UNIVERSITY OF TARTU
1632

Estonian Genome Center

Statistical dilemmas in personalized risk prediction: the Estonian Biobank example

Krista Fischer

Institute of Mathematics and Statistics,
Institute of Genomics,
University of Tartu

Estonian Academy of Sciences

Outline

- What is risk?
- Risk of death
- Basic methods for survival analysis
- Risk of a disease and competing risks
- Estonian Biobank and solutions for personalized medicine: an overview
- Risk prediction for common complex diseases in the Estonian Biobank

What is risk?

What is probability?

A quiz (D. Spiegelhalter)

https://tinyurl.com/yyeavsa4

How good we are at estimating probabilities?

How was the score calculated?

The Brier score:

Confidence	Score if you are right	Score if you are wrong
5	0	0
6	9	–11
7	16	–24
8	21	–39
9	24	–56
10	25	–75

Why?

- Know all the answers and are 100% confident in them (score 10): **25** points per question.
- Choose the answer randomly, but still choose the confidence score 10: on average, you will get half of the answers correct and half will be wrong: (25-75)/2= **-25**

- Know the answers with the probability of 80% and choosing 8 as the confidence. Average: 0,8*21-0,2*39= **9** points.
- Know the answers with the probability of 50% and choosing 8 as the confidence. Average: 0,5*21-0,5*39= **-9** points, etc.

Too much confidence will lead to a negative score: very often people overestimate their probability of knowing the answer!

https://www.nature.com/articles/d41586-024-04096-5

ESSAY | 16 December 2024 | Correction 18 December 2024

Why probability probably doesn't exist (but it is useful to act like it does)

All of statistics and much of science depends on probability — an astonishing achievement, considering no one's really sure what it is.

By David Spiegelhalter

Although a bit confusing, it is still useful to use probabilities to speak about risks (or do we have a better alternative?)

- What is the risk of death?
- The probability of death (at some point) is 100% for each and every one of us.
- We always have to keep a time axis in mind – death in next 10 years, before the age of 80, before 2050, …

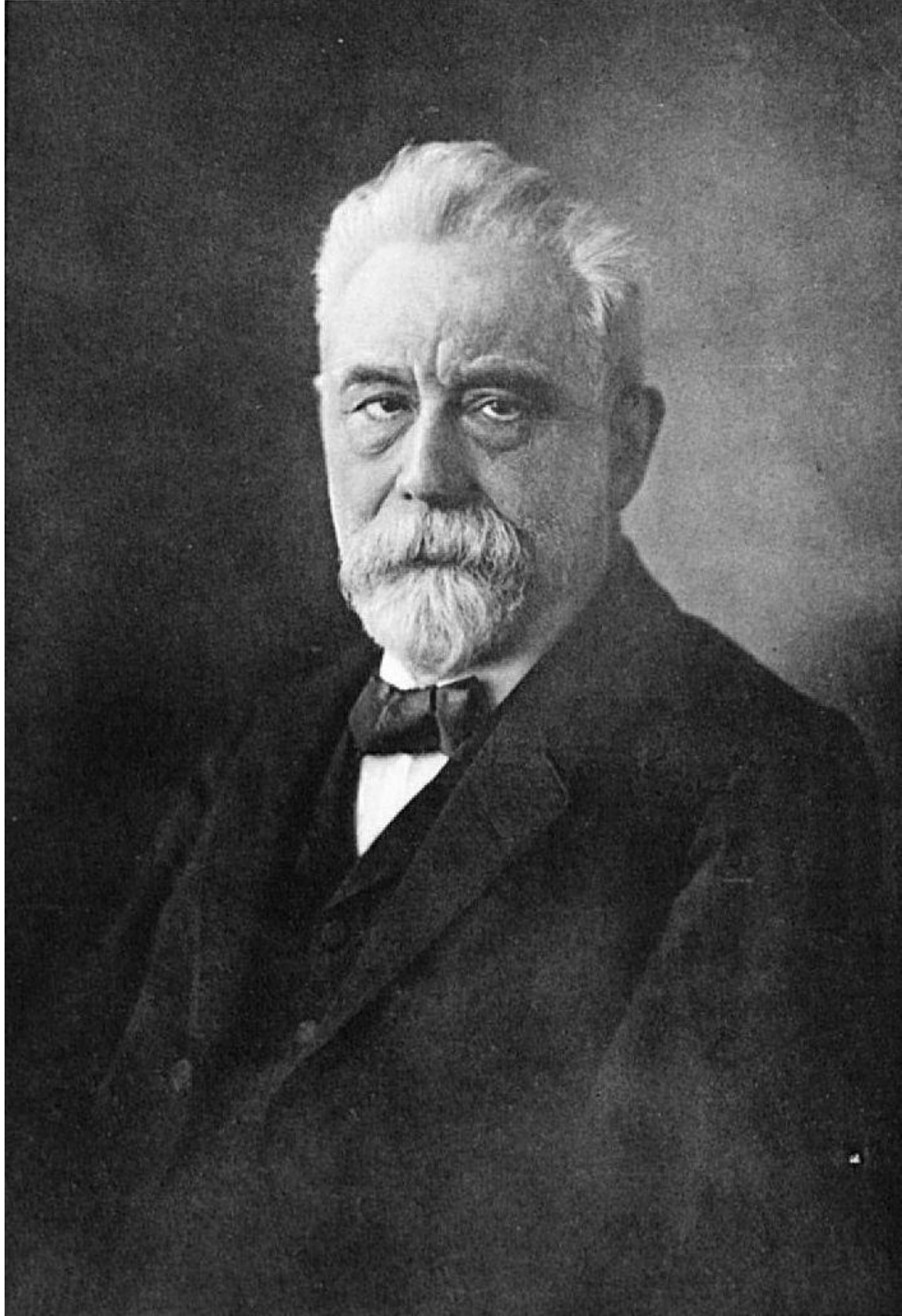
Although a bit confusing, it is still useful to use probabilities to speak about risks (or do we have a better alternative?)

- What is the risk of death?
- The probability of death (at some point) is 100% for each and every one of us.
- We always have to keep a time axis in mind – death in next 10 years, before the age of 80, before 2050, …

A bit of history…

Wilhelm Lexis (1837-1914)

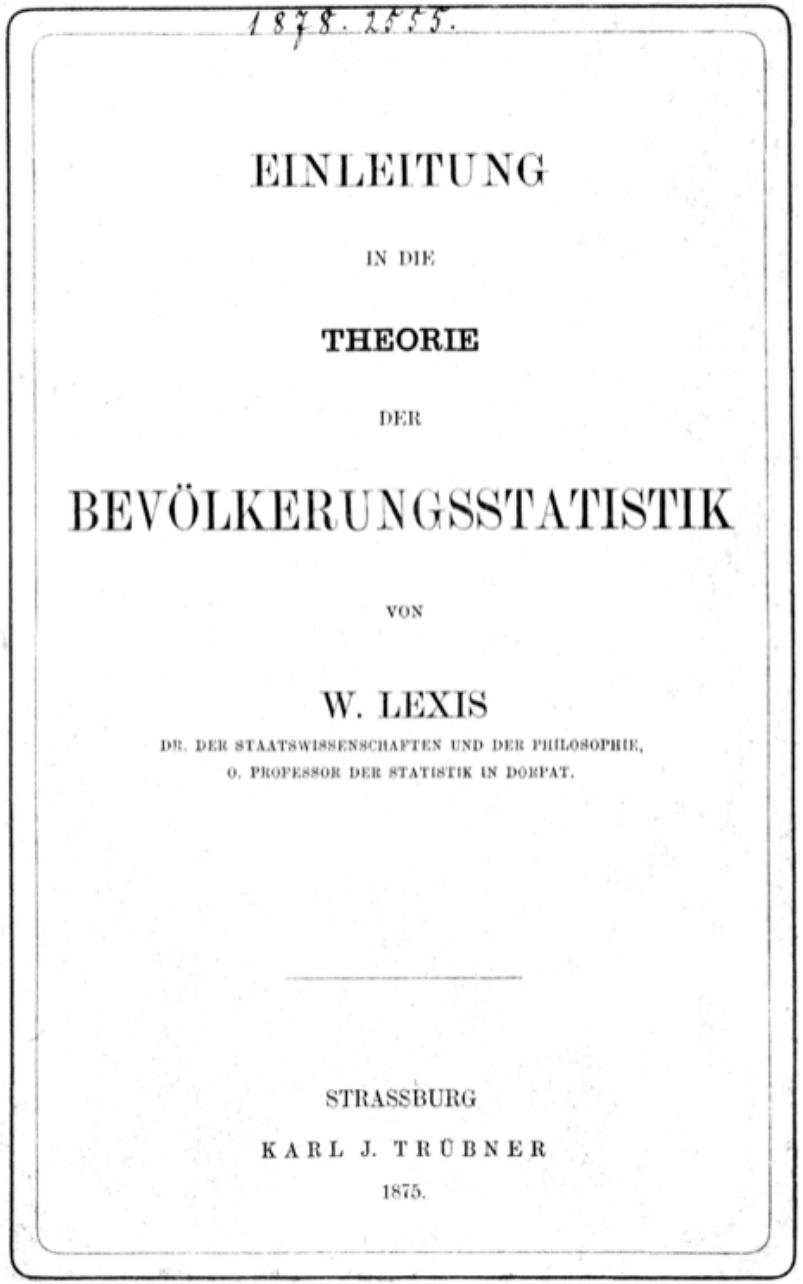
- German statistician, economist, and social scientist
- Pioneer of the analysis of demographic time series
- Professor at the University of Dorpat (now: Tartu University) in 1874-1876 (chair of geography, ethnography and statistics)



Introduction to the Theory of Population Statistics

by W. Lexis

- One of his most famous works
- Published 150 years ago
- Still known and cited by demographers, epidemiologists and also actuarians



The Lexis diagram (1875)

- Horizontal axis: birth cohorts (year of birth)
- Vertical axis: age
- Diagonal lines: calendar years

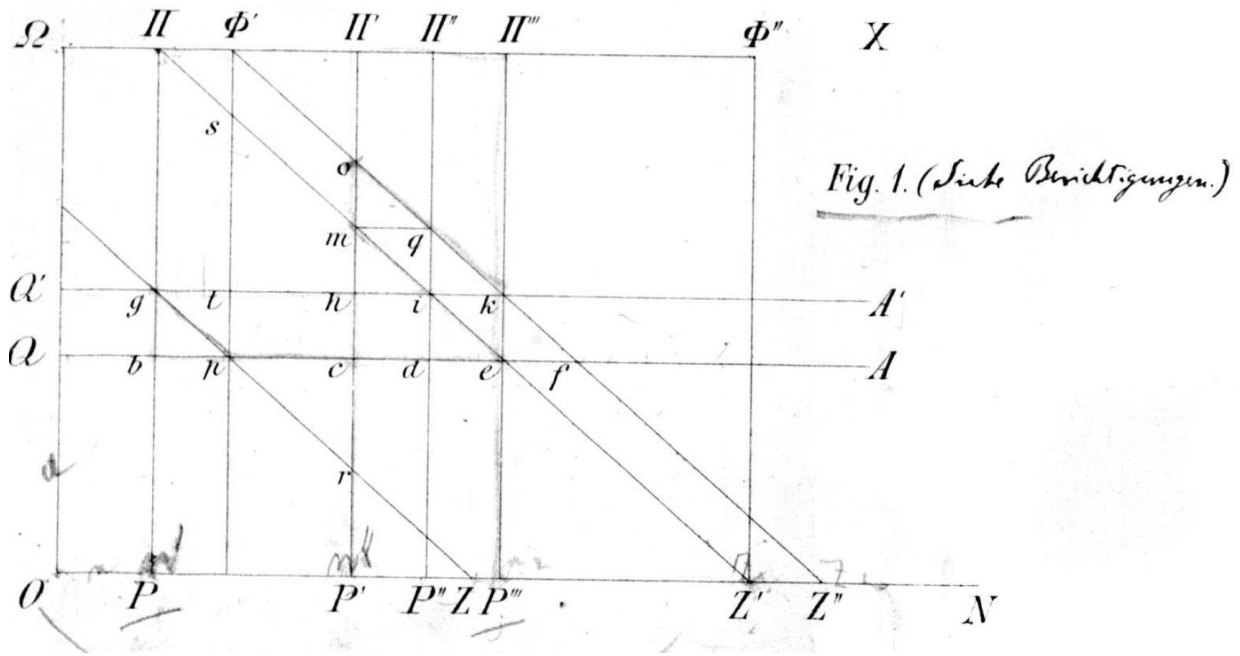


Fig. 1. (Siehe Berichtigungen.)

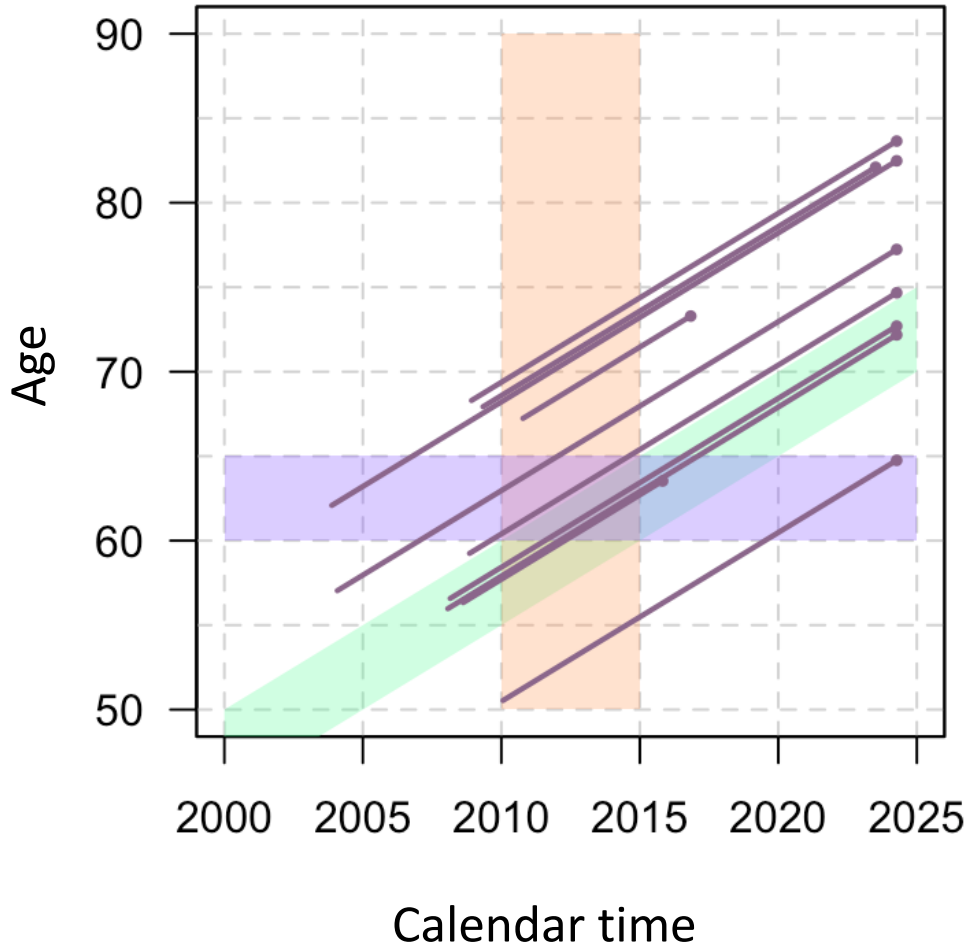
Lexis, 1875, Figure 1

A modern version of the Lexis diagram

- Introduced by R. Pressat (1961), now used widely in demography and epidemiology
- Basis of age-period-cohort modeling
- The risk can vary across each of the three time scales

- The problem if identifiablity: linear dependence

  Age = Calendar time – Birth time



Mai-Britt Meriloo, Bachelor's thesis 2025

A conference advert

150 years of Lexis diagram

- Tartu, October 17-18, 2025
- Organized by the Estonian Statistical Association
- (17.10: presentations in English, 18.10: Estonian)

The risk of death: time-to-event analysis approach

T: time to (age at) death.

Survival function: $P(T > t)$

- Probability of survival up to time (age) t.
- Easy to estimate (classical probability), if we have complete data: a cohort of individuals followed from birth to death.
- Usually, we don't have such cohorts!

The idea of Kaplan-Meier estimator (and other similar estimators)

A fictional study: duration 5 years. Suppose we start with individuals who are either 65, 70 or 75 in the beginning and record 5-year mortality.

Age	N	Alive in 5 years	5-year mortality	5-year survival
65	1000	850	0.15	0.85
70	700	560	0.2	0.8
75	600	450	0.25	0.75

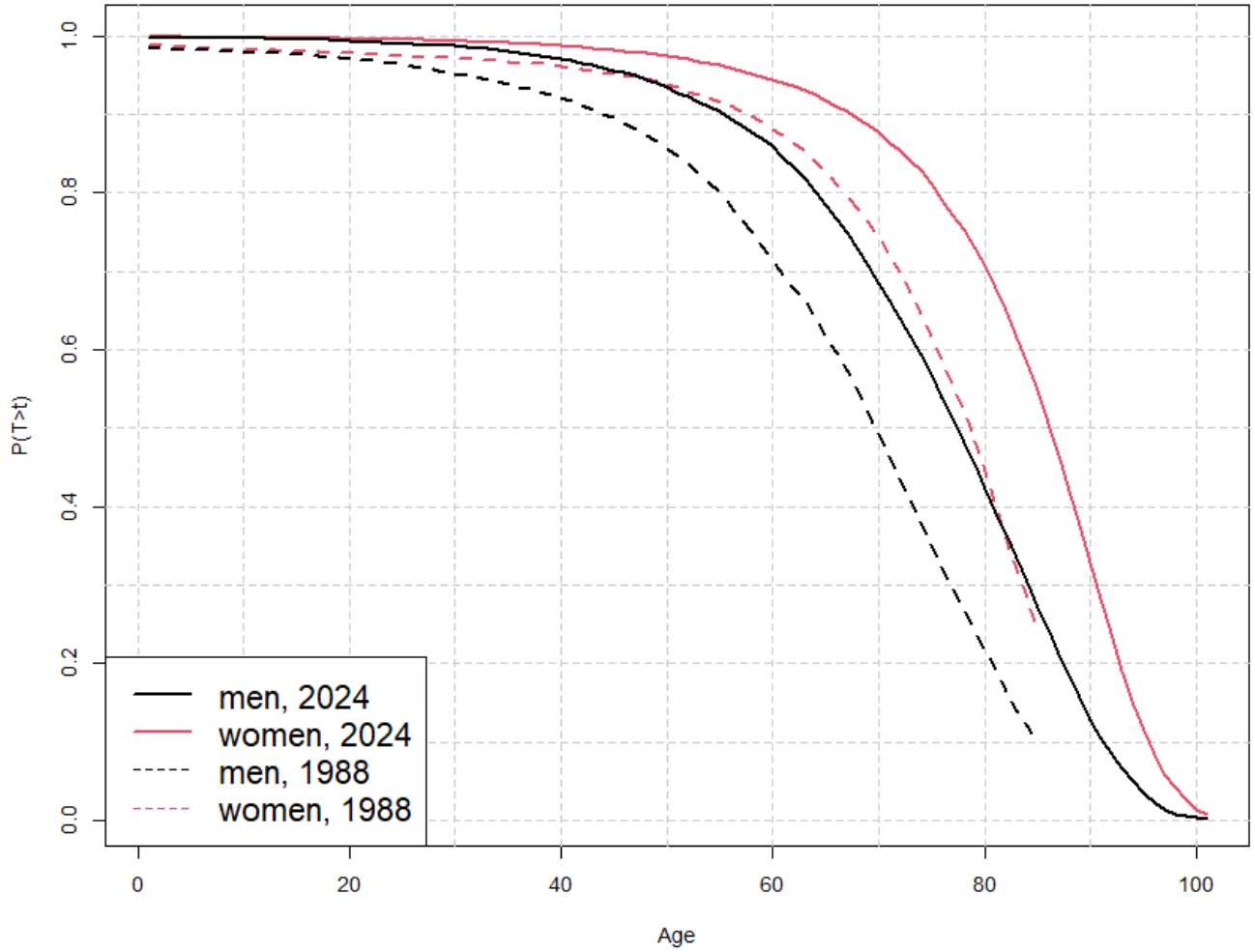
We can estimate the conditional probabilities $P(T > 70|T > 65)$ (5-year survival among 65-year-olds), $P(T > 75|T > 70)$ and $P(T > 80|T > 75)$.

Assuming that the survival rates stay the same in time, we can estimate the probability of a 65-year old to live up to 80, by combining these together:

$$P(T > 80|T > 65) = P(T > 80|T > 75) \cdot P(T > 75|T > 70) \cdot P(T > 70|T > 65).$$

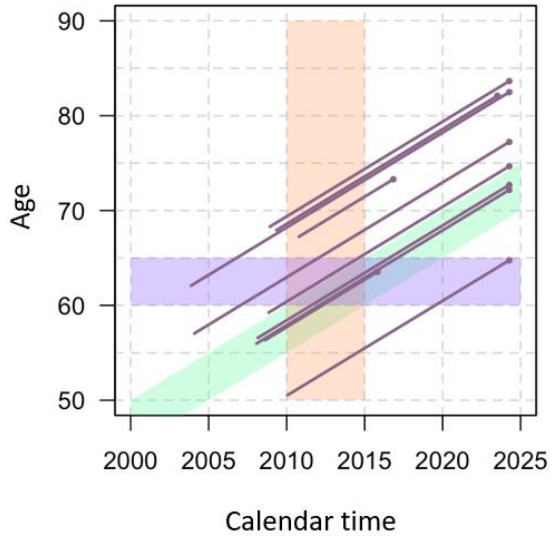
Thus, the probability here is $P(T > 80|T > 65) = 0.85 \cdot 0.8 \cdot 0.75 = 0.51.$

Survival curves based on one year of data (Statistics Estonia)

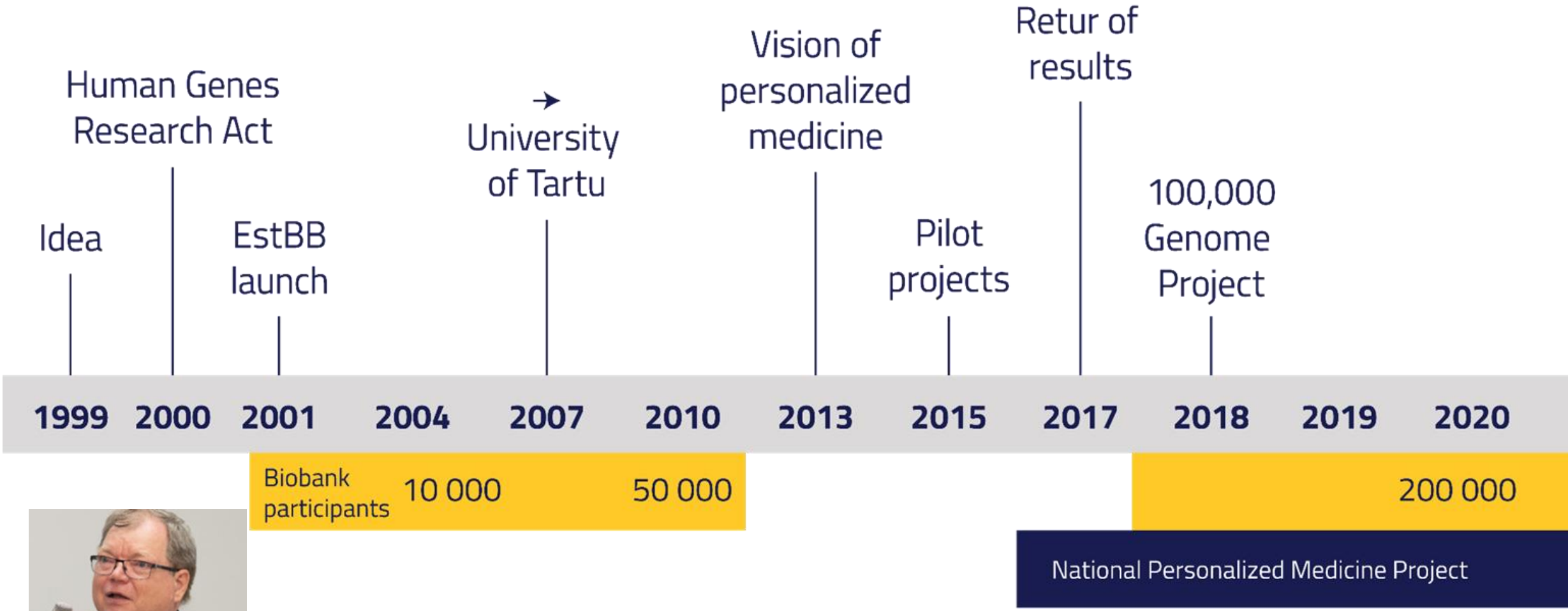


Useful for population statistics, but useless for individual survival prediction!

(Think of the Lexis diagram)



The Estonian Biobank: from population-based biobank to personalized medicine



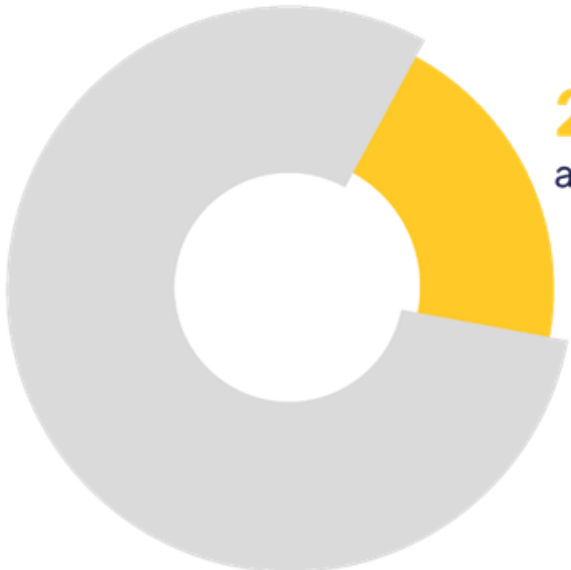
Prof. Andres Metspalu

Estonian
Genome Center

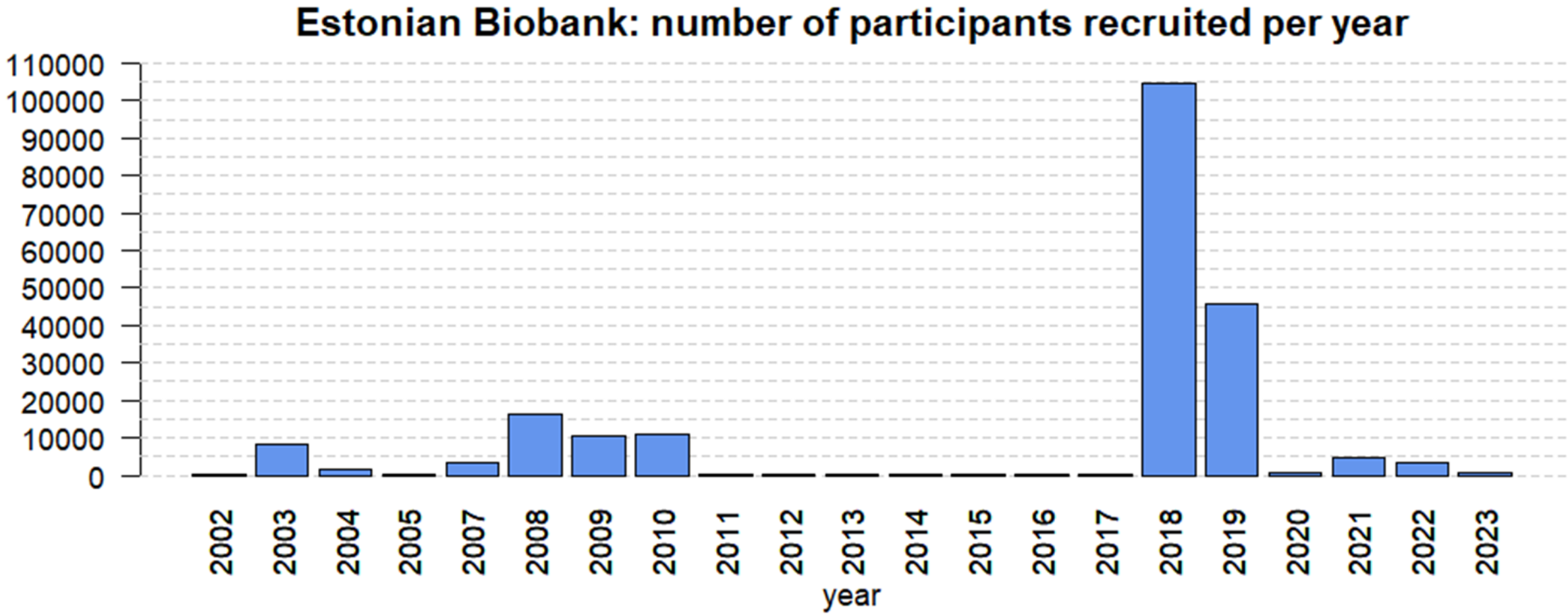
211 000+
biobank participants

Health records, diet,
physical activity etc

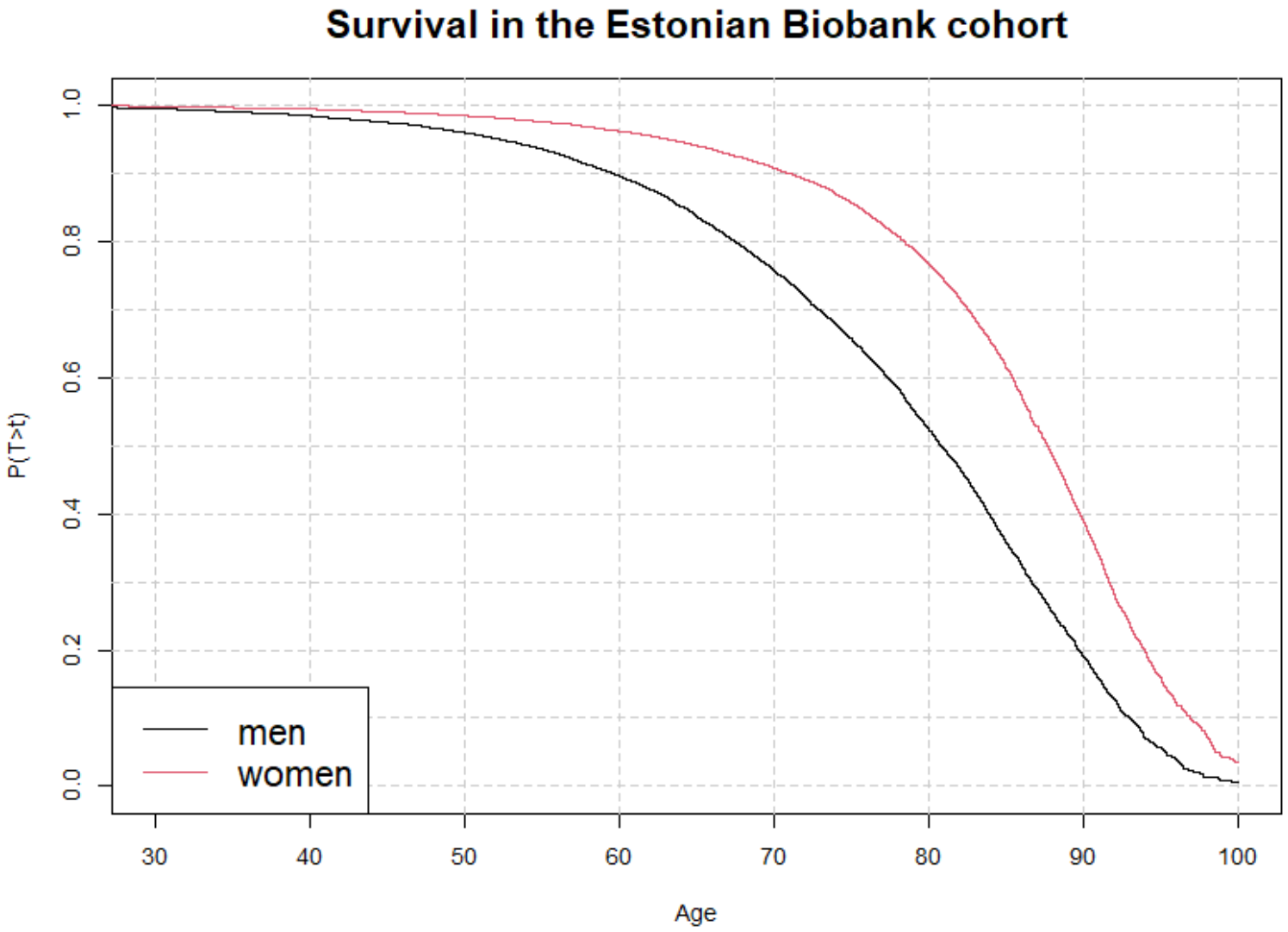
DNA, plasma and
cell samples



20% of Estonian
adult population

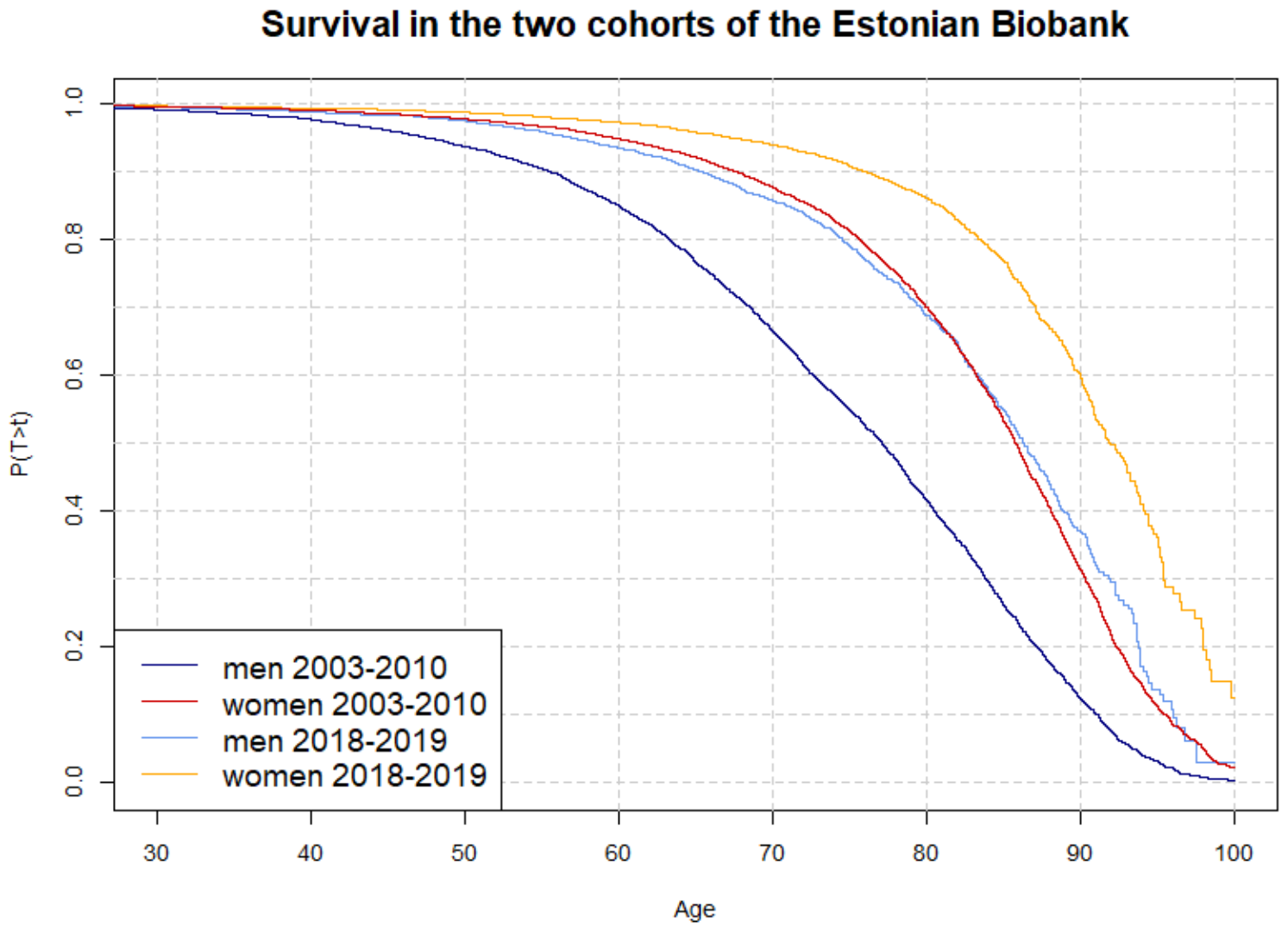

Estonian Biobank: number of participants recruited per year
110000
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
2002
2003
2004
2005
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
year

Survival curves based on cohort (biobank) data



This reflects average survival across different recruitment times (2002-2019)

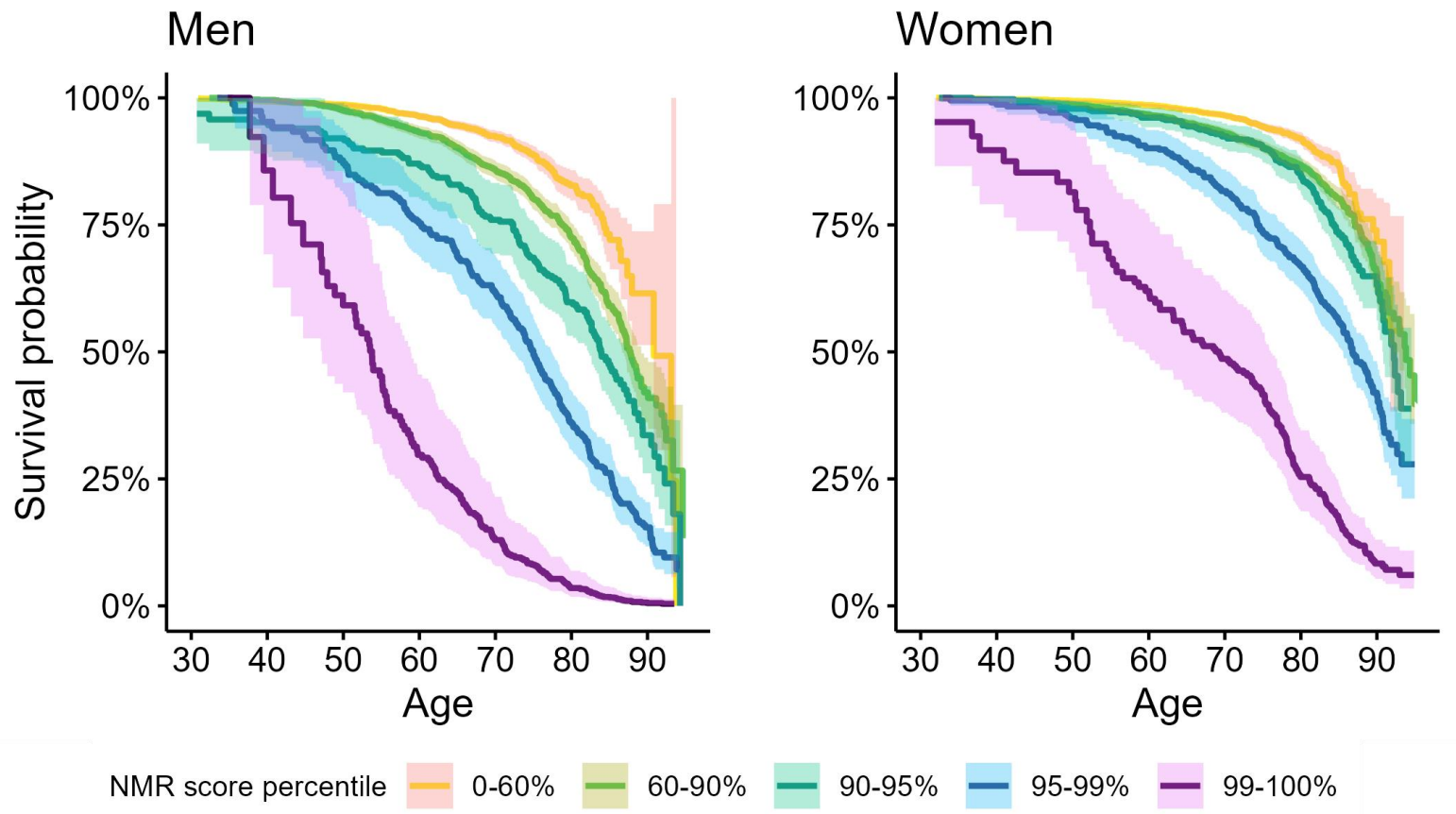
Survival curves based on cohort (biobank) data



Very much dependent on the cohort!

What data could be used for individual risk prediction and how?

Example: NMR-biomarkers and mortality (work with Mara Delesa-Velina)



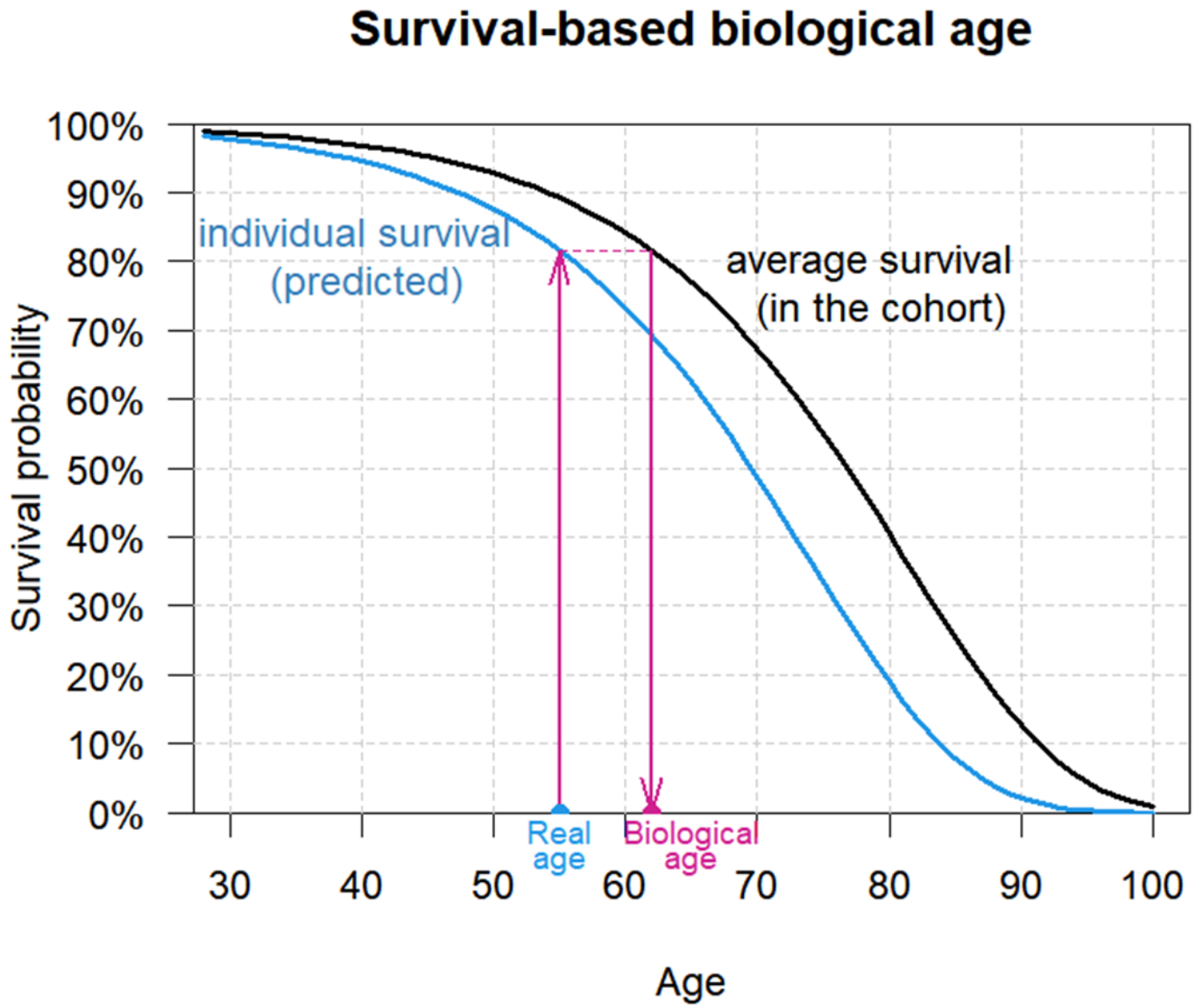
Difficult to use in the individual risk prediction

The risk score developed in the “old” cohort, predictions in the new cohort (validation set)

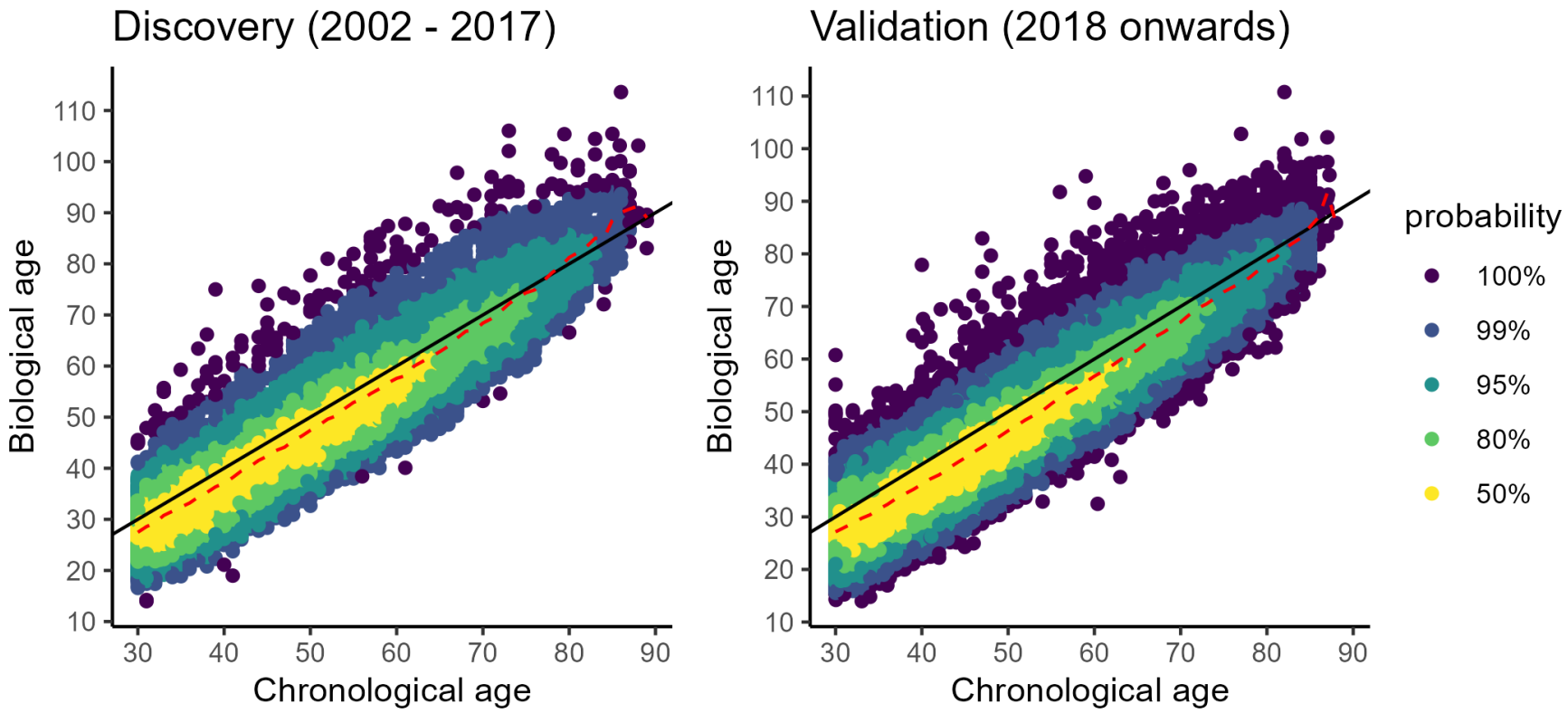
An alternative idea

Survival-based biological age: age, where the average survival probability in the population equals to the individual's current survival probability, given his/her covariate profile.

- How to compute? Parametric survival modeling (Gompertz, Weibull…)

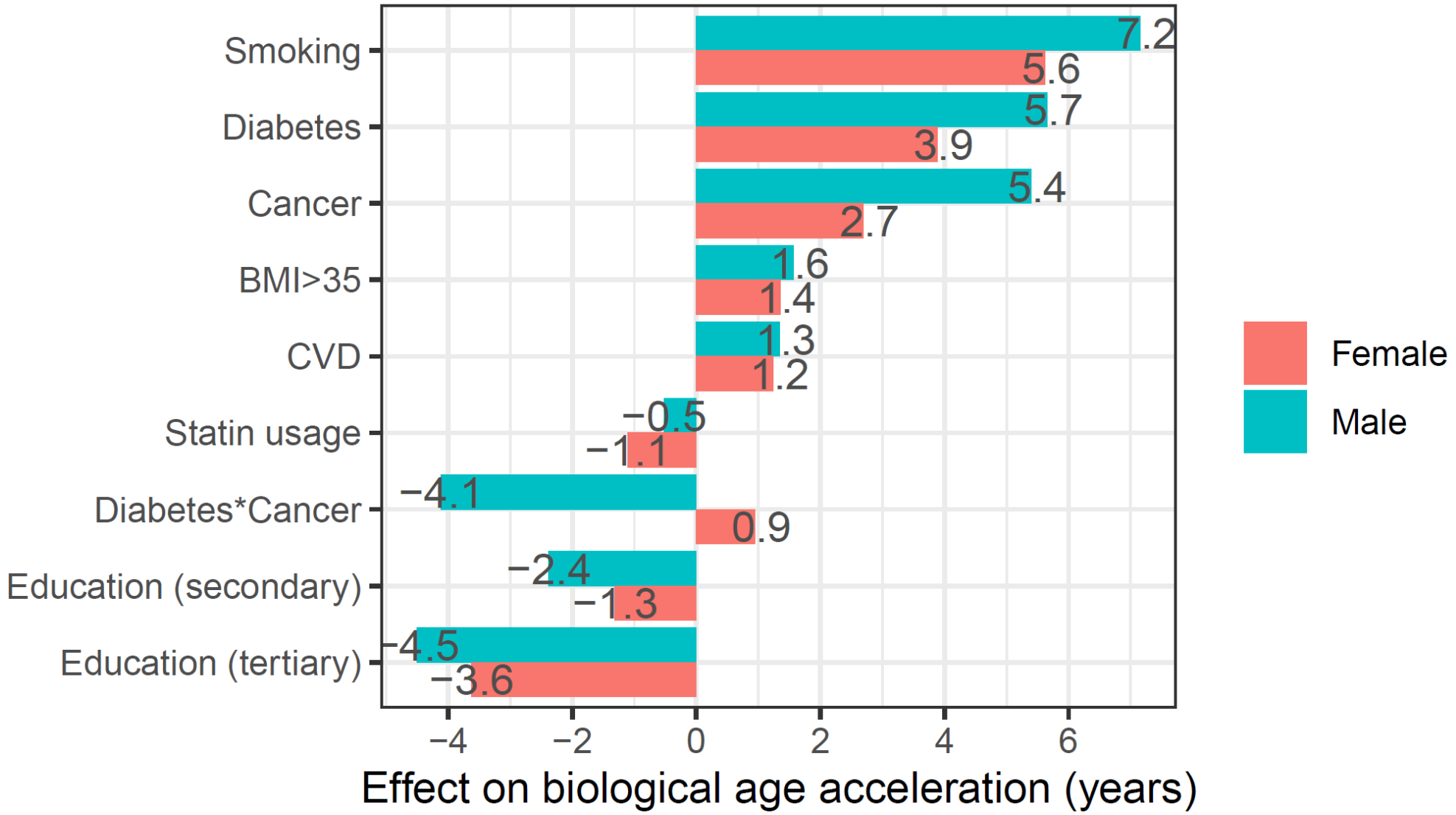


Biological age estimates



(Mara Delesa-Velina)

Effects of risk factors



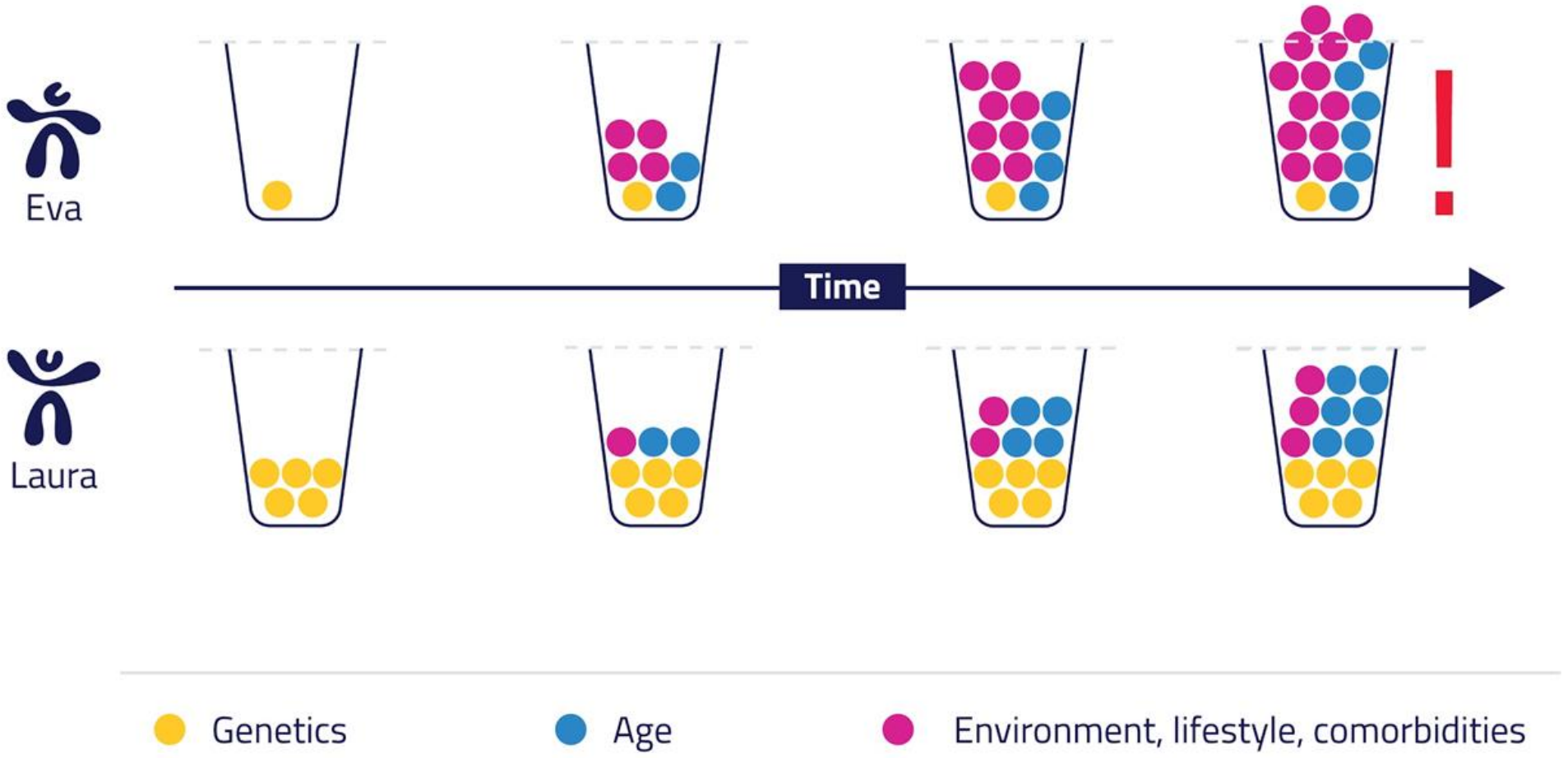
The risk of a disease

Are we interested in…

- Probability of ever getting the disease?
- Probability of getting the disease in X (5, 10, …) years?
- Probability of getting the disease before age A (60, 70, …)?

Probability of disease is often expressed as $P(T \leq t)$ and estimated as cumulative incidence:

Risk prediction for common complex diseases (the biobank view)



Personalized risk prediction for common complex diseases has existed long before biobanks

Common risk prediction algorithms:

- Coronary heart disease: SCORE, PCE, QRISK
- Diabetes: FinDRisc, QDiabetes, …
- Cancer: QCANCER…

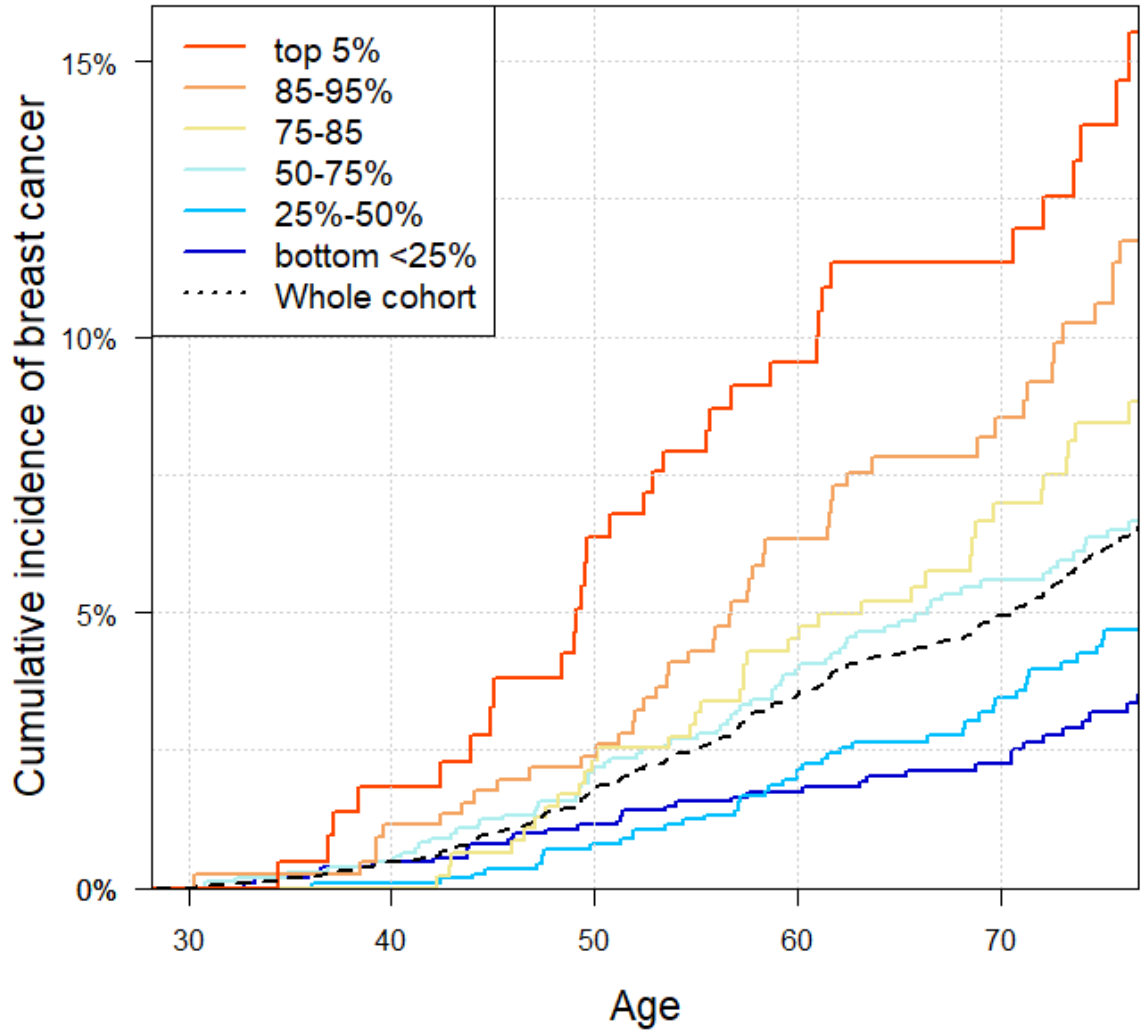
- Our aim is to add the genetic component to the established risk algorithms (calibrated for the Estonian data)

Is it worth adding the genetic component to the risk prediction algorithm?

- Genetic risk component summarized as the **Polygenic Risk Score (PRS)**
- Need to select the best PRS among alternatives (PGS Catalogue!)

- Does it explain a meaningful amount of variability?
- Need to validate in the (sub)cohort that was not included in the PRS development process!

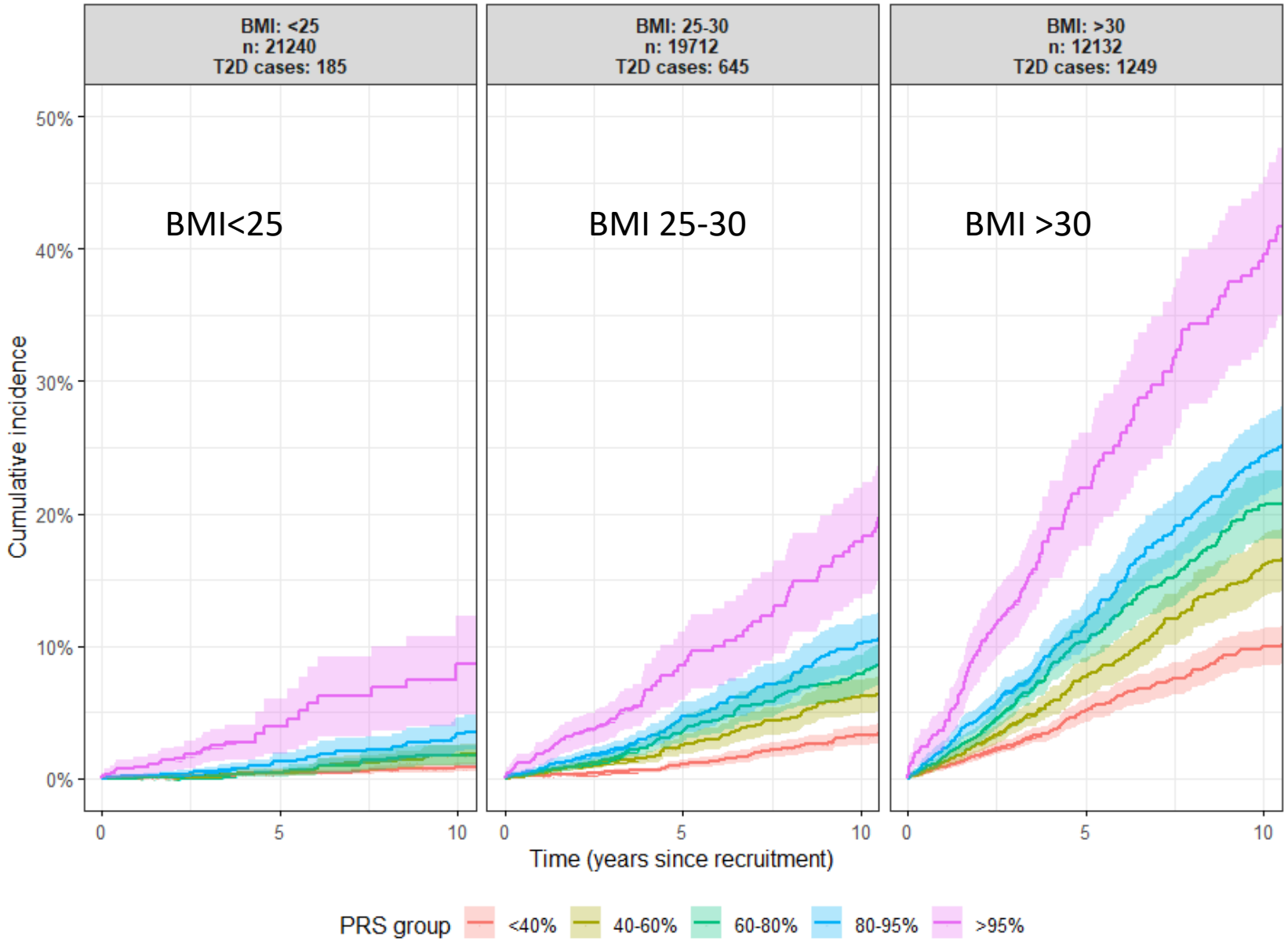
MetaGRS (combining 2 PRS-s) for incident Breast Cancer in the Estonian Biobank cohort



Cumulative incidence: $P(T \leq t)$

Läll et al, 2019, BMC Cancer

Type 2 Diabetes



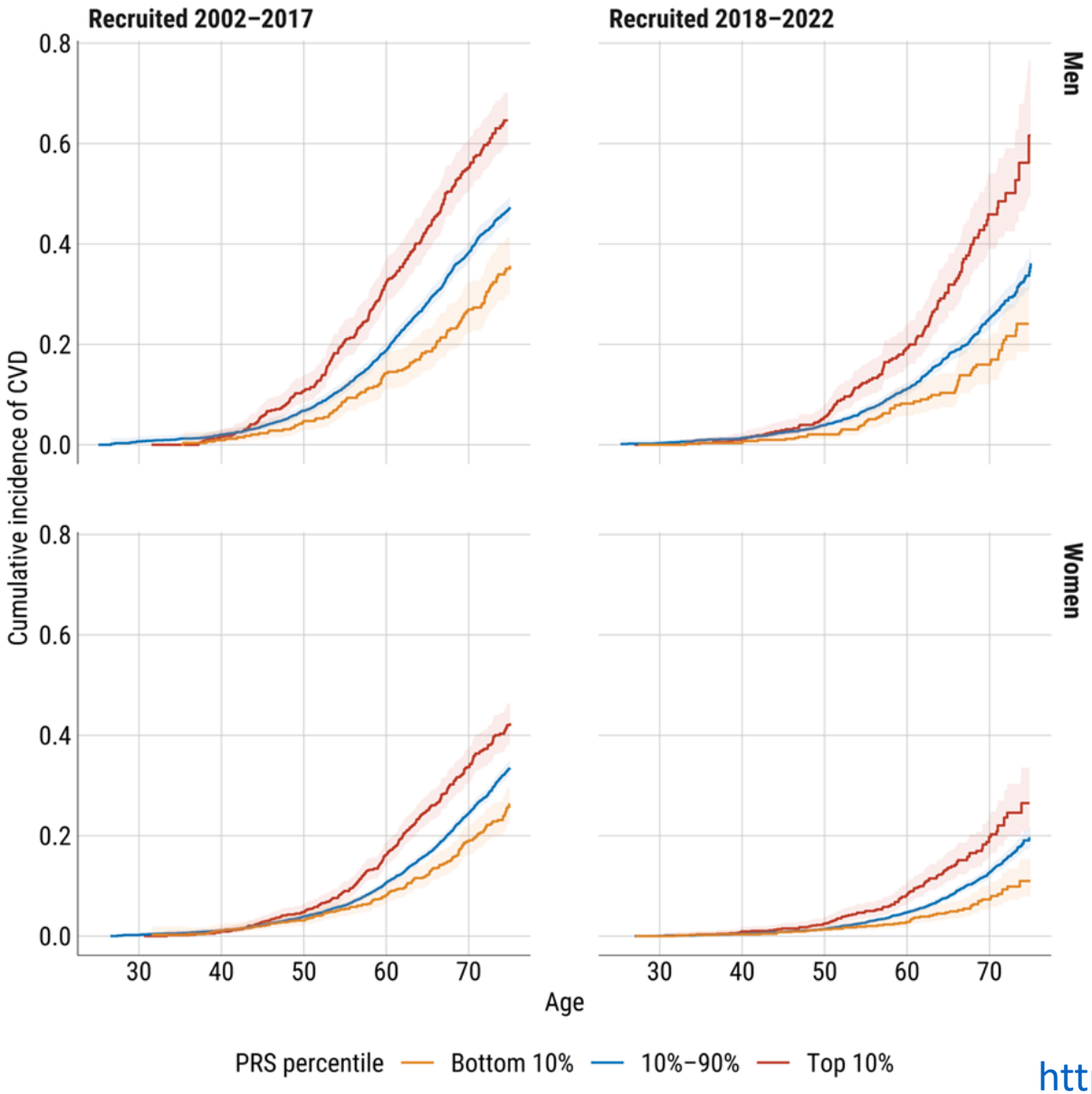
Top 5% PRS

Bottom 40% PRS

Also:
Läll 2017,
Gen.Med

Slide: Karmel Teder

Coronary Artery Disease



Again, we see different risks in the two cohorts

(Tuuli Puusepp)

https://www.medrxiv.org/content/10.1101/2025.04.02.25324383v1

Challenges

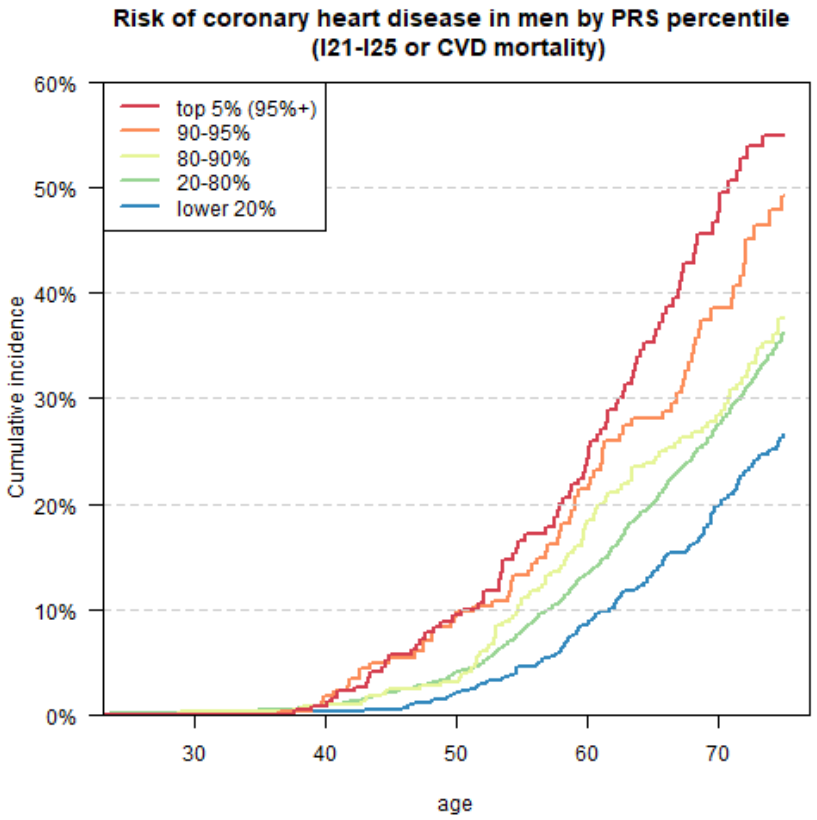
- Need to move from effect estimation and testing to absolute risk prediction

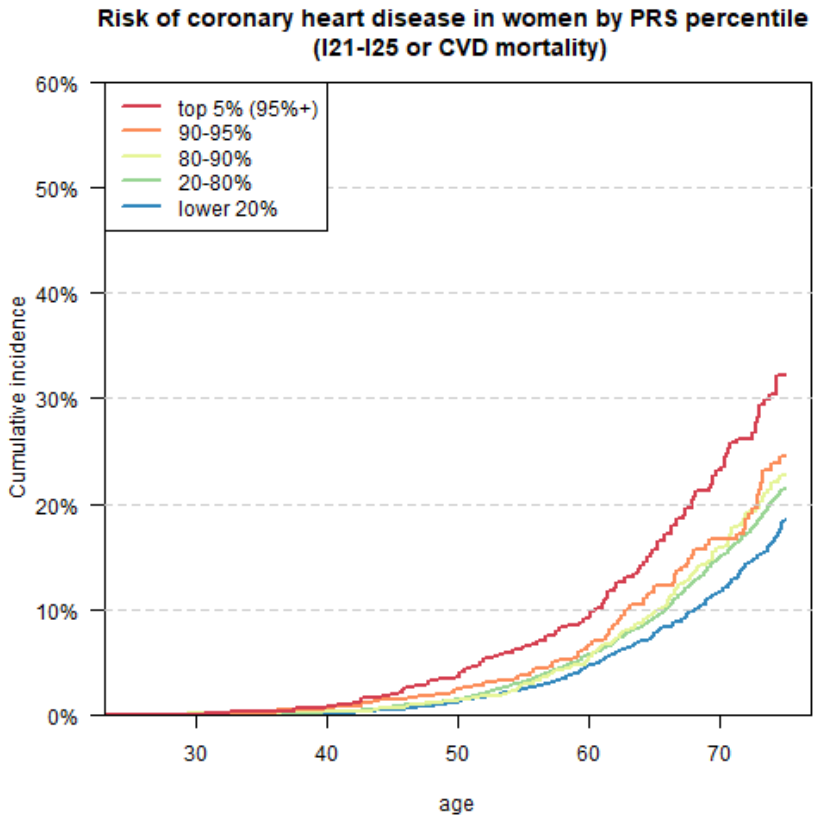
- What is the best way to communicate personalized risks?

Main steps of developing an algorithm for risk prediction and communication: the example of cardiovascular risk prediction in the Estonian Biobank

- Statistical modeling to assess the effects of risk factors and to develop the optimal predictive model
- Developing a predictive tool for absolute risk
- Finding the best way to communicate the risks

Assessing the effect of the PRS in the Estonian Biobank (33082 men and 74629 women of age 18-80) …





Kaplan-Meier curves illustrate the average effect size, but they are…

- not adjusted for other risk factors
- not easily implemented for out-of-sample predictions (nonparametric)

PRS: metaGRS from Innouye et al. (JACC, 2018)

The proportional hazards model

The hazard function:

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

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: when a covariate changes by a constant, the hazard is multiplied by a constant

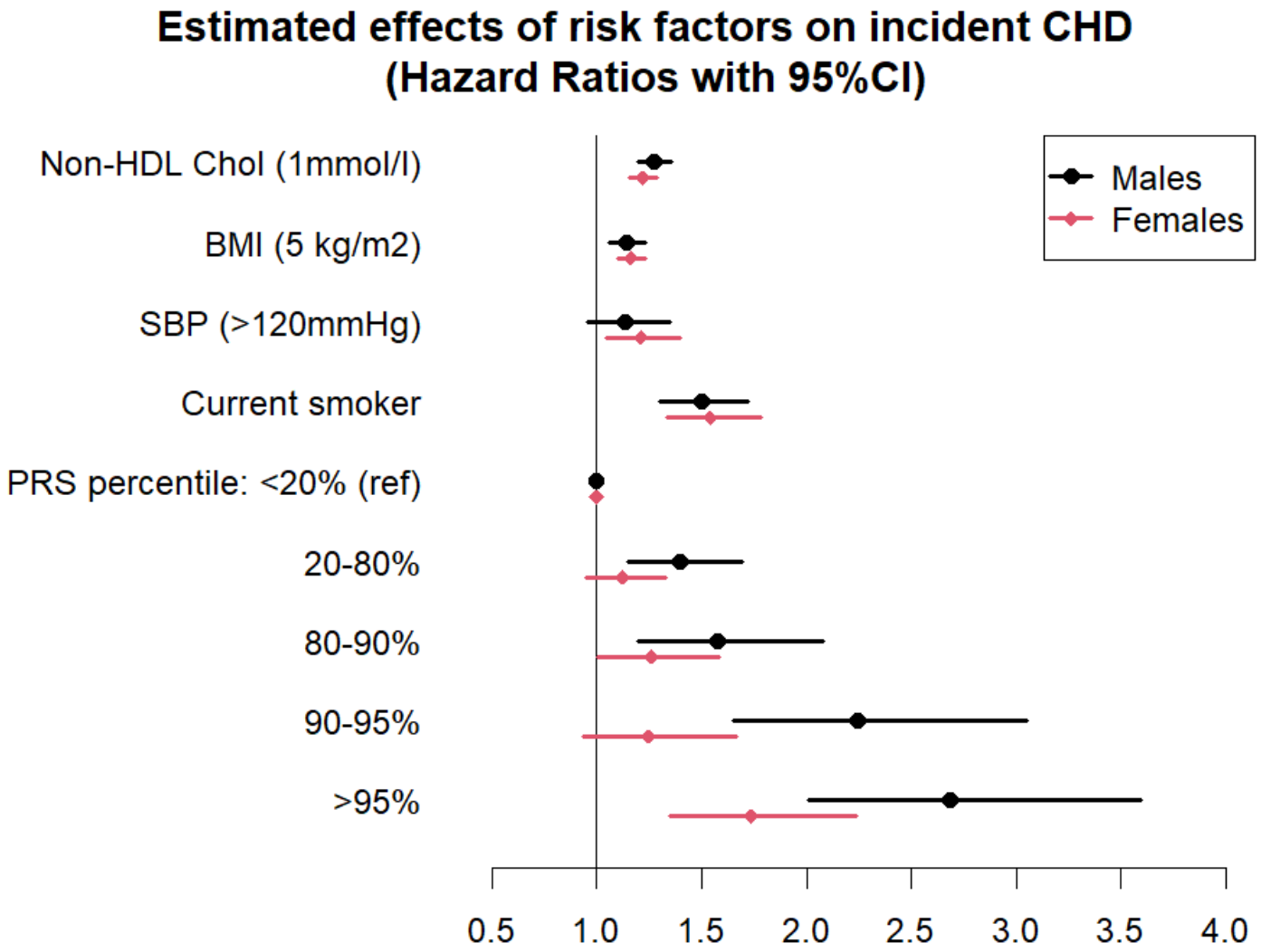
The Cox proportional hazards model

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

- A semiparametric model – does not use the real survival times, but the ranks of survival time and information on the individuals at risk at each event time.

- The baseline hazard h_0 will not be estimated – thus the model does only provide estimates of the covariate effects, but does not allow direct prediction of hazards or survival times

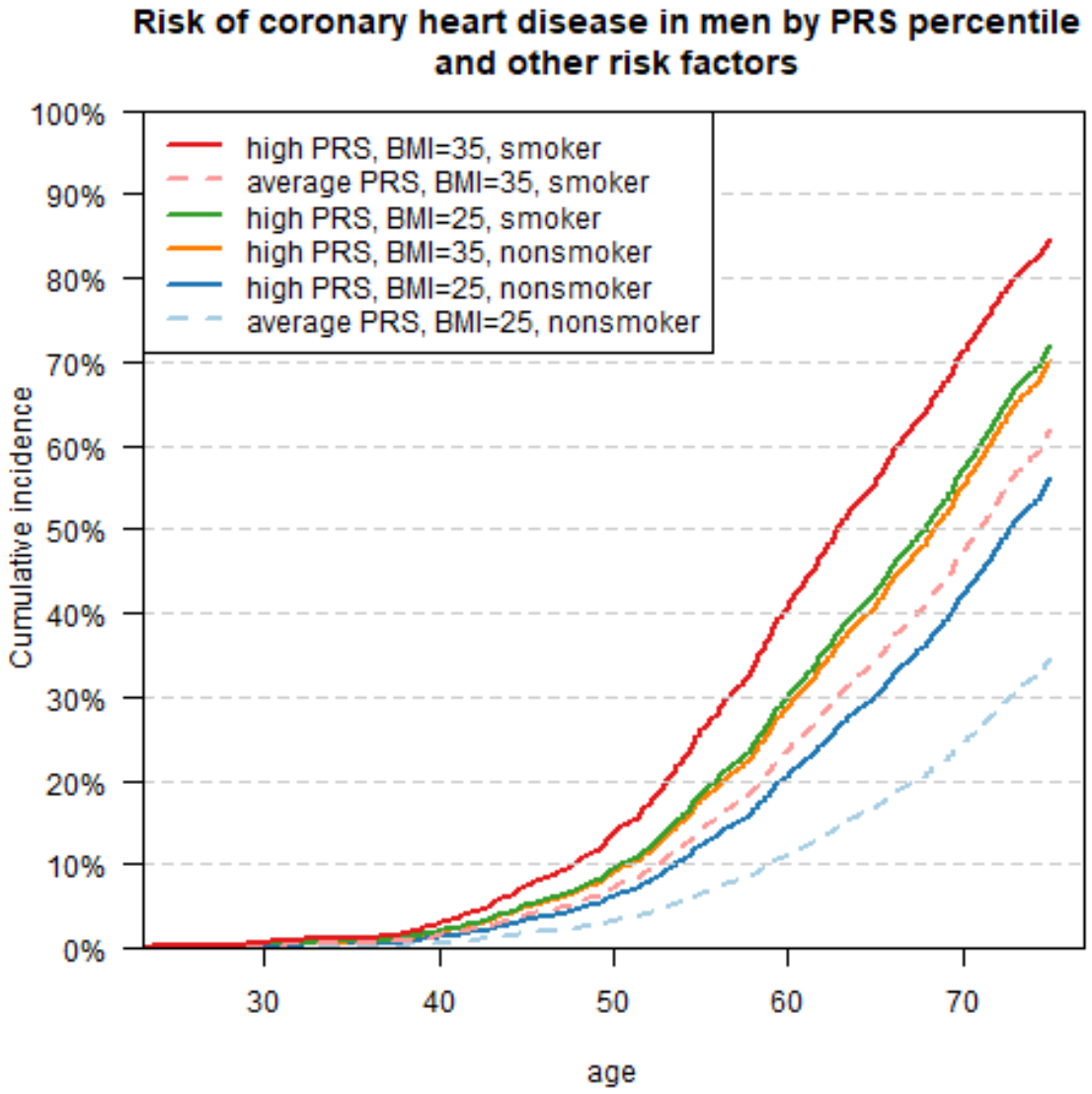
Sir David Cox
(Oxford, UK)
Paper on prop.haz models in 1972



We see the adjusted estimates and their range of uncertainty (CI-s), but …

- HR-s are not so easily understood by general public
- need baseline hazard estimates for actual risk prediction

Cumulate incidence estimates from the Cox model…



… good to illustrate the effects of risk factors on absolute risks, but…

- it still relies on nonparametric baseline hazard estimation
- does not account for the fact that an individual does not have CVD at baseline (varying age)

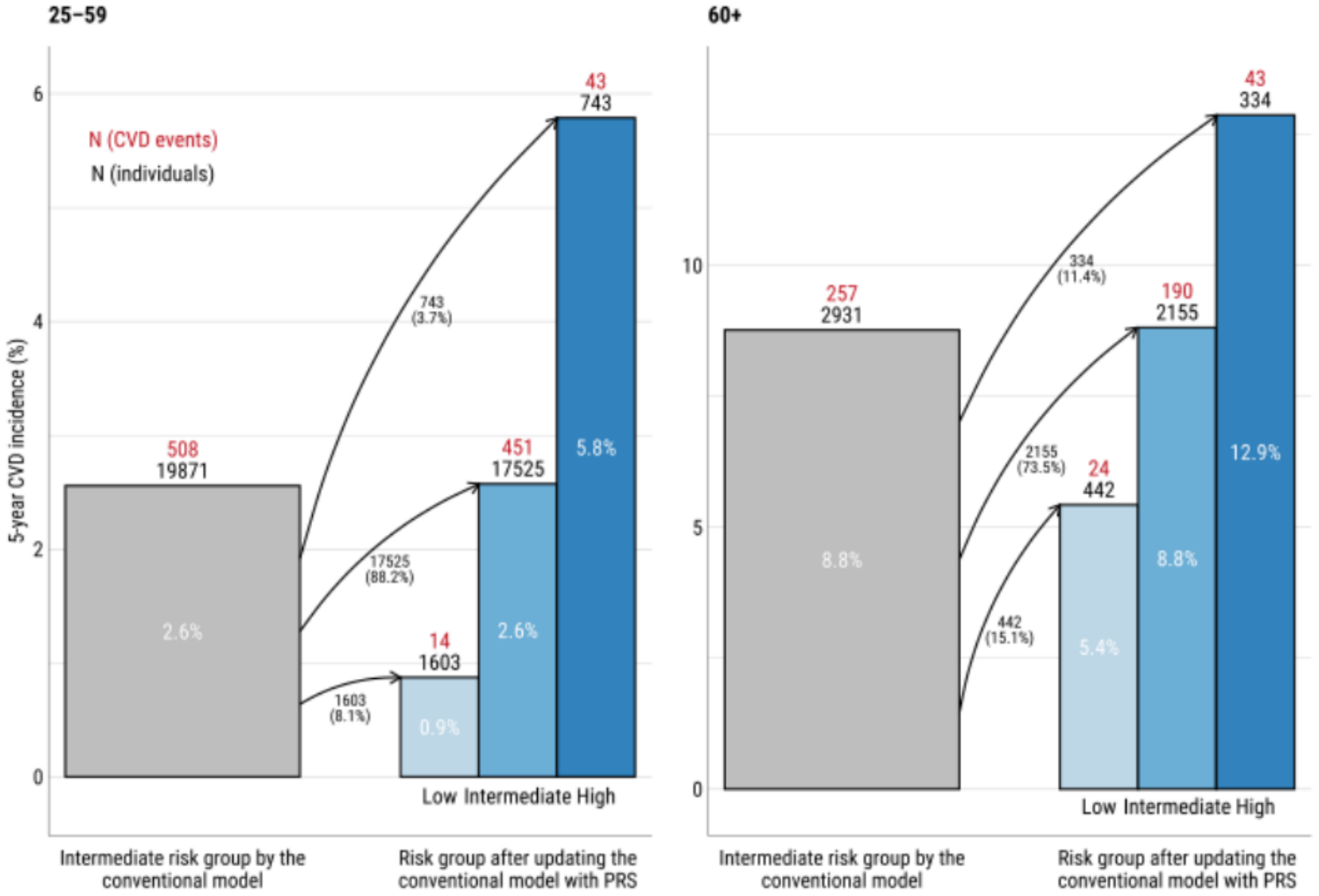


Figure 4. Reclassification of individuals initially categorized as intermediate risk for 5-year CVD incidence using the conventional model. Arrows indicate the movement of individuals between categories, with corresponding percentages representing the proportion of individuals reclassified.

Tuuli Puusepp et al. (2025)

What do we actually need to estimate for individual risk prediction?

- Feedback on risks is relevant for the individuals who are currently disease-free
- Often, a 10-year risk is a meaningful quantity to be estimated for risk stratification/feedback purposes
- Instead of the popular Cox model, parametric modeling of time-to-event censored outcomes deserves more attention, providing more straightforward tools for risk prediction

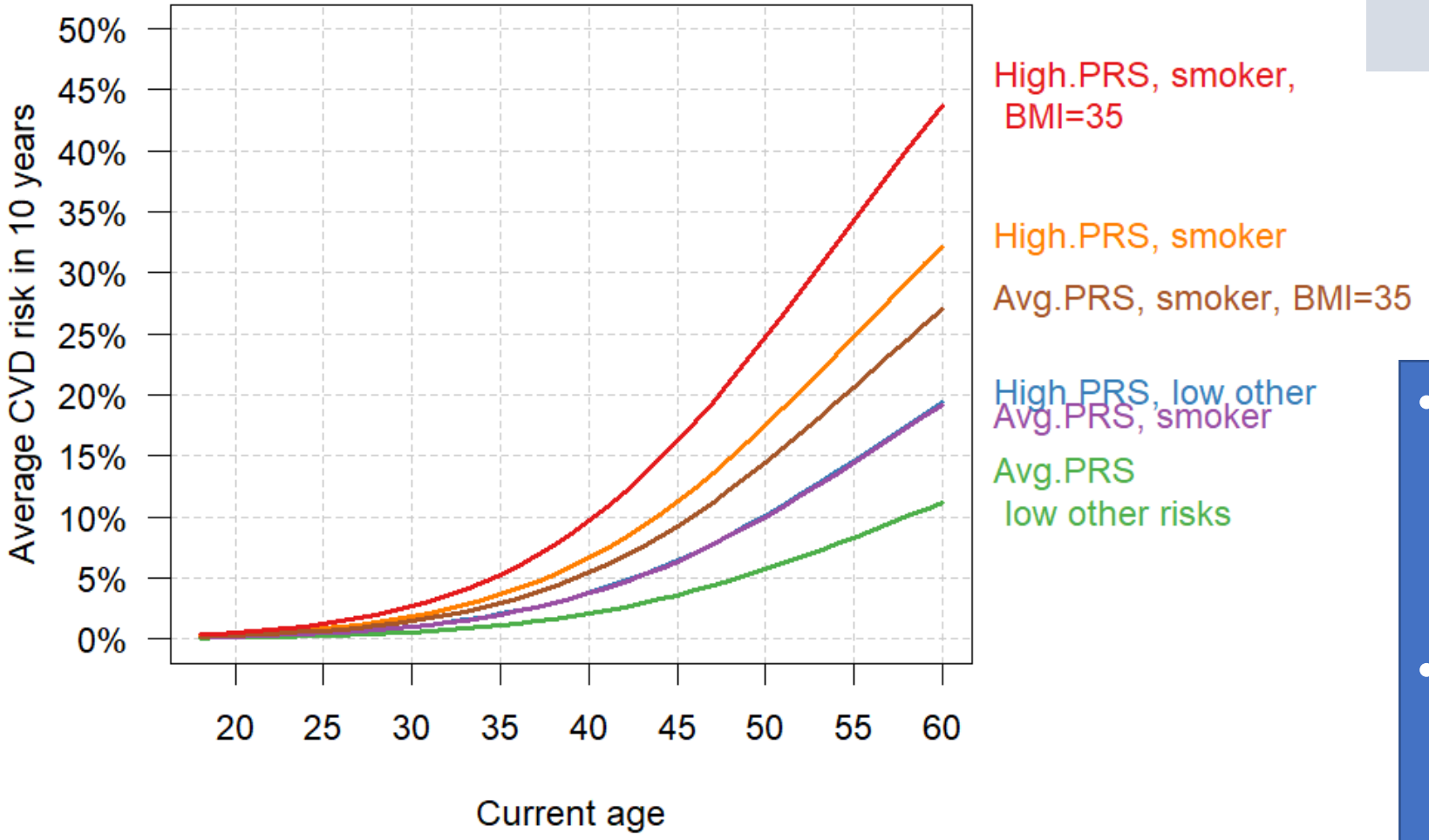
An alternative: a Weibull model for 10-year risk

10-year risk:

$$P(T_i < 10) = 1 - e^{-(10\lambda_i)^{\gamma}}$$

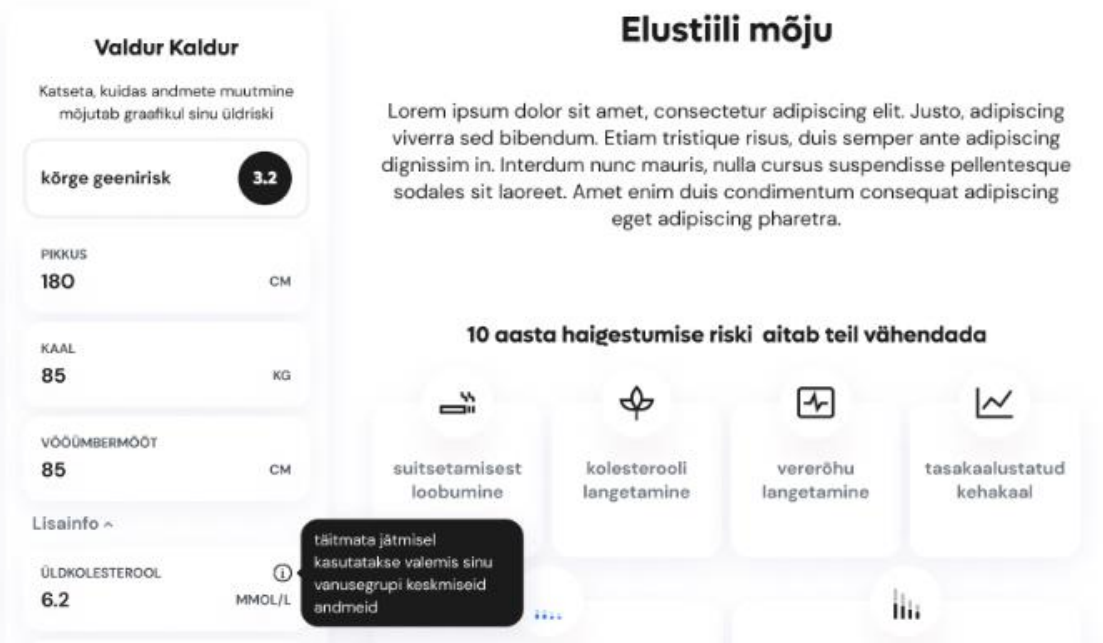
From the model

Predicted 10-year risks for men in the Estonian Biobank



- Validated in an independent cohort, also against robust (Cox) alternatives
- Currently being tested in a pilot study (INTERVENE)

Implementation in the portal



Mati Maasikas

Katseta, kuidas andmete muutmine mõjutab graafikul sinu üldriski

kõrge geenirisk **2**

PIKKUS
180 CM

KAAL
75 (85) KG

Lisainfo

ÜLDKOLESTEROOL
5.5 (7) MMOL/L

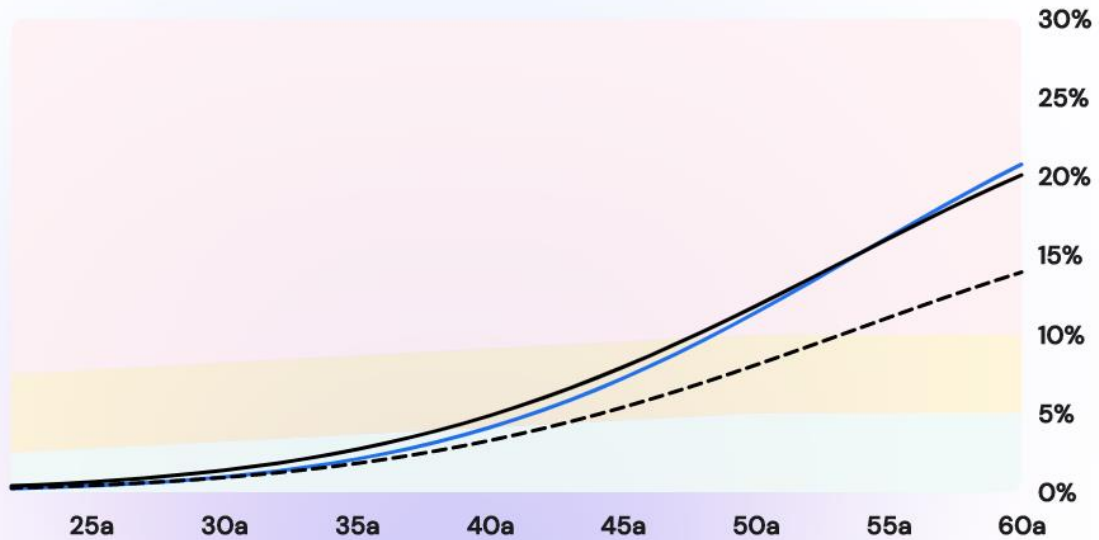
HDL - KOLESTEROOL
2 MMOL/L

SÜSTOOLNE VERERÕHK
115 (120) MMOL/L

SUITSETAN
EI JAH

Salvestamiseks mine profiilile →

Koguriski muutumine aastatega



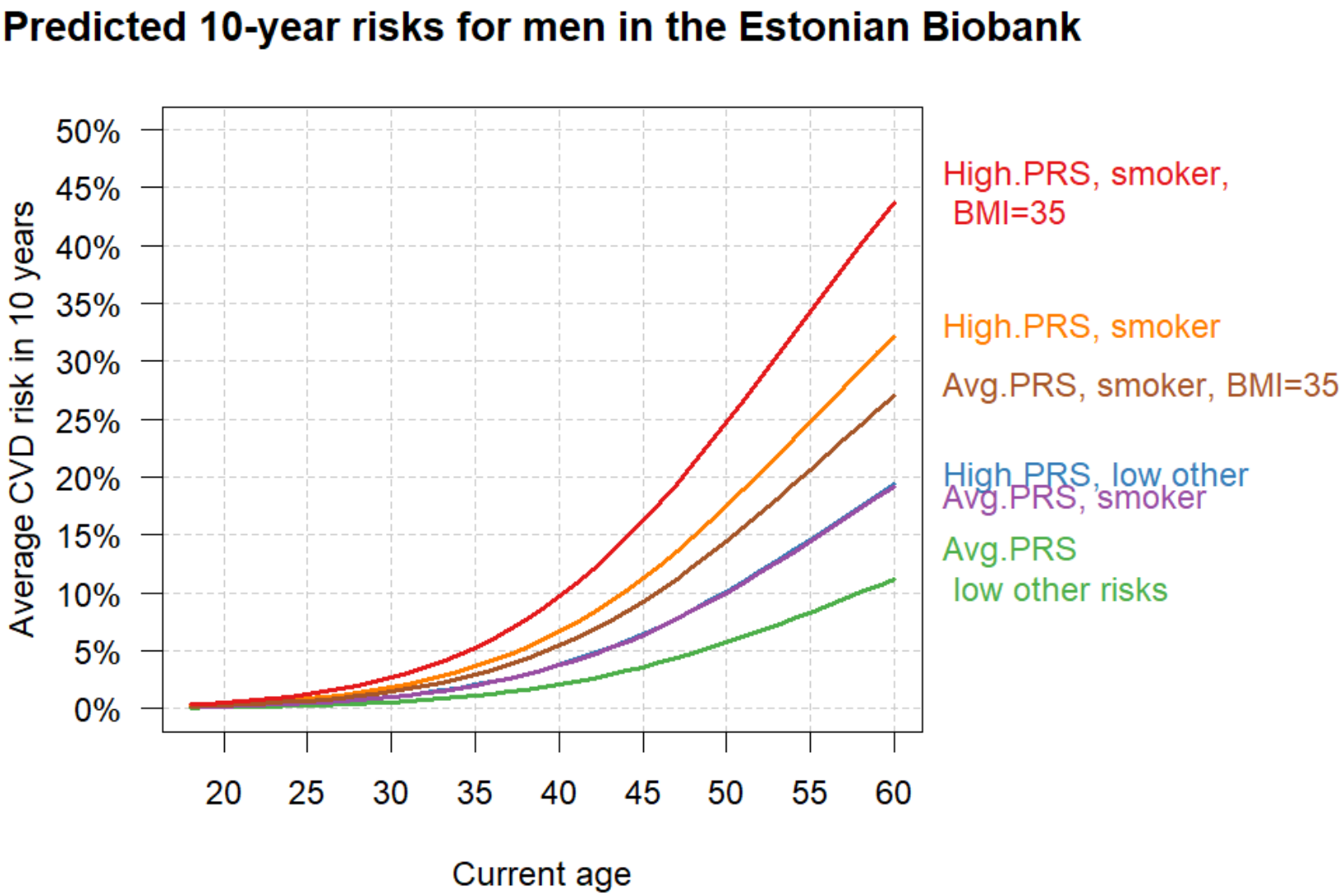
— Teiega samast soost ja samas vanuses inimeste riski hinnang
— Riski hinnang Teie praeguste tervisenäitajate alusel
--- Riski hinnang muudetud tervisenäitajate alusel

Koguriski saad maandada elustiili muutmisega

Kokku on Sul võimalik haigestumise riski vähendada	0.4% → 0.3%
Vähenda riski haigestuda 70. eluaastaks	20% → 14%

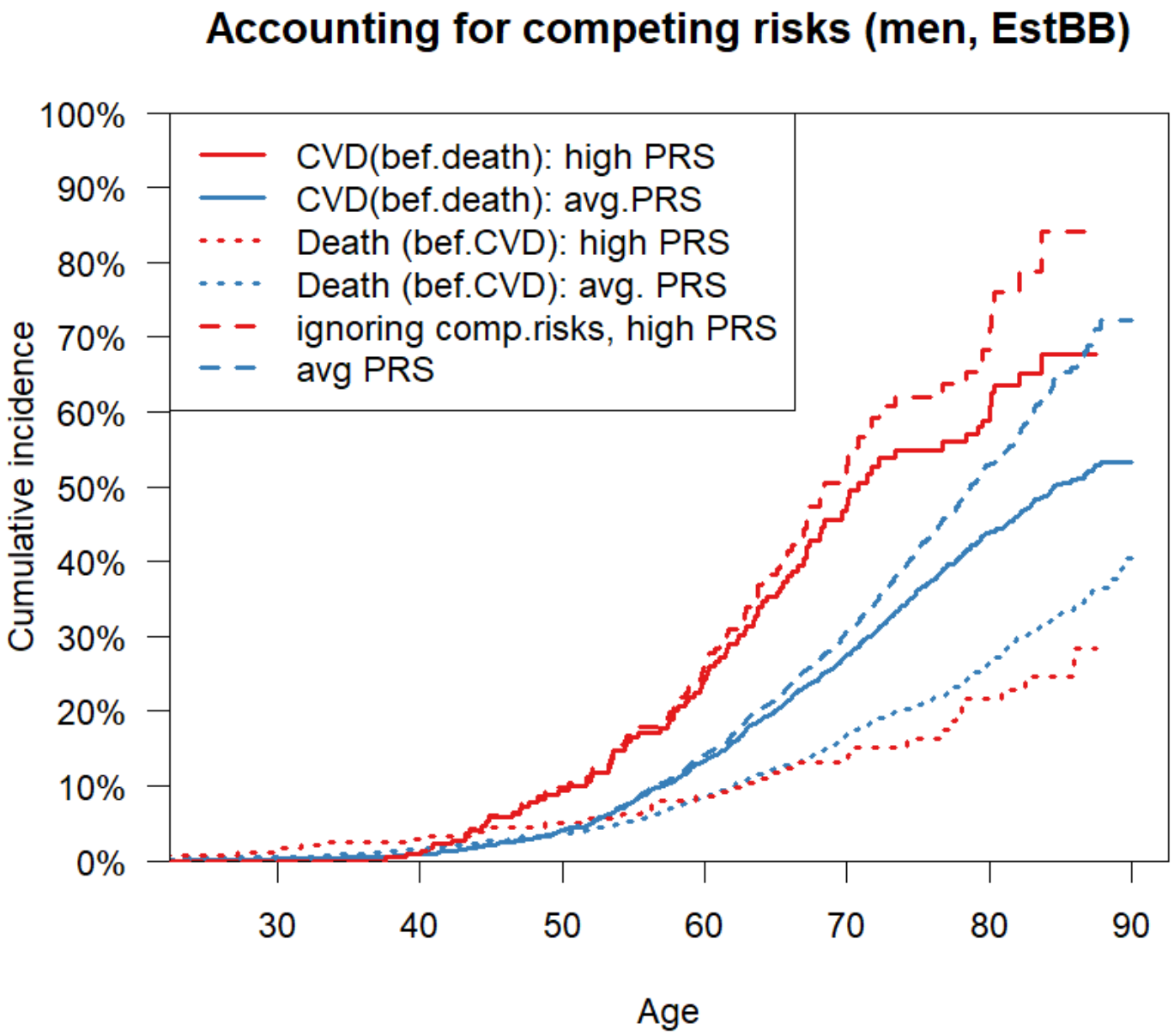
Further challenges: work in progress

The current model is not ideal for everyone…



Challenges:

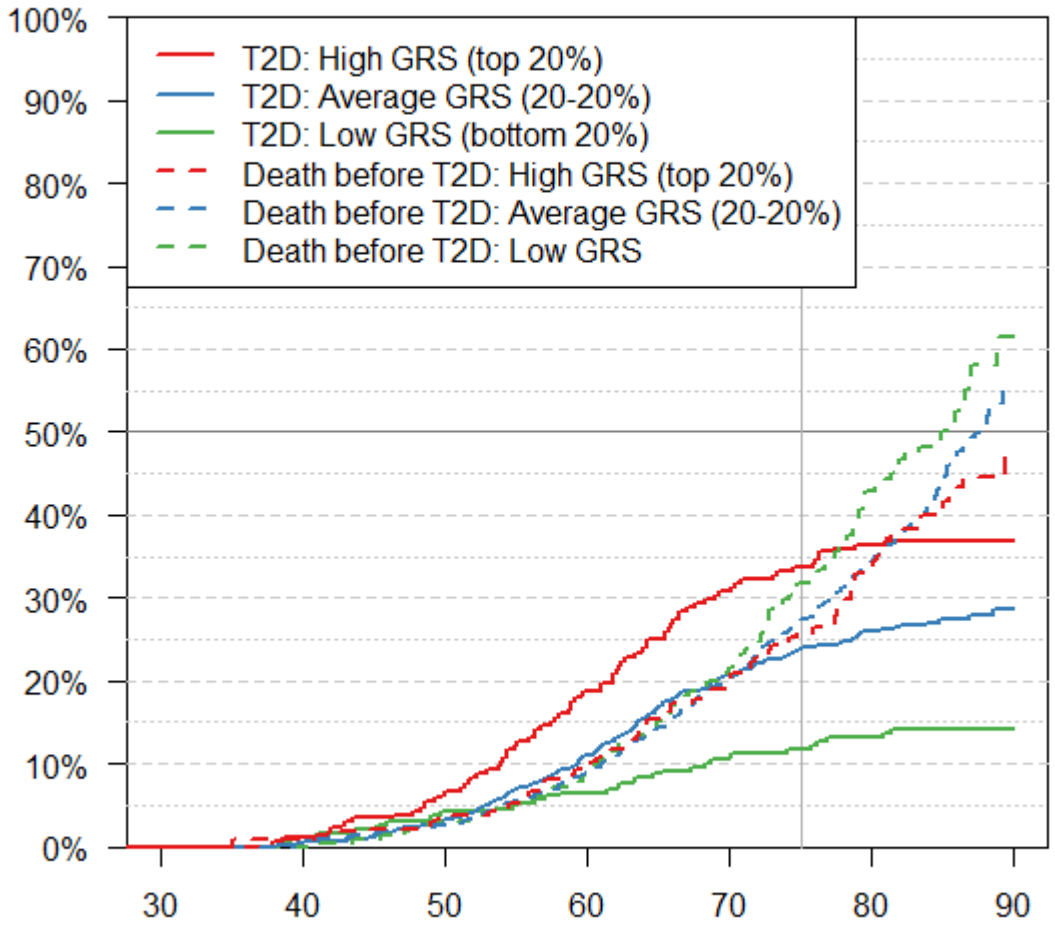
- Feedback to old people (competing risks)
- Feedback to young people (very low risks)



- Ignoring competing risks biases the risk predictions in old age
- However, communication of competing risks could be challenging
- Also – not so easy to implement in a Weibull model

Incident Type 2 Diabetes (T2D) and polygenic (genetic) risk score (GRS=PRS): cumulative incidence on age scale (men with BMI=25..30)

Treating death as a competing event



How to do it correctly?

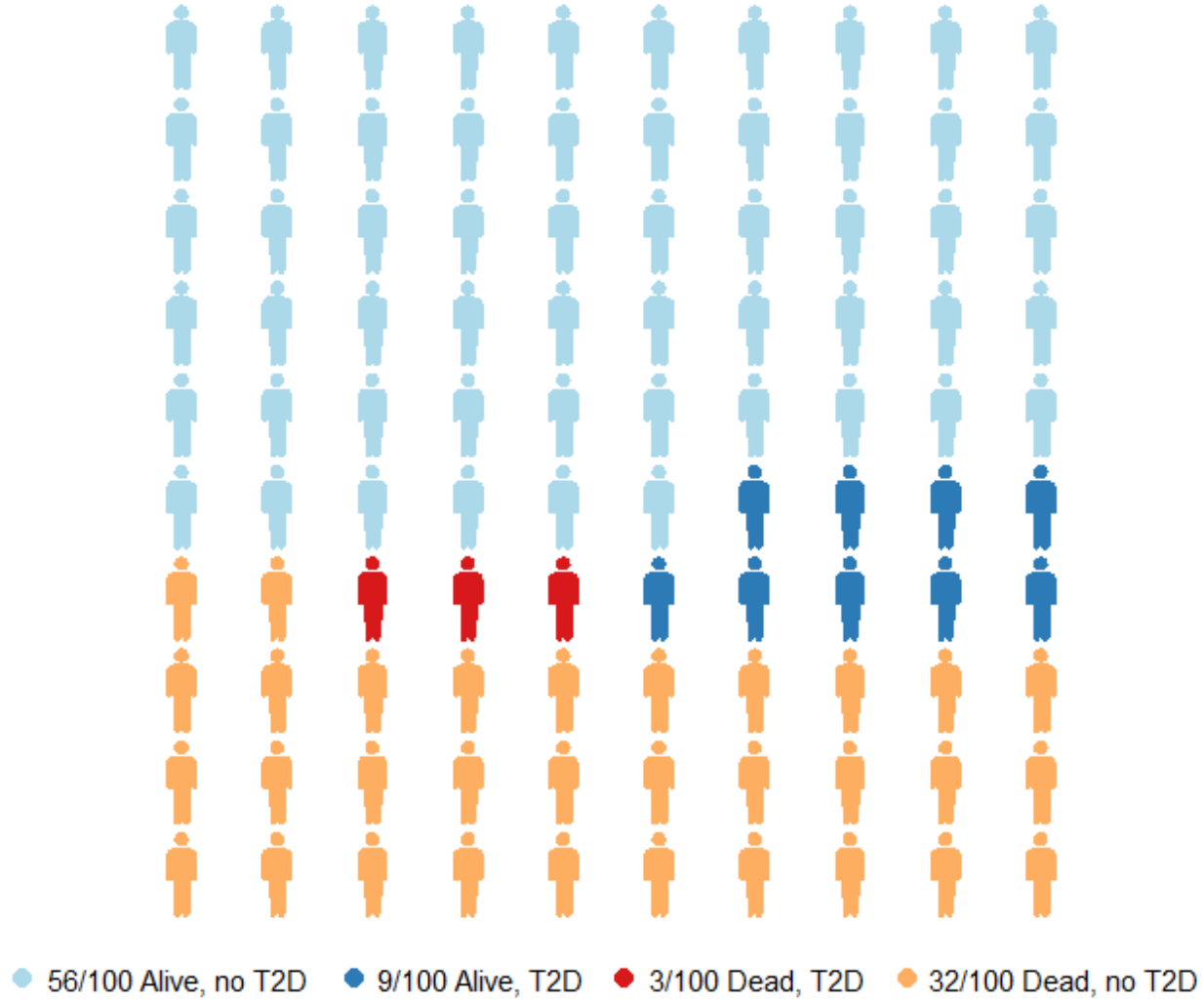
By age 75, in the high genetic risk group:
25% have died without a T2D diagnosis
34% have had a T2D diagnosis
41% would be alive and free of T2D
(note: sum=100)

In the low genetic risk group:
32% died without T2D
12% have had a T2D diagnosis
56% alive, free of T2D

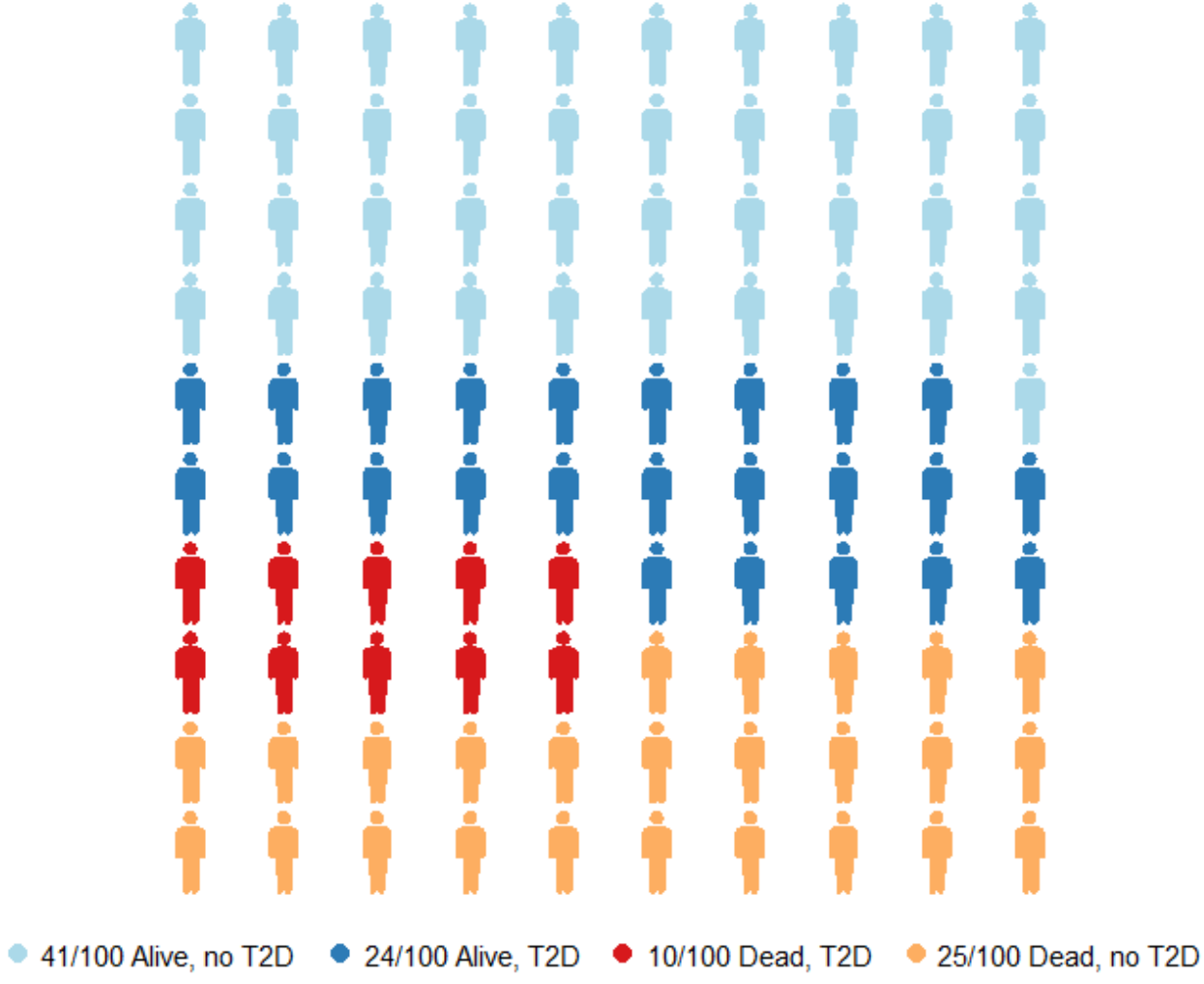
Is mortality higher in the low-PRS group?

An example with T2D: scenarios by the age of 75

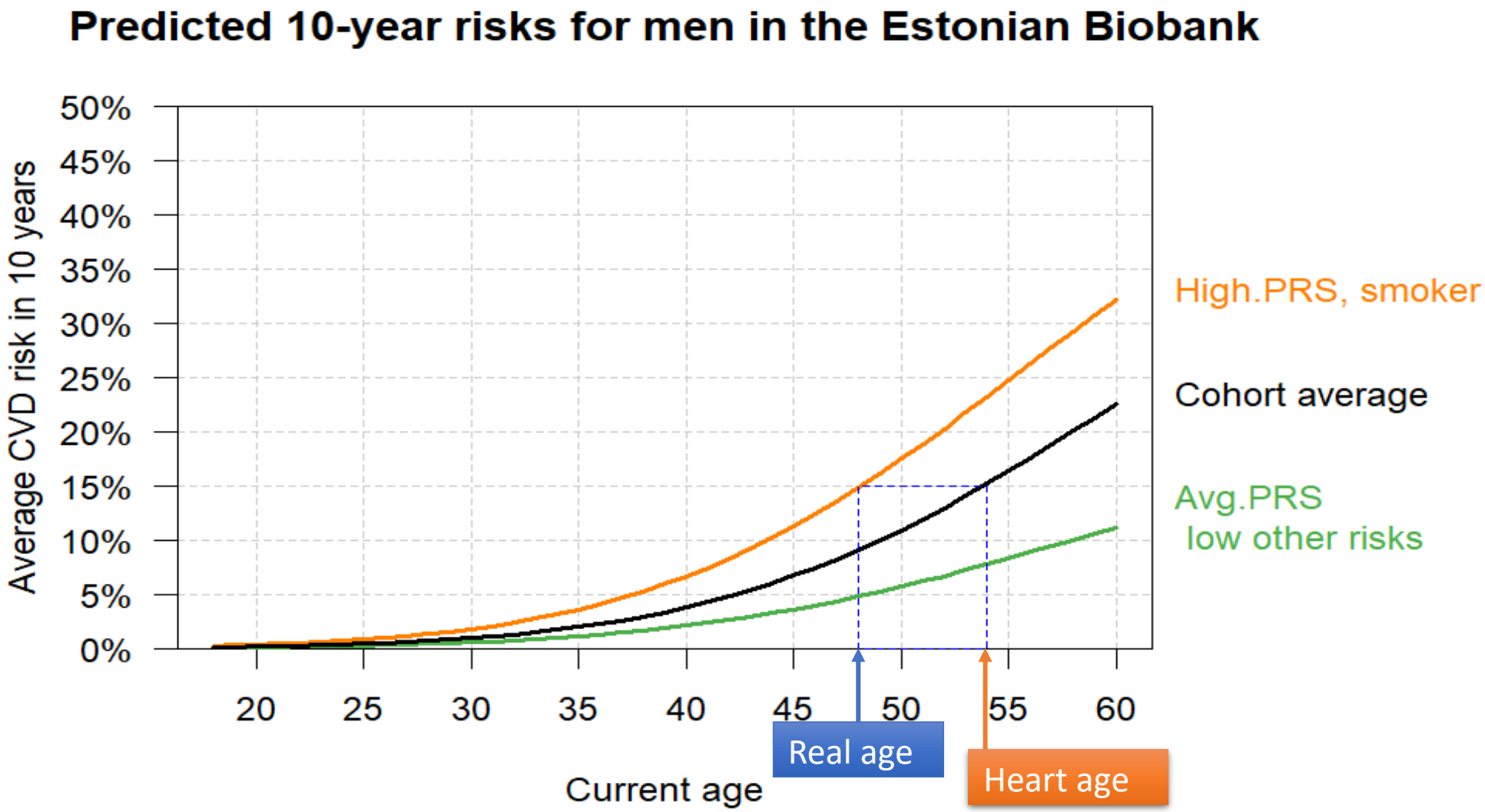
Low PRS



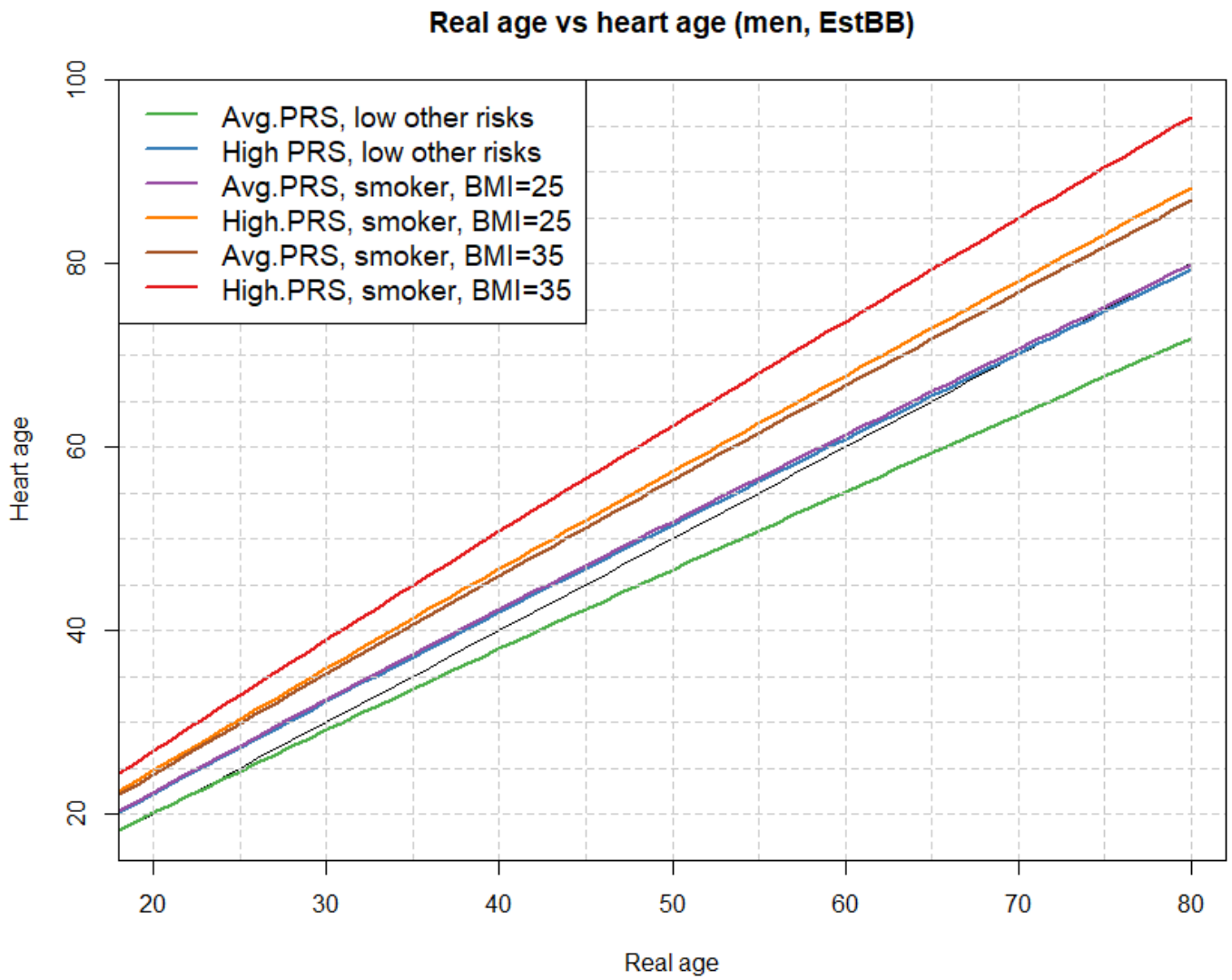
High PRS

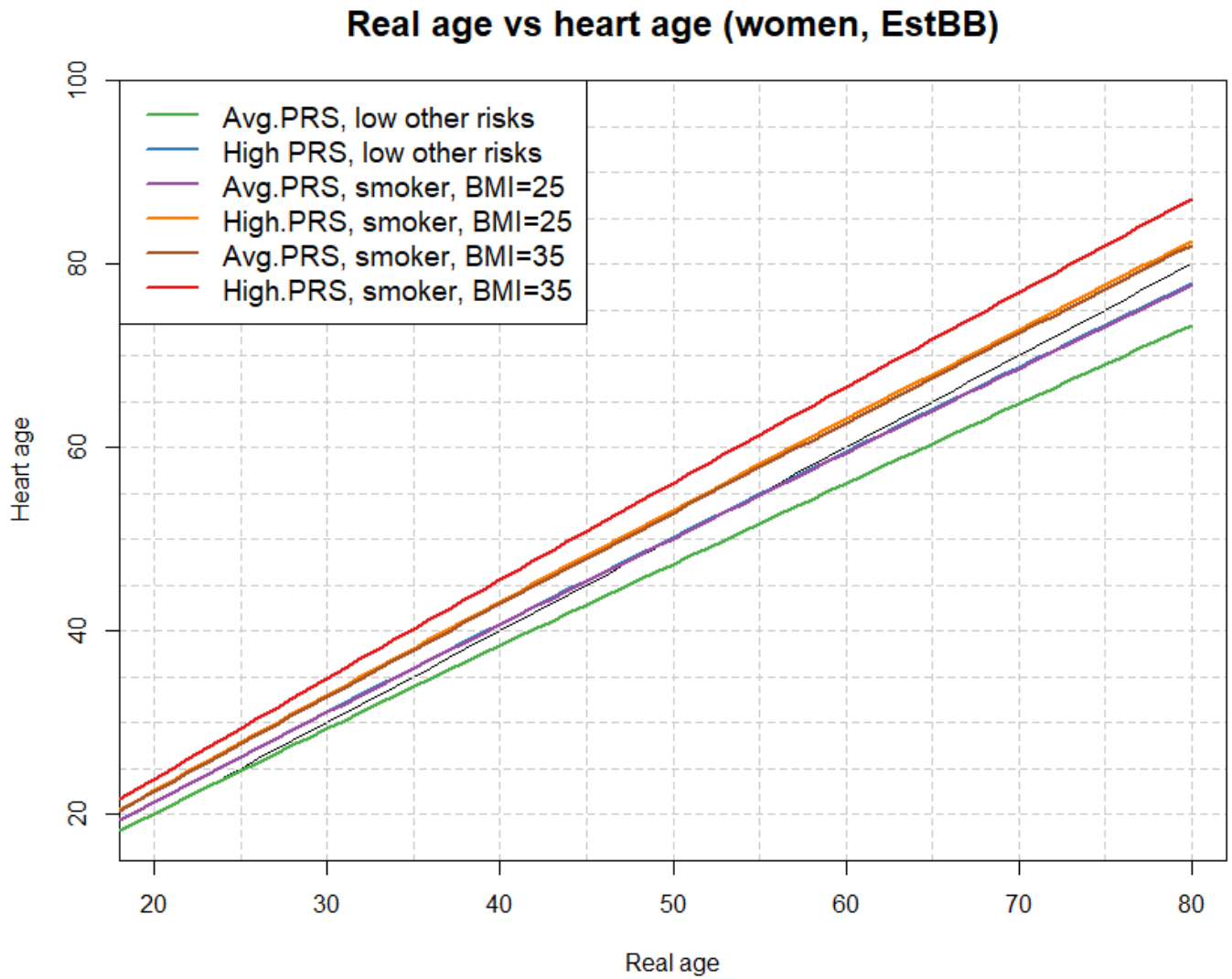


Application of the biological age for cardiovascular risk: heart age



Real age vs heart age (men, EstBB)




Real age vs heart age (women, EstBB)
Avg.PRS, low other risks
High PRS, low other risks
Avg.PRS, smoker, BMI=25
High.PRS, smoker, BMI=25
Avg.PRS, smoker, BMI=35
High.PRS, smoker, BMI=35
Heart age
Real age
20
40
60
80
100
20
30
40
50
60
70
80

Summary

- It is not easy to speak about risks, as risk itself is confusing concept
- Risk of death is “easier” to, as death occurs only once. Still, there is the problem of reference cohort and non-identifiable age-period-cohort effects (Lexis!)
- Biological age may be easier to communicate
- Disease risk estimation requires translation from model parameters to absolute risks. For short-term prediction, in case of no considerable competing risks, there are several approaches available.
- In case of competing risks – they can be taken into account in the analysis, but the communication task is still tricky.
- There are still many challenges ahead