MSK-CHORD Data: Overview of Clinical Data

MSK CHORD Study

- For this class, you will be conducting two projects using data from the MSK CHORD study.
- The data is available in the cBioPortal data repository for cancer genomics.
- You can download the MSK-CHORD data here: https://www.cbioportal.org/study/summary?id=msk_chord_2024

Data Files

- When you download the full dataset, it should be a **.tar.gz** file.
- When you unzip this file, it should create a folder that contains the following data files:
  - **data_clinical_patient.txt**
  - **data_timeline_diagnosis.txt**
  - **data_clinical_sample.txt**
  - **data_timeline_treatment.txt**
  - **data_timeline_tumor_sites.txt**
  - **data_timeline_progression.txt**
  - **data_timeline_surgery.txt**

`data_clinical_patient.txt`

- This file contains basic demographic information, clinical information, and outcomes for each patient in MSK CHORD.
- There are n = 24,950 patients in total.
- The variable `PATIENT_ID` uniquely identifies each patient.
  - This variable can be used to link patient information across data files.

Variables in `data_clinical_patient.txt`

- Key variables in `data_clinical_patient.txt` include:
  - `GENDER`
  - `CURRENT_AGE_DEID`
  - `STAGE_HIGHEST_RECORDED`
  - `SMOKING_PREDICTIONS_3_CLASSES`
  - `PRIOR_MED_TO_MSK`
  - `OS_MONTHS`
  - `OS_STATUS`

Variables in `data_clinical_patient.txt`

- `CURRENT_AGE_DEID` - I believe this is a “de-identified” age of the patient (in years) at **“time zero”**.
- `STAGE_HIGHEST_RECORDED` - The highest cancer stage observed for each patient.
- `PRIOR_MED_TO_MSK` - An indicator of whether or not the patient received treatment prior to starting treatment at MSK.
  - Many patients start their treatment at somewhere other than MSK.
  - Their treatment information in other files may not be as detailed at time points prior to starting treatment at MSK.

Variables in `data_clinical_patient.txt`

- `OS_MONTHS` - Time in months (from **time zero**) until date last observed.
- The last observation made on a patient can occur for **two reasons**:
  - Death
  - End of study or patient was lost to follow up
- `OS_EVENT`
  - Equals `0:LIVING` if patient was not dead at last follow up
  - Equals `1:DECEASED` if patient was dead at last follow up

Variables in `data_clinical_patient.txt`

- As an example, let's first load in the `data_clinical_patient.txt` file as a data frame called `ClinOutcomes`:

```
library(dplyr)
ClinOutcomes <- read.delim("~/Downloads/msk_chord_2024/data_clinical_patient.txt",
                           comment.char="#")
```

- Only keep the variables `PATIENT_ID`, `GENDER`, `CURRENT_AGE_DEID`, `OS_MONTHS`, `OS_STATUS`:

```
ClinOutcomes <- ClinOutcomes %>%
                select(PATIENT_ID, GENDER, CURRENT_AGE_DEID, OS_MONTHS, OS_STATUS)
```

Variables in `data_clinical_patient.txt`

- Look at the first few rows of `ClinOutcomes`

```
1 head(ClinOutcomes)
```

```
  PATIENT_ID GENDER CURRENT_AGE_DEID OS_MONTHS  OS_STATUS
1  P-0000012 Female               68 118.45466   0:LIVING
2  P-0000015 Female               45  13.90683 1:DECEASED
3  P-0000036 Female               68 115.46289   0:LIVING
4  P-0000041 Female               53  13.61094 1:DECEASED
5  P-0000066 Female               71  76.63553   0:LIVING
6  P-0000058 Female               54  60.75610 1:DECEASED
```

- For example, patient `P-0000036` was a Female, aged 68 at time zero, and was still alive at 115.46 months after time zero.
- Patient `P-0000058` is a Female patient, aged 54 at time zero, and deceased 60.75 months after time zero.

The file `data_timeline_diagnosis.txt`

- This file contains important information taken when the cancer was first diagnosed.
  - `PATIENT_ID`
  - `START_DATE`
  - `DX_DESCRIPTION` - description of initial diagnosis
  - `STAGE_CDM_DERIVED` - cancer stage at initial diagnosis
- A very important variable is `START_DATE`.
- This allows you to locate where **time zero** is relative to the **date of diagnosis**.

The file `data_timeline_diagnosis.txt`

- Let's load in the `data_timeline_diagnosis.txt` file as the data frame `Diagnosis`:

```
DiagnosisFull <- read.delim2("~/Downloads/msk_chord_2024/data_timeline_diagnosis.txt")
```

and only keep the variables `PATIENT_ID`, `START_DATE`, `STAGE_CDM_DERIVED`

```
Diagnosis <- DiagnosisFull %>%
             select(PATIENT_ID, START_DATE, STAGE_CDM_DERIVED)
head(Diagnosis)
```

```
  PATIENT_ID START_DATE STAGE_CDM_DERIVED
1  P-0078271       -232         Stage 1-3
2  P-0002243      -1299         Stage 1-3
3  P-0005017       -219           Stage 4
4  P-0038278      -5368         Stage 1-3
5  P-0005035      -1027           Stage 4
6  P-0021411      -5746         Stage 1-3
```

- For example, patient `P-0078271` was diagnosed with Stage 1-3 cancer 232 days prior to time zero.

The file `data_timeline_diagnosis.txt`

```
1 head(Diagnosis)
```

```
  PATIENT_ID START_DATE STAGE_CDM_DERIVED
1  P-0078271       -232         Stage 1-3
2  P-0002243      -1299         Stage 1-3
3  P-0005017       -219           Stage 4
4  P-0038278      -5368         Stage 1-3
5  P-0005035      -1027           Stage 4
6  P-0021411      -5746         Stage 1-3
```

- For every patient, **time zero** is the time at which their **tumor was first sequenced** at MSK.
- This means patient `P-0078271` was diagnosed with cancer **232 days before** their tumor was first sequenced at MSK.
- Patient `P-0021411` was diagnosed with cancer **5746 days before** their tumor was first sequenced at MSK.

The file `data_timeline_diagnosis.txt`

```
1 head(Diagnosis)
```

```
  PATIENT_ID START_DATE STAGE_CDM_DERIVED
1  P-0078271       -232         Stage 1-3
2  P-0002243      -1299         Stage 1-3
3  P-0005017       -219           Stage 4
4  P-0038278      -5368         Stage 1-3
5  P-0005035      -1027           Stage 4
6  P-0021411      -5746         Stage 1-3
```

- Time of first tumor sequencing is **not especially meaningful**.
- When doing an analysis, you will need to compute outcomes **relative to a different starting time**.
- **Overall survival** measured from **time of diagnosis** is a common outcome.
- **Overall survival** measured from **time starting a particular type of treatment** is often a meaningful outcome to look at.

The file `data_timeline_diagnosis.txt`

- Notice that the number of rows in `Diagnosis` is 25,145, but the number of patients in the study was 24,940

```
dim(DiagnosisFull)
```

```
[1] 25145    12
```

```
length( unique(DiagnosisFull$PATIENT_ID) )
```

```
[1] 24940
```

- In some cases, this looks to be the result of some patient with multiple, distinct cancer diagnoses:

```
## This patient appears to have a breast cancer diagnosis about 2 years after
## a lung cancer diagnosis
DiagEx1 <- subset(DiagnosisFull, PATIENT_ID=="P-0012895")
DiagEx1
```

```
    PATIENT_ID START_DATE STOP_DATE EVENT_TYPE SUBTYPE         SOURCE
178  P-0012895      -1016        NA  Diagnosis Primary Tumor Registry
179  P-0012895       -346        NA  Diagnosis Primary Tumor Registry
                                                  DX_DESCRIPTION AJCC
178 CARCINOID TUMOR, MALIGNANT | LUNG, MIDDLE LOBE (M8240/3 | C342)
179           LOBULAR CA + IFDC/DCIS | BREAST, UOQ (M8522/3 | C504)
    CLINICAL_GROUP PATH_GROUP STAGE_CDM_DERIVED
```

```
178              4          4           Stage 4
179             1A         1A         Stage 1-3
                                                 SUMMARY
178 Distant metastases/systemic disease
179 Localized
```

The file `data_timeline_diagnosis.txt`

- Another case of multiple, distinct cancer diagnoses:

```
## This patient appears to have a breast cancer diagnosis about 2 years after
## a lung cancer diagnosis
DiagEx2 <- subset(DiagnosisFull, PATIENT_ID=="P-0000012")
DiagEx2
```

```
    PATIENT_ID START_DATE STOP_DATE EVENT_TYPE SUBTYPE         SOURCE
319  P-0000012      -9641        NA  Diagnosis Primary Tumor Registry
320  P-0000012        -42        NA  Diagnosis Primary Tumor Registry
                                                DX_DESCRIPTION AJCC
319 INFILTRATING DUCT CARCINOMA | BREAST, NOS (M8500/3 | C509)
320           ADENOCARCINOMA, NOS | LUNG, NOS (M8140/3 | C349) IIIB
    CLINICAL_GROUP PATH_GROUP STAGE_CDM_DERIVED
319             99         99         Stage 1-3
320             3B         99         Stage 1-3
                                              SUMMARY
319 N/A
320 Distant
```

The file `data_timeline_diagnosis.txt`

- In other cases, it looks like essentially the same diagnosis is entered twice (possibly due to how the diagnoses are entered and extracted in their system)

```
## This patient appears to have one rectal cancer diagnosis
DiagEx3 <- subset(DiagnosisFull, PATIENT_ID=="P-0027479")
DiagEx3
```

```
    PATIENT_ID START_DATE STOP_DATE EVENT_TYPE SUBTYPE         SOURCE
293  P-0027479       -172        NA  Diagnosis Primary Tumor Registry
294  P-0027479       -172        NA  Diagnosis Primary Tumor Registry
                                                 DX_DESCRIPTION AJCC
293 ADENOCARCINOMA, NOS | RECTOSIGMOID JUNCTION (M8140/3 | C199)
294           ADENOCARCINOMA, NOS | RECTUM, NOS (M8140/3 | C209)
    CLINICAL_GROUP PATH_GROUP STAGE_CDM_DERIVED
293             3A         99         Stage 1-3
294             3A         99         Stage 1-3
                                            SUMMARY
293 Distant metastases/systemic disease
294 Distant metastases/systemic disease
```

The file `data_clinical_sample.txt`

- This dataset has 25040 rows. + A few patients have multiple samples, but most have only one sample.

```
1  Samples <- read.delim("~/Downloads/msk_chord_2024/data_clinical_sample.txt", comment.char="#")
```

- This contains information about how each **tumor specimen** was collected and results from sequencing.
- For now, I just want to look at the following variables:
  - `SAMPLE_ID`
  - `PATIENT_ID`
  - `CANCER_TYPE`
  - `SAMPLE_TYPE`
  - `METASTATIC_SITE`
  - `PRIMARY_SITE`

The file data_clinical_sample.txt

```r
Samples <- Samples %>%
           select(SAMPLE_ID, PATIENT_ID, CANCER_TYPE, SAMPLE_TYPE,
                  METASTATIC_SITE, PRIMARY_SITE)
head(Samples)
```

```
          SAMPLE_ID PATIENT_ID                CANCER_TYPE SAMPLE_TYPE
1 P-0000012-T03-IM3  P-0000012 Non-Small Cell Lung Cancer  Metastasis
2 P-0000012-T02-IM3  P-0000012              Breast Cancer     Primary
3 P-0000015-T01-IM3  P-0000015              Breast Cancer  Metastasis
4 P-0000036-T01-IM3  P-0000036 Non-Small Cell Lung Cancer     Primary
5 P-0000041-T01-IM3  P-0000041              Breast Cancer     Primary
6 P-0000066-T01-IM3  P-0000066              Breast Cancer  Metastasis
   METASTATIC_SITE PRIMARY_SITE
1             Neck         Lung
2   Not Applicable       Breast
3            Liver       Breast
4   Not Applicable         Lung
5   Not Applicable       Breast
6 Parasternal Mass       Breast
```

The file `data_clinical_sample.txt`

- The variable `CANCER_TYPE` gives a more "**direct**" classification of the type of cancer a patient has.
- This is a pan-cancer study, but there are only a few (major) cancer types represented

```
1 table(Samples$CANCER_TYPE)
```

```
             Breast Cancer          Colorectal Cancer
                      5368                       5543
Non-Small Cell Lung Cancer          Pancreatic Cancer
                      7809                       3109
           Prostate Cancer
                      3211
```

The file `data_clinical_sample.txt`

- The `SAMPLE_TYPE` tells you whether or not the sample was taken from "**Primary**" tissue or "**Metastatic**" tissue.
  - If Primary, that means the tissue was taken from the origin of the cancer (this should be the same as `CANCER_TYPE`).

```
1 table(Samples$SAMPLE_TYPE)
```

```
Local Recurrence       Metastasis          Primary          Unknown
              98             8878            15928              136
```

- A few samples are labeled as "**Local Recurrence**".
  - These cases are where someone previously had surgery or radiation and the cancer had later returned.
  - These samples will have been taken at the primary site or near the primary site.

The file `data_clinical_sample.txt`

- `METASTATIC_SITE` tells you where the specimen was collected (if the sample was not taken from the primary site)
- There are many different metastatic sites

```
met_site <- table(Samples$METASTATIC_SITE)
length(met_site)  ## 192 Metastatic sites listed
```

```
[1] 192
```

```
met_site[1:10]
```

```
                                Abdomen   Abdomen,Soft Tissue
                      1                                    34                                    1
       Abdominal wall                        Abdominal Wall                           Adnexa or
                      4                                    17                                    1
Adnexa or Endometrium                Adnexa or Endomterium                              Adrenal
                      1                                     5                                   24
        Adrenal Gland
                     76
```

The file `data_timeline_treatment.txt`

- This contains **longitudinal** treatment information for many patients.

```
Treatment <- read.delim("~/Downloads/msk_chord_2024/data_timeline_treatment.txt")
```

- Key variables in this dataset
  - `START_DATE` - times are relative to time of first tumor sequencing
  - `SUBTYPE` - the classification of the treatment
  - `AGENT` - the particular name of the treatment
- Not everyone has treatment information: 21,473 of the 24.940 patients have treatment information

```
length( unique(Treatment$PATIENT_ID) )
```

```
[1] 21473
```

The file `data_timeline_treatment.txt`

- It's probably better to refer to the data in this file as "**Systemic Treatment Information**"
  - *Systemic Treatment* - absorbed in the blood and can **spread throughout** the body
  - *Local Treatment* - surgery, radiation, (some types of chemo)
- Each systemic treatment is classified as 1 of 8 types of systemic treatment in the SUBTYPE variable

```
1 table(Treatment$SUBTYPE)
```

```
      Biologic  Bone Treatment          Chemo        Hormone         Immuno 
          8816            5635           73036          26670           4030 
Investigational          Other       Targeted 
          7284              12           9460 
```

The file `data_timeline_treatment.txt`

- `START_DATE` has the start date of the particular treatment.
- The data file is not sorted by `START_DATE`, so let's sort it by date within patient first

```
1  Treatment <- Treatment %>%
2               select(-RX_INVESTIGATIVE, -FLAG_OROTOPICAL) %>%
3               arrange(PATIENT_ID, START_DATE)
4
5  head(Treatment)
```

```
  PATIENT_ID START_DATE STOP_DATE EVENT_TYPE SUBTYPE            AGENT
1  P-0000012      -5437     -5369  Treatment   Chemo CYCLOPHOSPHAMIDE
2  P-0000012      -5437     -5326  Treatment   Chemo     FLUOROURACIL
3  P-0000012      -5437     -5327  Treatment   Chemo     METHOTREXATE
4  P-0000012         33        40  Treatment   Chemo        CISPLATIN
5  P-0000012         33        65  Treatment   Chemo        ETOPOSIDE
6  P-0000012         61        68  Treatment   Chemo      CARBOPLATIN
```

- Many patients start **multiple treatments at the same time**.
  - Patient P-0000012 received three chemotherapies at time -5437

The file `data_timeline_treatment.txt`

- If starting multiple treatments at the same time (or within a few days of each other), this is often labeled as receiving a “combination therapy”

```
1 TrtEx <- subset(Treatment, PATIENT_ID=="P-0000216")
2 head(TrtEx)
```

	PATIENT_ID	START_DATE	STOP_DATE	EVENT_TYPE	SUBTYPE	AGENT
431	P-0000216	-7	75	Treatment	Hormone	LETROZOLE
432	P-0000216	90	90	Treatment	Investigational	INVESTIGATIONAL
433	P-0000216	146	147	Treatment	Targeted	EVEROLIMUS
434	P-0000216	146	147	Treatment	Hormone	EXEMESTANE
435	P-0000216	210	211	Treatment	Hormone	MEGESTROL
436	P-0000216	238	640	Treatment	Chemo	CAPECITABINE

- Patient P-0000216 started a targeted therapy and a hormone therapy at the same time.
  - 146 days before first tumor sequencing

Data from Radiology Reports

- A major feature of this data resource is the extracted information from **radiology reports**
- Structured information from the radiology reports was **not extracted** by a human.
- Instead, they use **natural language processing** algorithms to construct summary measures from each radiology report.
- The files which contain this information are:
  - `data_timeline_cancer_presence.txt`
  - `data_timeline_cancer_presence.txt`
  - `data_timeline_progression.txt`

data_timeline_cancer_presence.txt

```
CancerPresence <- read.delim("~/Downloads/msk_chord_2024/data_timeline_cancer_presence.txt")
CancerPresence <- CancerPresence %>%
                  select(PATIENT_ID, START_DATE, PROCEDURE_TYPE,
                         NLP_HAS_CANCER_PROBABILITY, HAS_CANCER)
```

- This file contains NLP “predictions” about the **presence of cancer** based on text from a radiology report.
- Some of the key variables are:
  - PATIENT_ID
  - START_DATE
  - PROCEDURE_TYPE
  - NLP_HAS_CANCER_PROBABILITY
  - HAS_CANCER

data_timeline_cancer_presence.txt

- PROCEDURE_TYPE is the type of scan performed. This is eiter a **CT scan**, a **PET scan**, or an **MRI**

```
1 table(CancerPresence$PROCEDURE_TYPE)
```

```
    CT     MR    PET 
245839  85881  66808 
```

- For each scan, you get an overall, NLP-derived, probability of cancer presence

```
1 head(CancerPresence)
```

```
  PATIENT_ID START_DATE PROCEDURE_TYPE NLP_HAS_CANCER_PROBABILITY HAS_CANCER
1  P-0000012       -255             MR                   0.0010828          N
2  P-0000012       -253             MR                   0.0004522          N
3  P-0000012        -68            PET                   0.9997510          Y
4  P-0000012         10            PET                   0.9997148          Y
5  P-0000012         10             MR                   0.0003389          N
6  P-0000012         10             CT                   0.9996852          Y
```

`data_timeline_cancer_presence.txt`

- I'm not 100% sure yet what the variables `CHEST`, `ABDOMEN`, `PELVIS`, `HEAD`, and `OTHER` represent yet.
- This file, by itself, might not be that useful.
  - You can always look at `data_timeline_diagnosis.txt` to see when someone had an official cancer diagnosis.
  - The file `data_timeline_tumor_sites.txt` can give more information about particular tumor sites.
- This file does illustrate how they record the information from their NLP models.

The file `data_timeline_tumor_sites.txt`

- This file contains imaging results made over time.
- Key variables in this data file:
  - `PATIENT_ID`
  - `START_DATE`
  - `SOURCE_SPECIFIC`
  - `TUMOR_SITE`
  - `CHEST`, `ABDOMEN`, `PELVIS`, `HEAD`, `OTHER`
- This dataset has 506,647 rows from 23,362 patients:

```
TumorSites <- read.delim("~/Downloads/msk_chord_2024/data_timeline_tumor_sites.txt")
TumorSites <- TumorSites %>%
              select(-EVENT_TYPE, -SUBTYPE, -STOP_DATE, -SOURCE) %>% ## remove these variables
              arrange(PATIENT_ID, START_DATE, SOURCE_SPECIFIC) ## sort by START_DATE and scan type
```

The file `data_timeline_tumor_sites.txt`

```
1 head(TumorSites)
```

```
  PATIENT_ID START_DATE SOURCE_SPECIFIC       TUMOR_SITE CHEST ABDOMEN PELVIS
1  P-0000012        -68             PET      Lymph Nodes     0       0      0
2  P-0000012         10              CT      Lymph Nodes     1       0      0
3  P-0000012         10              CT            Other     1       0      0
4  P-0000012         10             PET      Lymph Nodes     0       0      0
5  P-0000012        149              CT      Lymph Nodes     1       0      0
6  P-0000012        281             PET Intra-Abdominal     0       0      0
  HEAD OTHER
1    0     1
2    0     0
3    0     0
4    0     1
5    0     0
6    0     1
```

- This dataset has 506,647 rows.
- The thing that can be confusing about this dataset is that there are often **multiple rows** associated with **one scan**.
- For example, rows 2 and 3 are the **same CT scan** at Day 10

The file `data_timeline_tumor_sites.txt`

```
1 head(TumorSites[,1:6])
```

```
  PATIENT_ID START_DATE SOURCE_SPECIFIC      TUMOR_SITE CHEST ABDOMEN
1  P-0000012        -68             PET     Lymph Nodes     0       0
2  P-0000012         10              CT     Lymph Nodes     1       0
3  P-0000012         10              CT           Other     1       0
4  P-0000012         10             PET     Lymph Nodes     0       0
5  P-0000012        149              CT     Lymph Nodes     1       0
6  P-0000012        281             PET Intra-Abdominal     0       0
```

- **Interpretation**: They filled a separate row for **each tumor site** identified (out of 10 possible tumor locations) by applying their NLP algorithm to that scan.
- For the CT scan at Day 10, the NLP algorithm predicted that there was tumor presence in the **Lymph Nodes** and some **Other** location.
- For the CT scan at Day 149, the NLP algorithm predicted that there was tumor presence in the **Lymph Nodes** but none of the other 9 locations.

The file `data_timeline_tumor_sites.txt`

- As another example, look at patient with ID P-0050196:

```
1  TumorEx <- subset(TumorSites, PATIENT_ID=="P-0050196")
```

- This is the patient depicted in **Fig. 1(d)** of the Jee et al. (2024) paper.

```
1  head(TumorEx[,1:6], 12)
```

```
       PATIENT_ID START_DATE SOURCE_SPECIFIC     TUMOR_SITE CHEST ABDOMEN
367744  P-0050196       -379             PET    Lymph Nodes     0       0
367745  P-0050196       -379             PET         Pleura     0       0
367746  P-0050196       -360              CT           Lung     1       1
367747  P-0050196       -360              CT    Lymph Nodes     1       1
367748  P-0050196       -360              CT         Pleura     1       1
367749  P-0050196       -311              CT           Lung     1       0
367750  P-0050196       -311              CT    Lymph Nodes     1       0
367751  P-0050196       -127              CT    Lymph Nodes     1       1
367752  P-0050196        -99              CT Adrenal Glands     1       1
367753  P-0050196        -99              CT    Lymph Nodes     1       1
367754  P-0050196        -93              CT           Lung     0       0
367755  P-0050196        -93              CT    Lymph Nodes     0       0
```

The file `data_timeline_tumor_sites.txt`

- Patient “P-0050196” had tumor presence in:
  - 2 locations at time -376
  - 3 locations at time -360
  - 2 locations at time -311
  - 1 location at time -127
  - 2 locations at time -99
  - 2 locations at time -93

The file `data_timeline_progression.txt`

```
1 Progression <- read.delim("~/Downloads/msk_chord_2024/data_timeline_progression.txt")
```

- This file contains NLP-derived measures of **cancer progression** at each scan.

```
1 Progression <- Progression %>%
2                select(PATIENT_ID, START_DATE, NLP_PROGRESSION_PROBABILITY,
3                       PROCEDURE_TYPE) %>%
4                arrange(PATIENT_ID, START_DATE)
5 head(Progression)
```

```
  PATIENT_ID START_DATE NLP_PROGRESSION_PROBABILITY PROCEDURE_TYPE
1  P-0000012       -255                  0.03717812             MR
2  P-0000012       -253                  0.38813508             MR
3  P-0000012        -68                  0.28478700            PET
4  P-0000012         10                  0.90068585             CT
5  P-0000012         10                  0.82265690            PET
6  P-0000012         10                  0.32794630             MR
```

- The variable `NLP_PROGRESSION_PROBABILITY` is an estimated probability that an expert radiologist would label the cancer as having "**progressed**" based on the information obtained from that scan.