing more sophisticated and possibly more appropriate methods in their analyses

Statistical methodology – problems are well known

The severeness of problems is even discussed in the public press:

The Economist ‘Unreliable research: Trouble at the lab.’ (October 2013):

“Scientists’ grasp of statistics has not kept pace with the development of complex mathematical techniques for crunching data. Some scientists use **inappropriate techniques** because those are the ones **they feel comfortable with**; others latch on to **new ones without understanding their subtleties**. Some just rely on the **methods built into their software**, even if they **don’t understand them.**”

The Lancet Research: Increasing Value, Reducing Waste Series

Comment (Introduction 1)

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.

Comment (Introduction 2)

Biomedical research: increasing value, reducing waste

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.

....

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

Improvement

At least two tasks are essential

- **Experts** in specific methodological areas have to work towards **developing guidance documents**
- An ever-increasing need for **continuing education** at all stages of the career
- For busy applied researchers it is often difficult to follow methodological progress even in their principal application area
 - Reasons are diverse
 - Consequence is that analyses are often deficient
- **Knowledge** gained through research on statistical methodology needs to be **transferred** to the broader community
- Many **analysts** would be **grateful for** an overview on the current **state of the art** and for **practical guidance documents**

Aims of the initiative

- **Provide guidance documents** for highly relevant issues in the design and analysis of observational studies
- As the statistical **knowledge** of the analyst **varies** substantially, guidance has to keep this background in mind. **Guidance** documents have to be provided **at several levels**
- For the **start** we will concentrate on **state-of-the-art** documents and the necessary evidence
- Help to identify questions requiring much more primary research

The overarching long-term aim is to improve key parts of design and statistical analyses of observational studies in practice

STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative

Willi Sauerbrei,^{a*†} Michal Abrahamowicz,^b
Douglas G. Altman,^c Saskia le Cessie,^d and[‡] James Carpenter^e
on behalf of the STRATOS initiative

Statistics in Medicine 2014

2011	ISCB Ottawa, Epidemiology Sub-Comm.	Preliminary ideas
2012	ISCB Bergen	Discussions, SG
2013	ISCB Munich	Initiative launched
2014-16	ISCB	Invited Sessions

<http://www.stratos-initiative.org/>

Basic information

Topic Group		Chairs and further members	
1	Missing data	Chairs:	James Carpenter, Kate Lee
		Members:	Melanie Bell, Els Goetghebeur, Joe Hogan, Rod Little, Andrea Rotnitzky, Kate Tilling, Ian White
2	Selection of variables and functional forms in multivariable analysis	Chairs:	Michal Abrahamowicz, Aris Perperoglou, Willi Sauerbrei
		Members:	Heiko Becher, Harald Binder, Frank Harrell, Georg Heinze, Patrick Royston, Matthias Schmid
3	Initial data analysis	Chairs:	Marianne Huebner, Saskia le Cessie, Werner Vach
		Members:	Maria Blettner, Dianne Cook, Heike Hofmann, Hermann-Josef Huss, Lara Lusa
4	Measurement error and misclassification	Chairs:	Laurence Freedman, Victor Kipnis
		Members:	Raymond Carroll, Veronika Deffner, Kevin Dodd, Paul Gustafson, Ruth Keogh, Helmut Küchenhoff, Pamela Shaw, Janet Tooze
5	Study design	Chairs:	Mitchell Gail
		Members:	Doug Altman, Gary Collins, Luc Duchateau, Neil Pearce, Peggy Sekula, Elizabeth Williamson, Mark Woodward
6	Evaluating diagnostic tests and prediction models	Chairs:	Gary Collins, Carl Moons, Ewout Steyerberg
		Members:	Patrick Bossuyt, Petra Macaskill, Ben van Calster, Andrew Vickers
7	Causal inference	Chairs:	Els Goetghebeur
		Members:	Bianca De Stavola, Saskia le Cessie, Niels Keiding, Erica Moodie, Ingeborg Waernbaum, Michael Wallace
8	Survival analysis	Chairs:	Michal Abrahamowicz, Per Kragh Andersen, Terry Therneau
		Members:	Richard Cook, Pierre Joly, Torben Martinussen, Maja Pohar-Perme, Jeremy Taylor
9	High-dimensional data	Chairs:	Lisa McShane, Joerg Rahnenfuehrer
		Members:	Axel Benner, Harald Binder, Anne-Laure Boulesteix, Tomasz Burzykowski, W. Evan Johnson, Lara Lusa, Stefan Michiels, Sherri Rose

Cross-cutting panels

Panels		Chairs
1	Glossary (GP)	Simon Day, Marianne Huebner, Jim Slattery
2	Data Sets (DP)	Saskia Le Cessie, Aris Perperoglou, Hermann Huss
3	Publications (PP)	Stephen Walter
		Co- Chairs: Bianca De Stavola, Mitchell Gail, Petra Macaskill
4	New Membership (MP)	James Carpenter, Willi Sauerbrei
5	Website (WP)	Joerg Rahnenfuehrer, Willi Sauerbrei
6	Literature Review (RP)	Gary Collins, Carl Moons
7	Simulation Studies (SP)	Michal Abrahamowicz, Harald Binder
8	Contact with Other Societies and Organizations (OP)	Willi Sauerbrei
9	Knowledge Transfer (TP)	Suzanne Cadarette

On requirements for an evidence supported guidance document

—

Issues in variable and function selection

(consider low dimensional data and not 'too small' sample sizes)

TG2: Selection of variables and functional forms in multivariable analysis

In multivariable analysis, it is common to have a **mix of binary, categorical (ordinal or unordered) and continuous variables** that may influence an outcome. While **TG6** considers the situation where the **main task is predicting the outcome** as accurately as possible, the main focus of **TG2** is to **identify influential variables** and gain insight into their individual and joint relationship with the outcome. Two of the (interrelated) **main challenges** are **selection of variables** for inclusion in a multivariable explanatory model and **choice of the functional forms for continuous variables**.

[...] The effects of **continuous predictors are typically modeled by either categorizing** them (which raises such issues as the number of categories, cutpoint values, implausibility of the resulting step-function relationships, local biases, power loss, or invalidity of inference in case of data-dependent cutpoints) **or assuming linear relationships** with the outcome, possibly after a simple transformation (e.g. logarithmic or quadratic). Often, however, the reasons for choosing such conventional representation of continuous variables are not discussed and the **validity of the underlying assumptions is not assessed**.

To address these limitations, statisticians have developed flexible modeling techniques based on various types of smoothers, including **fractional polynomials** and **several 'flavors' of splines**.

[...] collaborations with other TGs to account for such **complexities** as **missing data, measurement errors, time-varying confounding** or issues specific to modeling continuous predictors in survival analyses.

TG2: Part 1 – Selection of variables

- A large number of methods proposed (for many decades)
- High-dimensional data triggered the development of further proposals
- Many issues

The following slides are taken from the ‘Statistics in Practice’ presentation at the meeting of the German Region of the Biometric Society, March 2016

<http://www.biometrische-gesellschaft.de/arbeitsgruppen/weiterbildung/education-for-statistics-in-practice.html>

Education for Statistics in Practice, DAGStat 2016

Variable selection – a review and recommendations for the practicing statistician


Updated version!

Georg Heinze & Daniela Dunkler
Medical University of Vienna
CeMSIIS – Section for Clinical Biometrics

georg.heinze@meduniwien.ac.at, daniela.dunkler@meduniwien.ac.at


Focus of this presentation

- Methods and consequences of variable selection



Complexity is your enemy. Any fool can make something complicated. It is hard to keep things simple.

Sir Richard Branson
founder of Virgin Group

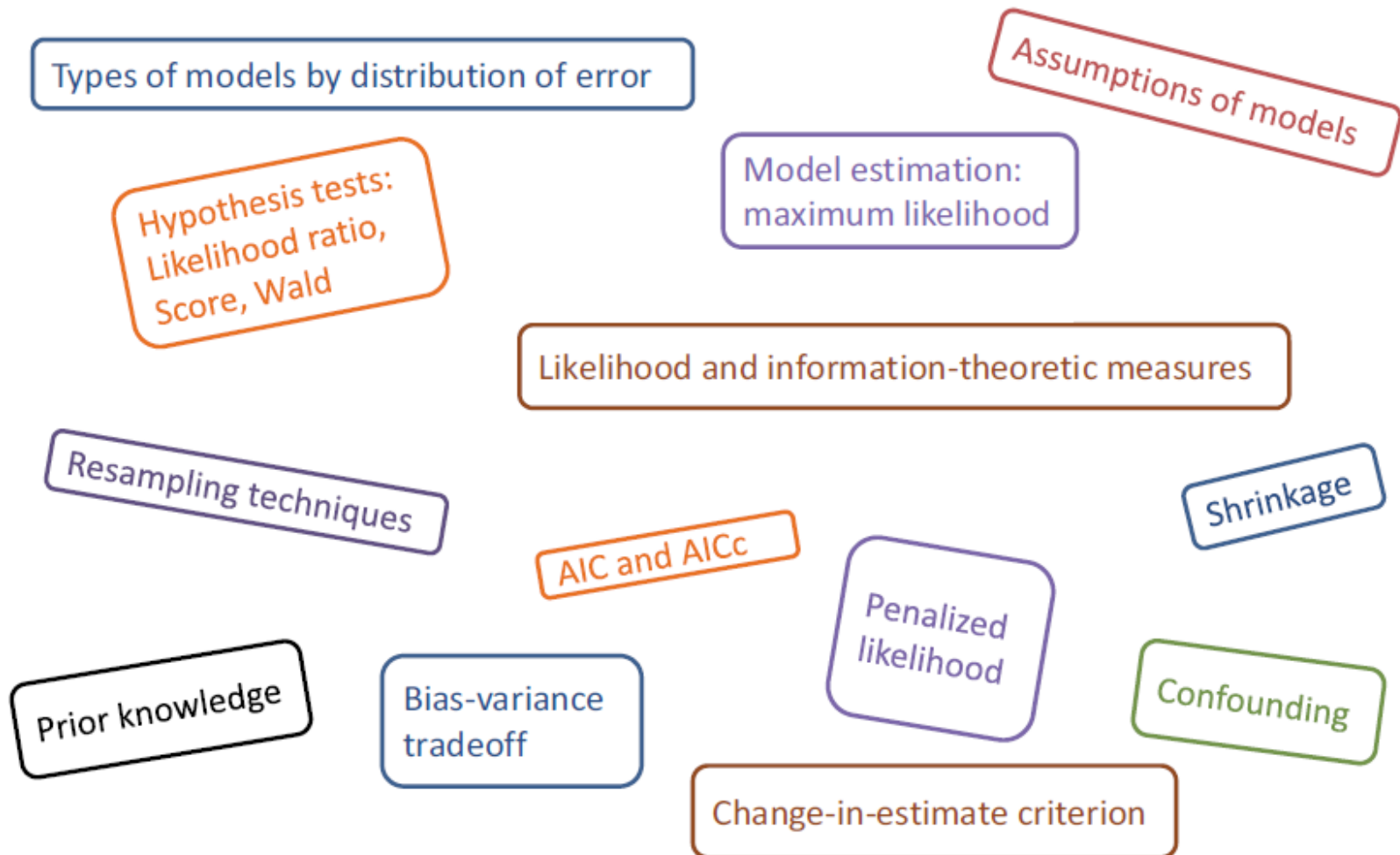


When I do my own makeup, I keep things pretty simple.

(Jordana Brewster)

izquotes.com

Statistical prerequisites



Basic algorithms

- 'Full' model
- Univariable filtering
- Best subset selection
- Forward selection
- Backward elimination
- Change-in-estimate: Purposeful variable selection and augmented backward selection
- Information-theoretic approach
- Directed acyclic graph (DAG)-based selection

TG2: Part 2 - Continuous variables

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

Rosenberg PS et al, Statistics in Medicine 2003; 22:3369-3381

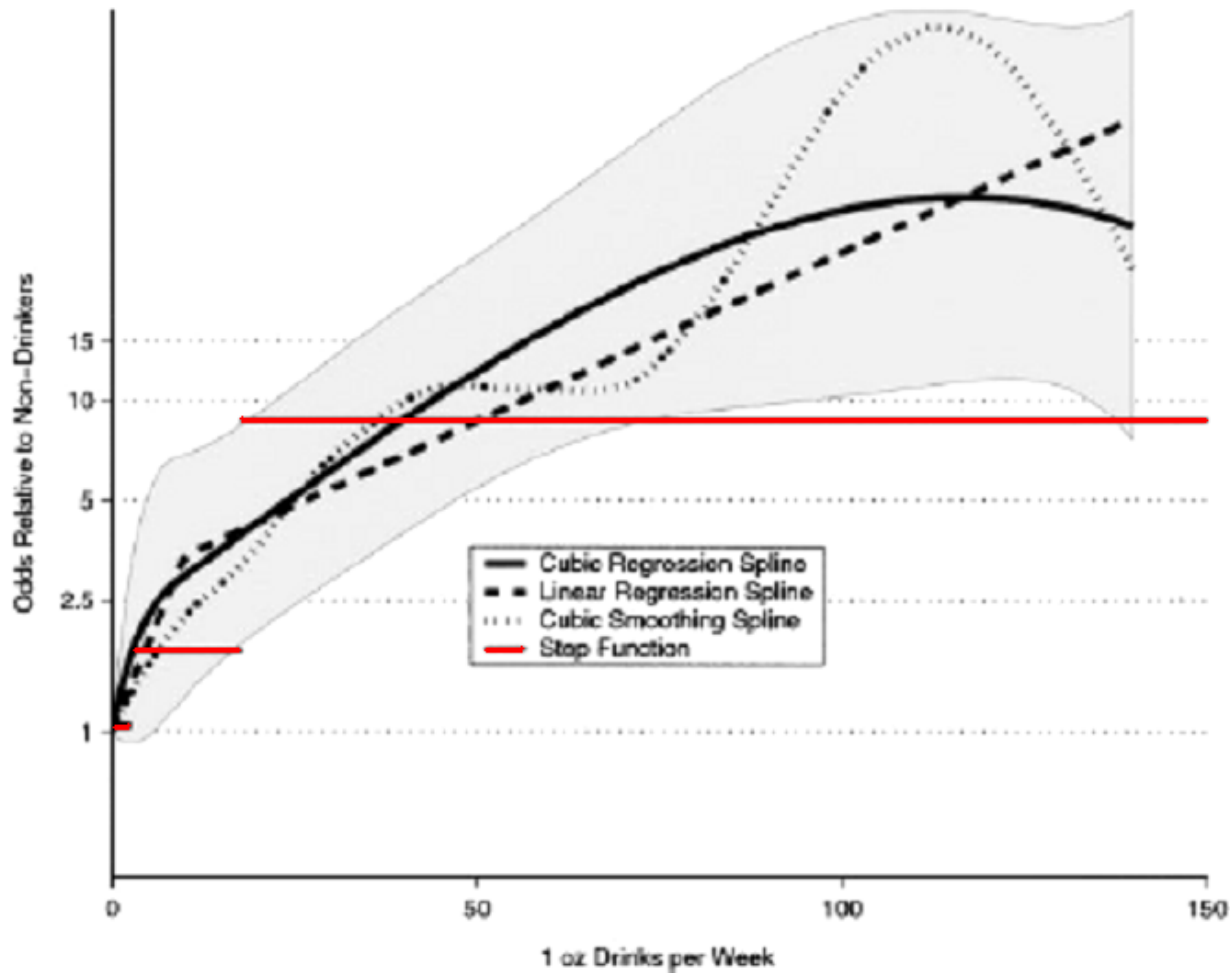
Discussion of issues in (univariate) modelling with splines

Trivial nowadays to *fit* almost any model

To *choose* a good model is much harder

Continuous risk factor different analyses – different results

Alcohol consumption as risk factor for oral cancer



Rosenberg et al, StatMed 2003

Continuous variables – which functional form?

1) Traditional approaches

a) Linear function

- may be inadequate functional form
- misspecification of functional form may lead to wrong conclusions

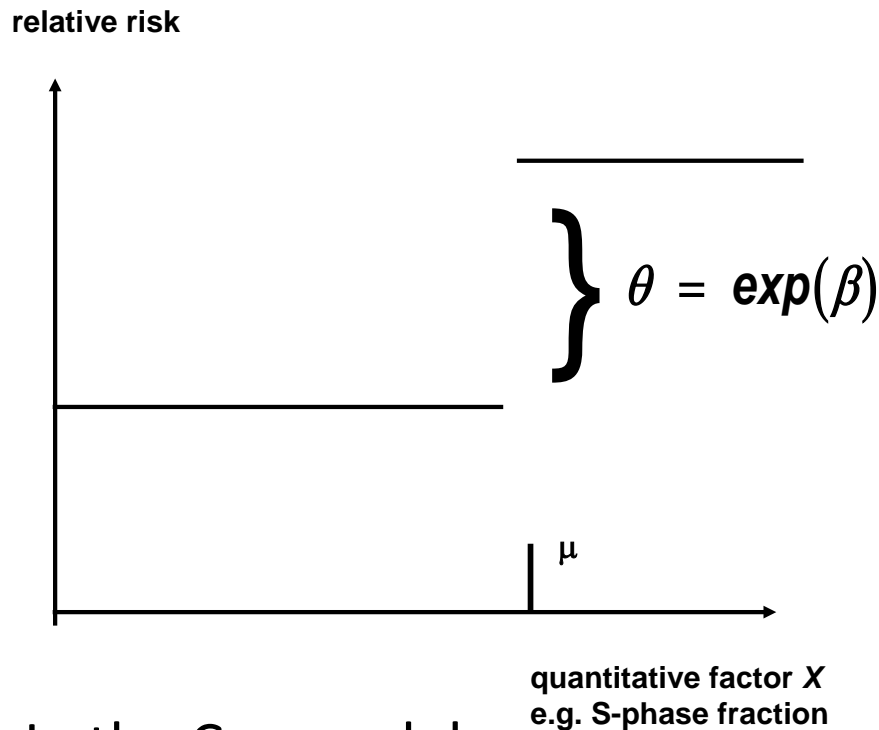
b) 'best' 'standard' transformation

c) Step function (categorical data)

- Loss of information
- How many cutpoints?
- Which cutpoints?
- Bias introduced by outcome-dependent choice

2) Flexible modeling techniques

Step function – the cutpoint problem



In the Cox model

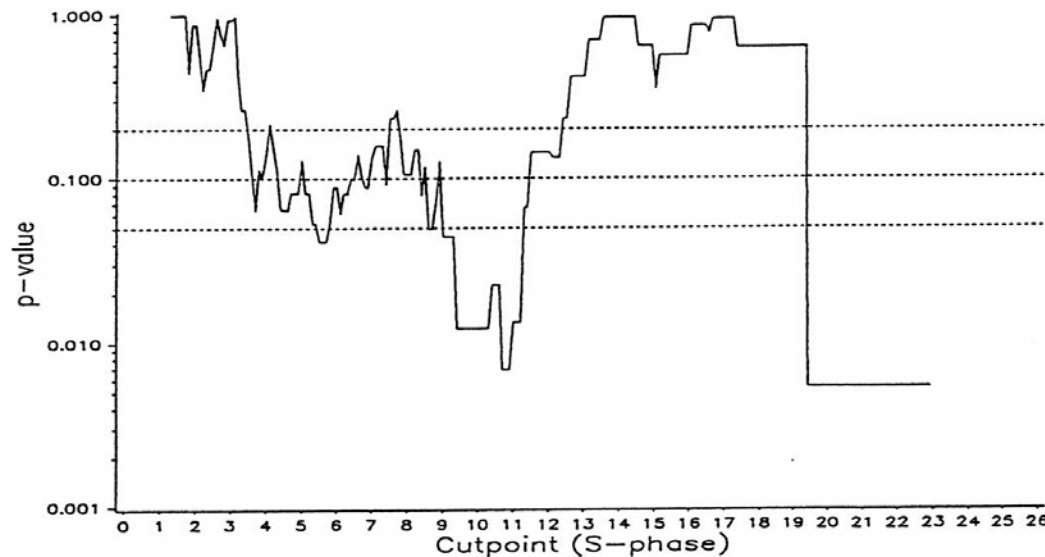
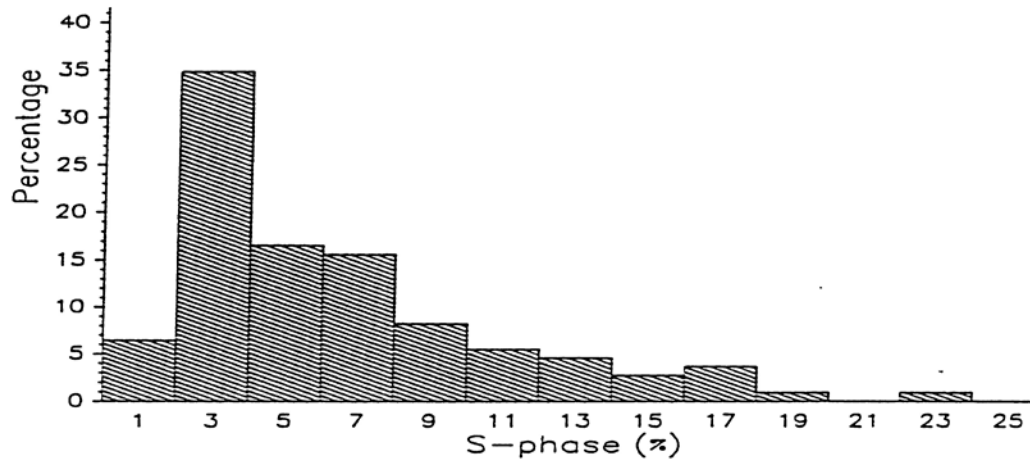
$$\lambda(t|X > \mu) = \exp \beta \lambda(t|X \leq \mu)$$

$\hat{\mu}$: estimated cutpoint for the comparison
of patients with X above and below μ .

Step function – biologically plausible?

Searching for optimal cutpoint minimal p-value approach

SPF in Freiburg DNA study



Problems
multiple testing
⇒ inflated type I error

biased estimates

different cutpoints in each
study

Example 1: Prognostic factors

GBSG-study in node-positive breast cancer

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

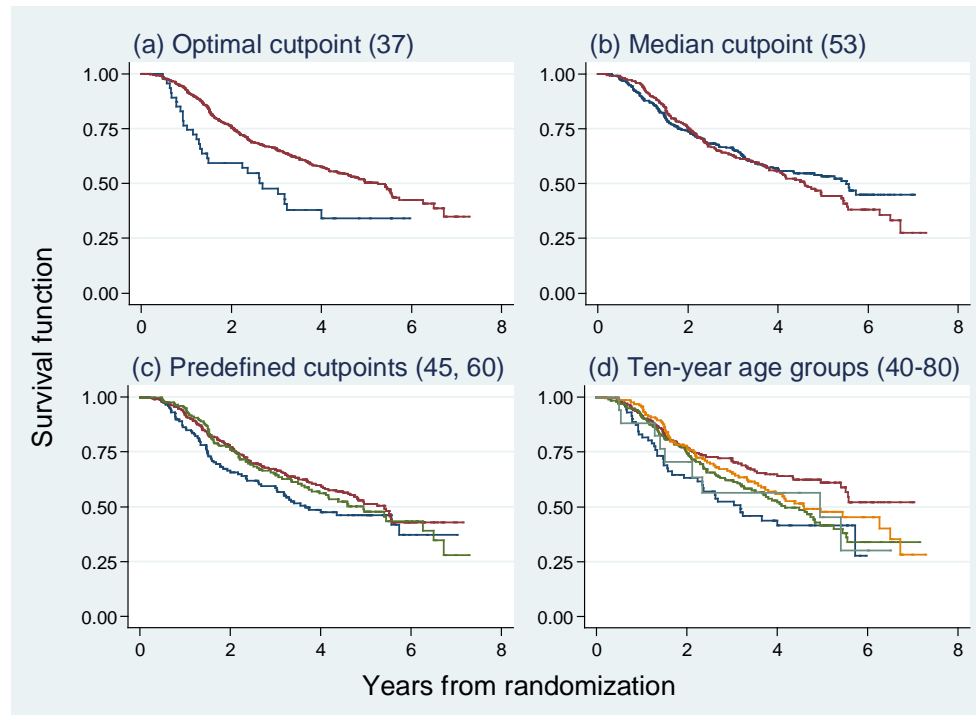
7 prognostic factors, of which **5** are continuous

Tamoxifen yes/no

We will consider

- age as prognostic factor
- estrogen receptor as predictive factor

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, 60 years);

(d) popular (10-year groups)

Dichotomizing continuous predictors in multiple regression: a bad idea

Patrick Royston^{1,*†}, Douglas G. Altman² and Willi Sauerbrei³

StatMed 2006, 25:127-141

Fractional polynomials

Fractional polynomials and the multivariable fractional polynomial (MFP) approach

Royston and Altman (1994)

Sauerbrei and Royston (1999)

Royston and Sauerbrei (2008)

The MFP approach combines

- *Selection of variables by using backward elimination (BE) with*
- *Selection of fractional polynomial (FP) functions of continuous variables*

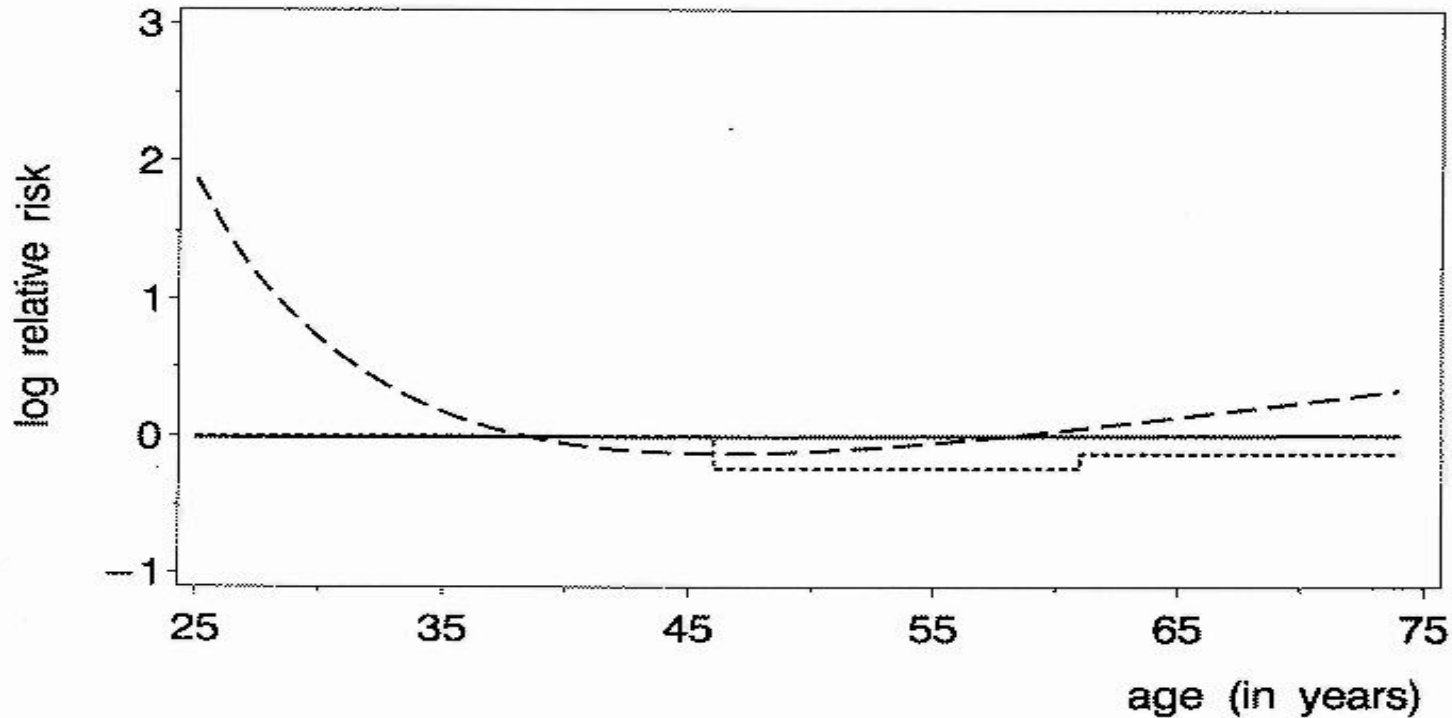
Although relatively simple and easily understood by researchers familiar with the basics of regression models, the selected models often extract most of the important information from the data. Models derived are **relatively easy to interpret and to report, a pre-requisite for transportability and general use in practice.**

Easy to use software is available.

<http://mfp.imbi.uni-freiburg.de/>

Continuous factors different analyses - different results

Age as prognostic factor in breast cancer (adjusted)



—	linear function	step function	- - -	fract. polyn.
---	-----------------	-------	---------------	-------	---------------

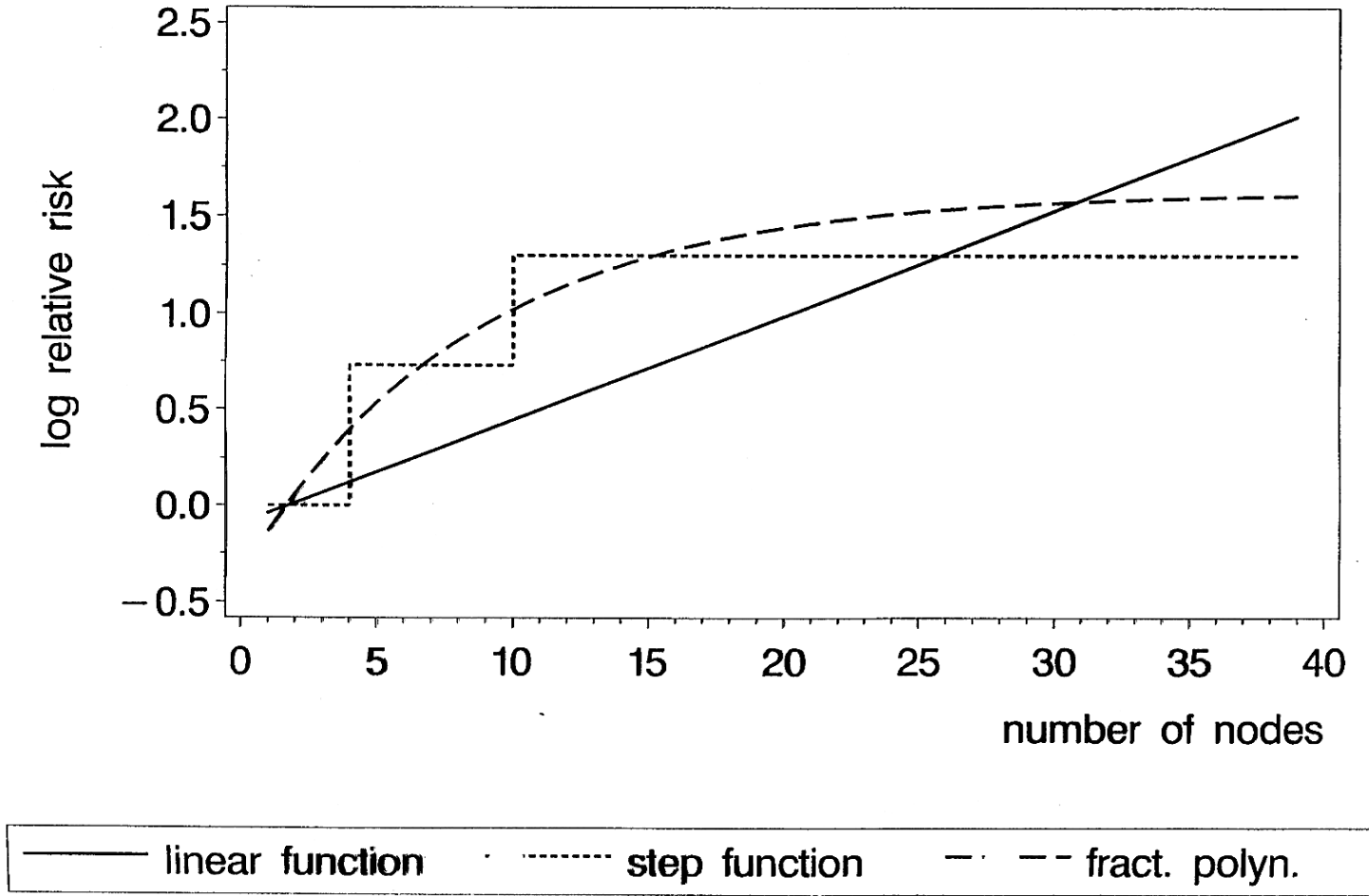
P-value 0.9

0.2

0.001

Results similar?

Nodes as prognostic factor in breast cancer (adjusted)



P-value 0.001

0.001

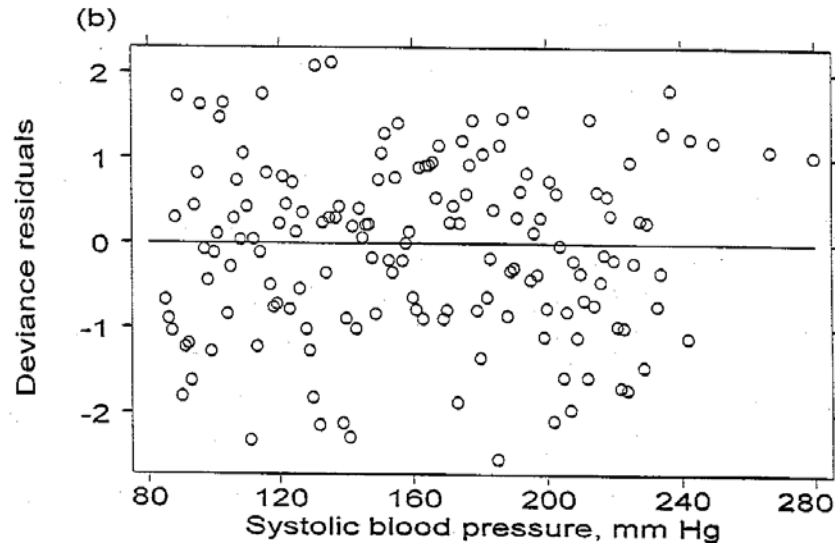
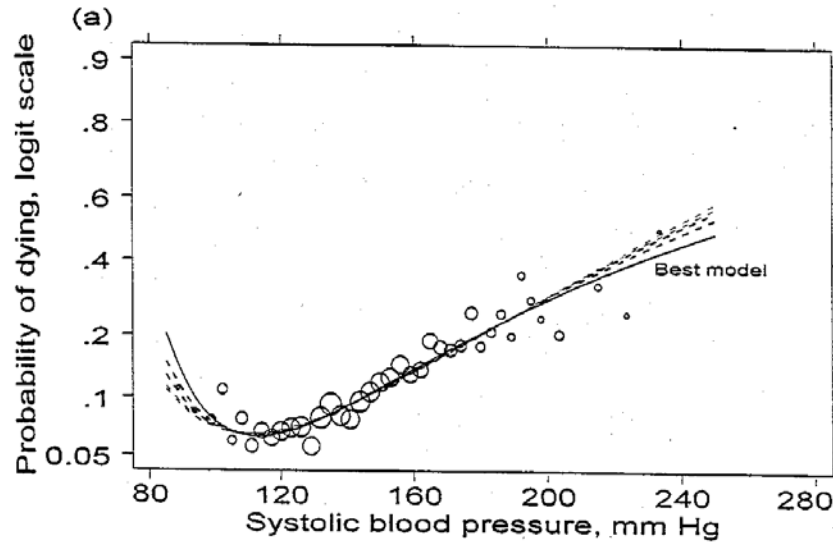
0.001

Example 2: Risk factors

- Whitehall 1
 - 17,370 male Civil Servants aged 40-64 years, 1670 (9.7%) died
 - Measurements include: age, cigarette smoking, BP, cholesterol, height, weight, job grade
 - Outcomes of interest: all-cause mortality at 10 years
⇒ logistic regression

FP analysis for systolic BP

Similar fit of several functions – no problem



Continuous risk factors - Presentation in categories

Whitehall 1 - Systolic blood pressure

Odds ratio from final FP(2) model

$$\text{LogOR} = 2.92 - 5.43X^{-2} - 14.30 * X^{-2} \log X$$

Presented in categories

Systolic blood pressure (mm Hg)		Number of men		OR (model-based)	
Range	ref. point	at risk	dying	Estimate	95%CI
≤ 90	88	27	3	2.47	1.75, 3.49
91-100	95	283	22	1.42	1.21, 1.67
101-110	105	1079	84	1.00	-
111-120	115	2668	164	0.94	0.86, 1.03
121-130	125	3456	289	1.04	0.91, 1.19
131-140	135	4197	470	1.25	1.07, 1.46
141-160	150	2775	344	1.77	1.50, 2.08
161-180	170	1437	252	2.87	2.42, 3.41
181-200	190	438	108	4.54	3.78, 5.46
201-240	220	154	41	8.24	6.60, 10.28
241-280	250	5	4	15.42	11.64, 20.43

Steps towards guidance documents

Selection of multivariable models for explanation (TG2)

- **Strategies for variable selection**
 - Better understanding of advantages and disadvantages
 - Role of model complexity, stability and shrinkage
- **Review of the literature about methods**
 - Strategies used in practice
 - Comparison of strategies for model building
- **Comparison of spline procedures**
- **Specific role of 'spike at zero' variables?**
- **Comparison of approaches for variable selection and choices of functional form**
- **Guidance documents for variable and function selection**

Summary

Many analyses have severe weaknesses – missing guidance is one of the main reasons

Variable and function selection - many issues

- A large number of variable selection strategies has been proposed
- There are several spline based procedures
- Hardly any informative comparisons

How to derive evidence to support guidance documents??

- Theoretical investigations?
- Large and meaningful simulation studies!!!
- Good examples