Unlocking the potential of large prospective biobank cohorts for -omics data analysis:

aspects of study design, prediction and causality

Krista Fischer, Ph.D

Barcelona, May 2016

estonian genome center
university of tartu

Machine learning vs Statistics

From R-Bloggers (www.r-bloggers.com)

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant = $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

The potential in large prospective biobank cohorts

- Sample sizes of 20000 – 500000
- Availability of molecular data: DNA genotyping, transcriptomics, metabolomics, etc.
- Follow-up information from electronic health records, national registries, etc.

- Biobank-based studies are paving the path towards the implementation of personalized medicine!

estonian genome center
university of tartu

Biobank cohorts have brought a new era…

Conventional epidemiology

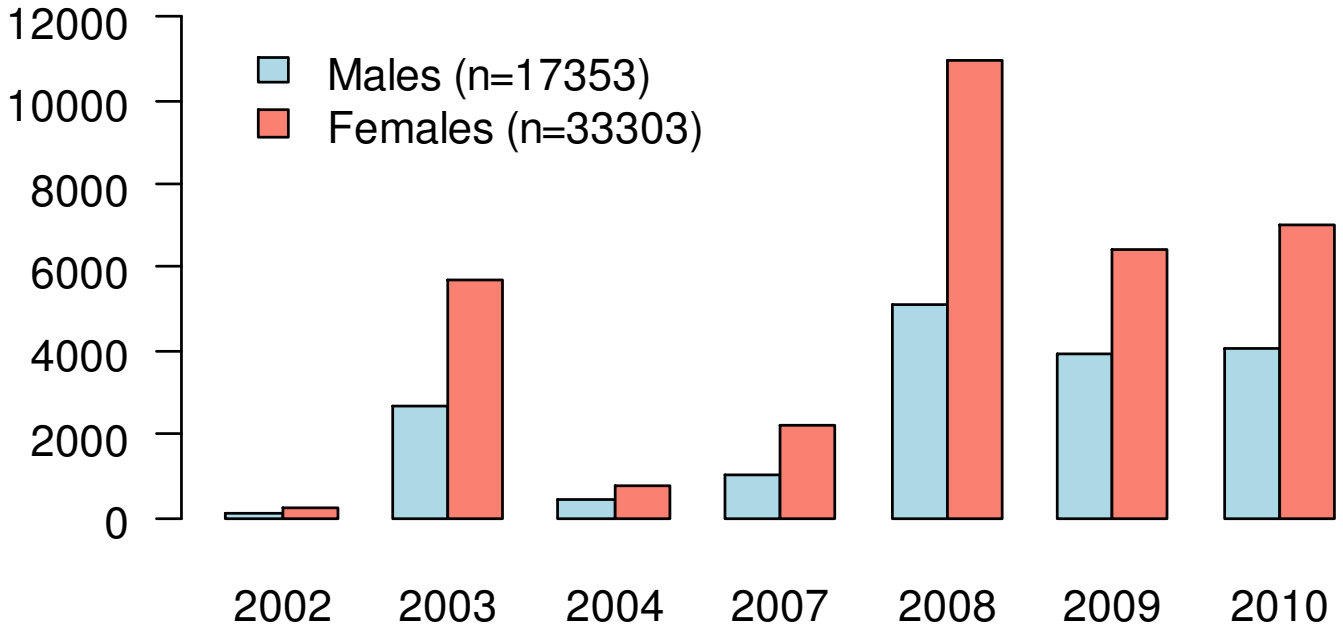
- Sample sizes: 1000+
- Studies on lifestyle-related risk factors on diseases and mortality
- Almost always subject to unmeasurable confounding

Genetic epidemiology based on biobanks

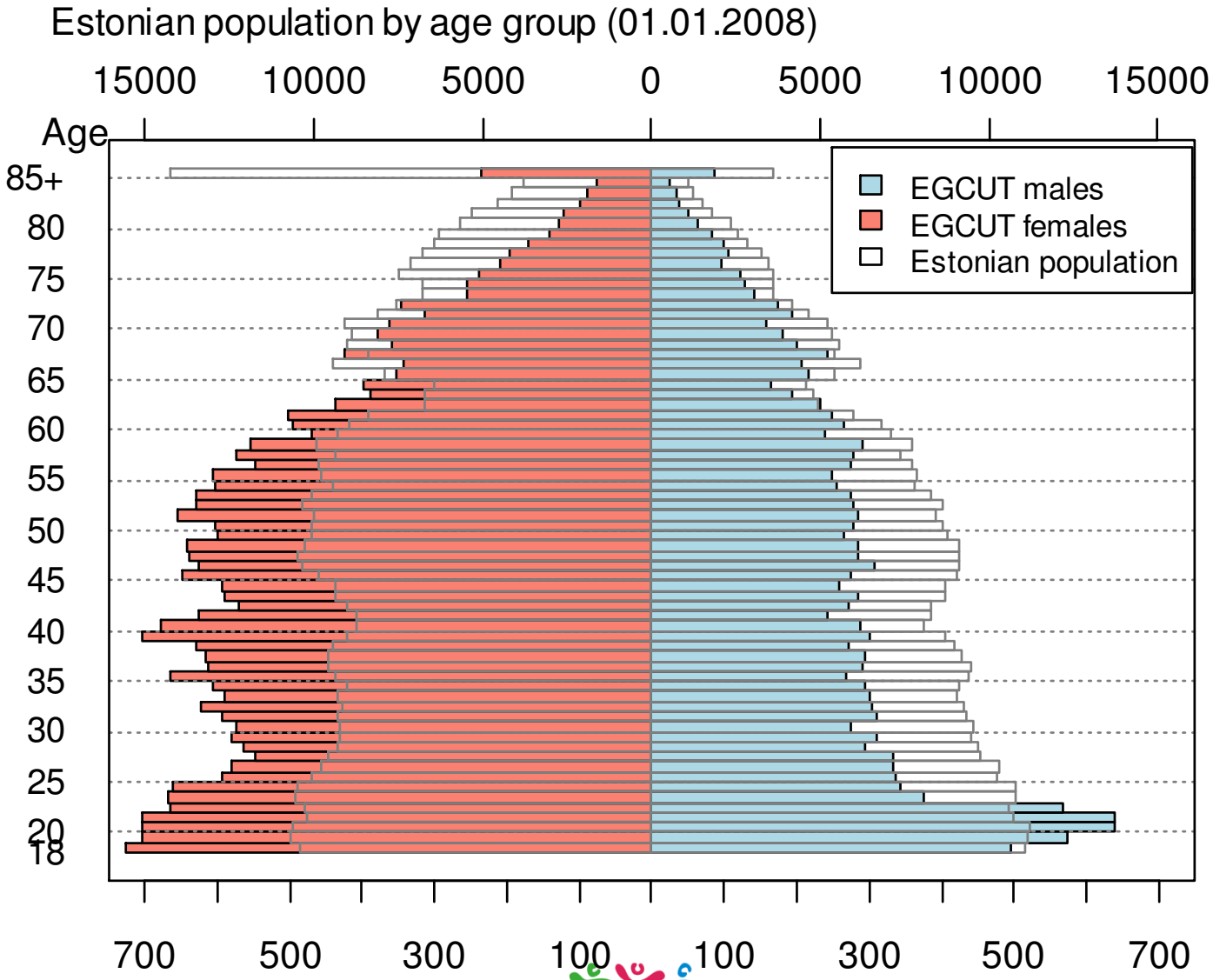
- Sample sizes: 10000+
- Studies on molecular biomarkers, sometimes combined with lifestyle factors
- The effects of genetic predictors are not confounded in the traditional way
- Possible confounding by population structure

Estonian Biobank
Estonian Genome Center, University of Tartu (EGCUT)

- About 52000 Gene Donors (GD) recruited in 2002-2010
- Age at recruitment: 18-103
- Recruited individuals per year:



EGCUT cohort vs Estonian population



estonian genome center
university of tartu

A prospective cohort of 50000+ participants („Gene Donors“)

51795 GD:

- 17795 males, 34000 females
- 81% ethnic Estonians

- About 5% of the Estonian adult population belong to the cohort
- Largest epidemiological cohort in Estonia (and in the Baltic States)
- **Follow-up for mortality, incident diseases, etc via registry and electronic health record linkages: median follow-up time by January 2016: 7.2 years**

estonian genome center
university of tartu

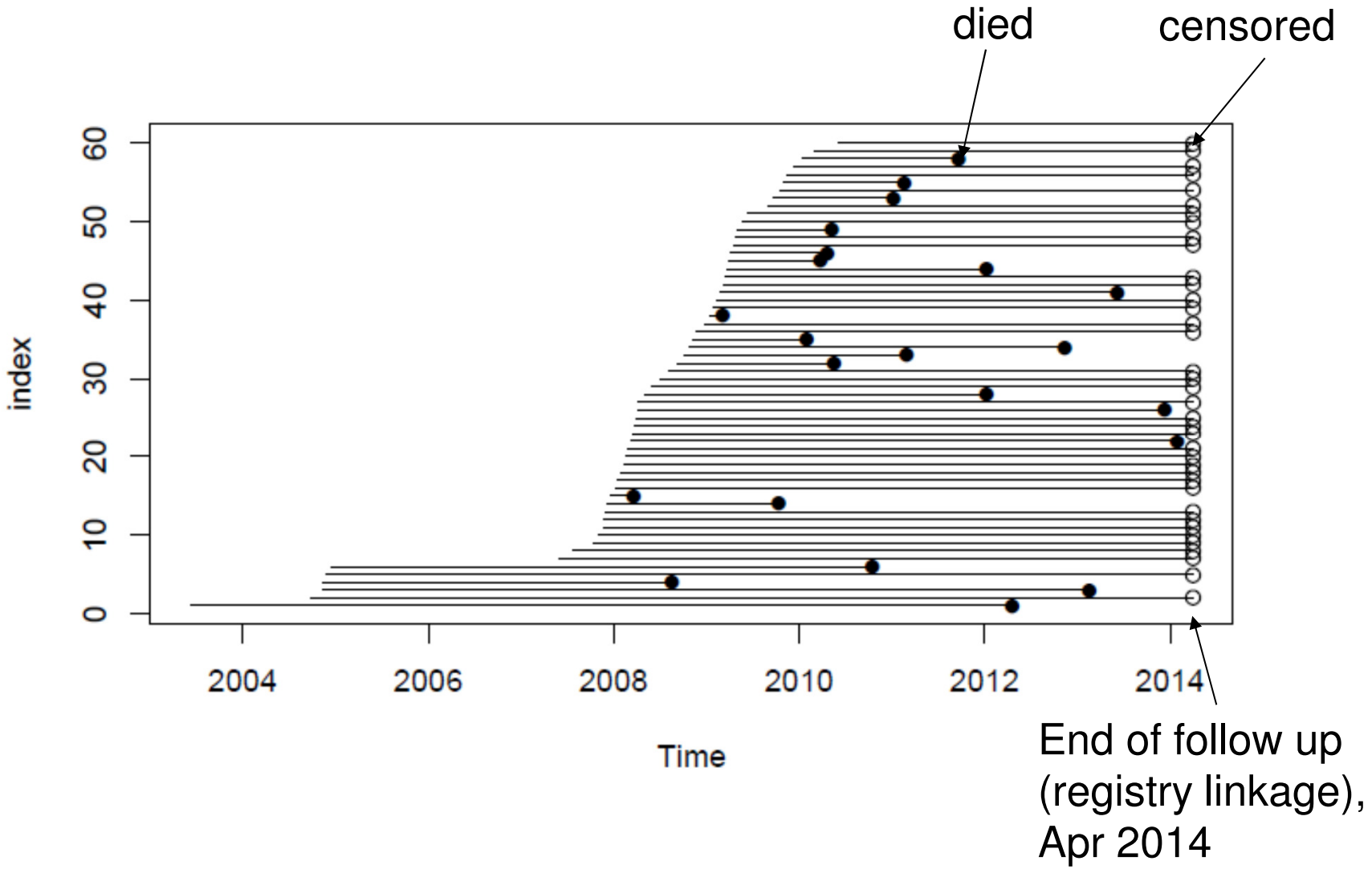
Follow-up studies: statistical aspects

- Follow-up studies usually gather time-to-event data: one is interested in outcome events that occur after recruitment of study subjects.

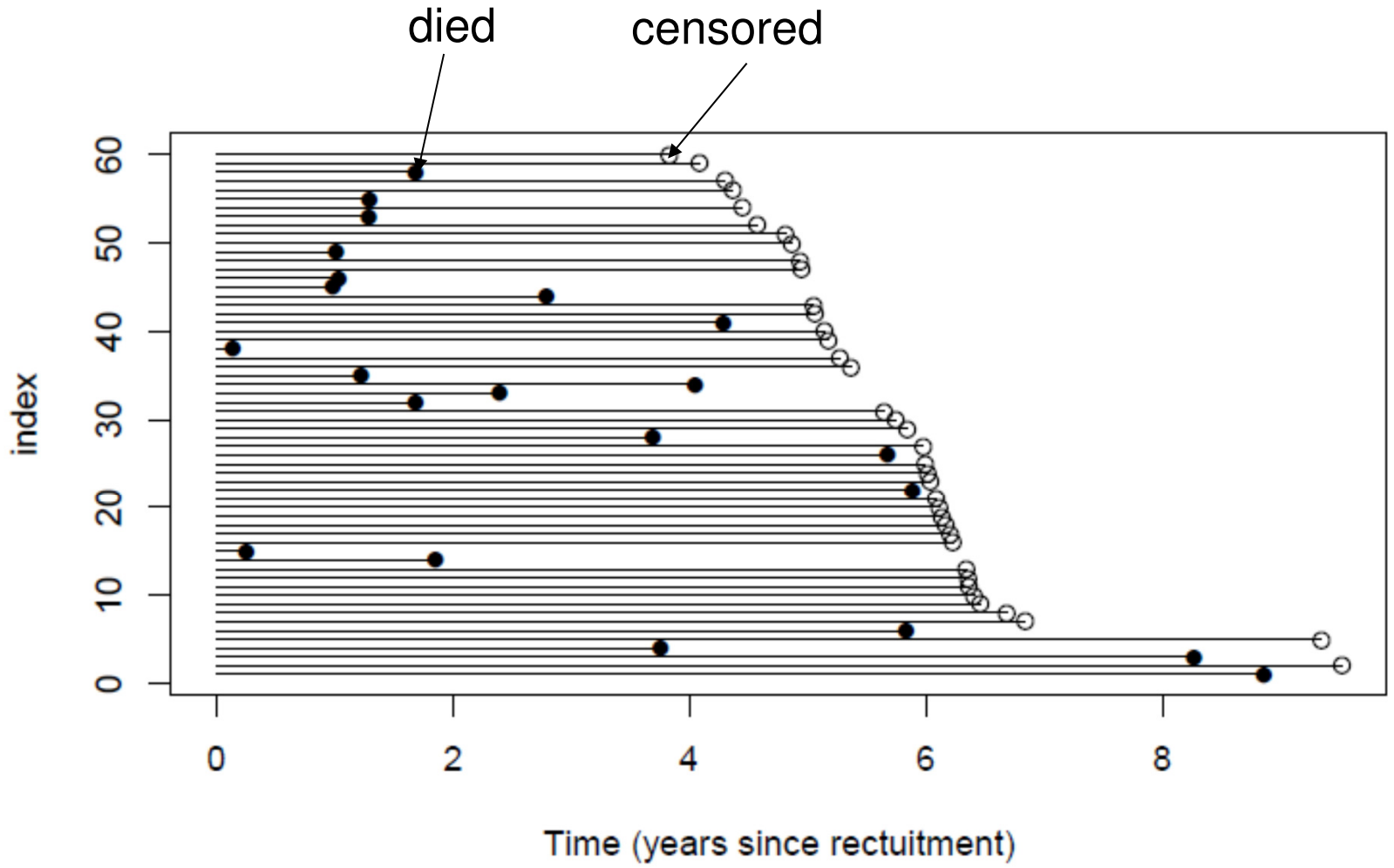
- The analysis is complicated by censoring – by the end of follow-up, the outcome event has only occurred for a subset of participants.

- Analysis depends on the choice of time scale (study time, calendar time, age)

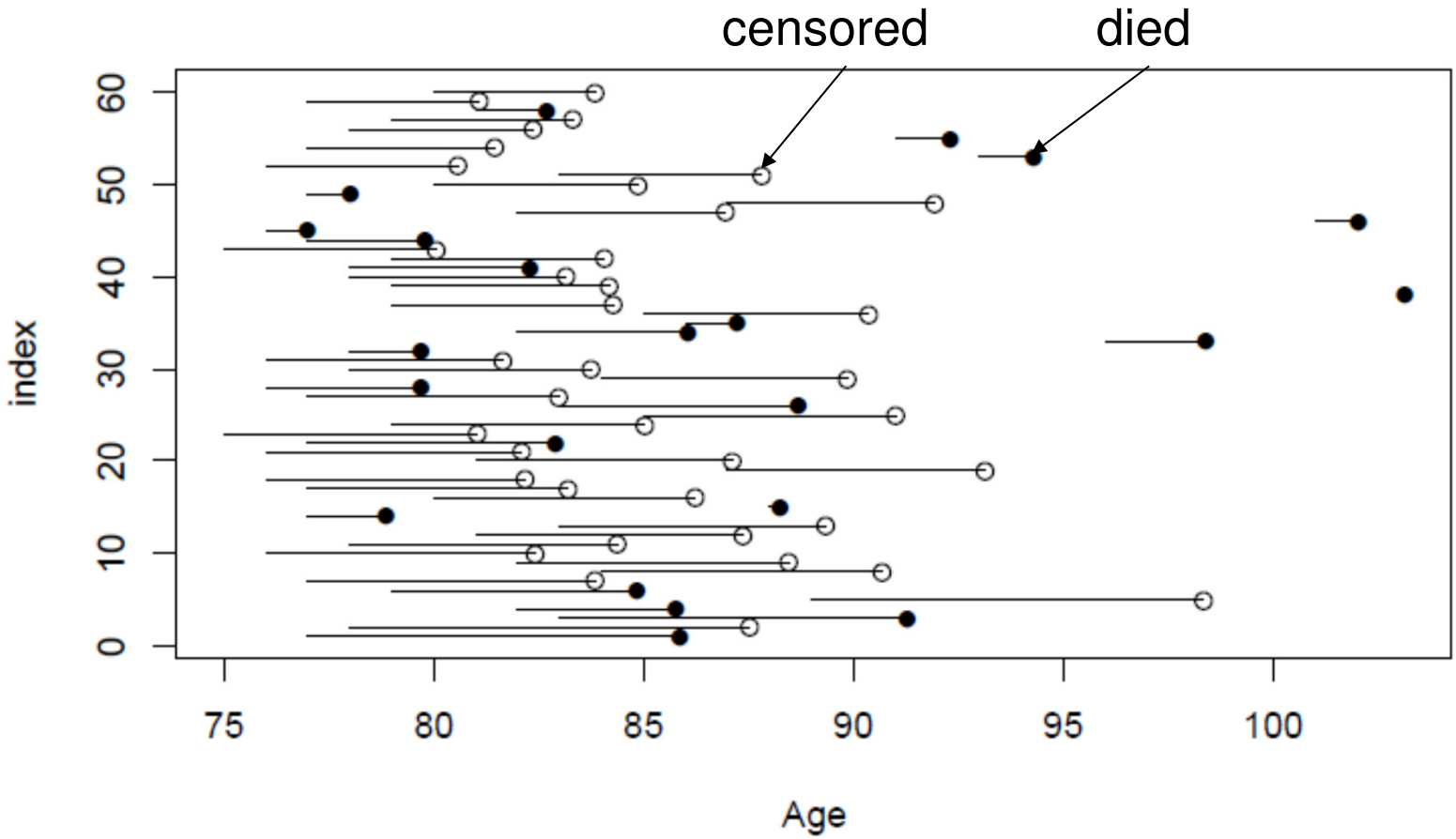
Example: follow-up of 60 individuals of age 75+ at the Estonian Biobank cohort on the calendar time scale



Example: follow-up of 60 individuals of age 75+ at the Estonian Biobank cohort on the study time scale



Example: follow-up of 60 individuals of age 75+ at the Estonian Biobank cohort on the age scale



Genetic predictors for survival/mortality – why needed?

- Better understanding of biological mechanisms
- Causality of risk factors – (Mendelian randomization and other IV approaches)

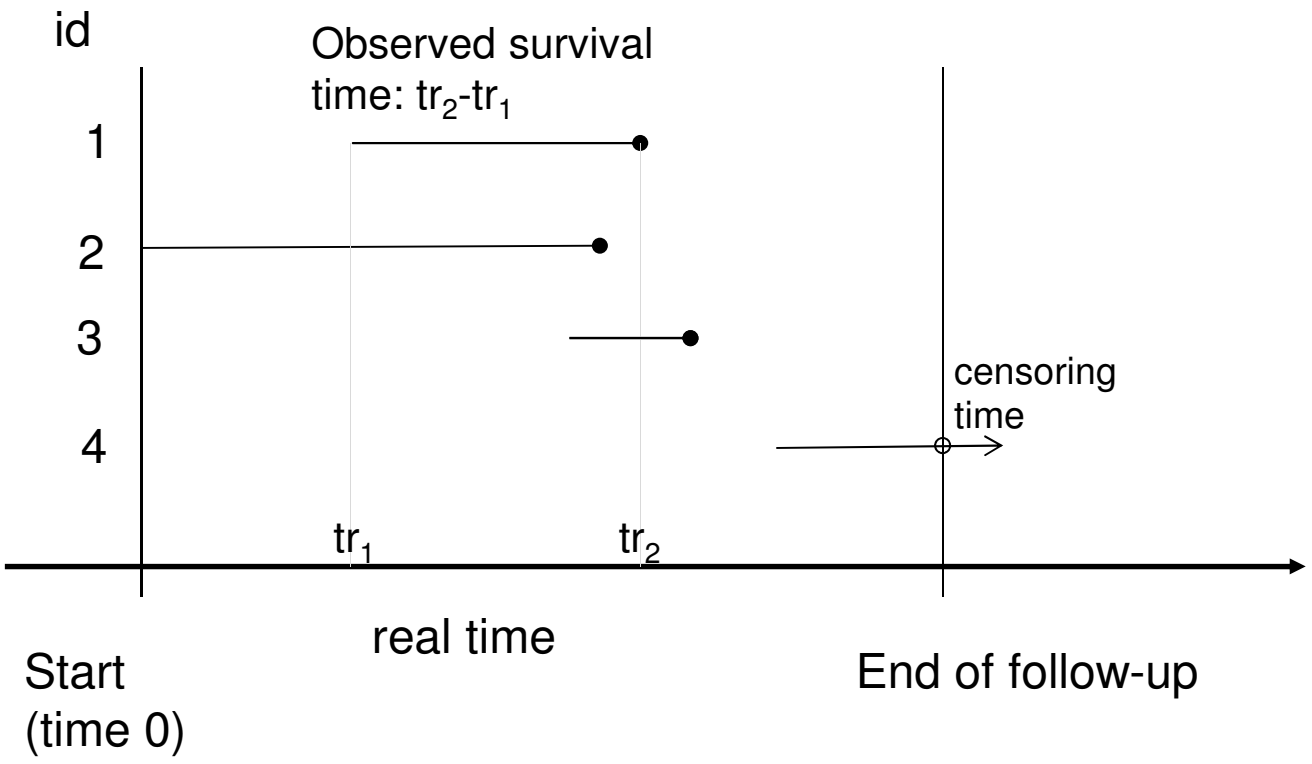
- Lead to more efficient prevention!

Mortality studies in population-based biobank cohorts – sampling and timescales

- Recruitment time is not a clearly defined event in participant's lifecourse.
- Time since recruitment is not a meaningful timescale

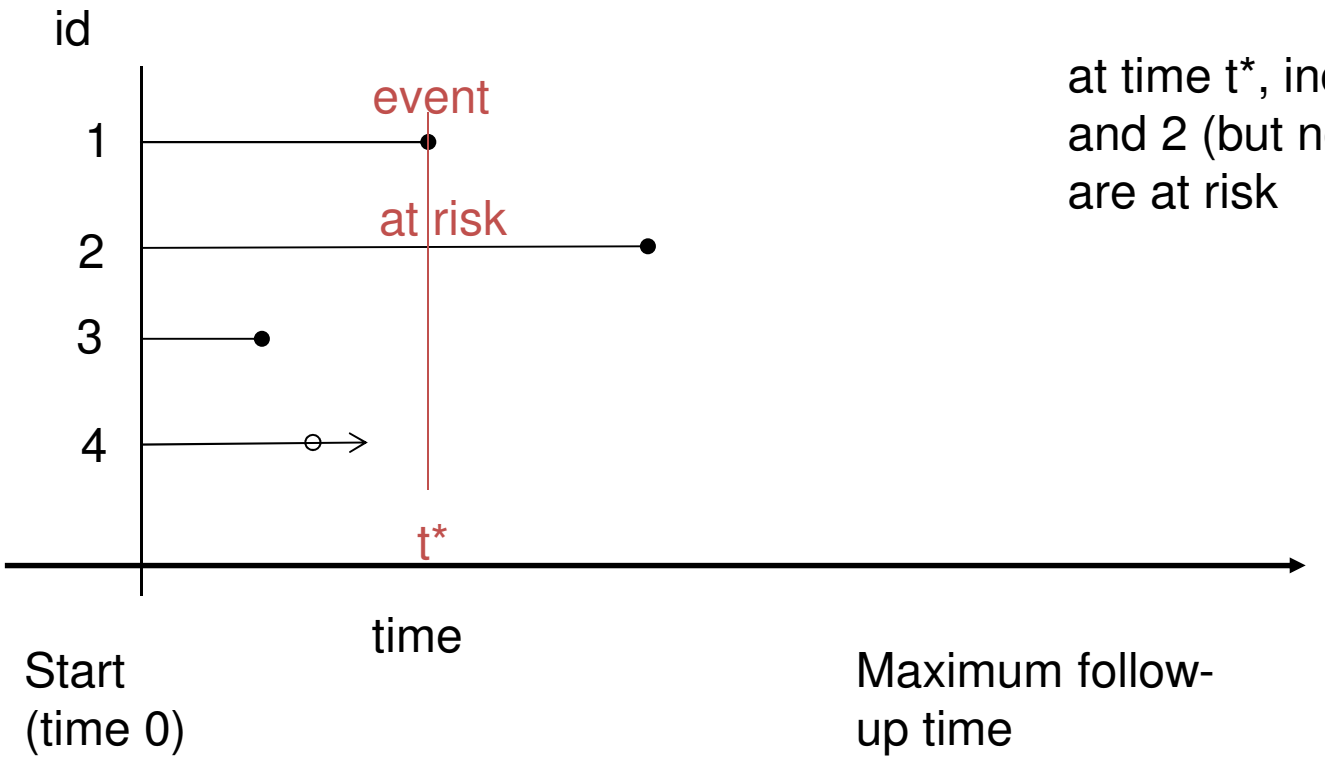
Biobank cohorts

- Usually individuals are recruited at different timepoints in *real time*



Standard survival analysis approach

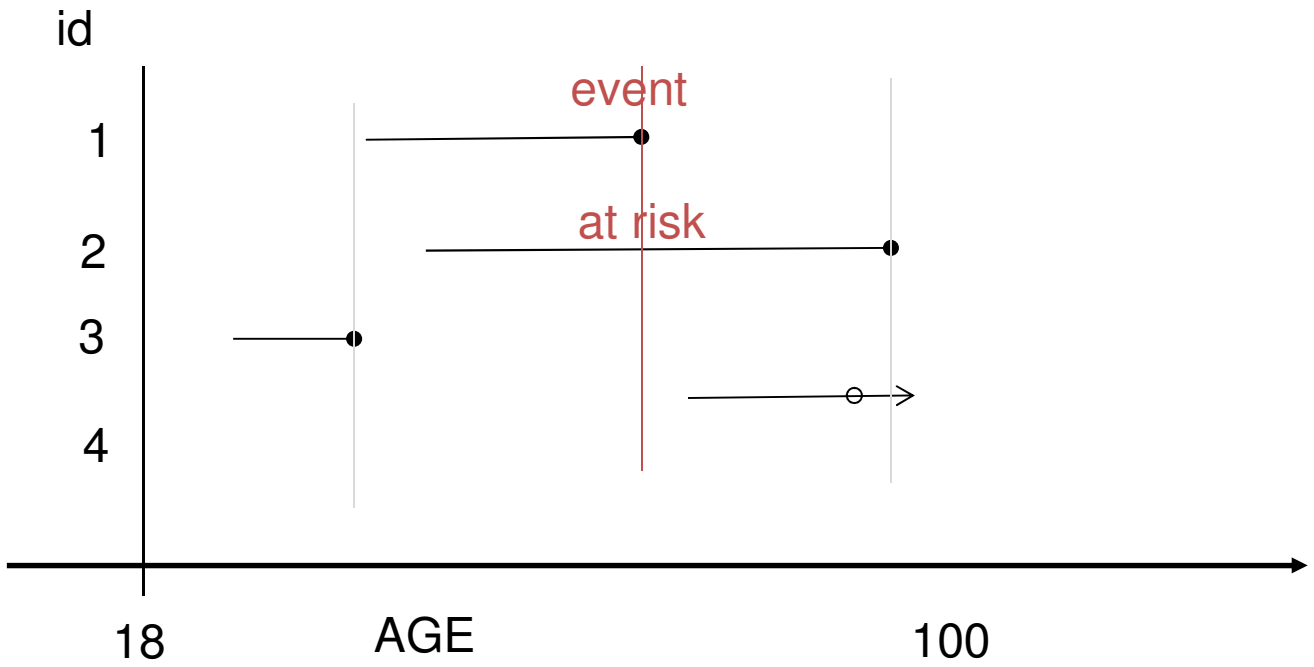
- At each *event time*, compare the individual who had an event with the individuals *at risk*



R: `Surv(time, event)`

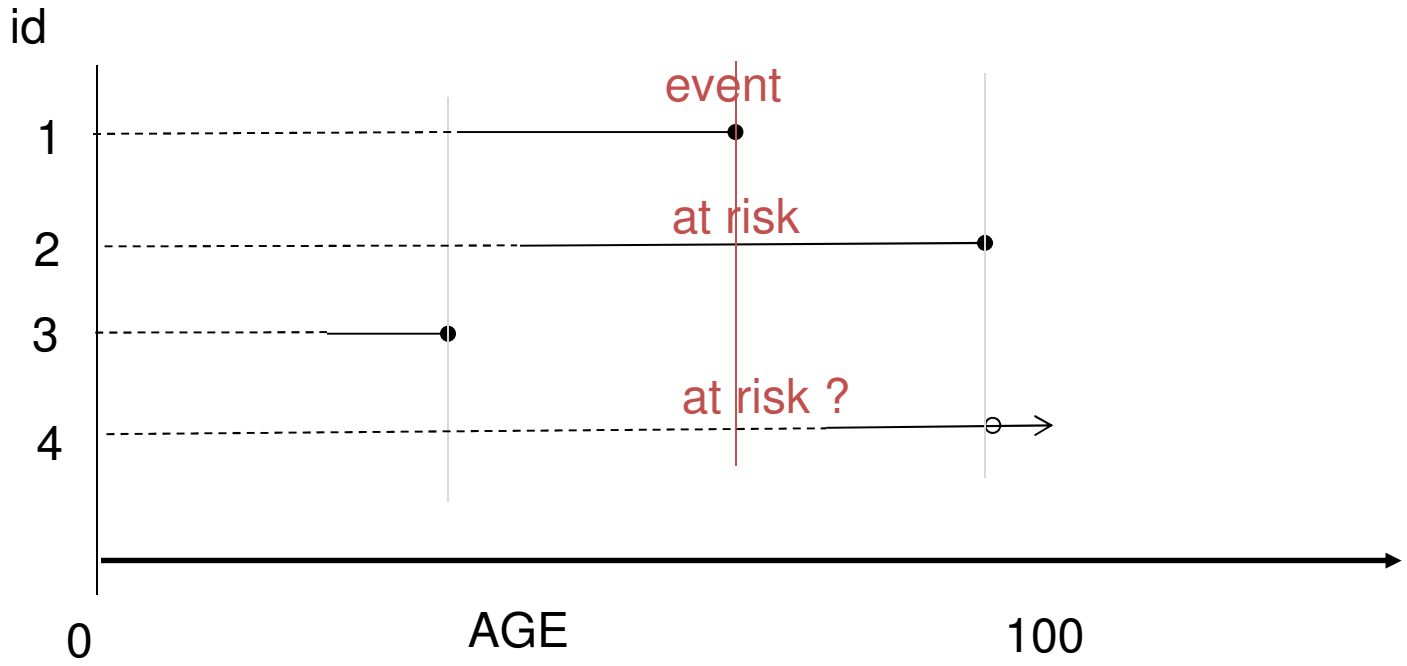
Age as time scale

- At each *event time*, compare the individual who had an event with the individuals who were still at risk while being at the same age



R: `Surv(age_entry, age_exit, event)`

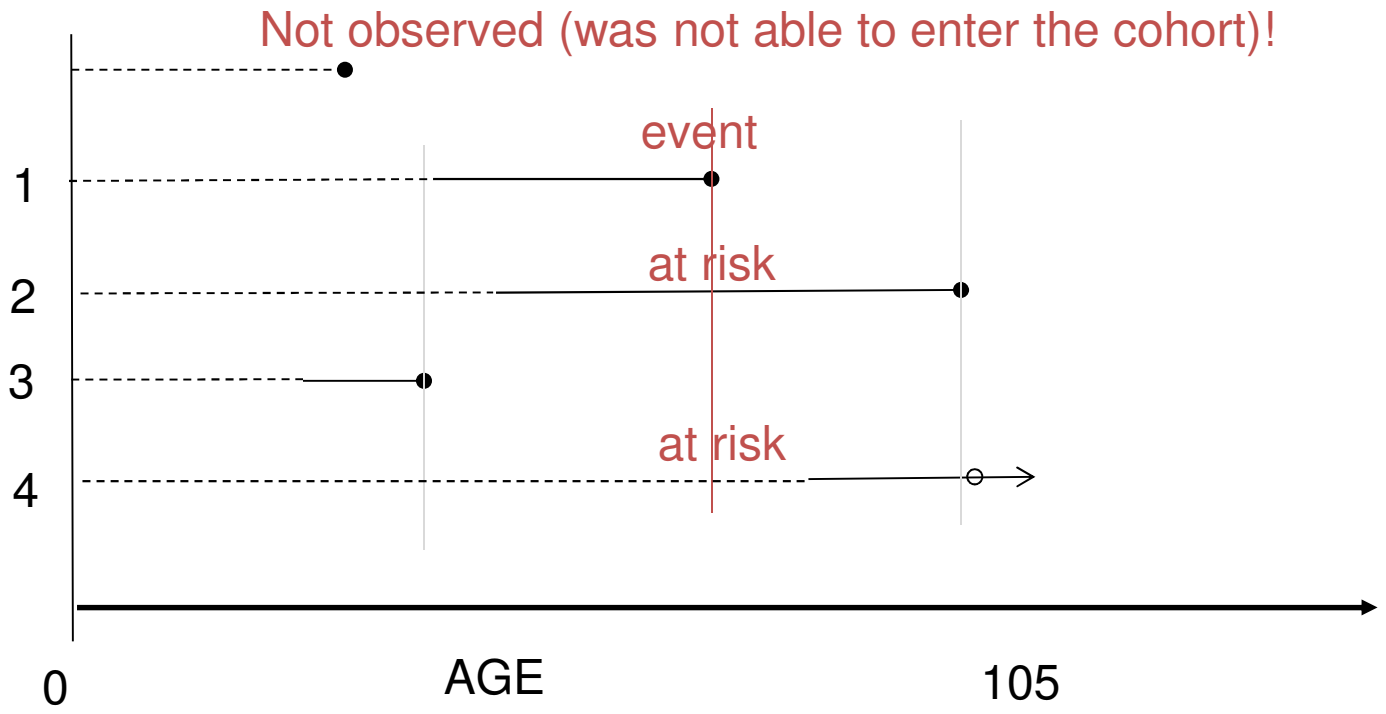
Genetic predictors affect from birth on, should we start the age scale at 0?



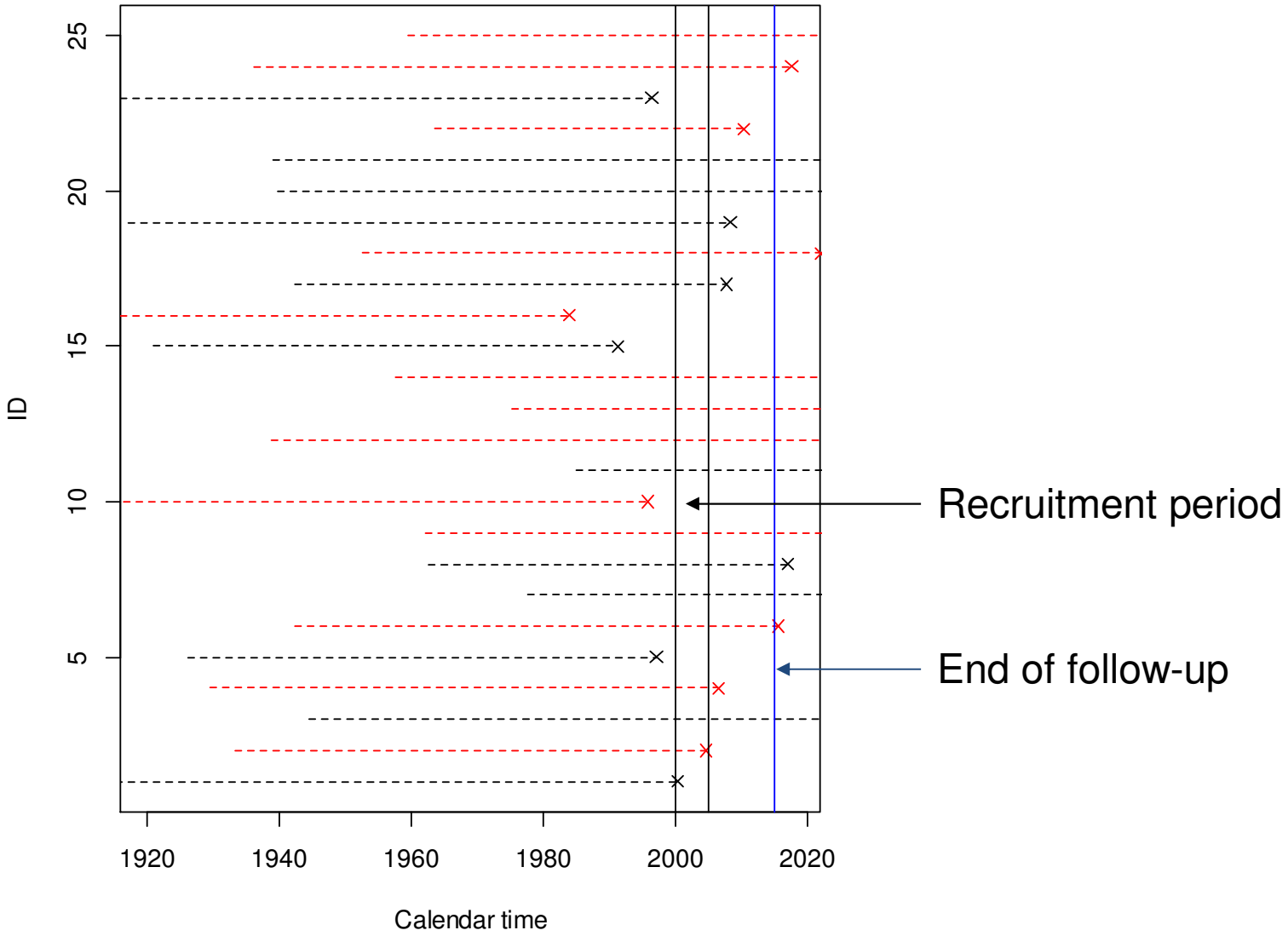
R: `Surv(age_exit, event)`

Genetic predictors affect from birth on, should we start the age scale at 0?

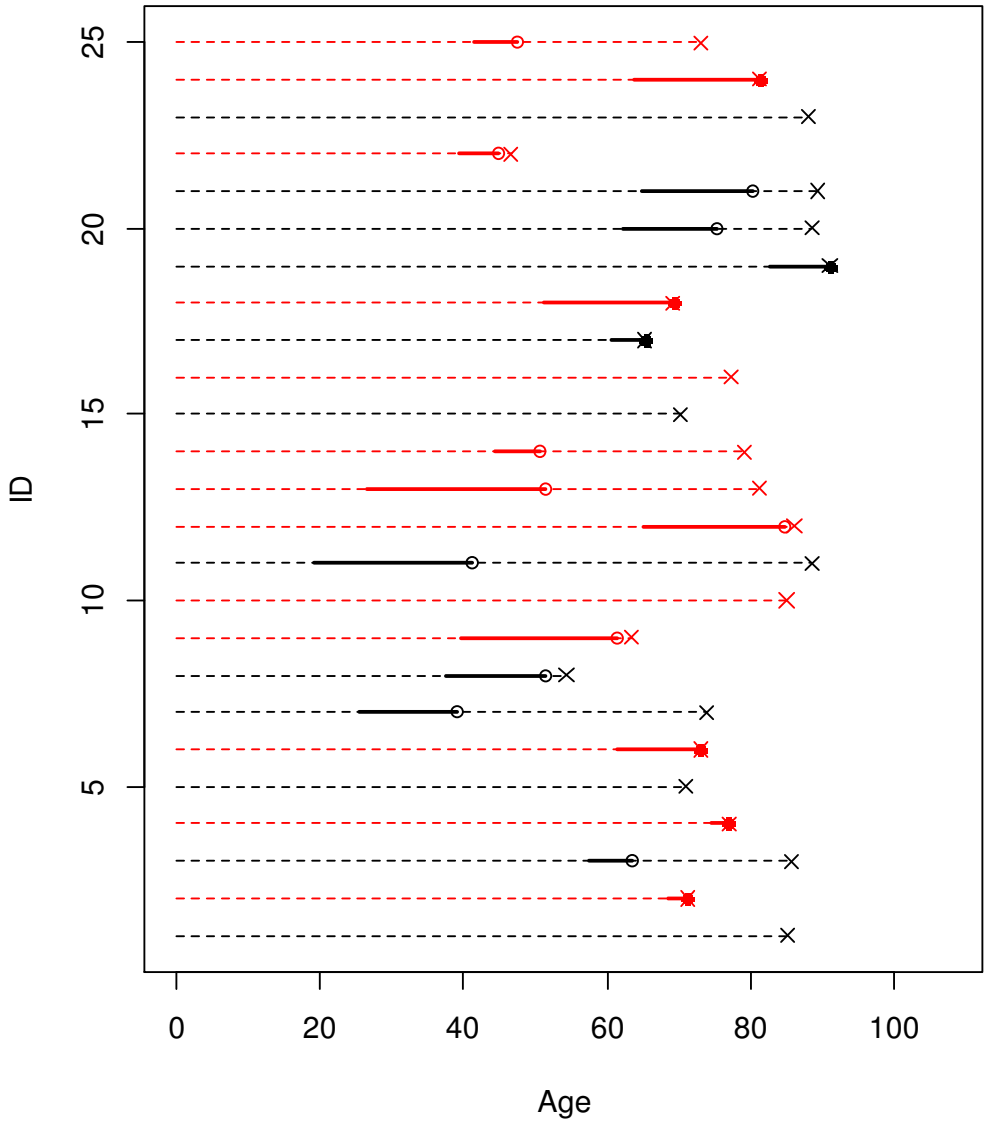
- Beware of left-truncation! (high-risk subjects are less likely to survive until potential recruitment time, if an individual is recruited at old age, he/she is likely to be at low risk)



Biobank recruitment and follow-up



Observed follow-up times on age scale



Most common analysis method: the proportional hazards model

The hazard function:

$$h(t) = \lim_{dt \to 0} P(t \leq T < t + dt | T > t)$$

Interpretation: probability („risk“) of the event occurring at the moment t for the ones at risk at time t.

Hazard for an individual with covariates $X_1, \ldots, X_k$

Baseline hazard (the same for everyone)

Covariate effects

$$h(t|X_1, \ldots, X_k) = h_0(t) e^{X_1\beta_1 + \cdots + X_k\beta_k}$$

$$= h_0(t) \times e^{X_1\beta_1} \times \cdots \times e^{X_k\beta_k}$$

…is a multiplicative model for hazard

Partial likelihood for the Cox model

To estimate the parameters, we need to maximize:

$$L = \prod_{j=1}^{d} \frac{\psi(i)}{\sum_{k \in R(\tau_j)} \psi(k)}$$

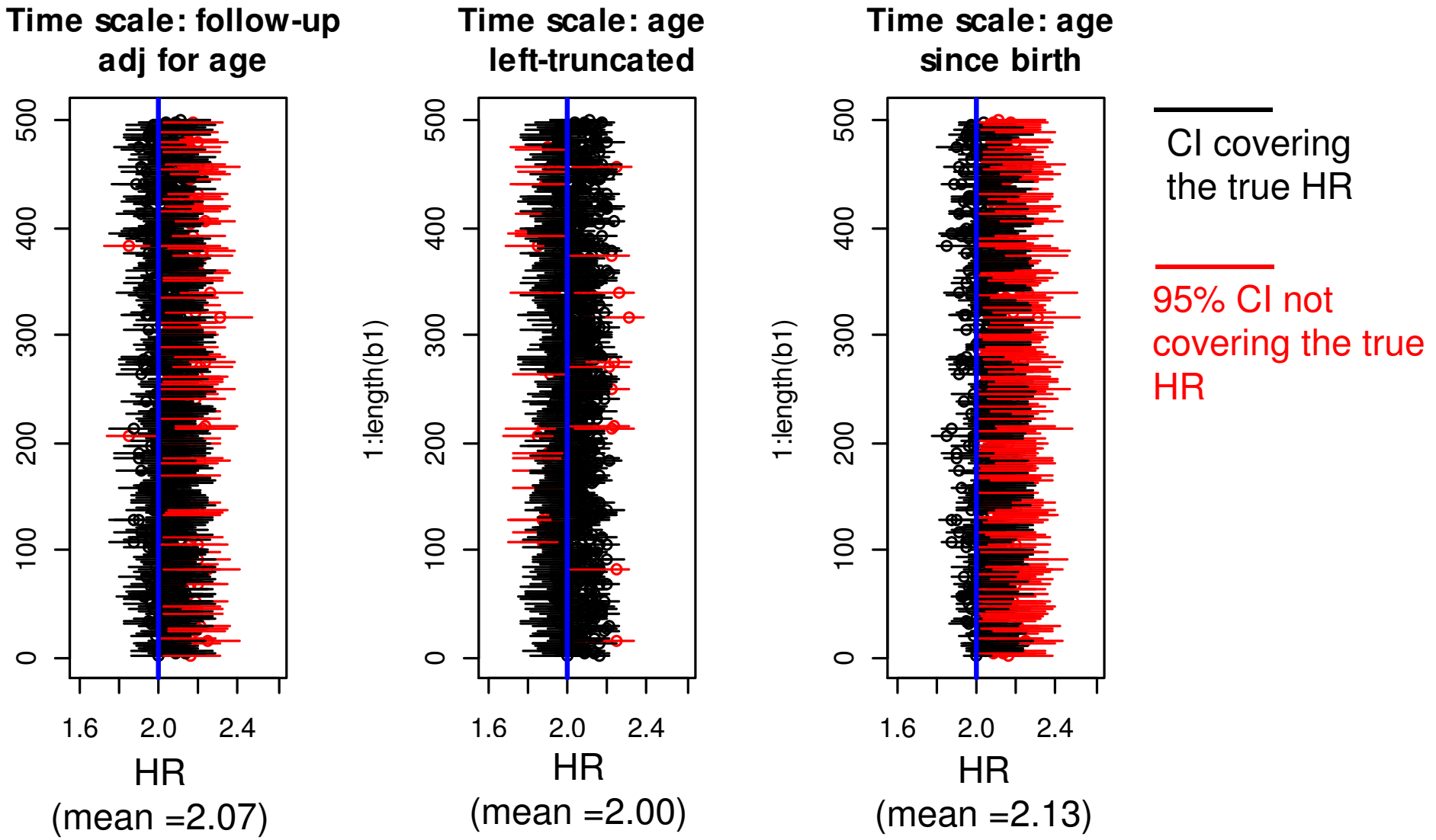
Where ith individual had an event (died) at time τ_j and $R(\tau_j)$ is the set of individuals at risk at time τ_j

And $\psi(i) = e^{X_{1i}\beta_1 + \cdots + X_{ki}\beta_k}$

Thus selecting a different at-risk set $R(\tau_j)$ may result in different estimates

Results of a simulation study (true HR=2)

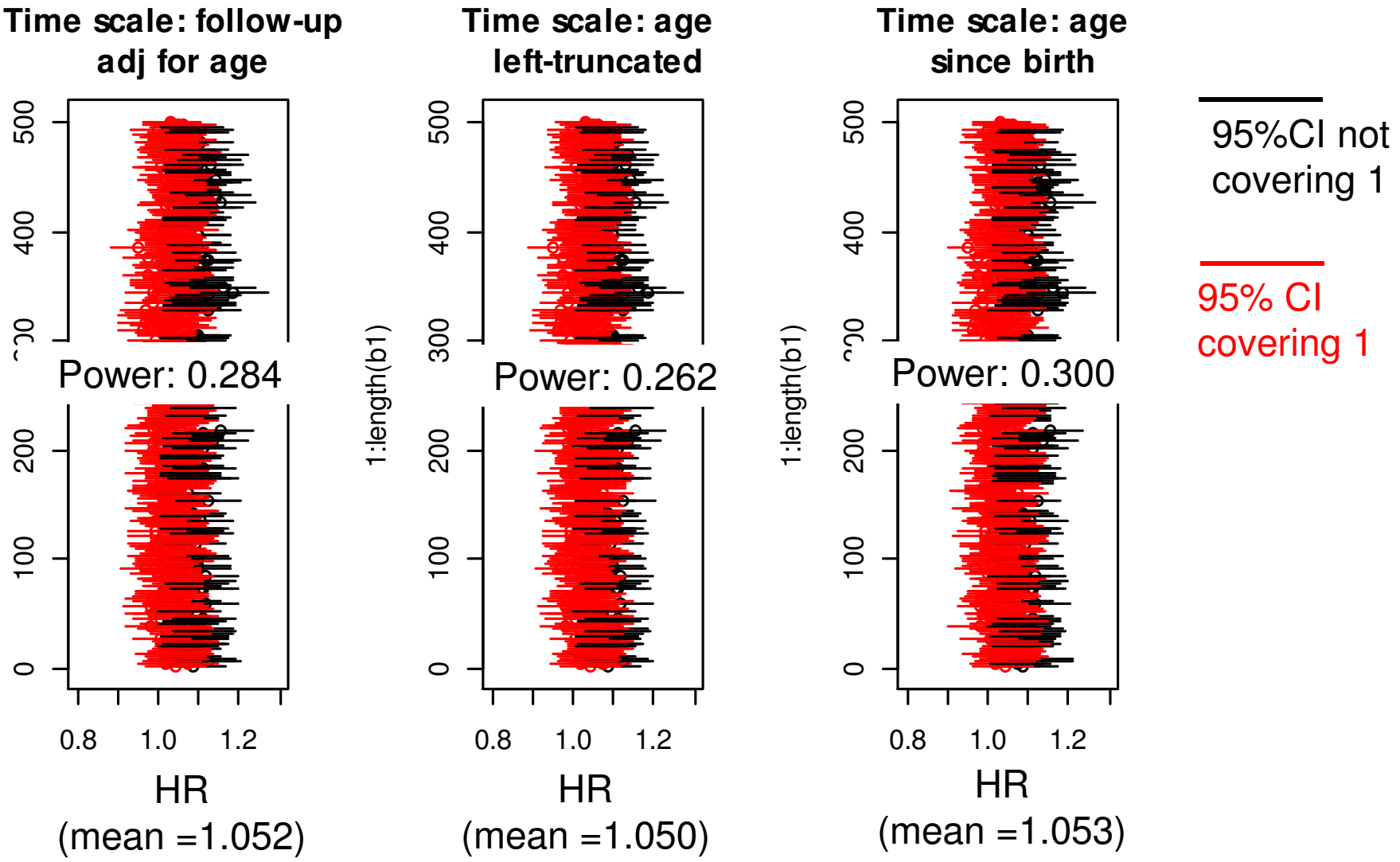
Estimated Hazard Ratios (with 95% CI) using different time scales



Including the ones not yet under follow-up in the risk set, creates upward bias!

But...simulation when HR is small (HR=1.05)

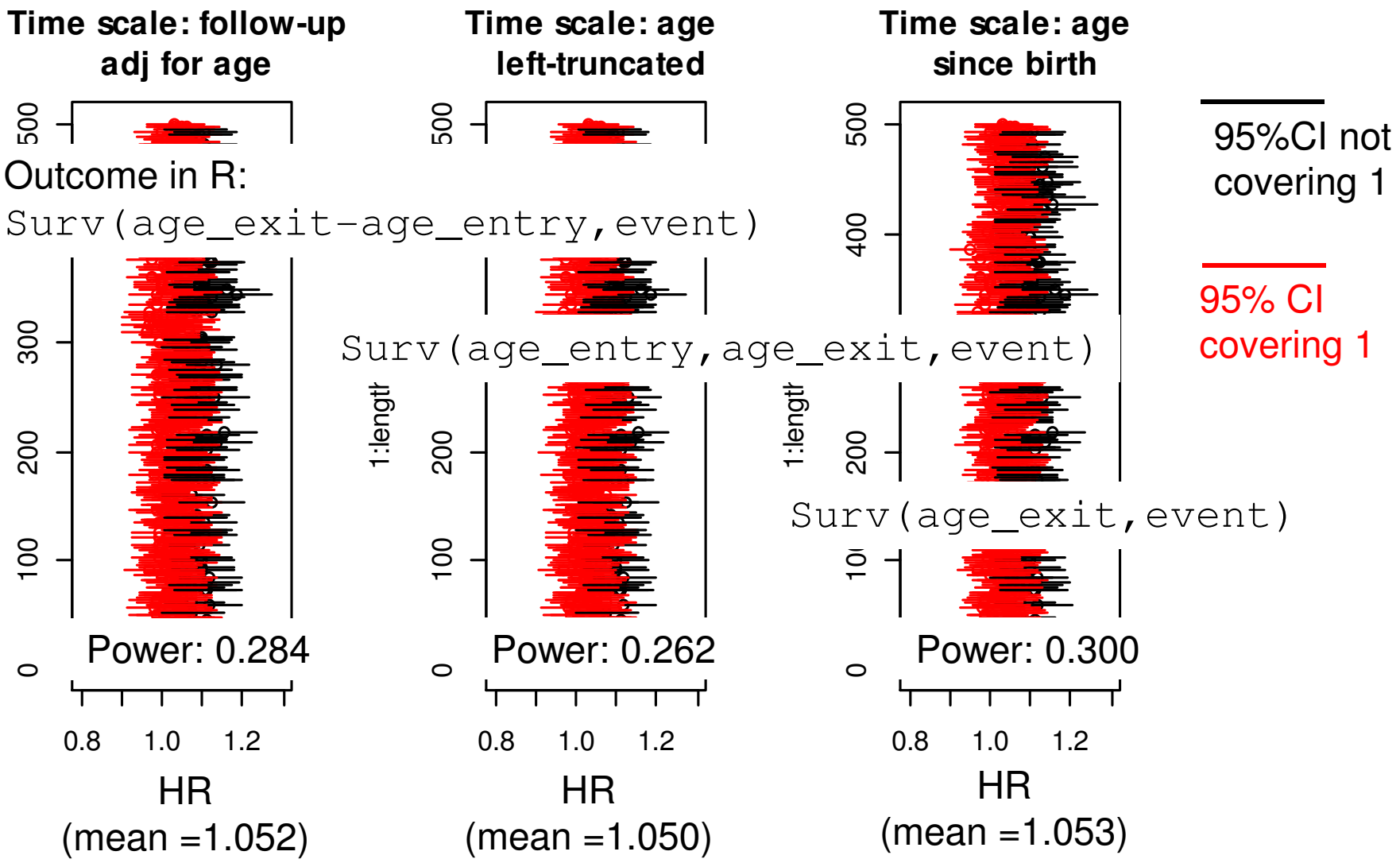
Estimated Hazard Ratios (with 95% CI) using different time scales



Using age since birth as timescale leads to highest power (but small bias)!

But...simulation when HR is small (HR=1.05)

Estimated Hazard Ratios (with 95% CI) using different time scales



Using age since birth as timescale leads to highest power (but small bias)!

Genetic predictors for mortality – more challenges in biobank data

- Biobank cohorts are relatively new
  - Problem: no of cases is often low – heavy censoring!
  - One possible solution: select cases and controls for genotyping, using case-cohort or nested case-control sampling

- Alternative: use data on parental survival

What happens if you use parental data?

Subject's genotype X is coded as the number of effect alleles:

$$X = a1 + a2$$

where a1 is received from one parent and a2 from the other

Parental genotypes X1 and X2 consist of the allele transferred to the child and another allele that is not transferred:

$X1 = a1 + b1$ $\qquad\qquad$ $X2 = a2 + b2$

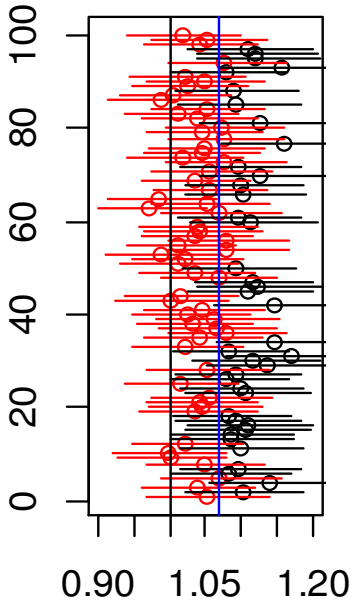
We use parental survival time (T1 or T2) and X (instead of X1 or X2) as a covariate!

Thus $cor(X,X1) = cor(X,X2) = 0.5$ and the parameter estimates are about half of the original parameters.

(Estimated HR = approx. square root of the original HR)

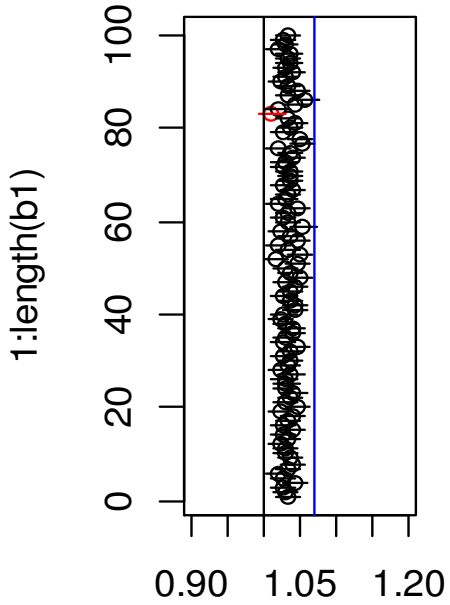
Some simulations:

Individual survival (1.7% have died)



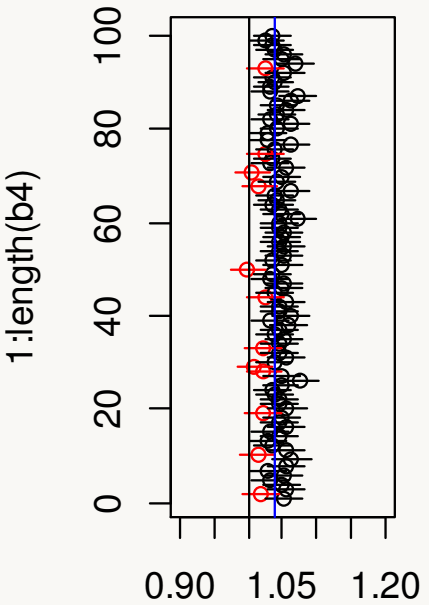
Power: 38%

Parental survival (35% have died)



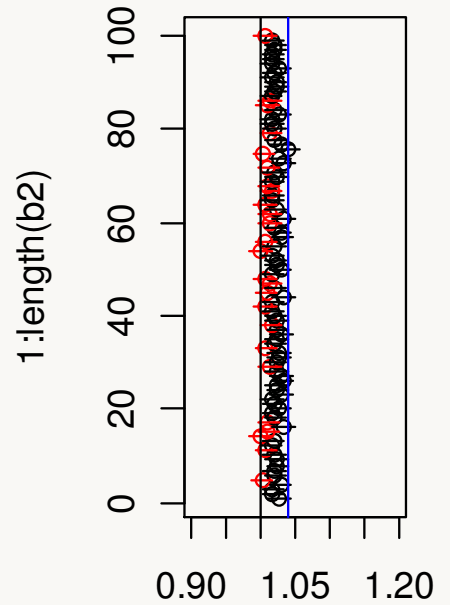
Power: 98%

Individual survival (18% have died)



Power: 88%

Parental survival (50% have died)



Power: 75%

Using parental survival will lead to better power if the parental event rate is at least 4 times higher than individual event rate!

Genetic predictors for mortality – methodological approaches?

Genome-wide Association Study (GWAS) for overall survival?

- Standard approach: Cox proportional hazards regression

- Complication: the algorithm is slow (e.g for a dataset of size >30000 individuals x 30000000 SNPs)

A two-step Cox modeling approach

- Fit a Cox model with non-genetic preditors

$$h(x) = h_0(x)e^{\gamma_1 Z_1 + \cdots + \gamma_k Z_k},$$

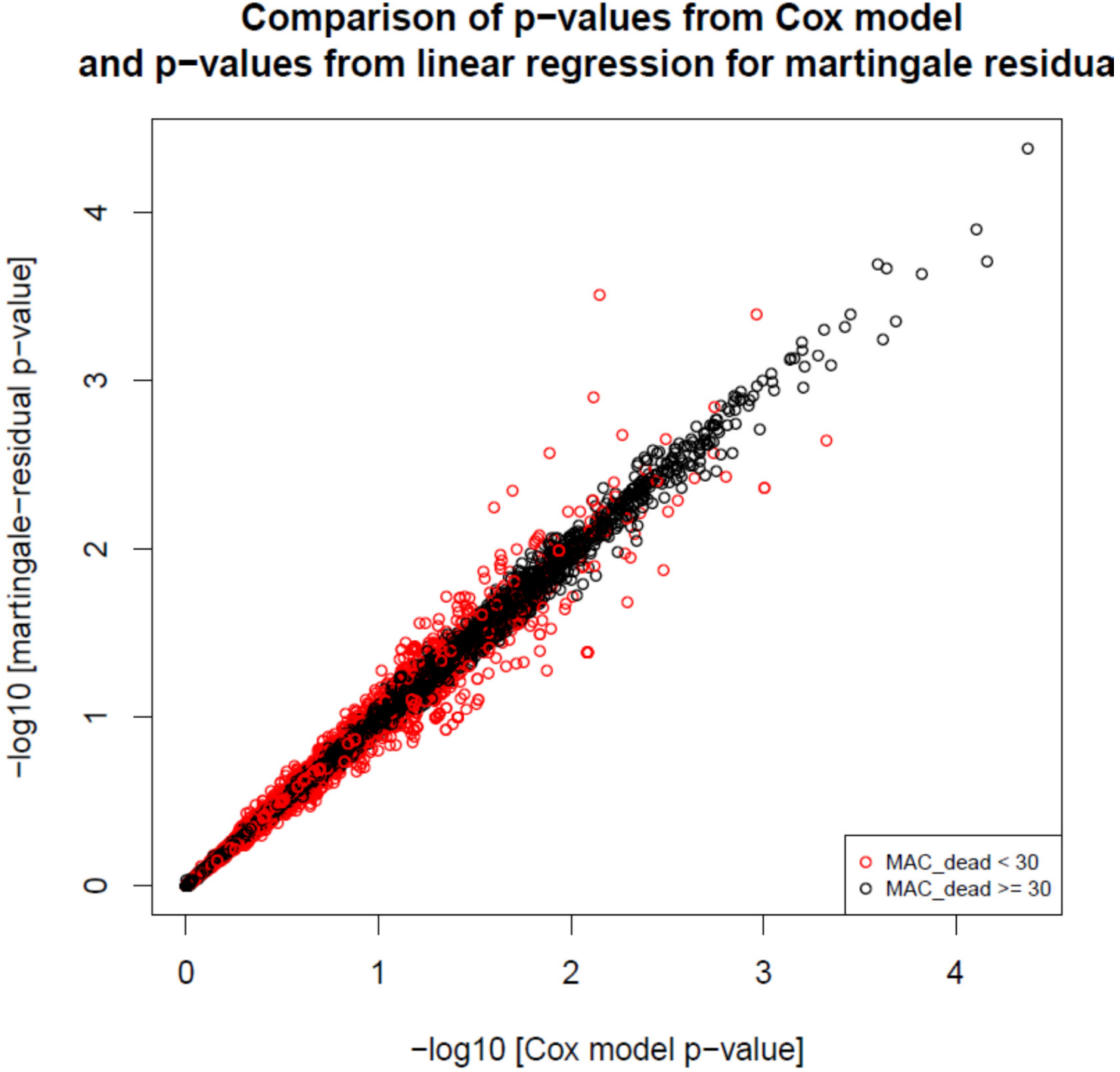
- Calculate Martingale residuals:

$$\widehat{M}_i = \delta_i - \widehat{\Lambda}_0(\tau_i)e^{\widehat{\gamma}_1 Z_1 + \cdots + \widehat{\gamma}_k Z_k}$$

With δ_i - censoring indicator, $\widehat{\Lambda}_0(\tau_i)$ - estimated baseline cumulative hazard.

Martingale residuals are linearly associated with omitted covariates, thus…

- Run a linear regression GWAS on the martingale residuals to identify associated SNPs


Comparison of p−values from Cox model
and p−values from linear regression for martingale residua
−log10 [martingale−residual p−value]
0
1
2
3
4
−log10 [Cox model p−value]
0
1
2
3
4
MAC_dead < 30
MAC_dead >= 30

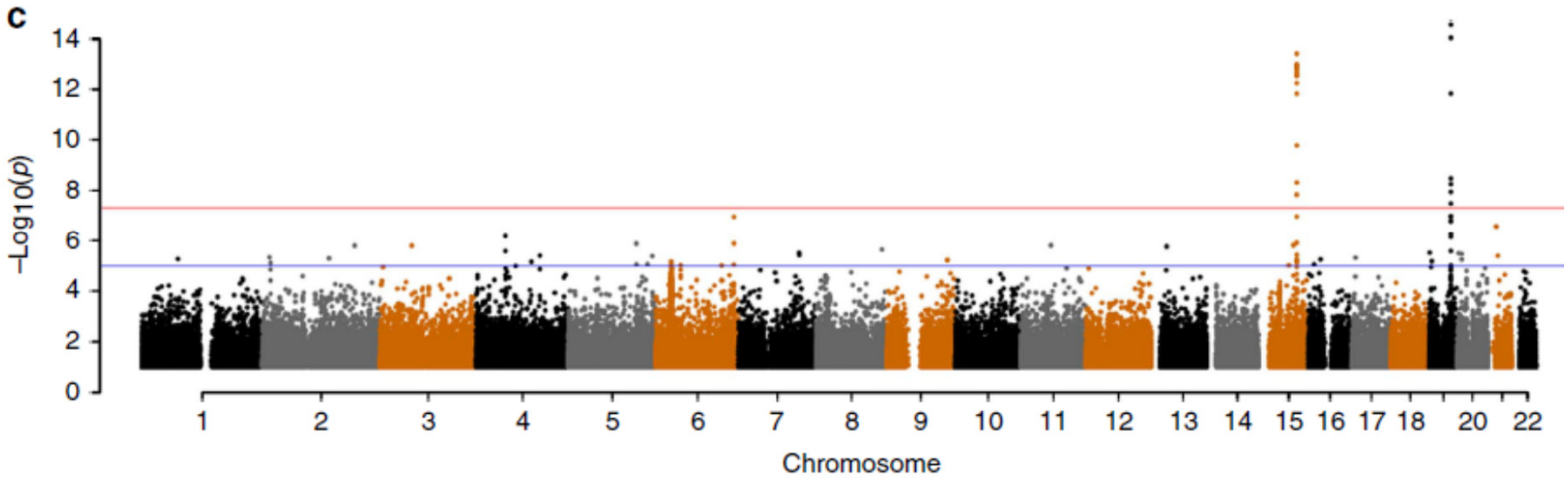
Some results…

Received 28 Sep 2015 | Accepted 29 Feb 2016 | Published 31 Mar 2016

DOI: 10.1038/ncomms11174 OPEN

Variants near *CHRNA3/5* and *APOE* have age- and sex-related effects on human lifespan

Peter K. Joshi[1], Krista Fischer[2], Katharina E. Schraut[1,3], Harry Campbell[1], Tõnu Esko[2,4,5,6] & James F. Wilson[1,7]



>270000 parental lifespans from the UK Biobank…

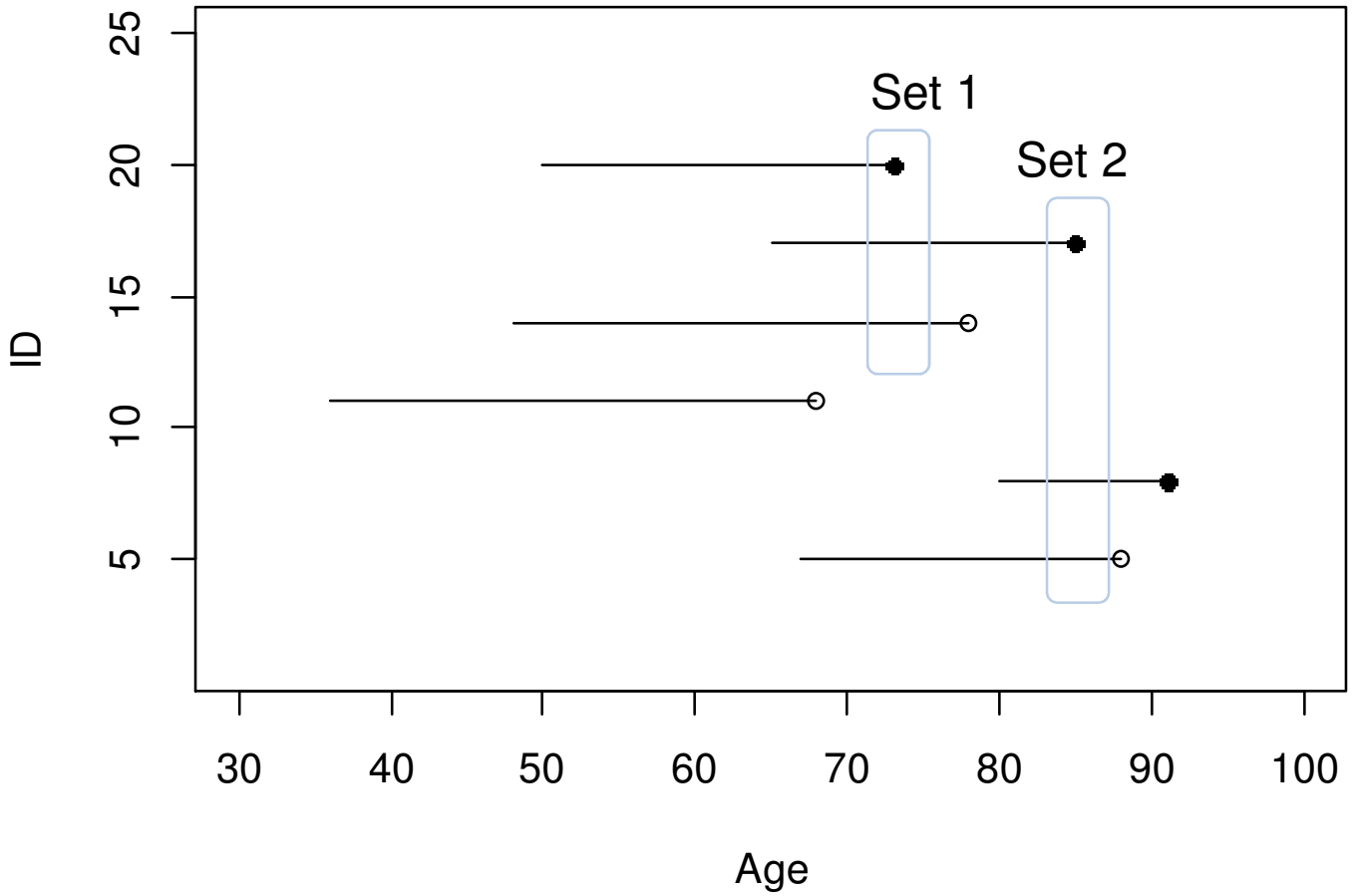
How to handle power issues? (low no of cases)

- Use parental lifespans

- Select samples for genotyping, using a case-control strategy
  - Nested case-control design
  - Simple case-control design

Nested case-control design

- For each case, select 2-4 controls that are under follow-up at the event time (according to chosen time scale), analyse using conditional logistic regression



Example of the Estonian Biobank analysis: effects on overall mortality

Full-cohort model (n=51621, 3355 events)

Variable	beta	Se	P-value
Years smoking	0.018	0.0011	$1*10^{-57}$
Educ: secondary	-0.40	0.037	$2*10^{-28}$
BMI>35	0.23	0.057	$4*10^{-5}$
T2D	0.43	0.048	$1*10^{-19}$
CAD	0.22	0.039	$1*10^{-8}$

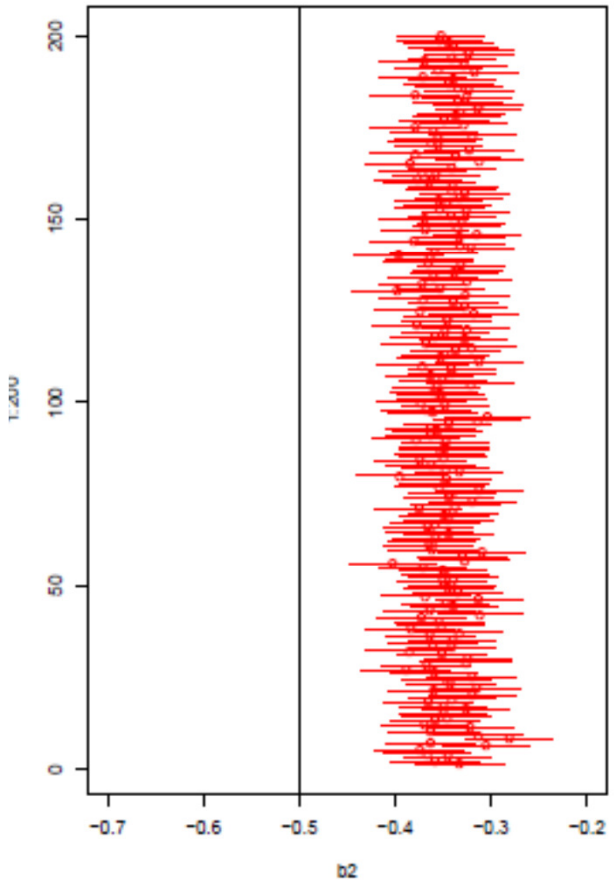
Nested CC-analysis (n=9317, 3351 events, 3 controls per case)

Variable	beta	se	P-value
Years smoking	0.019	0.0014	$2*10^{-42}$
Educ: secondary	-0.38	0.045	$2*10^{-19}$
BMI>35	0.21	0.069	$3*10^{-3}$
T2D	0.46	0.059	$2*10^{-14}$
CAD	0.22	0.047	$3*10^{-6}$

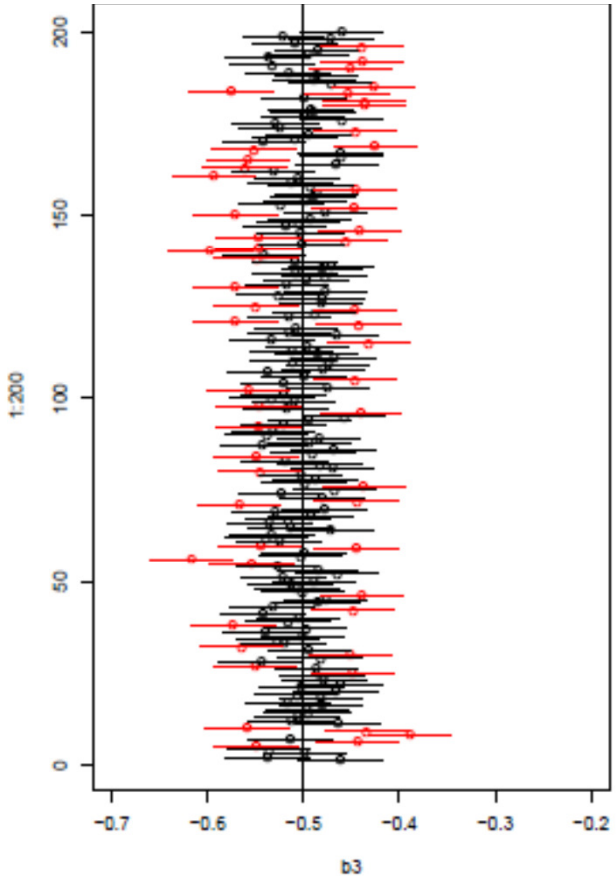
Often cases are over-sampled, but this is not a nested case-control design

- Use sampling weights! (Package „survey“ in R, for instance) – a simulation study

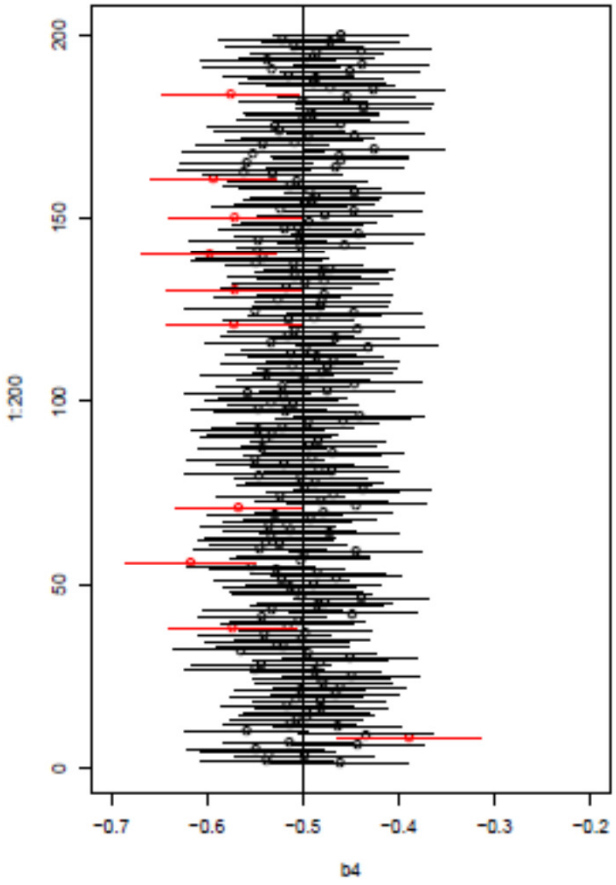
Unweighted



Frequency weights
coxph(…,weights=1/p)



Sampling weights
svycoxph(..) (survey)

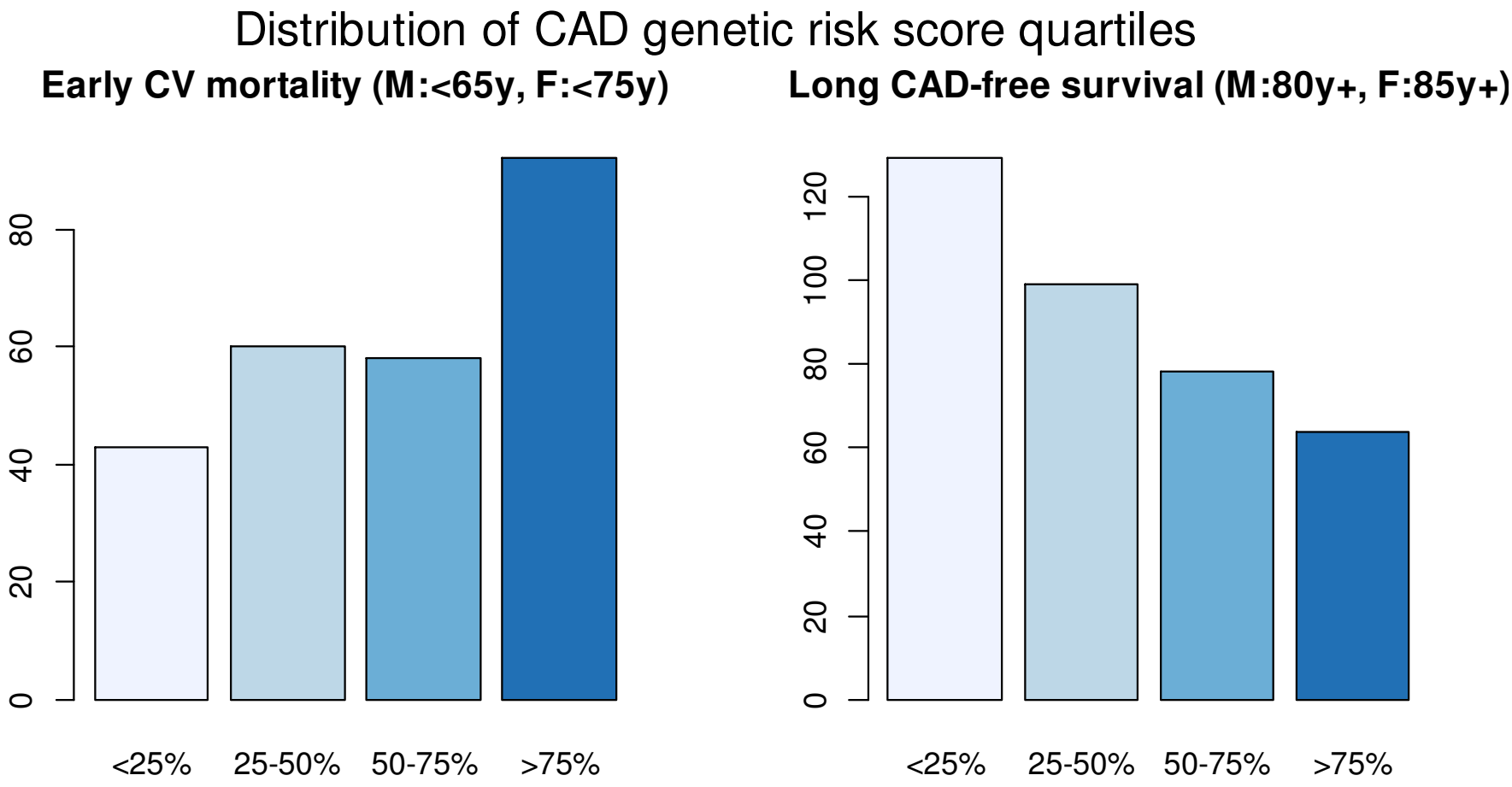


Other aspects to consider

- Proportionality of hazards
  - The main assumption of the Cox model
  - May be violated, if agespan in the cohort is wide
  - Right timescale choice may help

- Are we actually interested in the proportional hazards model?

Extreme cases and controls: 253 individuals who died early of cardiovascular causes vs 370 with long survival, free of coronary artery disease (CAD).

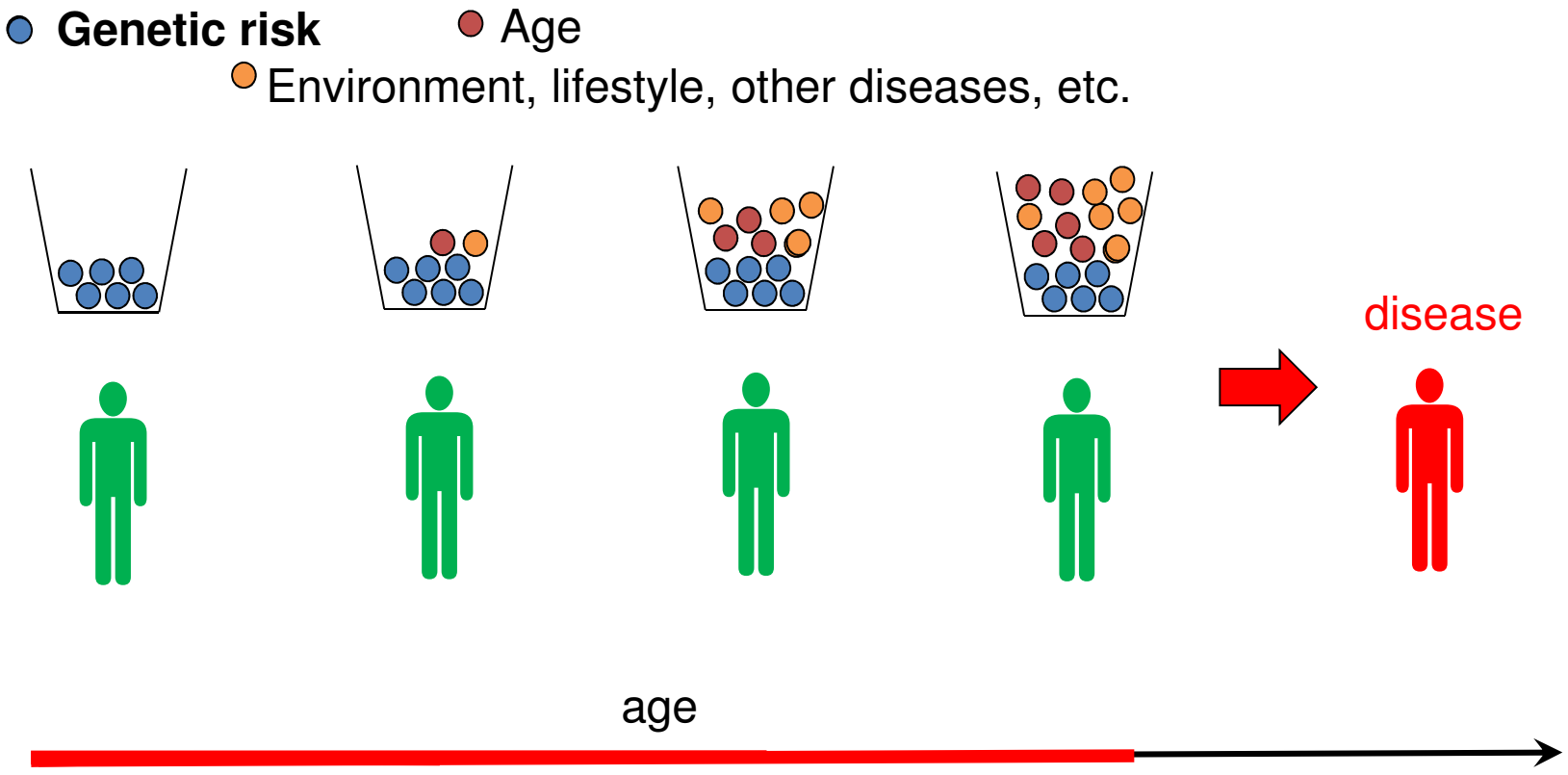


Part II
Genetic (polygenic) risk scores

- Most complex diseases are polygenic – there is a need for risk predictions that combine the effects of many variables

Why is genetic risk important?

Different risk factors contributing to the individual risk of a disease.



How to measure genetic risk?

Genome-wide association studies (GWAS)

Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes

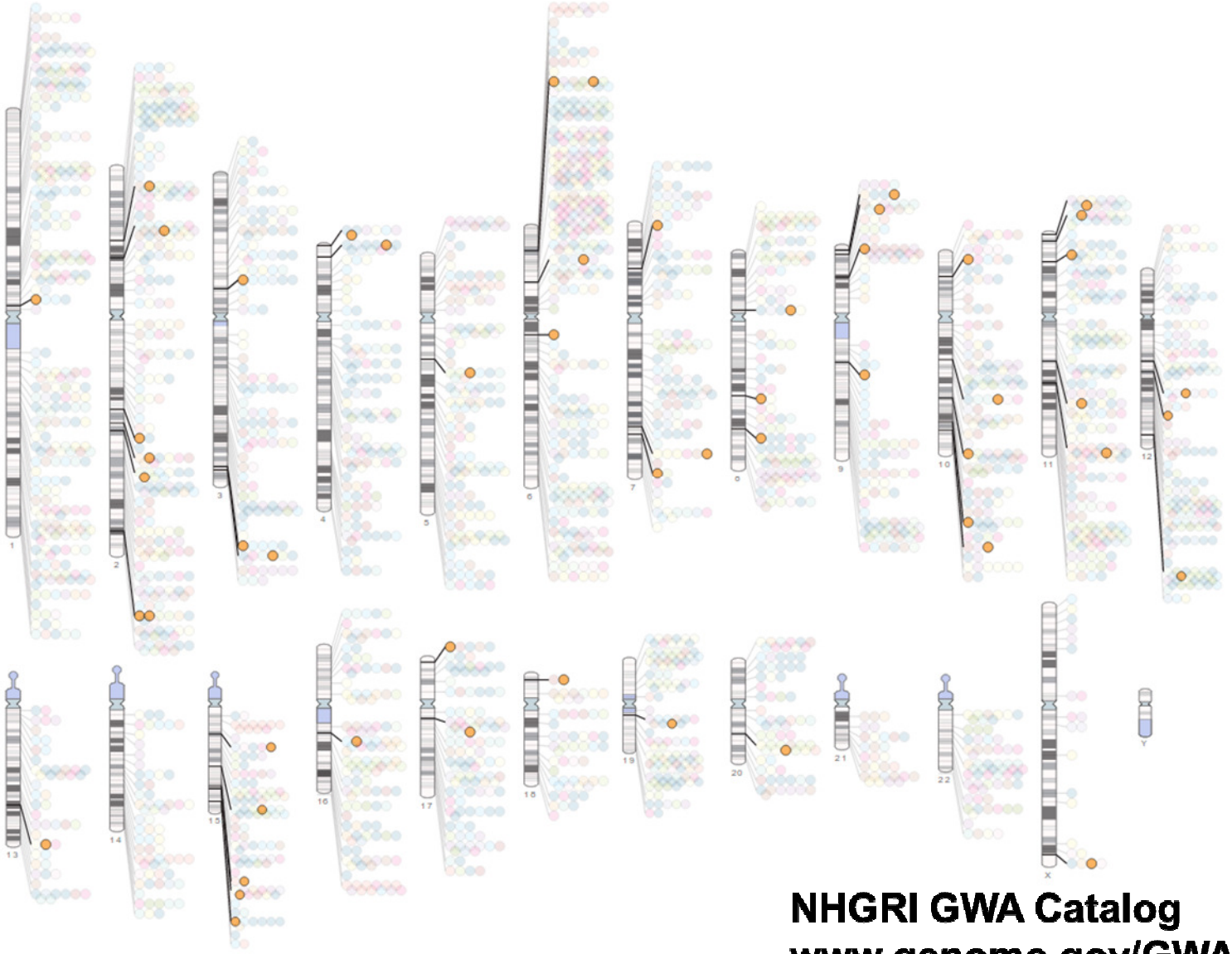
To extend understanding of the genetic architecture and molecular basis of type 2 diabetes (T2D), we conducted a meta-analysis of genetic variants on the Metabochip, including 34,840 cases and 114,981 controls, overwhelmingly of European descent. We identified ten previously unreported T2D susceptibility loci, including two showing sex-differentiated association. Genome-wide analyses of these data are consistent with a long tail of additional common variant loci explaining much of the variation in susceptibility to T2D. Exploration of the enlarged set of susceptibility loci implicates several processes, including CREBBP-related transcription, adipocytokine signaling and cell cycle regulation, in diabetes pathogenesis.

NATURE GENETICS VOLUME 44 | NUMBER 9 | SEPTEMBER 2012

By: A. Morris et al.

ecture and molecular basis of type 2 diabetes (T2D), we conducted a meta-analysis
ing 34,840 cases and 114,981 controls, overwhelmingly of European descent. We

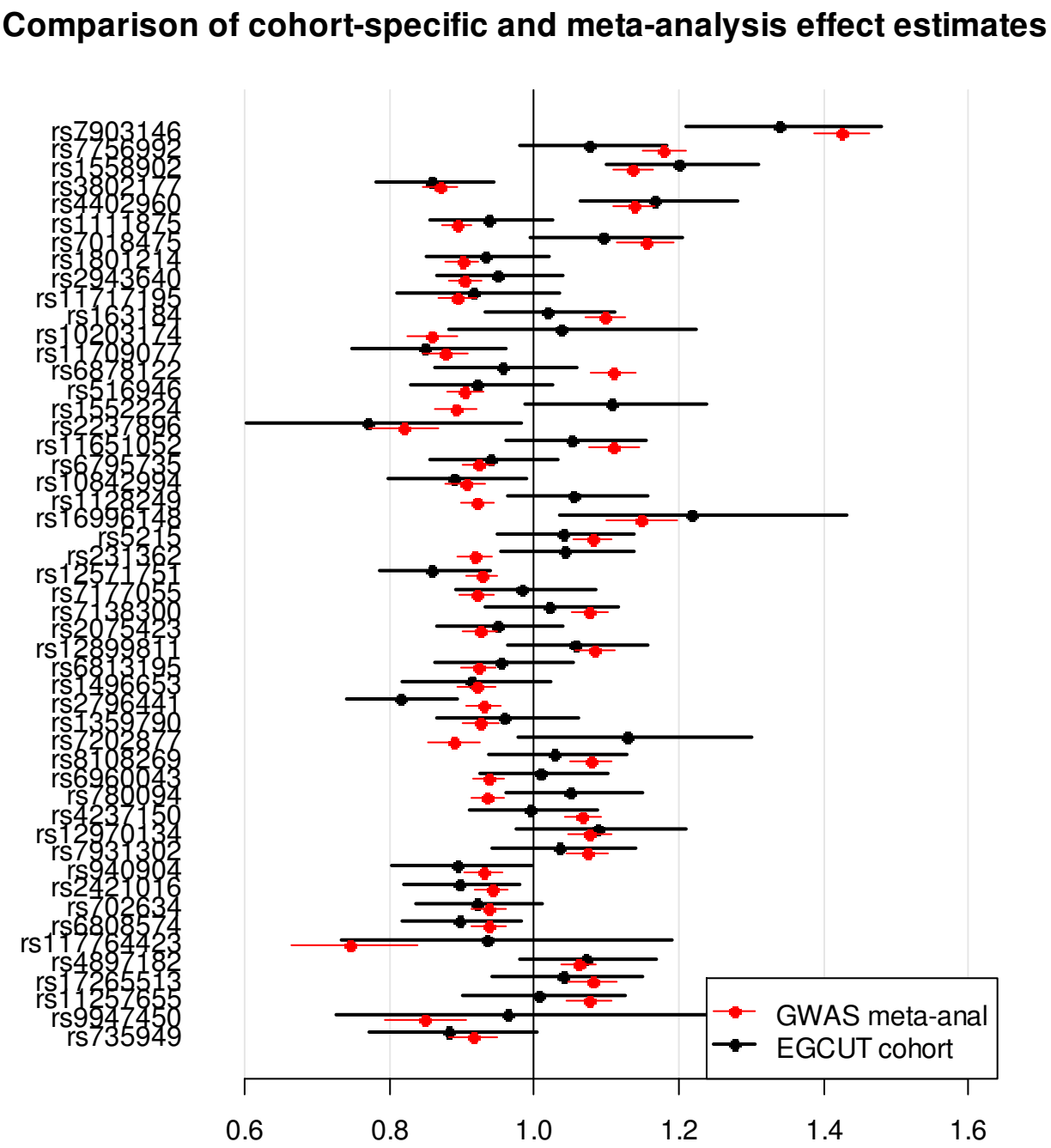
Type 2 Diabetes: what do we know about genetic risk? Known genetic loci (05/2014)



NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

Effect estimates (OR, 95% CI) from the large-scale GWAS meta-analysis and in the Estonian Biobank (Estonian Genome Center, University of Tartu, EGCUT) cohort (n=10200, incl 1200 T2D cases).

Effects of single markers relatively small – how to combine them to one predictor?



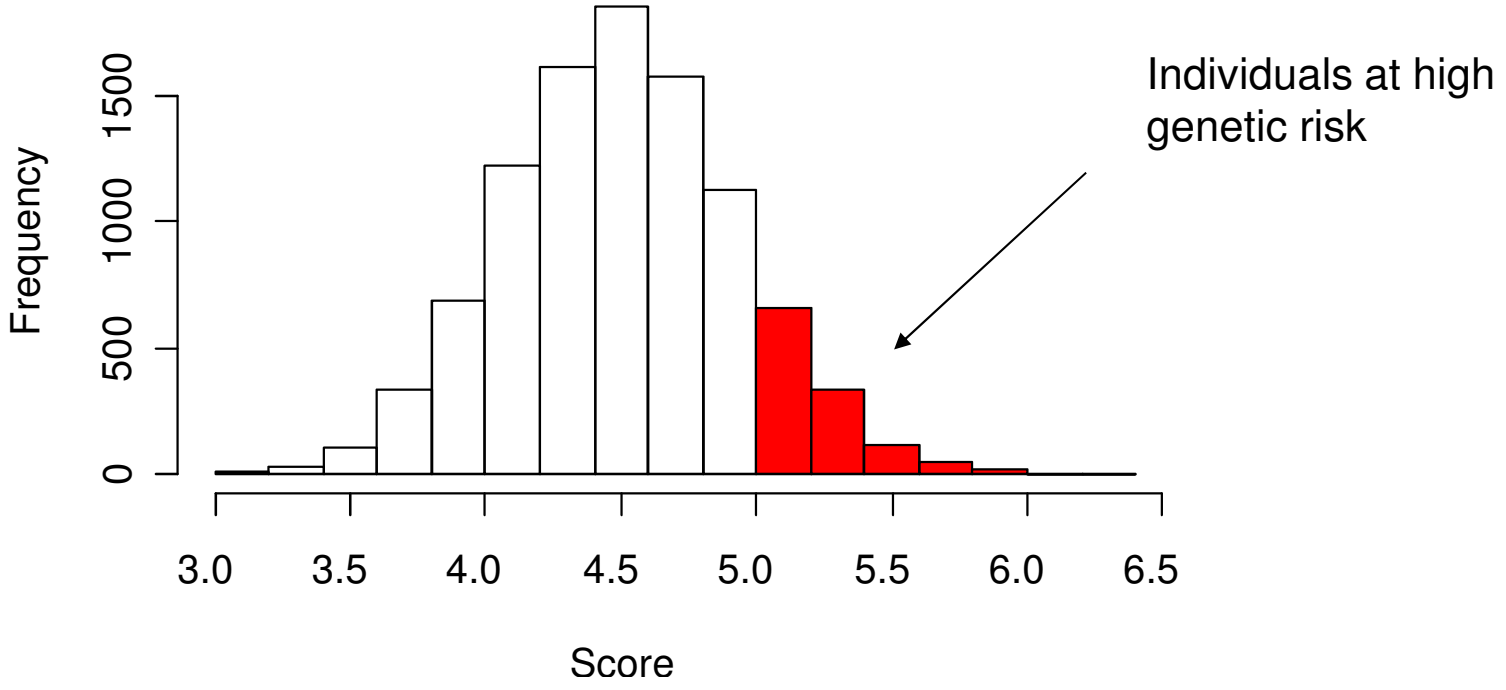
Genetic (polygenic) risk scores (GRS)

Calculated as $S = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$,

$X_2,\ldots, X_k$ - allele dosages for k independent markers (SNP-s),

$\beta_1, \beta_2, \ldots, \beta_k$ – weights

Polygenic risk score for type II diabetes:
histogram of the score in 7462 individuals (Estonian Biobank)



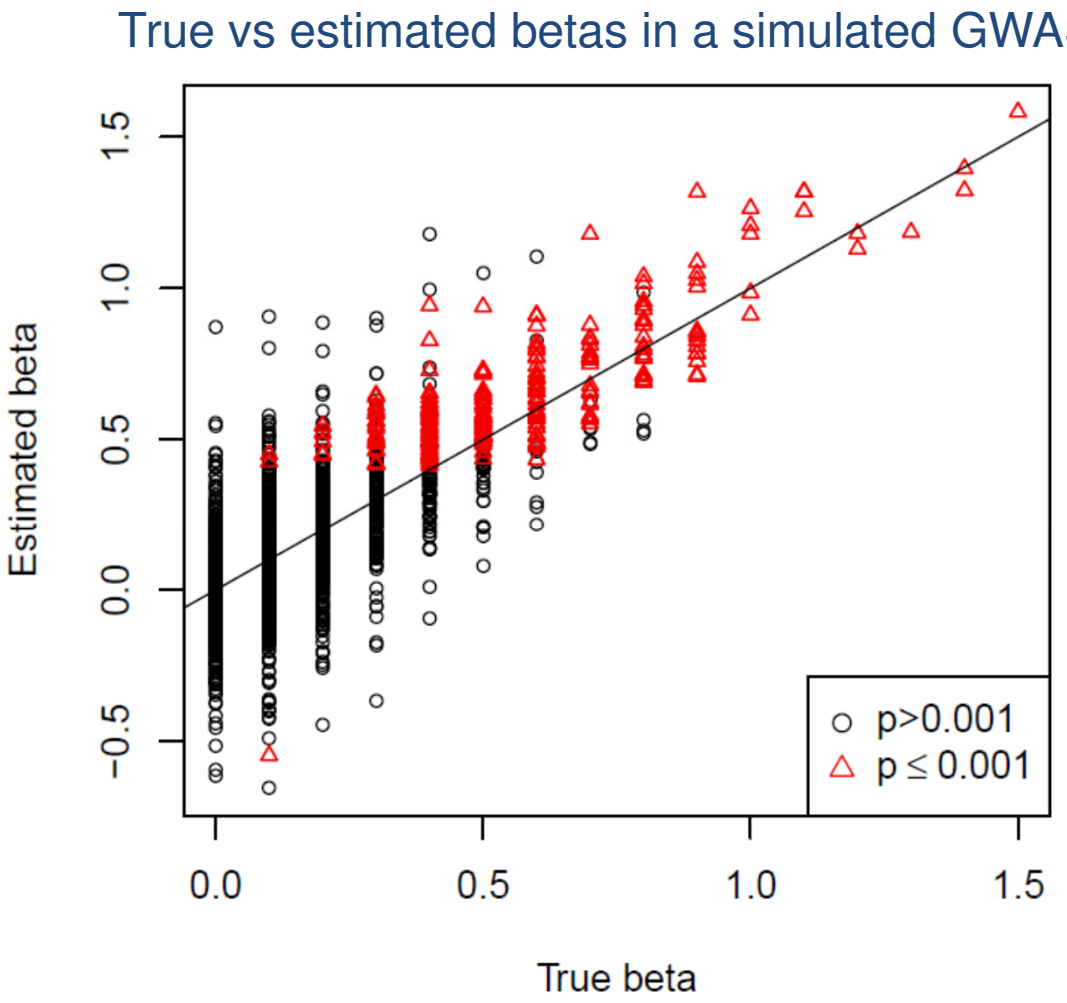
GRS: questions to address

- Which markers to include?
  - Only the most significant one?
  - All genome-wide significant markers ($p<5*10^{-8}$ in meta-analysis)
  - A larger number of markers

- Optimal weights?
  - All equal (allele counts)
  - Regression coefficients from GWAS meta-analysis?
  - Other weights?

Problem with p-value based selections: „winners curse“

One tends to select markers with effect overestimated by chance.



The „true GRS“…

The setting:

- K independent markers X_i, $i = 1 \ldots K$ are genotyped
- A subset $R(k)$, $k \leq K$ of markers having an effect on the disease risk

An additive polygenic risk score based on a subset of genetic variants, is defined as:

$$S_k = \sum_{i=1}^{K} I\left(X_i \in R(k)\right) w_i X_i$$

Weight (true association parameter)

Indicator (0/1) of whether the ith marker belongs to the subset of markers with a true effect

Both $R(k)$ and w_i are unknown and need to be estimated!

Doubly-weighted GRS:

We propose a doubly-weighted GRS as

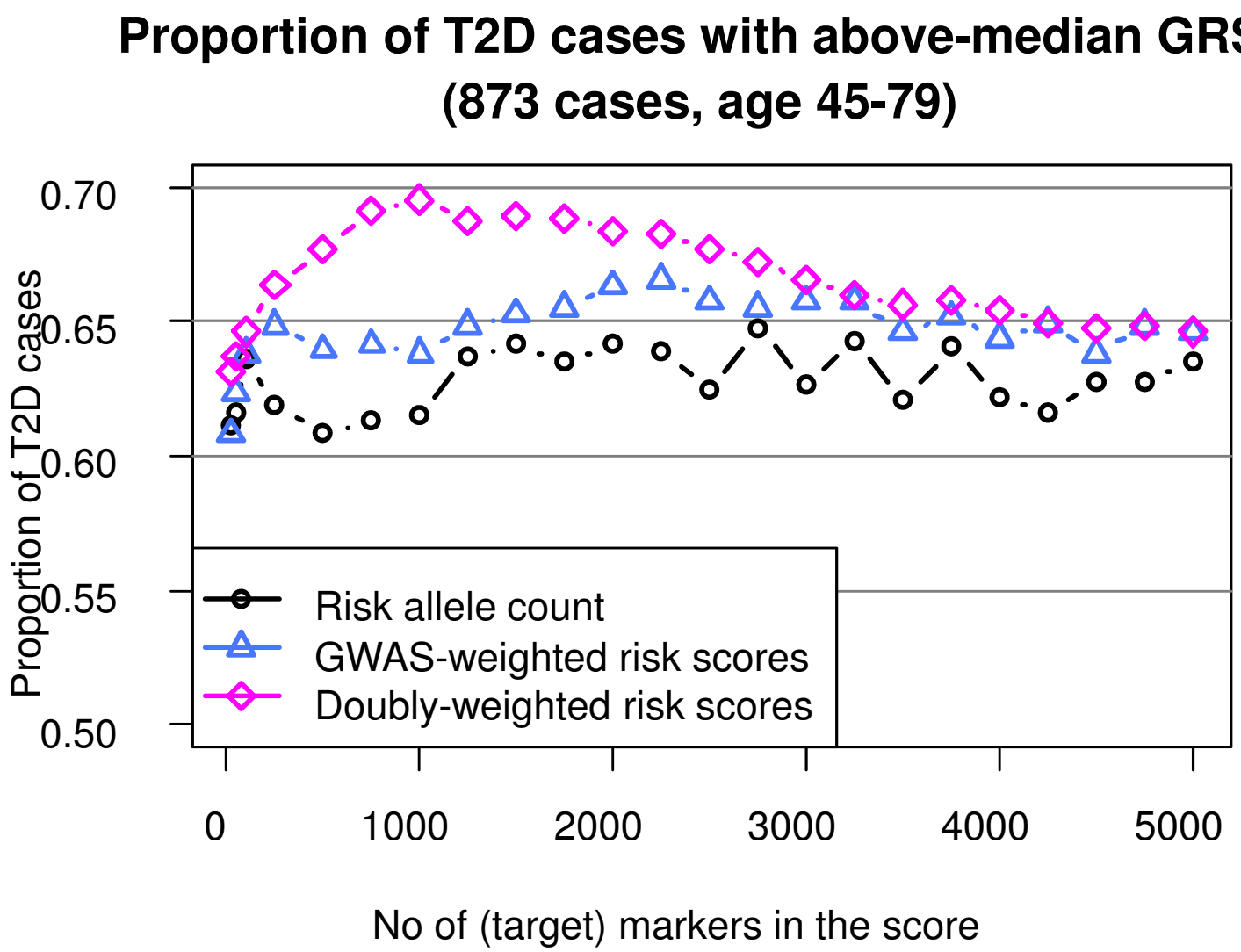
$$dGRS_k = \sum_{j=1}^{K} \hat{P}_j \times \hat{\beta}_j X_j$$

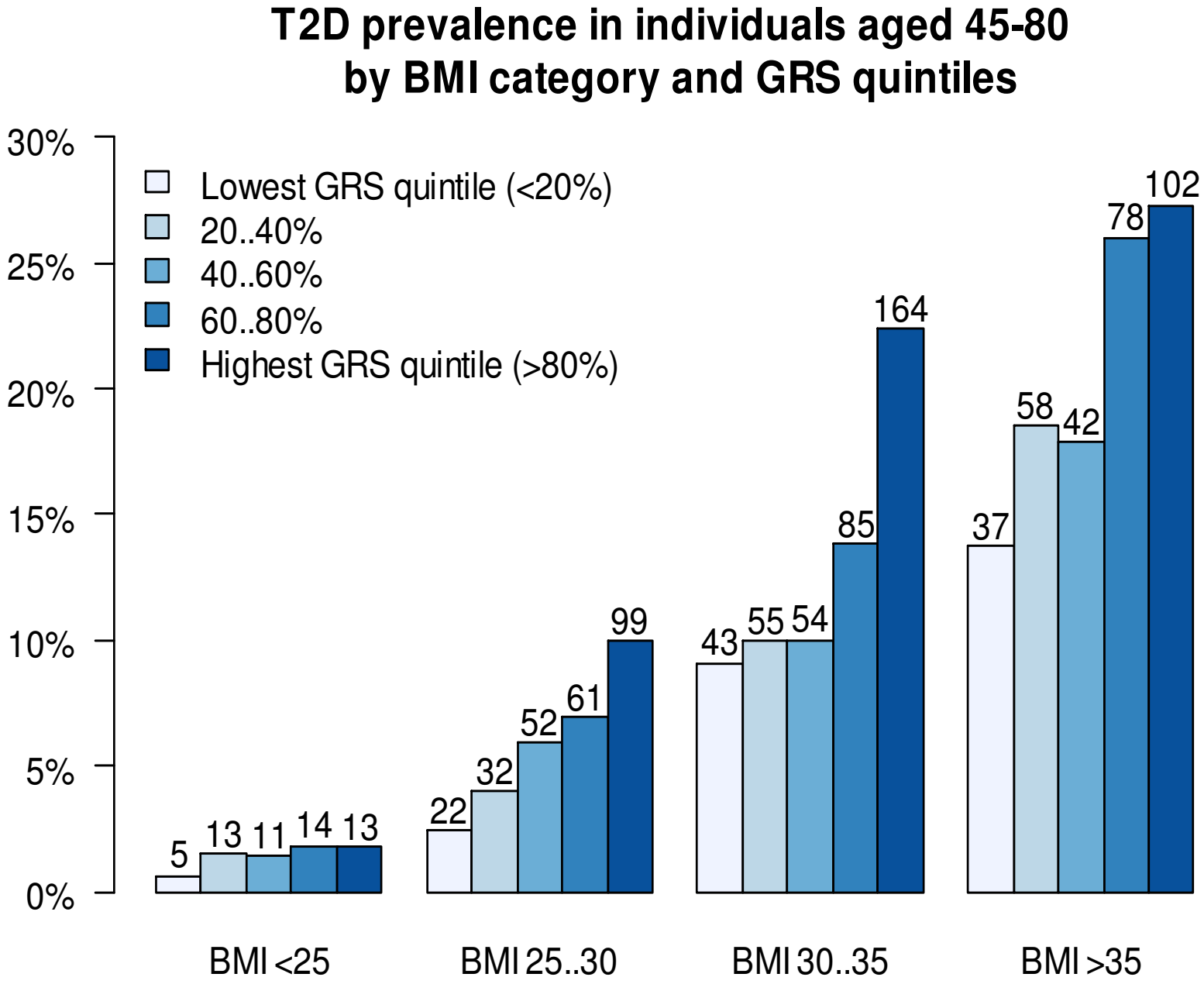
Estimated probability that the jth marker belongs to the set of k markers with strongest effect

(logistic) regression parameter estimate from GWAS

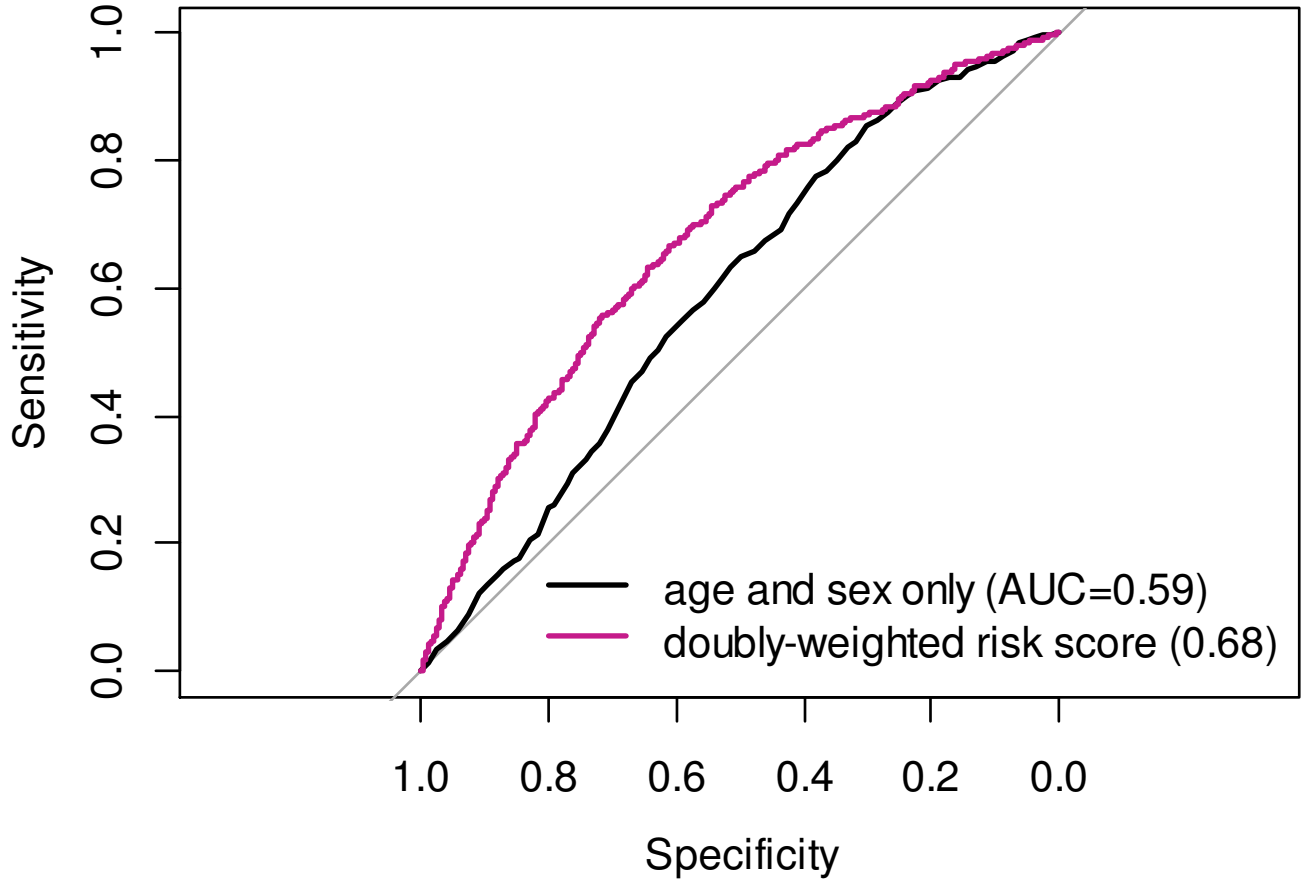
As $\hat{P}_j$ is still estimated from the estimated coefficients $\hat{\beta}_j$ and their standard errors, bias due to „winners curse“ is not complete removed, but partially corrected for

GRS for Type 2 Diabetes: allele count vs weighted scores

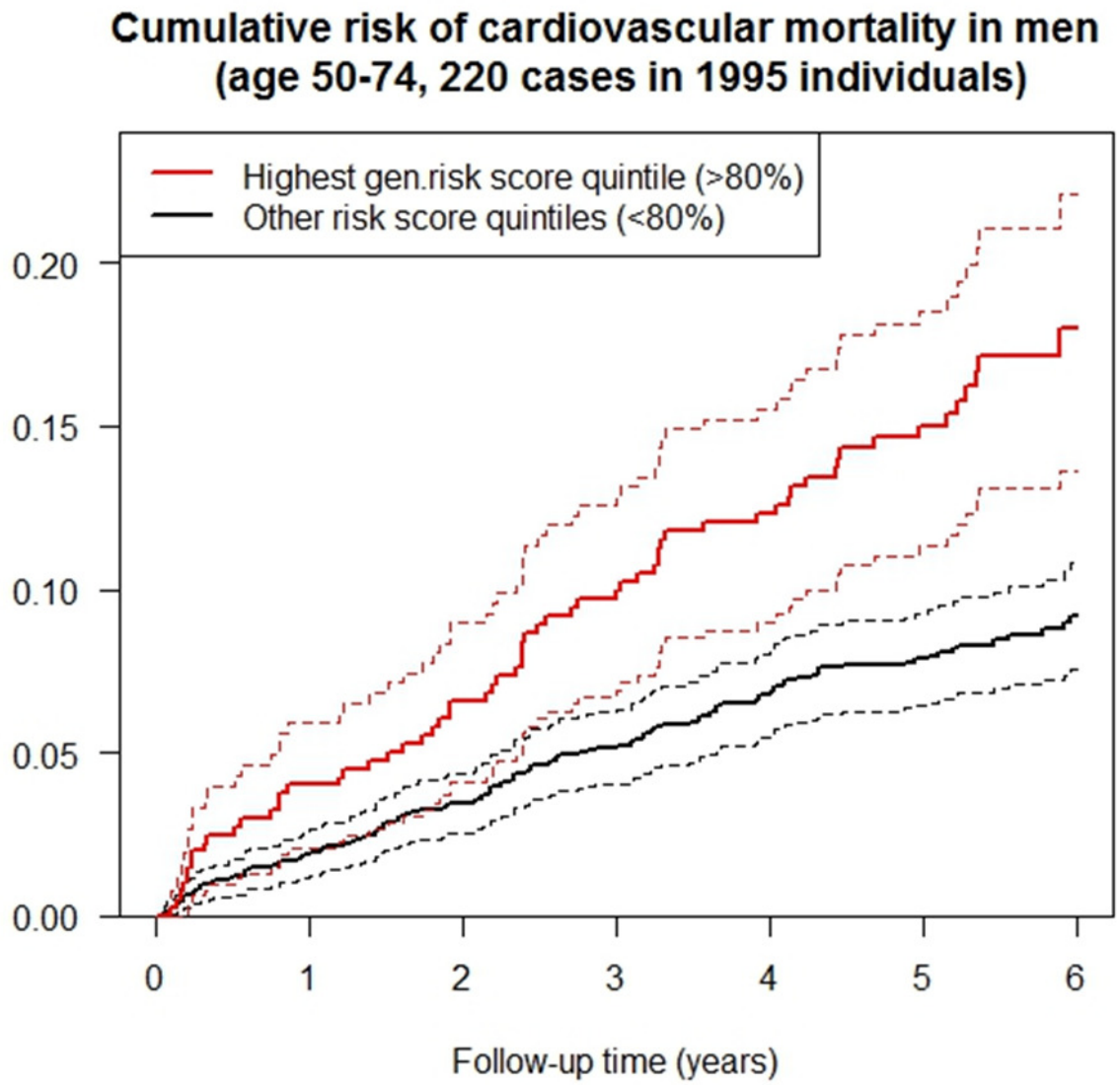



T2D prevalence in individuals aged 45-80
by BMI category and GRS quintiles
30%
25%
20%
15%
10%
5%
0%
Lowest GRS quintile (<20%)
20..40%
40..60%
60..80%
Highest GRS quintile (>80%)
5
13
11
14
13
22
32
52
61
99
43
55
54
85
164
37
58
42
78
102
BMI <25
BMI 25..30
BMI 30..35
BMI >35

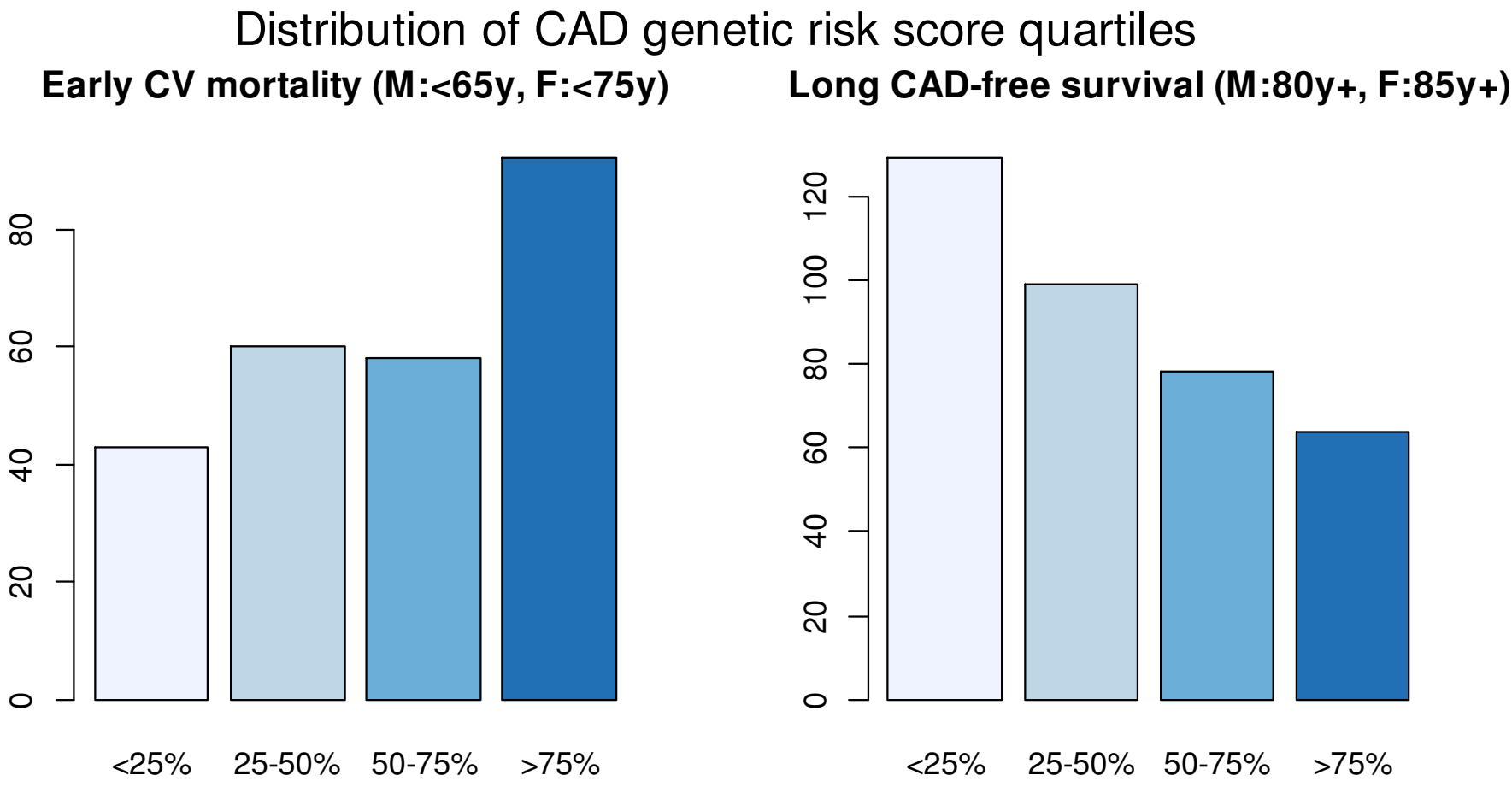
ROC curves (BMI=25..35)



Genetic risk score (GRS) for CAD and cardiovascular mortality in men



Extreme cases and controls: 253 individuals who died early of cardiovascular causes vs 370 with long survival, free of coronary artery disease (CAD).



Part 3: aspects of causality

Why needed?

- Epidemiology is always seeking for causality: if an exposure has a causal effect on the outcome, this means an intervention on exposure would affect the outcome

What is a causal effect?

Questions in epidemiology:

Is the exposure (smoking level, obesity, …) associated with the outcome (cancer diagnosis, mortality,…)?

is not the same question as

Does the exposure have a causal effect on the outcome?

Statistical analysis will answer the first question, but not necessarily the second one…

How to estimate causal effects?

First we need to define them!

…for instance, using counterfactual variables:

$Y_0 = Y(X=0)$ - potential outcome in case of no exposure

$Y_1 = Y(X=1)$ - potential outcome in case of exposure

$Y_1 - Y_0$ - causal effect of the exposure

Complication: Y_1 and Y_0 are never jointly observed in the same individual

How to estimate causal effects?

$Y_1 - Y_0$ - causal effect of the exposure

We can observe E(Y|X=1) and E(Y|X=0) - average outcomes in the exposed and unexposed subpopulations

However, in most cases,

$$E(Y|X=1) - E(Y|X=0) \neq E(Y_1) - E(Y_0)$$

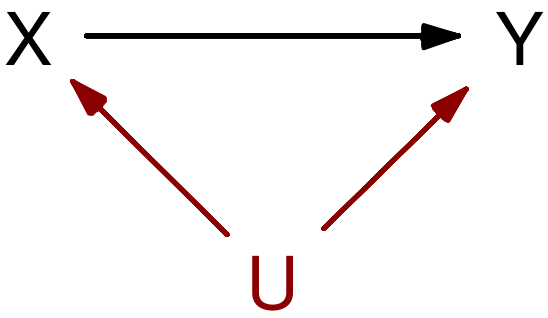
Main reason: (unmeasurable) confounders

Causal graphs (DAGs)

X – (observed) exposure

Y – (observed) outcome,

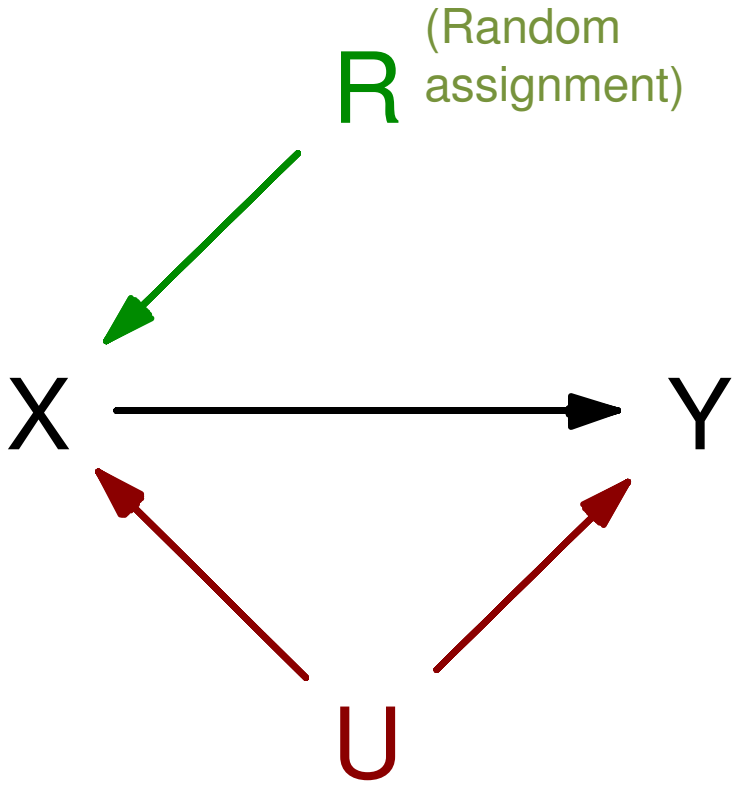
U – unmeasured confounders



If U affects both X and Y, it can create a correlation between X and Y even when there is no causal effect of X on Y, or it can bias the estimates of an actual causal effect

How to estimate causal effects?

One possible solution: randomized study – random allocation of the exposures

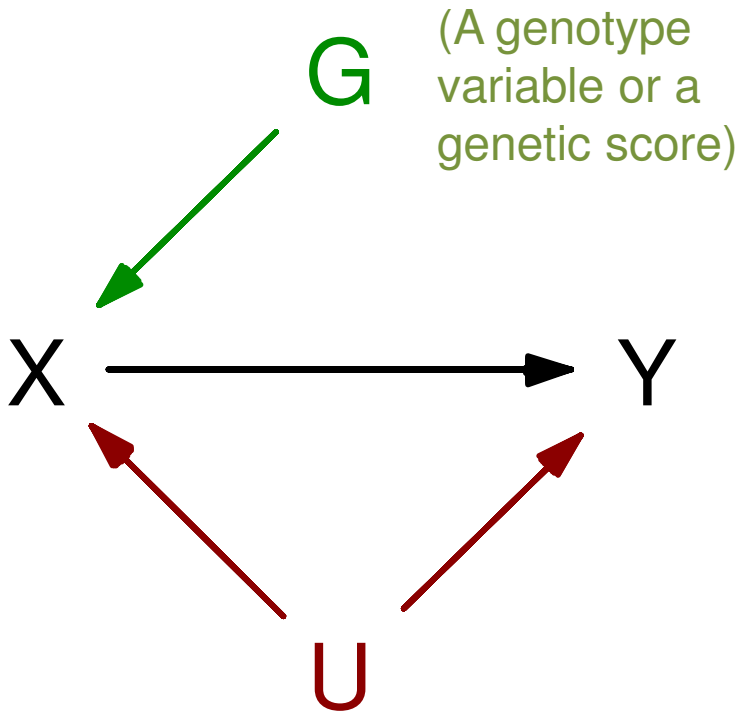


Association between R and Y can only be observed in case there is a causal effect of X on Y – thus even when R does not uniquely determine the exposure, a test of association between R and Y is a valid test of the causal effect

Not all exposures can be randomized!

Can genetics help us? The idea of Mendelian Randomization

- Idea: genetic determinant of an exposure would act like random assignment

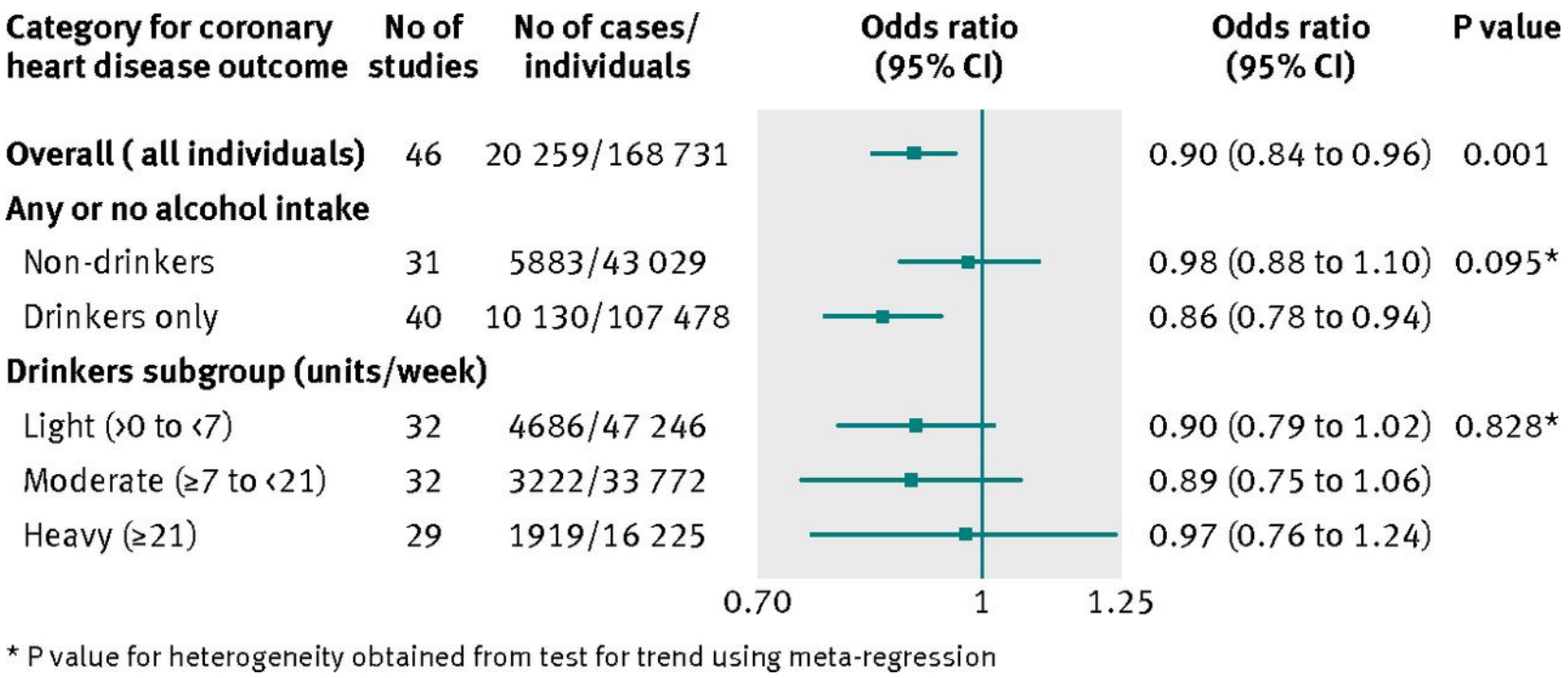


Is an association test between G and Y a valid test for a causal effect?

Example from recent literature:

Associated with 17.5% less alcohol intake

Fig 2 Meta-analysis pooled estimates of the association between ADH1B rs1229984 (A-allele carriers v non-carriers) and coronary heart disease overall, and stratified by alcohol intake.



Category for coronary heart disease outcome	No of studies	No of cases/ individuals	Odds ratio (95% CI)	Odds ratio (95% CI)	P value
Overall (all individuals)	46	20 259/168 731		0.90 (0.84 to 0.96)	0.001
Any or no alcohol intake					
Non-drinkers	31	5883/43 029		0.98 (0.88 to 1.10)	0.095*
Drinkers only	40	10 130/107 478		0.86 (0.78 to 0.94)	
Drinkers subgroup (units/week)					
Light (>0 to <7)	32	4686/47 246		0.90 (0.79 to 1.02)	0.828*
Moderate (≥7 to <21)	32	3222/33 772		0.89 (0.75 to 1.06)	
Heavy (≥21)	29	1919/16 225		0.97 (0.76 to 1.24)	

* P value for heterogeneity obtained from test for trend using meta-regression

Holmes M V et al. BMJ 2014;349:bmj.g4164

Mendelian randomization (MR)

Parameter of interest:

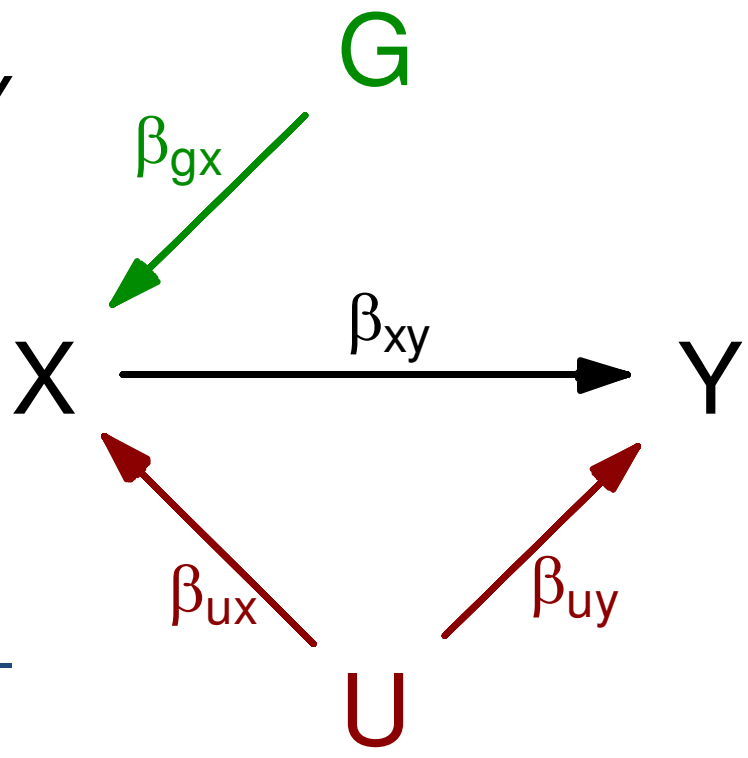
– β_{xy}, the causal effect of X on Y

Complication:

Association between X and Y is confounded by U

Solution:

genotype G serves as an instrument – correlation between G and Y provides evidence on β_{xy}

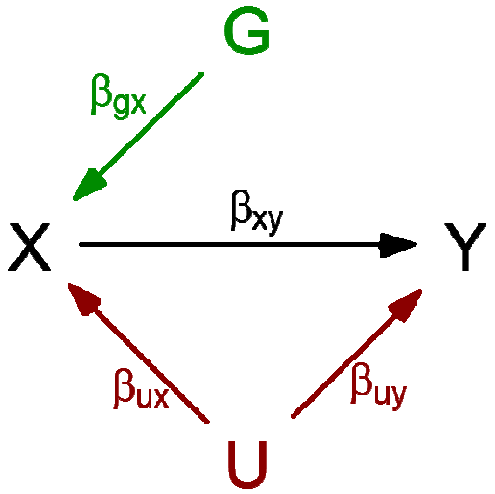


Untestable assumptions:

no direct effect of G on Y; no association between G and U

MR– how does it work?

With linear models, the following equations correspond to the assumed structure:



$$Y = c_y + \beta_{xy} X + \beta_{uy} U + \varepsilon_y \qquad E(\varepsilon_y | X,U)=0$$

and:

$$X = c_x + \beta_{gx} G + \beta_{ux} U + \varepsilon_x, \qquad E(\varepsilon_x | G,U)=0$$

As $U \perp G$, $E(X|G) = c_x + \beta_{gx} G$

and $E(Y|G) = c_y + \beta_{xy} E(X|G) = c_y + \beta_{xy}\beta_{gx} G$

c_x, c_y - some constants

MR – how does it work?

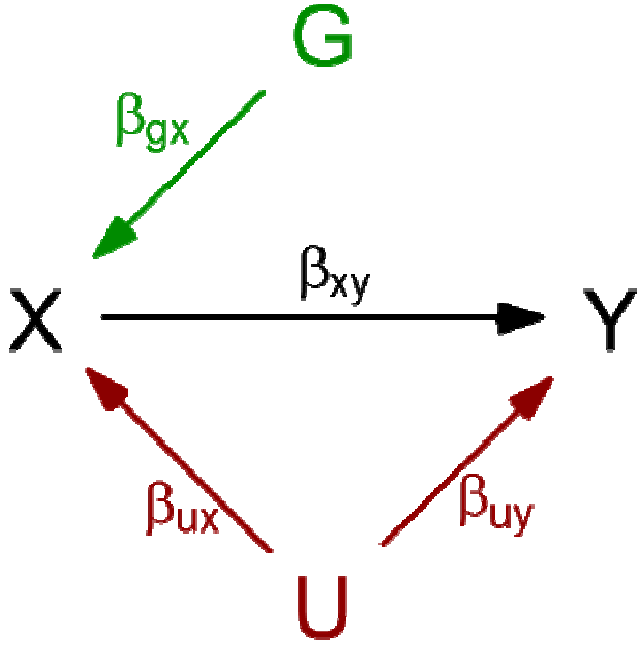
As

Regressing X on G, we estimate:

$E(X|G) = c_x + \beta_{gx}G$

Regressing Y on G we estimate:

$E(Y|G) = c_y + \beta_{xy}\beta_{gx}G,$



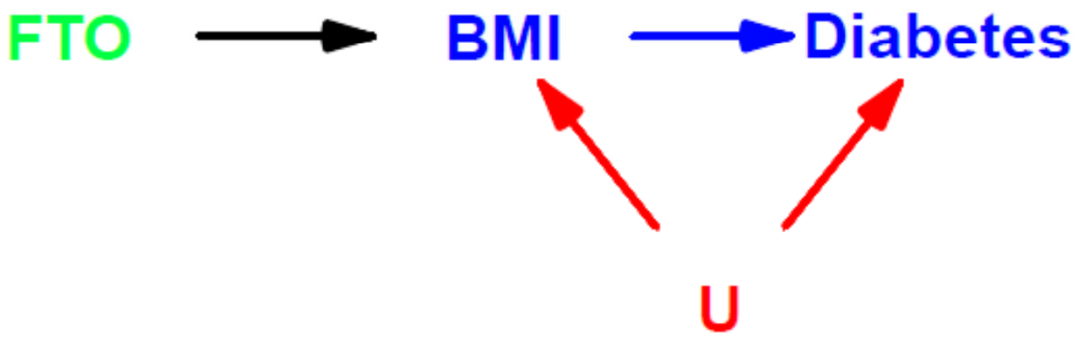
Thus also β_{xy} is estimable as the ratio of the two estimated coefficients of G

– the technique is called Instrumental Variables (IV) estimation, where G is an instrument

…provided there is no direct effect of G on Y!

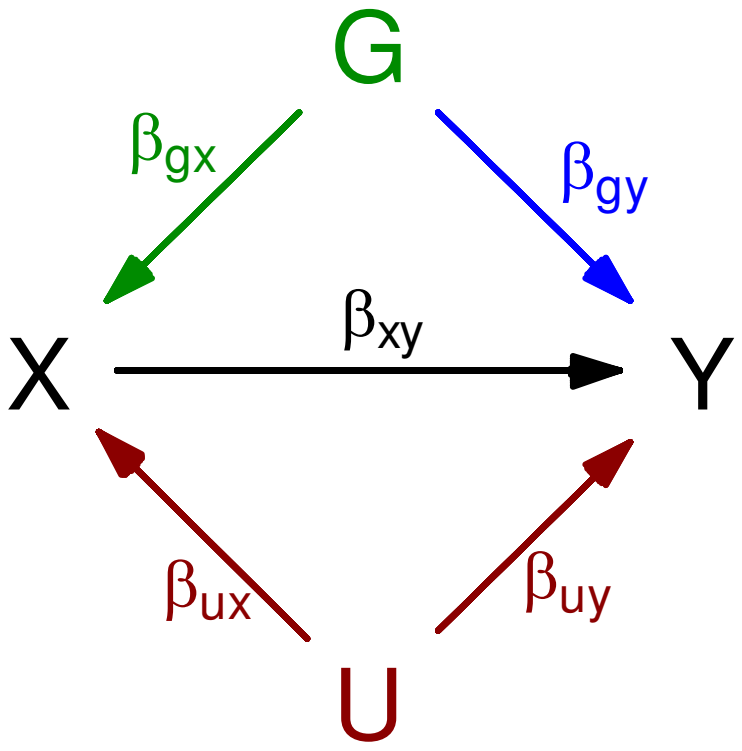
Mendelian randomization example

FTO genotype, BMI and Blood Glucose level (related to Type 2 Diabetes risk; Estonian Biobank, n=3635, aged 45+)



- Average difference in Blood Glucose level (Glc, mmol/L) per BMI unit is estimated as 0.085 (SE=0.005)
- Average BMI difference per FTO risk allele is estimated as 0.50 (SE=0.09)
- Average difference in Glc level per FTO risk allele is estimated as 0.13 (SE=0.04)
- Instrumental variable estimate of the mean Glc difference per BMI unit is 0.209 (se=0.078)

A general association structure with one genotype and two phenotypes



If $\beta_{gy} \neq 0$, the genotype G is said to have a pleiotropic effect on Y (violation of the MR assumption!)

What is estimated in the presence of pleiotropy?

As

Regressing X on G, we estimate:

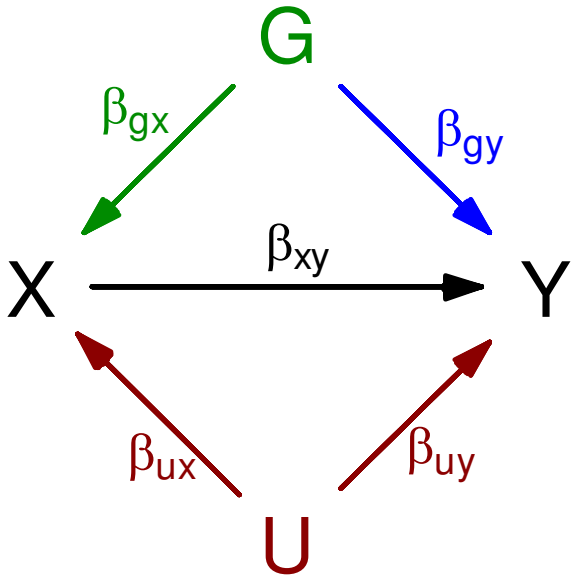
$$E(X|G) = c_x + \beta_{gx}G$$

Regressing Y on G we estimate:

$$E(Y|G) = c_y + \beta_{xy}\beta_{gx}G + \beta_{gy}G,$$

If we use the IV method, we estimate

$$\frac{\beta_{xy}\beta_{gx} + \beta_{gy}}{\beta_{gx}}$$



Thus a MR estimate in the presence of pleiotropy is biased by β_{gy}/β_{gx}

Can we test pleiotropy?

- Can we fit a regression model for Y, using both X and G as covariates?

- In this case we estimate:

$E(Y|X,G) = c_y + \beta_{xy}X + \beta_{gy}G + \beta_{uy}E(U|X,G)$

If X depends on U and G, one can also express U as a function of X and G!

It appears that:

$$E(Y|X,G) = \text{const} + \left[\beta_{xy} + \frac{\beta_{ux}\beta_{uy}}{1-\beta_{gx}^2}\right]X + \left[\beta_{gy} - \beta_{gx}\frac{\beta_{ux}\beta_{uy}}{1-\beta_{gx}^2}\right]G$$

Can we test pleiotropy?

$$E(Y|X,G) = \text{const} + \left[\beta_{xy} + \frac{\beta_{ux}\beta_{uy}}{1-\beta_{gx}^2}\right] X + \left[\beta_{gy} - \beta_{gx}\frac{\beta_{ux}\beta_{uy}}{1-\beta_{gx}^2}\right] G$$

So:

- Even when there is no causal effect of X, thus $\beta_{xy}=0$, we may still estimate a nonzero (and significant!) coefficient of X

- Even when there is no pleiotropy ($\beta_{gy}=0$), we may still estimate a nonzero (negative!) coefficient of G!

Conclusions

- Large prospective biobank cohorts make it possible to discover important pathways leading to diseases and premature mortality

- Biobank cohorts are different from standard „textbook-datasets“ for survival analysis: timescale choice and sampling design issues need to be considered

- Instead of single variants, polygenic scores are more likely to have potential for personalized preventive medicine

Conclusions II

- Association is not causality - old truth, but still needs to be reminded while analyzing –omics data

- Mendelian Randomization can be a useful tool to establish causality, but it relies on statistically untestable assumptions. The assumptions should be verified based on external knowledge (biology).

There are no „forbidden models“, but it is important to understand the interpretation of model parameters given realistic assumptions.

Collaborators

(Estonian Genome Center, University of Tartu)

www.biobank.ee

- Kristi Läll (PhD student)
- Reedik Mägi
- Tõnu Esko
- **Andres Metspalu (director)**

- Peter Joshi (Univ. of Edinburgh – survival analysis)

Krista.Fischer@ut.ee

estonian genome center
university of tartu

European Mathematical Genetics Meeting 2017
April 5-7th, 2017, Tartu, Estonia

estonian genome center
university of tartu