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"))

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

Exploratory data analysis

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%

Print data from python

```
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]
```
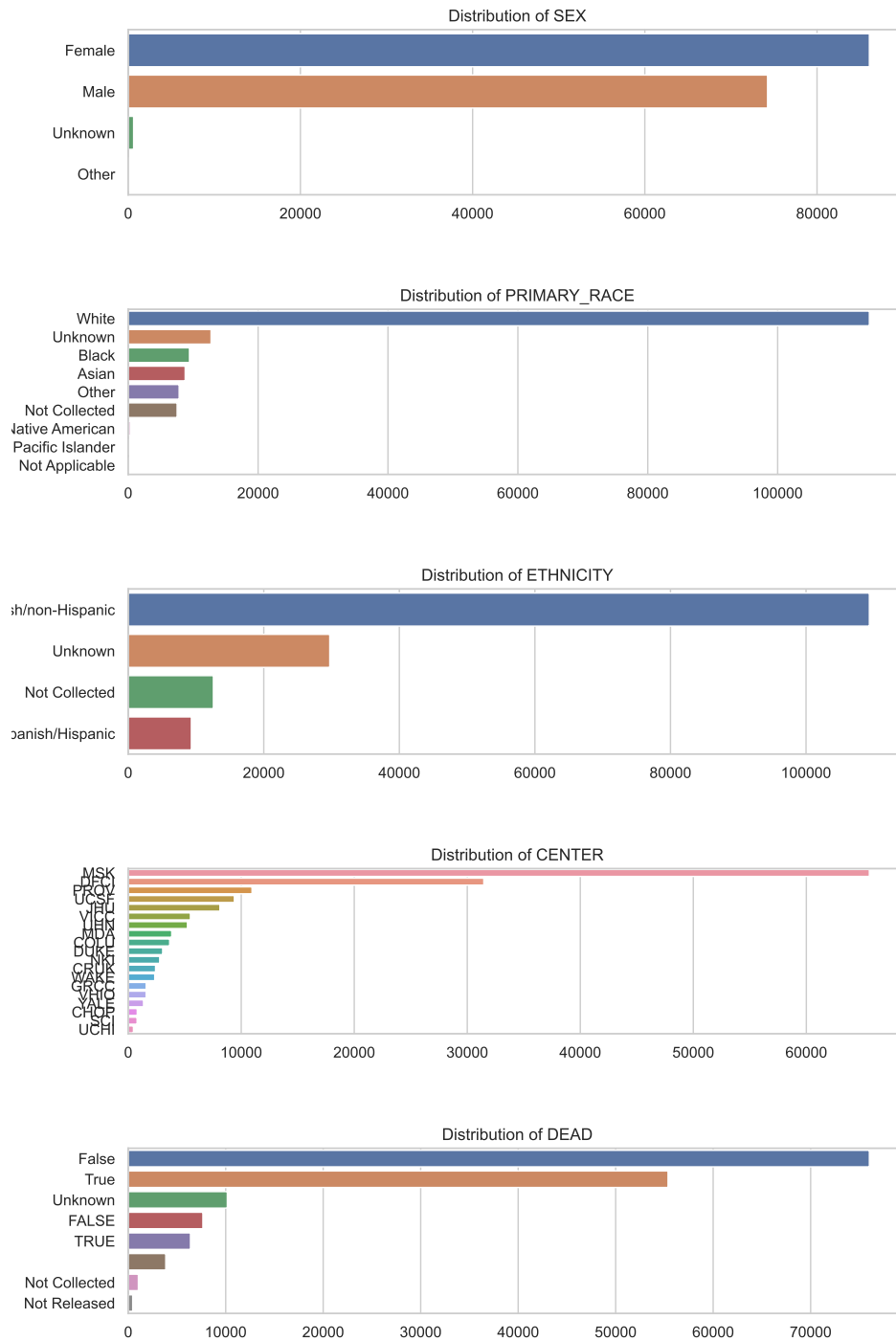
Plotting of categorical fields (from Chat GPT)

Distribution of SEX

Distribution of PRIMARY_RACE

Distribution of ETHNICITY

Distribution of CENTER

Distribution of DEAD

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
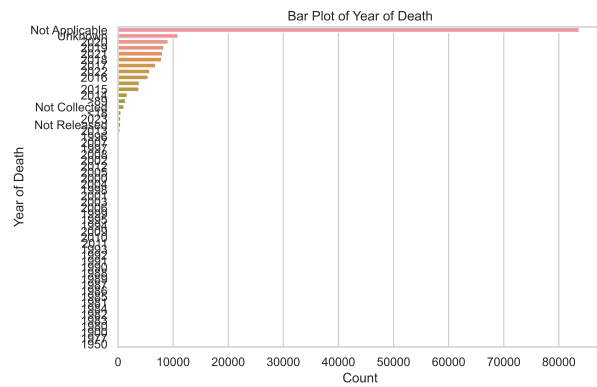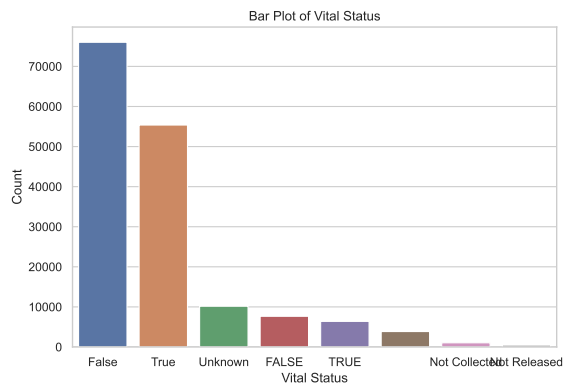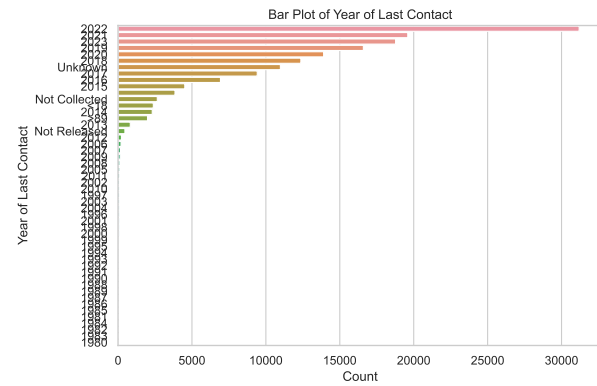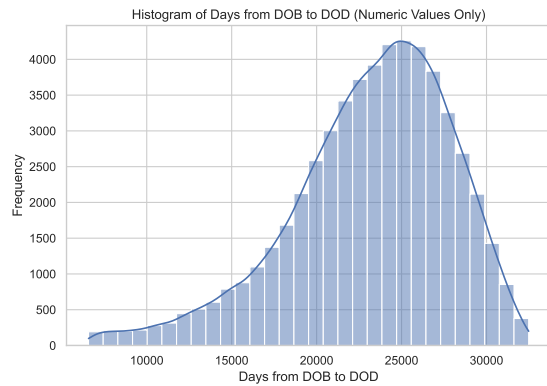

```r
#dod_summ = reticulate::py_to_r(int_dod_summary)
dod_summ = py$int_dod_summary
knitr::kable(head(dod_summ,10), )
```

	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",
```