Evaluation of dynamic risk prediction models

Thomas Alexander Gerds

Department of Biostatistics, University of Copenhagen

10 October 2013

Outline

Risk prediction models

Illustration: leukemia research

Dynamics

Evaluation

What is absolute risk

Absolute cancer risk is the probability that an individual with given risk factors and a given age will develop cancer over a defined period of time.

What is absolute risk

More generally: The absolute risk of an event is the probability that an individual with given risk factors and a given age will have the event within a defined period of time.

What is absolute risk

More generally: The absolute risk of an event is the probability that an individual with given risk factors and a given age will have the event within a defined period of time.

The absolute risk is a probability and has a direct interpretation for the single patient. The model is calibrated if we can expect that x out of 100 experience the event among all patients that receive a predicted risk of x%.

Conditional risk such as a hazard or a hazard ratio does not have an intuitive interpretation for prediction.

What is dynamic risk? (in cancer research)

Dynamic = changing, able to change and to adapt

For the patient

- Environment
- Treatment
- Disease

For the modeller

- Prediction time-point
- Prediction horizon
- Event status, measurements of biomarkers, treatment, questionnaire results, etc.

The purpose of a statistical model

Developing statistical models that estimate the probability of developing cancer over a defined period of time will help

- clinicians identify individuals at higher risk
- allowing for earlier or more frequent screening
- counseling of behavioral changes to decrease risk

These types of models also will be useful for designing future chemoprevention and screening intervention trials in individuals at high risk of specific cancers in the general population.[1]

$\mapsto$ personalized medicine

[1] National Cancer Institute

The making of a statistical risk prediction model

1. A statistical model specifies the relation between the absolute risk and all risk factors including biomarkers and treatment through mathematical functions and a priori unknown parameters such as regression coefficients.

2. Lists of potential risk factors are "screened" (dimension-reduction).

3. The model is "fitted" to a training data set which contains measurements of risk factors and outcome of earlier patients.

4. The model is "validated": internally via crossvalidation and externally using independent validation data.

Acute leukemia patients[2]

Eosinophilia is a condition in which the eosinophil count in the peripheral blood exceeds 0.45 x 109/L. Several studies have focused on the prognostic impact of eosinophilia on transplant outcome.

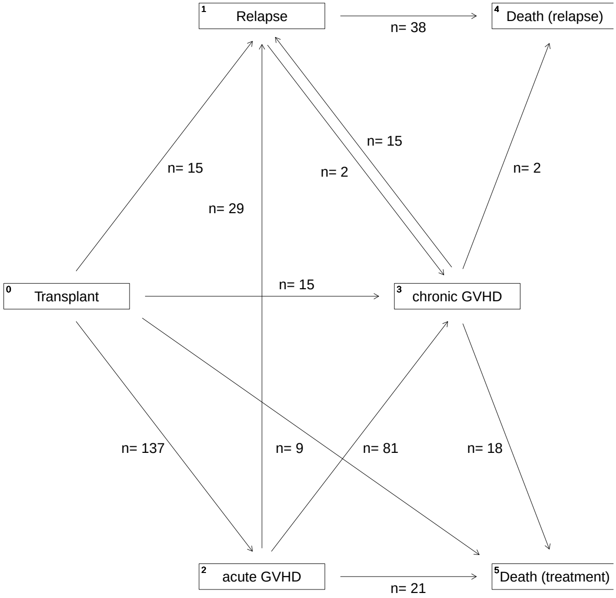
The prognostic significance of eosinophilia after myeloablative allogeneic stem cell transplantation and the relationship between chronic graft-versus-host disease (cGVHD) and concomitant eosinophilia remain to be established.

We retrospectively collected data from patients who developed cGVHD after having received allogeneic stem cell transplantation

We analysed times of events after the onset of cGVHD.

[2] Research with Katrine Mortensen and Christen Lykkegaard Andersen, Department of Hematology, Copenhagen University Hospital

Outcome


1
Relapse
n= 38
4
Death (relapse)
n= 15
n= 29
n= 2
n= 15
n= 2
0
Transplant
n= 15
3
chronic GVHD
n= 137
n= 9
n= 81
n= 18
2
acute GVHD
n= 21
5
Death (treatment)

Risk factors not changing over time

Information available at transplant date:

Variable	Level	Female	Male	Total	P-value
n		51	91	142	
Age groups	< 20	2(15.4)	11(84.6)	13	
	$20 - 40$	27(37.0)	46(63.0)	73	
	> 40	22(39.3)	34(60.7)	56	0.26008
Disease	MDS	6(46.2)	7(53.8)	13	
	CML	5(17.2)	24(82.8)	29	
	AML	26(46.4)	30(53.6)	56	
	ALL	10(27.0)	27(73.0)	37	
	SAA	2(50.0)	2(50.0)	4	
	Other	2(66.7)	1(33.3)	3	0.06172
Donor/recipient sex	FDFR/MDMR/MDFR	51(47.7)	56(52.3)	107	
	FDMR	0(0.0)	35(100.0)	35	< 0.0001
Donor relation	Matched unrelated	25(32.5)	52(67.5)	77	
	HLA identical	26(40.0)	39(60.0)	65	0.44930
HLA match	mismatch	0(0.0)	10(100.0)	10	
	match	51(38.6)	81(61.4)	132	0.03455
Type of Marrow	PBSCT	28(38.4)	45(61.6)	73	
	BMT	23(33.3)	46(66.7)	69	0.65375
...	...	...	...	...	...
...	...	...	...	...	...

Risk factors changing in time

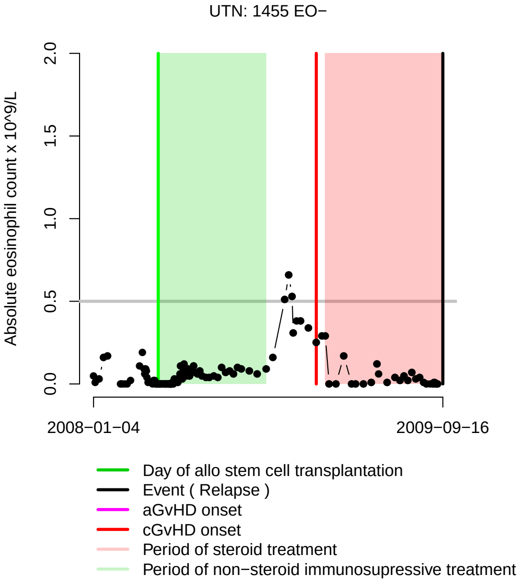
Blood test results (EOS= eosinophil count x 10^9/L)

UTN	Date	EOS	IGG	lymfo	THROM
1004	2003-03-11	0.01	7.5	1.1	297
1004	2003-03-18	0.01	7.1	1.1	274
1004	2003-03-24	0.01	7.8	1.5	216
1004	2003-03-26	0.03	6.4	1.6	224
1004	2003-04-03	0.03	5.9	0.4	200
1004	2003-04-09	0.10	5.5	0.2	188
...	...	...	...	...	...
...	...	...	...	...	...

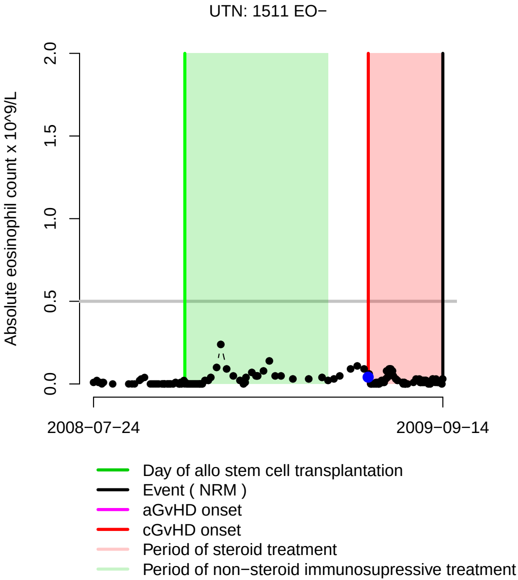
Treatment: conditioning regimes, steriod treatment, ...

Disease: relapse, graft-versus-host disease

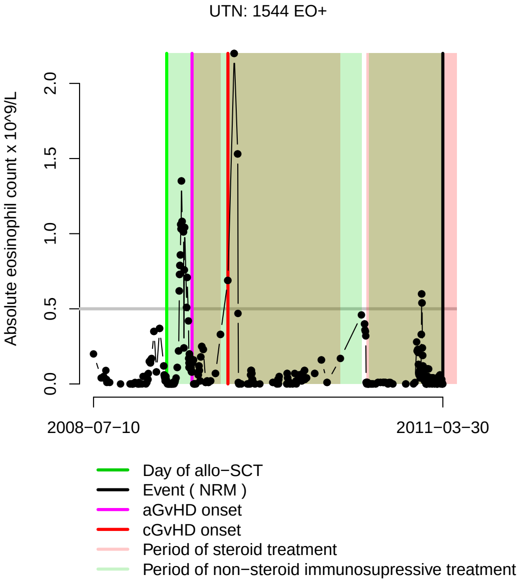
Sample patient: eosinophilia process



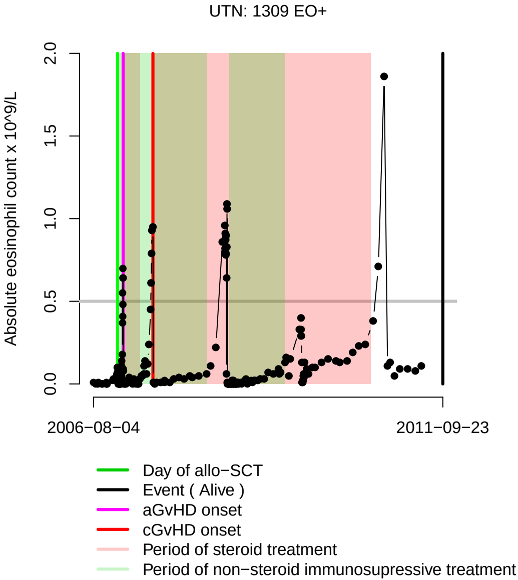
Sample patient: eosinophilia process



Sample patient: eosinophilia process



Sample patient: eosinophilia process



Time origin and prediction horizon

The time origin is the start of followup.

A landmark time is another day where the clinician/patient are interested in predicted risk of future events.

Examples:

- date of diagnosis
- date of treatment
- date when a new screening result or blood test becomes available
- xx-years after the initial treatment

The prediction horizon defines the length of the time period after the time origin (in which the predicted risk is calibrated).

Moving from the time-origin to a landmark time

```
|--------------+---------------+----------------->
0              s               E                 t
origin     landmark          event           horizon
```

The following changes at the landmark time s:

- sample size – dead patients are excluded
- inclusion criteria, e.g. only patients with diagnosed chronic graft versus host disease
- length of remaining followup
- the longitudinal marker history
- the age of non-updating measurements

Moving from the time-origin to a landmark time

```
|--------------+---------------+----------------->
0              s               E                 t
origin     landmark          event           horizon
```

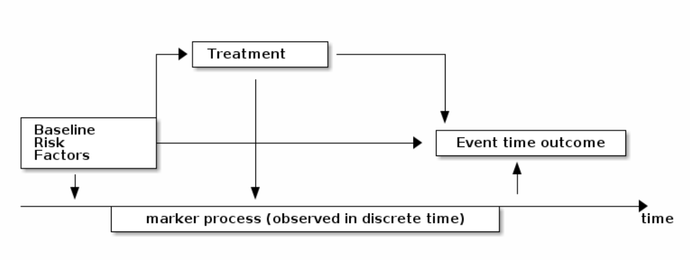
Modelling options at the landmark time s:

- new selection of important variables
- new evaluation of the modelling assumptions: proportional hazard, functional form for covariate effects, etc.
- re-estimation of regression parameters
- re-calibration of the model

A general picture

```
 _   _             _                                 _ 
| | | |_ __   ___ | |__  ___  ___ _ ____   _____  __| |
| | | | '_ \ / _ \| '_ \/ __|/ _ \ '__\ \ / / _ \/ _` |
| |_| | | | | (_) | |_) \__ \  __/ |   \ V /  __/ (_| |
 \___/|_| |_|\___/|_.__/|___/\___|_|    \_/ \___|\__,_|

                  __                       _               
  ___ ___  _ __  / _| ___  _   _ _ __   __| | ___ _ __ ___ 
 / __/ _ \| '_ \| |_ / _ \| | | | '_ \ / _` |/ _ \ '__/ __|
| (_| (_) | | | |  _| (_) | |_| | | | | (_| |  __/ |  \__ \
 \___\___/|_| |_|_|  \___/ \__,_|_| |_|\__,_|\___|_|  |___/
```



Modelling strategies

- Survival regression: combine baseline risk factors with history of updating risk factors, treatment and outcome up to the landmark (Cox, Fine-Gray, etc.).

- Two stage models: first summarize the history of each patient's longitudinal marker process, then include this summary as a new risk factor.

- Joint models: model for the joint distribution of the longitudinal markers, the event time outcome, and the baseline risk factors. Person-parameters: random effects, frailty . . . prediction?

- Multi-state models: model transitions between time-dependent
    - *intermediate events*, e.g. graft-versus-host disease, relapse
    - *terminal events*, e.g. death.

  Use the data of the training patients to learn about the likelihood of the possible pathways and apply this knowledge for predicting new patients.

Predictions

Obtaining predictions:

- Explicit formula: translate predictions of hazards and joint models into a prediction of the absolute risk of the event.
- Discrete event simulation: based on regression models for the transition intensities computer simulate the likelihood of possible pathways through the multi-state model.

Updating predictions: at the landmark time we may

- update modelling, including screening for risk factors, tuning, and estimation of parameters
- do not update the model, only the longitudinal markers, and use a formula:

  Prob(E in (s,t]) = Prob(E before t) - Prob(E before s).

Evaluating predictions:

- prediction error, discrimination, calibration

Performance measures

$$Y_i(s,t) = \begin{cases} 0 & \text{patient i is event free between landmark s and horizon t} \\ 1 & \text{patient i has the event between landmark s and horizon t} \end{cases}$$

$\hat{R}_i(s,t) =$ absolute risk of event before t predicted at s for patient i

---

Mean squared error (Brier score)

$$\text{expected BrierScore}(s,t) = \frac{1}{\tilde{n}} \sum_{i \in \text{testset}:\, T_i > s} \{Y_i(s,t) - \hat{R}_i(s,t)\}^2$$

measures both discrimination and calibration

Discrimination ability[3]

$$\text{AUC}(s,t) = \frac{\sum_i \sum_j I\{Y_i(s,t) = 0, Y_j(s,t) = 1, \hat{R}_i(s,t) < \hat{R}_j(s,t)\}}{\sum_i \sum_j I\{Y_i(s,t) = 0, Y_j(s,t) = 1\}}$$

[3] Interpretation? Conditioning on the future?

Technical complications

Censored data

If some (test set) patients are lost to followup[4] before the event of interest has happened then we can

- either use inverse probability of censoring weights
- or pseudo values

to achieve that (the expected value of) the prediction performance summary measure remains (asymptotically) unaffected.

Internal validation

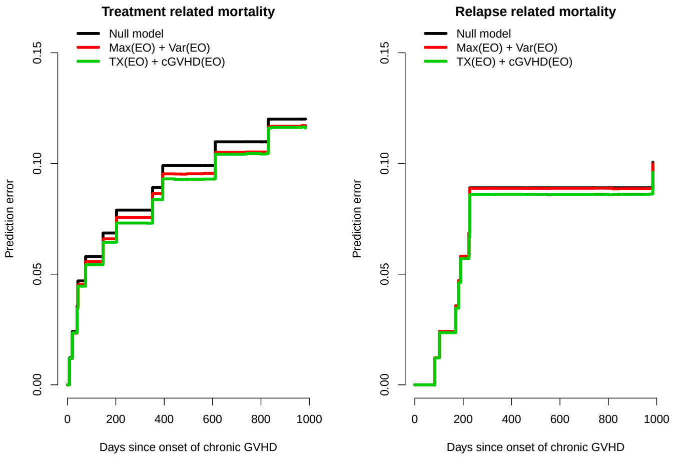
Data splitting can be applied and summarized via

- cross-validation
- bootstrap (.632+ method)

to estimate the expected performance of the model in new patients.

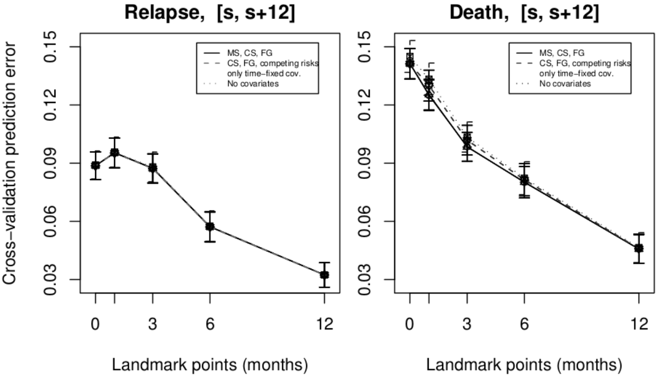
[4] Death is a competing risk and requires to be predicted.

Effect of eosinophilia: Brier score landmark analysis at date of cGVHD



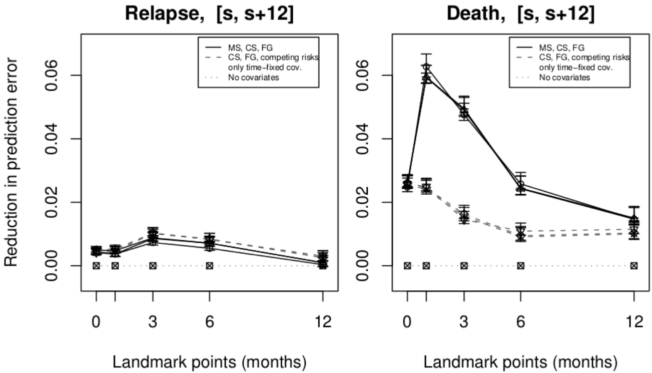
Results from Cortese et al. 2013 (1)

Comparison of multi-state model (MS) and landmark analyses (FG, CS) for fixed time-horizon and varying landmark times.



Results from Cortese et al. 2013 (2)

Relative gain in Brier score



Concluding comments (I)

- The definition and evaluation of prediction performance can quickly be adapted to landmarking and dynamic risk prediction.
- Risk, prediction and model performance depend on two time-scales: landmark time and time horizon
- Longitudinal measurements can potentially improve predictions. Changing treatments and feedback complicate this in non-randomized studies.
- We can (as usual) compare different modeling strategies with respect to the predictive performances of the resulting risk prediction models.

Concluding comments (II)

The usual question: What is the best model? has some new flavors

- Is it worth to update the risk factors at the landmark?
- Is it worth to screen and model the risk factors again at each landmark?
- Is it worth to re-estimate the regression parameters?
- Is a joint model able to outperform a two-stage model in terms of prediction ability?