Synapse ID: syn2813558 **DOI:** doi:10.7303/syn2813558 **Upload Destination:** Synapse Storage

2.2 - Challenge Data Description

Rules and Terms of Use

All Challenge participants must abide by the data Terms of use and Challenge rules.

General Description of Data Used in the Prostate Cancer DREAM Challenge

The data sets used in this Challenge are collated data based on **comparator arm** data sets of Phase III prostate cancer clinical trial hosted on the *Project Data Sphere®* platform. The data represents 4 cancer trials of first line metastatic Hormone Refractory Prostate Cancer (HRPC) patients, where all patients received docetaxel treatment in the comparator arm. These 4 sets of raw trial data are consolidated into one set of data tables with standardized format. Summary variables are created to capture clinically important covariates. The raw data used for this challenge was provided to the *Project Data Sphere* initiative by the following data providers:

StudyID	Data Provider	# of patients	Min, median and max of follow up times for all patients (days)
ASCENT2	Memorial Sloan Kettering Cancer Center	476	(4,357,796)
CELGENE	Celgene	526	(11,279,750)
EFC6546	Sanofi	598	(8,642.5,1594)
AZ	AstraZeneca	470	(6,463,1148)

Data from Celgene, Sanofi and Memorial Sloan Kettering Cancer Center were used to create the training data sets, while AstraZeneca data will be held back for the leaderboard and final scoring. Besides the dependent variables for the Subchallenges, the data also include clinical covariates such as patient demographics, lesion measure, medical history, prior surgery and radiation, prior medicine, vital sign, lab, etc.

Data Tables to be Released for the Challenge

Six data tables are released for this Challenge. The CoreTable is the main table which is summarized at patient level with dependent variables and clinical covariates. The rest of the tables are raw longitudinal tables used to create CoreTable which are at the event level and could be used for additional variable creation and/or exploration. Please see table below for a brief description of each data table.

Table Name	Level	Table Description
CoreTable	Patient level	Subject level summary table including dependent variables for the two Subchallenges, and clinical covariates
PriorMed	Patient-event level	Prior Medication table records medication patients took or had taken before 1st treatment date of the trial.
MedHistory	Patient-event level	Medical History table records patient reported diagnoses (co-existing disease) at time of patient screening to participate in the trial.
LesionMeasure	Patient-event level	Lesion table records target and non-target lesion measurement.
LabValue	Patient-event level	Lab test table includes all lab data (hematology and urinary lab)
VitalSign	Patient-event level	Vital Sign table records patient vital sign (height, weight, etc.)

All data can be viewed and downloaded from the Accessing Data (https://www.synapse.org/#!Synapse:syn2813558/wiki/209590) page, and description of the data fields can be found in our data dictionary.

General Notes on the Data

- Note that not all trials sampled the exact same clinical variables, so there will be missing values between trials for some clinical variables. It is up to the participants to determine which variables are useful in building their models.
- In SAS, in general, both blank and . means missing, except blank shows missing for character variable, while . shows missing for numeric variable. The challenge data set is prepared in SAS, then output into .csv format, which is why you will see blank and . for different variable. Data dictionary provides more explanation on each variable.
- Columns in the data files are delimited with quotes and a comma ("",)

CoreTable

The CoreTable is the main table for the Challenge and represents the core patient level data. It includes dependent variables for the 2 Subchallenges as well as summarized clinical covariates. These clinical covariates are created from sets of standardized longitudinal data tables (including but not limited to the 5 longitudinal data tables released for the Challenge) which cover information about demographics, co-existing disease conditions, prior treatment of the tumor and other co-existing conditions, important baseline lab results and vital sign, lesion measure and early response to therapy. The following is important information solvers need to know about this table to use it for the Challenge.

1) Dependent variable definition:

- Subchallenge 1a and 1b has two dependent variables: DEATH (death flag) as the survival outcome variable, and LKADT_P (last known alive day - in days) as time to event. When DEATH="YES", it means death is reported for this patient, blank means no death is reported.
- Subchallenge 2 has two dependent variables: DISCONT (discontinuation due to Adverse Event flag by 3 month or 93 days, 1/0), and ENTRT*PC (treatment end day - in days). ENDTRS*C is an intermediate variable created in the process of dependent variable creation reflecting large categories of discontinuation reason, detail of this variable please see Creation of dependent variable for Q2.docx in the Challenge Data folder.

2) Treatment variables (TRT1*ID, TRT2*ID, TRT3*ID): shows all drugs patient receive during the trial. TRT1*ID (treatment 1) represents study drug (since these are

comparator arm data, it will be placebo), TRT2*ID (treatment 2) represents docetaxel in this case, TRT3*ID (treatment 3) represents the third drug patient receive in this case prednisone (note: no third drug for AstraZeneca trial).

3) Covariates are either baseline information or static information for the patient. The following covariates are available only when StudyID="ASCENT2", they were kept because they are important for prognostic purpose: SMOKE (smoking flag), SMOKFREQ (smoking frequency), SMOKSTAT (smoking status), GLEAS*DX (gleason score at diagnosis), TSTAG*DX (Primary Tumor staging at diagnosis).

4) Target vs. non-target lesion (TARGET, NON_TARGET):
RECIST Criteria including target vs. non-target Lesion definition (https://www.eortc.be/Recist/documents/RECISTGuidelines.pdf)

5) Prostate Cancer Staging (TSTAG_DX):
AJCC Staging for Prostate Cancer (https://cancerstaging.org/references-tools/quickreferences/Documents/ProstateSmall.pdf)

6) Gleason Score (GLEAS_DX):
A system of grading prostate cancer tissue based on how it looks under a microscope. Gleason scores range from 2 to 10 and indicate how likely it is that a tumor will spread. A low Gleason score means the cancer tissue is similar to normal prostate tissue and the tumor is less likely to spread; a high Gleason score means the cancer tissue is very different from normal and the tumor is more likely to spread.

7) ECOG Performance Status: The ECOG score (published by Oken et al in 1982), also called the WHO or Zubrod score (after C. Gordon Zubrod), runs from 0 to 5, with 0 denoting asymptomatic for disease and 5 death.

ECOG Status	Description
0	Asymptomatic. (Fully active, able to carry on all pre-disease activities without restriction).
1	Symptomatic but completely ambulatory. (Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature. For example, light housework, office work).
2	Symptomatic, < 50% in bed during the day. (Ambulatory and capable of all self care but unable to carry out any work activities. Up and about more than 50% of waking hours).
3	Symptomatic, > 50% in bed, but not bedbound. (Capable of only limited self-care, confined to bed or chair 50% or more of waking hours).
4	Bed-bound. (Completely disabled. Cannot carry on any self-care. Totally confined to bed or chair).
5	Death

PriorMed

PriorMed is a longitudinal data table containing event level data. It includes all medications a patient took or has taken before first treatment of the trial. Therefore, **the entire PriorMed table contains only baseline information**.

1) Date variables to use for PriorMed table: CMSTDT*PC (medication start day) and CMENDT*PC (medication end day) are days relative to first treatment date (first treatment date is used as the reference date in majority of cases, very small number of patients use consent date as reference date).

2) Censoring: All records in PriorMed have CMSTDT*PC <=0, which means only records of medication patients started taking prior to first treatment are released. When CMENDT*PC>0, patients continue to take the medication after starting first treatment, the value of CMENDT*PC in this case could potentially give away survival information. Therefore, CMENDT*PC is censored at zero to avoid confounding.

3) Variable to use for prior medical treatment:

- CMTRT is the original verbatim captured for the medication, **CMDECOD is the standardized version and the version recommended for use in most cases**.
- CMDECOD (standardized term of medication) is available for all 4 trials except for StudyID="ASCENT2". There are two options to fill this gap. Option 1: CMTRT (original verbatim of medication) can be used in place of CMDECOD for ASCENT2, cons are potential inconsistency across trials due to raw data nature of CMTRT. Option 2: CMATC4 (Chemical Class of medication, also the lowest classification for treatment category) is assigned based on CMTRT (original verbatim of medication) for ASCENT2 trial, so that all 4 trials have CMATC4 populated. Cons are loss of some granularity compared to standardized term.

4) Treatment classification: CMATCn (n=1-4, 1 is highest category, 4 is lowest) are the classification for the medication in hierarchy. CMDICTV is the dictionary version of the MedDRA where the classification is sourced from.

- CMATC4 (Chemical Class of medication, also the lowest classification for treatment category) is available for 4 trials. CMATC1, CMATC2 and CMATC3 (higher level classification for medication) are only missing for StudyID="AZ". However, since AZ data will be used for leader-board and final scoring, these variables are advised to use only for exploratory analysis not for model submission.

LabValue

LabValue is a longitudinal data table containing event level data. It includes all lab tests a patient took from screening up to 84 days after first treatment date. Therefore, the LabValue table has both baseline and post baseline information about lab. Lab test records post baseline can be used to discover new prognostic factors, however only baseline lab can be used for model submission for leader-board and final scoring.

1) Date variable to use for LabValue table: LBDT_PC (Lab test day) is days relative to first treatment date (first treatment date is used as the reference date in majority of cases, very small number of patients use consent date as reference date).

2) **How baseline lab is defined: LBBLFL=='Y'.**

- LBBLFL variable creation logic: For STUDYID='AZ' and 'EFC6546', baseline flag already exists therefore is used to populate LBBLFL. For STUDYID='CELGENE' and 'ASCENT2', in general the last non-missing value immediately prior to first dose was the record flagged as baseline.
- Note any record where VISIT='Screening' or 'Pre-enrollment' could be flagged as baseline even if its date value might indicate it is post baseline. For example, a patient has a baseline lab record where LBDT_PC=71, those records are flagged baseline because VISIT="Pre-enrollment".

3) What does it mean when LBBLFL=='Y' (baseline lab) and LBDT_PC>0 (lab test day greater than 0)? There are less than 1% records in such situation, most of which has consent date as reference date instead of first treatment, but all of which are marked as "Pre-enrollment" or "Screening", therefore should be OK.

4) For LabValue table in the training data sets, LBDT_PC<=84 for all records, meaning all labs captured within 84 days of reference date (first treatment date) are available for use. For LabValue table in leader-board and final scoring data sets, only baseline lab will be available (LBBLFL=='Y') for solvers.

5) Recommendation on what variables to use for Lab test result:
There are two sets of lab value available in the data, the original and the standardized version, both provided by the data providers. **In general, the standardized version are recommended for use (e.g. LBSTRESC - standardized lab test result character value),** but it is always helpful to see the original value (LBORRES - original lab test result) as reference for user.

6) LBNRIND is Reference Range Indicator, which flags when the lab value is higher, lower or normal based on the range used in the study and is populated in most cases. Caution for the users: it is recommended to compare lab test result against a set of non study specific set of normal ranges to make sure values make sense and flagging

done OK if solver decides to use the information.

LesionMeasure

LesionMeasure is a longitudinal data table that includes event level data. It includes all lesion test result a patient has from screening up to 98 days or Cycle 4 after reference day (first treatment date for StudyID="CELGENE" and "AZ" and randomization date for StudyID="EFC6546"). Therefore, the LesionMeasurement table has both baseline and post baseline information about lab. Lab test records post baseline can be used to discover new prognostic factors, however only baseline lab can be used for model submission for leader-board and final scoring.

1) Missing Data: StudyID="ASCENT2" did not provide event level lesion data, it only provides summary level lesion location variables, which have been included in CoreTable. Therefore, there are no event level data for ASCENT2 in this table.

2) Date variable to use for LesionMeasurement table: LSDT*PC (lesion test measurement day) is days relative to reference day (AZ and CELGENE trial mostly use 1st treatment date as reference day for LSDT*PC, except a very small number of patients use consent date, while EFC6546 use randomization date as reference day).

3) **How to identify baseline lesion records: VISIT==’SCREENING’ or VISIT== ‘1’.**

4) For training data sets, LSDT_PC (lesion test measurement day)<=98 or VISIT<=CYCLE 4 for all records, meaning all labs are captured within 98 days of reference day or at most cycle 4 follow-up, because lesion gets at least one more measure after screening in 4 cycles.

5) Lesion test result: There are two sets of lesion result available in the data: the original and the standardized version, both provided by the data providers. In general, **the standardized version is recommended for use (LSSTRESC <standardized test result> and LSSTRESU <unit for standardized test result>)**, but it is always helpful to see the original value (LSORRES <original test result> and LSORRESU <unit for original test result>) as reference for user.

MedHistory

MedHistory is a longitudinal data table that includes event level data. It includes all medical diagnoses patients provided at screening, which covers co-existing conditions patients have. Therefore, **the entire MedHistory table contains only baseline information.**

1) Date variable to use in MedicalHistory table: MHSTDT_P (medical condition start day) is days from SCREENING

2) All medical history captured in the table is from SCREENING, therefore they are are baseline information.

3) ASCENT2 situation: ASCENT2 trial did not provide event level medical history table, it only provided summary level pre-specified co-morbidity variables, which have been included in CoreTable. Therefore, there are no event level data for ASCENT2 in MedicalHistory table. These co-morbidity terms are captured in MHTERM (original verbatim for medical condition).

4) Medical condition captured in medical history table:
> MHTERM are the original verbatim captured for the condition, **MHDECOD is the standardized version and the version recommended for use in most cases**. Note, ASCENT2 has missing MHDECOD, reason stated above.
> MHLLT (Low Level Term), MHHLT (High Level Term), MHHLGT (High Level Group Term), MHBODSYS (Body System or Organ code) represent classification of the medical condition, from lowest to the highest level.

VitalSign

VitalSign is a longitudinal data table that includes event level data. It includes all vital sign a patient took from screening up to 84 days after reference day. Therefore, the VitalSign table has both baseline and post baseline information. Vital sign records post baseline can be used to discover new prognostic factors, however only baseline vital sign can be used for model submission for leader-board and final scoring.

1) Date variables to use in VitalSign table: VSDT*PC (vital sign day) is days from a reference date; AZ and CELGENE use first treatment date as reference day for VSDT*PC (vital sign day), ASCENT2 use consent date, EFC6546 use randomization date.

2) VISIT variable shows the visit schedule for the patients. We did not try to standardize this variable because every trial is slightly different in the way visit is scheduled. For the 3 trials in the training data sets, VISIT variable will have value such as "SCREENING", "Pre-enrollment", "Cycle 1", "Cycle 1 day 1", where "SCREENING"/"Pre-enrollment" meant the visits for patient screened to enroll in the trial, and "Cycle" means the cycle of treatment patients receive. For AZ, VISIT variable has value "1", "2", "1.01", etc., where VISIT="1" means screening with very few exceptions, the decimal means un-scheduled visits in-between, e.g. "1.01" means unscheduled visit after screening.

3) **How baseline vital sign is defined: VSBLFL==’Y’. (vital sign baseline flag)**
>For studyid='ASCENT2', since all records were screening the last non-missing value for a test closest to first dose was selected to be flagged as baseline.
>For studyid='EFC6546' the provided database contained the VSBLFL flag.
>For studyid=AZ or studyid=CELGENE, in general the last non missing value prior to first dose will be flagged as baseline although a screening record that appears to be post baseline could also be flagged.

4) What does it mean when VSBLFL==’Y’ (baseline vital sign) and VSDT_PC>0 (vital sign capture day greater than 0)? Couple of situation it could happen:
>When cycle 1 day 1 is considered candidate for baseline during baseline flag creation, VSDT*PC (vital sign day) could be greater than 0, however VSDT*PC<=12, which is reasonable.
>For ASCENT2, some records have VSDT_PC>0 (~5% of total records). However, the reference day is consent date and all records were screening. It should be OK

5) For training data sets, VSDT*PC<=84 for almost all records, meaning all labs are captured within 84 days of reference date, except for 2 records with VSDT*PC=277 from ASCENT2, one patient height and weight taken at day 277 from consent, while VISIT="SCREENING".

6) Recommendation for what variable to use for vital sign value: There are two sets of vital sign value available in the data, the original and the standardized version, both provided by the data providers. In general, **the standardized version is recommended for use (e.g. VSSTRESC <standardized vital sign test result>)**, but it is always helpful to see the original value (VSORRES <original vital sign test result>) as reference for user.

Challenge Data Release By Phase

Data Sets For...	Trial Used	Information Released	Information Not Released
Training	All 3 trials	complete data sets for CoreTable, all 5 longitudinal data sets	None
Leader-board	1/3 AZ trial	CoreTable covariates, 5 longitudinal data sets baseline information	CoreTable dependent variable related information, Longitudinal data sets non-baseline information.

Final Scoring	2/3 AZ trial	CoreTable covariates, 5 longitudinal data sets baseline information	CoreTable dependent variable related information, Longitudinal data sets non-baseline information.

Created by J. Christopher Bare (chris.bare) on Monday, March 16, 2015 6:07 PM

Modified by James Costello (james.costello) on Friday, March 20, 2015 6:53 PM

Wiki History

Life Sciences DISCOVERY FUND (http://www.lsdfa.org/)

NIH National Heart, Lung, and Blood Institute (http://www.nhlbi.nih.gov/)

ALFRED P. SLOAN FOUNDATION (http://www.sloan.org/)

NIH National Institute on Aging (http://www.nia.nih.gov/)

NIMH National Institute of Mental Health (http://www.nimh.nih.gov/)

CHILDREN'S TUMOR FOUNDATION (http://www.ctf.org/)