MSK-CHORD: Clinical Outcomes

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

Common Survival Outcomes

- When defining survival outcomes, an important thing to choose is the “starting time”.
- For this type of data, three common ways to define the starting time are
  - **Time of Diagnosis**
  - **Time of First Metastasis**
  - **Time of First Treatment (for a specific treatment of interest)**

Time from Diagnosis

- Time from cancer diagnosis is a very common starting point.

OS from Diagnosis

- **Overall Survival (OS)** is probably the most common or important survival endpoint used.
- The most meaningful endpoint in many cases.
- A more “clear” endpoint.
    - Less dependent on details of how the study was conducted.
    - OS results are often more consistent across studies.

OS from Diagnosis

- Let’s try to quantify **OS from diagnosis** by using the `data_timeline_diagnosis.txt` and `data_clinical_patient.txt` files

```r
library(dplyr)
DiagnosisFull <- read.delim2("~/Downloads/msk_chord_2024/data_timeline_diagnosis.txt")
DiagnosisFull <- DiagnosisFull %>%
                 arrange(PATIENT_ID, START_DATE)

## For people with multiple diagnoses, only keep rows of first diagnosis
DiagnosisFirst <- DiagnosisFull[!duplicated(DiagnosisFull$PATIENT_ID),]
```

- Now, load `data_clinical_patient.txt` and **merge** the two files

```r
ClinOutcomes <- read.delim("~/Downloads/msk_chord_2024/data_clinical_patient.txt",
                           comment.char="#")
ClinDiag <- inner_join(DiagnosisFirst, ClinOutcomes, by="PATIENT_ID")
```

- Keep only the variables we need for now:

```r
ClinDiag <- ClinDiag %>%
            select(PATIENT_ID, START_DATE, OS_MONTHS, OS_STATUS)
```

OS from Diagnosis

```r
head(ClinDiag)
```

```
  PATIENT_ID START_DATE OS_MONTHS  OS_STATUS
1  P-0000012      -9641 118.45466   0:LIVING
2  P-0000015      -2559  13.90683 1:DECEASED
3  P-0000036       -315 115.46289   0:LIVING
4  P-0000041      -3334  13.61094 1:DECEASED
5  P-0000057      -1036  29.62189 1:DECEASED
6  P-0000058       -205  60.75610 1:DECEASED
```

- To quantify **OS from Diagnosis**, we must create a new variable which records **months of follow up from time of diagnosis**

```r
## Remember START_DATE measures things in days
ClinDiag$OS_FROM_DIAG <- (ClinDiag$OS_MONTHS*30.4375 - ClinDiag$START_DATE)/30.4375
head(ClinDiag)
```

```
  PATIENT_ID START_DATE OS_MONTHS  OS_STATUS OS_FROM_DIAG
1  P-0000012      -9641 118.45466   0:LIVING    435.20210
2  P-0000015      -2559  13.90683 1:DECEASED     97.98076
3  P-0000036       -315 115.46289   0:LIVING    125.81196
4  P-0000041      -3334  13.61094 1:DECEASED    123.14688
5  P-0000057      -1036  29.62189 1:DECEASED     63.65885
6  P-0000058       -205  60.75610 1:DECEASED     67.49121
```

OS from Diagnosis

- I use the conversion **1 month = 30.4375 days**.
  - I’m not sure this conversion matches how they constructed months in MSK-CHORD, but it should be pretty close.
- To estimate **OS survival curves**, you need to convert the variable `OS_STATUS` to either numeric or logical variable

```r
ClinDiag$OS_STATUS_NUM <- ifelse(ClinDiag$OS_STATUS=="1:DECEASED", 1, 0)
```

OS from Diagnosis

- To estimate a **survival curve** in **R**, you can use the `survfit` function from the `survival` package in R

```r
library(survival)

## Use the format: Surv(time, status) when using survfit
os_all_fit <- survfit(Surv(OS_FROM_DIAG, OS_STATUS_NUM) ~ 1, data=ClinDiag)
os_all_fit
```

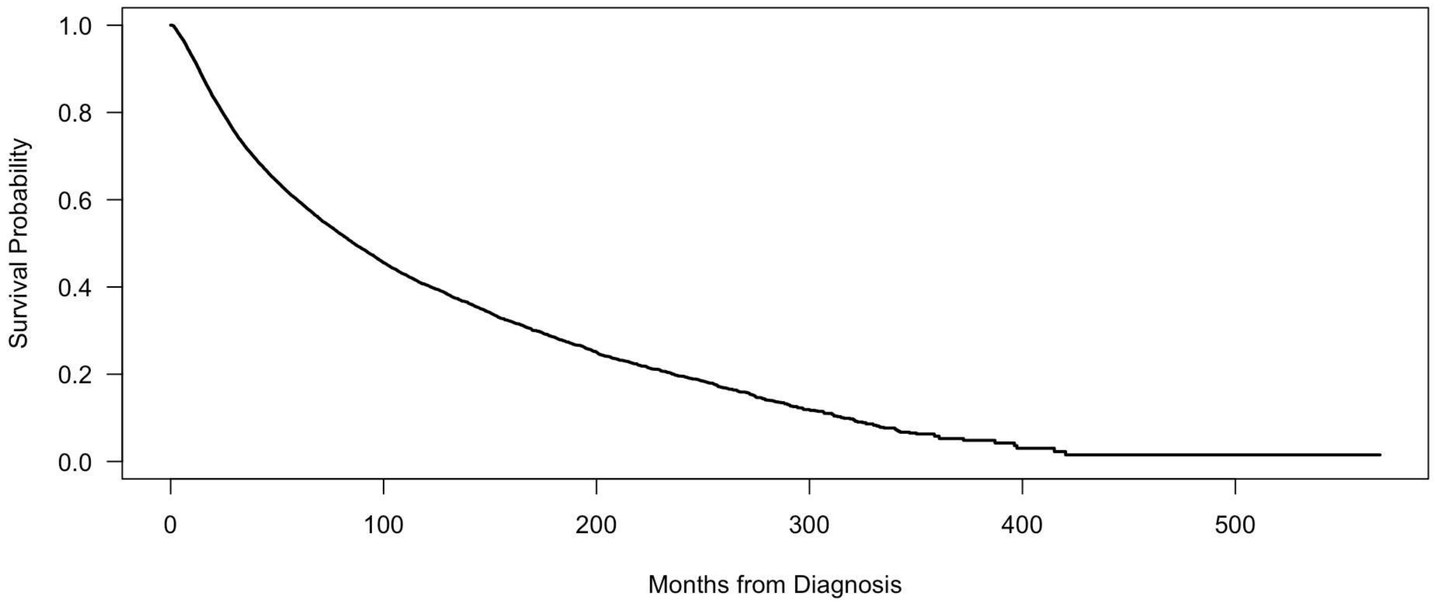
```
Call: survfit(formula = Surv(OS_FROM_DIAG, OS_STATUS_NUM) ~ 1, data = ClinDiag)

         n events median 0.95LCL 0.95UCL
[1,] 24940  11290     86    83.7    88.5
```

- Grouping all patients together, median survival was **86 months** from diagnosis, with a 95% confidence interval of (83.7 - 88.5).

OS from Diagnosis

- You can plot the full OS survival curve by just plotting the object returned by `survfit`

```r
plot(os_all_fit, xlab="Months from Diagnosis", ylab="Survival Probability",
     las=1, lwd=2, conf.int=FALSE)
```



OS from Diagnosis (by Cancer Type)

- To estimate survival for **different subgroups**, you can use the `Surv(time, status) ~ subgrp` syntax.
- I merged `ClinDiag` with the data from `data_clinical_sample.txt` to get cancer type information.
- To estimate OS across types, we can use `survfit` in the following way:

```r
os_by_type <- survfit(Surv(OS_FROM_DIAG, OS_STATUS_NUM) ~ CANCER_TYPE, data=ClinDiagSamp)
```

OS from Diagnosis (by Cancer Type)

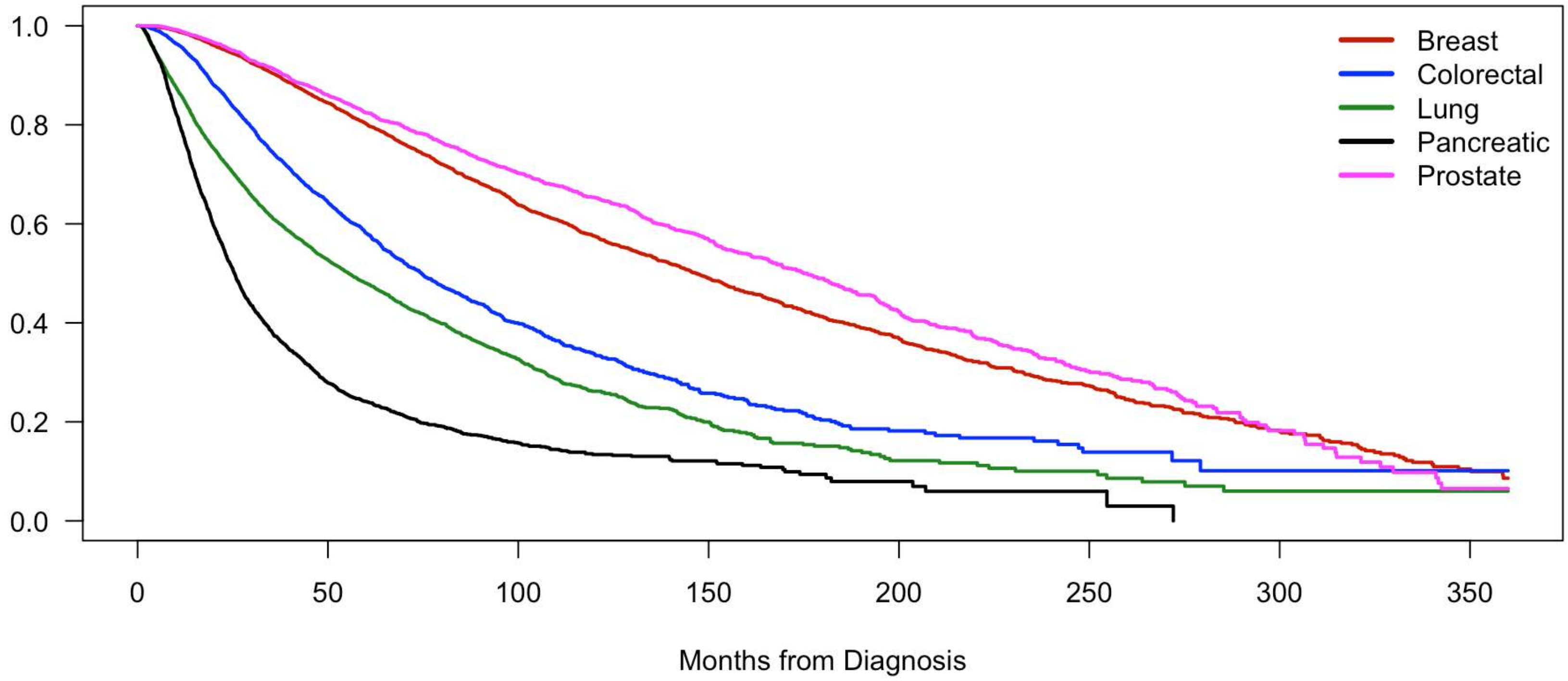
```r
os_by_type
```

```
Call: survfit(formula = Surv(OS_FROM_DIAG, OS_STATUS_NUM) ~ CANCER_TYPE,
    data = ClinDiagSamp)

                                           n events median 0.95LCL 0.95UCL
CANCER_TYPE=Breast Cancer               5354   1983  146.5   139.8   154.0
CANCER_TYPE=Colorectal Cancer           5527   2212   74.5    70.9    78.5
CANCER_TYPE=Non-Small Cell Lung Cancer 7760   3949   55.4    53.1    58.2
CANCER_TYPE=Pancreatic Cancer           3095   2110   25.5    24.4    26.6
CANCER_TYPE=Prostate Cancer             3204   1036  174.9   165.8   185.5
```

- There is considerable heterogeneity in median OS across cancer types:
    - Pancreatic Cancer: 25.6 months
    - Prostate Cancer: 174.6 months

OS from Diagnosis (by Cancer Type)

```r
plot(os_by_type, xlab="Months from Diagnosis", xlim=c(0, 360), las=1,
     col=c("red3", "blue", "forestgreen", "black", "magenta"), lwd=2)
legend("topright", legend=c("Breast", "Colorectal", "Lung", "Pancreatic", "Prostate"),
       col=c("red3", "blue", "forestgreen", "black", "magenta"), lwd=3, bty='n')
```



Time of Metastasis in MSK-CHORD

- Metastasis probably represents the major dividing point in cancer.
- Metastatic vs. non-metastatic (local) represents a major change in
  - Prognosis
  - Which treatment strategies are appropriate
  - How treatments are evaluated
  - Can represent major change in tumor biology.
- For many patients, cancer is already metastatic when cancer is first diagnosed (de novo metastatic disease).
- For many other patients, cancer is not metastatic but later metastasizes.

Time of Metastasis in MSK-CHORD

- Expected survival can be very different in metastatic vs. non-metastatic patients.
- Grouping these patients together combines **two quite different subgroups**.
- One way to make outcomes more comparable is to eliminate non-metastatic patients.
  - That is, those who were never metastatic for any time.
- Measure survival from **time of metastasis** for the remaining patients.

Time of Metastasis in MSK-CHORD

- There is not a single, structured variable in MSK-CHORD that represents a **clinician-verified time of metastasis.**
- Instead, you need to reconstruct it from the `data_timeline_tumor_sites.txt` file.

```r
TumorSites <- read.delim("~/Downloads/msk_chord_2024/data_timeline_tumor_sites.txt")
TumorSites <- TumorSites %>%
              select(PATIENT_ID, START_DATE, TUMOR_SITE, SOURCE_SPECIFIC) %>%
              arrange(PATIENT_ID, START_DATE)
TumorSites[20:26,]  ## Look at Patient P-0000015
```

```
   PATIENT_ID START_DATE TUMOR_SITE SOURCE_SPECIFIC
20  P-0000015        -90       Bone              CT
21  P-0000015        -90      Other              CT
22  P-0000015        -22       Bone              CT
23  P-0000015        -22      Liver              CT
24  P-0000015        -22      Other              CT
25  P-0000015        -22     Pleura              CT
26  P-0000015        -12       Bone             PET
```

Time of Metastasis in MSK-CHORD

- Patient `P-0000015` was originally diagnosed with **breast cancer**.

```r
TumorSites[20:26,]  ## Look at Patient P-0000015
```

```
   PATIENT_ID START_DATE TUMOR_SITE SOURCE_SPECIFIC
20  P-0000015        -90       Bone              CT
21  P-0000015        -90      Other              CT
22  P-0000015        -22       Bone              CT
23  P-0000015        -22      Liver              CT
24  P-0000015        -22      Other              CT
25  P-0000015        -22     Pleura              CT
26  P-0000015        -12       Bone             PET
```

- She had a scan at time -90, which indicates tumor presence in **Bone** and an **Other** site.
- If you check the file `data_timeline_diagnosis.txt`, this patient was diagnosed with **Stage 1 Breast Cancer** at time -2559.
- This indicates the time of first metastasis is at time -90.
- This is 2559 - 90 = 2469 days after first diagnosis.

Time of Metastasis in MSK-CHORD

- Basically, for those who **did not** have metastasis at time of diagnosis, you can find first date of metastasis by looking at
    - The `START_DATE` at which they had a scan which indicates tumor presence in some non-primary site.
    - Use `data_timeline_tumor_sites.txt` to find this.

Metastasis at time of diagnosis in MSK-CHORD

```r
head(DiagnosisFirst)
```

```
  PATIENT_ID START_DATE STAGE_CDM_DERIVED
1  P-0000012      -9641         Stage 1-3
3  P-0000015      -2559         Stage 1-3
4  P-0000036       -315           Stage 4
5  P-0000041      -3334         Stage 1-3
6  P-0000057      -1036           Stage 4
7  P-0000058       -205           Stage 4
                                                      SUMMARY
1 N/A
3 Localized
4 Distant
5 Localized
6 Distant
7 Distant
```

- **Stage** and **Diagnosis Summary** indicate if metastasis was present at time of diagnosis.

Metastasis at time of diagnosis in MSK-CHORD

```r
## Trim white space to right of the SUMMARY variable to clean it up
DiagnosisFirst$SUMMARY <- trimws(DiagnosisFirst$SUMMARY, which="right")

## Now, tabulate the summaries
table(DiagnosisFirst$SUMMARY)
```

```
                      Distant Distant metastases/systemic disease
                         2542                                7506
                      In situ                           Localized
                          200                                5742
                          N/A                Regional both 2and 3
                          486                                1921
 Regional by direct extension                        Regional nos
                         1318                                  18
      Regional to lymph nodes          Regional, direct extension
                         2082                                 745
Regional, extension and nodes          Regional, lymph nodes only
                          951                                1248
             Unknown/Unstaged                   Unstaged  unknown
                           93                                  88
```

Metastasis at time of diagnosis in MSK-CHORD

- Any summary labeled “**Distant**” is metastasis at time of diagnosis.

```r
## Create a variable which indicates a "Distant" summary at diagnosis
DiagnosisFirst$Mets_at_diag <- ifelse(DiagnosisFirst$SUMMARY=="Distant"
                       | DiagnosisFirst$SUMMARY=="Distant metastases/systemic disease",
                                      "Yes", "No")
table(DiagnosisFirst$Mets_at_diag)
```

```
   No   Yes 
14892 10048 
```

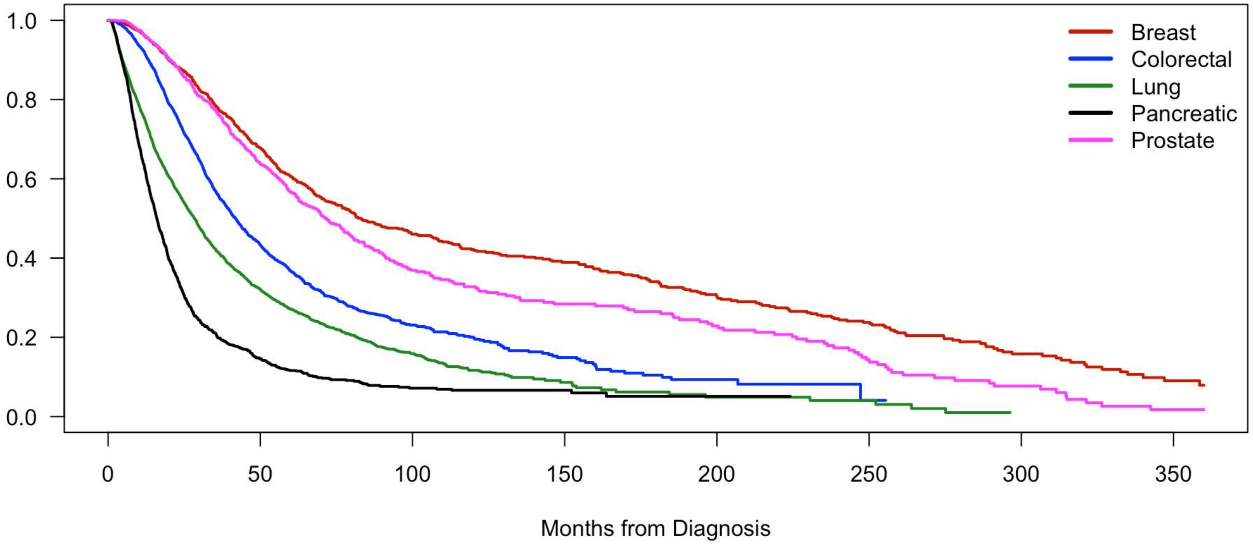
- About 41% of patients had metastasis at first diagnosis.

Possible Project Topics

- For prostate cancer, look at time from diagnosis to metastatic disease. What are the genomic characteristics of patients with a long time from initial diagnosis to development of metastasis.

OS from diagnosis (among de novo)

```
Call: survfit(formula = Surv(OS_FROM_DIAG, OS_STATUS_NUM) ~ CANCER_TYPE,
    data = ClinDiagSampMet)

                                          n events median 0.95LCL 0.95UCL
CANCER_TYPE=Breast Cancer              1311    657   83.0    76.8    98.8
CANCER_TYPE=Colorectal Cancer          2580   1529   41.9    39.9    43.9
CANCER_TYPE=Non-Small Cell Lung Cancer 3749   2623   28.4    26.8    29.6
CANCER_TYPE=Pancreatic Cancer          1424   1123   16.0    15.2    16.9
CANCER_TYPE=Prostate Cancer             984    488   71.7    64.9    79.4
```



Time from Treatment Start

- In some cases, you might want to compare outcomes after a particular type of treatment
- The start time for comparison should be the beginning of treatment
- The treatment initiation data is in the file `data_timeline_treatment.txt`

```r
Treatment <- read.delim("~/Downloads/msk_chord_2024/data_timeline_treatment.txt")
Treatment <- Treatment %>%
             arrange(PATIENT_ID, START_DATE) %>%
             select(PATIENT_ID, START_DATE, SUBTYPE, AGENT)
head(Treatment)
```

```
  PATIENT_ID START_DATE SUBTYPE            AGENT
1  P-0000012      -5437   Chemo CYCLOPHOSPHAMIDE
2  P-0000012      -5437   Chemo     FLUOROURACIL
3  P-0000012      -5437   Chemo     METHOTREXATE
4  P-0000012         33   Chemo        CISPLATIN
5  P-0000012         33   Chemo        ETOPOSIDE
6  P-0000012         61   Chemo       CARBOPLATIN
```

Time from Treatment Start

- You might be interested in a particular class of treatments

```r
table(Treatment$SUBTYPE)
```

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

- For example, **immunotherapies** are a more recent class of treatments

```r
ImmunoTrt <- subset(Treatment, SUBTYPE=="Immuno")
```

- All of the treatments classified as “Immuno” are really **immune checkpoint inhibitors**

```r
table(ImmunoTrt$AGENT)
```

```
 ATEZOLIZUMAB      AVELUMAB    CEMIPLIMAB    DURVALUMAB    IPILIMUMAB
          389             3             3           254           236
    NIVOLUMAB PEMBROLIZUMAB  TREMELIMUMAB
          886          2258             1
```

Time from Treatment Start

- They also classify certain treatments as “**Biologic**”
- Many of these could also be referred to as immunotherapies

```r
Biologic <- subset(Treatment, SUBTYPE=="Biologic")
table(Biologic$AGENT)[1:10]
```

```
ADO-TRASTUZUMAB EMTANSINE    AFLIBERCEPT OPHTHALMIC               ALDESLEUKIN 
                      352                        30                         1 
              ALEMTUZUMAB               AMIVANTAMAB                       BCG 
                        3                        16                       103 
     BELANTAMAB MAFODOTIN               BEVACIZUMAB              BLINATUMOMAB 
                        2                      3380                         1 
               BORTEZOMIB 
                       29 
```

Time from Treatment Start

- For the immune checkpoint inhibitors (ICI), let’s look at what the treatment data looks like:

```
head(ImmunoTrt, 8)
```

```
    PATIENT_ID START_DATE SUBTYPE        AGENT
9    P-0000012       1734  Immuno    NIVOLUMAB
171  P-0000113         85  Immuno   DURVALUMAB
355  P-0000165        449  Immuno    NIVOLUMAB
460  P-0000235        381  Immuno    NIVOLUMAB
466  P-0000239       1150  Immuno    NIVOLUMAB
574  P-0000302        929  Immuno    NIVOLUMAB
576  P-0000302       2178  Immuno   IPILIMUMAB
668  P-0000373        977  Immuno PEMBROLIZUMAB
```

- Most patients receive an only **one ICI**
    - But, about 17% (like P-0000302) received more than one ICI

```
dim( ImmunoTrt )
```

```
[1] 4030    4
```

```
length( unique(ImmunoTrt$PATIENT_ID) )
```

```
[1] 3341
```

Time from Treatment Start

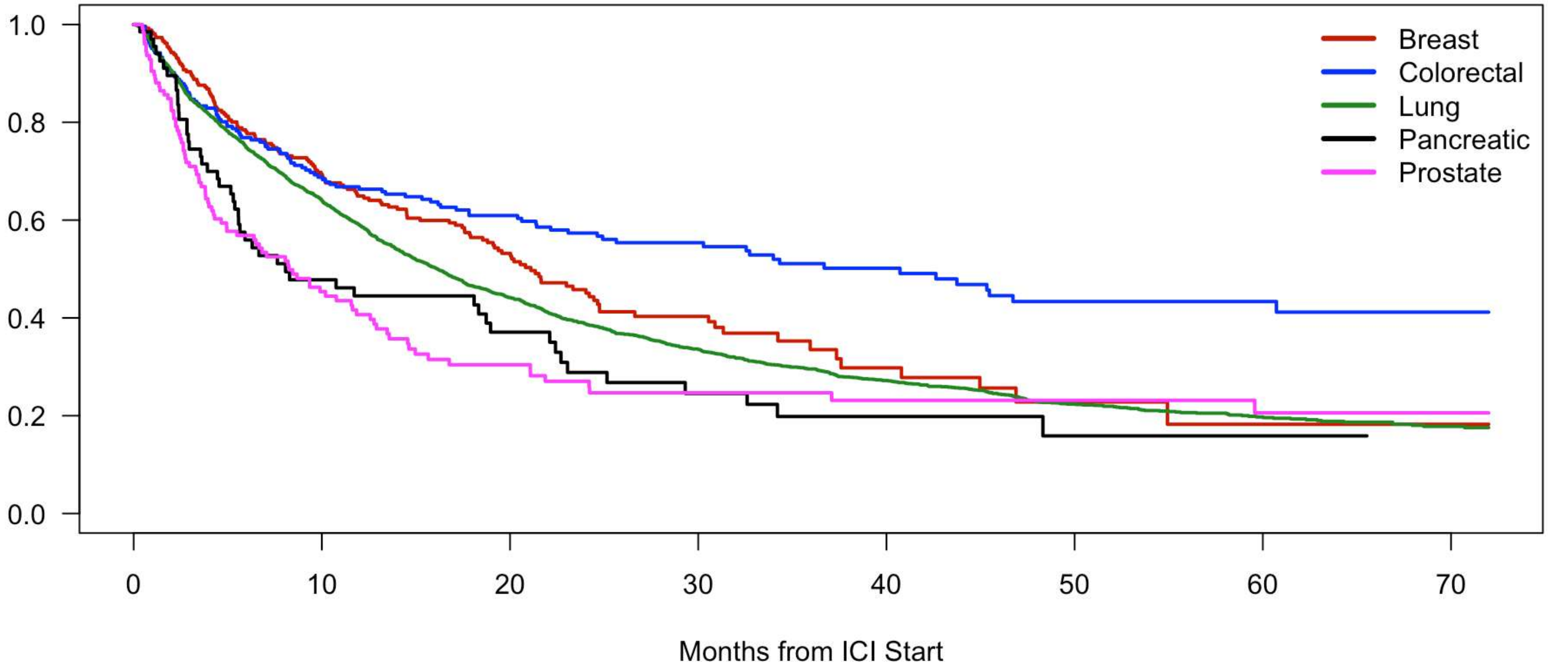
- For patients receiving **more than one ICI**, you would typically look at the time of **first ICI**:

```r
ImmunoTrtFirst <- ImmunoTrt[!duplicated(ImmunoTrt$PATIENT_ID),]
head(ImmunoTrtFirst, 8)
```

```
    PATIENT_ID START_DATE SUBTYPE         AGENT
9    P-0000012       1734  Immuno     NIVOLUMAB
171  P-0000113         85  Immuno    DURVALUMAB
355  P-0000165        449  Immuno     NIVOLUMAB
460  P-0000235        381  Immuno     NIVOLUMAB
466  P-0000239       1150  Immuno     NIVOLUMAB
574  P-0000302        929  Immuno     NIVOLUMAB
668  P-0000373        977  Immuno PEMBROLIZUMAB
953  P-0000449       1977  Immuno  ATEZOLIZUMAB
```

OS from Time of first ICI

- To look at OS from treatment start, you would need to:
  - Subtract time of ICI start from time of last follow up (`OS_MONTHS`)



Progression-free Survival

- **Progression-free survival (PFS)** is a common clinical endpoint used in many cancer studies.
- Goal: quantify how long a patient can live **without** having their cancer "progress".
- Outcome: time to death **OR** time to progression (whichever comes first)
- $Y_i = \min\{T_i, U_i\}$.
  - T_i - time of progression
  - U_i - time of death
- A median PFS of 5 years means that the median of Y_i is 5 years.

Progression-free Survival

- PFS is often measured from the start of a certain treatment.
    - For example, you might report median PFS from the time of starting an ICI.
    - PFS is also commonly measured from time of metastatic disease.

PFS in MSK-CHORD

- In MSK-CHORD, you are really going to be looking at a **“real-world”** measure of **radiographic PFS**.
- **“Real-world” PFS**: This is PFS where scans are not taken at regular, protocol-determined intervals.
  - Scans are taken whenever clinicians/patients think they are appropriate.
- **Radiographic PFS**: Progression is only determined by imaging-based criteria.
- Clinical/Radiographic PFS is used in many cancer studies.
  - Here, progression can include additional factors not captured by imaging.
  - We cannot capture clinical/radiographics PFS using the data in MSK-CHORD.

Radiographic PFS in MSK-CHORD

- To capture radiographic PFS in MSK-CHORD, you will need to use the `data_timeline_progression.txt` file.
    - Merging with the "OS from treatment start" data would give you:

```
    PATIENT_ID START_DATE_SCAN START_DATE_TRT NLP_PROGRESSION_PROBABILITY
156  P-0000239            1008           1150                 0.022726787
157  P-0000239            1106           1150                 0.044055530
158  P-0000239            1161           1150                 0.991534200
159  P-0000239            1204           1150                 0.998716100
160  P-0000302             -17            929                 0.005758766
161  P-0000302              47            929                 0.002286533
162  P-0000302             119            929                 0.676131840
    PROCEDURE_TYPE     AGENT OS_From_Trt OS_STATUS_NUM
156             CT NIVOLUMAB    2.162821             1
157             CT NIVOLUMAB    2.162821             1
158             CT NIVOLUMAB    2.162821             1
159             CT NIVOLUMAB    2.162821             1
160             CT NIVOLUMAB   75.275584             0
161             CT NIVOLUMAB   75.275584             0
162             CT NIVOLUMAB   75.275584             0
```

Radiographic PFS in MSK-CHORD

```
   PATIENT_ID START_DATE_SCAN START_DATE_TRT NLP_PROGRESSION_PROBABILITY
32  P-0000012            1702           1734                 0.002305086
33  P-0000012            1862           1734                 0.128094140
34  P-0000012            1974           1734                 0.015662400
35  P-0000012            2086           1734                 0.002995548
36  P-0000012            2269           1734                 0.088399045
37  P-0000012            2444           1734                 0.111479566
38  P-0000012            2445           1734                 0.001614616
39  P-0000012            2627           1734                 0.001661092
40  P-0000012            2809           1734                 0.003464612
   PROCEDURE_TYPE     AGENT OS_From_Trt OS_STATUS_NUM
32            PET NIVOLUMAB    61.48547             0
33             CT NIVOLUMAB    61.48547             0
34             CT NIVOLUMAB    61.48547             0
35             CT NIVOLUMAB    61.48547             0
36             CT NIVOLUMAB    61.48547             0
37             MR NIVOLUMAB    61.48547             0
38             CT NIVOLUMAB    61.48547             0
```

Time to Next Treatment

- Time to next treatment (TTNT) is another clincal endpoint that provides a measure of time to progression.

https://pmc.ncbi.nlm.nih.gov/articles/PMC8085844/

- Often used as another **"real-world"** measure of progression.
  - Most of the time, it can **reliably** be obtained from health records
- TTNT (usually) captures the period of time under which the disease is stable or improving.
  - Clinicians will not usually recommend the start of a new treatment if the current treatment is working well.
- TTNT can be thought of as a surrogate for **duration of clinical benefit**.

Time to Next Treatment Example

```
    PATIENT_ID START_DATE        SUBTYPE                    AGENT
364  P-0000175      -1422        Hormone               LEUPROLIDE
365  P-0000175      -1366 Bone Treatment          ZOLEDRONIC ACID
366  P-0000175       -199        Hormone               EXEMESTANE
367  P-0000175       -129       Targeted               EVEROLIMUS
368  P-0000175        420        Hormone               EXEMESTANE
369  P-0000175        537          Chemo             CAPECITABINE
370  P-0000175        707        Hormone              FULVESTRANT
371  P-0000175        707       Targeted              PALBOCICLIB
372  P-0000175        832          Chemo               PACLITAXEL
373  P-0000175        860          Chemo PACLITAXEL PROTEIN-BOUND
374  P-0000175       1064        Hormone                TAMOXIFEN
```

- For example, starting from the combination therapy at date 707, the TTNT would be defined as 832 - 707 = 125 days.