State-of-the-art in variable and functional form selection: research required!

Georg Heinze

Medical University of Vienna & TG2 of STRATOS Initiative

TG2 is about multivariable model-building

- Descriptive models:
  Capture the association of explanatory and outcome variables

- Predictive modeling:
  ‚Transparent' (as opposed to black-box) prediction models,
  often with superior performance
  background knowledge can be easily inserted

  TG 6

- Explanatory modeling:
  Designed to estimate an identifiable causal effect of interest directly
  or for prediction of counterfactual outcomes

  TG 7

TG2 aims

- Level-3: to evaluate what are the recommendable strategies and procedures for multivariable modeling building
- Level-2: to summarize state-of-the-art and key issues
  to give recommendations
- Level-1: to teach multivariable model building to non-statisticians
  to give recommendations

Ongoing work: State of the art

State-of-the-art in selection of variables and functional forms in multivariable analysis – outstanding issues

Willi Sauerbrei[1], Aris Perperoglou[2], Matthias Schmid[3], Michal Abrahamowicz[4], Heiko Becher[5], Harald Binder[1], Daniela Dunkler[6], Frank E. Harrell Jr[7], Patrick Royston[8], and Georg Heinze[6] for TG2 of the STRATOS initiative

arXiv:1907.00786 [stat.ME]

Crisis? What Crisis? (1)

- Two examples of problems in variable or functional form selection:

Example 1:

- European Heart Journal (IF 20.8)
- Linear regression, N=86
- 12 explanatory variables
- Univariate selection with $\alpha = 0.05$ to determine a ‚starting set'
- Then backward elimination with $\alpha = 0.05$
- Two variables survived this torture
- One of them was not mentioned in the list of candidate variables

Crisis? What Crisis? (2)

Example 2:

- JAMA Internal Medicine (IF 15)
- N=666,137
- Main exposure: ‚metabolic equivalent training' (MET) in hours/week
- For the main analysis, MET was categorized into
  0 h/w, 0.2-7.5, 7.7-15, 15.2-22.5, 22.7-40, 40.2-75, 75.2+

Comment & Response

November 2015

Physical Activity and Successful Aging

Even a Little Is Good

David Hupin, MD, MSc[1]; Frédéric Roche, MD, PhD[1]; Pascal Edouard, MD, PhD[1]

Author Affiliations | Article Information

JAMA Intern Med. 2015;175(11):1862-1863. doi:10.1001/jamainternmed.2015.4744

Figure. Hazard Ratios (HRs) and 95% CIs for Leisure Time Moderate- to Vigorous-Intensity Physical Activity and Mortality

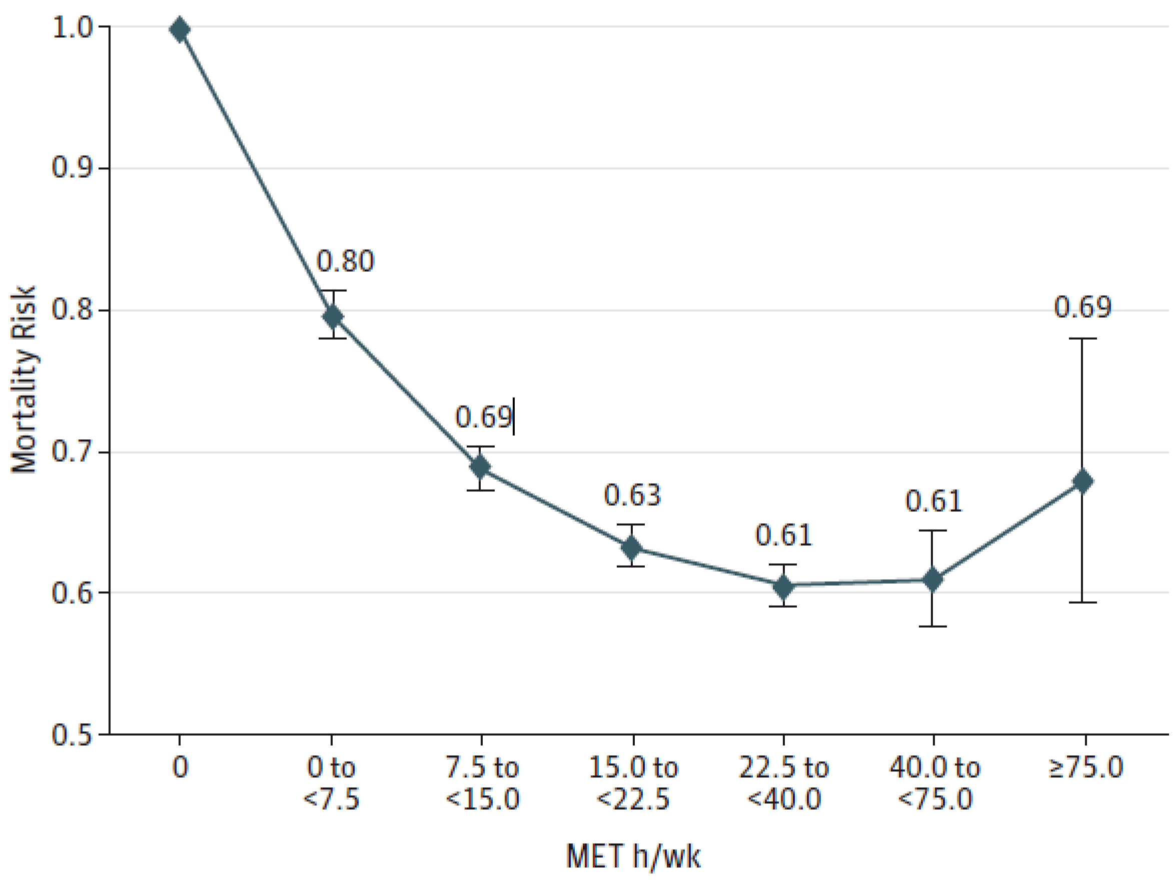
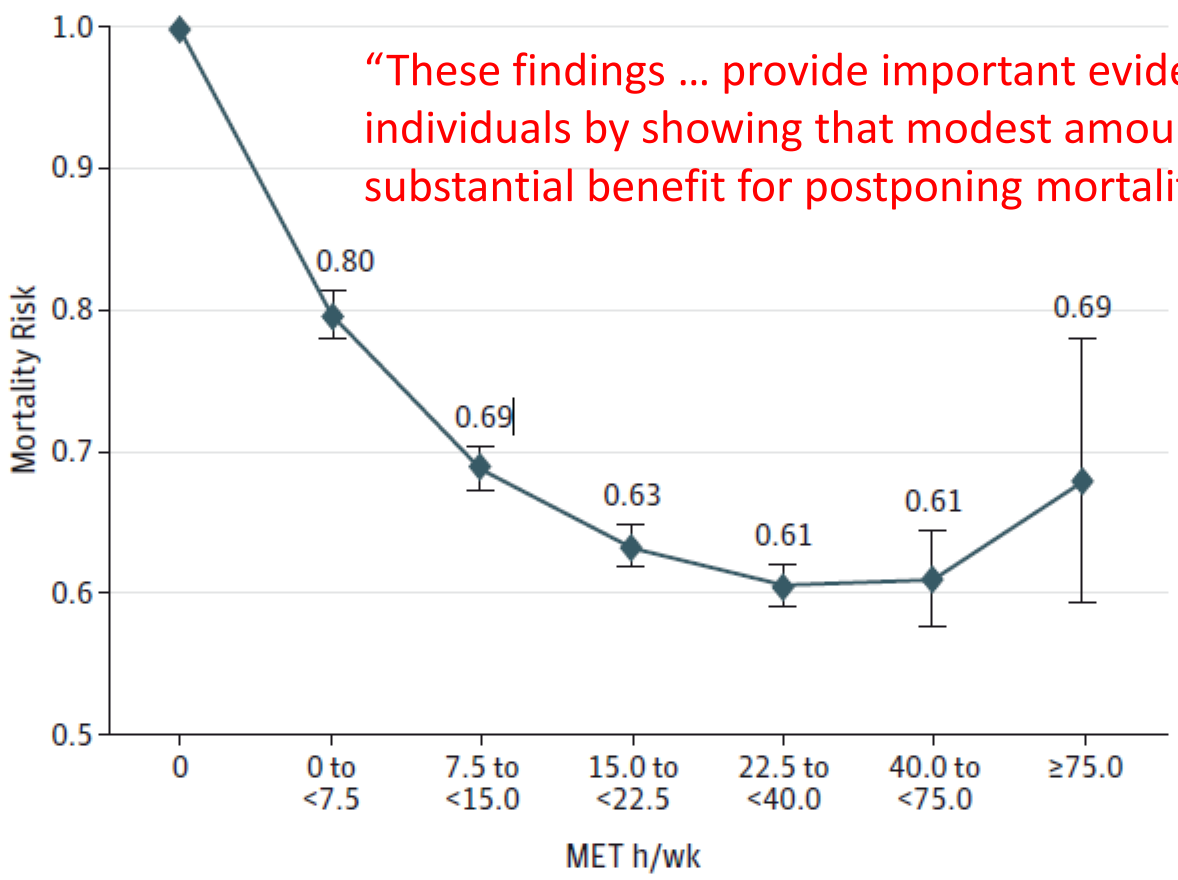
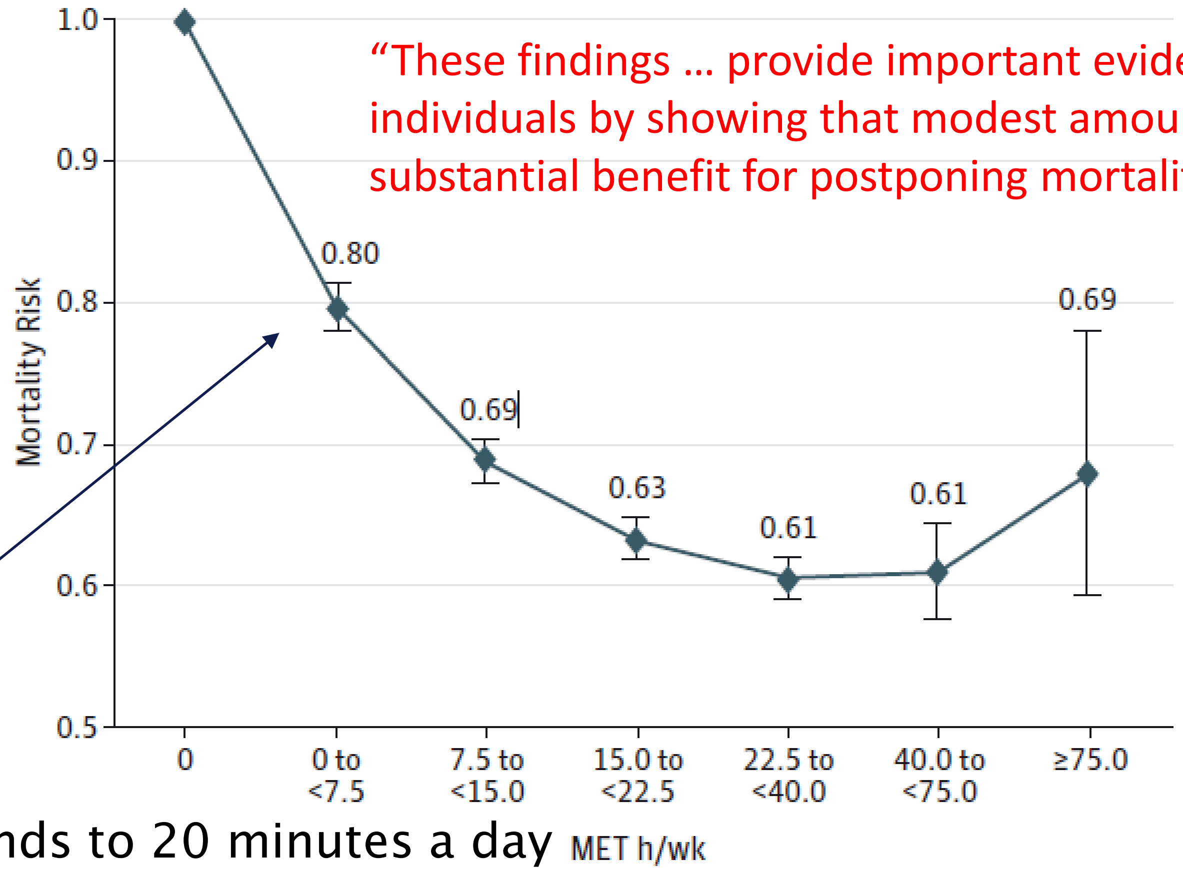


Figure. Hazard Ratios (HRs) and 95% CIs for Leisure Time Moderate- to Vigorous-Intensity Physical Activity and Mortality



"These findings … provide important evidence to inactive individuals by showing that modest amounts of activity provide substantial benefit for postponing mortality"

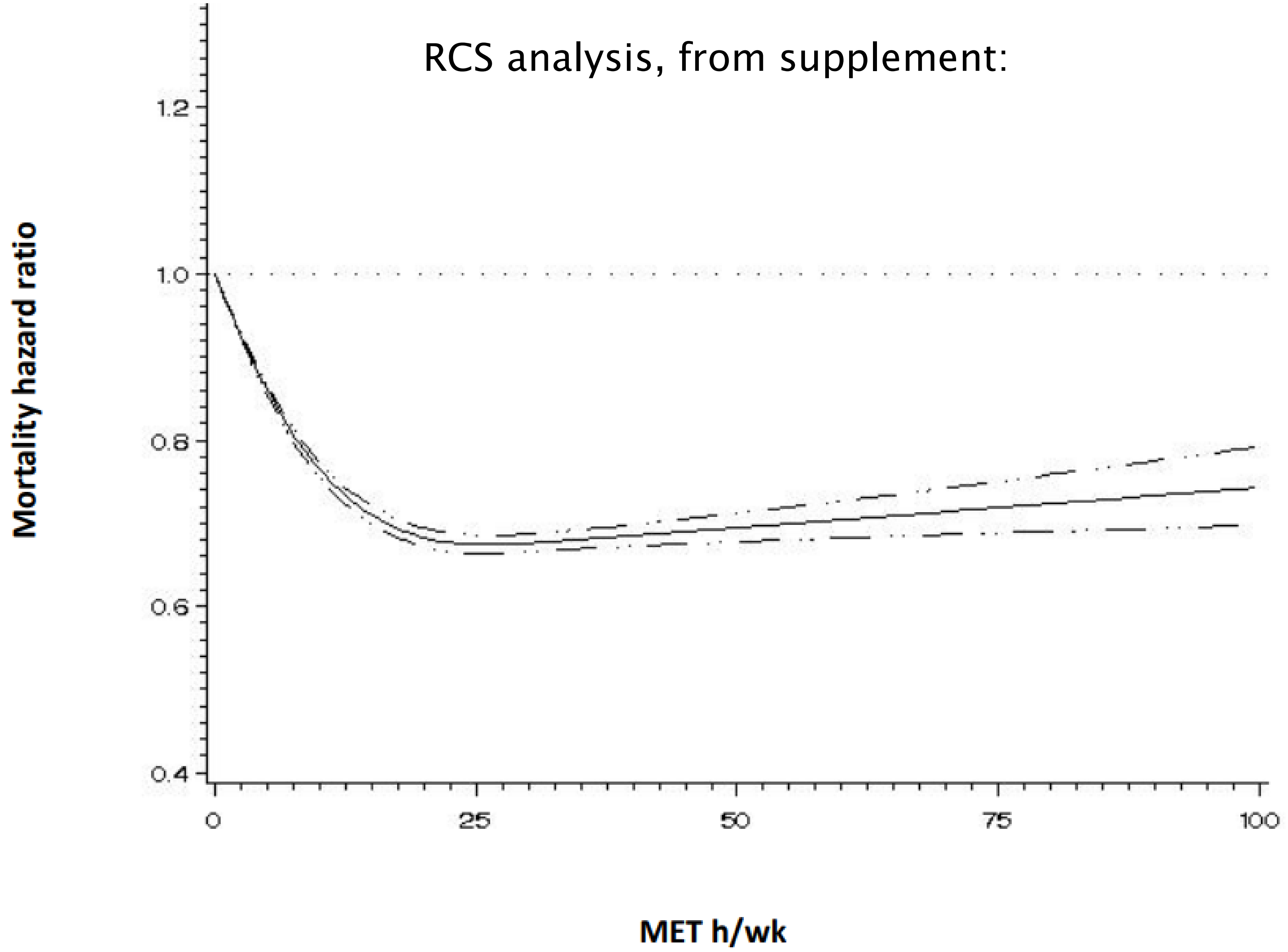
Figure. Hazard Ratios (HRs) and 95% CIs for Leisure Time Moderate- to Vigorous-Intensity Physical Activity and Mortality



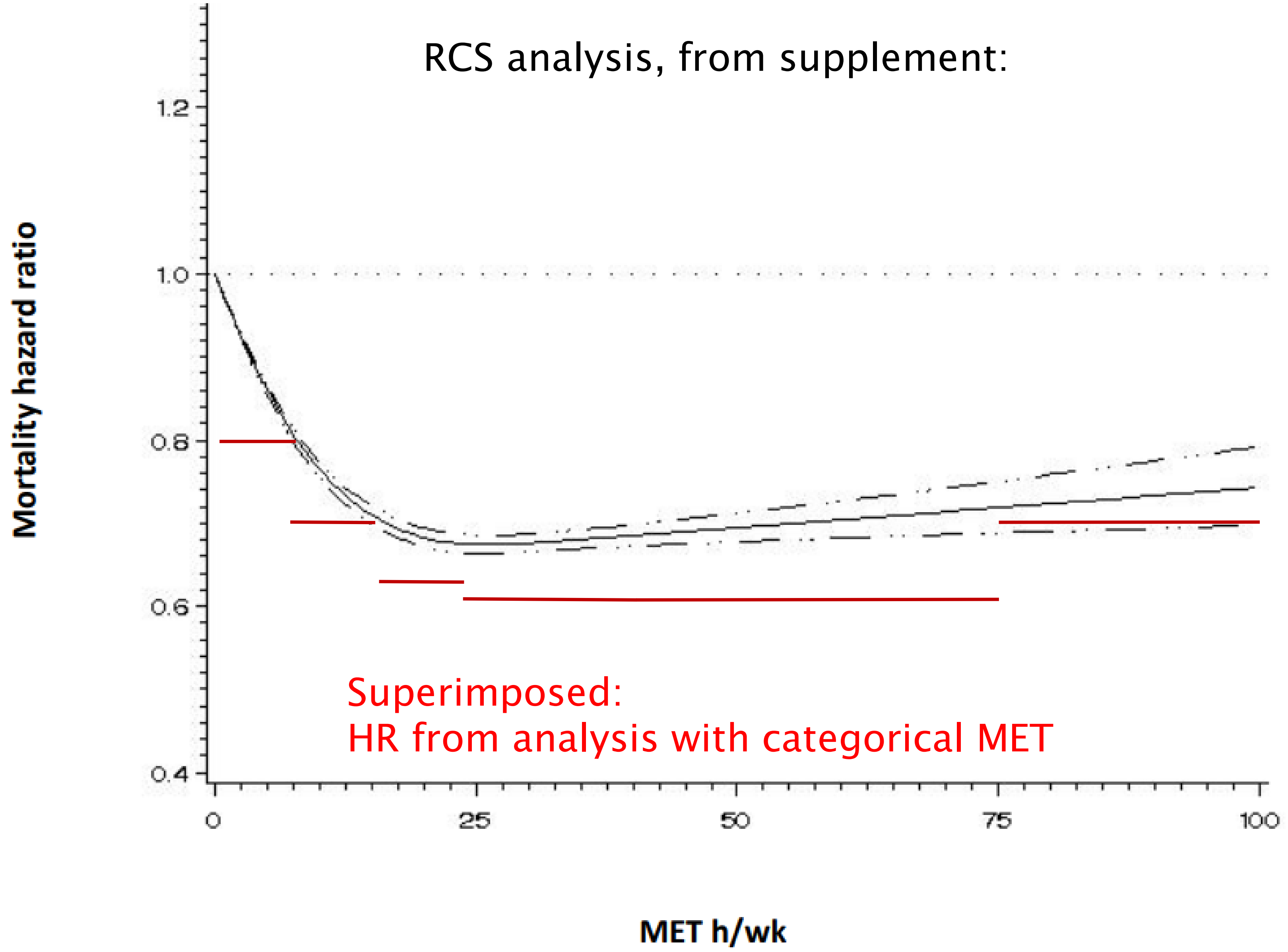
"These findings … provide important evidence to inactive individuals by showing that modest amounts of activity provide substantial benefit for postponing mortality"

Effect of walking 16 seconds to 20 minutes a day

RCS analysis, from supplement:



RCS analysis, from supplement:



RCS vs. categorized analysis

- Questions:
  - If both analyses were correct - > why do they give different results?
  - Back-confounding by the categorized analysis?
  - Or wrong treatment of ‚spike at zero' in RCS analysis?

Comment & Response

November 2015

Physical Activity and Successful Aging

Even a Little Is Good

David Hupin, MD, MSc[1]; Frédéric Roche, MD, PhD[1]; Pascal Edouard, MD, PhD[1]

» Author Affiliations | Article Information

JAMA Intern Med. 2015;175(11):1862-1863. doi:10.1001/jamainternmed.2015.4744

What is ‚a little'?
What is ‚good'?

Towards recommendations – research required!

1. Investigation and comparison of the properties of **variable selection strategies**
2. **Comparison of spline procedures** in univariable and multivariable contexts
3. How to model one or more variables with a ‚**spike-at-zero**'?
4. Comparison of **multivariable procedures for model and function selection**
5. **Role of shrinkage** to correct for bias introduced by data-dependent modelling
6. Evaluation of new approaches for **post-selection inference**
7. Adaptation of procedures for **very large sample sizes** needed?

1. Properties of variable selection strategies (pre-STRATOS)

Biometrical Journal

REVIEW ARTICLE | Open Access

Variable selection – A review and recommendations for the practicing statistician

Georg Heinze, Christine Wallisch, Daniela Dunkler

First published: 02 January 2018 | https://doi.org/10.1002/bimj.201700067 | Cited by: 29

Level-2

transplant international WILEY

Level-1

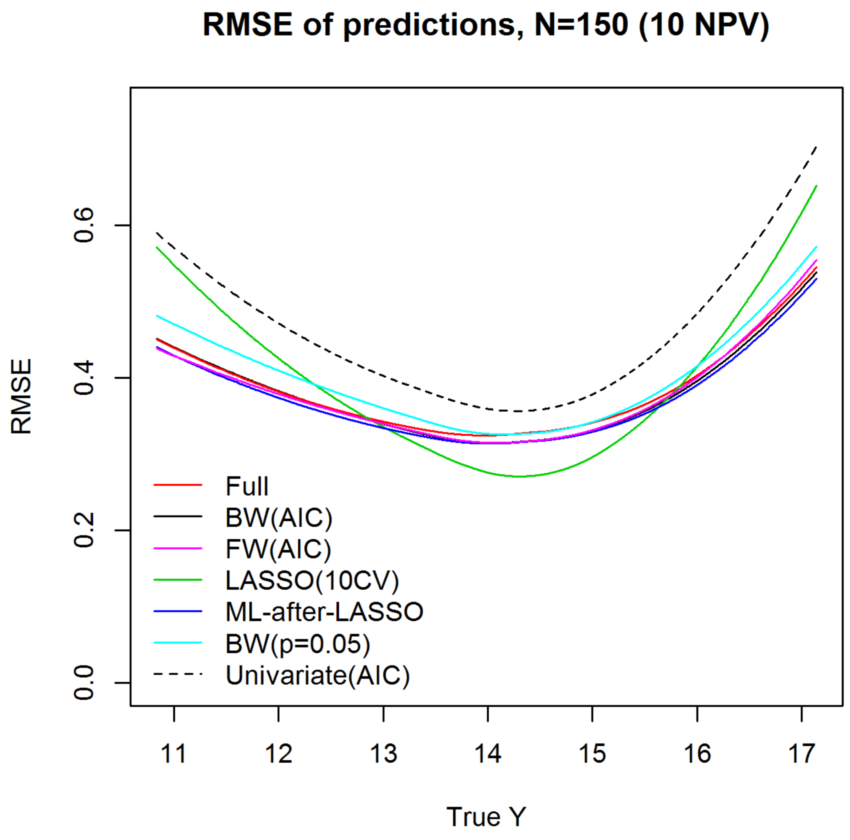
Review | Free Access

Five myths about variable selection

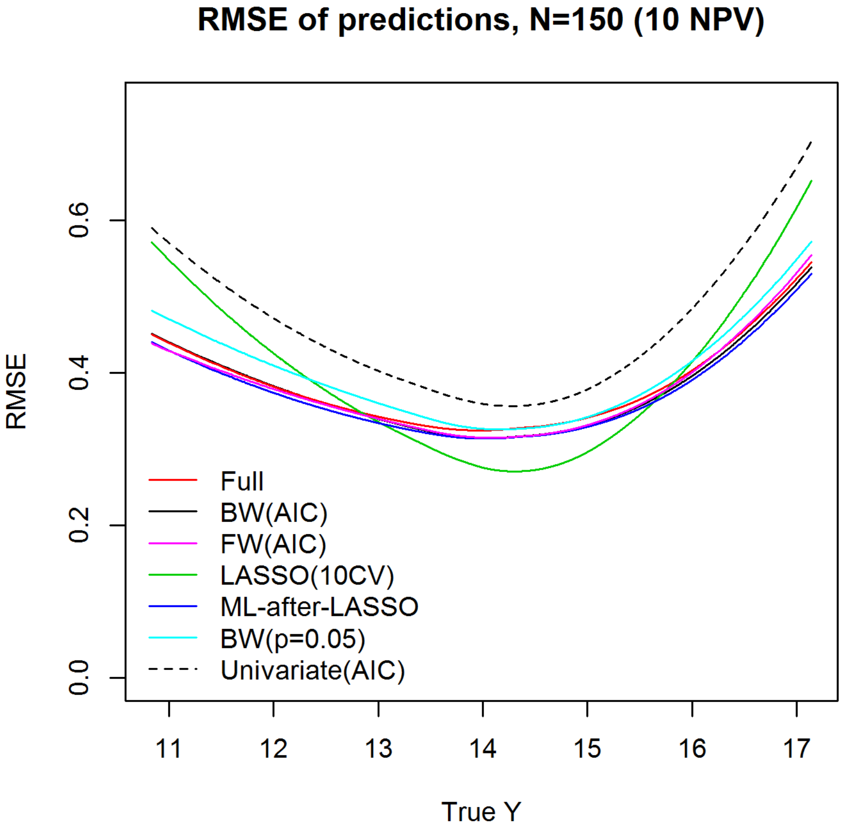
Georg Heinze, Daniela Dunkler

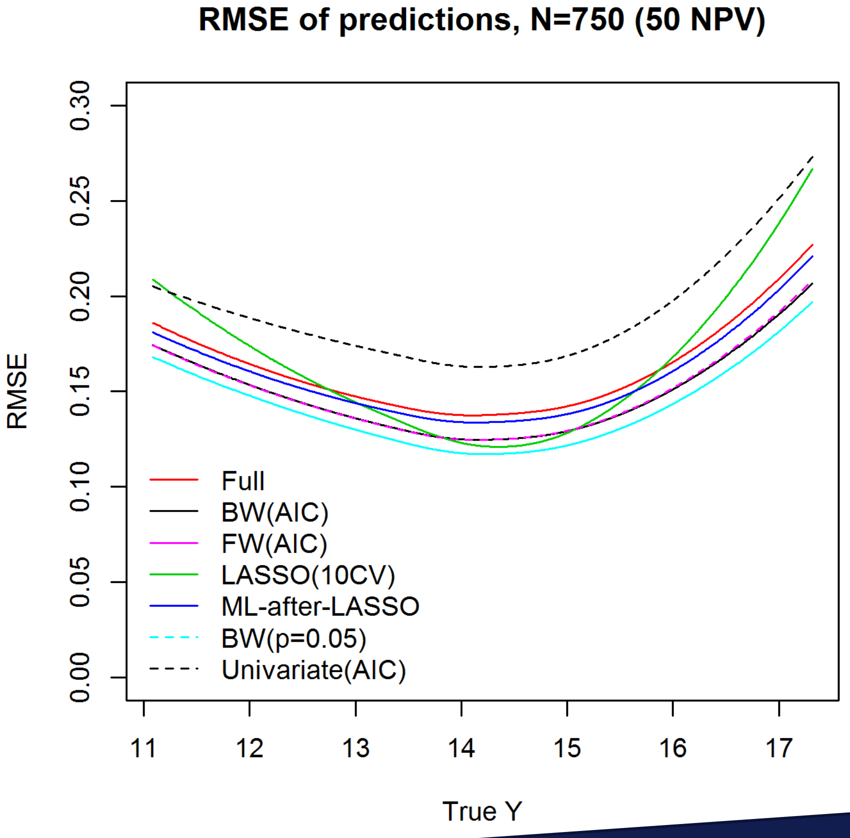
First published: 29 November 2016 | https://doi.org/10.1111/tri.12895 | Cited by: 23

1. Properties of variable selection strategies



1. Properties of variable selection strategies





One size does not fit all!

1. Properties of variable selection strategies

- What about modern techniques:
  - Adaptive Lasso, Garotte
  - Boosting
  - SCAD
  - Kullback-Leibler projection (`projpred`, Goutis&Robert 1998)

or Bayesian ones:

- Bayesian Lasso (Laplace prior) ,
- Spike-and-slab priors
- Horseshoe priors

- Are they useful for low-dimensional situations?
- In which situations do they improve over traditional approaches?
- Do they improve instability of selection methods?
- Do they improve the accuracy of estimates?
- Are there pitfalls in their application for non-expert users?

2. Comparison of spline procedures

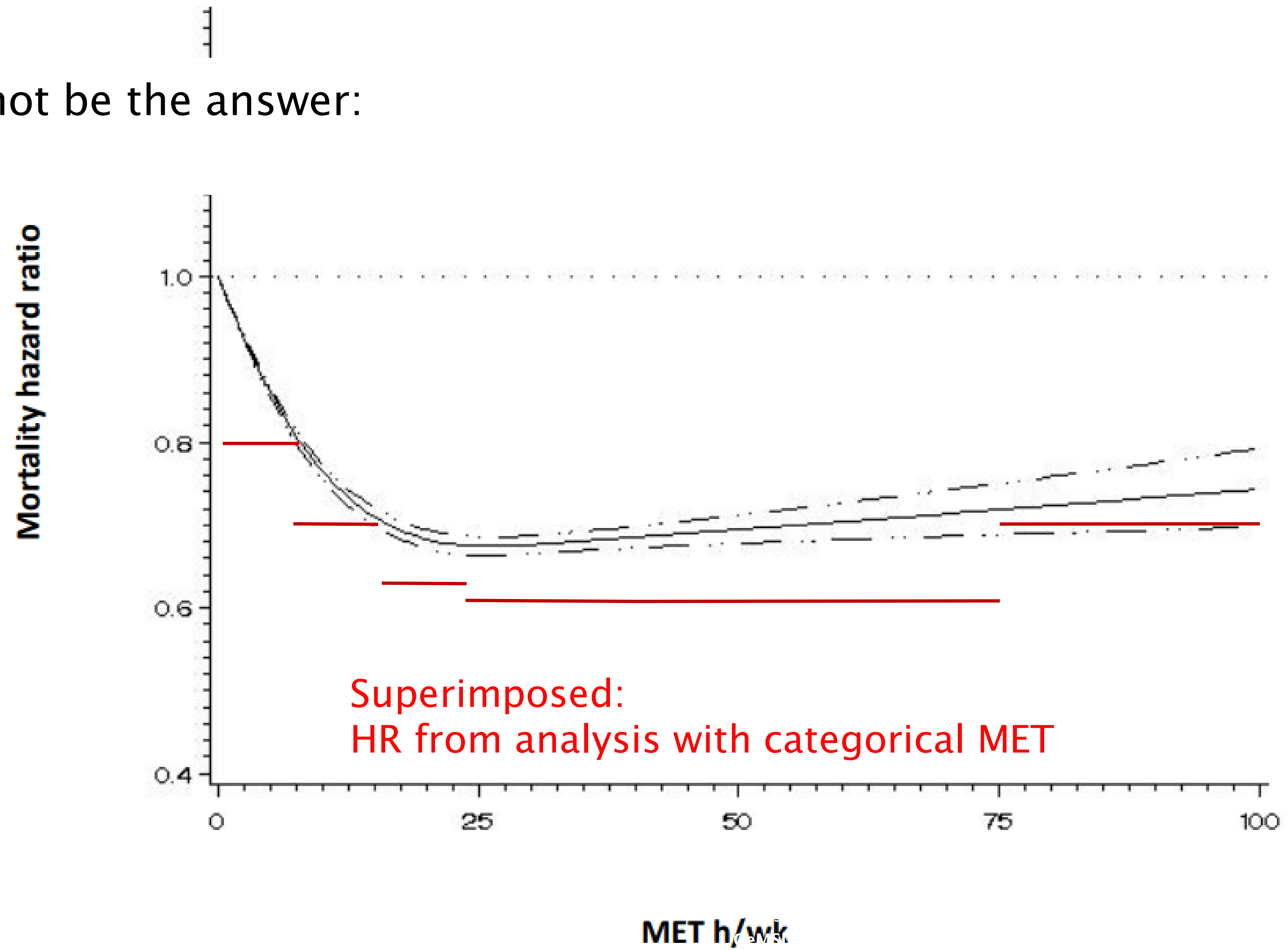
- Results from various spline procedures differ from true function?
- How does this depend on relevant parameters (number and location of knots)?
- Permitted complexity, usability for non-experts

- Multivariable context – multiple variables of mixed types

- For level-1: **How to report results** in a clinical paper?
  - Just a supplementary figure, or main result?
  - Recommendations for typical contrasts to report?
  - Dependency on knot locations, permitted complexity

3. How to model a variable with spike-at-zero?

- This cannot be the answer:



4. Comparison of multivariable procedures for model and function selection

- Several procedures possible:
  - Multivariable fractional polynomials: designed for that purpose
  - Splines-based:
    - mgcv and gamlss implement forward and backward selection
    - Penalty term on second order derivative -> linearity
    - Extra penalty for each smooth term -> null space penalization
    - LASSO-based approaches (COSSO, SpAM, …)
    - GAMSEL
    - Extensions of the nonnegative garotte
    - …

5. Role of shrinkage to correct for bias introduced by data-dependent modelling

- Post-selection shrinkage:
  - Global shrinkage factor
  - Parameterwise shrinkage factors (resampling-based; Sauerbrei 1999)
  - Joint shrinkage factors (Dunkler et al 2016)
  - Garotte – can be used for selection, but also for post-selection shrinkage

- The aim is not to correct unconditional bias, but to correct the overestimation bias with respect to the *selected model, had that model been prespecified*

6. Evaluation of new approaches for post-selection inference

- Selective inference (Taylor and Tibshirani, 2015)
- Model confidence bounds (Li et al, Biometrics 2019)

- Bayesian approaches delivering credible intervals from posteriors
  - Bayesian Lasso (Laplace prior)
  - Spike and slab priors
  - Horseshoe priors

Also for not-selected variables!

  - Solve the problem of selection uncertainty in an elegant way
  - But still are cumbersome to implement and conduct

7. Adaptation of procedures for very large sample sizes

- N and P growing
  - Pharmacoepidemiology, Electronic Health Records, Registries, Big IPDMA, …
- In large samples, ‚everything becomes significant'
- Procedures based on cross-validation or AIC may not be so well suited
- Usual tuning criteria valid?
- *Does* BIC *do the trick?*
- Alternative approaches
  combining background knowledge with statistical learning?

Regarding these issues...

- Mathematical theory is unlikely to help
- Simulation studies are key (see e.g., Binder et al, StatMed 2013)
- However, simulation studies are biased towards the proposed method (Boulesteix et al, BiomJ 2018)
- Simulation studies are often poorly designed, conducted and reported (Morris et al, StatMed 2019)
- Simulation panel of STRATOS may provide guidance (B. & M. are members)
- Experience from comparative analyses with real data sets
- Translation to level-1 is needed!

TG2 projects

Completed:

- Perperoglou et al. A Review of spline procedures in R. *BMC Med Res Meth*, 2019.
- Sauerbrei et al. State-of-the-art in selection of variables and functional forms in multivariable analysis: outstanding issues. **arXiv:1907.00786**, 2019.

Ongoing:

- Literature review of the practice of variable and functional form selection (VFFS)
- Literature review of VFFS in statistical series in medical journals
- Initial data analysis for regression analyses
  (‚regression without regrets' TG2&TG3-collaboration) →next talk!!!

REVIEW Open Access

A review of spline function procedures in R

Aris Perperoglou[1*], Willi Sauerbrei[2], Michal Abrahamowicz[3], Matthias Schmid[4] on behalf of TG2 of the STRATOS initiative

Perperoglou et al. BMC Medical Research Methodology (2019) 19:46
https://doi.org/10.1186/s12874-019-0666-3

BMC Medical Research Methodology

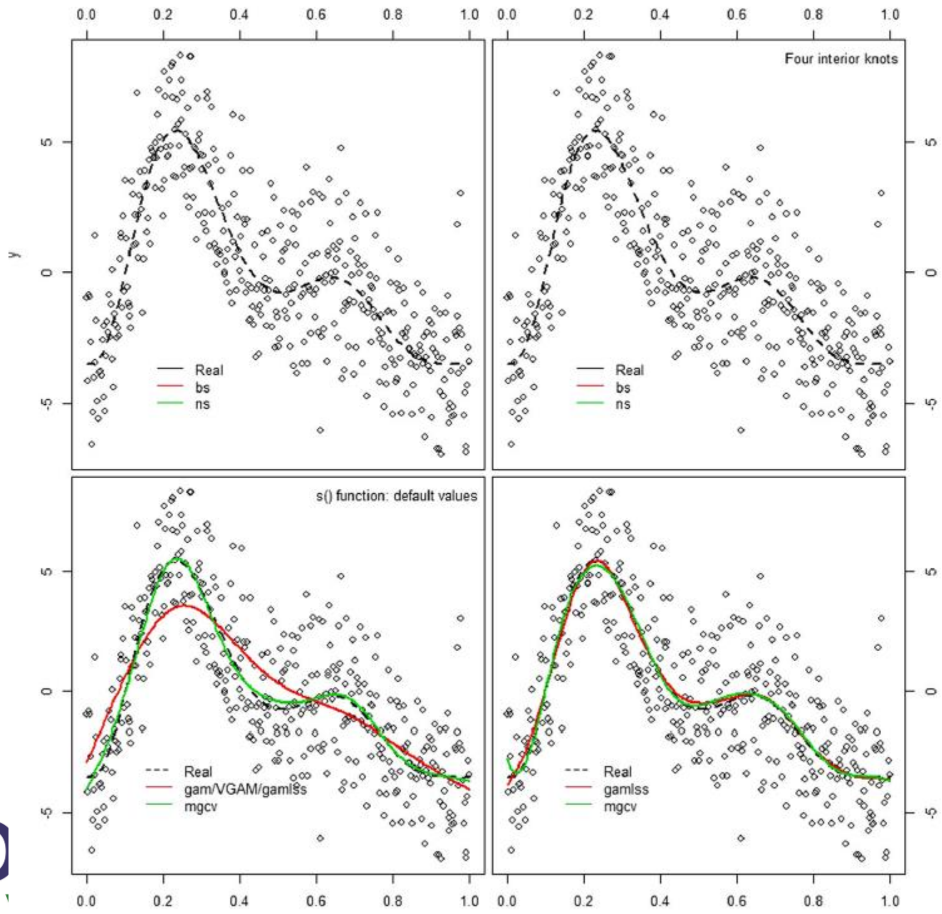


Table 1 R packages used for the creation of splines

Package	Downloaded	RD	Description	Authors
gss	632212	9	General smoothing splines	Chong Gu
rms	598185	63	Regression modeling strategies	Frank Harrell Jr
polspline	406661	11	Polynomial spline routines	Charles Kooperberg
pspline	146939	11	Penalized smoothing splines	Brian Ripley
logspline	130048	10	Logspline density estimation routines	Charles Kooperberg
cobs	58533	6	Constrained B-splines	PT Ng and M Maechler
crs	58347	2	Categorical regression splines	JS Racine, Z Nie, BD Ripley
splines2	31031	4	Regression spline functions and classes	Wenjie Wang and Jun Yan
bigsplines	25940	1	Smoothing splines for large samples	Nathaniel E. Helwig
bezier	18483	1	Bezier curve and spline toolkit	Aaron Olsen
pbs	17794	1	Periodic B splines	Shuangcai Wang
freeknotsplines	13761	0	Free-knot splines	S Spiriti, P Smith, P Lecuyer
orthogonalsplinebasis	13436	1	Orthogonal B-spline functions	Andrew Redd
ConSpline	10565	0	Partial linear least-squares regression using constrained splines	Mary Meyer
episplineDensity	9375	0	Density estimation exponential	S Buttrey, J Royset, R Wets

The number of times of time each package was downloaded is measured from 01/10/2012 to 15/11/2018. Number of downloads does not correspond to unique users. Reverse dependencies (RD) stands for the number of other packages that call each one