AACR GENIE clinical and sample data summary report

Larson Hogstrom

2024-01-29

Overview

The goal of this document is to analyze patient and sample information from the AACR GENIE data set. A number of fields are available for each patient including high-level demographic information, academic center, contact date, and death status.

Loading data

```
bDir <- "../../data/processed/balderResultsDb"
figDir <- "../../output/actionability_db_curration_20231220"
#mydb <- DBI::dbConnect(RSQLite::SQLite(), paste0(bDir,"/actionable-biomarker-db.sqlite"))
mydb <- DBI::dbConnect(RSQLite::SQLite(), paste0(bDir,"/balder-compiled-raw-data-v20240311

#genie <- RSQLite::dbGetQuery(mydb, 'SELECT * FROM GeniePatientVarients')
sample.genie <- RSQLite::dbGetQuery(mydb, 'SELECT * FROM GenieClinicalSampleData')
patient.genie <- RSQLite::dbGetQuery(mydb, 'SELECT * FROM GeniePatientData')
```

Description of fields

The following fields are contained in the patient data file:

1. **Patient Identifier (PATIENT_ID)**: A unique identifier for each patient, following the convention GENIE-CENTER-XXXX.
2. **Sex (SEX)**: The sex of the patient.

3. **Primary Race (PRIMARY_RACE)**: The primary race recorded for the patient.
4. **Ethnicity (ETHNICITY)**: The ethnicity of the patient, particularly indicating if they are Spanish/Hispanic or not.
5. **Center (CENTER)**: The center of sequencing.
6. **Interval in days from DOB to date of last contact (INT_CONTACT)**: The number of days from the patient's date of birth to their last contact date.
7. **Interval in days from DOB to date of death (INT_DOD)**: The number of days from the patient's date of birth to their date of death (if applicable).
8. **Year of last contact (YEAR_CONTACT)**: The year in which the patient was last known to be alive.
9. **Vital Status (DEAD)**: Indicates whether the patient is deceased.
10. **Year of Death (YEAR_DEATH)**: The year in which the patient died (if applicable).

Null field summary

Summary of columns and proportion of null entries

	Count of null entries	Proportion of null
PATIENT_ID	0	0%
SEX	629	0.39%
PRIMARY_RACE	20291	12.61%
ETHNICITY	42306	26.28%
CENTER	0	0%
INT_CONTACT	14367	8.93%
INT_DOD	12483	7.76%
YEAR_CONTACT	14101	8.76%
DEAD	11672	7.25%
YEAR_DEATH	12451	7.74%

Defining an inferred duration of treatment

```
# define infered treatment duration

#### death
iNumericDoD <- as.numeric(patient.genie$INT_DOD)
```

```
Warning: NAs introduced by coercion
```

```
  isNullDoD <- is.na(iNumericDoD)
  # which proportion don't have a proper date for DoD
  table(isNullDoD)
```

```
isNullDoD
 FALSE   TRUE
 54747 106218
```

```
  # identity of non-numeric entries
  table(patient.genie$INT_DOD[isNullDoD])
```

```
                     <6570          >32485 Not Applicable  Not Collected
         3852         4615            1587          83681           1092
  Not Released      Unknown
          465         10926
```

```
  # all of the not dead and Unknown death samples have non-null death entry
  #table(isNullDoD,patient.genie$DEAD_status)

  ### contact days
  iNumericContact <- as.numeric(patient.genie$INT_CONTACT)
```

```
Warning: NAs introduced by coercion
```

```
  isNullContact <- is.na(iNumericContact)
  table(isNullContact)
```

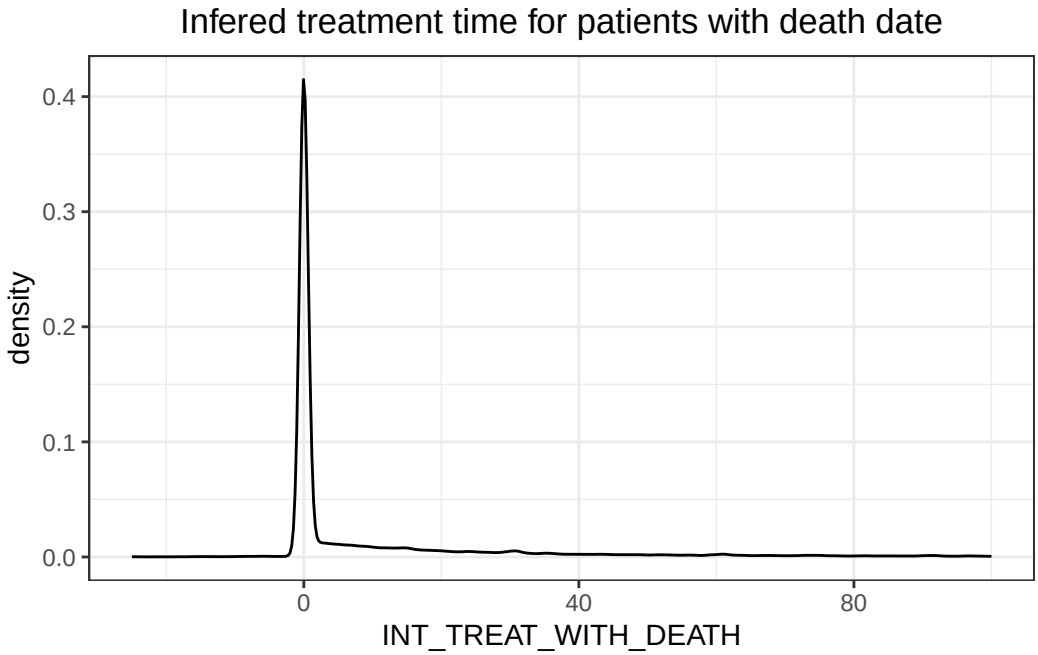
```
isNullContact
 FALSE   TRUE
132791  28174
```

```
  # non-null death or contact days
  iNullDoDAndContact <- isNullContact | isNullDoD
```

```
patient.genie[!iNullDoDAndContact,"INT_TREAT_WITH_DEATH"] <- iNumericDoD[!iNullDoDAndConta
```

```
#outF <-  paste0(outDir,"/cancer_type_clinical_annotation_match_summary.pdf")
ggplot(patient.genie,aes(x=INT_TREAT_WITH_DEATH))+
  geom_density()+
  theme_bw()+
  xlim(-25,100)+
  ggtitle(paste0("Infered treatment time for patients with death date"))+
  theme(plot.title = element_text(hjust = 0.5))
```

```
Warning: Removed 113315 rows containing non-finite values (`stat_density()`).
```




```
#ggsave(outF,height = 8,width = 7)
```

```
# define infered treatment duration (years)
iNumericDoD <- as.numeric(patient.genie$YEAR_DEATH)
```

```
Warning: NAs introduced by coercion
```

```r
isNullDoD <- is.na(iNumericDoD)
# which proportion don't have a proper date for DoD
table(isNullDoD)
```

```
isNullDoD
 FALSE   TRUE
 59067 101898
```

```r
# identity of non-numeric entries
table(patient.genie$YEAR_DEATH[isNullDoD])
```

```
                      <18            >89 Not Applicable  Not Collected
         3852           546          1368          83681           1092
  Not Released       Unknown
          465         10894
```

```r
# all of the not dead and Unknown death samples have non-null death entry
#table(isNullDoD,patient.genie$DEAD_status)

### contact days
iNumericContact <- as.numeric(patient.genie$YEAR_CONTACT)
```

```
Warning: NAs introduced by coercion
```

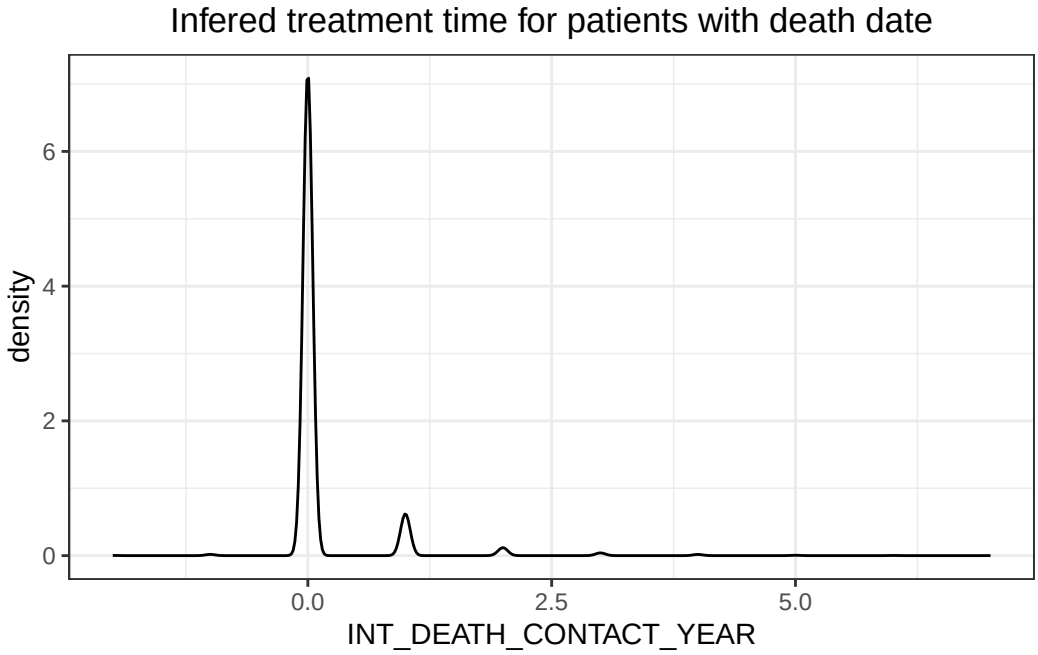
```r
isNullContact <- is.na(iNumericContact)
table(isNullContact)
```

```
isNullContact
 FALSE   TRUE
138631  22334
```

```r
# non-null death or contact days
iNullDoDAndContact <- isNullContact | isNullDoD
patient.genie[!iNullDoDAndContact,"INT_DEATH_CONTACT_YEAR"] <- iNumericDoD[!iNullDoDAndCon
```

```r
ggplot(patient.genie,aes(x=INT_DEATH_CONTACT_YEAR))+
  geom_density()+
  theme_bw()+
  xlim(-2,7)+
  ggtitle(paste0("Infered treatment time for patients with death date"))+
  theme(plot.title = element_text(hjust = 0.5))
```

```
Warning: Removed 103685 rows containing non-finite values (`stat_density()`).
```



Python exploratory analysis

Print data from python

```r
patient_py = reticulate::r_to_py(patient.genie)
```

Summary of pandas fields

Summary of categorical fields

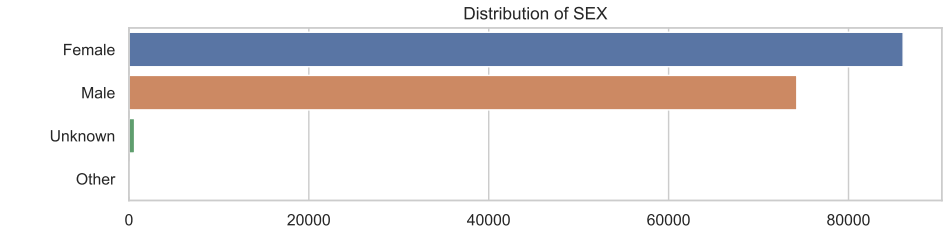
```
knitr::kable(py$categorical_summary,format="latex", align="l")
```

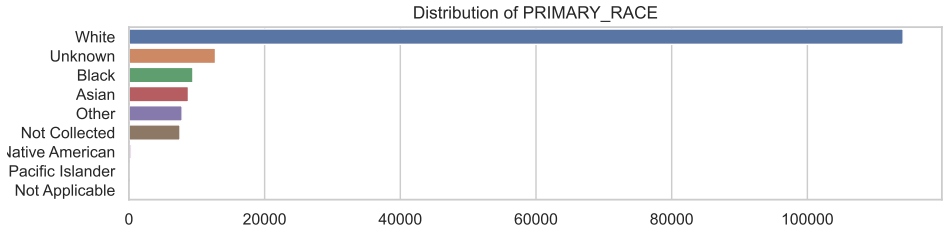
	PATIENT_ID	SEX	PRIMARY_RACE	ETHNICITY	CENTER	IN
count	160965	160965	160965	160965	160965	160
unique	160965	4	9	4	19	228
top	GENIE-VICC-101416	Female	White	Non-Spanish/non-Hispanic	MSK	Un
freq	1	86078	114120	109341	65577	112

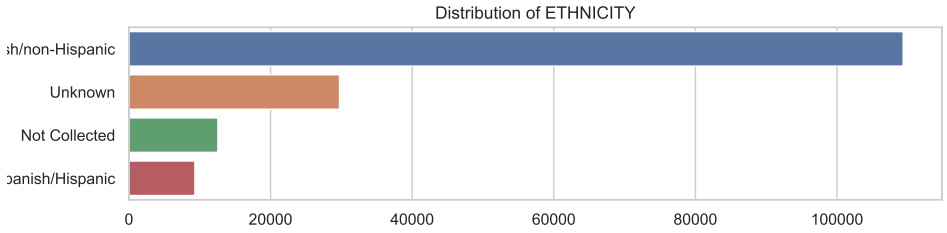
```
categorical_summary = data.describe(include=['object'])
categorical_summary
```

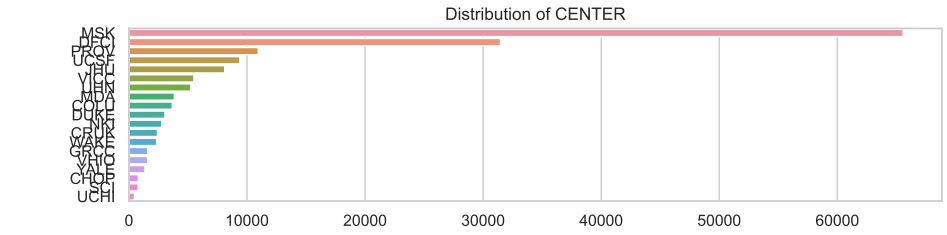
```
              PATIENT_ID     SEX  ...    DEAD      YEAR_DEATH
count             160965  160965  ...  160965          160965
unique            160965       4  ...       8              54
top     GENIE-VICC-101416  Female  ...   False  Not Applicable
freq                   1   86078  ...   76021           83681

[4 rows x 10 columns]
```

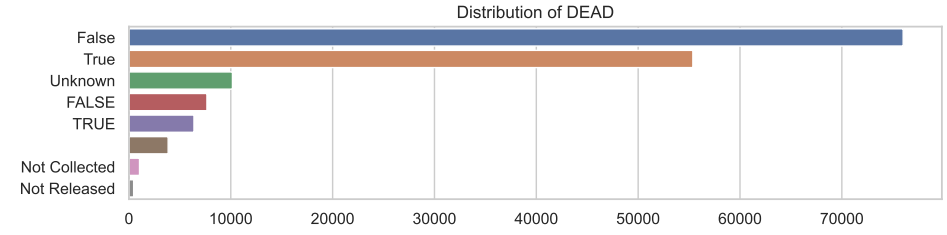
Plotting of categorical fields (from Chat GPT)











Review of survival data

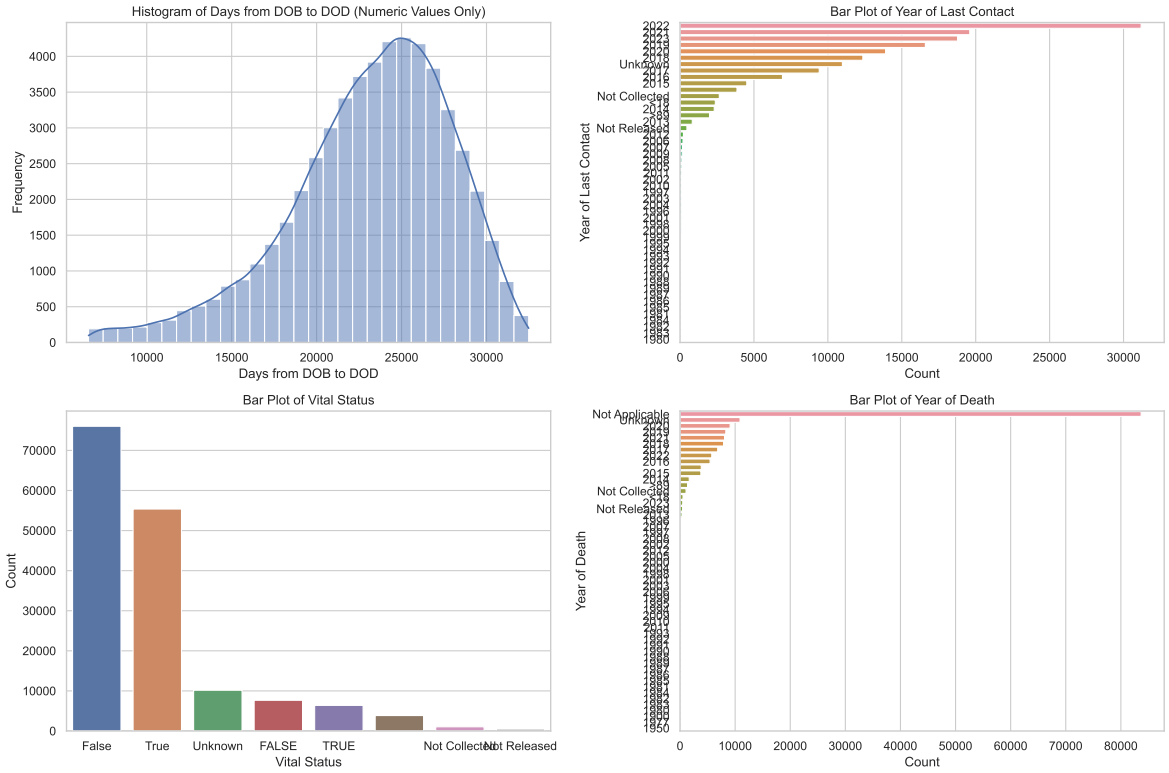
Count tble of year of patient death

```
table(as.numeric(patient.genie$YEAR_DEATH),exclude=NULL)
```

```
Warning in table(as.numeric(patient.genie$YEAR_DEATH), exclude = NULL): NAs
introduced by coercion
```

```
  1900   1950   1977   1980   1981   1982   1983   1984   1985   1986   1987
     1      1      1      1      4      3      2      3      6      6      7
  1988   1989   1990   1991   1992   1993   1994   1995   1996   1997   1998
     9      8     14     16     33     36     51     57     79     77     64
  1999   2000   2001   2002   2003   2004   2005   2006   2007   2008   2009
    59     67     62     72     62     64     68     62     78     73     50
  2010   2011   2012   2013   2014   2015   2016   2017   2018   2019   2020
    44     36     72    352   1669   3781   5453   6835   7901   8310   9092
  2021   2022   2023   <NA>
  8075   5758    493 101898
```

Plots of DOB, Year of last contact, and survival

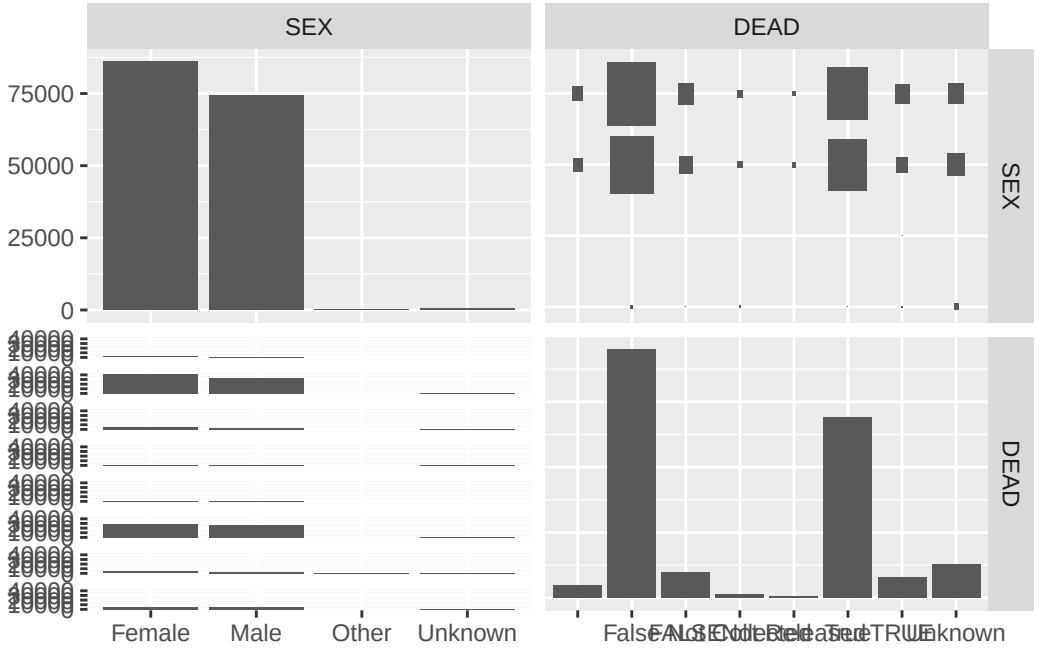



```
#dod_summ = reticulate::py_to_r(int_dod_summary)
dod_summ = py$int_dod_summary
knitr::kable(head(dod_summ,10), )
```

	x
count	54747.000
mean	23170.952
std	4774.158
min	6573.000
25%	20445.500
50%	23809.000
75%	26632.000
max	32482.000

Review of patient data pairwise

```
ggpairs(patient.genie[,c("SEX","DEAD")]) #"PRIMARY_RACE",
```



Binary Death assignments

```
{r}
#| echo: true

patient.genie$DEAD_status <- patient.genie$DEAD
patient.genie[is.na(patient.genie$DEAD),"DEAD_status"] <- "Unknown"
patient.genie[patient.genie$DEAD=="","DEAD_status"] <- "Unknown"
patient.genie[patient.genie$DEAD=="Not Collected","DEAD_status"] <- "Unknown"
patient.genie[patient.genie$DEAD=="Not Released","DEAD_status"] <- "Unknown"
patient.genie[patient.genie$DEAD=="True","DEAD_status"] <- "TRUE"
patient.genie[patient.genie$DEAD=="False","DEAD_status"] <- "FALSE"
print(table(patient.genie$DEAD_status))

table(patient.genie$YEAR_DEATH)
```