‘Give me six hours to chop down a tree ...’

Statistical tools to meet the challenges of modern medical data

ACC Coolen

King’s College London

April 20th 2016

1982:

Commodore 64

1.02 MHz

64 Kb RAM

1982:

Commodore 64

1.02 MHz
64 Kb RAM

Regression Models and Life-Tables

D. R. Cox

Journal of the Royal Statistical Society. Series B (Methodological), Volume 34, Issue
(1972), 187-220.

Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281972%2934%3A2%3C187%3ARMAL%3E2.0.CO%3B2-6

Cox's proportional hazards regression

brilliant compromise between statistical demands in medicine and computing power limitations

of the 1970s

Using 1970s techniques in 2016 is perfectly acceptable if

- they can handle statistical tasks of modern medicine
- or it is not possible to develop more powerful ones

otherwise we are not serving our patients as well as we could and should

Overview

The bigger picture ...

- Maths meets cancer medicine
- General pitfalls in statistics
- Regression for survival data, why and how
- What has changed since the 1970s

Cohort heterogeneity and competing risks

- Consequences and fingerprints of latent heterogeneity
- Bayesian latent class models
- Applications to epidemiological cancer data
- Applications to data from failed cancer trials

Overfitting in multivariate survival analysis

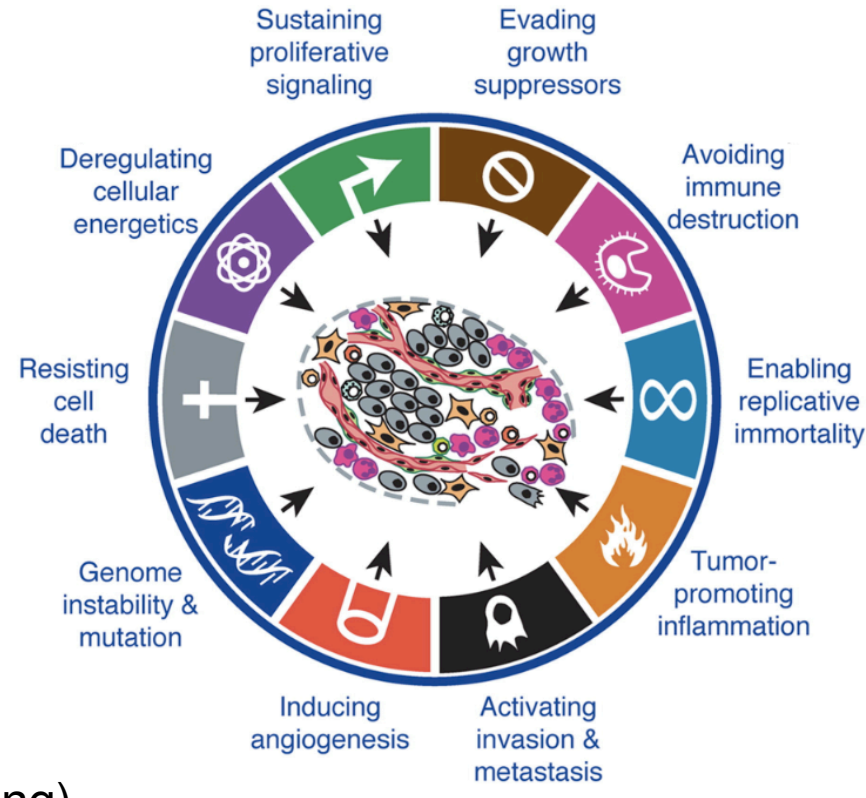
- Sample size and covariate selection
- Eliminate redundant information
- Overfitting correction protocols

The bigger picture

Maths meets cancer medicine

potential of quantitative innovation
in 21st century cancer medicine ...

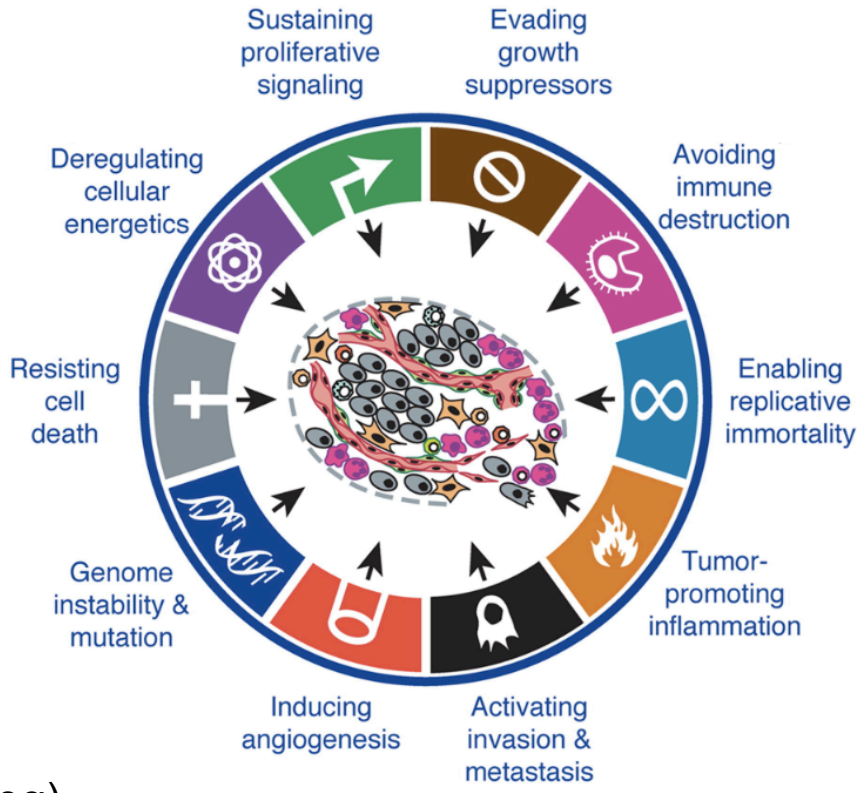
- *Reverse engineering*
  - reprogram circuits in cells
  - reprogram the immune system
- *Adaptive medical trials*
  - faster, fewer patients
  - response-triggered intervention
- *Predict clinical outcome and treatment response*
  - high-dimensional data (genetic, imaging)
  - heterogeneity of cancers and patients
  - risk correlations, comorbidities
  - confounding factors, batch effects



Maths meets cancer medicine

potential of quantitative innovation
in 21st century cancer medicine ...

- *Reverse engineering*
  - reprogram circuits in cells
  - reprogram the immune system
- *Adaptive medical trials*
  - faster, fewer patients
  - response-triggered intervention
- *Predict clinical outcome and treatment response*
  - high-dimensional data (genetic, imaging)
  - heterogeneity of cancers and patients
  - risk correlations, comorbidities
  - confounding factors, batch effects



General pitfalls in statistics

- *Often counterintuitive*

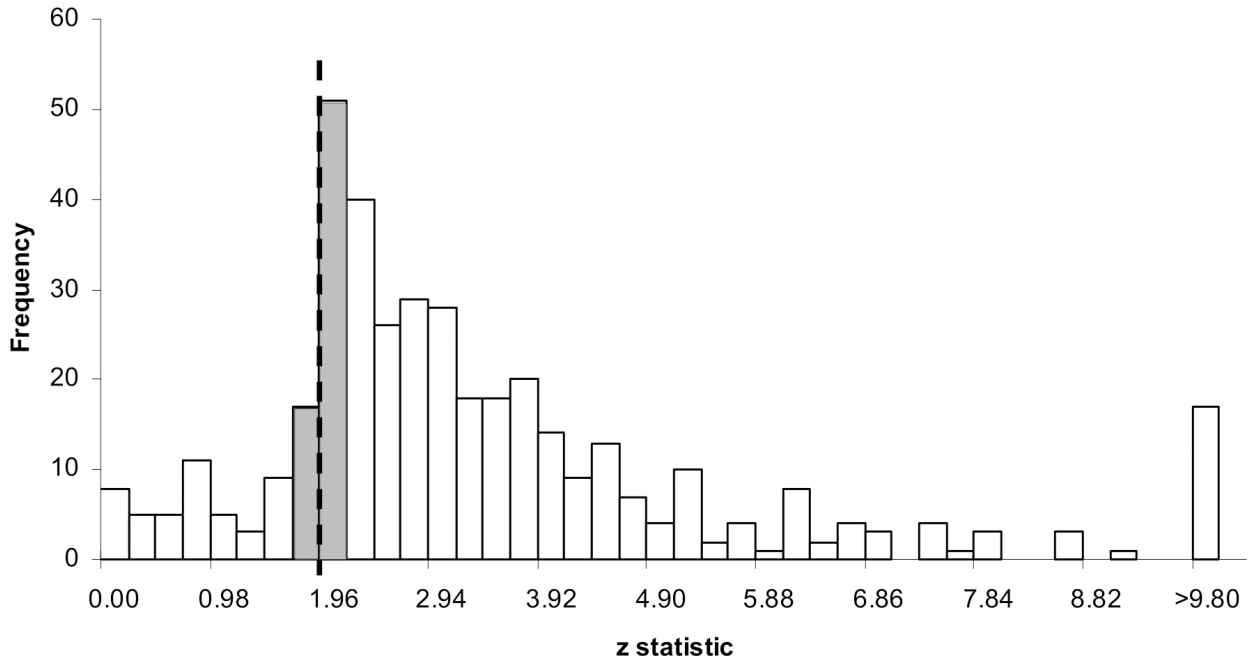
  Monty Hall problem, gambling, ...

  I have just thrown 10 successive sixes!
  Prob $\approx 16.5.10^{-8}$

  how likely am I to throw yet another six?

- *Selective reporting*
  (aka cheating)

  z-scores reported in PLoS



+AllTrials

All Trials Registered | All Results Reported

Home | Find out more | Get involved | Supporters | News | **Sign the petition** | **Donate**

Donate ›

Get involved ›

Latest news ›

- *'Probability' can mean different things ...*

our ignorance of

(a) something *that cannot be known*
(Russian roulette, we will spin the cylinder)

(b) something *that is known, but not by us*
(Russian roulette, cylinder has already been spun)

relevant in medicine?

Suppose we find survival function $S(t) = \mathrm{e}^{-t/\tau}$

explanation I: homogeneous cohort, *random* death times,
each individual i has hazard rate $1/\tau$

explanation II: heterogeneous cohort, *deterministic* death times t_i,
distributed according to $p(t) = \tau^{-1}\mathrm{e}^{-t/\tau}$
(potential for stratification!)

Regression for survival data, why and how

- *Objective*

  find and quantify patterns (if any) that relate covariates to event times, in order to:

  1. *predict clinical outcome for individuals*
  2. *discover disease mechanisms*
  3. *design interventions (modifiable covariates)*

- *How can we know that what we find is real?*

  only one way: predict outcome for *unseen* cases

- *When do we need*

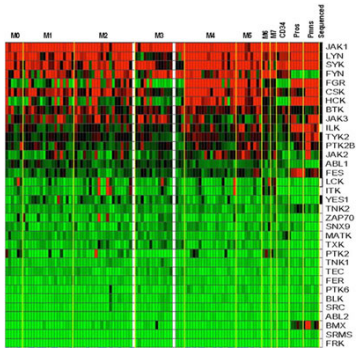
  parameter interpretation: (2,3)
  multivariate regression: (1,2,3)

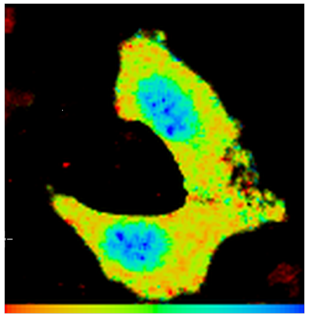
- *Choice of regression models*

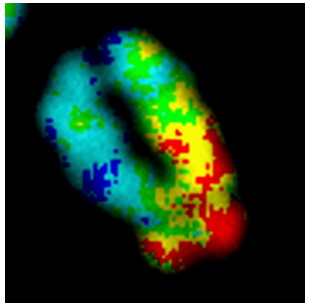
  primitive models can only reveal primitive patterns ...
  primitive models often make brutal assumptions ...
  complex models require many more data ...

What has changed since the 1970s?

- *Medical data have evolved*







  - sheer *volume* ...
  - *diversity* of data sources
    (clinical, genomic, biomarkers, health records, imaging, ...)
  - *complexity* of experimental pipelines
    (confounders, batch effects, variability between centres, ...)
  - *dimension* mismatch

    then: 500 samples, 10 covariates
    now: 500 samples, 10^6 covariates

- *Statistical thinking has evolved*

  away from maximum likelihood estimators
  towards Bayesian methods:

  quantify uncertainty in *parameters and models*

example:

Tsiatis' identifiability problem (1975)
(how to disentangle competing risks)

- if hazard rate for risk 1 is low:

  (i) event 1 is intrinsically unlikely, or
  (ii) it is often preceded by event 2

- eliminating one risk can *change* hazard rate of the other ...
  to disentangle: need joint event time stats $p(t_1, t_2)$ ...

  $p(t_1, t_2)$ cannot be inferred from survival data ...

- simplest way out:

  assume *all risks statistically independent*
  (required to interpret KM curves, Cox regression ...)

The Bayesian view:

– multiple hypotheses H may explain our data
– but not all are equally probable ...
– calculate each $Prob(H|data)$ from Bayes' formula

illustration:
identifiability problem

- true data:

$$p(t_2) = a\mathrm{e}^{-at_2}, \quad \begin{cases} \text{with prob } \epsilon: & t_1 = t_2 + \tau \\ \text{with prob } 1-\epsilon: & \text{draw } t_1 \text{ from } p(t_1) = b\mathrm{e}^{-bt_1} \end{cases}$$

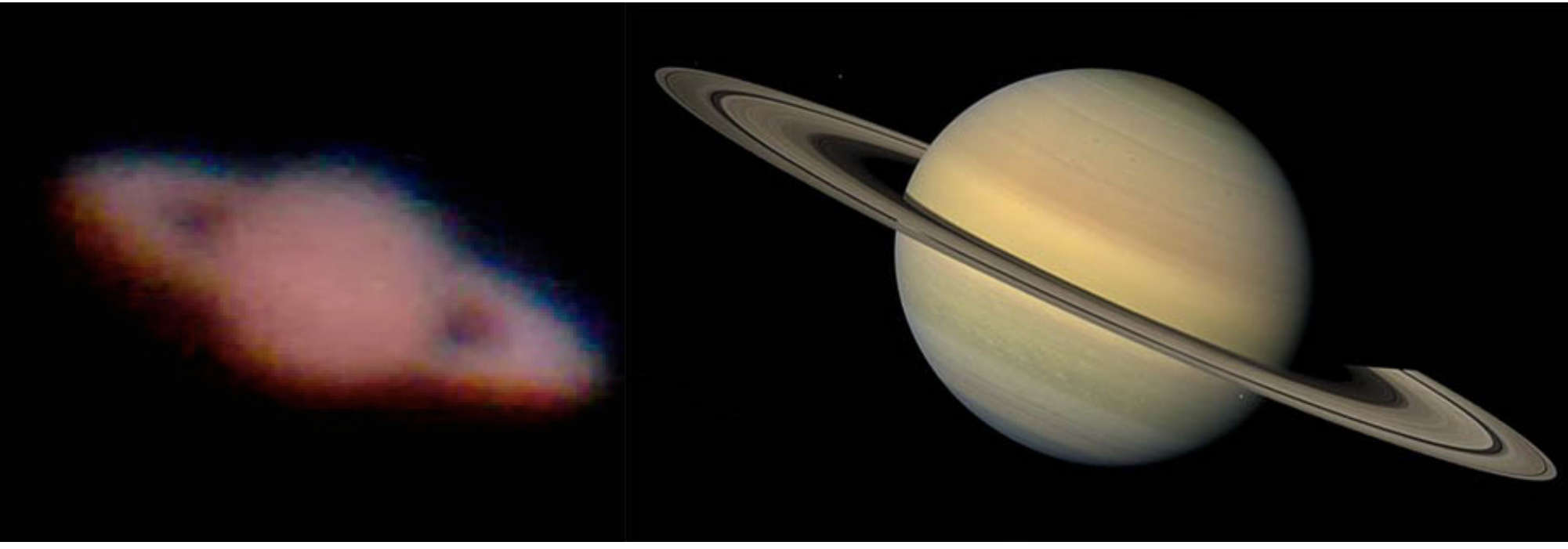
- explanation assuming risk independence:

$$p(t_2) = a\mathrm{e}^{-at_2}, \qquad p(t_1) = \underbrace{-\left(\epsilon + (1-\epsilon)\mathrm{e}^{-bt_1}\right) \log\left(\epsilon + (1-\epsilon)\mathrm{e}^{-bt_1}\right)}_{\text{with prob } \epsilon\text{: event 1 never happens}}$$

implausible if e.g. risk 2 is cancer, risk 1 is death ...

Cohort heterogeneity and competing risks

primitive tools
can only reveal
primitive patterns



David Cox:

'The proportion of my life that I spent working on the proportional hazards model is, in fact, very small. I had an idea of how to solve it but I could not complete the argument and so it took me about four years on and off...'

conventional methods

for time-to-event data

Kaplan-Meier estimators
Cox regression
.....

- not designed to handle disease/host heterogeneity, beyond variability in covariates
- to allow interpretation:

  have to assume different risks are uncorrelated, dangerous when many censoring events ...

random effects and latent class models

- usually constructed for primary risk only, so still cannot handle correlated risks

Consequences and fingerprints of latent heterogeneity

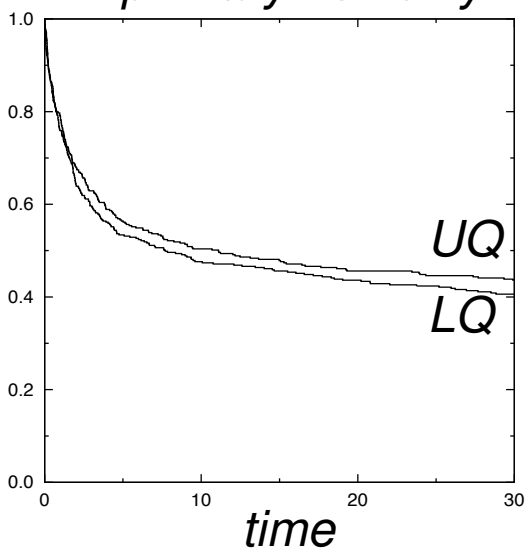
- *Violation of proportional hazards assumption*
- *Interpretation of time dependencies tricky*

  even if all *individual* hazard rates h_i are time-independent, cohort hazard rate will be time-dependent:

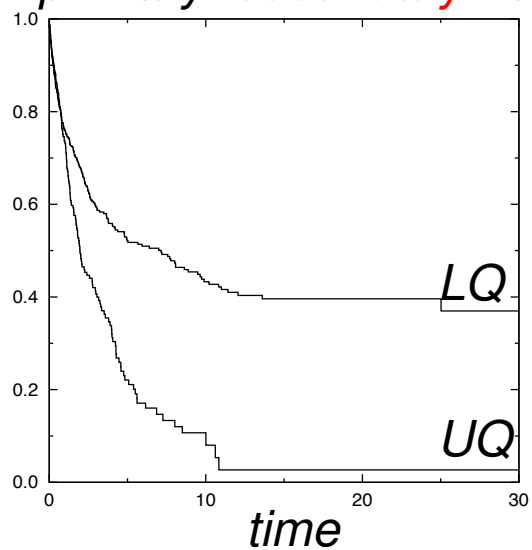
$$h(t) = \frac{\sum_{i=1}^{n} h_i e^{-h_i t}}{\sum_{i=1}^{n} e^{-h_i t}}$$

- *Interpreting cause-specific survival curves (KM, Cox) no longer possible ...*

primary risk only



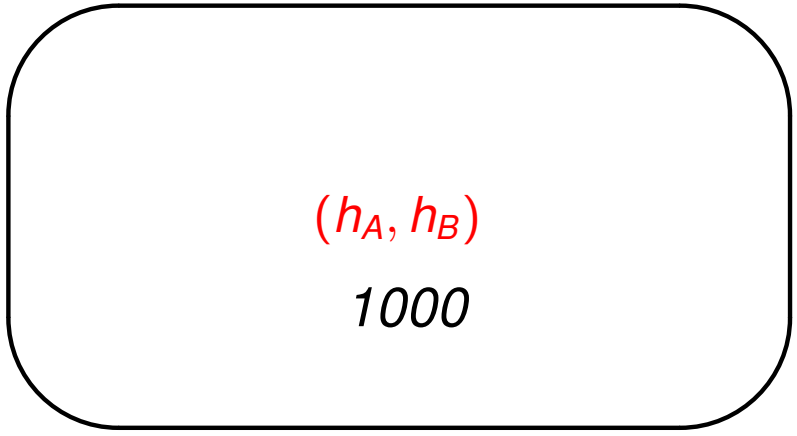
primary+secondary risk



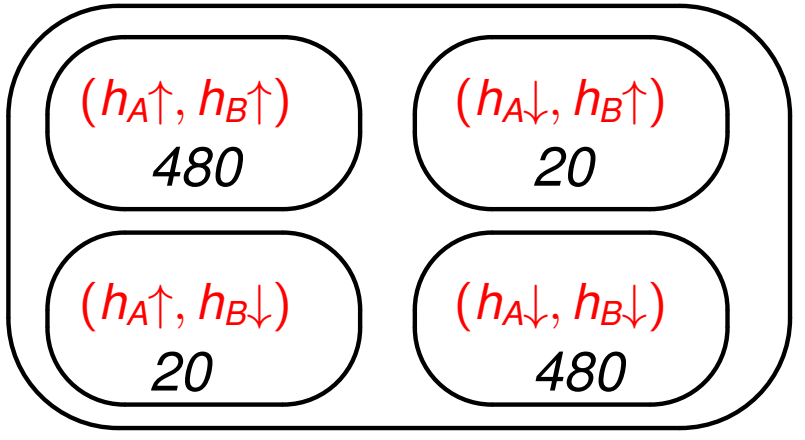
Link between cohort heterogeneity and informative censoring

Say 1000 people,
two risks, hazard rates h_A and h_B

- homogeneous cohort:
  all *individuals* have (h_A, h_B)



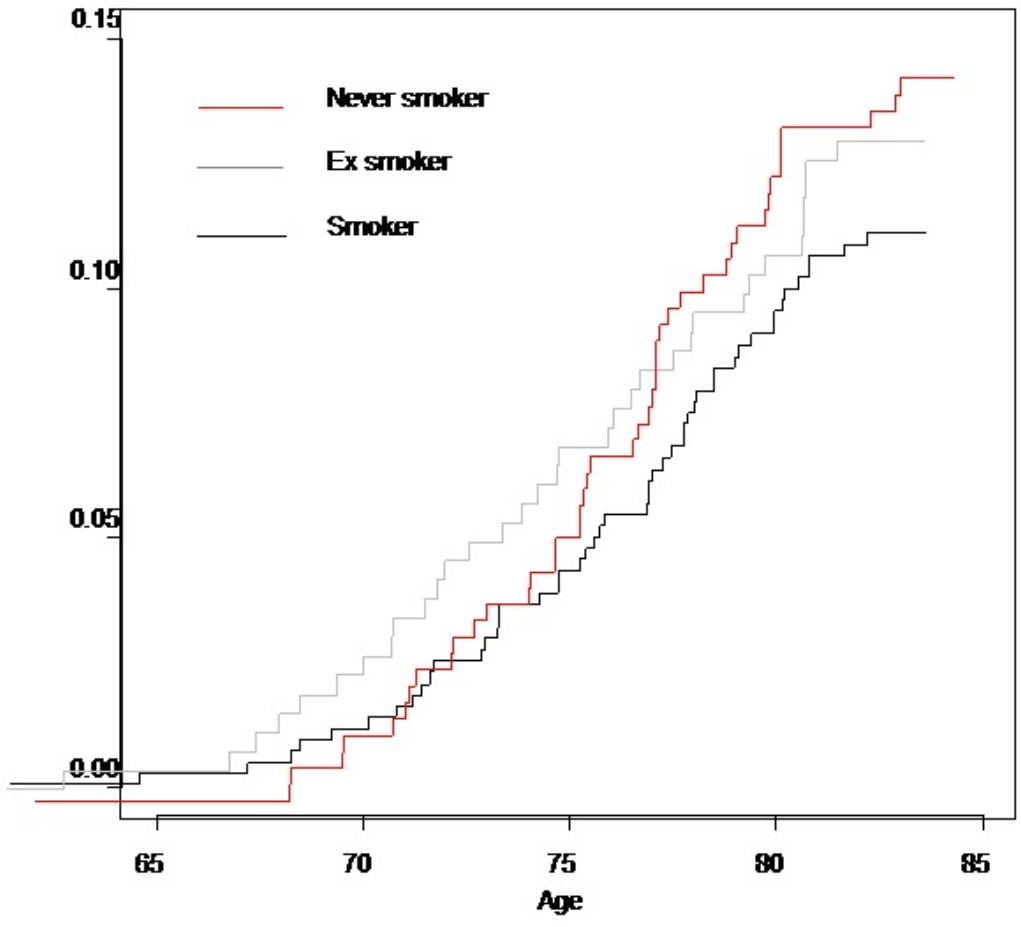
- heterogeneous cohort,
  informative cohort filtering



If our tools assume *censoring risks are <u>uncorrelated</u> with primary risk*

censoring by competing risks
can give nonsensical results ...

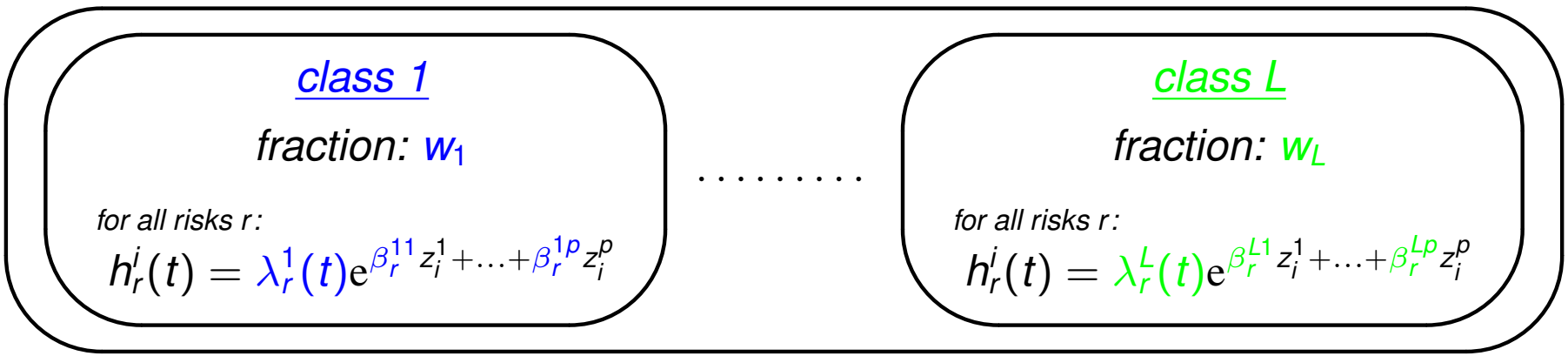
– harmful drugs look beneficial
– beneficial drugs look harmful
– false protectivity of covariates



(ULSAM cancer data)

Bayesian latent class methods

- model all risks simultaneously
- individuals with *same* covariates can have *distinct* associations and *distinct* base hazard rates
- competing risks, informative censoring: reflect correlated association parameters of different risks



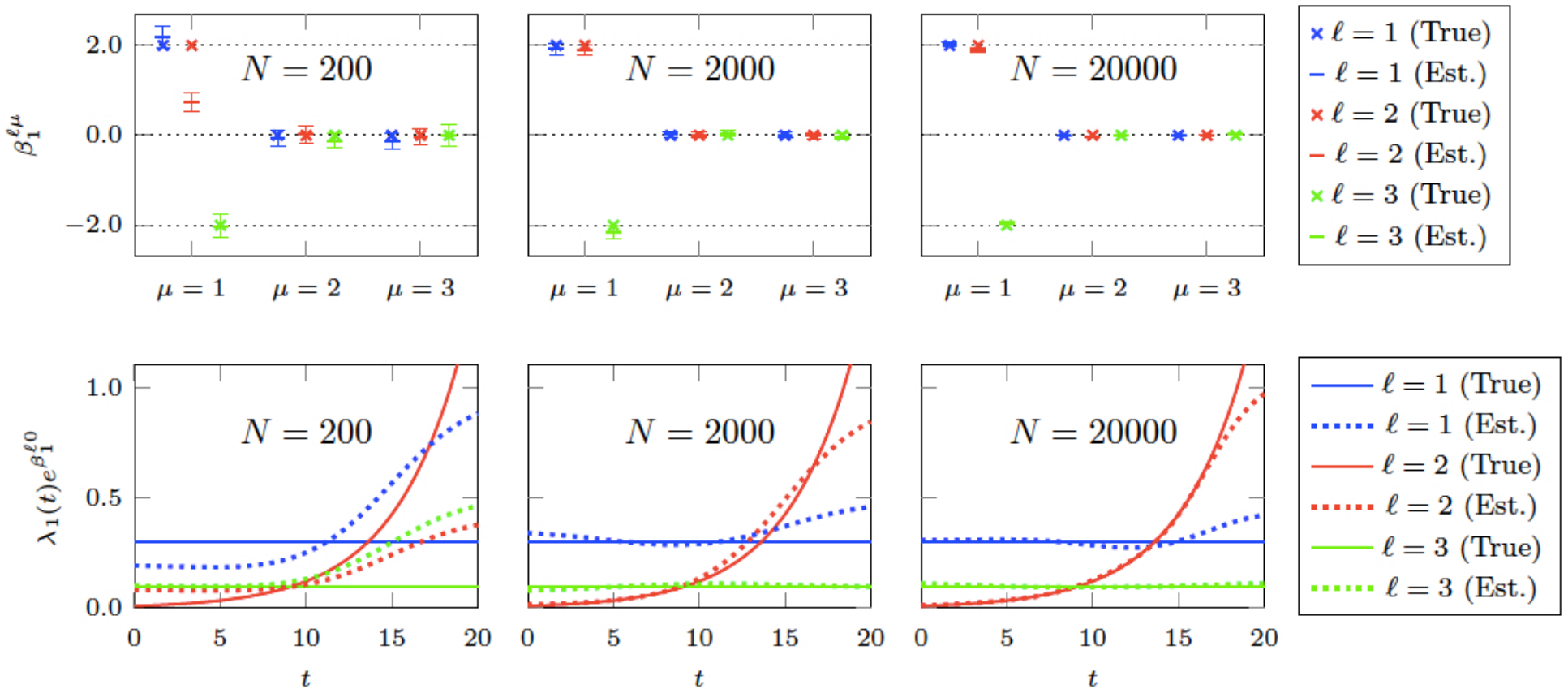
Personalised cause-specific hazard rate model variants		
$M = 1$	Heterogeneous frailties Homogeneous associations Homogeneous base hazard rates	$h_r^i(t) = \lambda_r(t) \mathrm{e}^{\beta_r^{\ell 0} + \sum_\mu \beta_r^\mu z_i^\mu}$
$M = 2$	Heterogeneous frailties Heterogeneous associations Homogeneous base hazard rates	$h_r^i(t) = \lambda_r(t) \mathrm{e}^{\beta_r^{\ell 0} + \sum_\mu \beta_r^{\ell\mu} z_i^\mu}$
$M = 3$	Heterogeneous frailties Heterogeneous associations Heterogeneous base hazard rates	$h_r^i(t) = \lambda_r^\ell(t) \mathrm{e}^{\beta_r^{\ell 0} + \sum_\mu \beta_r^{\ell\mu} z_i^\mu}$

- Bayesian analysis and model selection:
  reliable error bars, and multiple classes *only if data demand it*
- reduces to standard Cox regression if no heterogeneity
  (Occam's Razor action of Bayesian model selection)
- formulae for survival curves *decontaminated* for informative censoring,
  and *retrospective class allocation* of individuals

(Rowley et al, 2016)

Synthetic data

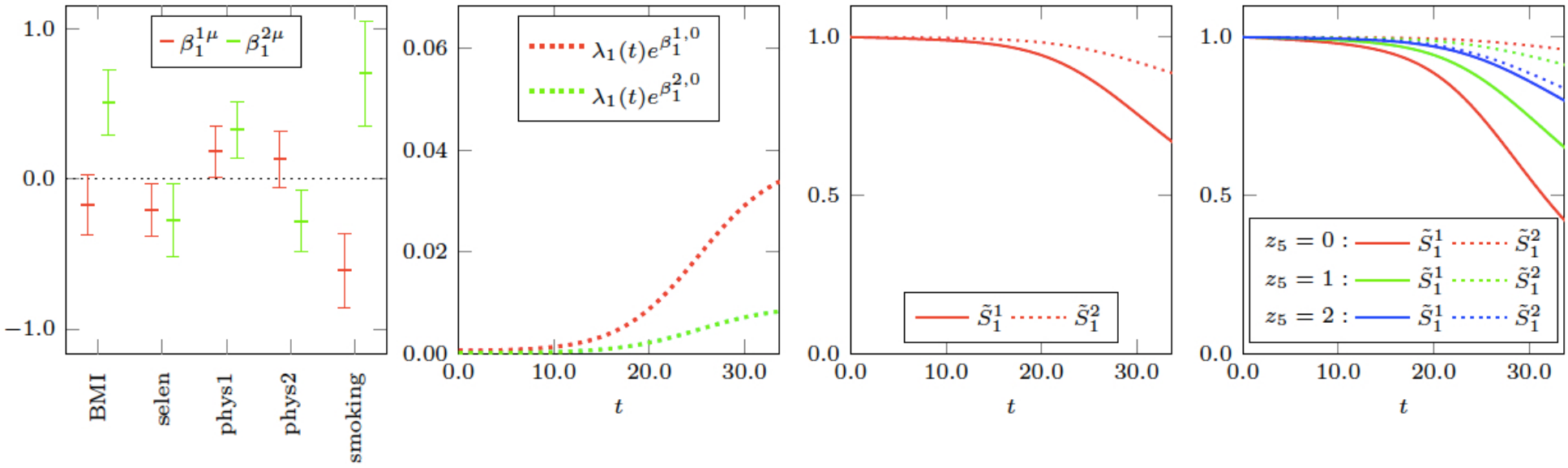
3 classes:
red, blue, green



Prostate cancer data

(ULSAM data base, $n = 2047$)

Cox regression:
smoking is protective against PC



negative association with smoking *only* in *extremely frail* subgroup of patients

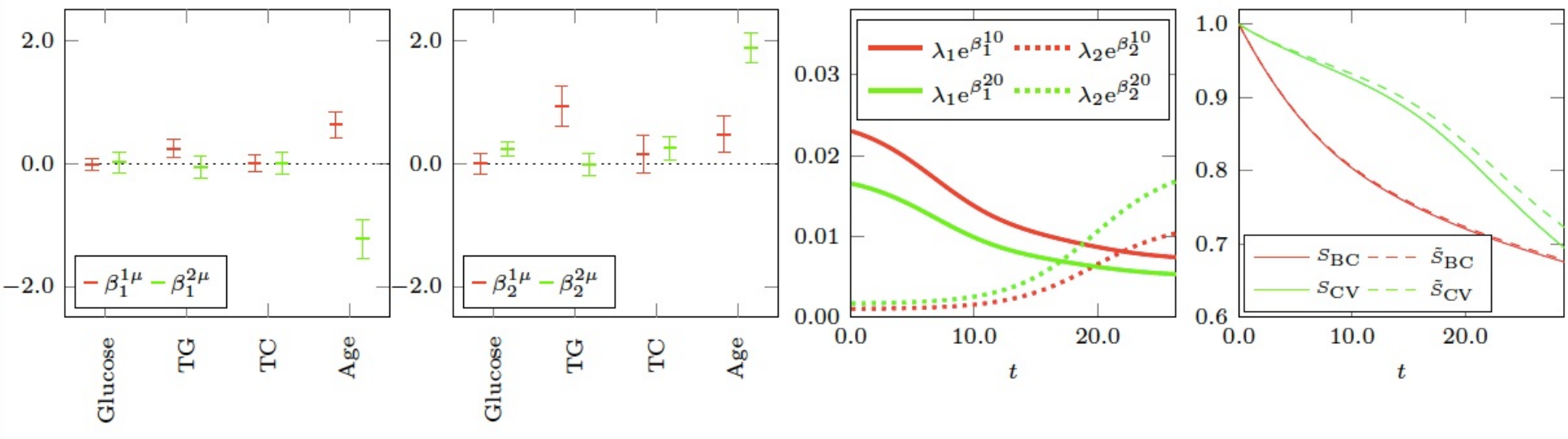
red class: high overall frailty
green class: low overall frailty

Breast cancer data

(AMORIS data base, $N = 1798$)

Cox regression finds no significant associations
(proportional hazards violated)

BC death *CV death* *solid: BC dashed: CV*



red class: predominantly younger women
green class: predominantly older women

(Wulaningsih et al, BMC Cancer 2015)

Addition of cetuximab to oxaliplatin-based first-line combination chemotherapy for treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial

Timothy S Maughan, Richard A Adams, Christopher G Smith, Angela M Meade, Matthew T Seymour, Richard H Wilson, Shelley Idziaszczyk, Rebecca Harris, David Fisher, Sarah L Kenny, Edward Kay, Jenna K Mitchell, Ayman Madi, Bharat Jasani, Michelle D James, John Bridgewater, M John Kennedy, Bart Claes, Diether Lambrechts, Richard Kaplan, Jeremy P Cheadle, on behalf of the MRC COIN Trial Investigators

Lancet 2011; 377: 2103–14
Published Online
June 4, 2011

Summary

Background In the Medical Research Council (MRC) COIN trial, the epidermal growth factor receptor (EGFR)-targeted antibody cetuximab was added to standard chemotherapy in first-line treatment of advanced colorectal cancer with the aim of assessing effect on overall survival.

outcome:

Interpretation This trial has not confirmed a benefit of addition of cetuximab to oxaliplatin-based chemotherapy in first-line treatment of patients with advanced colorectal cancer. Cetuximab increases response rate, with no evidence of benefit in progression-free or overall survival in *KRAS* wild-type patients or even in patients selected by additional mutational analysis of their tumours. The use of cetuximab in combination with oxaliplatin and capecitabine in first-line chemotherapy in patients with widespread metastases cannot be recommended.

The COIN trial (colorectal cancer)

1st batch, $n = 154$

HR [95% CI]	$\beta(0)$	FRET eff	Her2-Her3	Cetux	KRASmut
Cox (M1L1K1), *lnZ=-946.63*					
	0.02 pm 0.09	0.7 [0.4-1.1]	1.7 [1.1-2.8]	0.7 [0.5-0.9]	1.6 [1.1-2.2]
		p=0.1	p=0.03	p=0.02	p=0.008
Model M2L2K3(R), *lnZ=-941.21*					
class I, W=64%	*-1.14*	0.5 [0.2-1.0]	1.1 [0.5-2.5]	1.1 [0.6-1.8]	1.4 [0.8-2.3]
alloc[$p_1 > 0.5$]: N=128		p=0.047	p=0.8	p=0.8	p=0.2
class II, W=36%	*-2.16*	0.8 [0.06-9.6]	3.4 [0.2-71]	0.3 [0.09-0.8]	3.0 [1.3-7.0]
alloc[$p_2 > 0.5$]: N=26		p=0.8	p=0.4	p=0.01	p=0.01

- two sub-cohorts, with similar base hazard rates, but distinct overall frailties and associations.
- method provides retrospective class assignment
- new tools to identify *a priori* the responders to Cetuximab?

(Ng et al, ASCO 2015)

The COIN trial (colorectal cancer)

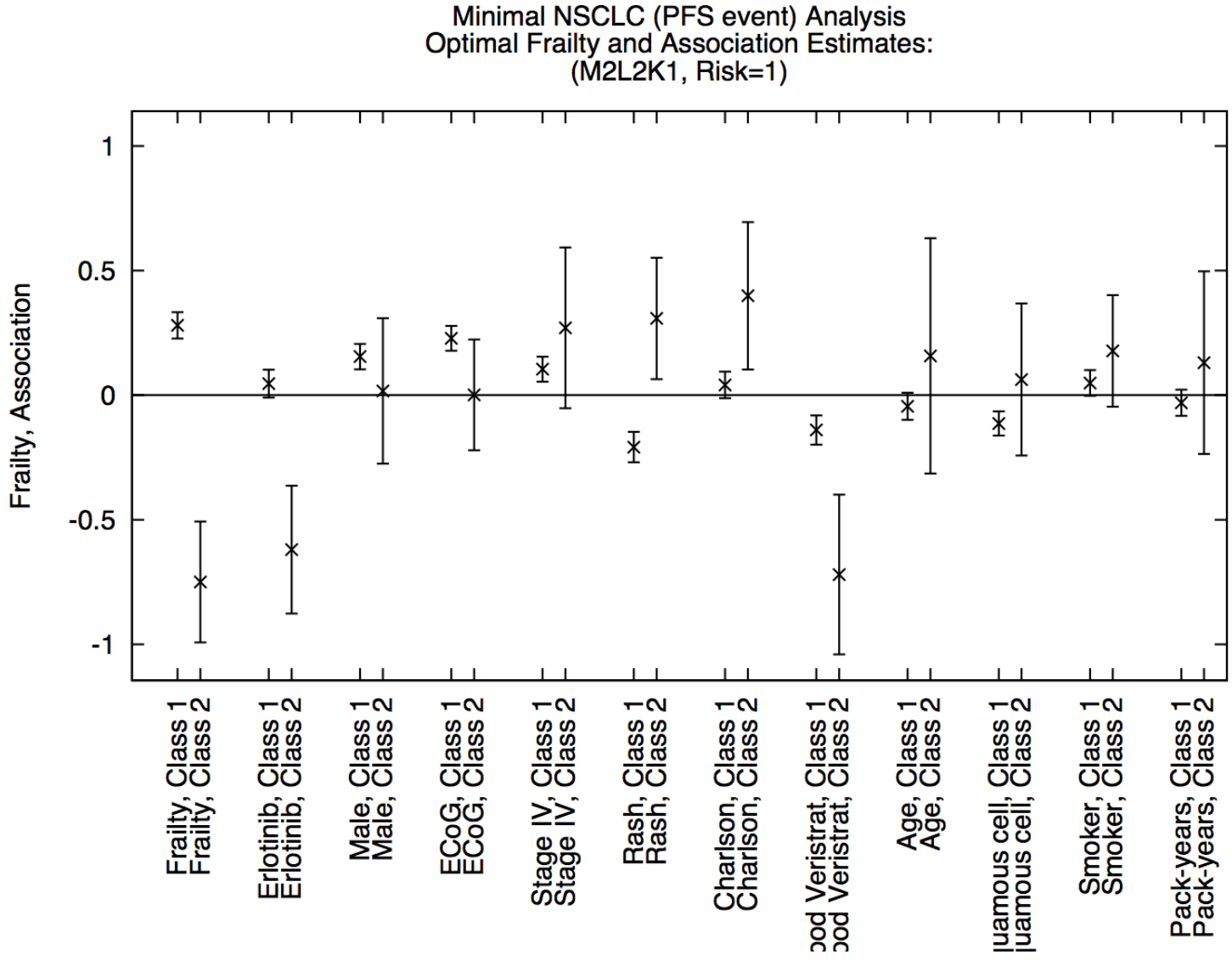
1st+2nd batch, $n = 398$

HR [95% CI]	$\beta(0)$	FRET eff	Her2-Her3	Cetux	KRASmut
Cox (M1L1K5), *lnZ=-2419.82*					
	-1.89	0.9 [0.7-1.0]	1.1 [0.9-1.5]	0.8 [0.7-0.9]	1.3 [1.1-1.7]
		p=0.3	p=0.4	p=0.03	p=0.006
Model M2L2K4(R), *lnZ=-2418.064*					
class I, W=31%	*-2.57*	1.8 [0.8-4.6]	0.8 [0.4-1.7]	0.5 [0.3-1.0]	1.5 [0.9-2.6]
alloc[$p_1 > 0.5$]: N=59		p=0.2	p=0.6	p=0.05	p=0.2
class II, W=69%	*-1.56*	0.5 [0.4-0.8]	1.4 [0.9-2.1]	1.0 [0.7-1.4]	1.3 [0.9-1.9]
alloc[$p_2 > 0.5$]: N=339		p=0.006	p=0.1	p=0.8	p=0.1

- two sub-cohorts, with similar base hazard rates, but distinct overall frailties and associations.
- method provides retrospective class assignment
- new tools to identify *a priori* the responders to Cetuximab?

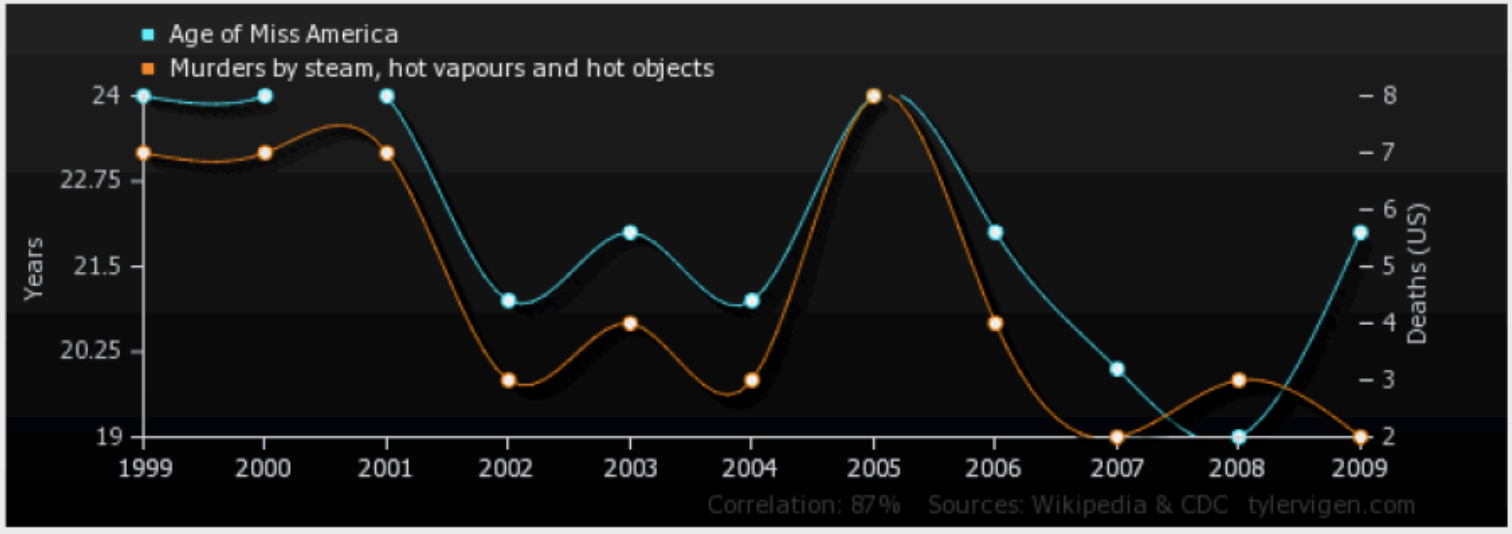
The TOPICAL trial (lung cancer)

$n = 670$



Overfitting

Age of Miss America
correlates with
Murders by steam, hot vapours and hot objects



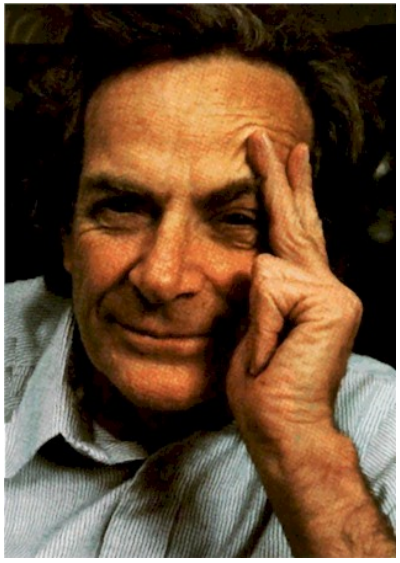
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Age of Miss America *Years (Wikipedia)*	24	24	24	21	22	21	24	22	20	19	22
Murders by steam, hot vapours and hot objects *Deaths (US) (CDC)*	7	7	7	3	4	3	8	4	2	3	2

Correlation: 0.870127

common sense:

if you look for long enough, you will always find some signal,
the problem is how to distinguish between true and fluke ones

Pearson correlation: 0.87
surely statistically significant?



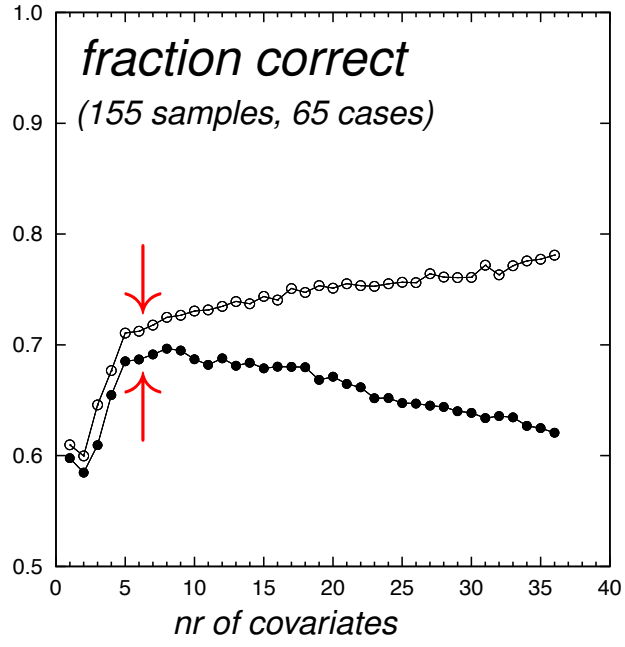
Feynman would suddenly interrupt himself in the middle of a statistics lecture, and excitedly say something like:

'On my way to campus today, I saw a car with the licence plate XRT-375 in the parking lot - isn't that amazing? What are the odds of seeing that exact licence?'

'shoot randomly at a wall,
then draw target circles around the bullet holes ...'

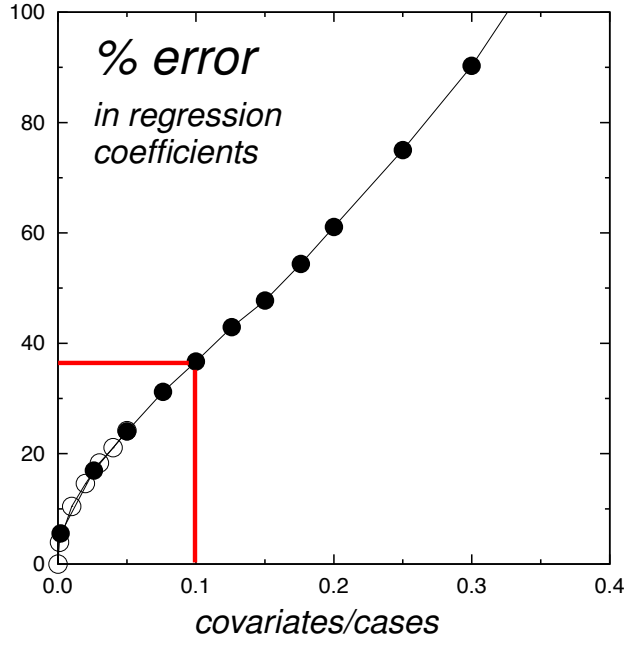
overfitting in multivariate
Cox regression

p-values, z-scores,
confidence intervals
don't measure overfitting!



rule of thumb:
'10 cases per covariate'
too optimistic ...

uncorrelated covars
o: *1000 samples*
●: *500 samples*



Strategies to deal with overfitting

in covariate-to-outcome analysis

- *Know when to 'back off'*

  'safe' ratio covariates/samples
  for Cox regression?

- *Eliminate redundant information*

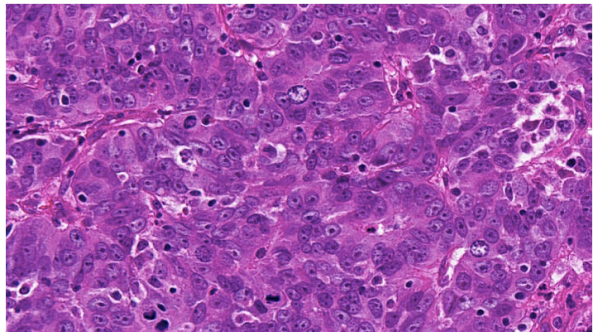
  improve covariates/samples ratio
  nonlinear dimension reduction (information theory),
  'true' data dimension?

- *Model (avoid?) overfitting effects*

  statistics of full parameter uncertainty,
  while keeping computation feasible

Empirical approach

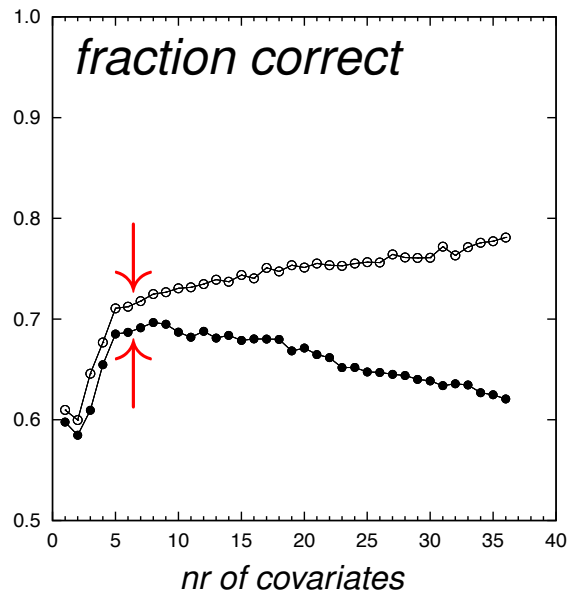
multivariate Bayesian Cox,
(MAP, Gaussian priors $\rightarrow$ L2 regularisation)
+ outcome prediction (Breslow's base rate)
+ cross-validation



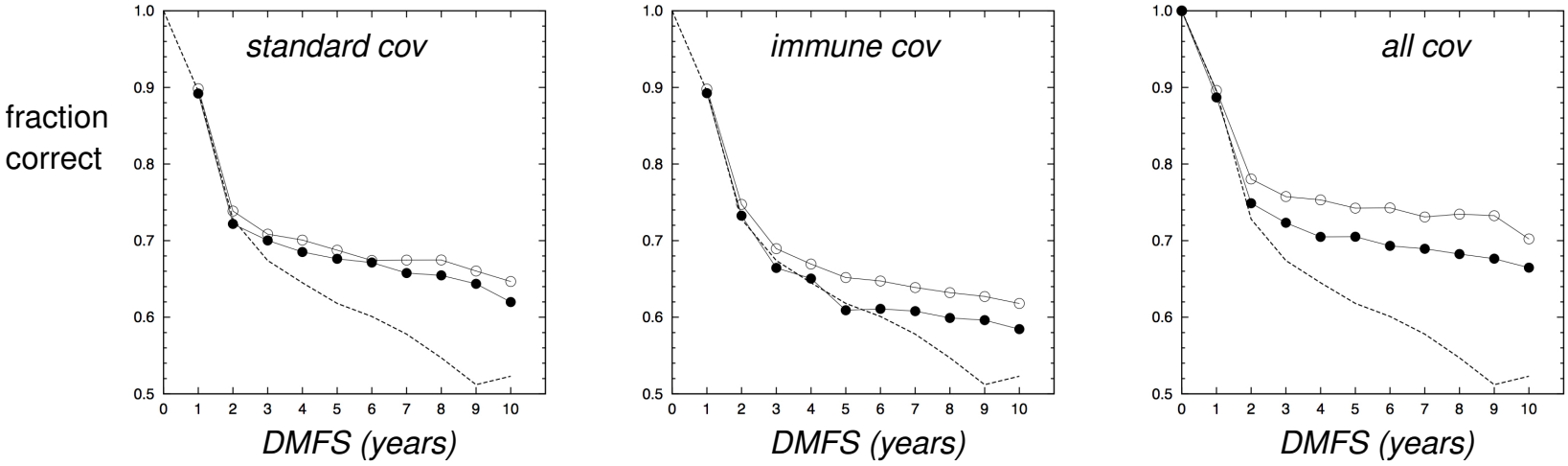
lymphocyte infiltration markers
to predict clinical outcome in BC
(Gazinska, Grigoriadis, Tutt, Pinder)

- split data set:
  50% training, 50% validation
- regression on training set,
  with all covariates

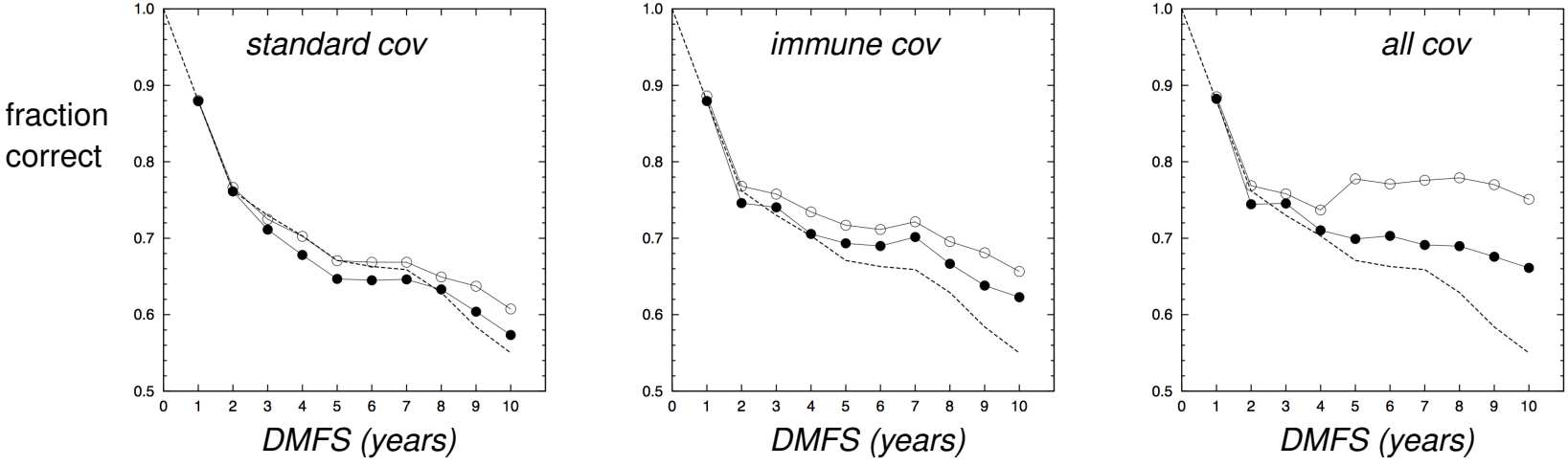
  prediction performance on training set?
  prediction performance on validation set?
- remove least informative covariate and repeat,
  (many random separations into
  training/validation sets, many cutoff times)
- identify optimal set of covariates



All BC
n=309



TNBC only
n=170



Bayesian latent variable methods

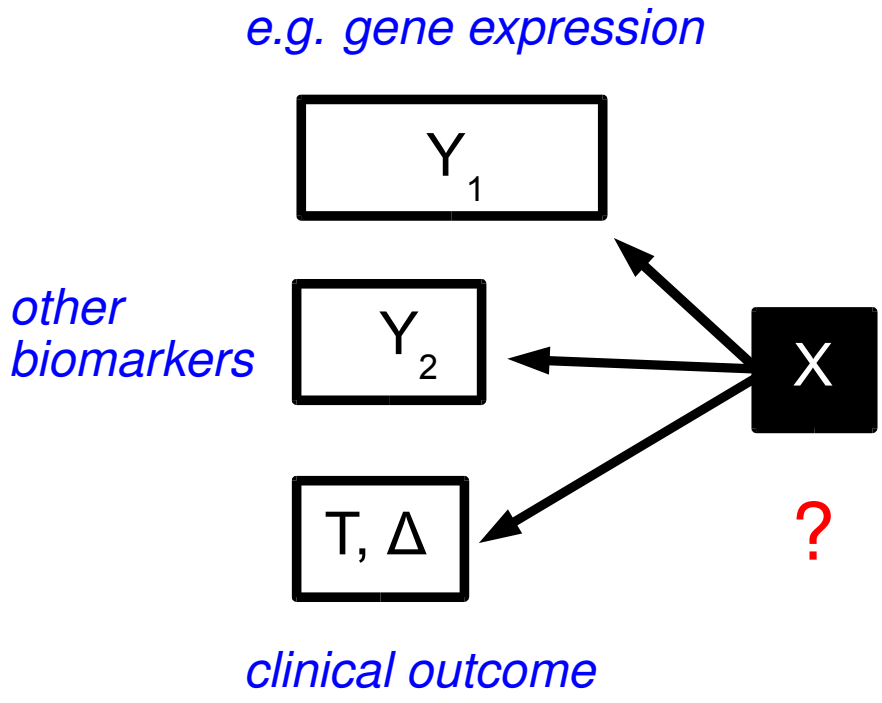
for survival analysis

Assume:

(a) data Y_k are *high-dim windows* on q-dim latent variables X

(b) X actually drives outcome

(c) dimension of X less than dimension of Y_k



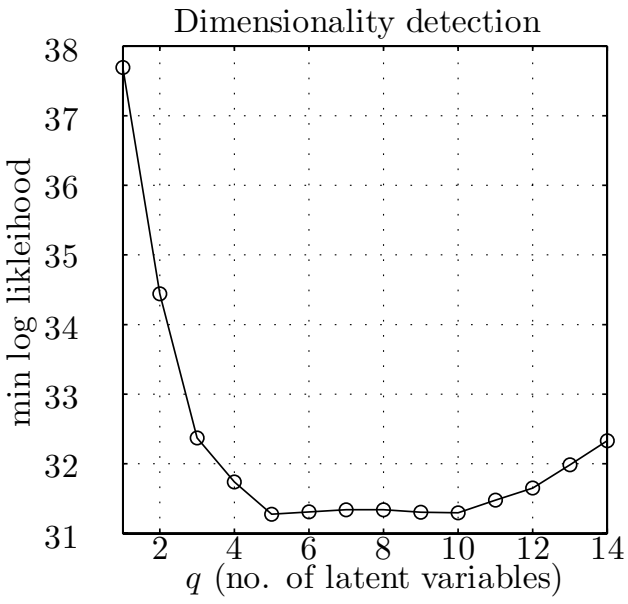
- nonlinear stochastic relations $Y_k = f_k(X) + \textit{noise}$
- dimension detection: optimal q?
- find most probable latent variables X
- use X to predict clinical outcome

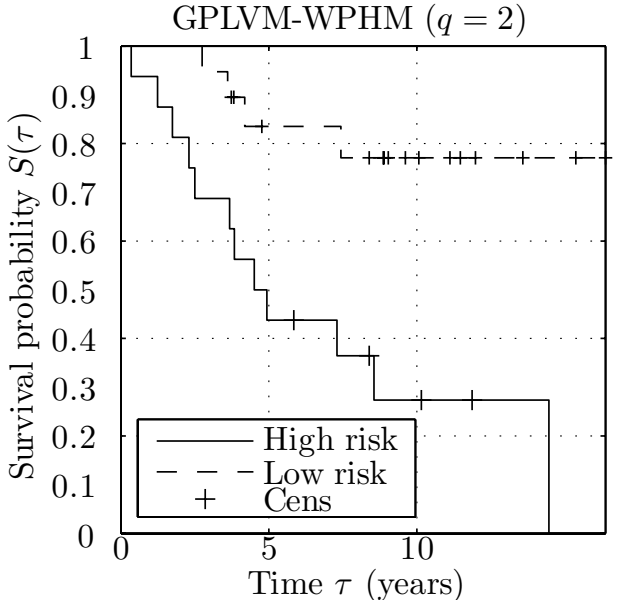
Gaussian process latent variable model (GPLVM) combined with Weibull proportional hazards model (WPHM)

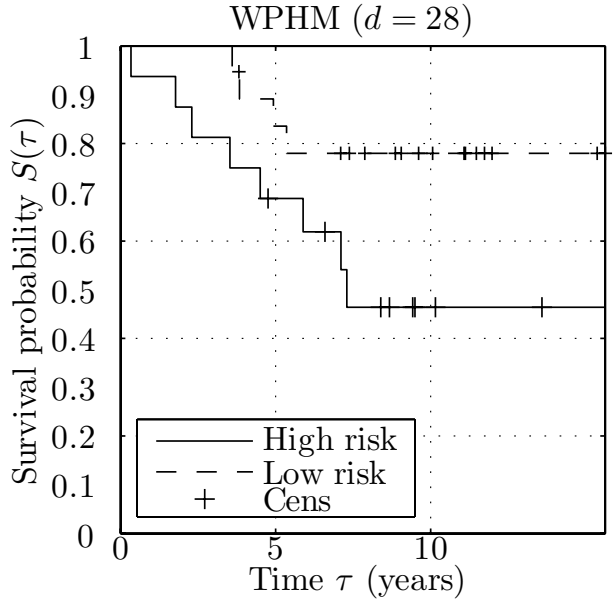
Application to METABRIC

BC gene signature data

data Y : scores of 28 gene signatures
outcome: overall survival time







left: extract dimension of X
from training set ($n = 74$)

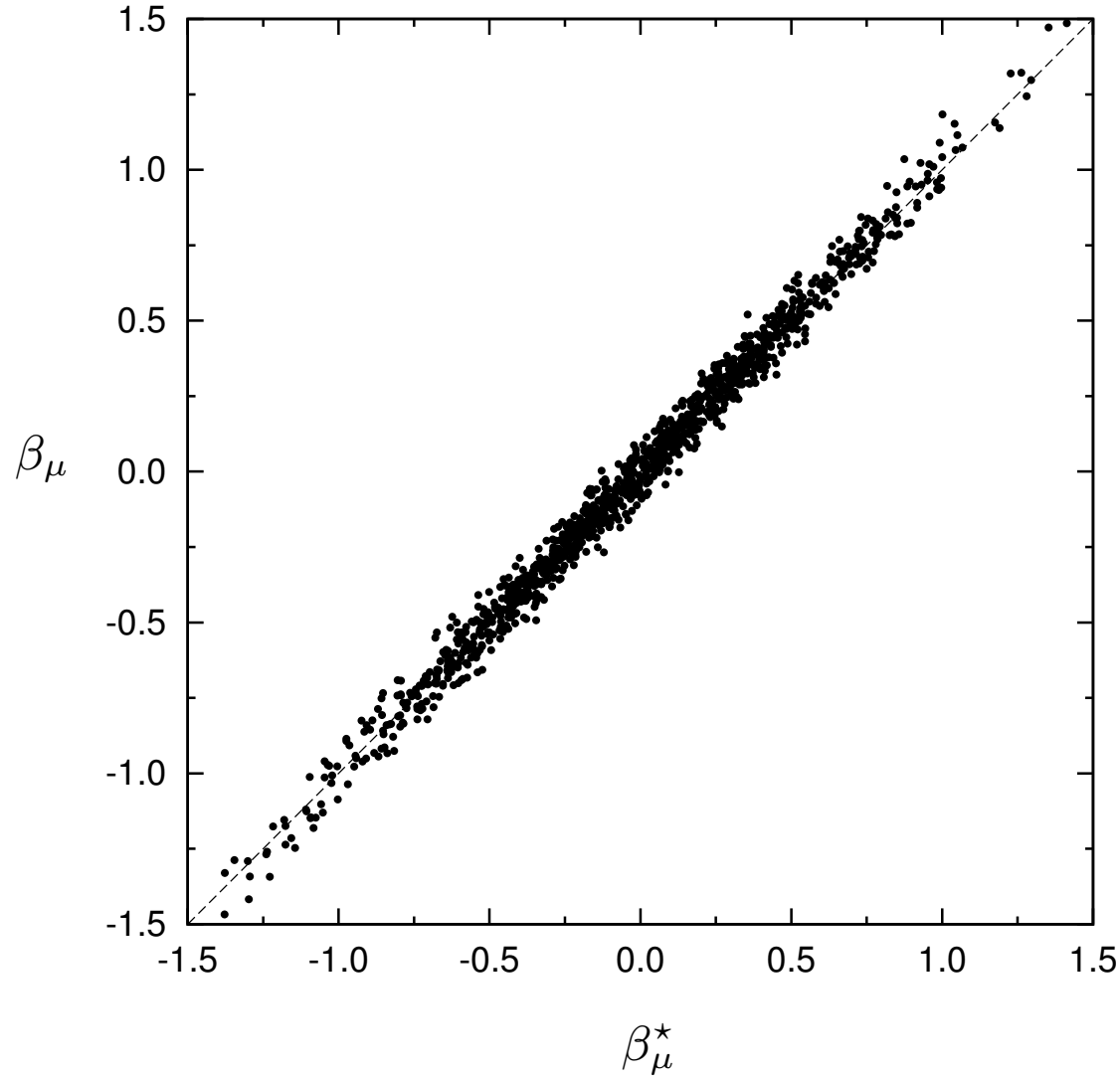
right: predict low/high risk groups directly from Y
(tested in validation set, $n = 74$)

middle: predict low/high risk groups from X ($q = 2$)
(tested in validation set, $n = 74$)

(Barrett & Coolen,
Stats in Medicine 2015)

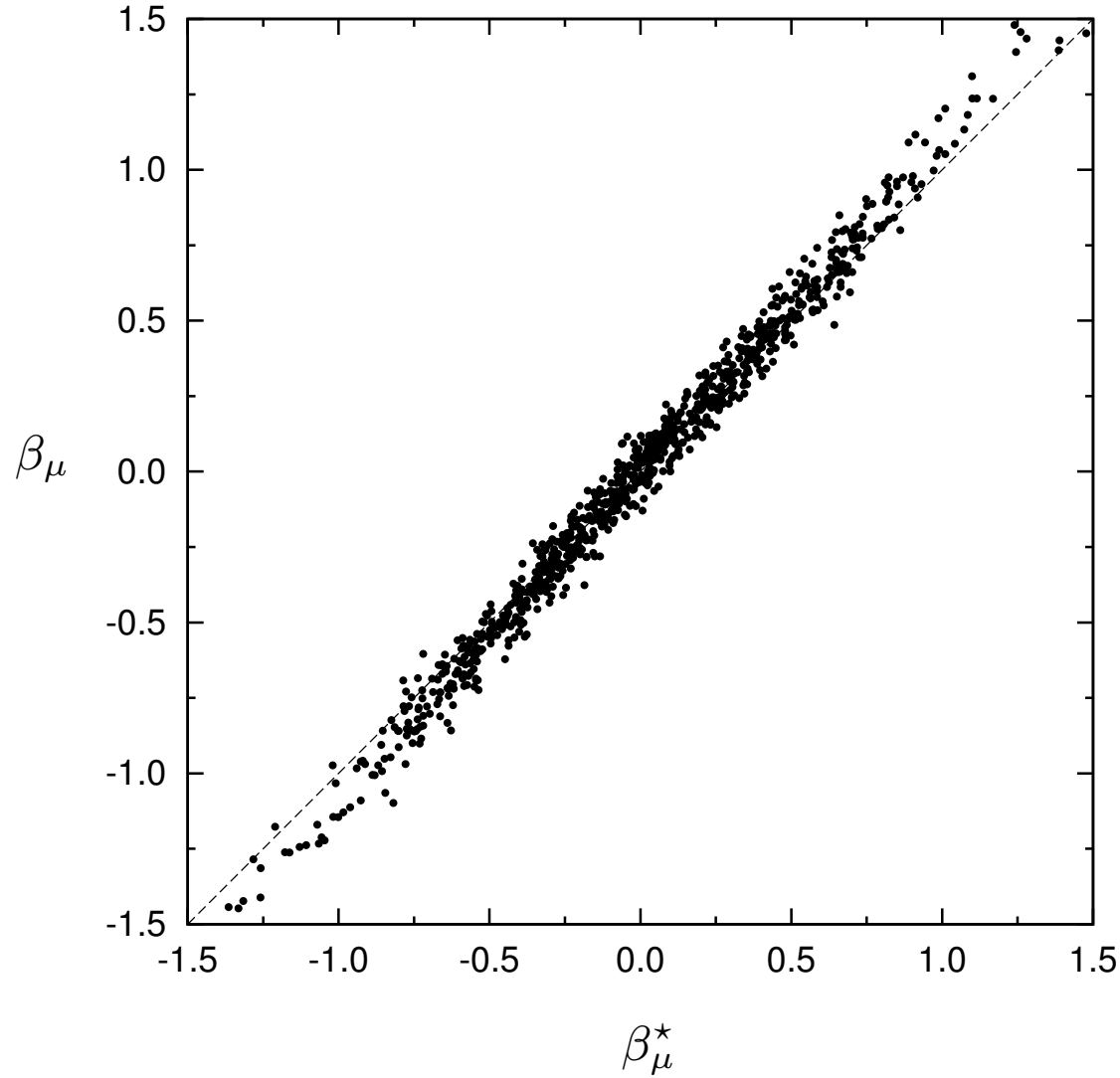
$n = 500$,
predicted versus true regression coefficients
$\mathrm{HR}_\mu = \exp(2\beta_\mu)$

$d/n = 0.002$



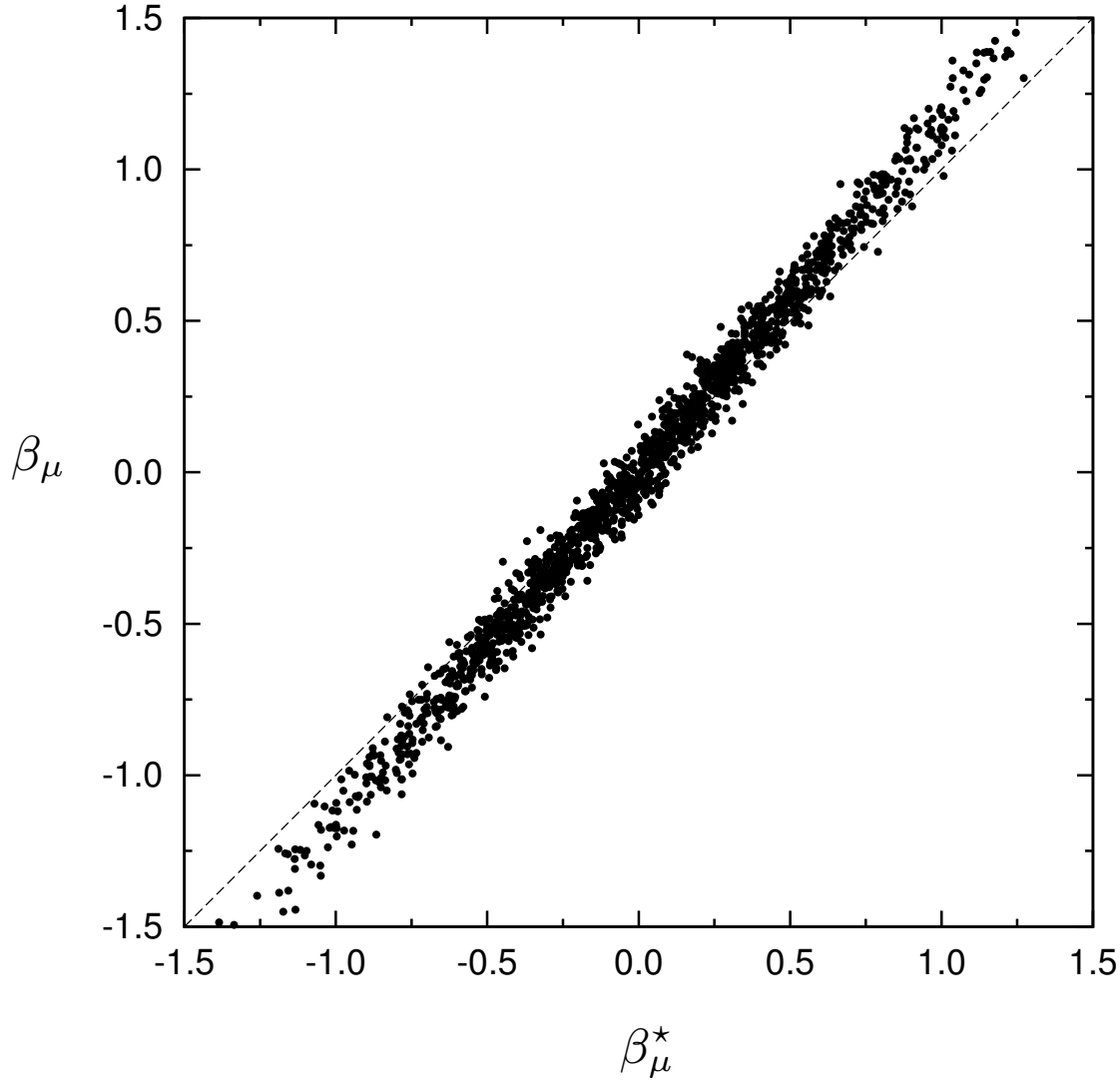
$n = 500$,
predicted versus true regression coefficients
$\mathrm{HR}_\mu = \exp(2\beta_\mu)$

$d/n = 0.10$



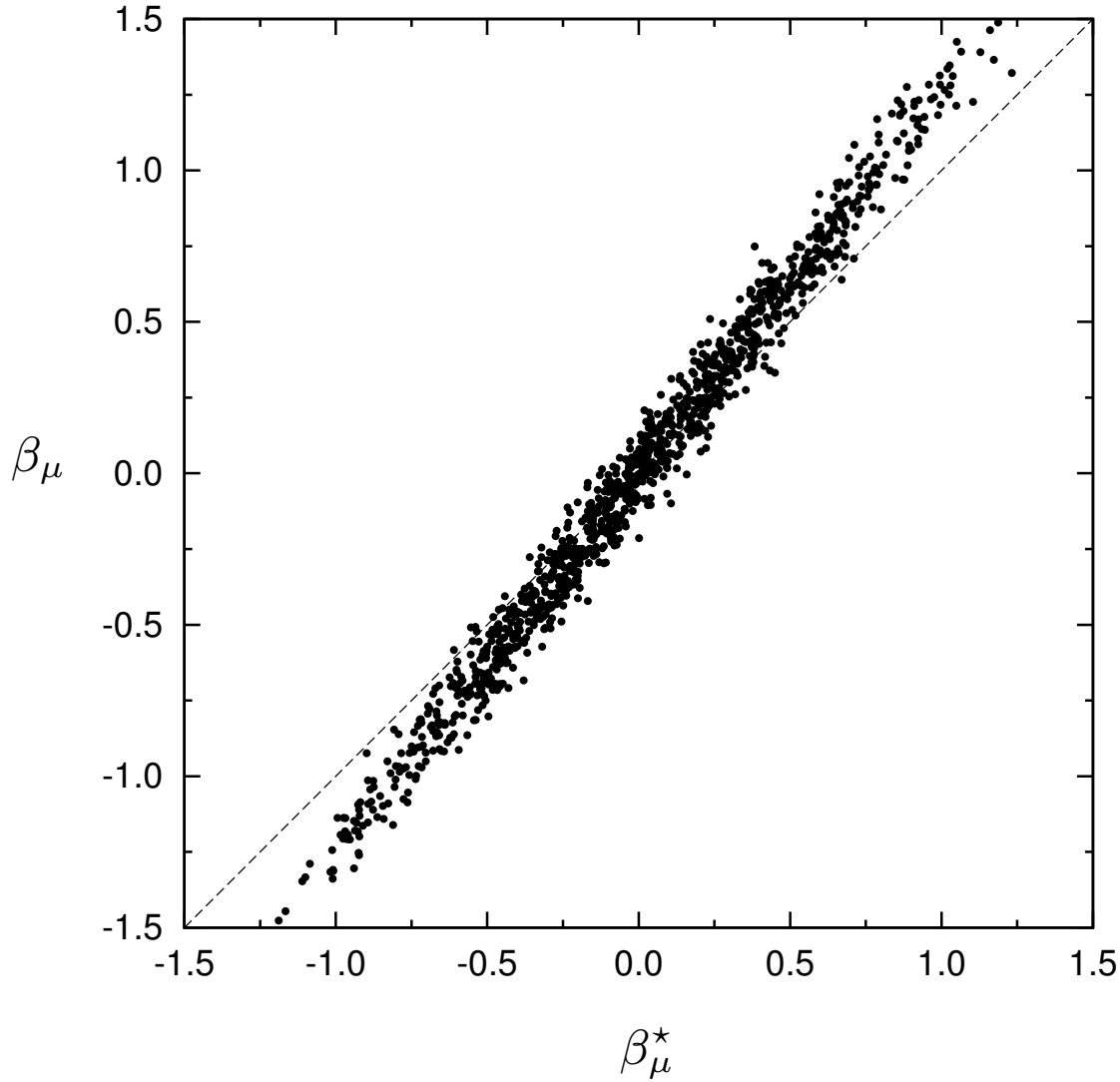
$n = 500$,
predicted versus true regression coefficients
$\mathrm{HR}_\mu = \exp(2\beta_\mu)$

$d/n = 0.20$



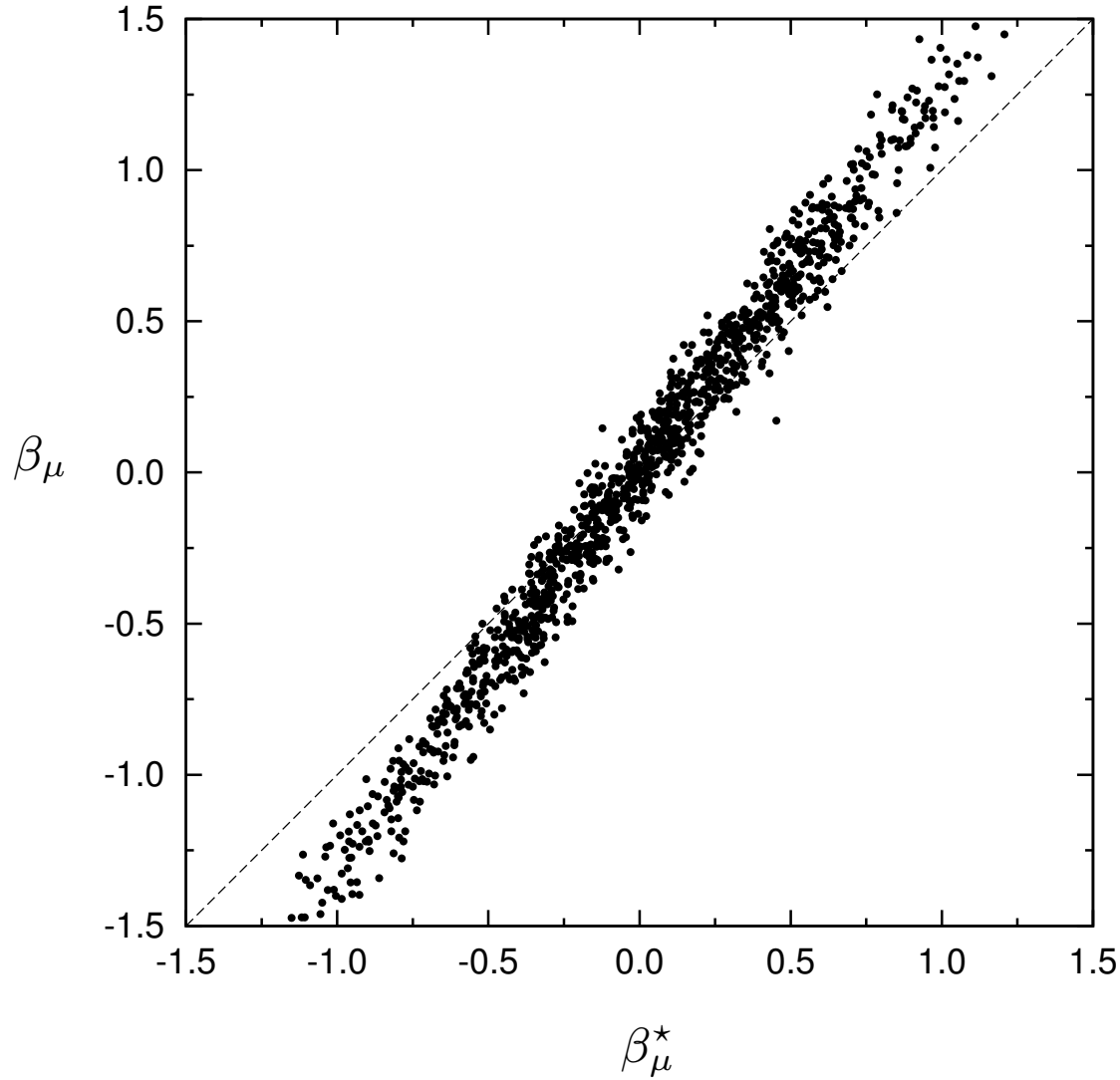
$n = 500$,
predicted versus true regression coefficients
$\mathrm{HR}_\mu = \exp(2\beta_\mu)$

$d/n = 0.30$



$n = 500$,
predicted versus true regression coefficients
$\mathrm{HR}_\mu = \exp(2\beta_\mu)$

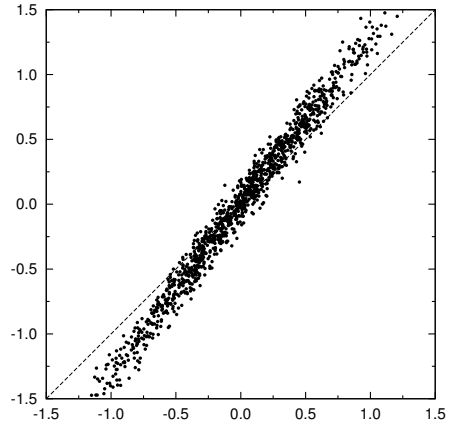
$d/n = 0.40$



Bad news

Overfitting is more dangerous than pure noise,
because it is causes *deterministic bias*

> *we always overestimate the*
> *strength of associations*
> *(whether positive or negative)*



Good news

Unlike pure noise,
deterministic bias may be predictable ...

New possibilities, roadmap for research ...

- Predict impact of overfitting, in terms of
  - sample size, nr of covariates
  - correlations among covariates
  - true association strengths
- Overfitting correction of Cox parameters
  - reliable regression at ratios covariates/samples ~ 0.5?
  - Bayesian Cox regression with unbiased estimates?

Overfitting in Cox model – intuition

- Empirical distribution of covariates and event times, generated from Cox model with parameters $\beta^\star$ in cohort of size n:

$$\hat{P}_{\beta^\star}(t, \mathbf{z}|) = \frac{1}{n} \sum_{i=1}^{n} \delta(t - t_i)\delta(\mathbf{z} - \mathbf{z}_i)|_{\beta^\star}$$

  Cox regression minimises Kullback-Leibler distance between parametrisation P_β and empirical distribution:

$$\beta_{\text{Cox}} = \text{argmin}_\beta \, D(\hat{P}_{\beta^\star} || P_\beta)$$

- If $n \to \infty$ for fixed d:

$$\lim_{n\to\infty} \hat{P}(\beta^\star) = P(\beta), \qquad \beta_{\text{Cox}} = \text{argmin}_\beta D(P_{\beta^\star} || P_\beta) = \beta^\star$$

  If $d \sim n$:

$$\lim_{n\to\infty} \hat{P}(\beta^\star) \neq P(\beta) \ \ldots$$

Overfitting in Cox model – analysis

- Regression performance:

$$E(\boldsymbol{\beta}^\star) = \min_{\boldsymbol{\beta}} D(\hat{P}_{\boldsymbol{\beta}^\star}||P_{\boldsymbol{\beta}}) - \overbrace{D(\hat{P}_{\boldsymbol{\beta}^\star}||P_{\boldsymbol{\beta}^\star})}^{\textit{not zero}}$$

$$E(\beta^\star) > 0: \quad \text{underfitting}$$
$$E(\boldsymbol{\beta}^\star) = 0: \quad \text{optimal regression}$$
$$E(\boldsymbol{\beta}^\star) < 0: \quad \text{overfitting}$$

- Typical behaviour:

$$\begin{aligned} E &= \left\langle \min_{\boldsymbol{\beta}} \left\{ \frac{1}{N} \sum_i \log \left[\frac{P(t_i|\mathbf{z}_i, \boldsymbol{\beta}^\star)}{P(t_i|\mathbf{z}_i, \boldsymbol{\beta})} \right] \right\} \right\rangle_D \\ &= -\lim_{\gamma\to\infty} \frac{1}{\gamma} \left\langle \log \int \mathrm{d}\boldsymbol{\beta}\, \mathrm{e}^{-\gamma \left\{ \frac{1}{N} \sum_i \log \left[P(t_i|\mathbf{z}_i,\boldsymbol{\beta}^\star)/P(t_i|\mathbf{z}_i,\boldsymbol{\beta}) \right] \right\}} \right\rangle_D \end{aligned}$$

- Replica method:
  use $\langle \log Z \rangle = \lim_{m\to 0} m^{-1} \log \langle Z^m \rangle$,
  analytical continuation from integer to non-integer m

result:
explicit eqns from which to solve E and overfitting correction factor

Epilogue

- We should not adapt our medical questions to
  the limitations of current statistical tools,
  but *build statistical tools* that answer our questions

- Two of the main challenges in modern survival analysis
  are *latent cohort heterogeneity* and *overfitting*

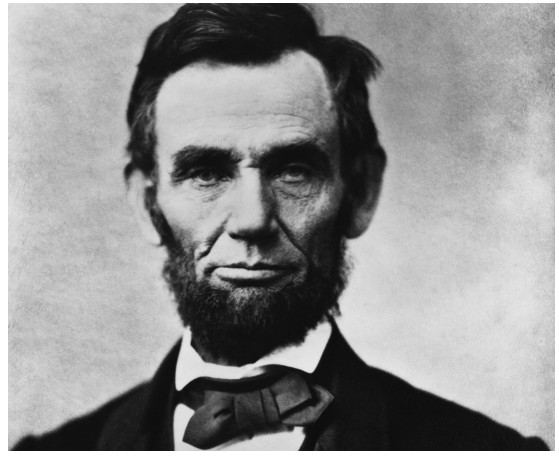
- There is no *scientific* obstacle that prevents us from developing
  new statistical tools tailored to the challenges of modern medicine

- Cultural obstacles

  – aversion to statistical innovation (journals, grant awarding bodies)

  – uncritical attitude towards appropriateness of standard methods

  – undue focus on univariate hazard ratios and p-values,
    instead of multivariate *outcome prediction*, with error bars



‘Give me six hours to chop down a tree
and I will spend the first four sharpening the axe’

with thanks to

IMMB @ KCL

Paul Barber, James Barrett,
Katherine Lawler, Mark Rowley

Cancer Division @ KCL

Hans Garmö, Anita Grigoriadis, Tony Ng,
Mieke van Hemelrijck, Lars Holmberg,
Wahyu Wulaningsih

Cancer Institute @ UCL

James Barrett

Waseda University, Tokyo

Masato Inoue

Funding:

BBSRC, EPSRC, MRC, EU FP-7,
Ana Leaf Foundation, Prostate UK

papers, seminars, notes:
www.mth.kcl.ac.uk/∼tcoolen