

eTRIKS : Tutorial for clinical data selection

Follow up details:

Name	Organization	Date	Details
N. Jullian	CNRS	25/11/2013	version 1.0
N. Jullian	CNRS	05/11/2013	version 2.0

Introduction

This tutorial focuses on clinical data analysis by using the eTRIKS training server. The objective of the tutorial is to familiarize you with some basic tools available in eTRIKS. The tutorial is based on tranSMART v1.2 (released september 2014) accessible via :

public.transmart.etriks.org

The tutorial is divided into 3 lessons :

- Part 1 : Browsing the data tree Page 4
- Part 2 : Comparison of 2 subsets Page 7
- Part 3 : Survival Data analysis Page 10

The dataset used for this example is from **Desmedt et al** published in 2007 in Clinical Cancer Research (Clin Cancer Res 2007, Vol 13, n°11, June 1, 2007). The study reports gene expression profiling of frozen samples from 198 untreated breast cancer patients. The endpoints are defined as:

- Time to Distant Metastasis (TDM)
- Overall Survival (OS)

Launch eTRIKS by typing the following URL http://public.transmart.etriks.org in your favourite Internet browser. Go to the Dataset Explorer Tab and open the tree under **Public Studies**. Public datasets that are uploaded into eTRIKS are labelled with pathology name, author name and GSE code. Expand the tree under the **GEO Studies** and then the **Breast_Cancer** node. The Desmedt study refers to GSE7390. For more details on the study, right click on the study name and select « **Show Definition** » as shown in Figure 1. A popup window shows the summary of the study as deposited in the NCBI repository.

tranSMART		Browea	Analy	70 Cono Signat	uro /Liete	CWAS UP	load Data	Admin	Utilitie	
vi.2 All	Comparison	Summany Statistics	Crid View	Advanced Workflow	Data Export	Evnort John	Workspace	Cenome	Browser	uə
Active Filters and R		Summary Statistics		Advanced Norkhow	Dutu Export	Export jobs	×	Save Sub	oset 🗵 Cle	ear
		Subset 1	(Exclude X		Subse	<u>t 2</u>	Exclude	x	
Navigate Terms										
9 Across Trials		AND)	Exclude		А	ND	Exclude	x	
Gancer Cell Line Encyclopedia (1070)										
Asthma		AND		Exclude		A	ND	Exclude	x	
Breast Cancer										
Biomarker Data (198) B □ Clinical Data (198)										
🗄 🦳 Subjects (198)										
😐 🗐 Sorlie(2003) GSE4382 (167)										
Inflammatory Bowel Disease										
Multiple Myeloma										
H C Multiple Sclerosis										
🗄 🦳 Rheumatoid Arthritis										
B GSK Cell Lines (950)										
TCGA Studies									De	uhea

Figure 1: Dataset Explorer view

A popup window shows the summary of the study as deposited in the NCBI repository as shown in Figure 2.

Show Concept Definition-Desmedt(2007)_GSE7390 (198)				
TITLE:	Strong Time Dependence of the 76-Gene Prognostic Signature			
SUMMARY:	Background: Recently a 76-gene prognostic signature able to predict distant metastases in lymph node-negative (N-) breast cancer patients was reported. The aims of this study conducted by TRANSBIG were to independently validate these results and to compare the outcome with clinical risk assessment. Materials and Methods: Gene expression profiling of frozen samples from 198 N- systemically untreated patients was performed at the Bordet Institute, blinded to clinical data and independent of Veridex. Genomic risk was defined by Veridex, blinded to clinical data. Survival analyses, done by an independent statistician, were performed with the genomic risk and adjusted for the clinical risk, defined by Adjuvant!Online. Results: The actual 5- and 10-year time to distant metastasis (TDM) were 98% (88%-100%) and 94% (83%-98%) respectively for the good profile group. The actual 5- and 10-year overall survival (OS) were 98% (88%-100%) and 87% (73%-94%) respectively for the good profile group and 76% (68%- 82%) and 73% (65%-79%) for the poor profile group. The actual 5- and 10-year overall survival (OS) were 98% (88%-100%) and 87% (73%-94%) respectively for the good profile group and 84% (77%-89%) and 72% (63%-78%) for the poor profile group and 84% (77%-89%) and 72% (63%-78%) for the poor profile group. We observed a strong time-dependency of this signature, leading to an adjusted HR of 13.58 (1.85-99.63) and 8.20 (1.10-60.90) at 5 years, and 5.11 (1.57-16.67) and 2.55 (1.07-6.10) at 10 years for TDM and OS respectively. Conclusion: This independent validation confirmed the performance of the 76-gene signature and adds to the growing evidence that gene expression signatures are of clinical relevance, especially for identifying patients at high risk of early distant metastases.			
CONTRIBUTOR:	Christine,,Desmedt; Fanny Piette; Sherene Loi; Yixin Wang; Francoise Lallemand; Benjamin Halbe-Kains; Giuseppe Viale; Mauro Delorenzi; Yhi Zhang; Mahasti Saghatchian; Jonas Bergh; Rosette Lidereau; Paul Ellis; Adrian Harris; Jan,G,Klijn; John,A,Foekens; F			
CONTRIBUTOR:	Christine,,Desmedt; Fanny Plette; Sherene Lol; Yixin Wang; Francoise Lallemand; Benjamin Haibe-Kains; Giuseppe Viale; Mauro Delorenzi; Yhi Zhang; Mahasti Saghatchian; Jonas Bergh; Rosette Lidereau; Paul Ellis; Adrian Harris; Jan,G,Klijn; John,A,Foekens; F			
CONTACT:	Name: Benjamin Haibe-Kains; Email: bhaibeka@ulb.ac.be; Phone: +3225413428; Laboratory: Functional Genomics Unit; Institute: Institut Jules Bordet; Address: Bld de Waterloo 127; City: Bruxelles; Zip/postal_code: 1000; Country: Belgium; Web_link: http://www.ulb.ac.be/di/map/bhaibeka/index.html			
TYPE:	Expression profiling by array			
OVERALL_DESIGN:	dataset of microarray experiments from primary breast tumors used to validate the 76-gene signature (VERIDEX). No replicate, no reference sample.			
STATUS:	Public on Jun 11 2007			
SUBMISSION_DATE:	2007-03-28			
LAST_UPDATE_DATE:	2013-05-31			
PUBMED_ID:	17545524			
ORGANISM(S):	Homo sapiens			
PLATFORM:	GPL96			
		Close		

Figure 2: Details of the study

Lesson 1 : Browsing the data tree

Expand the Desmedt(2007)_GSE7390 node by clicking on the + sign on the left. The right number in parenthesis represents the number of patients in the study that is 198 here. Now you have 3 nodes listed:

- Biomarker Data (198)
- Clinical Data (198)
- Subjects (198)

Each node is related to a type of data and we are going to browse through each node to see what kind of data is available for this study (Figure 3).

A label is assigned to each data node:

- « 123 » for continuous numerical variables
- « abc » for categorical text variables
- « ^{IOI} » for high dimensional data.

First, expand the **Biomarker Data** node to access the details of the micro-array based profiling: HG-U133 Affymetrix Human Genome U133A Array. Again the number in parenthesis indicates that the experiment has been performed for 198 patients.

Then, expand the **Clinical Data** node. There are 4 groups of data collected in this study. The entire tree related to Clinical Data is displayed in the screenshot below (Figure 3). The Disease Free Survival, Distant Metastasis Free Survival and Overall Survival are related to the endpoint (clinical outcome) of the study. The Risk Assessment tree is made of 4 components that have been evaluated in the study with the aim to compare their predictive power with the signature outcome.



Figure 3: View of the Clinical Data tree

Lastly, expand the **Subjects** node. The data is divided into 2 nodes: Demographics and Medical History (Figure 4). The Age variable is the only Demographics variable available. A number of measures are listed under the Medical History, among which is the Estrogen Receptor status that we will further explore in Lesson 2.



Figure 4: Details of the Subject tree

Lesson 2: Analysis Subset Selection

Expand the Desmedt(2007)_GSE7390 node by clicking on the + sign on the left. We are going to select two subsets of subjects and assess whether there exist a dependencies between the clinical outcome and a set of selected variables, namely **Age** and **Estrogen Receptor Status**.

Question: How many subjects are contained in the high risk group (subjects that developed distant metastasis within 10 years, as defined in the paper) and how many subjects are contained in the low risk group (subjects which remained metastasis free during the first 10 years).

As a first step, we define 2 subsets of subjects by selecting the outcome which is the TDM (time to distant metastasis) variable, expressed as the number of days to disease progression (distant metastasis). The « **123** » label at the left of the node indicates this is a continuous numerical variable. Drag this variable into the top box of Subset 1. A popup window is displayed and you will now define the intervals that you want to explore further. You might click on the Show Histogram box for further details on the repartition among subjects (Figure 5).



Figure 5: Histogram of TDM values

We are going to start with a 10 years interval as discussed in the paper. Click on « By numeric value », select « LESS THAN » as an operator and enter 3650 as the threshold value for 10 years (the unit is days here).

Repeat similar actions for Subset 2 and define the complementary second group by entering numeric values ≥3650 (GREATER THAN OR EQUAL TO) as shown on Figure 6.



Figure 6: window for defining intervals for subset definition

The subset 1 represents the high risk group with subjects that developed distant metastasis within 10 years while the subset 2 represents the low risk group with subjects which remained metastasis free during the first 10 years. Click on **Summary Statistics** in order to proceed the comparison of the pre-defined groups (Figure 7). This is the standard comparison provided for any pair of subsets.



Figure 7: Summary statistics for comparison of 2 pre-defined subsets

The top line of the window describes the 2 subsets with the definition of the threshold values that were entered: TDM < 3650 for Subset 1 and TDM \ge 3650 for Subset 2. The Subject Totals box gives details on the number of subjects for each subset: **75 for Subset 1** (later displayed in red colour) and **123 for Subset 2** (displayed in blue colour). The comparison of the Age variable is then displayed as a Histogram view (left) and as a Box plot graph (right).

The similarity between the 2 subsets highlights that there is no difference in Age repartition between the high risk of disease recurrence and the low risk group.

Conclusion: The disease recurrence is independent of the Age of the subjects. This analysis allows you to compare the impact of the demographic variables (Age and Sex if it is defined) onto the variable of interest, here the clinical outcome which is the disease recurrence.

Clinical question : does the **Estrogen Receptor Status** have an impact on the clinical outcome TDM (Time to Distant Metastasis) ?

We are going to bring an extra parameter in the analysis now. Expand the Medical History node under the Subjects folder and then the Tumour characteristics node. The Estrogen Receptor Status is a categorical binary variable defined by 2 values: Positive or Negative corresponding to ER+ and ER- tumours respectively. Select the Estrogen Receptor Status and drag the variable towards the Results/Analysis window on the right. The Results/Analysis window is updated with a comparison of the repartition of ER+ and ER- status between the 2 previously defined subsets (Figure 8).



Figure 8: Comparison of ER status (subset 1 versus subset 2)

We observe here that there is a significant difference (pValue <0.05) between the 2 subsets : there is a higher incidence of ER- (41%) in the high risk group (Subset 1) as compared to the low risk group (27%, Subset 2).

Conclusion : in this cohort, there is a higher proportion of subjects with ER negative status in the high risk group (incidence of metastasis within 10 years) as compared to the low risk group and the difference is statistically significant at a 95% confidence level. Indeed, there are conflicting studies associating the tumour ER status and the incidence of distant metastasis (Berman et al, 2013).

Lesson 3 : Survival Analysis

Go to the **Comparison** Tab. Click on the Clear button if your selection is not empty at that stage. Drag the entire study node into Subset 1, and click on Advanced Workflow and select Survival Analysis in the Analysis pulldown menu. This workflow estimates the proportion of subjects living for 2 distinct categories over a period of time. The purpose of the Kaplan Meier plot is to report the comparison of the 2 conditions.

The « Variable Selection » window pops up and you must enter the variables that you wish to examine (Figure 9).

							-				
tranSMART All :						Browse	Analyze	Gene Signature/Lists	GWAS	Upload Data	Utilitie
«)	Comparison	Summary Statistics	Grid View	Advanced Workflow	Data Export	Export Jobs	Workspace	Genome Browser			
Active Filters and m	🕄 Analysis 🕶										
	Analysis: Survi	val Analysis									
	Cohorts: Subset 1: (\Pu	blic Studies\GEO Studie	s\Breast Cancer	\Desmedt(2007) GSE73	90\)						
Navigate Terms	Variable Se	election 🕜									
a Across Trais → Tubic Studies b Cancer Cell Line Encyclopedia (1070) → ECD Studies a CED Studies → Prost Cancer → Breast Cancer → Bre	Time Select time v drag it into t variable is re	variable from the Data S he box. For example, "S quired.	et Explorer Tree urvival Time". T	and Select a vertice of the box. For exercise of the box For variable is not re- variable is not re- clear	on which you w example, "Cane is should be "bin quired.	would like to sort cer Stage". If this ned" using the o	the cohort and c s variable is cont ption below. This sional Data	Censoring Variable trag it. Select the appropriate of incous into the box. For examp s This variable is not requ Clear	ensoring varia le, "Survival (f ired.	ble and drag it Censor) -> Yes". Clear	

Figure 9: Survival Analysis window

There are 3 categories of variables here :

- Time
- Category
- Censoring Variable

The **Survival Analysis** refers to a time dependent variable. It might be any time related measure like TDM in this study. Next, we want to define 2 categories of subjects and the objective is to compare their impact on the time dependent endpoint. This parameter must be entered as a Category variable and we must define which categories we want to compare. The Censored variable allows you to take into account the subjects that are censored along the study, i.e. the subjects with no record for the clinical measure.

Select TDM node and drag it into the Time box. Select the Veridex Signature Risk Goup under Risk Assessment and drag it into the Category box. The Veridex signature corresponds to the 76-gene signature that is evaluated in this validation cohort. It has been applied to classify the 198 subjects of the dataset into 2 categories, namely « Poor » and « Good » defining respectively the disease progression with incidence of distant metastasis within 10 years or not. As this variable is already categorical, there is no need to bin.

Before proceeding with the calculations, we are going to fill the Censor box. The censored variable related to the TDM measure is DMCENS. Expand the Censored, DMCENS node under the « Distant Metastasis Free Survival » folder node and drag the **No** category into the Censor box. As explained in the « Censoring Variable » box, only the node corresponding to the No values must be registered here. Your final window should look like the view in Figure 10.

tranSMART	Browse Analyze Gene Signature/Lists GWAS Unload Data Utilities
v1.2	Comparison Summary Statistics Grid View Advanced Workflow Data Evont Evont John Workspace Genome Rowser
Active Filters and Clear	
	Analysis
	Aitaiyais, Sui vivai Autaiyais
	Cohorts: Subset 1: (\Public Studies\GEO Studies\Breast Cancer\Desmedt(2007) GSE7390\)
	Variable Selection 🕖
ravigate terms	1
Across Trials	Time Colours Constitute Verifield
Cancer Cell Line Encyclopedia (1070)	Select time variable from the Data Set Explorer Tree and Select a variable on which you would like to sort the cohort and drag it. Select the appropriate censoring variable and drag it
GEO Studies	drag it into the box. For example, "Survival Time". This into the box. For example, "Cancer Stage". If this variable is continuous into the box. For example, "Survival (Censor) -> Yes".
😟 🦳 Asthma	Vanable is required. (ex. Age, then it should be "binned" using the option below. This is not required. Variable is not required.
🕀 🗁 Breast Cancer	
Desmedt(2007) GSE7390 (198)	VTime to Distant Metastasis (Days)\
Biomarker Data (198)	
🖃 🔂 Clinical Data (198)	
Disease Free Survival (198) Disease Meteodesis Free Survival (198)	
Concored DMCENS (198)	
abc No. (51)	Clear High Dimensional Data Clear Clear
abc Yes (147)	
Element Censoreds, DMFSCENS (198)	Enable binning
123 Distant Metastasis Free Survival (Days) (198)	Run
123 Time to Distant Metastasis (Days) (198)	
Overall Survival (198)	
Risk Assessment (198)	
Adjuvant! Online (AoL) Risk Group (198)	
O St. Gallas Criteria Rick Group (19)	
Veridex Signature Risk Group (198)	
abc Good (55)	
abc Poor (143)	
123 Adjuvant! Online (AoL) 10-Year Survival Probability (198)	
-122 Nottingham Prognostic Index (NPI) Score (196)	

Figure 10: Parameters to enter for Kaplan Meier graphs

Then click on the Run button. The window is updated and the Kaplan Meier graph is displayed (Figure 11).



Figure 11: Kaplan Meier graph for TDM

This graph is similar to Figure 1 in the original publication, except for the X axis which is cut at 10 years, i.e. 3650 days in the paper. As indicated by the label, the blue line corresponds to the Veridex good profile group while the red line is associated with the poor profile group.

When focusing on the first 10 years of the study (Days <3650), we observe here a higher incidence of distant metastasis for the subjects of the « poor » Veridex profile category as compared to those of the « good » profile group. Values measured after 4000 days are difficult to comment as the number of Censored patients increases significantly (represented as + signs on the plot).

Conclusion: This study validates the 76-gene signature as a good predictor for the incidence of distant metastasis within 10 years.