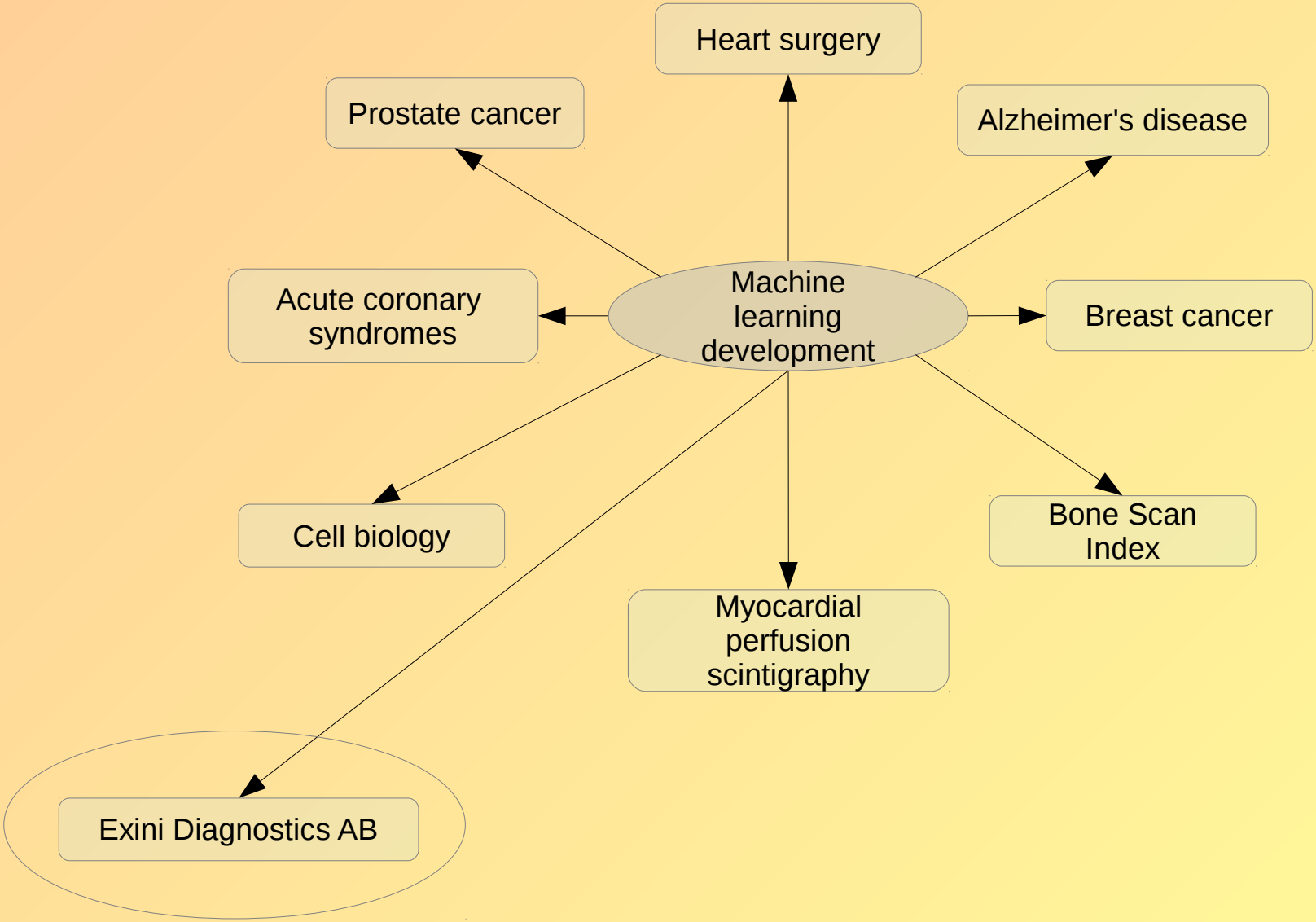
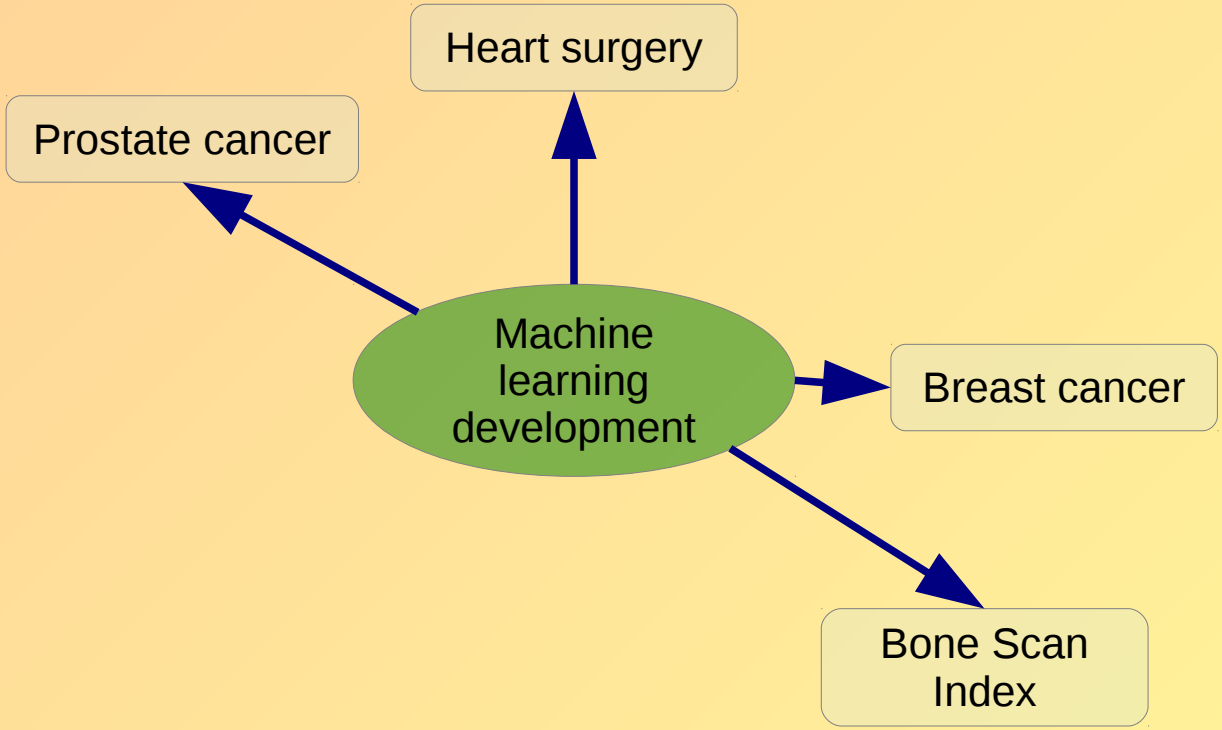
Modeling of survival data and some nice applications in clinical medicine

Mattias Ohlsson

Overview of activities

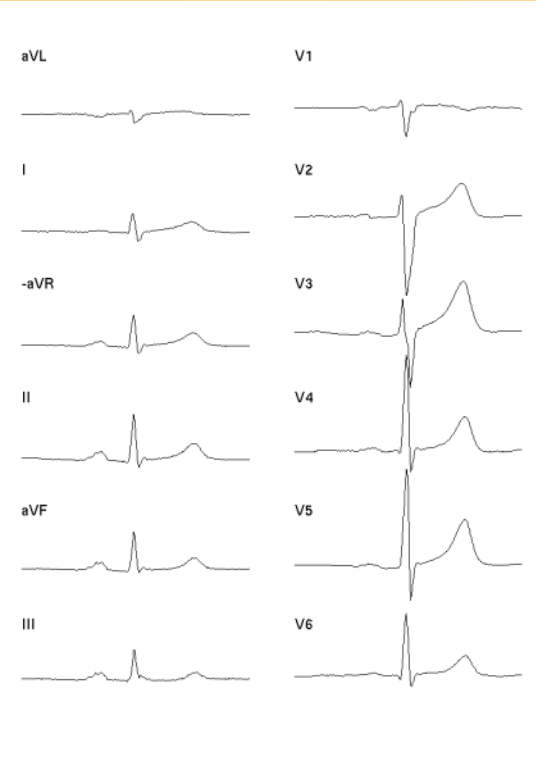


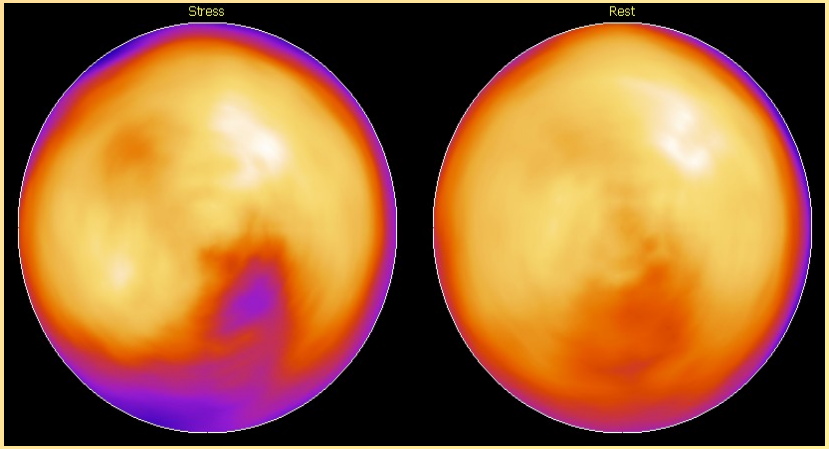
Common theme: survival analysis



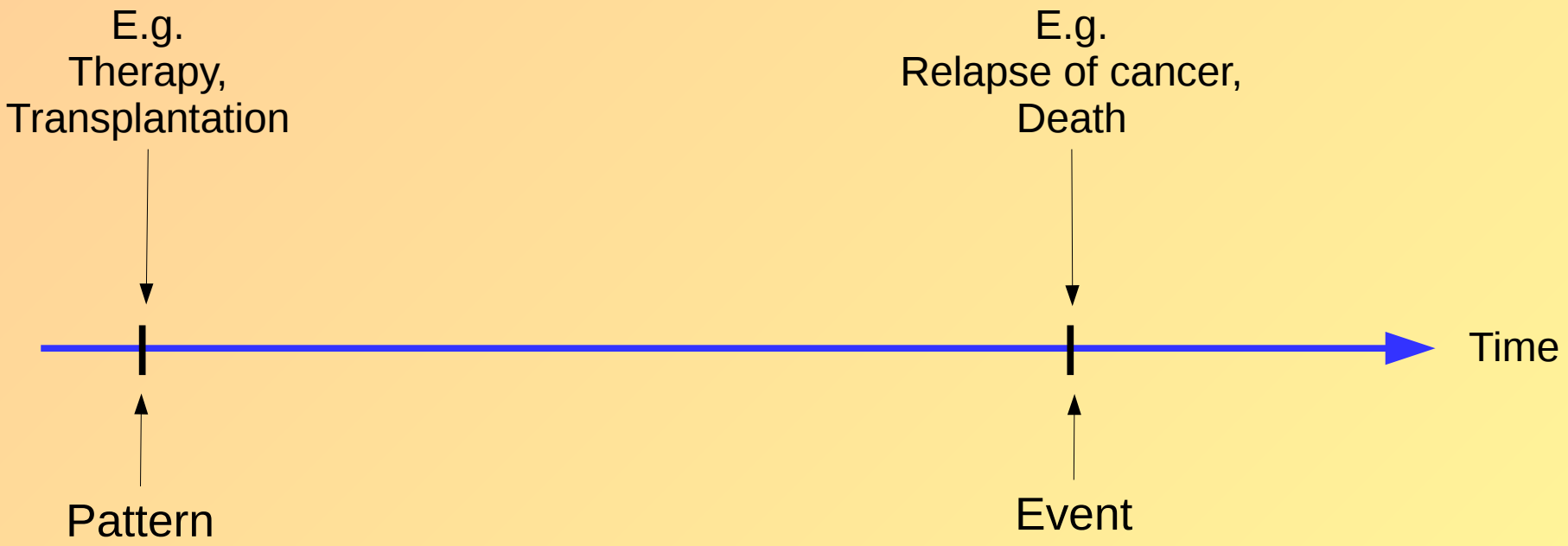
Survival analysis ~ adding a temporal information to pattern recognition problems

“Classical” pattern recognition





Survival analysis ~ analysis of time duration until an event occurs

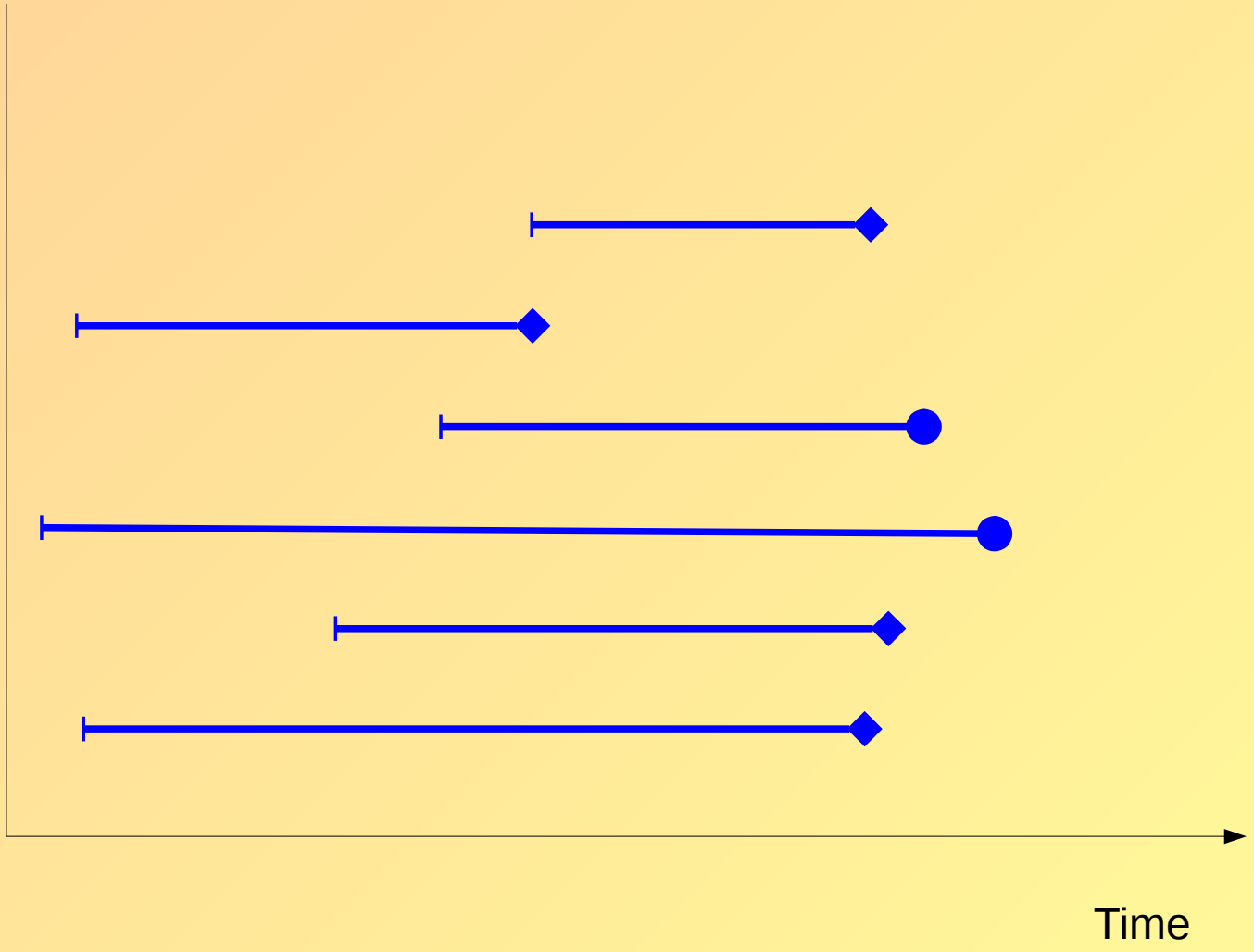


Survival analysis – the data

Pattern information

Event

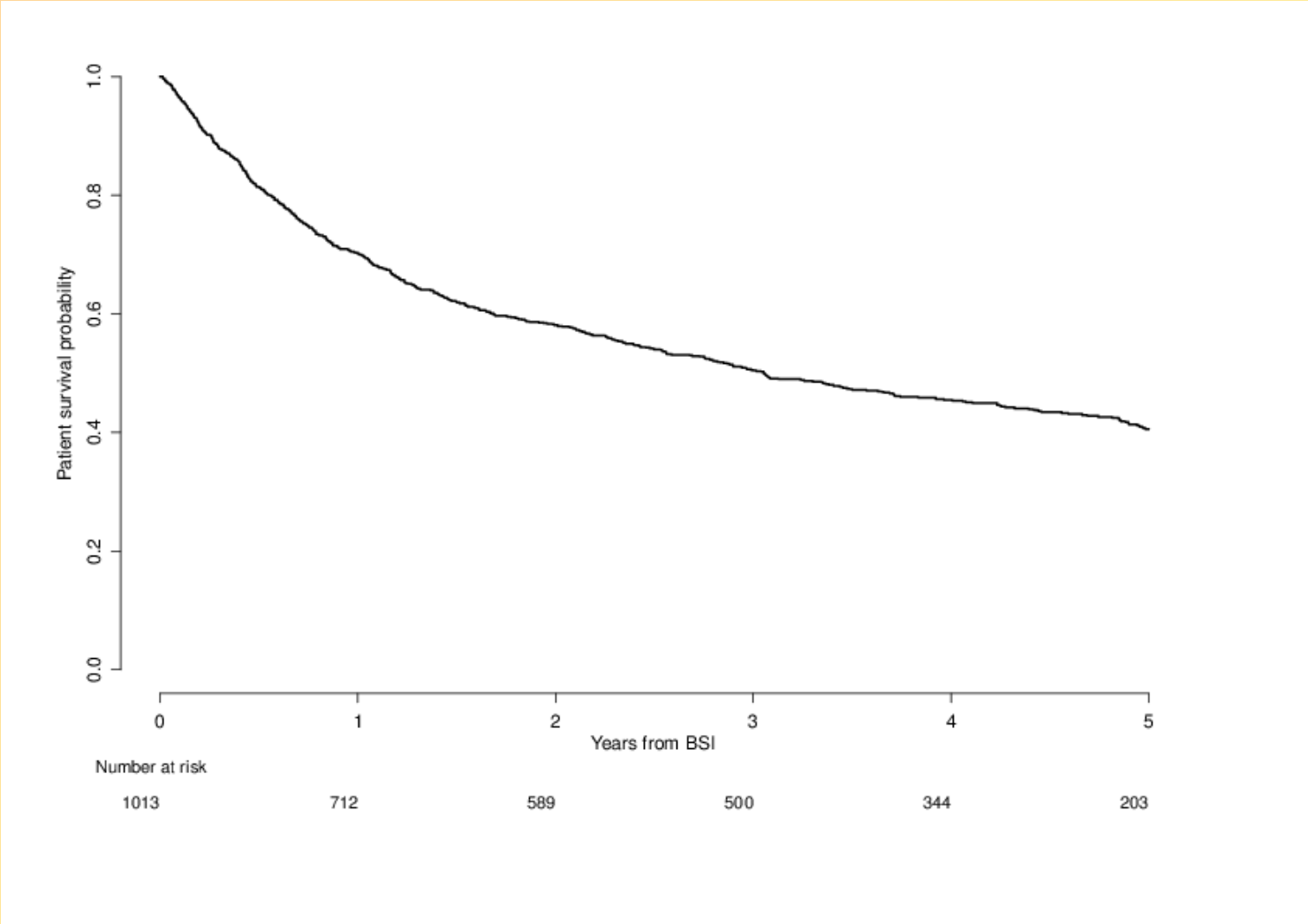
Censored event, eg. still alive, left the study, death because of other reasons etc

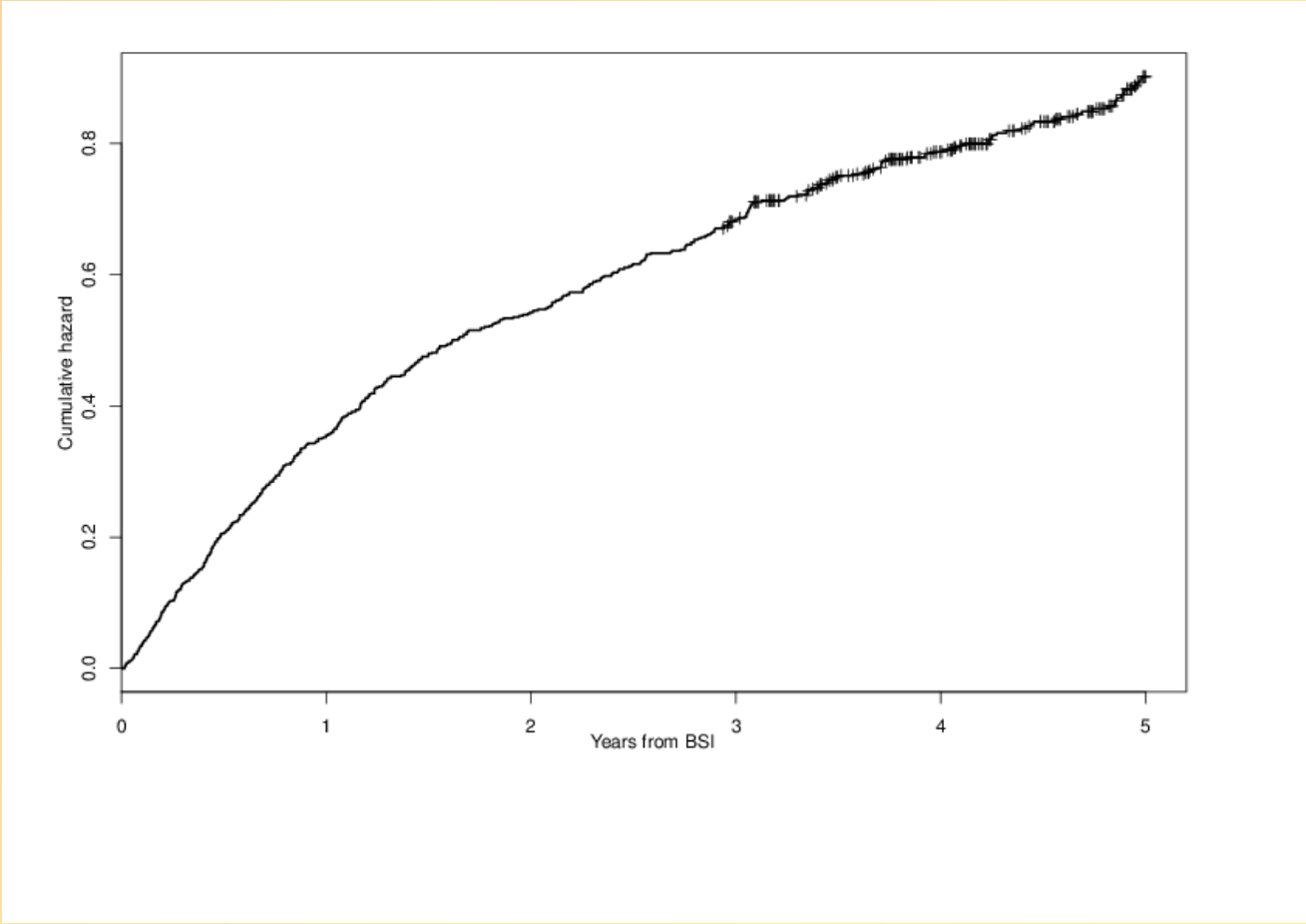


Survival analysis – what do we want to model?

The survival function

$$S(t) = P(T > t)$$





The hazard function

$$h(t) = \lim_{\Delta t \to 0} \frac{P(T < t + \Delta t | T \geq t)}{\Delta t} = -\frac{S'(t)}{S(t)}$$

(event rate at time t conditional on survival until time t or later.
Interpretation: risk of dying at time t)

Imaging biomarker for prostate cancer

Evaluation of the Bone Scan Index

In collaboration with Lars Edenbrandt, Clinical Physiology, Malmö

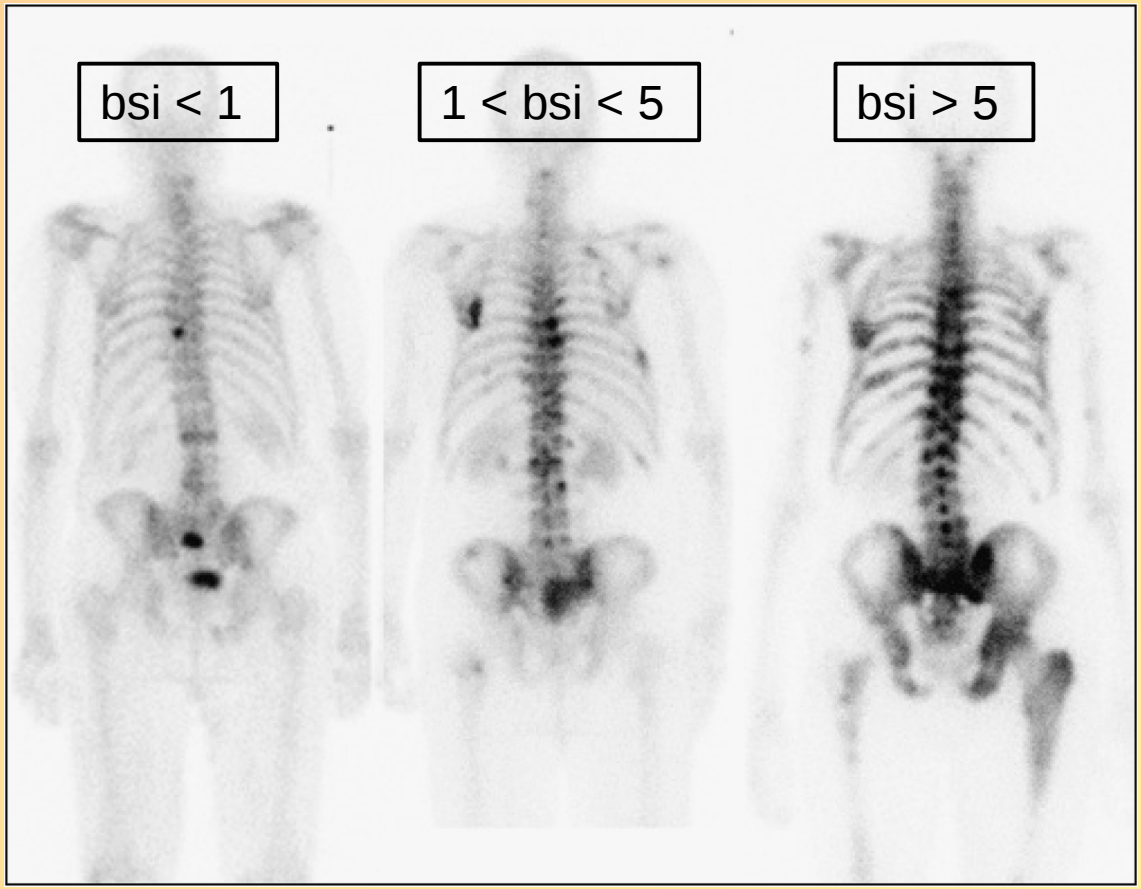
This project deals with later stages of prostate cancer, specifically when bone metastases occur.

Aim: Characterize the BSI imaging biomarker

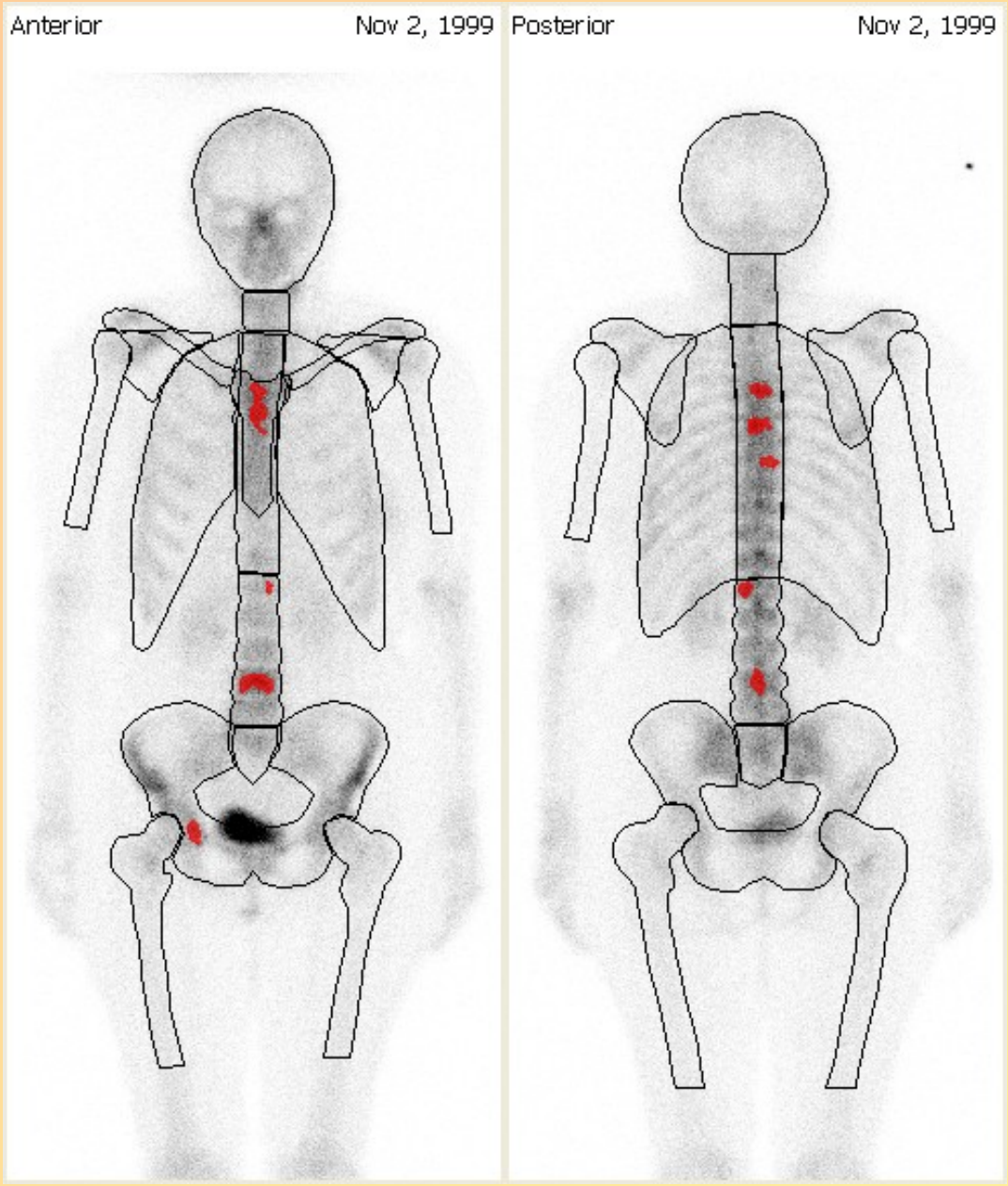
- prognosis
- treatment response

Other common biomarkers or “scores” may not be optimal (e.g. PSA, Gleason score)

BSI = Bone Scan Index



BSI is a method of expressing the tumor burden in the bone as a percentage of the total skeletal mass.



How to calculate BSI?

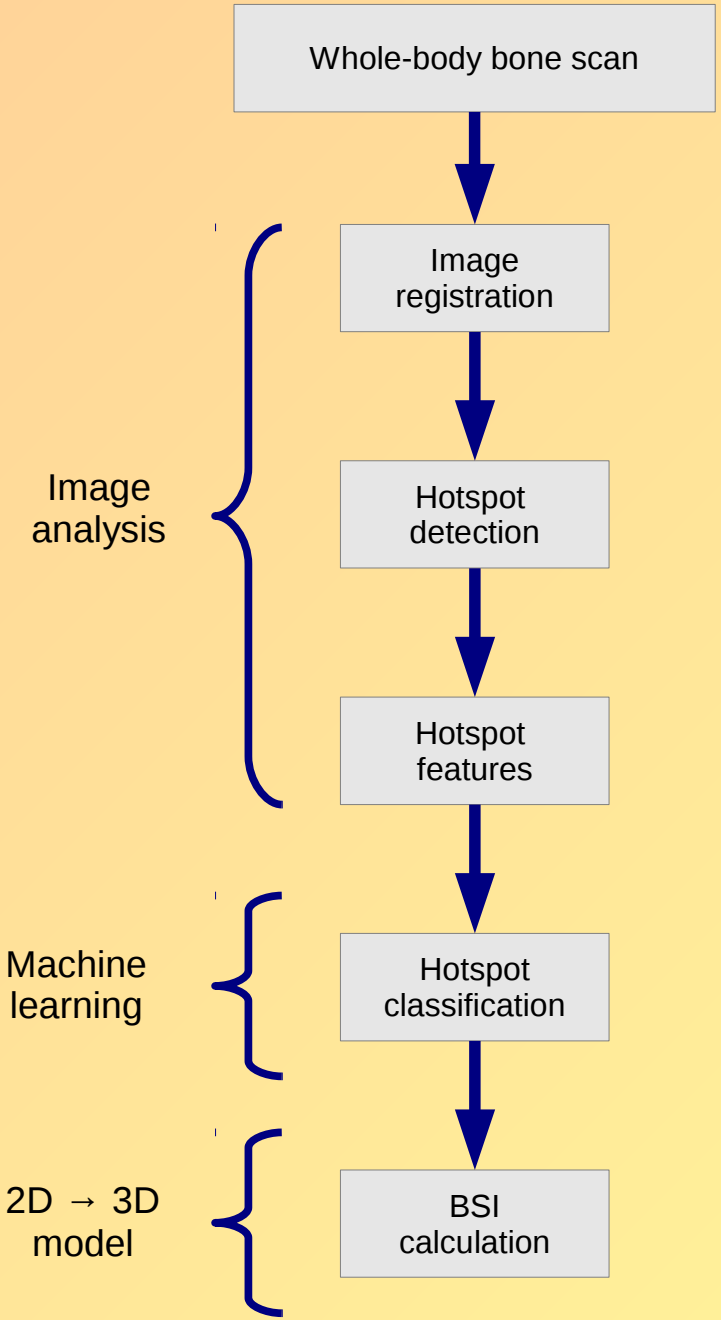
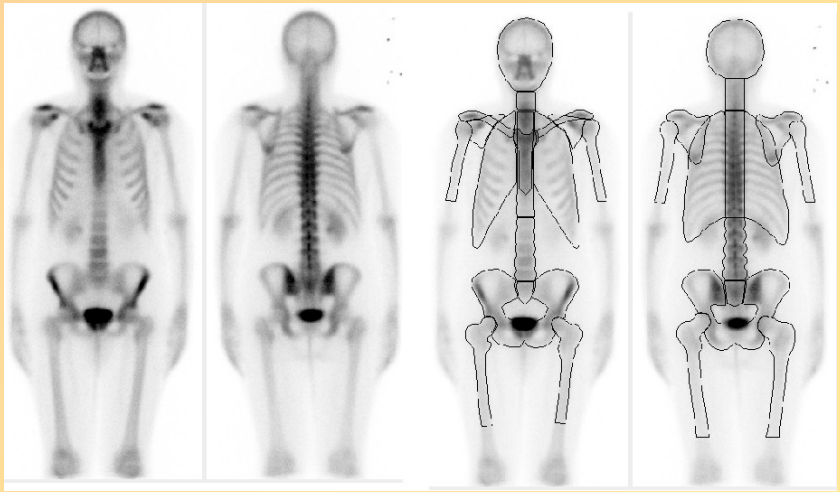

Whole-body bone scan
Image registration
Hotspot detection
Hotspot features
Hotspot classification
BSI calculation
Image analysis
Machine learning
2D → 3D model

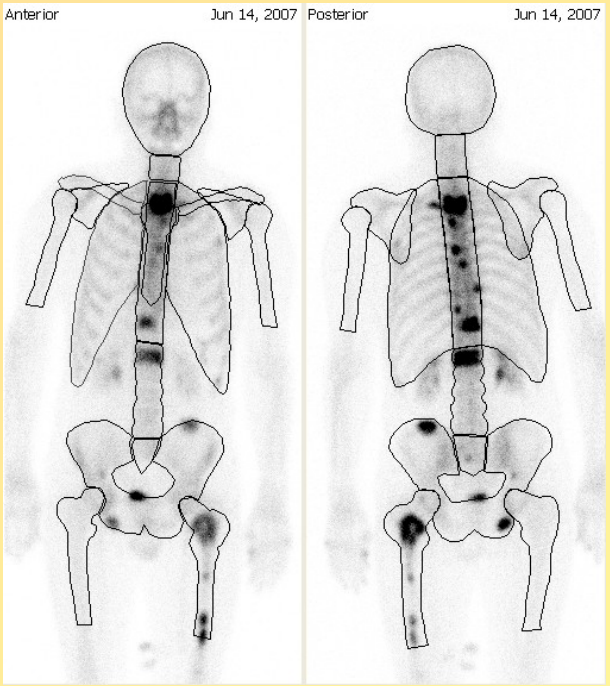
Image registration

Atlas

- Based on ~10 images
- Manual delineation
- 12 anatomical regions



The atlas is registred to the new image using the Morphon method



Hotspot detection

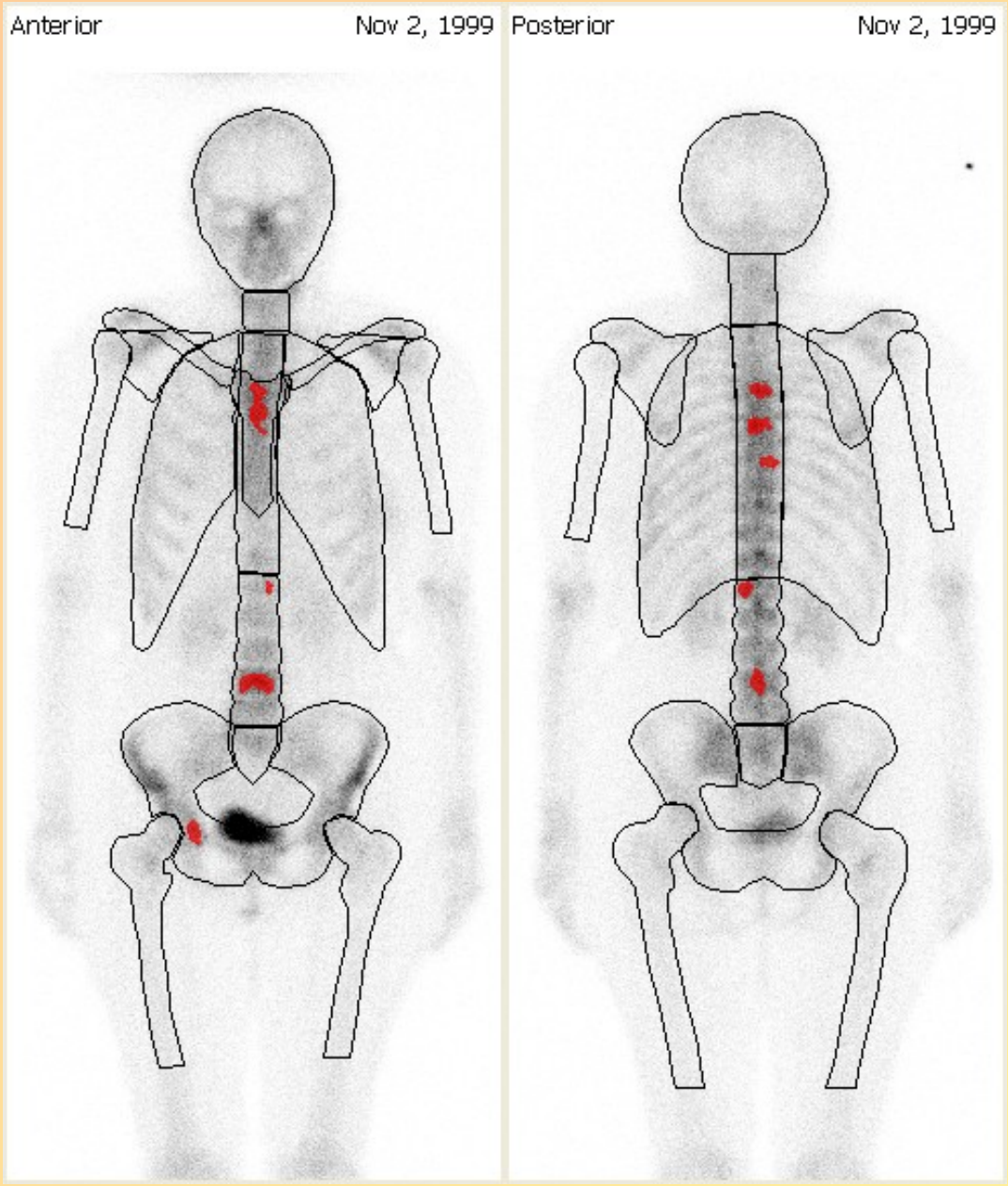
- Find average intensity in ***healthy*** bone tissue
- Normalize using the above average intensity

- Bandpass filtering
- Thresholding

Hotspot features

- Geometry features
- Localization features
- Intensity distribution features
- Other global features capturing the density of hotspots. Both regional and global.

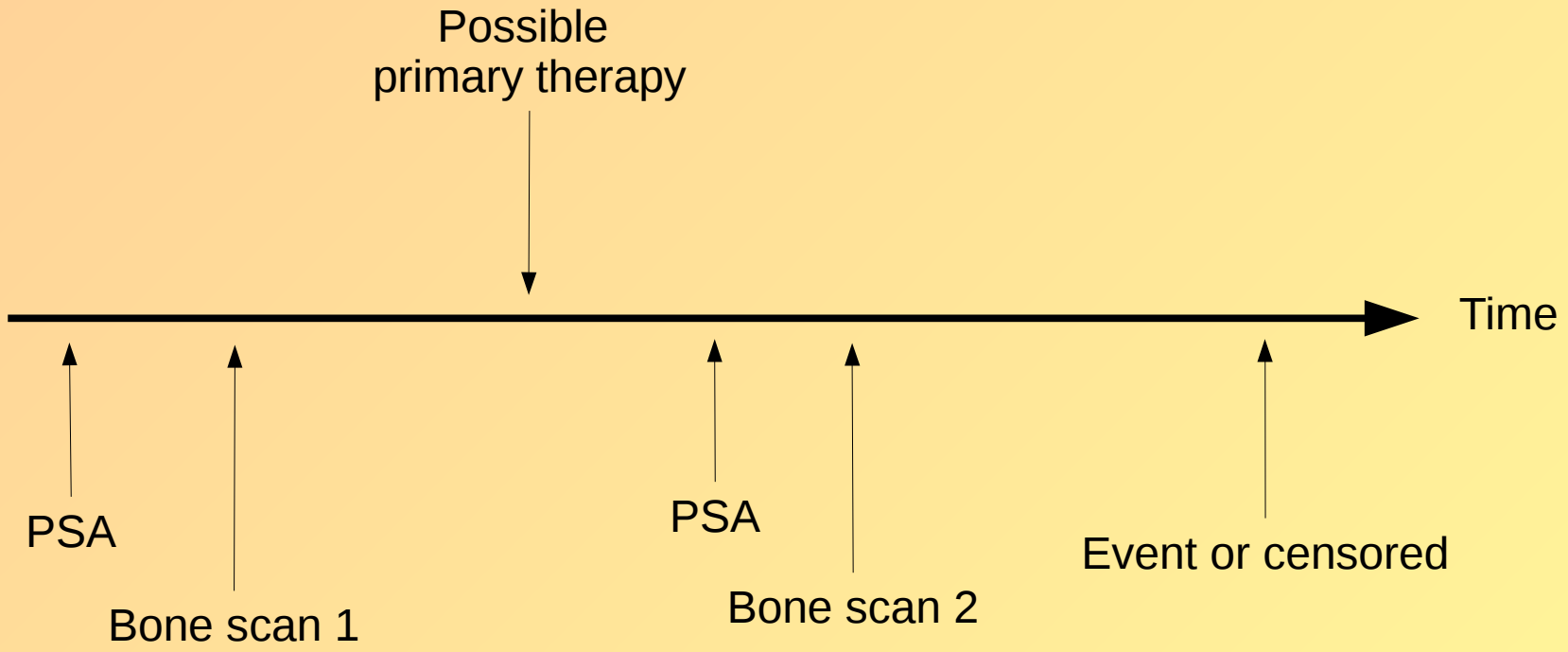
Machine learning



Now we can!

Exini Diagnostics

Data overview



How to model survival?

COX proportional hazard model of survival data –
very common method.

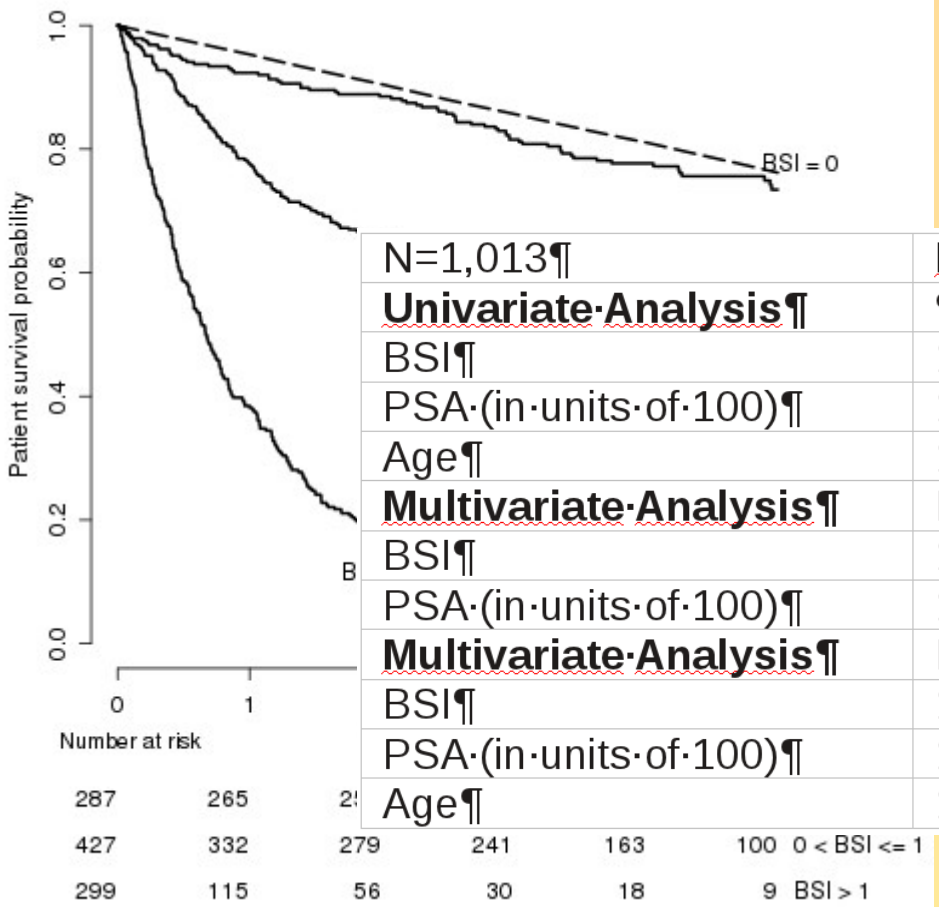
$$h(t, \mathbf{x}) = h_o(t) e^{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n}$$

$\beta_1 \cdots \beta_n$ can be estimated using “maximum partial likelihood”

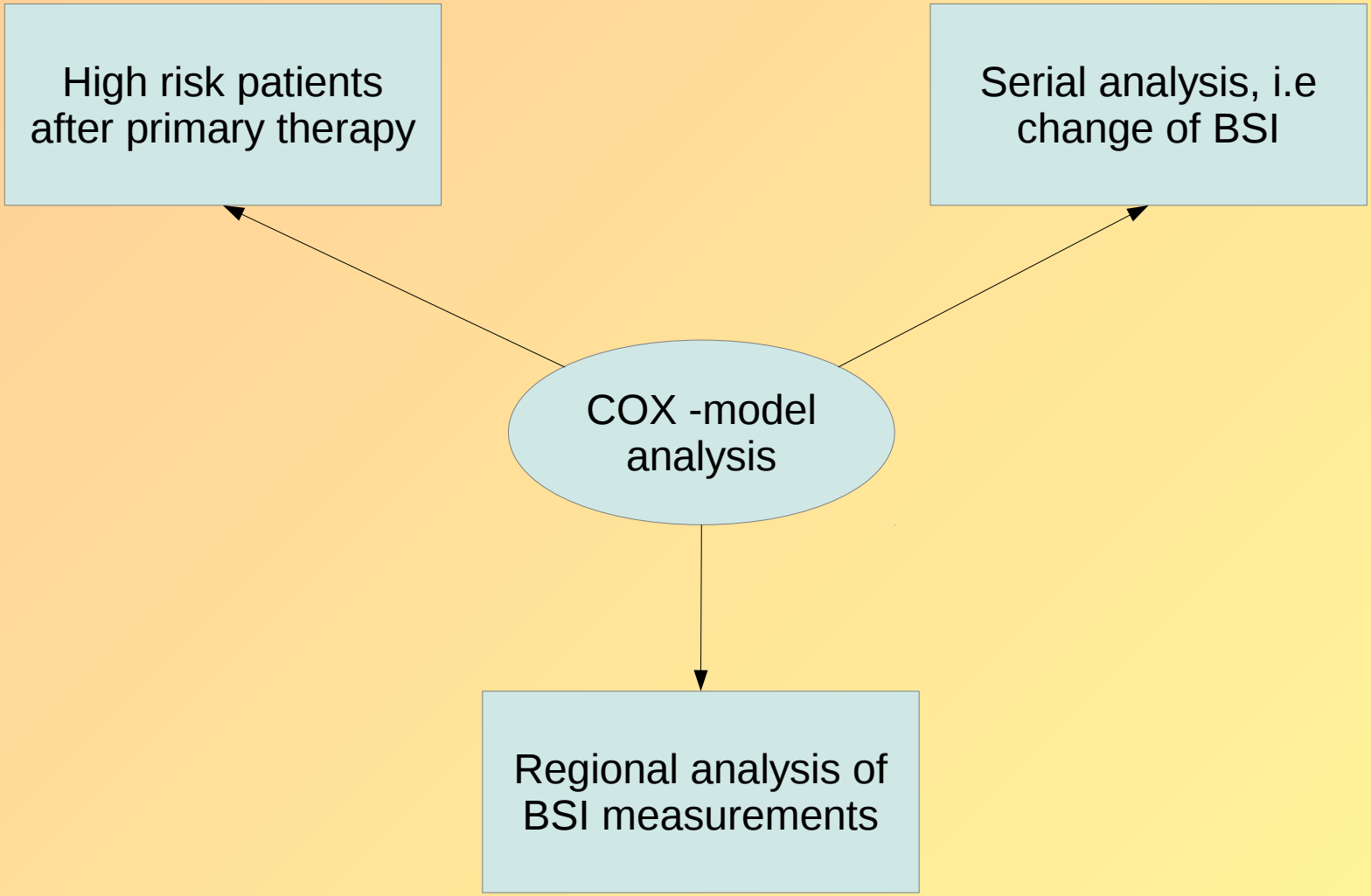
Relative risk (hazard ratio) becomes simple. For example comparing a unit change of one covariate:

$$\frac{h(t, x)|_{x_i+1}}{h(t, x)|_{x_i}} \equiv HR_{x_i} = e^{\beta_i}$$

Some results

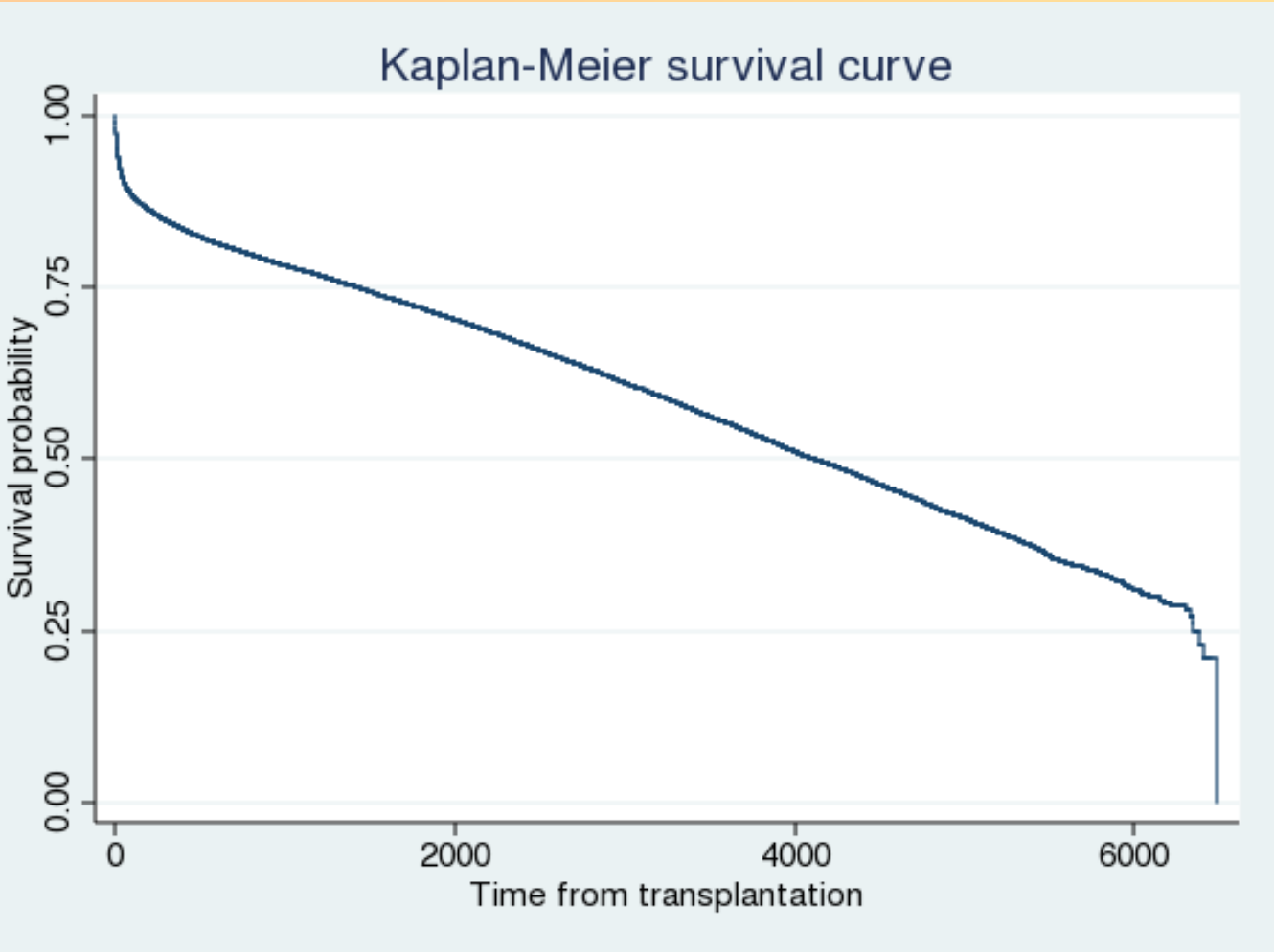


N=1,013¶	**Hazard·Ratio¶**	**p-value¶**	**c-index¶**
Univariate·Analysis¶	¶	¶	¶
BSI¶	1.23·(1.21·–·1.26)¶	<·0.0001¶	0.75¶
PSA·(in·units·of·100)¶	1.04·(1.03·–·1.05)¶	<·0.0001¶	0.68¶
Age¶	1.03·(1.02·–·1.04)¶	<·0.0001¶	0.58¶
Multivariate·Analysis¶	BSI·+·PSA¶		
BSI¶	1.22·(1.19·–·1.25)¶	<·0.0001¶	0.75¶
PSA·(in·units·of·100)¶	1.02·(1.01·–·1.03)¶	<·0.0001¶	
Multivariate·Analysis¶	BSI·+·PSA·+·Age¶		
BSI¶	1.22·(1.20·–·1.25)¶	<·0.0001¶	0.73¶
PSA·(in·units·of·100)¶	1.02·(1.00·–·1.03)¶	0.0006¶	
Age¶	1.03·(1.02·–·1.04)¶	<·0.0001¶	


High risk patients after primary therapy
Serial analysis, i.e change of BSI
COX -model analysis
Regional analysis of BSI measurements

Risk evaluation before heart transplantation

In collaboration with Johan Nilsson. hjärtkirurgi, LU



Overall survival for ~ 56 000 transplantations

Recipient-Donor matching problem

Many

Recipient
Age (years)
Female gender
Height (cm)
Weight (kg)
Ischemic cardiomyopathy
Non-ischemic cardiomyopathy
Insulin-treated diabetes
Hypertension
Antiarrhythmic
Amiodarone
Previous blood transfusion
Previously transplanted
Previous cardiac surgery
ECMO
Blood group (A,B,AB,O)
Creatinine (μmol/l)
Serum bilirubin (mg/dl)
….

Optimal match?

Donor
Demographic data
Age (year)
Female gender
Weight (kg)
Duration of ischemia (min)
CODD: Head trauma
CODD: Cerebrovascular event
Blood group (A,B,AB,O)

One

In total about 140 available “features”

Today

Possible recipients

1. Compatible blood group match
2. Recipient donor weight match ± 20%

Prioritize according to

1. Identical blood group
2. If young donor, select young recipient (< 35 years) or donor age - recipient age < 15 years
3. If PVR > 3.0 then 0-15% larger weight for the donor

Two or more recipients have the same priority then random selection

Aim: Better selection → improved survival

For discrete data $0 < t_1 < t_2 < \cdots < t_L \quad , a_l = (t_{l-1}, t_l]$

Survival function $S(t_l) = P(T > t_l)$

Discrete hazard rate $h_l = -\dfrac{S(t_l) - S(t_{l-1})}{S(t_{l-1})}$

$$S(t) = \prod_{l:t_l \leq t} (1 - h_l)$$

As usual “maximize the likelihood function”

PLANN = Partial logistic regression with ANN

$$h_l(\vec{x}, a_l) \rightarrow y(\vec{x}, a_l)$$

Model these by neural networks

The output from the neural network will provide smoothed estimates of the discrete hazard rates.

A more flexible modeling!

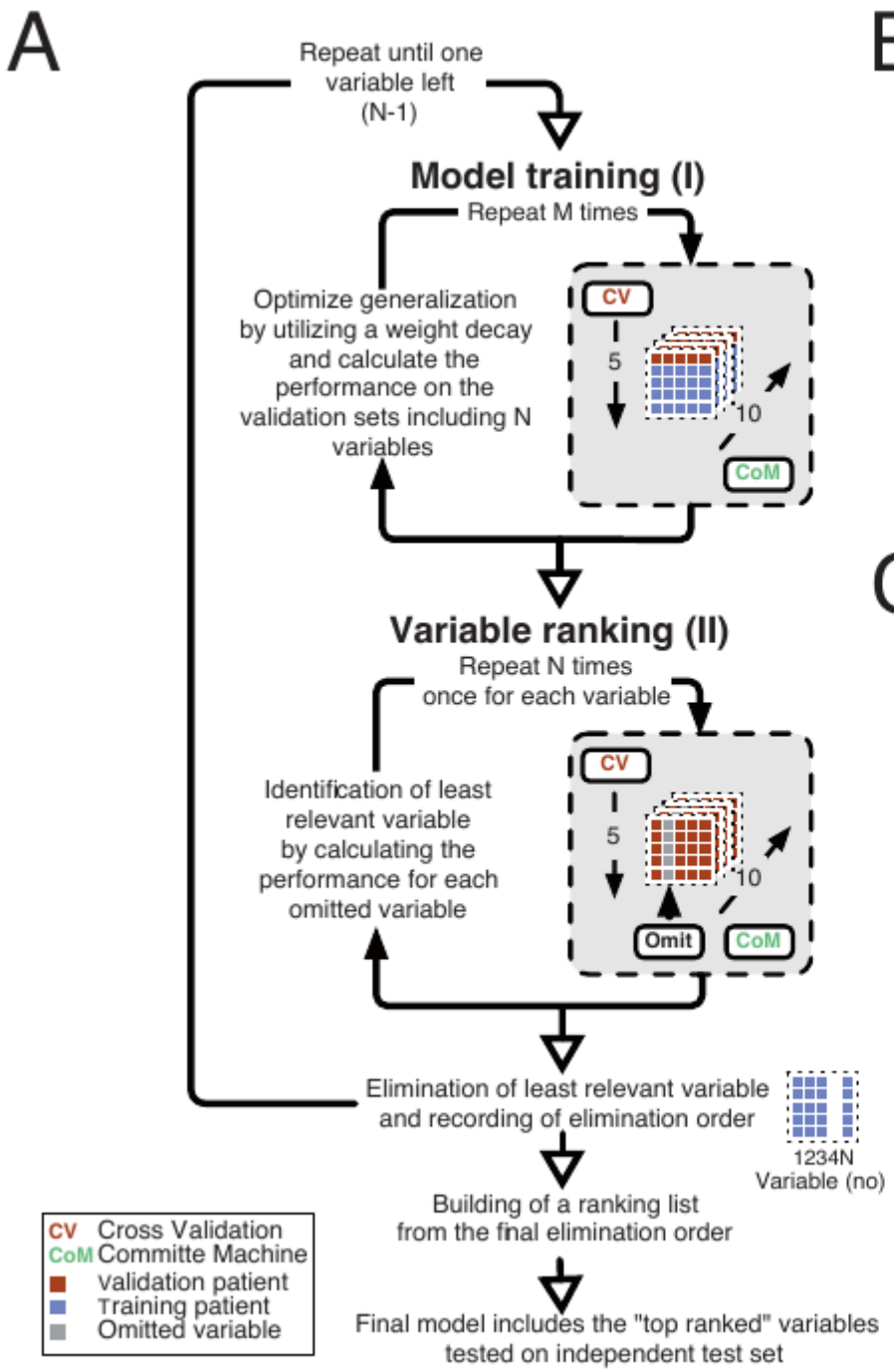
Also add: regularization, ensemble approaches, multiple random imputations etc.

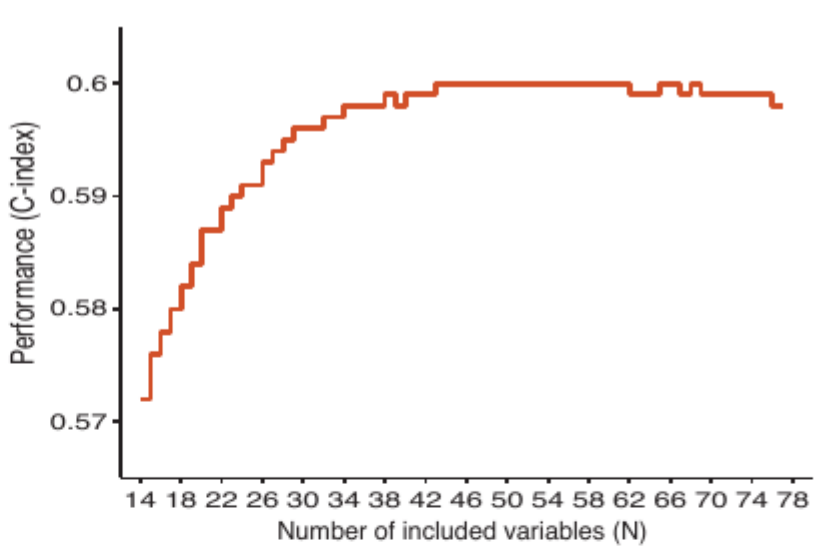
Study Population – ISHLT database

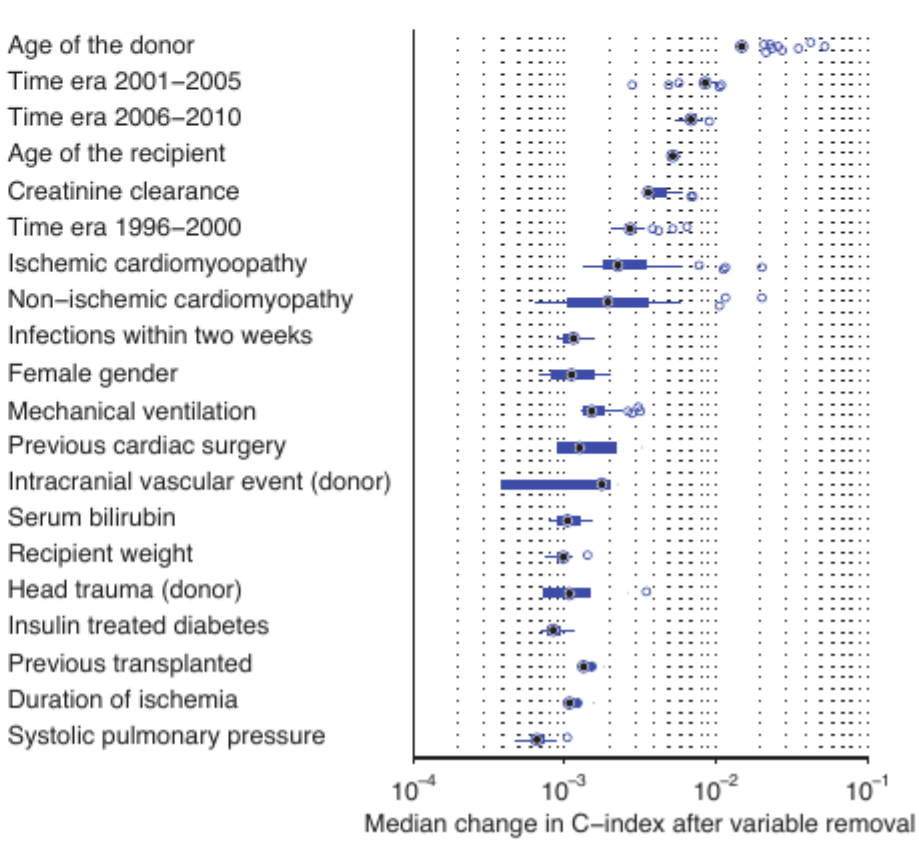
- 56 625 patients that have undergone a heart transplantation
- Mean age was 51 and 21% women.
- Mean follow-up duration of 5.2 years
- Overall 30-day mortality was 9% (n=5010)
- One-year mortality was 18% (n=9380)
- A total of 21 502 patients (38%) died during follow-up.
- Main cause of death was

    - late graft failure (3215)
    - major adverse cardiovascular events (2993)
    - infections (2656)

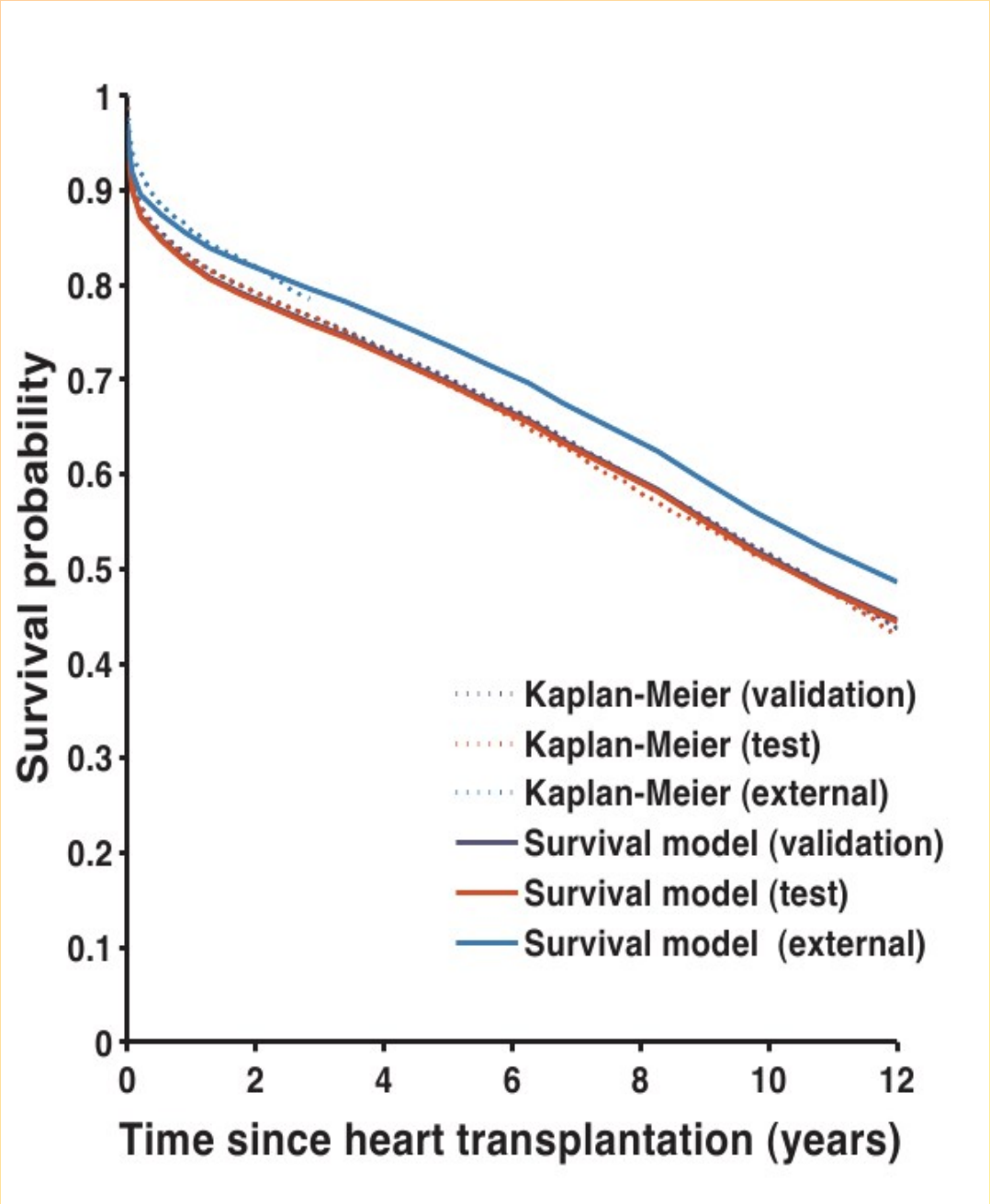
Also: Scandiatransplant, ~1300 patients, external validation


A
Repeat until one variable left (N-1)
Model training (I)
Repeat M times
Optimize generalization by utilizing a weight decay and calculate the performance on the validation sets including N variables
CV
5
10
CoM
Variable ranking (II)
Repeat N times once for each variable
Identification of least relevant variable by calculating the performance for each omitted variable
CV
5
10
Omit
CoM
Elimination of least relevant variable and recording of elimination order
1234N
Variable (no)
Building of a ranking list from the final elimination order
Final model includes the "top ranked" variables tested on independent test set
CV Cross Validation
CoM Committe Machine
validation patient
Training patient
Omitted variable


B
Performance (C-index)
0.6
0.59
0.58
0.57
14 18 22 26 30 34 38 42 46 50 54 58 62 66 70 74 78
Number of included variables (N)


C
Age of the donor
Time era 2001–2005
Time era 2006–2010
Age of the recipient
Creatinine clearance
Time era 1996–2000
Ischemic cardiomyoopathy
Non–ischemic cardiomyopathy
Infections within two weeks
Female gender
Mechanical ventilation
Previous cardiac surgery
Intracranial vascular event (donor)
Serum bilirubin
Recipient weight
Head trauma (donor)
Insulin treated diabetes
Previous transplanted
Duration of ischemia
Systolic pulmonary pressure
10^{-4}
10^{-3}
10^{-2}
10^{-1}
Median change in C–index after variable removal

We now have a model!



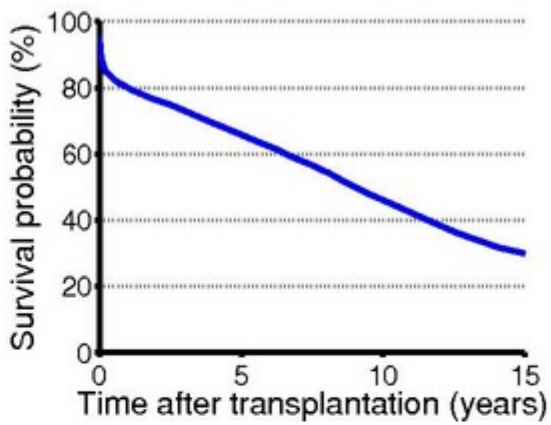
IHTSA

An International Heart Transplantation Survival Algorithm

Results

	1 year	5 years	10 years
Survival	82 %	68 %	48 %
Mortality	18 %	32 %	52 %

Median life expectancy 9.1 years



Recipient data

Diagnosis	Non-ischemic cardiomyopathy
Age	65
Gender	Male
Height	175
Weight	80
Insulin treated diabetes	
Hypertension	
Infection within two weeks	
Antiarrhythmic	
Amiodarone	
Recipient blood group	A
Previous blood transfusion	
Previously transplanted	
Previous cardiac surgery	
Intensive care unit	
Mechanical ventilation	
ECMO	
Ventricular assist device	
Transplant era	2006-
SPP (mmHG)	44
PVR (wood units)	2.6
Creatinine (µmol/l)	120
Serum bilirubin (mg/dl)	1.4
PRA > 10 %	
HLA-DR 2 mismatch	

Donor data

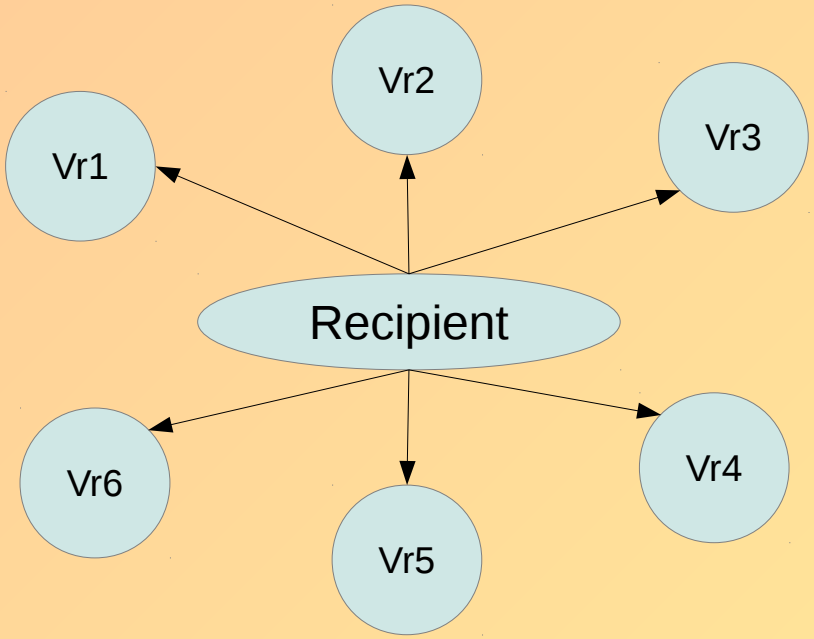
Age	65
Gender	Male
Height	175
Weight	80
Duration of ischemia (min)	186
Donor blood group	A
Donor cause of death	Head trauma

Submit Defaults

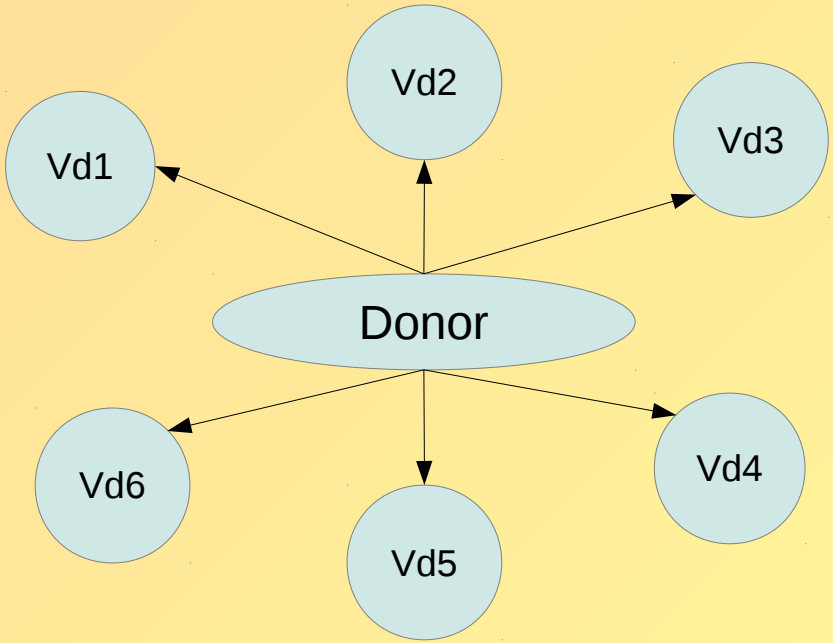
Can we learn something new?

- The model gives you a predicted survival curve for a donor-recipient pair.
- We can measure the “performance” for any given pair (both real and virtual).
- The area under *S(t)* is our measure of performance.

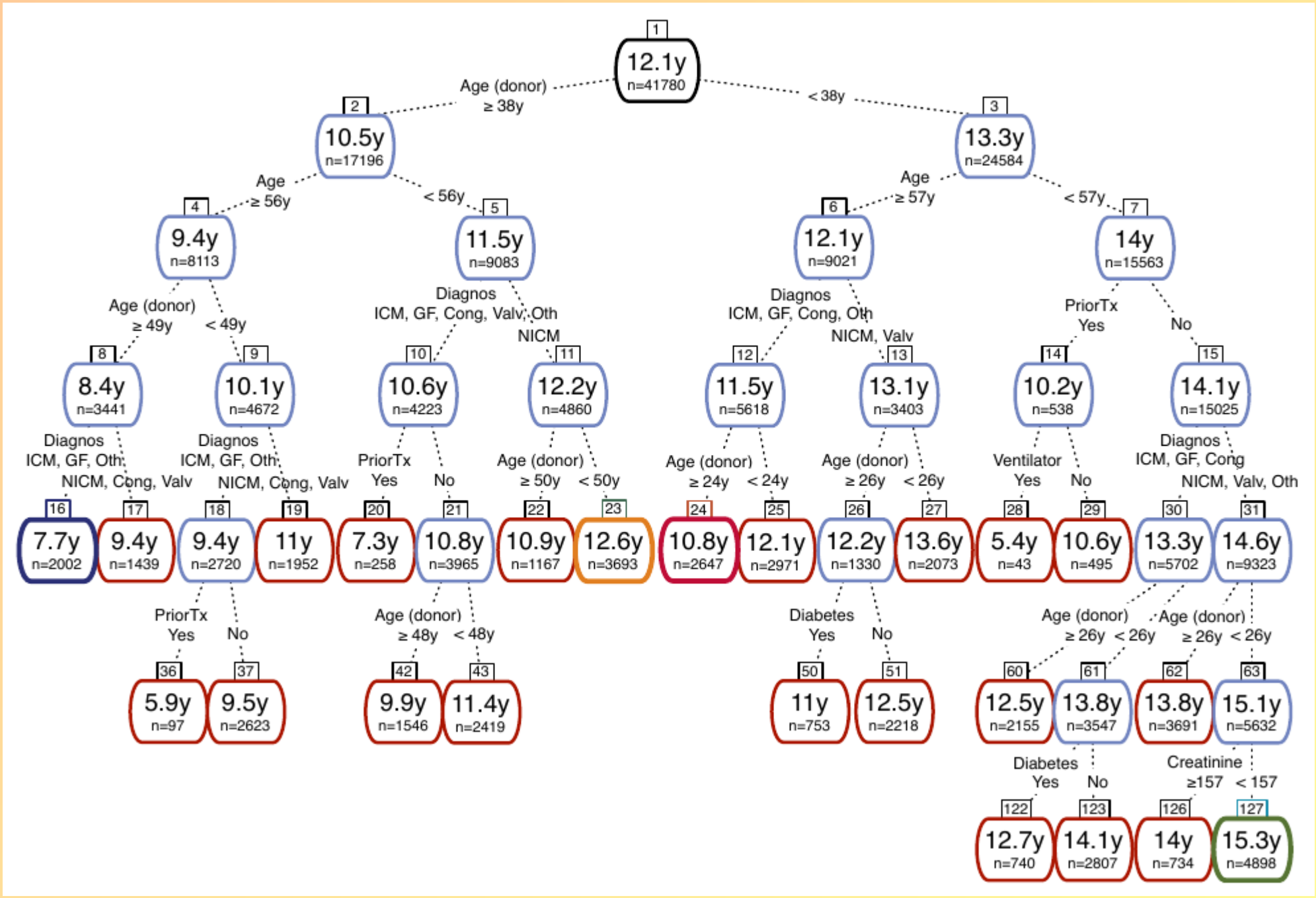
Virtual recipient-donor matching



Match all
combinations



Visualize the important combinations
using a regression tree


1
12.1y
n=41780
Age (donor)
≥ 38y
< 38y
2
10.5y
n=17196
3
13.3y
n=24584
Age
≥ 56y
< 56y
Age
≥ 57y
< 57y
4
9.4y
n=8113
5
11.5y
n=9083
6
12.1y
n=9021
7
14y
n=15563
Age (donor)
≥ 49y
< 49y
Diagnos
ICM, GF, Cong, Valv, Oth
NICM
Diagnos
ICM, GF, Cong, Oth
NICM, Valv
PriorTx
Yes
No
8
8.4y
n=3441
9
10.1y
n=4672
10
10.6y
n=4223
11
12.2y
n=4860
12
11.5y
n=5618
13
13.1y
n=3403
14
10.2y
n=538
15
14.1y
n=15025
Diagnos
ICM, GF, Oth
NICM, Cong, Valv
Diagnos
ICM, GF, Oth
NICM, Cong, Valv
PriorTx
Yes
No
Age (donor)
≥ 50y
< 50y
Age (donor)
≥ 24y
< 24y
Age (donor)
≥ 26y
< 26y
Ventilator
Yes
No
Diagnos
ICM, GF, Cong
NICM, Valv, Oth
16
7.7y
n=2002
17
9.4y
n=1439
18
9.4y
n=2720
19
11y
n=1952
20
7.3y
n=258
21
10.8y
n=3965
22
10.9y
n=1167
23
12.6y
n=3693
24
10.8y
n=2647
25
12.1y
n=2971
26
12.2y
n=1330
27
13.6y
n=2073
28
5.4y
n=43
29
10.6y
n=495
30
13.3y
n=5702
31
14.6y
n=9323
PriorTx
Yes
No
Age (donor)
≥ 48y
< 48y
Diabetes
Yes
No
Age (donor)
≥ 26y
< 26y
Age (donor)
≥ 26y
< 26y
36
5.9y
n=97
37
9.5y
n=2623
42
9.9y
n=1546
43
11.4y
n=2419
50
11y
n=753
51
12.5y
n=2218
60
12.5y
n=2155
61
13.8y
n=3547
62
13.8y
n=3691
63
15.1y
n=5632
Diabetes
Yes
No
Creatinine
≥157
< 157
122
12.7y
n=740
123
14.1y
n=2807
126
14y
n=734
127
15.3y
n=4898

C-index modeling

In-house development (Patrik E & Jonas K)

C-index (*concordance index*) is a performance measure for survival modeling (with censored data)

$$C = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \mathbf{1}_{\mathbf{f(x_i)<f(x_j)}}$$

$f(x_i) =$ predicted survival time for patient *i*.

$\Omega =$ All pairs of patients (i,j) such that:

- Both *i* and *j* have events and $t_i < t_j$
- Patient *i* have and event and t_i is smaller that patient *j*'s censor time

You do not have to predict the survival time. A prognostic index that can order the patients correctly is sufficient.

$$C = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \mathbf{1}_{\mathbf{f(x_i)<f(x_j)}}$$

The c-index is rank based measure = problems when optimizing the prediction model.

Our approach:

- Use neural networks to compute a prognostic index.

$$p(\mathbf{x}) = \sum_{j=1}^{J} \omega_j \cdot \varphi \left(\sum_{k=1}^{K} \tilde{\omega}_{jk} x_k + \tilde{\omega}_{j0} \right) + \omega_0$$

- Maximize the C-index with respect to model parameters

$$C = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \mathbf{1}_{\mathbf{p(x_i)>p(x_j)}}$$

- We use *genetic algorithms* for the optimization.

- Use en ensemble of networks rather than just a single one.

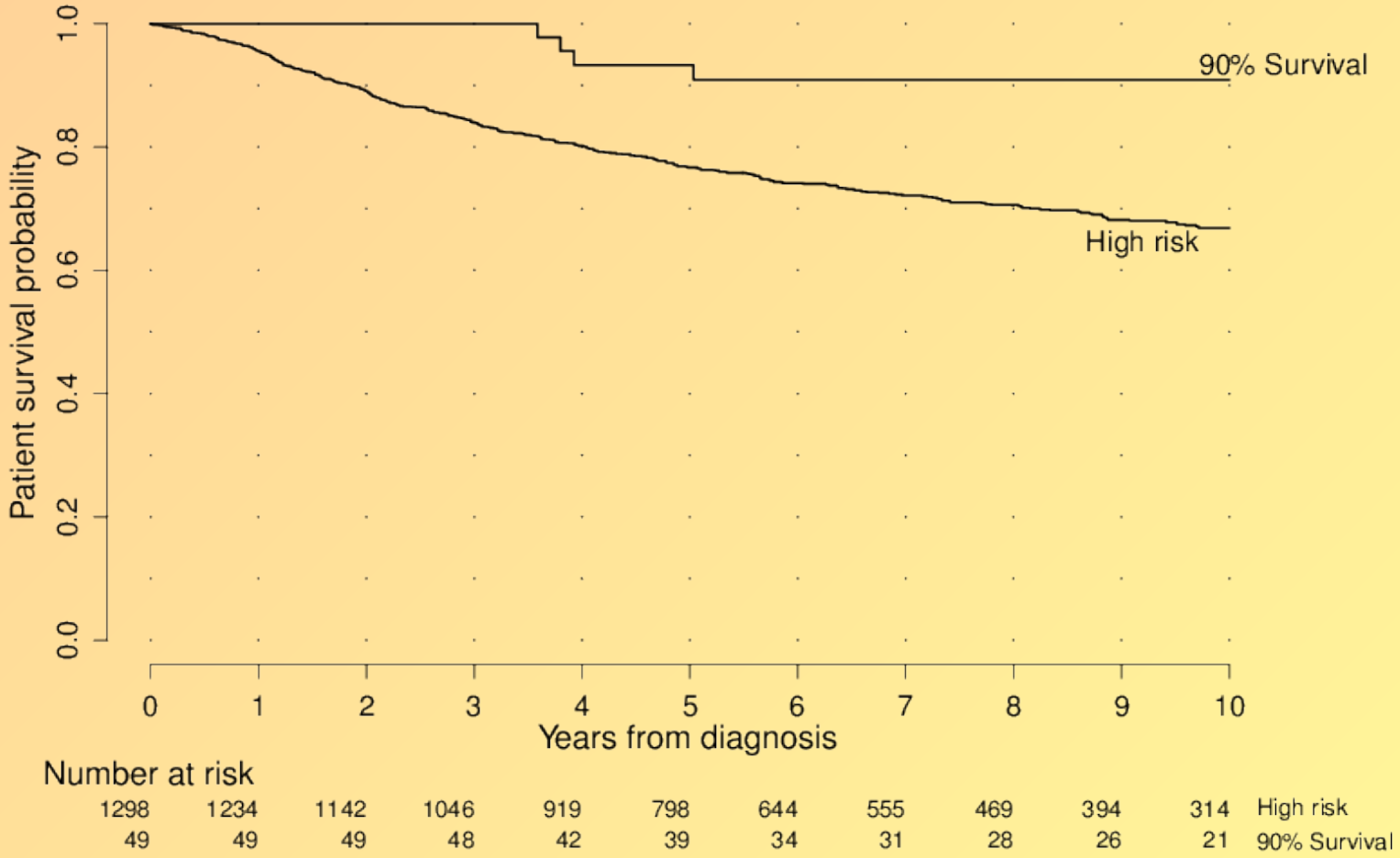
Clinical application

- ~ 4000 female patients with breast cancer, that have had removal of primary tumor.
- Recurrence for about 21% (after 5 years)
- Median age ~ 60 years.

Example of covariates: Age, tumor size, number of positve lymph nodes, HER2-status, histological grade.

Aim: Construct a prognostic index for recurrence

Split into high risk and low risk group based on the index



Extension: Predict actual survival times.

$$E = \frac{1}{N} \sum_n q_n (\hat{t}(\mathbf{x}_n) - t_n)^2$$

Machine learning model

Correct for censored cases $q_n \equiv \delta_n + (1 - \delta_n) \cdot I(t_n > \hat{t}_n)$

More extension: Use more information from the censored cases