EXCEPTIONAL SURVIVAL MODEL MINING

Juliana B. Mattos[1], Eraylson G. Silva[1],
Paulo S. G. de Mattos Neto[1], and Renato Vimieiro[2]

[1] Centro de Informática, Universidade Federal de Pernambuco, Recife-PE, Brasil
{jbm4,egs,psgmn}@cin.ufpe.br

[2] Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil
rvimieiro@dcc.ufmg.br

Context & Problem

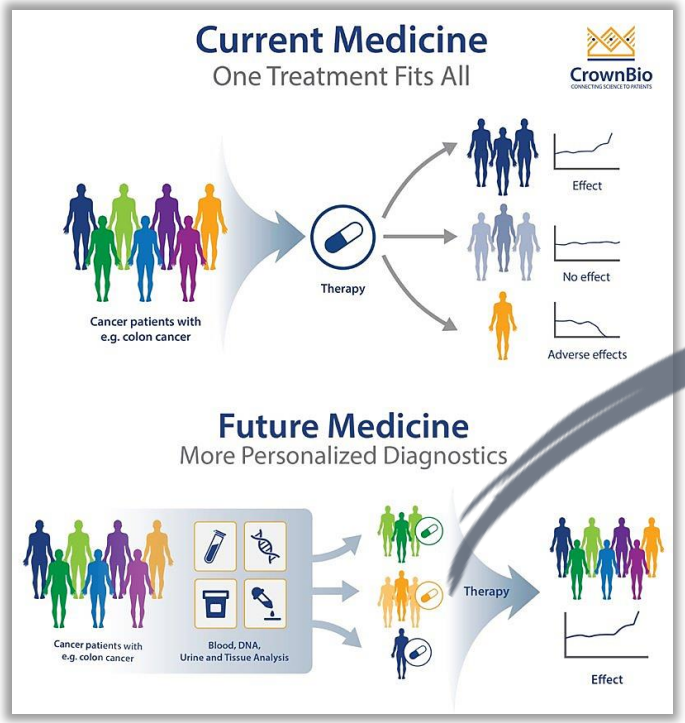


Image source: https://blog.crownbio.com/pdx-personalized-medicine# (access: September 11th, 2020)

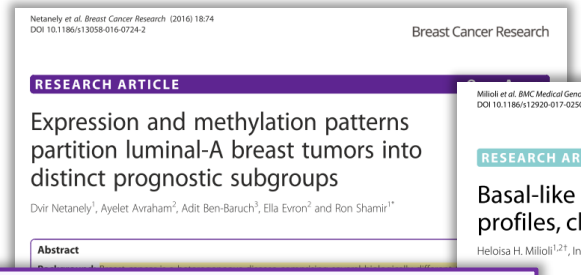
WHO ARE SUCH SUBGROUPS?

- Large-scale biologic databases
- Methods for characterizing patients
- Strong computational tools

WHICH CHARACTERISTICS DELINEATE GROUPS OF PATIENTS WITH **DISTINCT SURVIVAL EXPERIENCE?**

WHICH FACTORS ARE ASSOCIATED WITH **DIFFERENT PROGNOSTICS**?

The Problem in the Literature


Netanely et al. Breast Cancer Research (2016) 18:74
DOI 10.1186/s13058-016-0724-2

Breast Cancer Research

RESEARCH ARTICLE

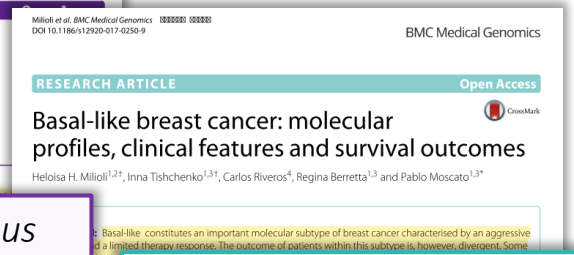
Expression and methylation patterns partition luminal-A breast tumors into distinct prognostic subgroups

Dvir Netanely[1], Ayelet Avraham[2], Adit Ben-Baruch[3], Ella Evron[2] and Ron Shamir[1*]

Abstract

Breast cancer is a heterogeneous disease comprising several biologically different types, […] ***precise identification of breast cancer subtypes, especially within the largest and highly variable luminal-A class, remains a challenge.***

Netanely, D., Avraham, A., Ben-Baruch, A., Evron, E., & Shamir, R. (2016). Expression and methylation patterns partition luminal-a breast tumors into distinct prognostic subgroups. Breast Cancer Research, 18(1):74.


Milioli et al. BMC Medical Genomics
DOI 10.1186/s12920-017-0250-9

BMC Medical Genomics

RESEARCH ARTICLE Open Access

Basal-like breast cancer: molecular profiles, clinical features and survival outcomes

Heloisa H. Milioli[1,2†], Inna Tishchenko[1,3†], Carlos Riveros[4], Regina Berretta[1,3] and Pablo Moscato[1,3*]

Basal-like constitutes an important molecular subtype of breast cancer characterised by an aggressive and a limited therapy response. The outcome of patients within this subtype is, however, divergent. Some

Basal-like constitutes an important molecular subtype of breast cancer characterised by an aggressive behaviour and a limited therapy response. ***The outcome of patients within this subtype is, however, divergent. Some individuals show an increased risk of dying in the first five years, and others a long-term survival of over ten years after the diagnosis.***

Milioli, H. H., Tishchenko, I., Riveros, C., Berretta, R., & Moscato, P. (2017). Basal-like breast cancer: molecular profiles, clinical features and survival outcomes. BMC medical genomics, 10(1):19.

Shivakumar, M., Lee, Y., Bang, L., Garg, T., Sohn, K.A. and Kim, D., 2017. Identification of epigenetic interactions between miRNA and DNA methylation associated with gene expression as potential prognostic markers in bladder cancer. *BMC medical genomics*, *10*(1), p.30.

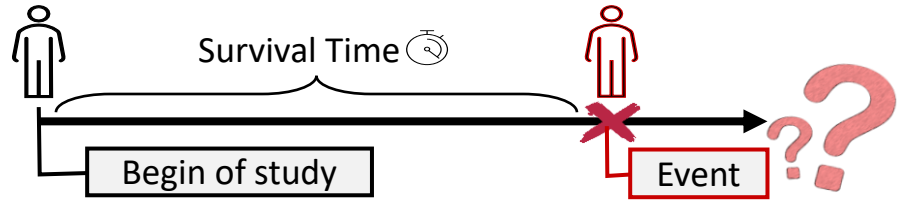
Pepke, S. and Ver Steeg, G., 2017. Comprehensive discovery of subsample gene expression components by information explanation: therapeutic implications in cancer. *BMC medical genomics*, *10*(1), pp.1-18.

Smith, M.K., Stein, G., Cheng, W., Miller, W.C. and Tucker, J.D., 2019. Identifying high risk subgroups of MSM: a latent class analysis using two samples. *BMC infectious diseases*, *19*(1), p.213.

Lu, T.P. and Chen, J.J., 2015. Subgroup identification for treatment selection in biomarker adaptive design. *BMC medical research methodology*, *15*(1), p.105

Survival Analysis

Collection of methods and techniques designed to analyse data in which the target variable is the **time until a given event** occurs.



GOAL

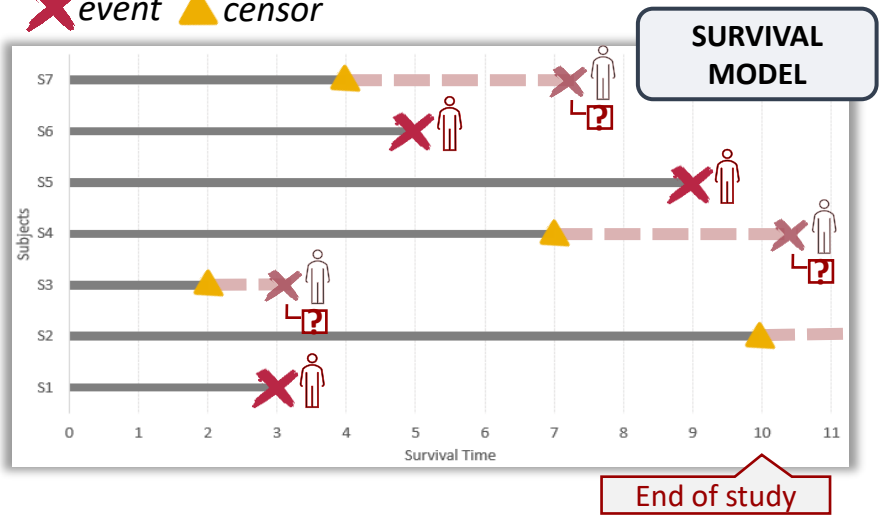
To **estimate** the
time to the event
for a new instance,
based on feature predictors.

SURVIVAL DATA:

- Set of descriptive attributes
- Event Information:
  T: *survival/censor times*
  δ: event status

Observed times:

event *censor*



Survival Analysis Methods

STATISTICAL METHODS

- Non-Parametric
- Semi-Parametric
- Parametric

→ distributional and restrictive assumptions

MACHINE LEARNING METHODS

- Survival trees
- Bayesian methods
- Neural Network
- Support Vector Machine
- Ensemble
- Active/Transfer/Multi-task learning

→ ~~distributional and restrictive assumptions~~

→ modelling non-linear relationships

→ high quality results

RULE-BASED METHODS

- Rough sets
- Bump hunting
- Logical Analysis of Data (LAD)
- Survival tree
- Sequential covering
- ….

→ Simple and understandable results

Survival Analysis Methods

STATISTICAL METHODS

MACHINE LEARNING METHODS

RULE-BASED METHODS

PREDICTIVE APPROACHES

GLOBAL MODELS

Prediction of T
Output: scores or probabilities

GOAL

Classification of new instances
Output: partitions of the data

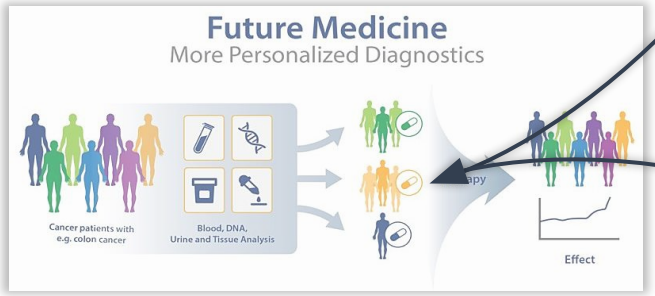
CHARACTERIZATION

Discretization or **stratification** of the time variable

Covariates' **split criterion** and **stratifications**

Impel observations to fit **pre-defined classes**

Research Opportunity



WHO ARE SUCH SUBGROUPS?

Characteristics for distinct survival experience?

Factors associated with different prognostics?

GLOBAL MODELS X LOCAL PATTERNS

EXISTENT APPROACHES

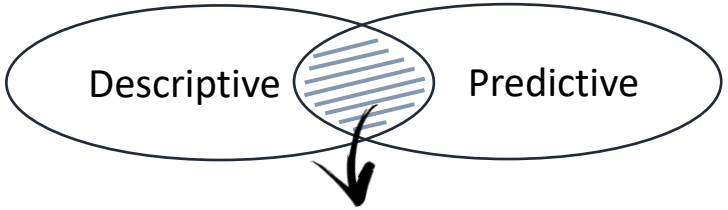
- previously known variable's interactions
- lack the ability to shed light into new interactions

GOAL

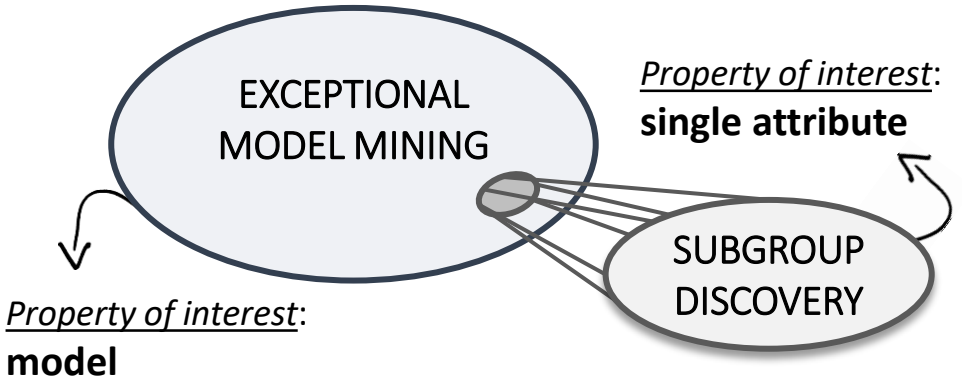
Discover and **describe** multiple (and potentially overlapping) interesting **subgroups** with relation to the survival response

Exceptional Model Mining (EMM)

DATA MINING PERPECTIVES



SUPERVISED DESCRIPTIVE PATTERN MINING
Understand the underlying **phenomena**
– according to a property of interest (target).



PROBLEM STATEMENT

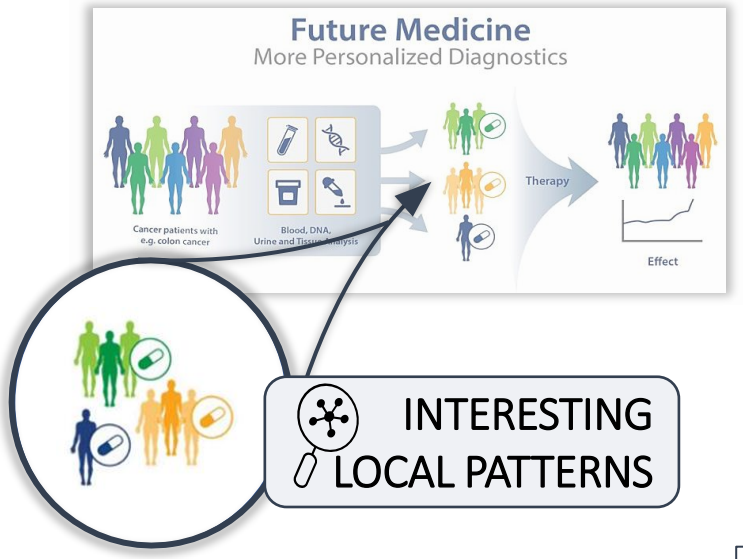
The task of EMM is:

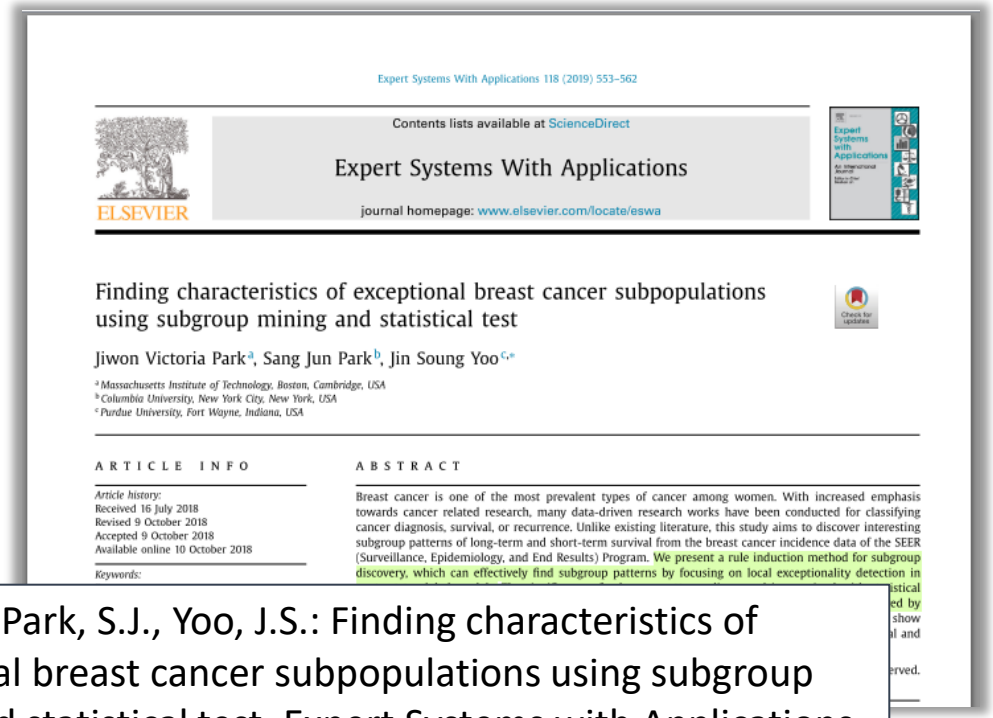
*To discover the subgroups of the population that are statistically "**most interesting**", i.e. are as large as possible and have the most unusual statistical characteristics with respect to the model of interest.*

Heuristic Search Strategies
→ Beam search
→ Evolutionary Computing

No EMM works exploring bio-inspired meta-heuristic

EMM & Survival Analysis





Park, J.V., Park, S.J., Yoo, J.S.: Finding characteristics of exceptional breast cancer subpopulations using subgroup mining and statistical test. Expert Systems with Applications 118, 553–562 (2019)

No works striving to uncover subgroups with unusual survival models

SUBGROUP DISCOVERY FRAMEWORK

Tree-based rule induction approach

***Target**: mean survival time*

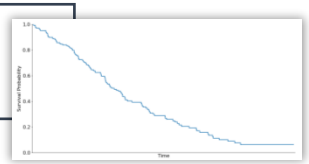
Exceptional Survival Model Ant Miner – ESM-AM

EMM FRAMEWORK:

Search for subgroups with exceptional **survival functions**

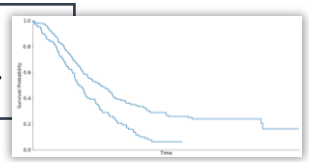
Model:

KAPLAN MEIER (KM) ESTIMATES



Interestingness measure:

LOGRANK STATISTICAL TEST



Search strategy:

ANT-COLONY OPTIMIZATION

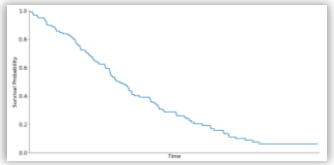
***Output**: SET OF RULES*

IF <conditions> THEN <KM model>

Antecedent: Consequent

IF $\boldsymbol{term1}$ AND $\boldsymbol{term2}$ AND …

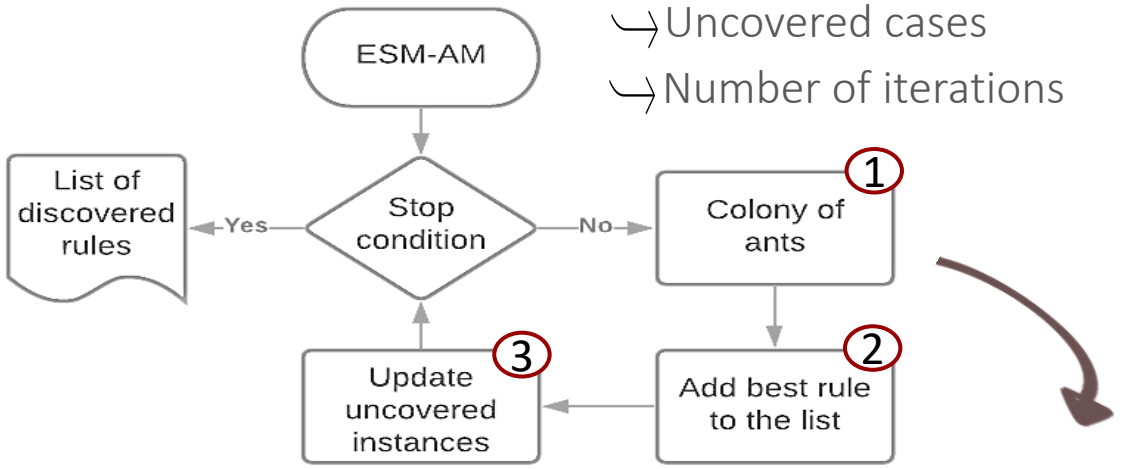
$term$: < attribute = value >



EXCEPTIONAL SUBGROUP CHARACTERISATION → **SUBGROUP'S SURVIVAL MODEL**

Exceptional Survival Model Ant Miner – ESM-AM

Stop condition:
↳ Uncovered cases
↳ Number of iterations



Adaptation of the **Ant-Miner** algorithm
to discover subgroups with exceptional survival functions

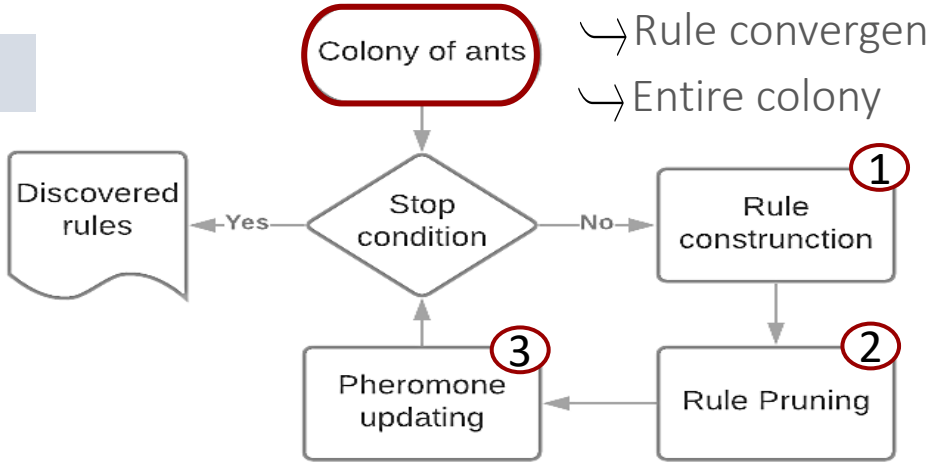
Parpinelli, R.S., Lopes, H.S., Freitas, A.A.: Data mining with an ant colony optimization algorithm. IEEE Transactions on Evolutionary Computation 6(4), 321–332 (2002)

COVERING-BASED APPROACH

User-defined parameters:

(1) no_of_ants
(2) max_uncovered_cases
(3) no_rules_converg
(4) min_cases_per_rule
(5) α

Stop condition:
↳ Rule convergence
↳ Entire colony



Exceptionality:
Subgroup *versus* **Complement**

Non-significant rules are discarded at a level of significance of α

If no significant rules are discovered, the algorithm is finalized

ESM-AM Results

14 real-world survival data sets

- Removal of observations containing missing values
- Feature selection
- Discretization with K-Means into five interval categories

Rule-models' evaluation metrics

- Number of rules
- Rule length
- Rule coverage
- Ruleset coverage
- Integrated Brier Score (IBS)

#rules & rule length

Compact models

Rule coverage

Neither cover most cases nor very small groups

IBS

Homogeneous subgroups

Ruleset coverage

Variability

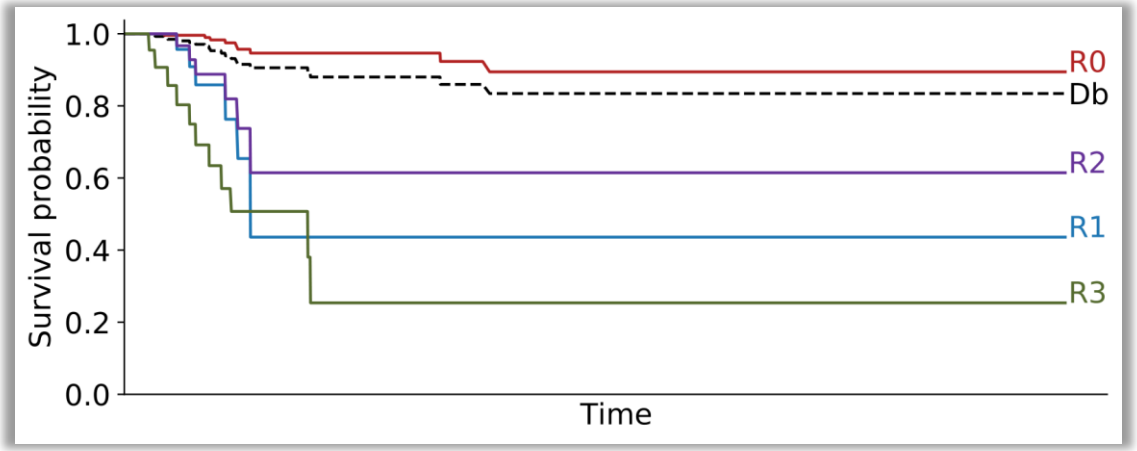
Discovery of exceptional survival behaviour

Discovery of significant subgroups and identification of data characteristics that interfere in survival experience

ESM-AM Results

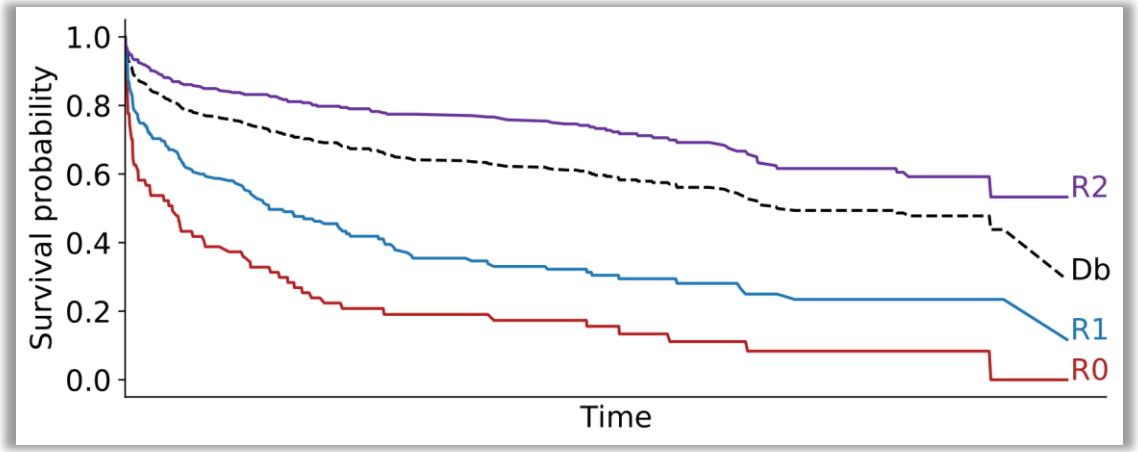
PTC: Papillary thyroid carcinoma

Event: recurrence/progression



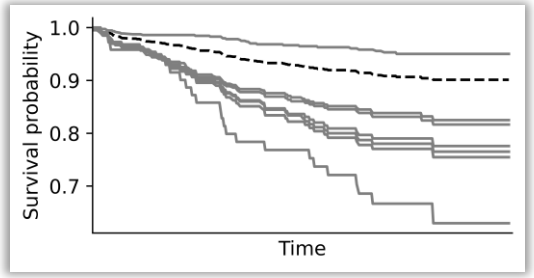
Whas500: Worcester Heart Attack

Event: death



Local patterns with significant distinct survival response

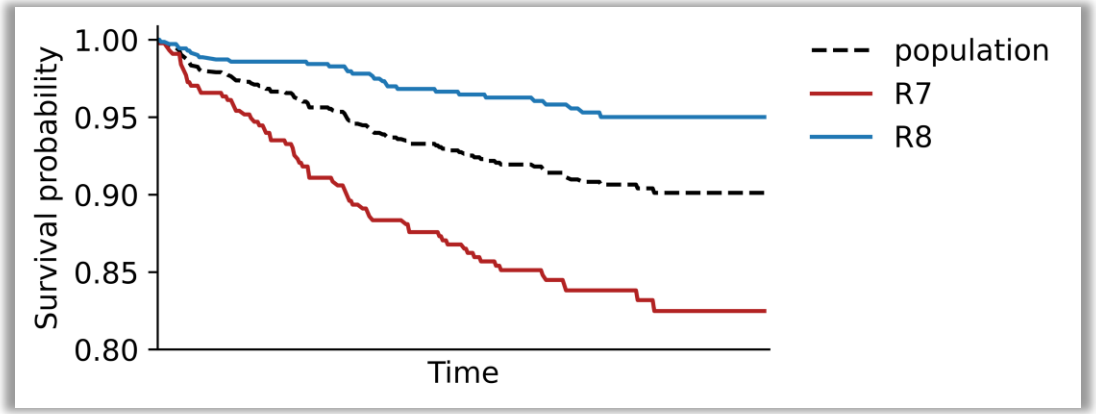
ESM-AM Results



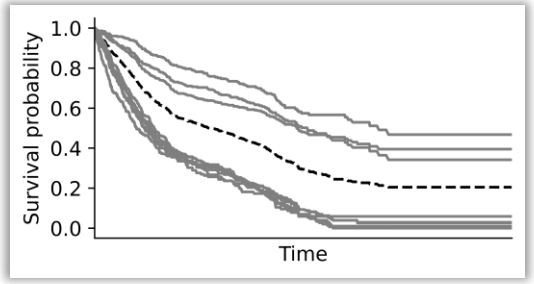
ACTG320 DATA SET
HIV infected patients

Discovered sugbroups for *actg320* dataset
R7: IF **strat2 = 0** THEN average survival = 226.94
R8: IF **strat2 = 1** THEN average survival = 232.19

Strat2 = {0,1} (low/high)
counting of cells with **CD4 protein expression**



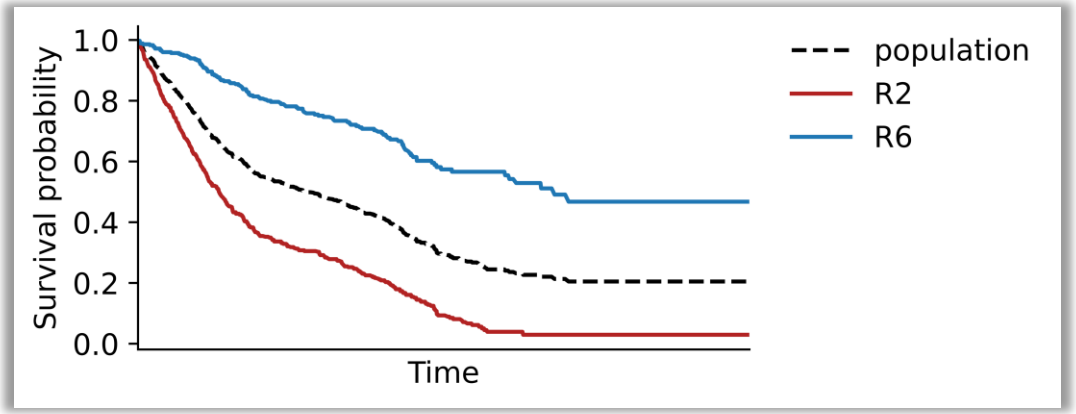
ESM-AM Results



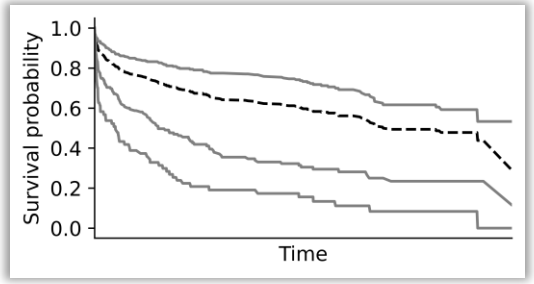
LUNG DATA SET
Early lung cancer

Discovered sugbroups for *lung* dataset
R2: IF **stage-1 = 3** THEN average survival = 835.80
R6: IF **stage-1 = 1** THEN average survival = 1523.17

stage1 = {1,2,3} **(1 < 3)**
general stage of lung cancer



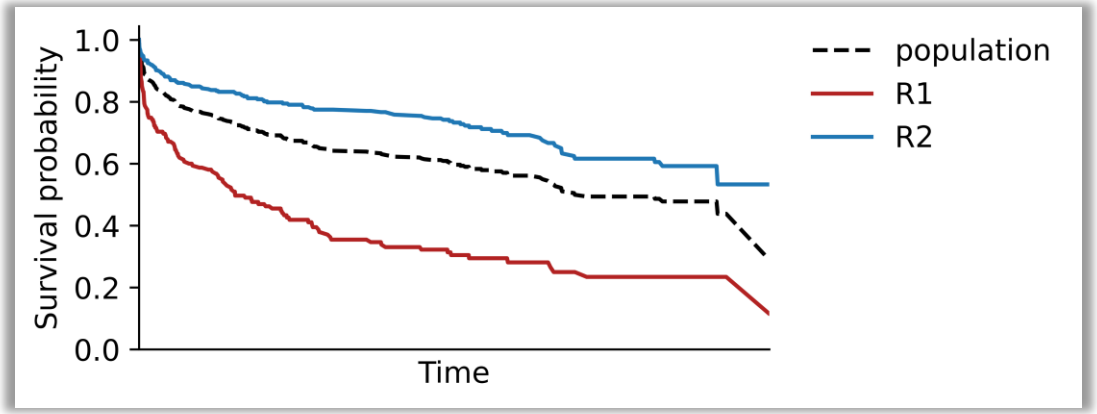
ESM-AM Results



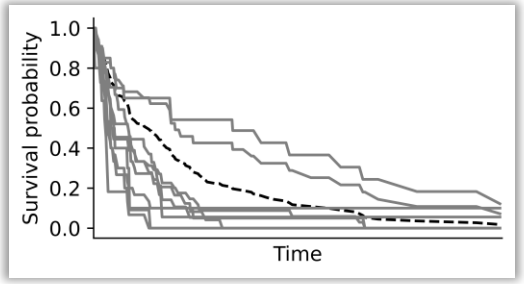
WHAS500 DATA SET
Worcester Heart Attack

Discovered sugbroups for *whas500* dataset
R1: IF **chf = True** THEN average survival = 593.59
R2: IF **chf = False** THEN average survival = 1012.21

Chf = {True, False}
congestive heart complications



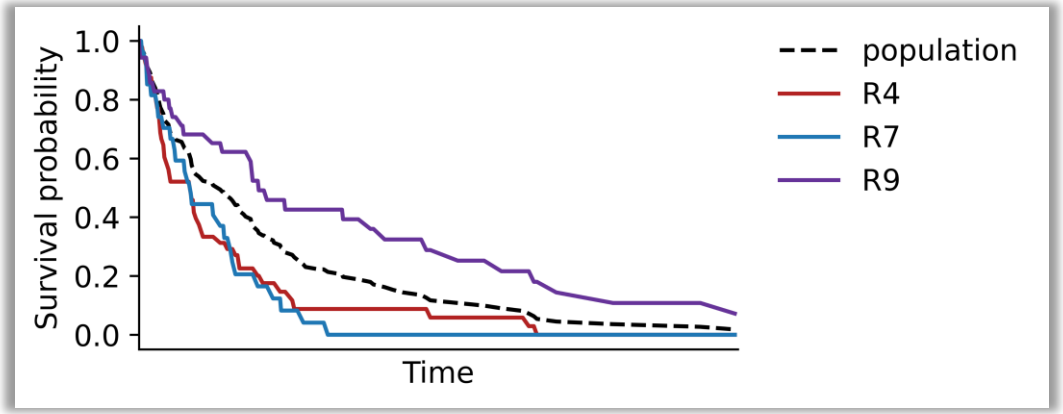
ESM-AM Results



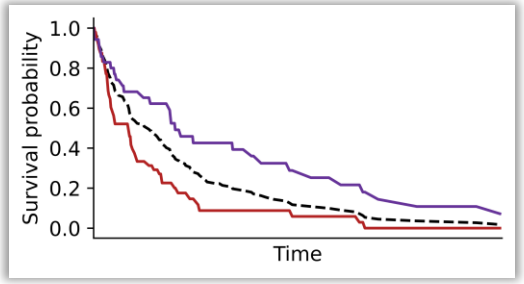
VETERAN DATA SET
Lung cancer

Induced rules for *veteran* dataset
R4: IF **cell-type** = **small** THEN average survival = 71.67
R7: IF **cell-type** = **adeno** THEN average survival = 64.11
R9: IF **cell-type** = **aquamous** THEN average survival = 200.20

Cell-type:
small, adeno, aquamous



ESM-AM Results



VETERAN DATA SET
Lung cancer

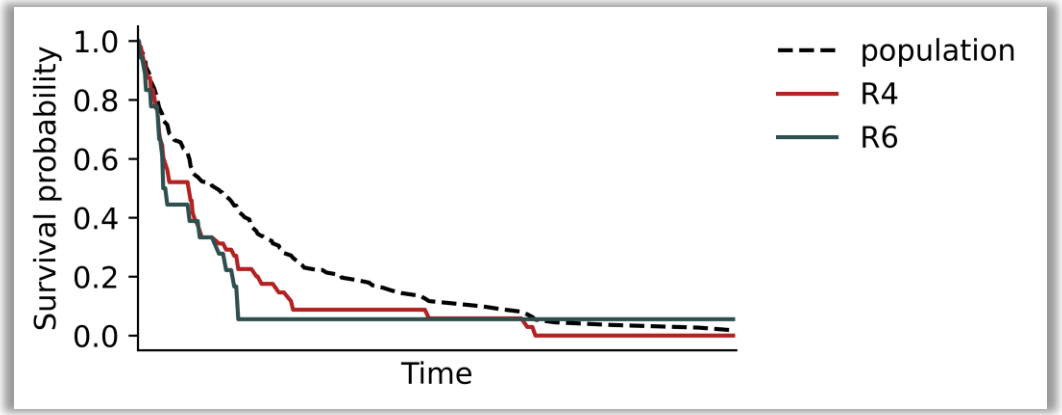
Induced rules for *veteran* dataset

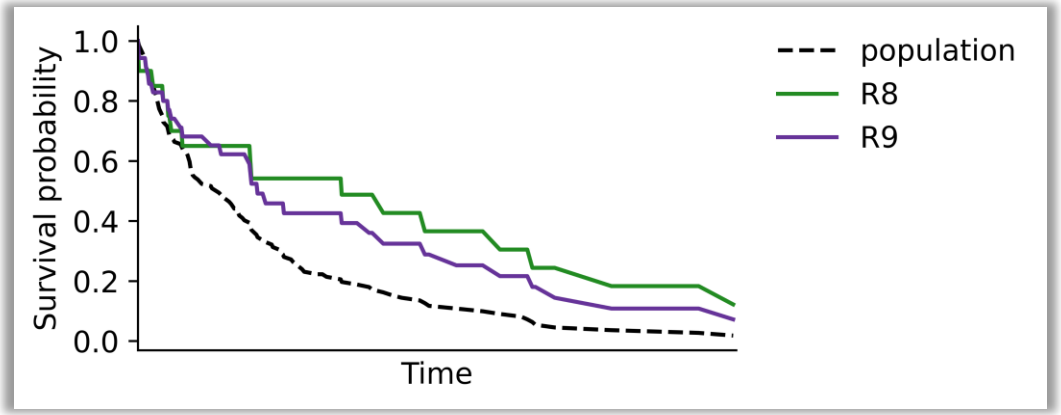
R4: IF cell-type = **small** THEN average survival = **71.67**

R6: IF cell-type = **small** AND **treat = test** THEN average survival = **47.17**

R8: IF cell-type = **aquamous** AND **treat = test** THEN average survival = **260.30**

R9: IF cell-type = **aquamous** THEN average survival = **200.20**





Future work

1. Cope with **numerical attributes**
2. Other quality measures: consider **exceptionality** and **coverage**
3. Tackle problems: **pattern's redundancy**, **high-dimensionality** and **false statistical discoveries**
4. Investigate **new heuristic functions** and **new pheromone updating procedure**
5. Expand the results' analysis: further experimental **statistical procedures** and more detailed **exploratory data analysis**