Next weeks

- Final Exam: April 11 2:00 – 5:00 pm SS 1085
  - 4 questions
  - one theory question
  - one applied question
  - one question from HW
  - one question about a study
  - one question with computer output
- SM: 9.1, 9.2.1, 9.2.2 (to end p.431), 9.3.1, 9.4.2; 10.1, 10.2, 10.3, 10.4, 10.6, 10.7.1 (skip p.529-530), 10.7.2, 10.7.3, 10.8.1, 10.8.2
  (skip log rank test, time-dependent covariates)
- C& D: from slides only – Ch 1, 2, 7.2, 7.3, 6.5
- Office Hours: April 8, 9, 10; 3 – 5
- HW 4: due April 11

Today

- Model formulation CD Ch. 6
- Regression with survival data
- In the news: biomarkers and death; PLOS 1 paper

- April 4: Questions re any HW, re grading, re 2012 final test; Summary of course notes

Today

- Regression with survival data
- In the news: biomarkers and death; PLOS 1 paper
- Model formulation CD Ch. 6
- April 4: Questions re any HW, re grading, re 2012 final test; Summary of course notes

Today

- Regression with survival data
- In the news: biomarkers and death; PLOS 1 paper
- Model formulation CD Ch. 6

- April 4: Questions re any HW, re grading, re 2012 final test; Summary of course notes

David A Sprott 1930 – 2013

Today

- Regression with survival data
- In the news: biomarkers and death; PLOS 1 paper
- Model formulation CD Ch. 6

- April 4: Questions re any HW, re grading, re 2012 final test; Summary of course notes

Empirical models

CD Ch. 6.5

- in many fields of study the models used as a basis for interpretation do not have a special subject-matter base
- rather represent broad patterns of haphazard variation quite widely seen
- this is typically combined with a specification of the systematic part of the variation
- which is often the primary focus
- modelling then often reduces to a choice of distributional form
- and of the independence structure of the random components

... empirical models CD Ch. 6.5

- functional form of the probability distribution sometimes critical, for example where an implicit assumption is involved of a relationship between variance and mean: geometric, Poisson, binomial
- the simple situations that give rise to binomial, Poisson, geometric, exponential, normal and log normal are some guide to empirical model choice in more complex situations
- In some specific contexts there is a tradition establishing the form of model likely to be suitable
- illustration: financial time series – $Y(t) = \log\{P(t)/P(t-1)\}$ has a long-tailed distribution, small serial correlation, large serial correlation in $Y^2(t)$
- illustration: a common type of response arises as the time from some clearly defined origin to a critical event
- often have a long tail of large values; exponential distribution is a natural staring point
- extensions may be needed, including Weibull, gamma or log-normal

... empirical models

CD Ch. 6.5

- often helpful to develop random and systematic parts of the model separately
- models should obey natural or known constraints, even if these lie outside the range of the data
- example $P(Y=1)=\alpha+\beta x$
- often use instead $\log\frac{P(Y=1)}{P(Y=0)}=\alpha'+\beta' x$
- however, β measures the change in probability per unit change in x
- in many common applications, relationship between y and several variables $x_1,\ldots x_p$ is involved
  - unlikely that the system is wholly linear
  - impractical to study nonlinear systems of unknown form
  - therefor reasonal to begin with a linear model
  - and seek isolated nonlinearities

... empirical models

CD Ch. 6.5

- often helpful to develop random and systematic parts of the model separately
- naive approach: one random variable per study individual
- values for different individuals independent
- more realistic: possibility of structure in the random variation
- dependence in time or space, or a hierarchical structure corresponding to levels of aggregation
- ignoring these complications may give misleading assessments of precision, or bias the conclusions

... empirical models

CD Ch. 6.5

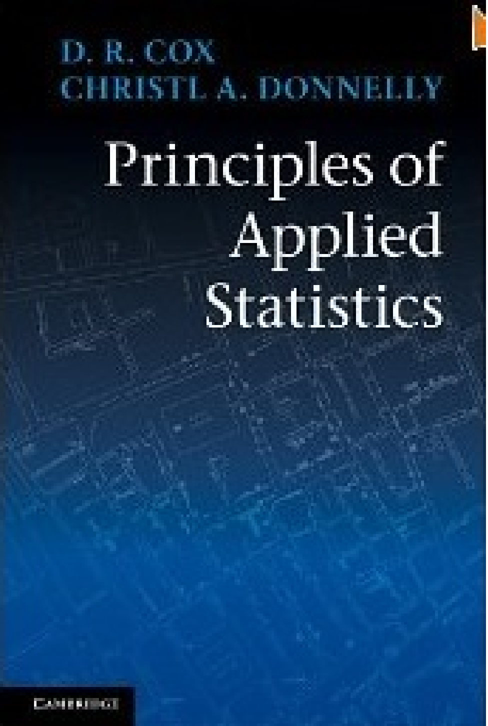
- example: standard error of mean $\sigma/\sqrt{n}$
- but, under mutual correlation, becomes $(\sigma/\sqrt{n})(1+\Sigma\rho_{ij})^{1/2}$
- if each observation correlated with k others, at same level, $(\sigma/\sqrt{n})(1+k\rho)^{1/2}$

-
  ```
  0.1  0.2  0.4  0.8
  --------------------------
  1.14 1.26 1.48 1.84
  1.18 1.34 1.61 2.05
  1.22 1.41 1.73 2.24
  1.26 1.48 1.84 2.41
  1.30 1.55 1.95 2.57
  1.34 1.61 2.05 2.72
  ```

... empirical models

CD Ch. 6.5

- important to be explicit about the unit of analysis
- has a bearing on independence assumptions involved in model formulation
- example: if all patients in the same clinic receive the same treatment
- then the clinic is the unit of analysis
- in some contexts there may be a clear hierarchy
- assessment of precision comes primarily from comparisons between units of analysis
- modelling of variation within units is necessary only if of intrinsic interest
- when relatively complex responses are collected on each study individual, the simplest way of condensing these is through a number of summary descriptive measures
- in other situations it may be necessary to represent explicitly the different hierarchies of variation



D. R. COX
CHRISTL A. DONNELLY

Principles of Applied Statistics

CAMBRIDGE

Regression with survival data

SM §10.8

- response y^0 is time to 'failure', or 'survival' time $y^0 \geq 0$
- density function $f(\cdot)$, distribution function $F(\cdot)$
- survivor function $S(\cdot) = 1 - F(\cdot)$
- hazard function $h(\cdot) = f(\cdot)/\{S(\cdot)\}$
- cumulative hazard function $H(y) = \int_0^y h(u)du = -\log S(y)$
- parametric models: exponential, Weibull, Gamma, log-normal, log-logistic SM §5.4
- random censoring: $C \sim G(\cdot)$, independently of Y^0
- observe $(y_j, d_j), j = 1, \ldots, n$
- $y_j = \min(y_j^0, c_j), \qquad d_j = \mathbb{1}(y_j^0 \leq c_j)$

... regression with survival data

SM §10.8

- $y_j = \min(y_j^0, c_j), \qquad d_j = \mathbb{1}(y_j^0 \leq c_j)$
- data $(y_j, d_j, x_j), j = 1, \ldots, n$; $\quad x_j$ explanatory variables
- survivor $S(\cdot; x, \beta)$, density $f(\cdot; x, \beta)$, hazard $h(\cdot; x, \beta)$
- log-likelihood

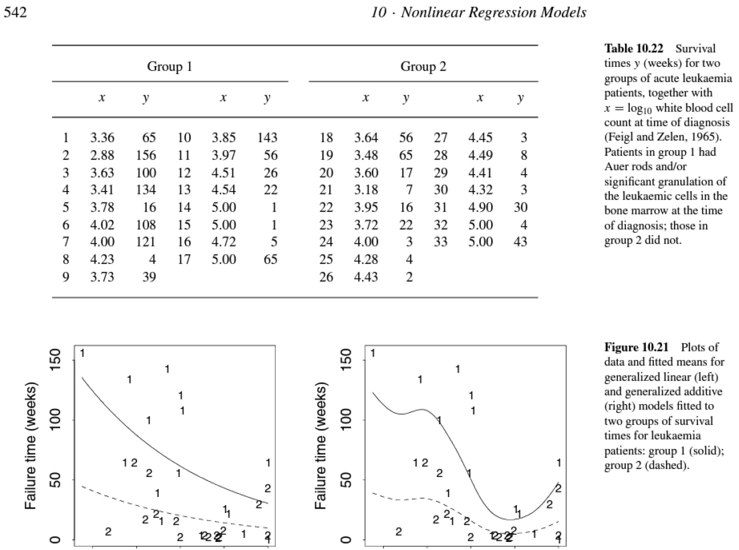
$$\ell(\beta; y, d) = \sum_{j=1}^{n} \{d_j \log h(y_j; x_j, \beta) - H(y_j; x_j, \beta)\}$$

SM (5.26)

- maximum likelihood estimates $\hat{\beta}$
  observed information function $-\ell''(\hat{\beta})$
- residuals

$$r_j = H(y_j; x_j, \hat{\beta}) + 1 - d_j$$

Example 10.36



542 *10 · Nonlinear Regression Models*

	Group 1						Group 2				
	x	y		x	y		x	y		x	y
1	3.36	65	10	3.85	143	18	3.64	56	27	4.45	3
2	2.88	156	11	3.97	56	19	3.48	65	28	4.49	8
3	3.63	100	12	4.51	26	20	3.60	17	29	4.41	4
4	3.41	134	13	4.54	22	21	3.18	7	30	4.32	3
5	3.78	16	14	5.00	1	22	3.95	16	31	4.90	30
6	4.02	108	15	5.00	1	23	3.72	22	32	5.00	4
7	4.00	121	16	4.72	5	24	4.00	3	33	5.00	43
8	4.23	4	17	5.00	65	25	4.28	4			
9	3.73	39				26	4.43	2			

Table 10.22 Survival times y (weeks) for two groups of acute leukaemia patients, together with $x = \log_{10}$ white blood cell count at time of diagnosis (Feigl and Zelen, 1965). Patients in group 1 had Auer rods and/or significant granulation of the leukaemic cells in the bone marrow at the time of diagnosis; those in group 2 did not.

Figure 10.21 Plots of data and fitted means for generalized linear (left) and generalized additive (right) models fitted to two groups of survival times for leukaemia patients: group 1 (solid); group 2 (dashed).

... example 10.36

```
> library(SMPracticals)
> library(survival)
> data(leuk)
> head(leuk)
    wbc      ag time
1  2300 present   65
2   750 present  156
3  4300 present  100
4  2600 present  134
5  6000 present   16
6 10500 present  108
> with(leuk,log10(wbc[1:5]))
[1] 3.361728 2.875061 3.633468 3.414973 3.778151
leuk.glm <- glm(time ˜ ag + log10(wbc), data = leuk, family = Gamma(link = "log"))
> summary(leuk.glm, dispersion = 1)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.8154     1.2932   4.497 6.89e-06 ***
agpresent     1.0176     0.3492   2.914  0.00357 **
log10(wbc)   -0.7009     0.3036  -2.308  0.02097 *
---

(Dispersion parameter for Gamma family taken to be 1)

    Null deviance: 58.138  on 32  degrees of freedom
Residual deviance: 40.319  on 30  degrees of freedom

> summary(leuk.glm)
...
(Dispersion parameter for Gamma family taken to be 1.087715)
```

... example 10.36

```
> leuk.surv <- survreg(Surv(time, rep(1,length(time))) ~ log10(wbc) + ag, data = leuk, dist=
> summary(leuk.surv)

Call:
survreg(formula = Surv(time, rep(1, length(time))) ~ log10(wbc) +
    ag, data = leuk, dist = "exponential")
             Value Std. Error     z        p
(Intercept)  5.815      1.263  4.60 4.15e-06
log10(wbc)  -0.701      0.286 -2.45 1.44e-02
agpresent    1.018      0.364  2.80 5.14e-03

Scale fixed at 1

> leuk.surv2 <- survreg(Surv(time,rep(1,length(time)))~pspline(log10(wbc),df=0) + ag,
+ data = leuk, dist= "exponential" )

## see help file for survreg

> leuk.gam <- gam(time ~ s(log10(wbc)) + ag, data = leuk, family = Gamma(link = "log") )
> summary(leuk.gam, dispersion = 1)
...
Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.7270     0.2524   10.80  < 2e-16 ***
agpresent     1.1424     0.3547    3.22  0.00128 **
...
Approximate significance of smooth terms:
                edf Ref.df Chi.sq p-value
s(log10(wbc)) 3.236  3.967  14.59 0.00553 **
```

Proportional hazards model

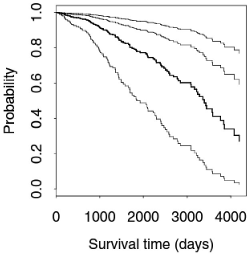
- hazard function $h(y; x, \beta) = h_0(y) \exp(x^{\mathrm{T}}\beta)$
- survivor function $S(y; x, \beta) = S_0(y)^{\exp(x^{\mathrm{T}}\beta)}$
- log-likelihood $\sum_{j=1}^{n} \{d_j x_j^{\mathrm{T}}\beta + \log h_0(y_j) - H_0(y_j) \exp(x_j^{\mathrm{T}}\beta)\}$
- partial likelihood

$$L_{\text{part}}(\beta) = \prod_{j=1}^{n} \left\{ \frac{\exp(x_j^{\mathrm{T}}\beta)}{\sum_{i \in \mathcal{R}_j} \exp(x_i^{\mathrm{T}}\beta)} \right\}^{d_j}$$

- derived in SM §10.8 as profile likelihood, treating $h_0(y_1), \ldots, h_0(y_n)$ as n nuisance parameters
- $\mathcal{R}_j$ risk set at time y_j^-
- all observations available to fail just before the time of the jth failure
- adjustments for ties, see p.544

... proportional hazards model

- estimation of the hazard function and survival probability
- $$\widehat{H}_0(y) = \sum_{j:y_j \leq y} \frac{d_j}{\sum_{i \in \mathcal{R}_j} \exp(x_i^{\mathrm{T}} \hat{\beta})}$$
- $$\widehat{S}_0(y) = \prod_{j:y_j \leq y} \left(1 - \frac{d_j}{\sum_{i \in \mathcal{R}_j} \exp(x_i^{\mathrm{T}} \hat{\beta})}\right)$$



... example 10.36

```
> leuk.ph <- coxph(Surv(time,rep(1,length(time)))~  ag + log10(wbc) , data = leuk)
> summary(leuk.ph)
...
              coef exp(coef) se(coef)      z Pr(>|z|)
agpresent  -1.0691    0.3433   0.4293 -2.490  0.01276 *
log10(wbc)  0.8467    2.3318   0.3132  2.703  0.00687 **
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

           exp(coef) exp(-coef) lower .95 upper .95
agpresent     0.3433     2.9126     0.148     0.7964
log10(wbc)    2.3318     0.4288     1.262     4.3083
```

```
> leuk.ph2 <- coxph(Surv(time,rep(1,length(time)))~  ag + log10(wbc) ,
+ ties="breslow", data = leuk)
> summary(leuk.ph2)
...
              coef exp(coef) se(coef)      z Pr(>|z|)
agpresent  -1.0176    0.3614   0.4235 -2.403  0.01626 *
log10(wbc)  0.8296    2.2924   0.3120  2.659  0.00785 **
---
```

```
> plot(survfit(leuk.ph))
```

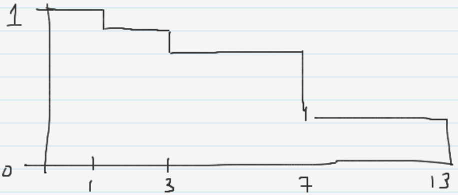
Kaplan-Meier estimation of $S(\cdot)$

- nonparametric estimation of survivor function
- censored data analogue of empirical cumulative distribution function
- $$\widehat{S}(y) = \prod_{i:y_i \leq y} \left(1 - \frac{d_i}{r_i}\right)$$
- with grouped data:
  - d_i number of items failing at time y_i
- with continuous data
  $$\widehat{S}(y) = \prod_{i:y_i \leq y} \left(1 - \frac{1}{r_i}\right)^{d_i}$$
- $$\widehat{\text{var}}\{\log \widehat{S}(y)\} = \sum_{i:y_i \leq y} \frac{d_i}{r_i(r_i - d_i)}$$

$\text{Obs}^{\underline{n}}$: 1, 3, 5^+, 7, 7^+, 13

$$\hat{S}(t) = \prod_{y_j \le t} \left(1 - \frac{1}{r_j}\right)^{d_j}$$

$$
= \begin{cases}
1 & t < 1 \\
\frac{5}{6} & 1 < t < 3 \\
\frac{5}{6} \times \frac{4}{5} = \frac{4}{6} & 3 < t < 7 \\
\frac{4}{6} \times \frac{2}{3} = \frac{4}{9} & 7 < t < 13 \\
\frac{4}{9} \times \frac{0}{1} = 0 & t > 13
\end{cases}
$$



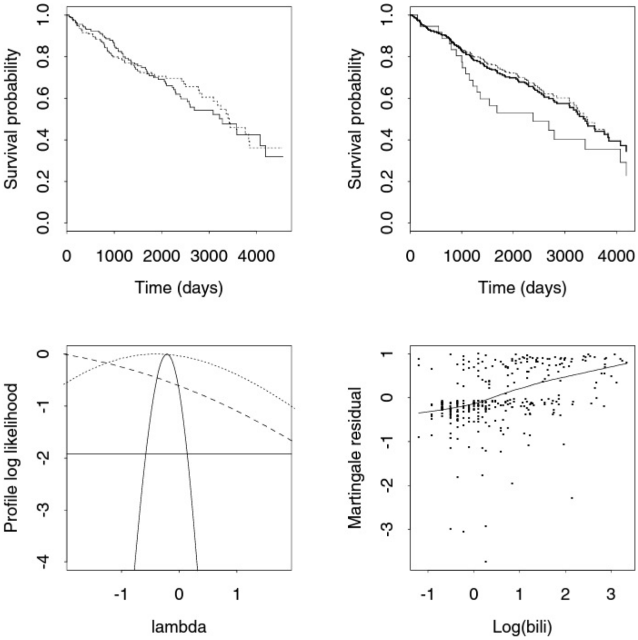
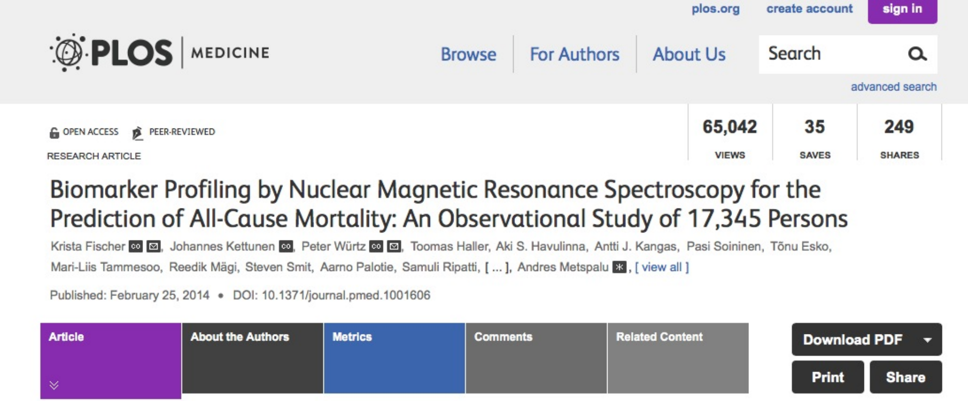


Figure 10.22 PBC data analysis (Fleming and Harrington, 1991). Top left: product-limit estimates for control (solid) and treatment (dots) groups. Top right: estimates of baseline survivor function for data stratified by sex, men (dots), women (solid). The heavy line shows the unstratified estimate. Lower left: profile likelihood for Box–Cox transformations of bilirubin (solid), albumin (dots), and prothrombin time (dashes); the horizontal line indicates 95% confidence limits for the transformation parameter. Lower right: martingale residuals from the model with terms `age`, `log(alb)`, `edtrt`, `log(protime)` against log bilirubin, and lowess smooth with $p = 2/3$.

Biomarkers and survival time

- News report (LA Times, Feb 28/14): “A blood test to predict imminent death? Would you want to take it?
- “the study suggests that several potentially deadly conditions – cancer, cardiovascular disease and a welter of non-vascular causes of death – may share signs, and even origins, that have been hidden in plain sight”
- “That said, a blood test to predict death is far from ready for prime time.”
- “They ran the carefully collected blood samples of 9,482 Estonians between the age of 18 and 101 through a scan that used nuclear magnetic resonance spectroscopy, to make measurements of the 106 biomarker candidates in each”
- “Over a median follow-up period of just over five years, 508 of the randomly chosen Estonian subjects died of various causes. The study’s authors compared the biomarker levels of both groups in an effort to identify those that were more common in the dead and less common in the living.”
- “The researchers then repeated their test of biomarkers on a separate population–8,444 Finnish men and women between 24 and 74. The biomarkers were equally predictive of death in this ‘validation group.’ ”



OPEN ACCESS PEER-REVIEWED

65,042 VIEWS | 35 SAVES | 249 SHARES

RESEARCH ARTICLE

Biomarker Profiling by Nuclear Magnetic Resonance Spectroscopy for the Prediction of All-Cause Mortality: An Observational Study of 17,345 Persons

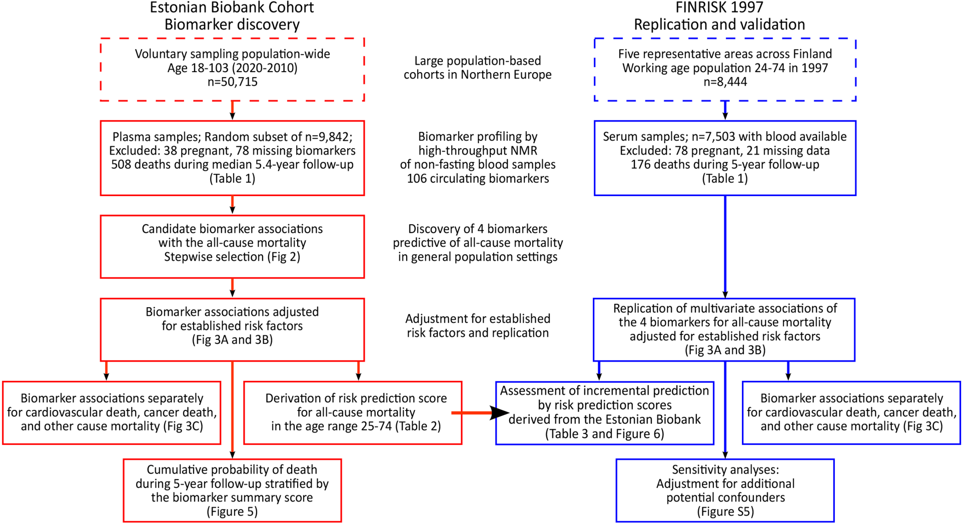
Krista Fischer, Johannes Kettunen, Peter Würtz, Toomas Haller, Aki S. Havulinna, Antti J. Kangas, Pasi Soininen, Tõnu Esko, Mari-Liis Tammesoo, Reedik Mägi, Steven Smit, Aarno Palotie, Samuli Ripatti, [...], Andres Metspalu, [view all]

Published: February 25, 2014 • DOI: 10.1371/journal.pmed.1001606

link

Study populations

- "In this observational study, two population-based cohorts in Estonia and Finland were followed for all-cause mortality via population registries"
- "The Estonian Biobank cohort included 50,715 individuals aged 18 – 103 y at recruitment (Oct 2002 – Feb 2011)
- "Biomarker profiling was conducted by NRM spectroscopy ... for a random subset of 9,842 individuals"
- "According to linkage with the Estonian population registry, 508 study participants had died during follow-up as of 1 June 2013."
- "The FINRISK 1997 study is a general population study ... persons aged 24 – 74 y"
- "In total, 8.444 individuals were recruited; biomarker profiling by NMR spectroscopy .. for 7,503 individuals"
- "... analyses in the validation cohort were confined to the first 5 y of follow-up; 176 ... died "
- Figure 1

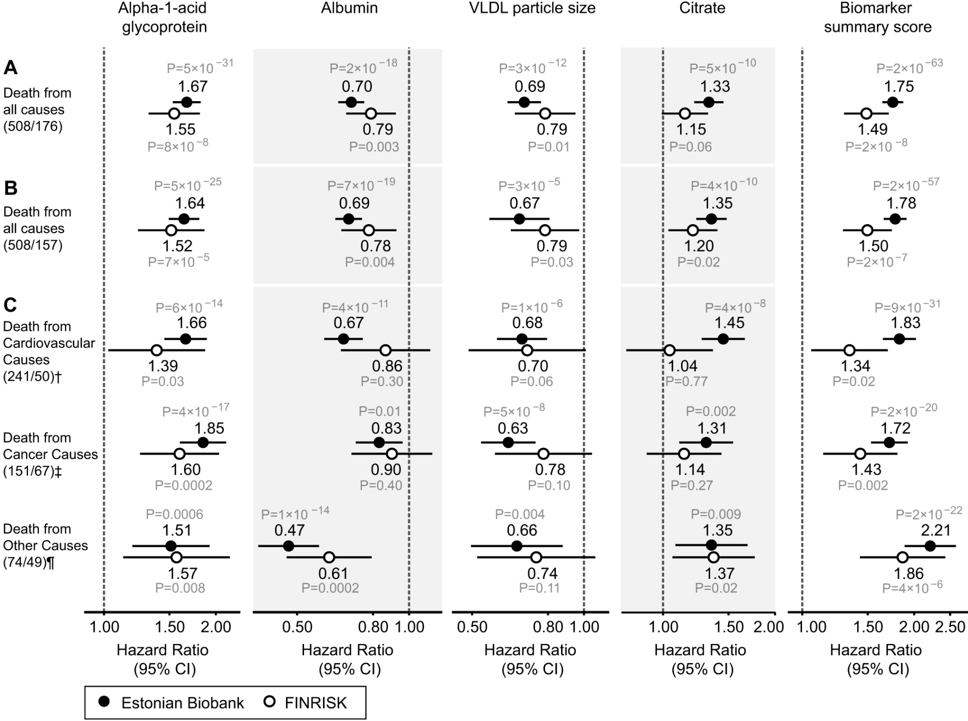

Estonian Biobank Cohort Biomarker discovery
FINRISK 1997 Replication and validation
Voluntary sampling population-wide Age 18-103 (2020-2010) n=50,715
Large population-based cohorts in Northern Europe
Five representative areas across Finland Working age population 24-74 in 1997 n=8,444
Plasma samples; Random subset of n=9,842; Excluded: 38 pregnant, 78 missing biomarkers 508 deaths during median 5.4-year follow-up (Table 1)
Biomarker profiling by high-throughput NMR of non-fasting blood samples 106 circulating biomarkers
Serum samples; n=7,503 with blood available Excluded: 78 pregnant, 21 missing data 176 deaths during 5-year follow-up (Table 1)
Candidate biomarker associations with the all-cause mortality Stepwise selection (Fig 2)
Discovery of 4 biomarkers predictive of all-cause mortality in general population settings
Biomarker associations adjusted for established risk factors (Fig 3A and 3B)
Adjustment for established risk factors and replication
Replication of multivariate associations of the 4 biomarkers for all-cause mortality adjusted for established risk factors (Fig 3A and 3B)
Biomarker associations separately for cardiovascular death, cancer death, and other cause mortality (Fig 3C)
Derivation of risk prediction score for all-cause mortality in the age range 25-74 (Table 2)
Assessment of incremental prediction by risk prediction scores derived from the Estonian Biobank (Table 3 and Figure 6)
Biomarker associations separately for cardiovascular death, cancer death, and other cause mortality (Fig 3C)
Cumulative probability of death during 5-year follow-up stratified by the biomarker summary score (Figure 5)
Sensitivity analyses: Adjustment for additional potential confounders (Figure S5)

Statistical analysis

- all biomarker concentrations were scaled to standard deviation units
- Cox proportional hazards models were used to assess the association of each candidate biomarker with the risk of all-cause mortality
- age at blood sampling was used as the time scale
- first, the biomarker leading to the smallest *p*-value in the Cox model adjusted for age and sex only was included
- subsequently, the biomarker leading to the smallest *p*-value in the model adjusted for age, sex and the first biomarker was included
- the process was repeated until no additional biomarkers were significant ... $p < 0.0005$
- Figure 1

... statistical analysis

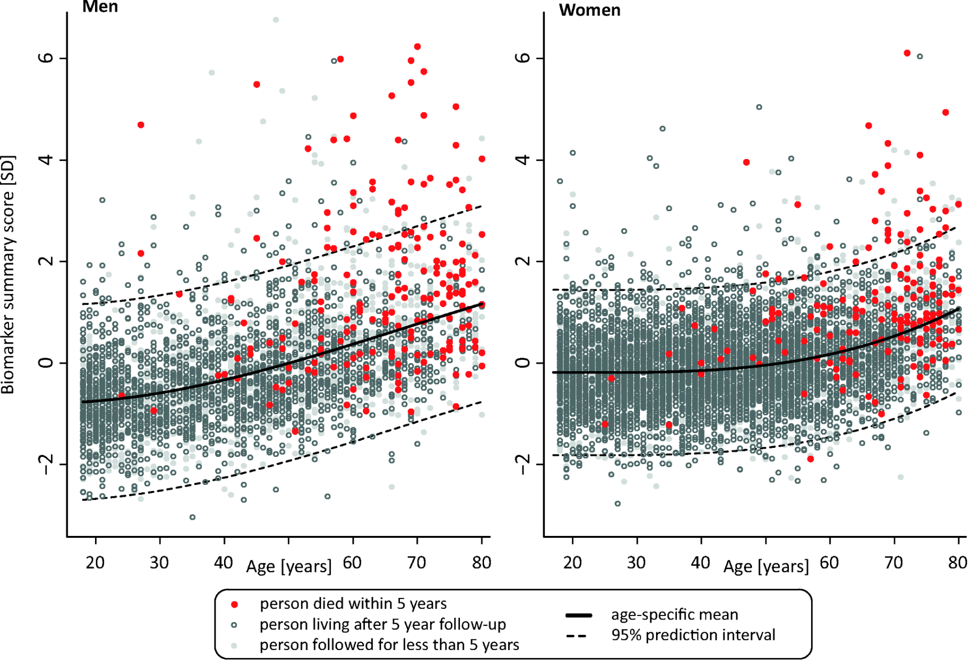
- the hazard ratios of the four identified biomarkers for all-cause mortality were subsequently examined in a multivariate model adjusted for age, sex, and conventional risk factors that were significant predictors of mortality in the Estonian Biobank cohort: high-density lipoprotein (HDL) cholesterol, current smoking, prevalent diabetes, prevalent cardiovascular disease, and prevalent cancer (Model A)
- Proportional hazards assumptions of the regression models were confirmed by Schoenfeld's test. `cox.zph`
  MASS, CH.13
- Figure 3



Alpha-1-acid glycoprotein
Albumin
VLDL particle size
Citrate
Biomarker summary score
A
Death from all causes (508/176)
$P=5\times10^{-31}$
1.67
1.55
$P=8\times10^{-8}$
$P=2\times10^{-18}$
0.70
0.79
P=0.003
$P=3\times10^{-12}$
0.69
0.79
P=0.01
$P=5\times10^{-10}$
1.33
1.15
P=0.06
$P=2\times10^{-63}$
1.75
1.49
$P=2\times10^{-8}$
B
Death from all causes (508/157)
$P=5\times10^{-25}$
1.64
1.52
$P=7\times10^{-5}$
$P=7\times10^{-19}$
0.69
0.78
P=0.004
$P=3\times10^{-5}$
0.67
0.79
P=0.03
$P=4\times10^{-10}$
1.35
1.20
P=0.02
$P=2\times10^{-57}$
1.78
1.50
$P=2\times10^{-7}$
C
Death from Cardiovascular Causes (241/50)†
$P=6\times10^{-14}$
1.66
1.39
P=0.03
$P=4\times10^{-11}$
0.67
0.86
P=0.30
$P=1\times10^{-6}$
0.68
0.70
P=0.06
$P=4\times10^{-8}$
1.45
1.04
P=0.77
$P=9\times10^{-31}$
1.83
1.34
P=0.02
Death from Cancer Causes (151/67)‡
$P=4\times10^{-17}$
1.85
1.60
P=0.0002
P=0.01
0.83
0.90
P=0.40
$P=5\times10^{-8}$
0.63
0.78
P=0.10
P=0.002
1.31
1.14
P=0.27
$P=2\times10^{-20}$
1.72
1.43
P=0.002
Death from Other Causes (74/49)¶
P=0.0006
1.51
1.57
P=0.008
$P=1\times10^{-14}$
0.47
0.61
P=0.0002
P=0.004
0.66
0.74
P=0.11
P=0.009
1.35
1.37
P=0.02
$P=2\times10^{-22}$
2.21
1.86
$P=4\times10^{-6}$
1.00 1.50 2.00
0.50 0.80 1.00
0.50 0.80 1.00
1.00 1.50 2.00
1.00 1.50 2.00 2.50
Hazard Ratio (95% CI)
Hazard Ratio (95% CI)
Hazard Ratio (95% CI)
Hazard Ratio (95% CI)
Hazard Ratio (95% CI)
● Estonian Biobank
○ FINRISK

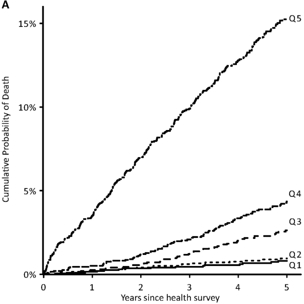
... statistical analysis

- A biomarker summary score was derived by adding the concentrations of the biomarkers weighted by the regression coefficients (natural logarithm of HR) observed in Model A
- $\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{\beta}_4 x_{4i}$
- Scatter plots of age versus the biomarker score were constructed for men and women, and the associations were examined by third degree polynomial regression fits
- The biomarker score was moderately correlated with age (r = 0.38), yet extreme biomarker score values were seen across all age groups.
- Excess mortality within 5 y of follow-up was observed for higher age, but in particular in combination with an elevated biomarker score
- Figure 4


Men
Women
Biomarker summary score [SD]
6
4
2
0
−2
Age [years]
20
30
40
60
70
80
● person died within 5 years
○ person living after 5 year follow-up
● person followed for less than 5 years
age-specific mean
95% prediction interval

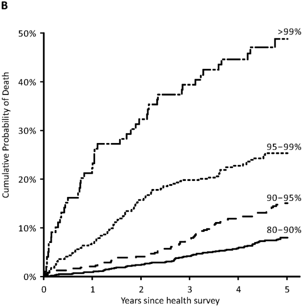
... statistical analysis

- To illustrate the strong association of the biomarker summary score in the Estonian Biobank cohort, the cumulative probability of death was derived across quintiles of the biomarker score
- The 5-y mortality for persons with a biomarker score within the highest quintile was 19 times higher than for those in the lowest quintile (288 versus 15 deaths during 5 y, corresponding to 15.3% versus 0.8%).
- Individuals within the highest quintile were further differentiated in terms of their short-term probability of dying according to their biomarker score percentiles
- Figure 5



Persons at risk by quintiles

Q1	1969	1966	1962	1960	1710	1397
Q2	1967	1963	1959	1955	1744	1430
Q3	1968	1963	1957	1943	1733	1417
Q4	1966	1956	1942	1925	1711	1345
Q5	1969	1900	1831	1773	1513	1083



Persons at risk in the highest quintile

80–90%	984	975	960	943	822	623
90–95%	492	482	472	455	381	272
95–99%	394	367	332	315	264	165
>99%	99	76	67	60	46	23

Risk score validation

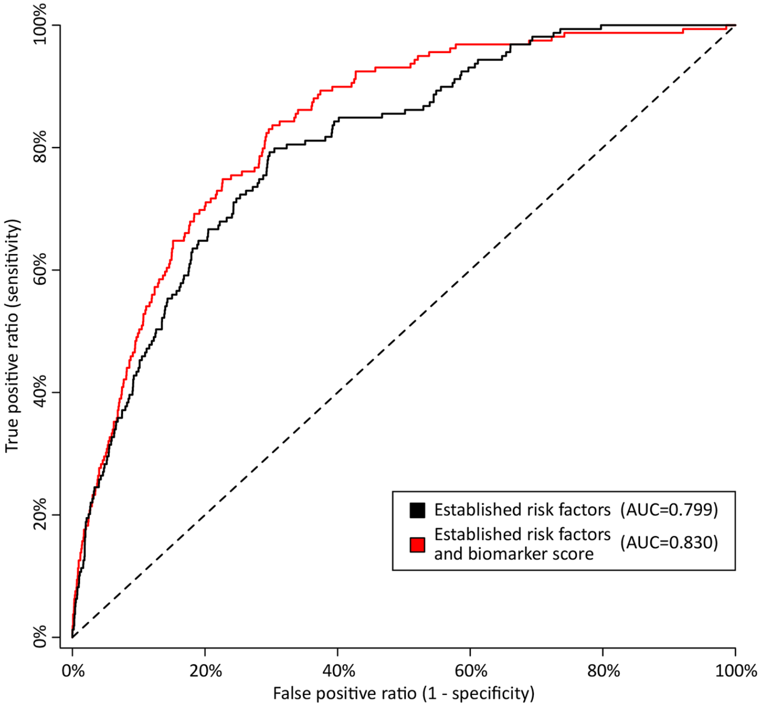
- risk prediction scores for all-cause mortality with and without the biomarkers were derived in the Estonian Biobank cohort and evaluated in the FINRISK validation cohort
- the regression coefficients used for calculating the two risk scores are listed in Table 2.
- Risk discrimination was significantly improved by including the biomarkers in the risk prediction
- The discrimination curves are illustrated in Figure 6.

Variable	Prediction Model without Biomarkers			Prediction Model with Biomarkers		
	HR	95% CI	*p*-Value	HR	95% CI	*p*-Value
Female gender	0.67	0.50–0.90	0.009	0.60	0.44–0.81	0.0008
Body mass index[a]	1.05	0.91–1.21	0.52	1.05	0.92–1.20	0.48
Systolic blood pressure[a]	0.96	0.85–1.09	0.51	1.04	0.92–1.18	0.55
Fasting duration (hours)	0.99	0.96–1.02	0.47	1.00	0.97–1.03	0.96
Total cholesterol[a]	1.05	0.91–1.21	0.50	1.15	0.97–1.36	0.11
HDL-cholesterol[a]	0.81	0.69–0.95	0.01	1.07	0.92–1.24	0.37
Triglycerides[a]	0.82	0.70–0.96	0.01	0.93	0.71–1.21	0.60
Creatinine[a]	1.10	1.03–1.18	0.005	1.04	0.96–1.12	0.31
Current smoking	1.86	1.26–2.75	0.002	1.56	1.05–2.33	0.03
Smoking duration (years)[a]	1.21	1.04–1.41	0.01	1.25	1.07–1.46	0.005
Cigarettes per day[a]	0.93	0.80–1.07	0.29	0.89	0.77–1.03	0.11
Alcohol[a]	1.09	0.98–1.21	0.11	1.04	0.94–1.16	0.43
Prevalent diabetes	1.58	1.15–2.15	0.004	1.49	1.09–2.03	0.01
Prevalent cardiovascular disease	1.38	1.05–1.82	0.02	1.42	1.08–1.87	0.01
Prevalent cancer	2.15	1.51–3.05	2×10^{-5}	2.26	1.59–3.20	5×10^{-6}
Alpha-1-acid glycoprotein[a]	—	—	—	1.76	1.57–1.97	9×10^{-23}
Albumin[a]	—	—	—	0.66	0.59–0.73	4×10^{-15}
VLDL particle size[a]	—	—	—	0.74	0.58–0.94	0.01
Citrate[a]	—	—	—	1.47	1.29–1.67	5×10^{-9}

Hazard ratios for all-cause mortality derived in the Estonian Biobank cohort in the age range matching the FINRISK cohort (25–74 y). The regression coeffients (natural logarithm of the HRs) from the Estonian Biobank cohort were used to derive two risk scores for the prediction of all-cause mortality: a reference risk score without biomarkers and a risk score including the four novel biomarkers. The two risk prediction scores were used to calculate the absolute risk estimates in the FINRISK cohort, and the incremental predictive utility of adding the four biomarkers to the risk prediction score was evaluated.

[a] Continuous variables were scaled to risk estimate per 1-SD increment in the variable.

doi:10.1371/journal.pmed.1001606.t002


True positive ratio (sensitivity)
100%
80%
60%
40%
20%
0%
0%
20%
40%
60%
80%
100%
False positive ratio (1 - specificity)
■ Established risk factors (AUC=0.799)
🟥 Established risk factors and biomarker score (AUC=0.830)

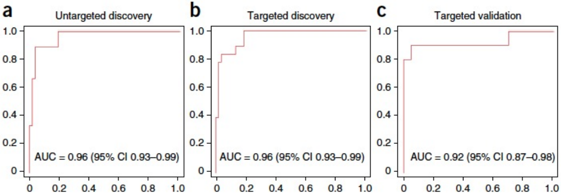


Figure 2 ROC results for the lipidomics analyses. (**a**–**c**) Plots of ROC results from the models derived from the three phases of the lipidomics analysis. Simple logistic models using only the metabolites identified in each phase of the lipidomics analysis were developed and applied to determine the success of the models for classifying the C_{pre} and NC groups. The red line in each plot represents the AUC obtained from the discovery-phase LASSO analysis (**a**), the targeted analysis of the ten metabolites in the discovery phase (**b**) and the application of the ten-metabolite panel developed from the targeted discovery phase in the independent validation phase (**c**). The ROC plots represent sensitivity (i.e., true positive rate) versus 1 – specificity (i.e., false positive rate).