Reports of
The Machine Learning and Inference Laboratory

The Side Out Foundation Metastatic Breast Cancer Database, an Open-access Portal for “multi-omics” Molecular Data and More

Sanja Avramovic

Negin Asadzadehzanjani

Reyhaneh Mogharabnia

Elisa Baldelli

Emanuel Petricoin

Mariaelena Pierobon

Janusz Wojtusiak

MLI 18-2
August 2018

RESEARCH AND EDUCATION IN MACHINE LEARNING

The Side Out Foundation Metastatic Breast Cancer Database, an Open-access Portal for "multi-omics" Molecular Data and More

Sanja Avramovic, Negin Asadzadehzanjani, Reyhaneh Mogharabnia , Elisa Baldelli, Emanuel Petricoin, Mariaelena Pierobon, Janusz Wojtusiak

Abstract

Although there are available databases of molecular data based on human malignancie, to the best of our knowledge there is no web-based publicly accessible database portal where broad "multi-omic" profiles are captured from metastatic breast cancer (MBC) patients along with demographic, clinical and pathological data. This report describes a database and a portal, which are primarily used to record information collected through the Side-Out Clinical Trials, a series of prospective Phase II clinical trials targeting refractory MBC where molecular information is used to drive treatment selection.

Introduction

Over the past few years, high volume molecular data collected from human malignancies has gained a lot of popularity. However, to better understand how biological information can be used to improve outcome for cancer patients, there is a need for creating user-friendly and easily accessible internet-based portals where these molecular data are captured and can be easily accessed by physicians, scientists, and the general public. At this time, there are a lot of databases containing the molecular characteristics of the primary tumors such as International Cancer Genome Consortium (ICGC), Cancer Genome Atlas (TCGA) and Clinical Proteomic Tumor Analysis Consortium (CPTAC) to name a few.[1] However, few resources are available that addresses the metastatic lesions including cBioPortal, Cancer RNA-seq Nexus, and human cancer metastasis database (HCMDB).[2] In this project, we developed a novel database containing demographic, clinical, and pathological information, outcome data, and multi-omics based molecular profiles of patients with metastatic breast cancer (MBC).

Database construction

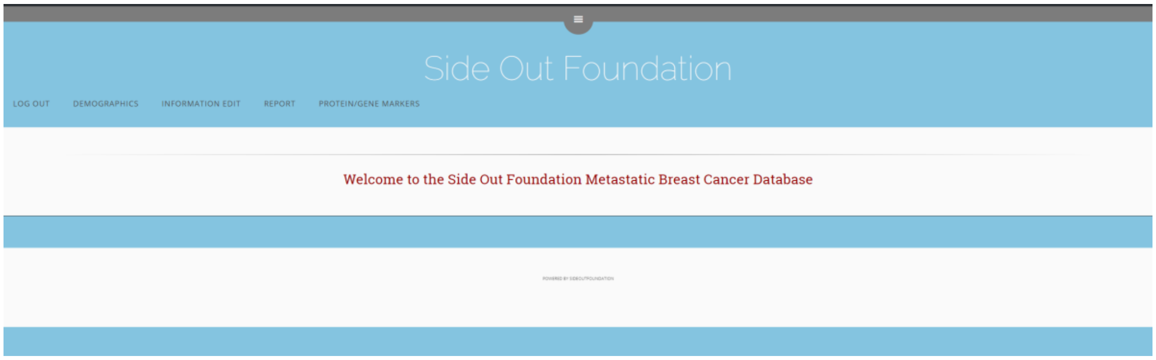
In order to create the portal, MySQL as an open-source relational database management system was used. The custom–codes were written by using the PHP server-side scripting language and the users can have access to the recorded data. The database is used to keep records of information collected from Side-Out Clinical Trials. These are a series of prospective Phase II clinical trials that are specifically designed for refractory MBC patients in which molecular information are used to drive treatment selection (NCT01074814, NCT01919749, NCT03195192).

Data storage

All data are de-identified and stored in secure environment within the George Mason University, Center for Discovery Science and Health Informatics.[3] The center is established in 2007 and housed within the Department of Health Administration and Policy, College of Health and Human Services, and it is dedicated to research on interrelated topics of discovery science and health informatics. Technical capacity of the center allows for secure and reliable hosting and analysis of health data in a secure environment. Access to the servers is governed by strict procedures and policies. The constructed database along with web-based interface are hosted on a dedicated virtual machine to provide additional levels of reliability and control.

Registration and login

The users can request access to the database using "Multi-OmicMBCPortal.gmu.edu" address. Once access is granted by the site administrator, users can create a profile and login to the database. Once logged in, they could have access to the information in the database.



Welcome to the Side Out Foundation Metastatic Breast Cancer Database

Different sections of the database

The database consists of a number of sections including Demographics, Report and Protein/Gene Markers tabs. In this section, over 700 different data fields are collected for each patient. These sections will be discussed in detail below. In the physical implementation of the database, these sections correspond to tables in relational database model.

Demographic Section

The demographic section contains patient's demographics including: Subject ID, Year of Birth, Sex, Side Out Trial, Age at initial diagnosis, Age at trial enrollment as well as an additional field to enter comments about the patient if necessary. Subject ID is a five-digit number separated by two dashes and is unique to each patient (For instance: 01-100). The first two digits on the left show the Side Out trial identifier. The side-out trial shows in which Side Out trial the patient was admitted. There are two age fields in this section which are used to import patient's age at diagnosis and enrollment, respectively.

Primary Tumor Characteristic Section

For each patient, tumor characteristics including tumor histotype and Hormone Receptor (HR) status and HER2 level of the primary lesion are recorded. There is also comment fields in this section where additional information regarding the tumor can be recorded.

Metastatic Lesion Characteristic Section

This section contains the following information: biopsy site of the metastatic lesion at trial enrollment; HR and HER2 status of the metastatic lesion, patient's previous treatments in terms of number of treatments, therapy received and time to progression (in days) on the last treatment before trial enrollment. Treatment recommendation based on patients' molecular profile collected through the Side Out trials are also recorded along with the time to progression.

Molecular Data Section

The molecular data tab includes the multi-omic information collected on the metastatic lesion as part of the Side Out clinical trial workflow. Molecular information was collected using immunohistochemistry (IHC), RNA expression, Fluorescence in situ hybridization (FISH), exome sequencing (not available for the Side Out 1 trial), and Reverse Phase Protein Microarray (RPPA)-based functional pathway activation mapping assay of FDA approved drug targets and downstream substrates.

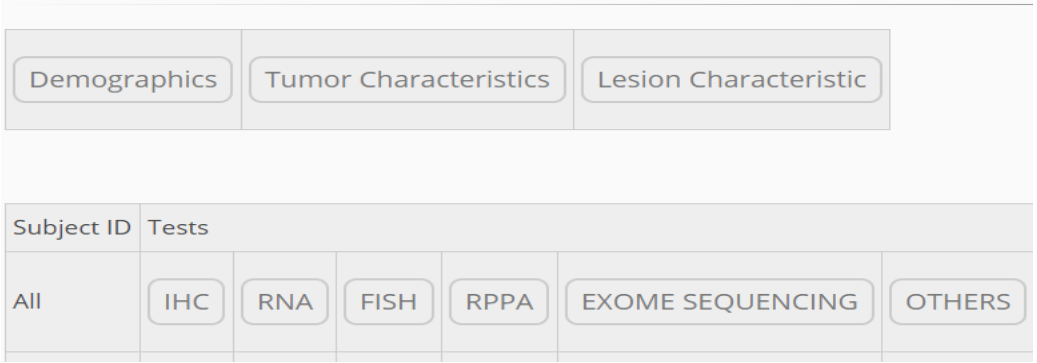
In the "IHC tab", the following information can be found for 26 therapeutically relevant proteins: staining intensity, percentage of positive cells, H score, and cut-off based on which patients were scored.

The "RNA Expression tab" contains information on the expression level of 106 transcripts. For each transcript, expression (under-expressed, no change, over-expressed, measured but not informative) and ratio between sample and reference control population is recorded.

The "FISH tab" includes information concerning 4 targetable genes (EGFR, HER2, cMyc, and TOP2A). Along with the gene name, gene amplification status, number of cells counted, number of positive cells, ratio between number of copies of the gene of interest versus a reference gene, and cut-off value are captured from each lesion.

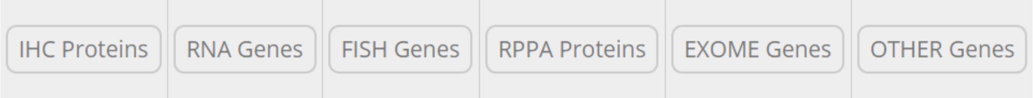
The "RPPA tab" collects information on the activation level of 26 FDA approved drug target and downstream substrates to capture pathway centered functional data. Pathway activation was established a priori on large cohorts of cancer samples. Patients with activation level above the 75th percentile of the reference population were considered positive. Finally, the "Exome Sequencing tab" includes a list of 600 targetable genes along with their characteristics (wild type, mutation, amplification, deletion, translocation, gene fusion, copy number variation, etc.). The approach we took in order to collect the data for the Exome sequencing and the RPPA was based on Pierobon et al. work. [4]

Report



Data can be downloaded by accessing the report tab. This page allows to visualize all the information recorded and could easily be used by interested third party. The recorded data can be exported by users through a number of options: whole dataset based on molecular analysis, by patients across all analytes, and by analyte across all patients. All clinical and pathological data as well as the multi-omic molecular data and outcome can be downloaded once access to the site is granted.

Protein/Gene Markers



In this section, the users can access the list of all molecular information captured by the database. By clicking on each of the tabs they can see the related subgroups. For instance, IHC protein tab includes subgroups such as Androgen Receptor, BCRP, c-Kit, CAV-1, CK 5/6, etc. RNA gene tab includes subgroups ABCC1, ABCG2, ADA, AR, AREG, etc. FISH gene tab has following subgroups: cMyc, EGFR, HER2/Neu, TOP2A. RPPA protein subgroups are AKT S473, c-Abl T735, c-Kit Y719, Cyclin D1, Cyclin D1 T286, etc., and exome gene tab includes subgroups such as 19Q, 1P, ABI1, ABL1, ABL2, etc. A new subgroup can be added, or an existing one can be deleted from any of the tabs.

Conclusion

To the best of our knowledge, the presented work represents the first publicly available database that keeps records of a wide variety of multi-omics profiles of metastatic lesions and outcome of MBC patients. The de-identified database also includes demographic, clinical and pathological information of each patient enrolled in the Side Out trials. This database captures unique aspects of metastatic breast cancers and can potentially be used for correlative analyses and hypothesis-generating studies. Finally, this dynamic web-based portal allows for the dissemination of data achieved from existing or upcoming clinical and translational studies targeting breast cancer patients.

References:

1. Pavlopoulou A, Spandidos DA, Michalopoulos I. Human cancer databases (Review). Oncology Reports. 2014;33(1):3–18.
2. Zheng G, Ma Y, Zou Y, Yin A, Li W, Dong D. HCMDB: the human cancer metastasis database. Nucleic Acids Research. 2017;46(D1).
3. Center for Discovery Science and Health Informatics at George Mason University [Internet]. Center for Discovery Science and Health Informatics at George Mason University. [cited 2018Jun25]. Available from: http://dshi.gmu.edu/
4. Pierobon M, Ramos C, Wong S, et al. Enrichment of PI3K-AKT–mTOR Pathway Activation in Hepatic Metastases from Breast Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 2017;23(16):4919-4928.

Table of contents

1 demographics Page number: 2
2 exome Page number: 5
3 exome_p Page number: 6
4 fish Page number: 7
5 fish_p Page number: 8
6 ihc Page number: 9
7 ihc_p Page number: 10
8 other Page number: 11
9 other_p Page number: 12
10 rna Page number: 13
11 rna_p Page number: 14
12 rppa Page number: 15
13 rppa_p Page number: 16
14 Relational schema Page number: 17

1 demographics

Creation: Nov 27, 2017 at 09:28 AM

Column	Type	Attributes	Null	Default	Extra	Links to	Comments	MIME
id	char(20)		No					
ini	char(3)		No					
dob	varchar(100)		No					
sot	char(20)		Yes	NULL				
age_d	char(20)		Yes	NULL				
age_t	char(20)		Yes	NULL				
sex	char(2)		No					
cmnt	text		No					
histo	char(20)		No					
cmnt2	text		No					
hr	char(20)		No					
her	char(20)		No					
cmnt3	text		No					
bio	char(10)		No					
cmnt4	char(10)		No					
hr2	char(10)		No					
her2	char(20)		Yes	NULL				
no_treat	int(10)		No					
all_treat1	char(50)		No					
no1	int(10)		No					
cmnt5_1	text		No					
cmnt6_1	text		No					
all_treat2	char(50)		No					
no2	int(10)		No					
cmnt5_2	text		No					
cmnt6_2	text		No					
all_treat3	char(50)		No					
no3	int(10)		No					
cmnt5_3	text		No					
cmnt6_3	text		No					
all_treat4	char(50)		No					
no4	int(10)		No					
cmnt5_4	text		No					
cmnt6_4	text		No					
all_treat5	char(50)		No					
no5	int(10)		No					
cmnt5_5	text		No					
cmnt6_5	text		No					

all_treat6	char(50)		No					
no6	int(10)		No					
cmnt5_6	text		No					
cmnt6_6	text		No					
all_treat7	char(50)		No					
no7	int(10)		No					
cmnt5_7	text		No					
cmnt6_7	text		No					
all_treat8	char(50)		No					
no8	int(10)		No					
cmnt5_8	text		No					
cmnt6_8	text		No					
all_treat9	char(50)		No					
no9	int(10)		No					
cmnt5_9	text		No					
cmnt6_9	text		No					
all_treat10	char(50)		No					
no10	int(10)		No					
cmnt5_10	text		No					
cmnt6_10	text		No					
all_treat11	char(50)		No					
no11	int(10)		No					
cmnt5_11	text		No					
cmnt6_11	text		No					
all_treat12	char(50)		No					
no12	int(10)		No					
cmnt5_12	text		No					
cmnt6_12	text		No					
all_treat13	char(50)		No					
no13	int(10)		No					
cmnt5_13	text		No					
cmnt6_13	text		No					
all_treat14	char(50)		No					
no14	int(10)		No					
cmnt5_14	text		No					
cmnt6_14	text		No					
all_treat15	char(50)		No					
no15	int(10)		No					
cmnt5_15	text		No					
cmnt6_15	text		No					
l_treat	char(50)		No					
cmnt5	text		No					
cmnt6	text		No					
l_time	int(10)		No					
so_treat	char(50)		No					
cmnt5_16	text		No					
cmnt6_16	text		No					
so_time	int(10)		No					

cmnt7	text		No					
ihc	char(20)		No					
rna	char(20)		No					
fish	char(20)		No					
exome	char(20)		Yes	NULL				
rppa	char(50)		Yes	NULL				
other	char(20)		No					

2 exome

Creation: Jun 20, 2017 at 12:53 PM

Column	Type	Attributes	Null	Default	Extra	Links to	Comments	MIME
gene	varchar(100)		No					

3 exome_p

Creation: Jul 01, 2018 at 08:20 PM

Column	Type	Attributes	Null	Default	Extra	Links to	Comments	MIME
id	varchar(200)		No					
gene3	varchar(100)		No					
status	varchar(50)		No					
descript	text		No					

4 fish

Creation: Jun 02, 2017 at 03:40 PM

Column	Type	Attributes	Null	Default	Extra	Links to	Comments	MIME
gene2	varchar(100)		No					

5 fish_p

Creation: Jul 01, 2018 at 08:27 PM

Column	Type	Attributes	Null	Default	Extra	Links to	Comments	MIME
id	varchar(50)		No					
gene2	varchar(100)		No					
res3	varchar(50)		No					
cell	int(100)		No					
ta1	varchar(100)		Yes	NULL				
percent	varchar(100)		Yes	NULL				
ta2	varchar(100)		Yes	NULL				
ratio2	varchar(50)		Yes	NULL				
cmnt3	text		No					
cut3	text		Yes	NULL				

6 ihc

Creation: May 31, 2017 at 03:31 PM

Column	Type	Attributes	Null	Default	Extra	Links to	Comments	MIME
protein	char(50)		No					

7 ihc_p

Creation: Jul 01, 2018 at 08:23 PM

Column	Type	Attributes	Null	Default	Extra	Links to	Comments	MIME
id	varchar(50)		No					
protein	varchar(50)		No					
res	varchar(20)		No					
si	int(20)		No					
ps	int(20)		No					
hscore	int(50)		No					
cmnt	text		No					
cut	text		Yes	NULL				

8 other

Creation: Jun 20, 2017 at 01:02 PM

Column	Type	Attributes	Null	Default	Extra	Links to	Comments	MIME
gene	varchar(100)		No					

9 other_p

Creation: Jul 01, 2018 at 08:25 PM

Column	Type	Attributes	Null	Default	Extra	Links to	Comments	MIME
id	varchar(100)		No					
gene4	varchar(100)		No					
cmnt5	text		No					

10 rna

Creation: May 31, 2017 at 03:34 PM

Column	Type	Attributes	Null	Default	Extra	Links to	Comments	MIME
gene	varchar(100)		No					

11 rna_p

Creation: Jul 01, 2018 at 08:28 PM

Column	Type	Attributes	Null	Default	Extra	Links to	Comments	MIME
id	varchar(50)		No					
gene	varchar(100)		No					
exp	varchar(50)		No					
ratio	varchar(50)		Yes	NULL				
res2	text		No					
cmnt2	text		No					

12 rppa

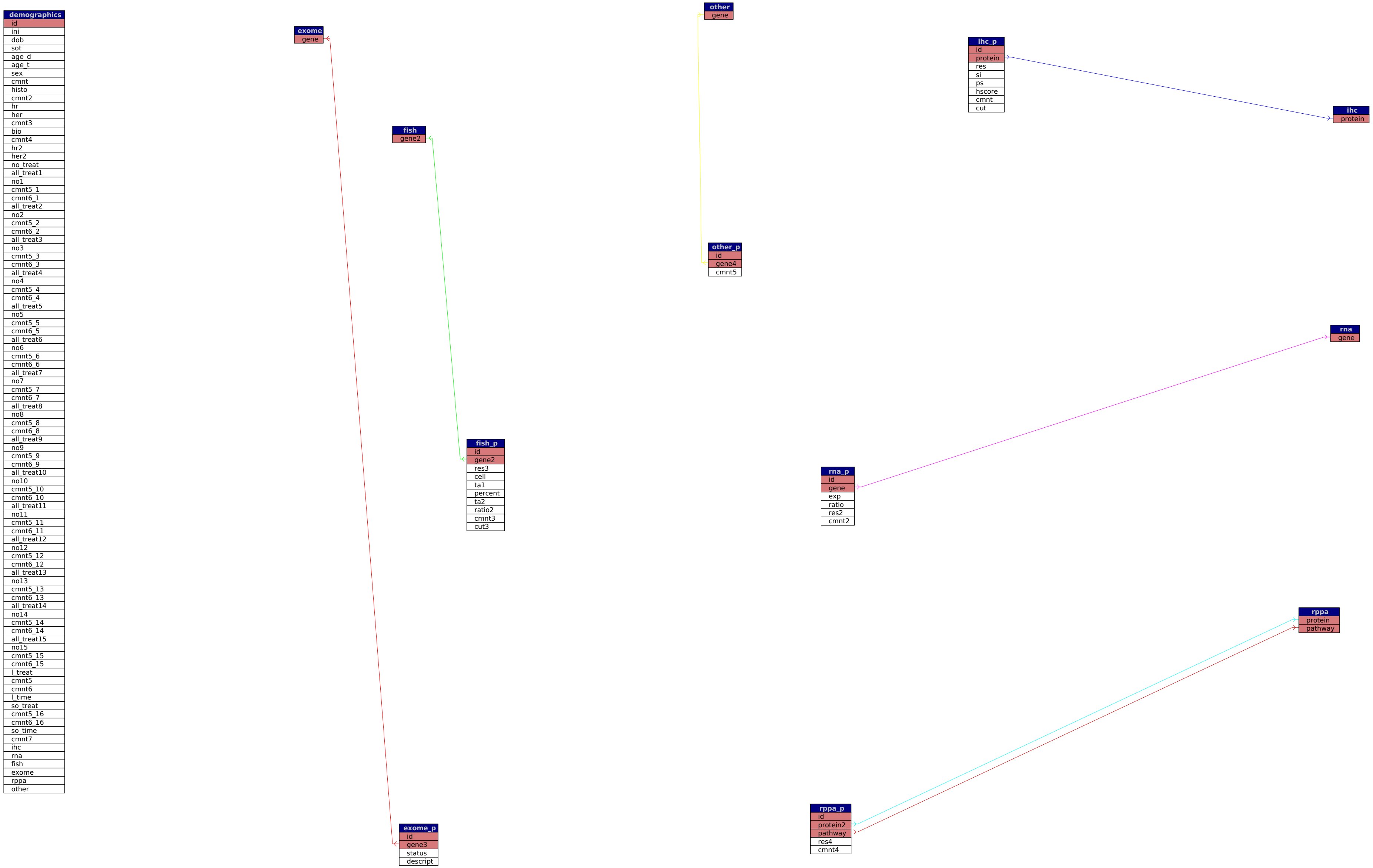
Creation: Jun 13, 2017 at 05:23 PM

Column	Type	Attributes	Null	Default	Extra	Links to	Comments	MIME
protein	varchar(50)		No					
pathway	varchar(100)		No					

13 rppa_p

Creation: Jul 01, 2018 at 08:29 PM

Column	Type	Attributes	Null	Default	Extra	Links to	Comments	MIME
id	varchar(50)		No					
protein2	varchar(100)		No					
pathway	varchar(100)		No					
res4	varchar(50)		Yes	NULL				
cmnt4	varchar(200)		No					


demographics
id
ini
dob
sot
age_d
age_t
sex
cmnt
histo
cmnt2
hr
her
cmnt3
bio
cmnt4
hr2
her2
no_treat
all_treat1
no1
cmnt5_1
cmnt6_1
all_treat2
no2
cmnt5_2
cmnt6_2
all_treat3
no3
cmnt5_3
cmnt6_3
all_treat4
no4
cmnt5_4
cmnt6_4
all_treat5
no5
cmnt5_5
cmnt6_5
all_treat6
no6
cmnt5_6
cmnt6_6
all_treat7
no7
cmnt5_7
cmnt6_7
all_treat8
no8
cmnt5_8
cmnt6_8
all_treat9
no9
cmnt5_9
cmnt6_9
all_treat10
no10
cmnt5_10
cmnt6_10
all_treat11
no11
cmnt5_11
cmnt6_11
all_treat12
no12
cmnt5_12
cmnt6_12
all_treat13
no13
cmnt5_13
cmnt6_13
all_treat14
no14
cmnt5_14
cmnt6_14
all_treat15
no15
cmnt5_15
cmnt6_15
l_treat
cmnt5
cmnt6
l_time
so_treat
cmnt5_16
cmnt6_16
so_time
cmnt7
ihc
rna
fish
exome
rppa
other
exome
gene
fish
gene2
exome_p
id
gene3
status
descript
fish_p
id
gene2
res3
cell
ta1
percent
ta2
ratio2
cmnt3
cut3
other
gene
other_p
id
gene4
cmnt5
ihc_p
id
protein
res
si
ps
hscore
cmnt
cut
ihc
protein
rna
gene
rna_p
id
gene
exp
ratio
res2
cmnt2
rppa
protein
pathway
rppa_p
id
protein2
pathway
res4
cmnt4

A publication of the *Machine Learning and Inference Laboratory*
College of Health and Human Services
George Mason University
Fairfax, VA 22030-4444 U.S.A.
http://www.mli.gmu.edu

Editor: J. Wojtusiak

The *Machine Learning and Inference (MLI) Laboratory Reports* are an official publication of the Machine Learning and Inference Laboratory, which has been published continuously since 1971 by R.S. Michalski's research group (until 1987, while the group was at the University of Illinois, they were called ISG (Intelligent Systems Group) Reports, or were part of the Department of Computer Science Reports).