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
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.\(^1\)

\[\rightarrow\text{ personalized medicine}\]

\(^1\text{National Cancer Institute}\)
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.
Illustration: leukemia research
Acute leukemia patients

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.

²Research with Katrine Mortensen and Christen Lykkegaard Andersen, Department of Hematology, Copenhagen University Hospital
## Risk factors not changing over time

Information available at transplant date:

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

...
Risk factors changing in time

Blood test results (EOS = eosinophil count $\times 10^9$/L)

<table>
<thead>
<tr>
<th>UTN</th>
<th>Date</th>
<th>EOS</th>
<th>IGG</th>
<th>lymfo</th>
<th>THROM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1004</td>
<td>2003-03-11</td>
<td>0.01</td>
<td>7.5</td>
<td>1.1</td>
<td>297</td>
</tr>
<tr>
<td>1004</td>
<td>2003-03-18</td>
<td>0.01</td>
<td>7.1</td>
<td>1.1</td>
<td>274</td>
</tr>
<tr>
<td>1004</td>
<td>2003-03-24</td>
<td>0.01</td>
<td>7.8</td>
<td>1.5</td>
<td>216</td>
</tr>
<tr>
<td>1004</td>
<td>2003-03-26</td>
<td>0.03</td>
<td>6.4</td>
<td>1.6</td>
<td>224</td>
</tr>
<tr>
<td>1004</td>
<td>2003-04-03</td>
<td>0.03</td>
<td>5.9</td>
<td>0.4</td>
<td>200</td>
</tr>
<tr>
<td>1004</td>
<td>2003-04-09</td>
<td>0.10</td>
<td>5.5</td>
<td>0.2</td>
<td>188</td>
</tr>
</tbody>
</table>

...  ...

Treatment: conditioning regimes, steroid treatment, ...

Disease: relapse, graft-versus-host disease
Sample patient: eosinophilia process

UTN: 1455 EO−

- Day of allo stem cell transplantation
- Event (Relapse)
- aGvHD onset
- cGvHD onset
- Period of steroid treatment
- Period of non-steroid immunosuppressive treatment
Sample patient: eosinophilia process

UTN: 1511 EO−

Day of allo stem cell transplantation
Event ( NRM )
aGvHD onset
cGvHD onset
Period of steroid treatment
Period of non−steroid immunosupressive treatment
Sample patient: eosinophilia process

UTN: 1544 EO+

Absolute eosinophil count x 10^9/L

2008−07−10 2011−03−30

0.0 0.5 1.0 1.5 2.0

●
●●●●
●●● ● ... )

Day of allo−SCT
Event ( NRM )
aGvHD onset
cGvHD onset
Period of steroid treatment
Period of non−steroid immunosupressive treatment
Sample patient: eosinophilia process

UTN: 1309 EO+

- Absolute eosinophil count x 10^9/L
- Day of allo-SCT
- Event (Alive)
- aGvHD onset
- cGvHD onset
- Period of steroid treatment
- Period of non-steroid immunosuppressive treatment

Date Range: 2006-08-04 to 2011-09-23
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

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>s</td>
</tr>
<tr>
<td>origin</td>
<td>landmark</td>
</tr>
<tr>
<td>event</td>
<td>horizon</td>
</tr>
</tbody>
</table>

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                |
| 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
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:

\[ \text{Prob}(E \text{ in } (s,t]) = \text{Prob}(E \text{ before } t) - \text{Prob}(E \text{ before } s). \]

Evaluating predictions:

▶ prediction error, discrimination, calibration
Performance measures

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

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

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\(^3\)

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

\(^3\)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

**Treatment related mortality**

- Null model
- Max(EO) + Var(EO)
- TX(EO) + cGVHD(EO)

**Relapse related mortality**

- Null model
- Max(EO) + Var(EO)
- TX(EO) + cGVHD(EO)
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

Relapse, [s, s+12]

Death, [s, s+12]
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?