

# Introduction of the STRATOS initiative and its main aims

Willi Sauerbrei  
for the STRATOS initiative

<sup>1</sup>Medical Center – University of Freiburg, Germany

**STRATOS**  
INITIATIVE



# Overview

- Necessity of guidance
- Aims and structure of the STRATOS initiative
- Issues in variable and function selection
  1. Selection of variables
  2. Selection of functional forms
  3. Combining the two parts

# 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.**”

## 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 evidence supported guidance** 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** has 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,<sup>a\*†</sup> Michal Abrahamowicz,<sup>b</sup>  
Douglas G. Altman,<sup>c</sup> Saskia le Cessie,<sup>d</sup> and<sup>‡</sup> James Carpenter<sup>e</sup>  
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
2016	Banff	Workshop
2016	IBC Victoria	Invited Session
2016	HEC Munich	Invited Session
2017	IBS-EMR Thessaloniki	Invited Session
2017	ISCB Vigo	Scientific topic
2017	CEN-ISBS Vienna	Invited Session

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

Basic information

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

Panels		Chairs and Co-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, 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, Douglas Altman
9	Knowledge Transfer (TP)	Suzanne Cadarette, Catherine Quantin

# On requirements for evidence supported guidance

Issues in variable and function selection

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

# General issue in observational Studies

Several variables, mix of continuous and (ordered) categorical variables, pairwise- and multicollinearity present

**Model selection required**

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

# Regression models

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

$g(\mathbf{x}) = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$  (assuming effects are linear)

## normal errors (linear) regression model

Y normally distributed

$$E(Y|X) = \beta_0 + g(X)$$

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

## logistic regression model

Y binary

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

## survival times

T survival time (partly censored)

Incorporation of covariates

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

# Aims of multivariable models

- Prediction of an outcome of interest
- Identification of 'important' predictors
- Adjustment for predictors uncontrollable by experimental design
- Stratification by risk
- ... and many more

# Key issues in model building

- Subject matter knowledge determines (parts) of a model. **Assumption** for the following:  
**data dependent model building required.**
- For medical decision making cutpoints are needed. We consider the **earlier step of deriving a model.**
- Important distinction between model for prediction and **model for explanation.** Our main interest is in the latter.

# Building multivariable regression models – some preliminaries

- ‚Reasonable‘ model class was chosen
- Comparison of strategies
  - Theory
    - only for limited questions, unrealistic assumptions
  - Examples or simulation
    - Examples based on published data
      - oversimplifies the problem
      - data clean
      - ‚relevant‘ predictors given
    - rigorous pre-selection → what is a full model?

## ... preliminaries continued

More problems are available,

see discussion on **initial data analysis** in Chatfield (2002) section ,*Tackling **real life** statistical **problems***'

see also Mallows (1998), The zeroth problem, Am. Stat.

TG3 – Initial Data Analysis

# 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

- Central issues:
  - Model with focus on prediction or explanation?
  - To select or not to select (full model)?
  - Which variables to include?
- A large number of methods proposed (for many decades)
- High-dimensional data triggered the development of further proposals
- Many critical issues

# (Traditional) methods for variable selection

## Full model

- variance inflation in the case of multicollinearity
  - Wald-statistic

Stepwise procedures  $\Rightarrow$  prespecified  $(\alpha_{in}, \alpha_{out})$  and actual significance level?

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

All subset selection  $\Rightarrow$  which criteria?

- $C_p$       Mallows
- AIC      Akaike Information Criterion
- BIC      Bayes Information Criterion

## Bayes variable selection

**MORE OR LESS COMPLEX MODELS?**

# Stepwise procedures

Central Issue:

- significance level  
choice depends on aim of the study

## Criticism

- **FS** and **StS** start with ,bad‘ univariate models (**underfitting**)
- **BE** starts with the full model (**overfitting**),  
less critical
- Multiple testing, P-values incorrect

Nevertheless very popular in practice

# Other procedures

- Bootstrap selection
- Change-in-estimate
- Variable clustering
- Incomplete principal components
- Penalized approaches (selection and shrinkage; Lasso, Garotte, SCAD, ...)
- Directed acyclic graph (DAG-) based selections
- 
- 
-

# Selection of variables: Statistical prerequisites

Types of models by distribution of error

Assumptions of models

Hypothesis tests:  
Likelihood ratio,  
Score, Wald

Model estimation:  
maximum likelihood

Likelihood and information-theoretic measures

Resampling techniques

AIC and AICc

Shrinkage

Prior knowledge

Bias-variance  
tradeoff

Penalized  
likelihood

Confounding

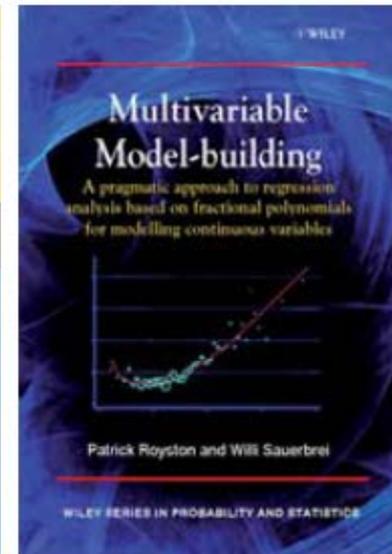
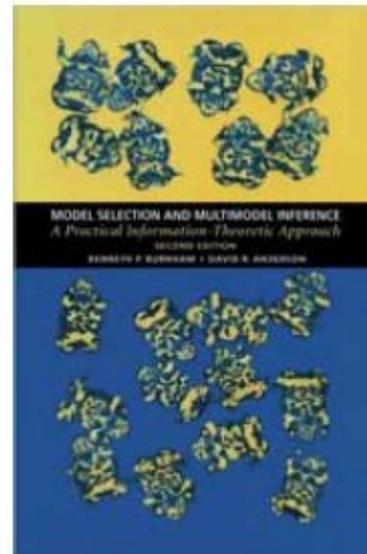
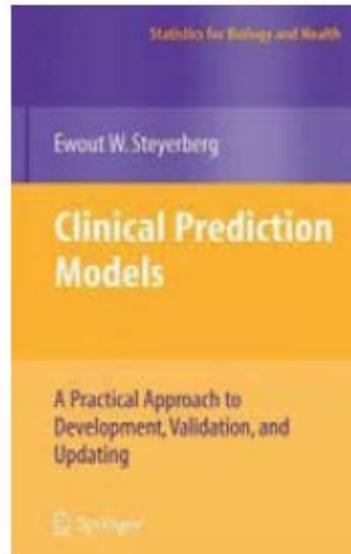
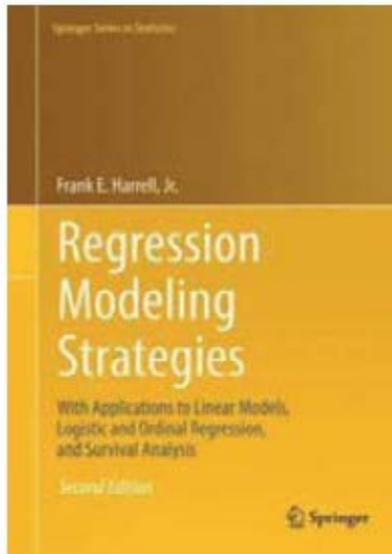
Change-in-estimate criterion

# Opinions on variable selection

for models with focus on prediction and explanation.



Variable selection



(Harrell, 2001; Steyerberg, 2009; Burnham & Anderson, 2002, Royston & Sauerbrei, 2008)

# "Recommendations" from the literature

We do **not know any** recommendation which is **supported by good evidence** from theory or meaningful simulation studies

# TG2: Part 2

## selection of functional forms

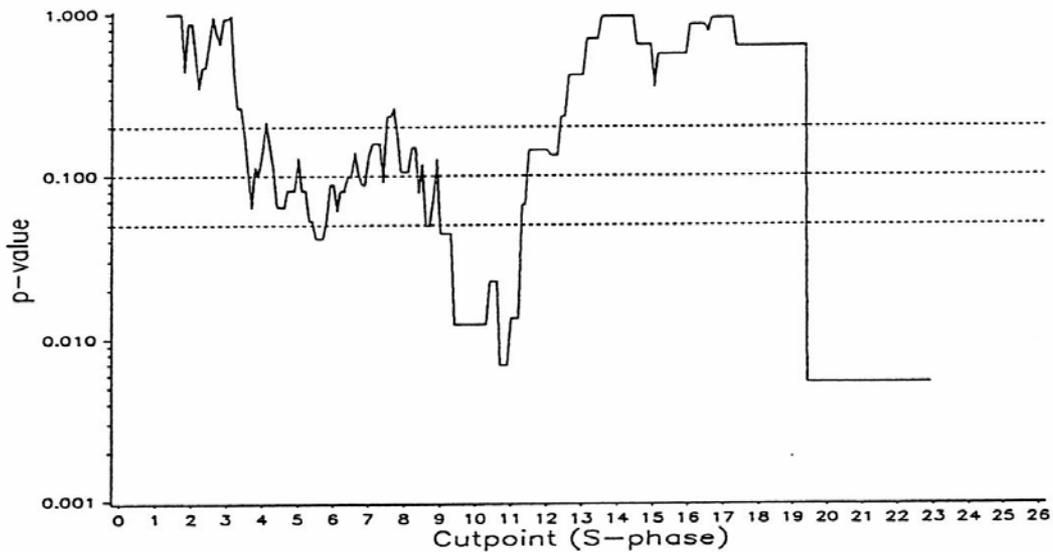
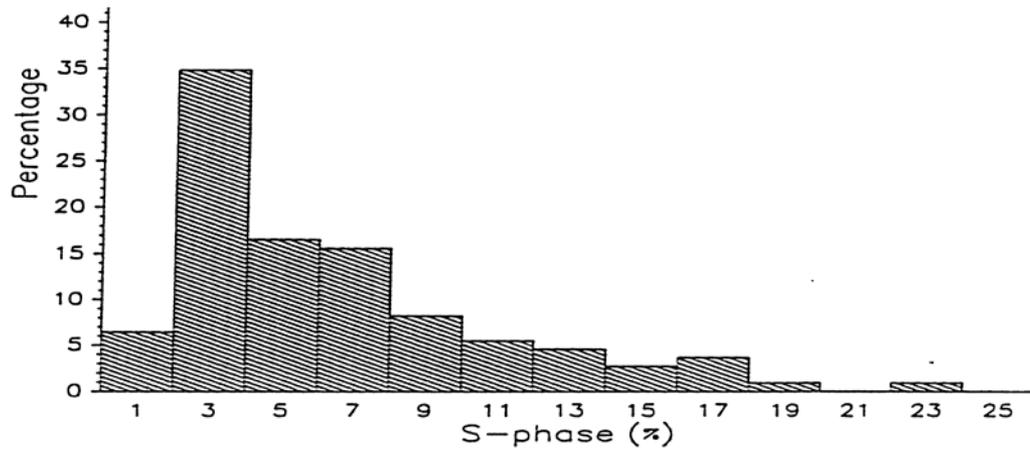
- Assume linearity
- Cut-points
- 'Optimal' cut-points
- Fractional polynomials
- Splines

# Functional forms:

## Models based on cut-points: problems!

- Cut-points are still popular in clinical and epidemiological research
- Use of cut-points in a model gives a step function
- How many cut-points?
- Where should the cut-points be put?
- Poor approximation to the true relationship
- Almost always fits the data less well than a suitable continuous function

# 'Optimal' cutpoint (better: minimal P-value approach)



# Optimal cutpoints: problems!

- Multiple testing  $\Rightarrow$  inflation of significance level
  - 40% instead of nominal 5%
- Inflated significance level does not disappear with increased sample size
- Large bias in estimate of difference between groups
- Results depend on chance
- Never reproducible – impossible to summarize across studies



# Continuous variables: modelling instead of categorization

- (Multivariable) Fractional Polynomials
- Spline based approaches

R packages

bs: B-Spline Basis for Polynomial Splines

ns: Generate a Basis Matrix for Natural Cubic Splines

early comparisons: see ISCB 2017 talk by Aris Perperoglou

# Other spline packages in R

Package	Description	Authors
gss	General Smoothing Splines	C Gu
polyspline	Polynomial spline routines	C Kooperberg
pspline	Penalized Smoothing Splines	B Ripley
cobs	Constrained B-Splines	PT Ng and M Maechler
crs	Categorical Regression Splines	JS Racine, Z Nie, BD R
bigsplines	Smoothing Splines for Large Samples	NE Helwig
bezier	Bezier Curve and Spline Toolkit	A Olsen
freesplines	Free-Knot Splines	S Spirti, P Smith, P Le
Orthogonal splinebasis	Orthogonal B-Spline Functions	A Redd
pbs	Periodic B Splines	S Wang
logspline	Logspline density estimation routines	C Kooperberg
episplineDensity	Density Estimation Exponential	S Buttrey, J Royset, R
Hmisc, rms	restricted cubic splines, plots	F Harrell

# A brief overview of regression packages in R

Package	Downloads	Vignette	Book	Website	Datasets
quantreg	2001231	X	X		7
mgcv	1438166	X	X		2
survival	1229305	X	X		33
VGAM	297308	X	X	X	50
gbm	271362			X	3
gam	168143		X	X	1
gamlss	78295	X	X	X	29

# TG2: Part 3 - Combining variable and function selection

**Two inter-related questions**, common to many multivariable explanatory models

Results of

- Data-dependent selections of independent variables may depend on
- decisions regarding functional forms of both
  1. the variable of interest (X)
  2. other variables, correlated with X

and *vice versa*



... for survival data (TG8)

... effects may vary in time

... **another interrelated issue**

# TG 2 – Some issues

- Which strategies for **variable selection** exist?  
What about their **properties**?
- **Data-dependent** modeling introduces **bias**.  
What about the role of **shrinkage** approaches?
- **Comparison of spline** procedures in a univariate context.  
**Which criteria** are relevant? Can we derive guidance for practice?
- What about variables with a '**spike-at-zero**'?
- **Multivariable procedures – selection of variables and functions**  
MFP well defined strategy  
Which of the spline based procedures?  
**Comparison in large simulation studies** needed
- **Multivariable procedures** and **correction for selection bias**  
How relevant? One step or two step approaches?  
E.g. **selection of variables and forms followed by shrinkage**
- **Big Data**  
Does it **influence properties** of procedures and their comparison?
- Role of model **validation**

**Much research required!**

# Conclusion

**We are far away from ‘state-of-the-art’ on selection of variables and functional forms**

Many more comparisons are urgently needed!

‘Exact distributional results are virtually impossible to obtain,  
even for simplest of common subset selection algorithms’

*Picard & Cook, JASA, 1984*

**=> Informative simulation studies are needed!**

## Stratos Initiative

The validity and practical utility of observational medical research depends critically on good study design, excellent data quality, appropriate statistical methods and accurate interpretation of results. Statistical methodology has seen substantial development in recent times. Unfortunately, many of these methodological developments are ignored in practice. Consequently, design and analysis of observational studies often exhibit serious weaknesses. The lack of guidance on vital practical issues discourages many applied researchers from using more sophisticated and possibly more appropriate methods when analyzing observational studies. Furthermore, many analyses are conducted by researchers with a relatively weak statistical background and limited experience in using statistical methodology and software. Consequently, even 'standard' analyses reported in the medical literature are often flawed, casting doubt on their results and conclusions. *An efficient way to help researchers to keep up with recent methodological developments is to develop guidance documents that are spread to the research community at large.*

These observations led to the initiation of the STRATOS (STRengthening Analytical Thinking for Observational Studies) initiative, a large collaboration of experts in many different areas of biostatistical research. The objective of STRATOS is to provide accessible and accurate guidance in the design and analysis of observational studies. The guidance is intended for applied statisticians and other data analysts with varying levels of statistical education, experience and interests (click to enlarge).

The Steering Group has decided to start with seven topics of general interest. Two topic groups were added later. Guidance documents will be developed by separate topic groups (TGs), each comprising experts in different area of statistical methodology, alongside applied researchers who may represent future end-users of the STRATOS documents. Each TG will start by developing guidance aimed primarily at level 2 statistical knowledge, which is perhaps slightly below state of the art. STRATOS structure and the initial road map (click to enlarge). The STRATOS initiative is closely connected to the International Society of Clinical Biostatistics (ISCB) and was launched at a half-day Mini-Symposium on the last day of the ISCB meeting in Munich, in August 2013.

### Panels

To co-ordinate the initiative, to share best research practices and to disseminate research tools and results from the work of the topic groups (TG), several cross-cutting panels have been started recently. They aim to develop recommendations (sometimes rather loose as for simulation studies, sometimes more strict as for STRATOS publications) and to provide the infrastructure for those aspects of the initiative that apply to all or most of the TGs, and to coordination of the efforts of the individual TGs. Recommendations aim to support, simplify and harmonize work within and across the TGs. They will also help increase transparency in deriving guidance documents in STRATOS.

The following Panels have been created to date:

MP	Membership	GP	Glossary	RP	Literature Review	BP	Bibliography
PP	Publications	SP	Simulation Studies	DP	Data Sets	TP	Knowledge Translation
WP	Website	CP	Contact Organizations				



### News

#### GMD 2017

Two sessions:  
September 17 - 21, 2017  
Oldenburg, Germany

#### DGEpl, DOM 2, GD SMP

Invited talk:  
September 5-6, 2017  
Lübeck, Germany

#### CEN IBS 2 joint conference

Two sessions:  
August 28 - September 1,  
2017  
Vienna, Austria

#### ISCB 2017

Two talks:  
July 9 - 13, 2017  
Vigo, Spain

[More news...](#)

### Topic groups

- 1 Missing data
- 2 Selection of variables and functional forms in multivariable analysis
- 3 Initial data analysis
- 4 Measurement error and misclassification
- 6 Study design
- 8 Evaluating diagnostic tests and prediction models
- 7 Causal Inference
- 5 Survival analysis
- 9 High-dimensional data

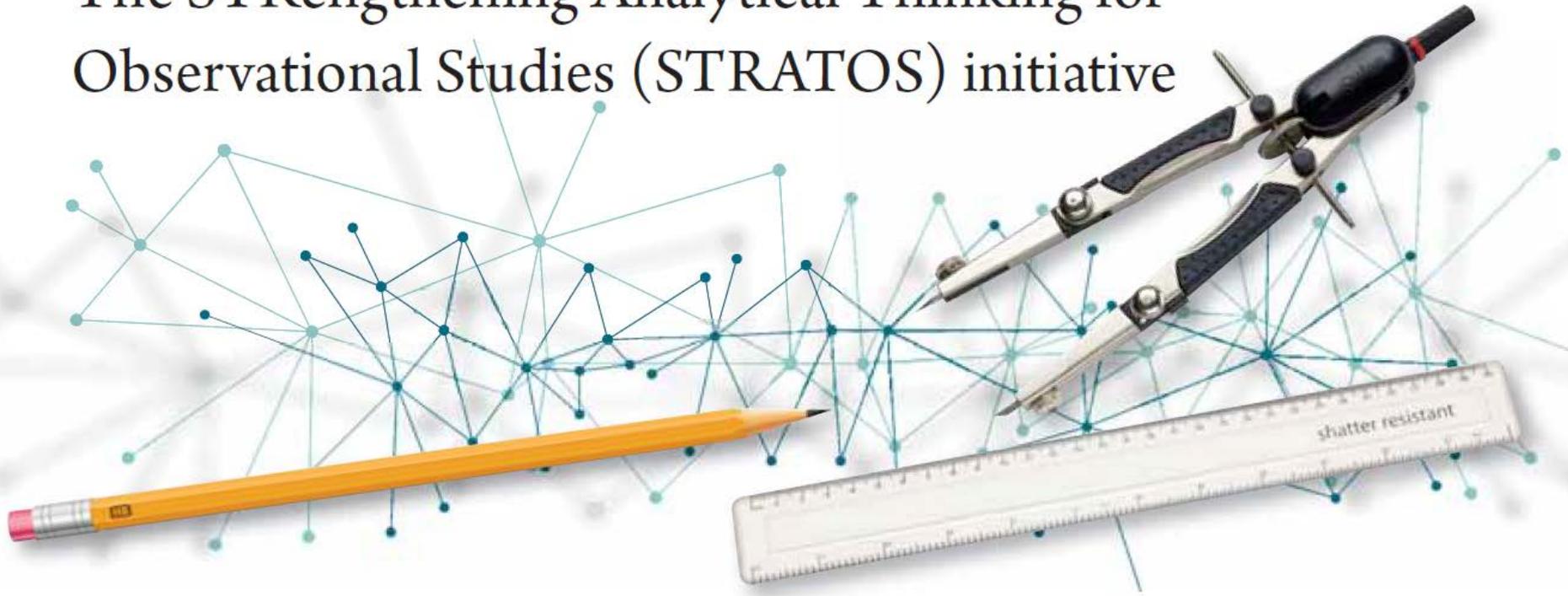
[All groups](#)

**Introductory Paper for Series in the IBS Biometric Bulletin**  
**STRATOS initiative – Guidance for designing and analysing observational studies**

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

# Guidance for designing and analysing observational studies:

The STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative



**Willi Sauerbrei<sup>1</sup>, Gary S. Collins<sup>2</sup>,  
Marianne Huebner<sup>3</sup>, Stephen D. Walter<sup>4</sup>,  
Suzanne M. Cadarette<sup>5</sup>, and  
Michal Abrahamowicz<sup>6</sup> on behalf of the  
STRATOS initiative**

# Thanks to all members of TG2 !

- Michal Abrahamowicz (Canada)
- Willi Sauerbrei (Germany)
- Aris Perperoglou (U.K.)
- Heiko Becher (Germany)
- Harald Binder (Germany)
- Frank Harrell (U.S.A.)
- Georg Heinze (Austria)
- Patrick Royston (U.K.)
- Matthias Schmid (Germany)