International initiative

Guidance for key issues of design and analysis of observational studies

TG 2: Selection of Variables and Functional Forms: flexible approaches improve estimation and inference

Michal Abrahamowicz

(McGill University, Montreal, Canada)

for the members of TG 2

Members of TG2

- **Chairpersons:**
  - Michal Abrahamowicz (McGill, Montreal, Canada)
  - Göran Kauermann (Munich, Germany)
  - Willi Sauerbrei (Freiburg, Germany)

- **Additional members so far:**
  - Harald Binder (Mainz, Germany)
  - Frank Harrell (Vanderbilt, Franklin, USA)

Main issues for the start

We focus on building

multivariable 'explanatory models' (*)

whose main goal is to identify influential predictors and gain insights into their relationships with the outcome, through the estimated model structure.

[Harrell 2001; Sauerbrei et al 2007].

(* for Prediction models see **TG 6**)

We address **2 inter-related questions**, common to all multivariable explanatory models :

1. Selection of 'relevant' Variables
2. Choice of the Functional Form for the effect of each Continuous variable

Important Restrictive Assumptions (for the 1st Phase of Guidance development)

- **Low dimensional data (number of potential predictors $5<p<30$) with 'sufficient' sample size ($n > 10\,p$)** *
  * Avoids problems specific for high-dimensional data (p>>n)
  - Ensures (i) adequate Stability of the estimated explanatory model, and (ii) accurate Inference

- **No interactions are assessed** (interactions are *priori* ignored, except for potential well established interactions, *a priori* identified and forced into all models considered)

- **No missing data** (analysis restricted to subjects with complete data on all relevant variables) (-> link to **TG1**)

- **Measurement errors are ignored** (-> link to **TG4**)

How to Start:
1st Issue: Selection of Predictors

- to Select Variables into the Final, Parsimonious Multivariable Model, from a larger set of available, "candidate predictors", most studies use then a Combination of the 2 complementary approaches:

(i) *A priori* inclusion of some, well established (in substantive literature) 'predictors' of the outcome of interest (***)

(ii) *A posteriori* use of Data-dependent procedures and criteria to select the 'useful' predictors among the remaining 'candidate variables'

(***) Some clinical/epidemiological studies prefer to select the predictors Exclusively on *A Priori* basis. This is justifiable when assessing the effect of a specific exposure/treatment (to ensure all 'confounders' are adjusted for), but NOT in Explanatory models.

1st Issue: Data-Dependent Strategies for *A Posteriori* selection of Predictors

- Several alternative strategies proposed and discussed in literature

  [Harrell 2001; Royston & Sauerbrei 2008; Steyerberg 2009]
- Strategies of Practical Interest involve mostly Iterative Stepwise (Sequential) Inclusion or Elimination (***)
- No theoretical reasons to expect some strategies to perform systematically better than others [Miller 2002]
- Yet, **Backward Elimination:** (a) reduces number of estimated models (important for flexible modeling and selection of functional forms); and may often (b) approximate the results of all-subsets regression; & (c) yield near-optimal AIC/BIC values [Sauerbrei et al 2007]

- (***) All-Subsets Regresssion Computationally Too Intensive (in the context of mutivariable flexible modeling)
- (***) More specialized techniques e.g. Lasso left for Later

How to start: 2nd Issue: Functional Forms for Continuous Predictors

- **CATEGORIZATION** of continuous predictors is still quite common in clinical/epi research [e.g. 1].
- **Several Drawbacks of Categorization [2]**:

(i) Implausibility of the Step-Function effect & 'Local Bias'

(ii) Arbitrary cut-offs for categories often vary wildly across studies of the same predictor-outcome association [3], inducing spurious differences

(iii) 'Bad' *a priori* selection of cut-offs results in worse fit to data and increased Type II error

(iv) If cut-offs selected *A Posteriori:* standard Inference is Not valid, and increased risk of Type I error and overfit bias [4]

Thus, we Focus on Modeling of Continuous Functions

[1] Riley RD, Abrams KR, Sutton AJ et al. Br J Cancer 2003, 88:1191-1198.

[2] Royston P, Altman DG, Sauerbrei W, Stat Med 2006, 25: 127-141.

[3] Malats N,Bustos A,Nascimento C et al. Lancet Onc 2005, 6:678-686.

[4] Schulgen G, Lausen B, Olsen JH, Schumacher M. AJE 1994, 140(2): 172-184 .

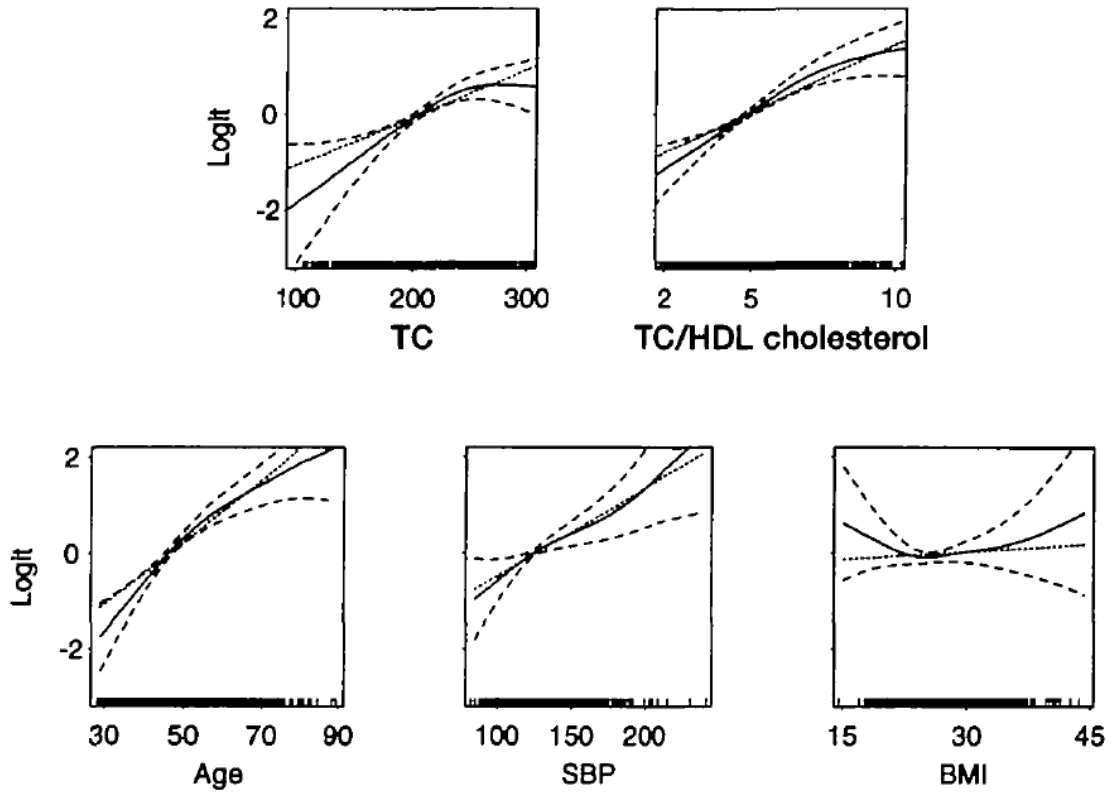
How to start: 2nd Issue: Functional Forms for Continuous Predictors

- 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 g(x) is Linear** & include Un-transformed X: **g(x) = βx**
- Linearity assumption is convenient (effect of X summarized by a single β, parsimony = improved power), and often adequate
- **Yet, Linearity should not be imposed *a priori*: numerous examples of Non-Linear or Non-Monotone effects,** e.g.:

(i) BMI -> all-causes mortality (both Obese and Too Thin subjects have Increased Risks),

(ii) Age at diagnosis -> mortality in different cancers (Youngest subjects have more aggressive disease, Oldest have increased risk of all-cause mortality)

GAM-estimated Non-linear effects of Risk Factors on logit of Coronary Heart Mortality [Abrahamowicz et al, AJE 1997]



How to start: FLEXIBLE MODELING of the Functional Forms for Continuous Predictors

- **Flexible Modeling techniques, proposed to estimate Non-linear (NL) effects of Continuous X's, with different Smoothers, include e.g.:**
- Fractional Polynomials (FP) [Royston&Sauerbrei2008;Royston&Altman 1994]
- Regression Splines

  [Ramsay 1988; Abrahamowicz & MacKenzie 2007]
- Restricted Cubic Splines

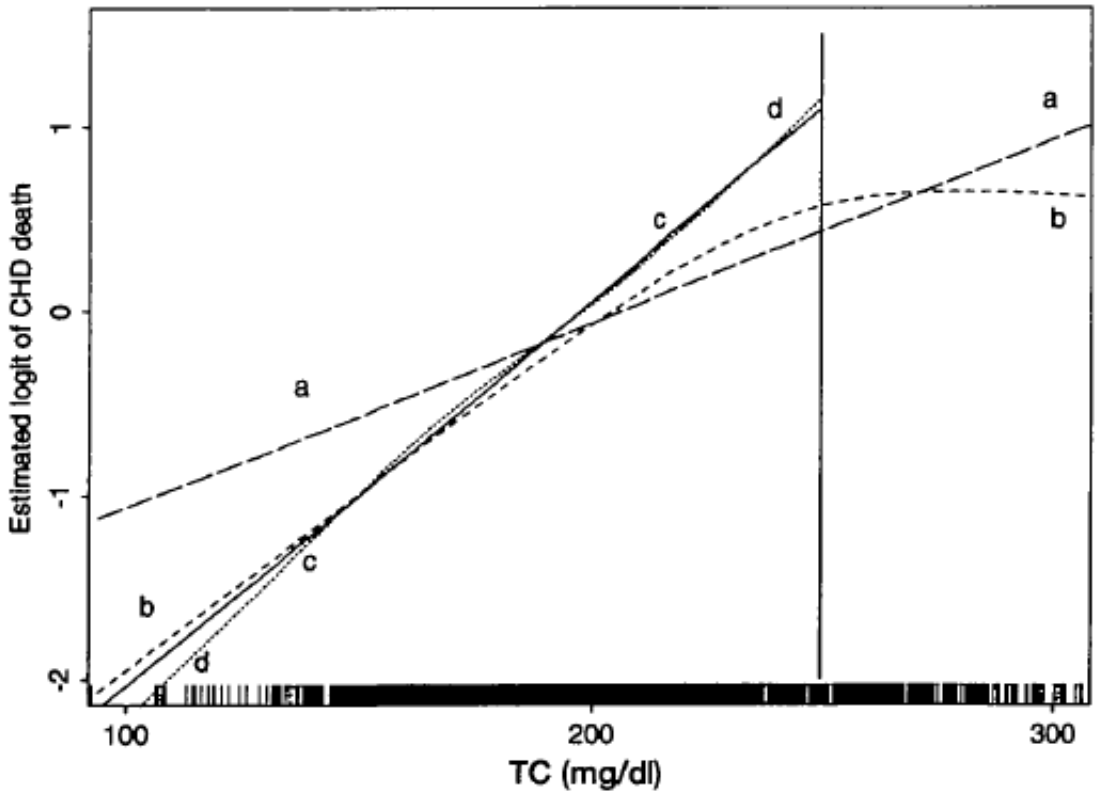
  [Harrell (2001)]
- Penalized Smoothing Splines

  [Gray JASA 1992, 87: 942-951]
- Generalized Additive Models (GAM)

  [Hastie & Tibshirani , 1990]
- ……+ several other types of (I- , P- …etc) -Splines

Flexible Modeling of g(x) avoids 'local biases' of a Linear Function: Cholesterol (X) vs logit of CVD death

- **(a) & (b): full range of X; (c) & (d) X<250; (a) & (c) linear (βx);**
- **(b) & (d) Smoothing Spline (GAM)**; [Abrahamowicz et al, AJE 1997]



How to start: FLEXIBLE MODELING of the Functional Forms: Which SMOOTHER ?

- In (limited) comparisons using Simulated & Real data:

> Different Smoothers yielded generally Similar NL (point) Estimates

[Binder et al 2013; Hastie & Tibshirani 1990]

Yet:

- FP's are more parsimonious than splines and, thus, reduce over-fit bias & improve stability of the estimates IF the True Dose-Response Function is relatively Simple

[Binder et al 2013]

Inter-Dependence of the Selections of (1) Variables vs (2) Functional Forms

- **The CHALLENGE is that the results of Data-dependent selections of (1) 'significant'/relevant Predictors may depend on (2) choices regarding Functional Forms of both, (2a) the Predictor of Interest (X) & (2b) Other Variables, correlated with X, and *vice versa***

[Rosenberg PS, Katki H, Swanson CA, Stat Med 2003, 22: 3369-3381]

Examples of Inter-dependence:

(1) Impact of Inaccurate Modeling on Variable Selection: Incorrect Linearity Assumption increases Type II error for testing the (truly NL) effect of X, resulting in its un-warranted exclusion

[e.g. Abrahamowicz et al 1997; Gagnon et al Br J Cancer 2010]

Impact of **Residual Confounding** (due to Incorrect Modeling of Confounders):

- **Further Examples of Inter-dependence:**

> (2) Failure to adjust for Important Confounders and their NL effects, increases either Type I or Type II error for testing:

- (2a) Linearity of the effect of a continuous X [Binder et al 2013];
- (2b) Association between a binary Z and the outcome [Benedetti & Abrahamowicz 2004] ;

> (3) in Survival analyses, a failure to account for NL effect of X increases type I error for a Time-dependent effect of X [Abrahamowicz & MacKenzie 2007]

How to Start: Towards recommendations

- Recommendations for building multivariable explanatory models must address Both inter-dependent issues.

- Recommendations should also consider: 'Transportability', as well as ease of both methods Implementation & Interpretation of results

- **Sauerbrei et al (2007) tentatively recommend** (under the restrictive assumptions of Slide 4) using **Multivariable Fractionals Polynomials (MFP) algorithm**

  that combines Backward Elimination (for issue 1) **with FP modeling of the effects of continuous predictors** (for issue 2). [Royston & Sauerbrei 2008]

How to Start: Towards recommendations

- **NEXT STEPS for TG2:**

(1) Comprehensive LITERATURE REVIEW to Identify (potential) other tentative Recommendations

(2) Developing Recommendations for Systematic, User-friendly SPLINE-based approaches that Integrate Flexible Modelling and Selection of Predictors & their Effects

(3) Designing further SIMULATION studies to COMPARE Alternative Approaches

How to Start:
ISSUES that require Further Attention

- **BOOTSTRAP** should be used to:
- (1) Correct Inference for Data-Dependent Model Selection [5, 6]
- (2) Investigate Model Stability [7,8] (***)

(***) 2 levels of Bootstrap analyses:

(a) (Simpler) re-estimate only the "final model" (selected based on the original data), to assess stability of the estimates of (i) regression coefficients and(ii) shapes of the non-linear functions

(b) (More complex and computer-intensive): re-run the entire model selection process, to assess the stability of the selection of variables and `non-linear effects of continuous variables

[5] Hurvich C.M and Tsai C.L. Am Statistician 1990; **44:** 214-217.

[6] Mahmud M, Abrahamowicz M, Leffondré K et al. Comm Stat 2006;35:27-45.

[7] Altman DG, Andersen PK. *Stat Med* 1989; **8**: 771-783.

[8] Sauerbrei W. *JRRS Series C,* 1999; **48**: 313-329.

CONCLUSION

- Paraphrasing Albert Einstein's credo about Scientific Theory :

  "**Statistical Models should be as Simple as Possible but NOT Simpler ....**"

(Selected) Relevant literature

- Abrahamowicz M, Berger Rd, Grover SA. Flexible Modeling of the Effects of Cholesterol on Coronary Heart Mortality. *Am J Epi (AJE)* 1997; 145: 714-729.
- Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Stat Med* 2007; 26: 392-408.
- Binder H, Sauerbrei W, Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates. *Stat Med* 2013; 32: 2262-2277.
- Benedetti A, Abrahamowicz M. Using generalized additive models to reduce residual confounding. *Stat Med* 2004; 23: 3781-3801.
- Gagnon B, Abrahamowicz M, Xiao Y, et al. Flexible modeling improves prognostic value of C-reactive protein. *Br J Cancer* 2010; 102: 1113-1122.
- Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, 2001.
- Hastie T., Tibshirani R. *Generalized Additive Models.* Chapman & Hall: NY, 1990.
- Kauermann G. *Nonparametric models and their estimation.* Allgem Stat Archiv 90, 135-150.
- Miller A. *Subset Selection in Regression.* Taylor & Francis, 2002.
- Ramsay JO. Monotone Regression Splines. *Stat Sciences* 1988; 3: 425-441.
- Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epi (IJE)* 1999; 28: 964-974.
- Royston P, Sauerbrei W. *Multivariable Model Building: A Pragmatic Approach to Regression Anaylsis Based on Fractional Polynomials for Modelling Continuous Variables.* Wiley, 2008.
- Sauerbrei W, Royston P, Binder H. Selection of important variables and functional form for continuous predictors in multivariable model building. *Stat Med* 2007; 26: 5512-5528.
- Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating.* Springer, 2009.