Predictive modelling of drug treatment responses

Tero Aittokallio, PhD, EMBL Group Leader
Professor of Statistics and Applied Mathematics
University of Turku, Finland

Institute for Molecular Medicine Finland (FIMM)
University of Helsinki, Finland
Aittokallio’s research group: systems medicine modeling and applications for revealing molecular mechanisms behind disease processes and predicting phenotypic responses to genetic and chemical perturbations

Genomic profiling
Drug screening

Network modeling
Predictive modeling

Genetic risk variants
Personalized therapies

Differential drug sensitivity
Effective drugs
Resistant drugs
Individualized treatment of leukemia patients

Ex vivo drug sensitivity and resistance testing (DSRT)

Compendia of drug responses

Genomics and molecular profiling

Integrated DSRT, molecular profile and clinical database

Patterns of response highlighting MOA, new drugs and biomarkers

Driver signals and pathways for cancer

Patient-specific treatment recommendations

Somatic mutations and cell signaling patterns

Technologies

Results

Patient samples & clinical data: DIAGNOSIS & RELAPSE
Benefits of ML-based predictive modelling (for personalized medicine)

- **Supervised machine learning** models aim to maximize the predictive (generalizability) power at the level of individuals, hence providing opportunities for e.g., individualized risk prediction based on personal genetic and clinical profiles.

- Machine learning models can effectively deal with interactions between model variables (e.g., *genetic interactions*), which are known to play a role in the development of many complex diseases, but are often missed by univariate association tests.

- Many applications of predictive modeling rely on effective *variable selection*, often implemented through model regularization, which penalizes the model complexity and enables predictions in individuals outside of training data.
Regularized Machine Learning in the Genetic Prediction of Complex Traits

Sebastian Okser, Tapio Pahikkala, Antti Airola, Tapio Salakoski, Samuli Ripatti, Tero Aittokallio

1 Department of Information Technology, University of Turku, Turku, Finland, 2 Turku Centre for Computer Science (TUCS), University of Turku and Åbo Akademi University, Turku, Finland, 3 Hjelt Institute, University of Helsinki, Helsinki, Finland, 4 Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland, 5 Wellcome Trust Sanger Institute, Hinxton, United Kingdom

A  T2D classification

B  Continuous trait regression

C  Area Under the Curve (AUC) vs. Number of Selected Features

D  Coefficient of Determination (R²) vs. Number of Selected Features
Accompanying pdf
12 page supplementary detailing implementation issues for many machine learning models for SNP-based predictors and beyond

Box 1. Synthesis of Learning Models for Genetic Risk Prediction

such as $L_1$ and $L_2$ norms for model weights (Figure 2A). Together with the squared loss function (Figure 2B), which is often used to measure the fit between the observed $y_j$ and estimated $\hat{y}_j$, these functional norms give rise to the optimization problem used in various types of linear genetic risk prediction models:

$$\sum_{j=1}^{n} (y_j - \hat{y}_j)^2 + \lambda_1 \sum_{j=1}^{p} |w_j| + \lambda_2 \sum_{j=1}^{p} w_j^2. \quad (2)$$

Ridge regression is the special case of Eq. 2, in which $\lambda_2 = 0$, and the regularization parameter $\lambda_1$ is used to shrink the variable weights toward zero to prevent any particular variable from having too large an effect on the model. However, the use of $L_2$ penalty alone tends to favor models that depend on all the variables. In Lasso, $\lambda_2 = 0$, and through adjusting the regularization parameter $\lambda_1$, it is possible to favor sparse models with only a few nonzero weights, leading to variable selection within the model fitting [82]. The Elastic Net model makes use of both penalty terms $L_1$ and $L_2$ to select also correlated features [83]; for instance, groups of variants within a pathway that together contribute to the predictive accuracy.

From these we can define the derivatives needed for finding the coordinate descent step directions for a number of loss functions (see table 4). Based on these and the various regularizers listed in Table 5 we can construct a number of well-known machine learning algorithms, as seen in Table 6. Note that, while the zero-norm is not differentiable, the corresponding constraint is still satisfied with the greedy coordinate descent based training methods. Different algorithms vary on their respective computational complexities (see Table 7).

<table>
<thead>
<tr>
<th>Name</th>
<th>$L(p, y)$</th>
<th>$\frac{\partial L(p, y)}{\partial p}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Squared</td>
<td>$(p - y)^2$</td>
<td>$2(p - y)$</td>
</tr>
<tr>
<td>Logistic</td>
<td>$\log(1 + e^{-yp})$</td>
<td>$\frac{-y}{1+e^{yp}}$</td>
</tr>
<tr>
<td>Hinge</td>
<td>$\max(0, 1 - yp)$</td>
<td>$0$ if $y \geq 1$, $-y$ else</td>
</tr>
</tbody>
</table>

Table 4: Common loss functions

<table>
<thead>
<tr>
<th>Name</th>
<th>$\Omega(w)$</th>
<th>$\frac{\partial \Omega(w)}{\partial w_j}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell_0$</td>
<td>$|w|_0$</td>
<td></td>
</tr>
<tr>
<td>$\ell_1$</td>
<td>$|w|_1$</td>
<td>$\sign(w_j)$</td>
</tr>
<tr>
<td>$\ell_2$</td>
<td>$|w|_2$</td>
<td>$2w_j$</td>
</tr>
</tbody>
</table>

Table 5: Common regularizers

<table>
<thead>
<tr>
<th>Method</th>
<th>S</th>
<th>L</th>
<th>H</th>
<th>$\ell_1$</th>
<th>$\ell_2$</th>
<th>$\ell_0$</th>
<th>R/C</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R/C</td>
<td>[19]</td>
</tr>
<tr>
<td>Elastic Net</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td>[20]</td>
</tr>
<tr>
<td>$\ell_1$ Logistic</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>[21]</td>
</tr>
<tr>
<td>$\ell_2$ Logistic</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>[22]</td>
</tr>
<tr>
<td>$\ell_1$ SVM</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>[23]</td>
</tr>
<tr>
<td>SVM</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>[24]</td>
</tr>
<tr>
<td>OLS</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td>[25]</td>
</tr>
<tr>
<td>Greedy RLS</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>[12]</td>
</tr>
<tr>
<td>Ridge Reg</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td>[26]</td>
</tr>
<tr>
<td>GLS</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td>[11]</td>
</tr>
</tbody>
</table>

6: Construction of various methods based on different loss functions and regularizers. $S$, $L$ and $H$ stand squared loss, logistic loss and hinge loss, respectively. R/C denotes whether the method is a (R)egression (C)lassification method only, all the regression methods can also be used for classification. OLS stands ordinary least squares and Ridge Reg for ridge regression. GLS represents greedy least squares. These notations may also be known by other names in the literature, for example ridge regression is also known as regularized least squares.

\textsuperscript{5} be exact, the hinge loss and $\ell_1$ norm are not differentiable everywhere, for these we provide subderivatives.

\textsuperscript{4} for details on these notations, see the relevant literature.
DREAM Challenges pose fundamental questions about systems biology and translational medicine – **crowdsourced collaborative challenges**

Designed and run by a community of researchers from a variety of organizations, invite participants to propose solutions — **machine learning and statistical modelling**

Expertise and institutional support are provided by Sage Bionetworks, e.g., the infrastructure providing training and testing data — **systematic development and objective validation of the predictive models**

Allowing individuals and groups to collaborate openly so that the “wisdom of the crowd” provides the greatest impact on science and human health.

http://dreamchallenges.org/
Part I - Mining of panels of molecular and genomic biomarkers that are predictive of drug (combination) responses in cancer cell line models \textit{(in vitro)} or in patient-derived samples \textit{(ex vivo)}
Single markers cannot predict drug response

(a) Ruxolitinib ex vivo response

$\gamma = 1.447$
$p = 0.024$

- Red: Highly sensitive AML samples
- Green: Controls

(b) Entinostat ex vivo response

$\gamma = 1.303$
$p = 0.033$

- Red: Highly sensitive AML samples
- Green: Controls

Legend:
- FLT3-ITD
- WT1
- TP53
- N/KRAS
- DNMT3A
- PTPN11
- RUNX1
- IDH1/IDH2
- NPM1
- c-KIT
- NUP98-NSD1
- MLL-X fusions
- ETV6-NTRK3
- FAB subtypes
- Disease stage
NCI/DREAM challenge: predicting exceptional drug responses using combinations of genomic features

**Machine learning models**

- Random forests
- Regression models
- Elastic net regression
- Regularized least squares
- Support vector machines
- Bayesian multilabel models
- Etc.

Improved machine learning algorithms for in-depth mining of the most predictive biomarkers and models for individual-level drug response (*in vitro or ex vivo*)
NCI/DREAM challenge: predicting exceptional drug responses using combinations of genomic features

Improved machine learning algorithms for in-depth mining of the most predictive biomarkers and models for individual-level drug response (ex vivo and in vivo)

Improved knowledge of the drug-target interactions and activity status of the target pathways should lead to more accurate models and marker combinations
A community effort to assess and improve drug sensitivity prediction algorithms

James C Costello\textsuperscript{1,2,3,4}, Laura M Heiser\textsuperscript{3,4}, Elisabeth Georgii\textsuperscript{4,5}, Mehmet Gönen\textsuperscript{4}, Michael P Menden\textsuperscript{5}, Nicholas J Wang\textsuperscript{3}, Mukesh Bansal\textsuperscript{6}, Muhammad Ammad-ud-din\textsuperscript{4}, PetteriHintsanen\textsuperscript{7}, Suleiman A Khan\textsuperscript{4}, John-Patrick Mpindi\textsuperscript{7}, Olli Kallioniemi\textsuperscript{7}, Antti Honkela\textsuperscript{8}, Tero Aittokallio\textsuperscript{7}, Krister Wennerberg\textsuperscript{7}, NCI DREAM Community\textsuperscript{9}, James J Collins\textsuperscript{1,2,10}, Dan Gallahan\textsuperscript{11}, Dinah Singer\textsuperscript{11}, Julio Saez-Rodriguez\textsuperscript{5}, Samuel Kaski\textsuperscript{4,8}, Joe W Gray\textsuperscript{3} & Gustavo Stolovitzky\textsuperscript{12}

DECEMBER 2014 NATURE BIOTECHNOLOGY

b

Cell lines (training data set)

Trastuzumab (ERBB2)
Disulfiram (ALDH2)
Methylglyoxal (pyruvate)
Cetuximab (EGFR)
GW5074 (RAF1)
Nelfinavir (protease)
TCS PIM-11 (PIM1)
Chloroquine (autophagy)
Imatinib (BCR-ABL)
PD184352 (MEK)
Nilotinib (BCR-ABL)
Mebendazole (tubulin)
Doxorubicin (TOP2A)
Baicalin (CYP2C9)
Olimomucine II (CDK1)
IKK 16 (IKK2)
PS-1145 (IKK)
MG-132 (proteasome)
MG-132b (proteasome)
MG-115 (proteasome)
4-HC (DNA alkylator)
4-HC+Dox (combination)
Bromopyruvate (glycolysis)
BS81 (FPhase)
FR182004 (ERK)
QNZ (NFkB)
Valproate (HDAC)
Everolimus (mTOR)

Cell lines (test data set)
**Winner:** Bayesian multitask multiple kernel learning

Modeling nonlinearities was a common in the top-performing methods
Use of prior information (e.g., biological pathways) helped predictions
Prediction accuracies only slightly better than random guessing in CV setup
Kernelized Bayesian Matrix Factorization (KBMF)

22 AML cell lines x 139 FIMM compounds

Task: Predict missing drug response values
– Microarray: gene expression (cont)
– Exome-seq: somatic mutations (bin)
– Compounds: structural properties (bin)
KBMF with component-wise multikernel learning

Prior information: cancer pathways from MSigDB database

doi: 10.1093/bioinformatics/btw433
Model validation and pathway-response links

Experimental validation of the model prediction in laboratory experiments at FIMM (6 AML cell lines and 8 drugs)

Sp. Cor = 0.44

Experimental validation of the model prediction in laboratory experiments at FIMM (6 AML cell lines and 8 drugs)
Feature selection - Bayesian Lasso regression (BLR)

**BLR**: Identifies the most predictive Biomarkers for Drug Screen

Drug sensitivity testing *ex vivo* cannot be implemented for each cancer patient -> need for response-predictive biomarkers

What are the most predictive omics assays (Exome/RNA-seq, CNV, MS/RPPA-based proteomics, metabolomics, etc)?

Mehreen Ali & Suleiman Khan
Predictive power of gene expression (RNA-seq) vs. somatic point mutations (Exome-seq)

Higher = Better

Prediction Correlation

Gene Expression
Mutation
Histology

MEK Inhibitors
EGFR Inhibitors

Suleiman Ali Khan & Astrid Murumägi
Part II - Mining of the clinical variables and their combinations most predictive of cancer patient clinical responses (*in vivo*)

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>DEATH</th>
<th>LKADT_P</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASC-002-0008</td>
<td>0</td>
<td>553</td>
</tr>
<tr>
<td>ASC-002-0011</td>
<td>1</td>
<td>467</td>
</tr>
<tr>
<td>ASC-002-0014</td>
<td>0</td>
<td>399</td>
</tr>
<tr>
<td>CELG-00024</td>
<td>1</td>
<td>129</td>
</tr>
<tr>
<td>CELG-00025</td>
<td>0</td>
<td>81</td>
</tr>
<tr>
<td>CELG-00026</td>
<td>0</td>
<td>231</td>
</tr>
<tr>
<td>VEN-336000702</td>
<td>1</td>
<td>736</td>
</tr>
<tr>
<td>VEN-336000801</td>
<td>1</td>
<td>451</td>
</tr>
<tr>
<td>VEN-336100401</td>
<td>1</td>
<td>409</td>
</tr>
<tr>
<td>AZ-00070</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AZ-00071</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AZ-00072</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>ALP</th>
<th>ALT</th>
<th>ANALGESICS</th>
<th>ALPxALT</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>6.18</td>
<td>2.40</td>
<td>0</td>
<td>14.83</td>
</tr>
<tr>
<td>...</td>
<td>5.08</td>
<td>2.71</td>
<td>0</td>
<td>13.76</td>
</tr>
<tr>
<td>...</td>
<td>4.45</td>
<td>2.30</td>
<td>0</td>
<td>10.26</td>
</tr>
<tr>
<td>...</td>
<td>6.38</td>
<td>2.64</td>
<td>0</td>
<td>16.85</td>
</tr>
<tr>
<td>...</td>
<td>4.60</td>
<td>4.36</td>
<td>1</td>
<td>20.02</td>
</tr>
<tr>
<td>...</td>
<td>5.48</td>
<td>3.18</td>
<td>1</td>
<td>17.42</td>
</tr>
<tr>
<td>...</td>
<td>6.07</td>
<td>3.04</td>
<td>1</td>
<td>18.48</td>
</tr>
<tr>
<td>...</td>
<td>5.08</td>
<td>3.43</td>
<td>1</td>
<td>17.43</td>
</tr>
<tr>
<td>...</td>
<td>4.43</td>
<td>2.40</td>
<td>1</td>
<td>10.62</td>
</tr>
<tr>
<td>...</td>
<td>5.78</td>
<td>1.61</td>
<td>1</td>
<td>9.31</td>
</tr>
<tr>
<td>...</td>
<td>5.24</td>
<td>2.64</td>
<td>0</td>
<td>13.83</td>
</tr>
<tr>
<td>...</td>
<td>4.64</td>
<td>2.30</td>
<td>0</td>
<td>10.69</td>
</tr>
</tbody>
</table>
• **Challenge:** predicting the overall survival of patients with metastatic castration resistant prostate cancer (mCRPC)

• **Material:** > 150 clinical variables from four phase III clinical trials with over 2,000 mCRPC patients treated with docetaxel

• **Project Data Sphere** provides one place where the cancer community can broadly share, integrate, analyze and discuss comparator arm data sets from multiple providers
Challenge Research Questions

Q1. Predict overall survival for metastatic, castrate-resistant prostate cancer patients based on patients’ clinical variables.

Q2. Predict treatment discontinuation of docetaxel treatment due to adverse events at early time points (< 3 months)

Chemotherapy naïve metastatic Castration Resistant Prostate Cancer (mCRPC) patients treated with docetaxel

Examples of clinical variables:

- Patient demographics, lesion measures
- Medical history, prior surgery and radiation,
- Vital sign, lab data (hematology and urinary lab)
- Prior medicine, adverse events, etc.

This Challenge did not include any genomic/molecular data, in contrast to most other DREAM challenges.
Subchallenge 1: Predict Overall Survival

mCRPC Phase III Clinical Trials

Memorial Sloan Kettering Cancer Center
476 patients

Celgene
526 patients

Sanofi
598 patients

AstraZeneca
470 patients

Test Leaderboard Scoring Rounds

Training Data
1600 patients

Team Predictions

Leaderboard Scoring Set
157 patients

Ranked Test Leaderboard Performance
(3 Rounds)

Final Scoring Set
313 patients

Final Team Ranked Performance

5 submissions per round, best score (integrated AUC) will be ranked on the leaderboard after each round. The leaderboard set will change on with 80% (of 157 patients) subsampled for scoring each round.

<table>
<thead>
<tr>
<th>Open Phase</th>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
<th>Final Round</th>
<th>Scoring/Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 weeks</td>
<td>8 weeks</td>
<td>3 weeks</td>
<td>3 weeks</td>
<td>3 weeks</td>
<td>3 weeks</td>
</tr>
</tbody>
</table>

March 16, 2015
Challenges open
Release Training Data
Release Leaderboard Scoring Set

April 1, 2015
Webinar
Leaderboard opens

April 8, 2015
Start of submission scoring with live email feedback

May 27, 2015

June 15, 2015
Release Final Scoring Set

July 6, 2015
Challenge closed

July 27, 2015
Final Results Announced

August 17, 2015
In the Halabi model, 8 factors are used to predict overall survival in mCRPC:

- EGOC performance
- Disease site
- Opioid analgesic use
- LDH
- Albumin
- Hemoglobin
- Alkaline phosphatase
- PSA
FIMM-UTU team

Team FIMM-UTU from Finland

Top row, left to right: Teemu Daniel Laajala$^{1,2}$, Tapio Pahikkala$^3$, Peddinti Gopalacharyulu$^2$, Suleiman Khan$^2$

Bottom row, left to right: Antti Airola$^3$, Tuomas Mirtti$^{2,4}$, Tero Aittokallio$^{1,2}$
Our approach

• Build more **multivariate machine learning models** in order to find clinically relevant variables that predict overall survival (shortened or extended) in docetaxel-treated mCRPC

• Extend the modeling framework by using **ensemble approach** and **additional variables and their interactions** in comparison with LASSO-based univariate predictors (Halabi model)

• Make use of wide array of both **unsupervised and supervised machine learning approaches** for predictive modelling and understanding the factors behind treatment response
(a) Manual extraction of additional variables; model-based imputation of missing data values
(b) Three ensemble components for CELGENE, EFC6546 and their combination. ASCENT2 study from Memorial Sloan Kettering Cancer Center was excluded from the final model.
(c) Ensemble prediction for the AZ study by averaging the ranks over the three components.
(d) Ensemble of penalized cox regression (ePCR) model provides accurate survival predictions.
Missing value imputation:
1) Imputation of missing at random (MAR)
2) Structured patterns (penalized regression)
Unsupervised exploration with PCA

All variables after imputation

Binary clinical variables (yes/no)
Unsupervised learning

Distributional differences
Introduce non-linear interactions

Unsupervised learning
PC2

Redundancy

<table>
<thead>
<tr>
<th></th>
<th>BONE lesion</th>
<th>ASC</th>
<th>CEL</th>
<th>EFC</th>
<th>AZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>131</td>
<td>87</td>
<td>69</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>345</td>
<td>439</td>
<td>529</td>
<td>470</td>
<td></td>
</tr>
</tbody>
</table>

Clinical relevance

<table>
<thead>
<tr>
<th>Variable</th>
<th>Yes</th>
<th>Maybe</th>
<th>No</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>X TURP</td>
<td></td>
<td>✓</td>
<td></td>
<td>Not relevant if for hyperplasia</td>
</tr>
<tr>
<td>X GASTREFL</td>
<td></td>
<td>✓</td>
<td></td>
<td>Not relevant</td>
</tr>
<tr>
<td>X MHEYE</td>
<td></td>
<td>✓</td>
<td></td>
<td>Not relevant</td>
</tr>
<tr>
<td>X SPINAL_CORD_SURGERY</td>
<td></td>
<td>✓</td>
<td></td>
<td>If due to metastases (also distrib.)</td>
</tr>
<tr>
<td>✓ ORCHIDECTOMY</td>
<td>✓</td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>✓ BONE</td>
<td>✓</td>
<td></td>
<td></td>
<td>Distribution vs. clinical relevance</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outcome variable \((y)\) and model predictors \((x)\)

- Initial data matrix \((X)\) included 124 variables. After removing clinically irrelevant ones, or highly-skewed variables, 101 predictor variables were used in our prediction model.
- Non-linearity was introduced by modeling a selected set of pairwise interactions, e.g. Alkaline Phosphatase (ALP) \(\times\) Alanine Transaminase (ALT), resulting in 3422 new features.
- All 1600 training dataset patients were initially used for modeling the survival (DEATH) of the 470 AZ patients. However, 476 ASCENT2 patients were excluded in the final predictions.

**Table: y**

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>DEATH</th>
<th>LKADT_P</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASC-002-0008</td>
<td>0</td>
<td>553</td>
</tr>
<tr>
<td>ASC-002-0011</td>
<td>1</td>
<td>467</td>
</tr>
<tr>
<td>ASC-002-0014</td>
<td>0</td>
<td>399</td>
</tr>
<tr>
<td>CELG-00024</td>
<td>1</td>
<td>129</td>
</tr>
<tr>
<td>CELG-00025</td>
<td>0</td>
<td>81</td>
</tr>
<tr>
<td>CELG-00026</td>
<td>0</td>
<td>231</td>
</tr>
<tr>
<td>VEN-336000702</td>
<td>1</td>
<td>736</td>
</tr>
<tr>
<td>VEN-336000801</td>
<td>1</td>
<td>451</td>
</tr>
<tr>
<td>VEN-336100401</td>
<td>1</td>
<td>409</td>
</tr>
<tr>
<td>AZ-00070</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AZ-00071</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AZ-00072</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

\(LKADT = \text{Last known day alive}\)

**Table: x**

<table>
<thead>
<tr>
<th>ID</th>
<th>ALP</th>
<th>ALT</th>
<th>ANALGESICS</th>
<th>ALPxALT</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>6.18</td>
<td>2.40</td>
<td>0</td>
<td>14.83</td>
</tr>
<tr>
<td>...</td>
<td>5.08</td>
<td>2.71</td>
<td>0</td>
<td>13.76</td>
</tr>
<tr>
<td>...</td>
<td>4.45</td>
<td>2.30</td>
<td>0</td>
<td>10.26</td>
</tr>
<tr>
<td>...</td>
<td>6.38</td>
<td>2.64</td>
<td>0</td>
<td>16.85</td>
</tr>
<tr>
<td>...</td>
<td>4.60</td>
<td>4.36</td>
<td>1</td>
<td>20.02</td>
</tr>
<tr>
<td>...</td>
<td>5.48</td>
<td>3.18</td>
<td>1</td>
<td>17.42</td>
</tr>
<tr>
<td>...</td>
<td>6.07</td>
<td>3.04</td>
<td>1</td>
<td>18.48</td>
</tr>
<tr>
<td>...</td>
<td>5.08</td>
<td>3.43</td>
<td>1</td>
<td>17.43</td>
</tr>
<tr>
<td>...</td>
<td>4.43</td>
<td>2.40</td>
<td>1</td>
<td>10.62</td>
</tr>
<tr>
<td>...</td>
<td>5.78</td>
<td>1.61</td>
<td>1</td>
<td>9.31</td>
</tr>
<tr>
<td>...</td>
<td>5.24</td>
<td>2.64</td>
<td>0</td>
<td>13.83</td>
</tr>
<tr>
<td>...</td>
<td>4.64</td>
<td>2.30</td>
<td>0</td>
<td>10.69</td>
</tr>
</tbody>
</table>

\(ANALGESICS = \text{Prior analgesics medical use}\)
(Elastic Net type) Penalized Cox regression

\[
\min_{\beta_0, \beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[ (1 - \alpha) \|\beta\|_2^2/2 + \alpha \|\beta\|_1 \right]
\]  \hspace{1cm} (1)

where the model coefficients \( \beta \) are effectively shrunk towards zero (embedded feature selection) by penalizing the model likelihood with non-zero coefficients with \( \lambda \). Furthermore, \( \alpha \) controls the proportion of L1/L2 penalization. The \textit{coxnet}-subfamily in the \textit{glmnet} R-package allowed us to estimate relative risks for individuals. Suitable sequence of \( \lambda \) is determined as a function of \( \alpha \). 10-fold cross-validation averaged over 3 random binnings were used to determine optimal \( \{\alpha, \lambda\} \). Cox proportional hazards regression allows comparisons of hazard ratios:

\[
\ln \left( \frac{H(t)}{H_0(t)} \right) = \beta X
\]  \hspace{1cm} (2)
Supervised learning of model parameters

Component-wise optimums

<table>
<thead>
<tr>
<th></th>
<th>ASC/EFC/CEL</th>
<th>ASC/CEL</th>
<th>ASC/EFC</th>
<th>EFC/CEL</th>
<th>ASC</th>
<th>CEL</th>
<th>EFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>0.00060</td>
<td>0.78458</td>
<td>0.00219</td>
<td>0.00031</td>
<td>0.01267</td>
<td>0.01081</td>
<td>0.00298</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>58.205</td>
<td>0.046</td>
<td>25.047</td>
<td>61.334</td>
<td>4.010</td>
<td>4.653</td>
<td>16.809</td>
</tr>
</tbody>
</table>
Test set results in the AstraZeneca study

Over 550 participants registered for the Challenge comprising over 50 teams. FIMM-UTU was clear winner at the end of the three leaderboard periods (Final).

Teams must perform better than Halabi model
Top teams must be within a Bayes Factor of <3
Post-challenge evaluation: independent comparison against the Halabi reference

Conclusions and future directions

• Modelling interactions was critical. Halabi misses all interactions.
• Incorporating clinical expertise through priori knowledge of variables proved useful. Multidisciplinary expertise important.
• Standardization of data elements and variables is important when comparing data from several clinical studies. Data harmonization.
• Making the model more practical for future clinical trials. Further testing ongoing with grant from National Cancer Institute (NCI).
• Auria /TYKS and HUB/HYKS biobanks collect many same clinical variables, which could help designing more effective clinical studies.
• Adding genetic variation by combining clinical treatment data (health care records) with population-level variant data (SISu).
Acknowledgements