

Jhet: Cleaning

Jhet Cabigas

2024-02-08

Github Repo: <https://github.com/Smoorad99/485-TT>

Research Questions:

The questions we aim to solve with our data are the following:

- How does race affect survival rate/months survived for testi cancer?
 - Exploring the role race plays in testicular cancer survival rates could help us address disparities between different races. If we find that survival rate of testicular cancer is impacted by race, it would push us to explore why this may be. For example, it may indicate inequality in healthcare received by different races.
- How does socioeconomic standing affect survival rate/months survived for testi cancer?
 - Socioeconomic status can impact the quality of healthcare an individual has (among other things), which we expect to impact testicular cancer survival rate. Exploring this could help us address if there is a significant relationship between socioeconomic status and testicular cancer survival rate. Suppose the difference is significant and the magnitude is great. In that case, it may suggest that a policy change could be beneficial (i.e., allocate more taxpayer money to healthcare for individuals in lower socioeconomic brackets).
- How does survival rate/months change based on treatment options for testi cancer?
 - Exploring the relationship between survival rate/months and treatment method helps us understand which treatment methods are most effective. Investigating the quality of life patients experience while undergoing different treatments may also help us better understand the effectiveness of each treatment.
- Has the survival rate of testi cancer increased/decreased over time?
 - Knowing the direction in which the survival rate moves allows us to question why it moves in said direction. If the survival rate has increased over time, we may ask: Has better medicine/treatment options led to this increase in survival rate?

Jhet: Data retrieval, Cols 1-11

Retrieval:

Using SEERStat, we retrieved a text file containing x observations for the following variables:

- race

- sex
- year of diagnosis
- primary site
- RX Summ (all 6)
- Reason no cancer-directed surgery
- Radiation recode
- Chemotherapy recode
- Scope of reg lymph nd surg
- Surgery of other reg/dis sites
- Site specific surgery
- Radiation to brain or cns recode?
- Months from diagnosis to treatment
- AFP Post-orchiectomy lab value recode?
- hCG post-orchiectomy range recode
- LDH post-orchiectomy range recode
- Number of Examined Pelvic Nodes
- Number of Positive Para-Aortic Nodes
- Number of Positive Pelvic Nodes
- Peritoneal Cytology
- Lymph-vascular Invasion
- CS tumor size (2004-2015)
- Regional nodes examined
- Regional nodes positive
- CS lymph nodes
- CS mets at dx
- CS Tumor Size/Ext Eval
- CS Reg Node Eval
- Tumor marker 1
- Tumor marker 2
- Tumor marker 3
- Survival months
- Survival months flags
- First malignant primary indicator

- Total number of in situ/malignant tumors for patient
- Total number of benign/borderline tumors for patient
- Race/ethnicity
- Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)
- Age recode with <1 year olds and 90+
- Age recode with single ages and 85+
- Age recode with single ages and 90+
- Race recode (W, B, AI, API)
- Marital status at diagnosis

Data cleaning:

The variables are relatively clean, and only require some recoding as factor or numeric variables instead of strings. Beyond that, NA's are recorded as "Blank(s)", which needed to be changed.

```
subset <- seer %>% select(1:11)

subset[subset == "Blank(s)"] <- NA
subset <- subset %>% mutate(
  Sex = as.factor(Sex),
  `Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)` = as.factor(`Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)`),
  `Age recode with <1 year olds` = as.factor(`Age recode with <1 year olds`),
  `RX Summ--Scope Reg LN Sur (2003+)` = as.factor(`RX Summ--Scope Reg LN Sur (2003+)`),
  `RX Summ--Surg Oth Reg/Dis (2003+)` = as.factor(`RX Summ--Surg Oth Reg/Dis (2003+)`),
  `RX Summ--Surg/Rad Seq` = as.factor(`RX Summ--Surg/Rad Seq`),
  `RX Summ--Reg LN Examined (1998-2002)` = as.numeric(`RX Summ--Reg LN Examined (1998-2002)`),
  `RX Summ--Systemic/Sur Seq (2007+)` = as.factor(`RX Summ--Systemic/Sur Seq (2007+)`),
  `Months from diagnosis to treatment` = as.numeric(`Months from diagnosis to treatment`)
)
```

Chase: Cols 12-22

Used tables and frequency tables to analyze the columns. One variable was factored, every entry is listed as "Blank(s)" in Radiation to Brain or CNS Recode (1988-1997) and Site specific surgery (1973-1997 varying detail by year and site).

```
names(seer)
```

```
## [1] "Sex"
## [2] "Year of diagnosis"
## [3] "Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)"
## [4] "Age recode with <1 year olds"
## [5] "RX Summ--Surg Prim Site (1998+)"
## [6] "RX Summ--Scope Reg LN Sur (2003+)"
## [7] "RX Summ--Surg Oth Reg/Dis (2003+)"
## [8] "RX Summ--Surg/Rad Seq"
## [9] "RX Summ--Reg LN Examined (1998-2002)"
```

```

## [10] "RX Summ--Systemic/Sur Seq (2007+)"
## [11] "Months from diagnosis to treatment"
## [12] "Radiation to Brain or CNS Recode (1988-1997)"
## [13] "Site specific surgery (1973-1997 varying detail by year and site)"
## [14] "Surgery of oth reg/dis sites (1998-2002)"
## [15] "Scope of reg lymph nd surg (1998-2002)"
## [16] "Chemotherapy recode (yes, no/unk)"
## [17] "Radiation recode"
## [18] "Reason no cancer-directed surgery"
## [19] "Primary Site"
## [20] "AFP Post-Orchiectomy Lab Value Recode (2010+)"
## [21] "hCG Post-Orchiectomy Range Recode (2010+)"
## [22] "LDH Post-Orchiectomy Range Recode (2010+)"
## [23] "Number of Examined Pelvic Nodes Recode (2010+)"
## [24] "Number of Positive Para-Aortic Nodes Recode (2010+)"
## [25] "Number of Positive Pelvic Nodes Recode (2010+)"
## [26] "Peritoneal Cytology Recode (2010+)"
## [27] "Lymph-vascular Invasion (2004+ varying by schema)"
## [28] "CS tumor size (2004-2015)"
## [29] "Regional nodes examined (1988+)"
## [30] "Regional nodes positive (1988+)"
## [31] "CS lymph nodes (2004-2015)"
## [32] "CS mets at dx (2004-2015)"
## [33] "CS Tumor Size/Ext Eval (2004-2015)"
## [34] "CS Reg Node Eval (2004-2015)"
## [35] "Tumor marker 1 (1990-2003)"
## [36] "Tumor marker 2 (1990-2003)"
## [37] "Tumor marker 3 (1998-2003)"
## [38] "Survival months"
## [39] "Survival months flag"
## [40] "Total number of in situ/malignant tumors for patient"
## [41] "Total number of benign/borderline tumors for patient"
## [42] "Race/ethnicity"
## [43] "Age recode with <1 year olds and 90+"
## [44] "Age recode with single ages and 85+"
## [45] "Age recode with single ages and 90+"
## [46] "Race recode (W, B, AI, API)"
## [47] "Marital status at diagnosis"

```

```
table(seer$`AFP Post-Orchiectomy Lab Value Recode (2010+)`)
```

```

##
##           0 nanograms/milliliter (ng/ml)
##                               249
##           1 - 19 ng/ml
##                               8676
##          100 - 199 ng/ml
##                               391
##         1000 - 1999 ng/ml
##                               134
##           20 - 29 ng/ml
##                               440
##          200 - 299 ng/ml
##                               174

```

##	2000 - 2999 ng/ml
##	73
##	30 - 39 ng/ml
##	260
##	300 - 399 ng/ml
##	105
##	3000 - 3999 ng/ml
##	36
##	40 - 49 ng/ml
##	174
##	400 - 499 ng/ml
##	82
##	4000 - 4999 ng/ml
##	26
##	50 - 59 ng/ml
##	120
##	500 - 599 ng/ml
##	42
##	5000 - 5999 ng/ml
##	11
##	60 - 69 ng/ml
##	84
##	600 - 699 ng/ml
##	29
##	6000 - 6999 ng/ml
##	18
##	70 - 79 ng/ml
##	78
##	700 - 799 ng/ml
##	36
##	7000 - 7999 ng/ml
##	12
##	80 - 89 ng/ml
##	62
##	800 - 899 ng/ml
##	28
##	8000 - 8999 ng/ml
##	16
##	90 - 99 ng/ml
##	56
##	900 - 999 ng/ml
##	34
##	9000 - 9999 ng/ml
##	5
##	Blank(s)
##	8189415
##	Greater than or equal to 10, 000 ng/ml
##	124
##	Not applicable; Information not collected for this case
##	4805
##	Test ordered, results not in chart
##	476
##	Unknown or no information; Test not done
##	9898

```
factor(seer$hCG Post-Orchiectomy Range Recode (2010+), labels=c("5,000-50,000", "Above normal and less
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```

## [99769] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99777] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99785] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99793] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99801] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99809] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99817] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99825] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99833] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99841] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99849] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99857] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99865] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99873] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99881] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99889] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99897] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99905] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99913] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99921] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99929] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99937] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99945] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99953] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99961] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99969] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99977] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99985] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [99993] Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s) Blank(s)
## [ reached getOption("max.print") -- omitted 8116170 entries ]
## 5 Levels: 5,000-50,000 Above normal and less than 5,000 MIU/mL ... normal

```

```
table(seer$`AFP Post-Orchiectomy Lab Value Recode (2010+)`)
```

```

##
##           0 nanograms/milliliter (ng/ml)
##                               249
##           1 - 19 ng/ml
##                               8676
##          100 - 199 ng/ml
##                               391
##         1000 - 1999 ng/ml
##                               134
##           20 - 29 ng/ml
##                               440
##          200 - 299 ng/ml
##                               174
##        2000 - 2999 ng/ml
##                               73
##           30 - 39 ng/ml
##                               260
##          300 - 399 ng/ml
##                               105
##        3000 - 3999 ng/ml

```

```

##                                     36
##                                40 - 49 ng/ml
##                                     174
##                                400 - 499 ng/ml
##                                     82
##                                4000 - 4999 ng/ml
##                                     26
##                                50 - 59 ng/ml
##                                     120
##                                500 - 599 ng/ml
##                                     42
##                                5000 - 5999 ng/ml
##                                     11
##                                60 - 69 ng/ml
##                                     84
##                                600 - 699 ng/ml
##                                     29
##                                6000 - 6999 ng/ml
##                                     18
##                                70 - 79 ng/ml
##                                     78
##                                700 - 799 ng/ml
##                                     36
##                                7000 - 7999 ng/ml
##                                     12
##                                80 - 89 ng/ml
##                                     62
##                                800 - 899 ng/ml
##                                     28
##                                8000 - 8999 ng/ml
##                                     16
##                                90 - 99 ng/ml
##                                     56
##                                900 - 999 ng/ml
##                                     34
##                                9000 - 9999 ng/ml
##                                     5
##                                Blank(s)
##                                8189415
##                                Greater than or equal to 10, 000 ng/ml
##                                124
## Not applicable; Information not collected for this case
##                                4805
##                                Test ordered, results not in chart
##                                476
##                                Unknown or no information; Test not done
##                                9898

```

```
table(seer$`Surgery of oth reg/dis sites (1998-2002)`)
```

```

##
##      0      1      2      3      4      5      6      7
## 972575 6244 7943 2661 6228 1225 3299 301
##      8      9 Blank(s)

```

```
##          22      26906  7188765
```

```
table(seer$`Scope of reg lymph nd surg (1998-2002)`)
```

```
##
##          0          1          2          3          4          5          6          9
##  557808   180164   18902   8244   10020   2734   403   249129
## Blank(s)
##  7188765
```

```
table(seer$`Chemotherapy recode (yes, no/unk)`)
```

```
##
## No/Unknown      Yes
##   5818277   2397892
```

```
table(seer$`Radiation recode`)
```

```
##
##                               Beam radiation
##                               1855978
##      Combination of beam with implants or isotopes
##                               80566
##                               None/Unknown
##                               5878338
##      Radiation, NOS  method or source not specified
##                               26307
## Radioactive implants (includes brachytherapy) (1988+)
##                               126320
##                               Radioisotopes (1988+)
##                               95513
##      Recommended, unknown if administered
##                               74566
##                               Refused (1988+)
##                               78581
```

```
table(seer$`Reason no cancer-directed surgery`)
```

```
##
##      Not performed, patient died prior to recommended surgery
##                               8437
##      Not recommended
##                               3076494
## Not recommended, contraindicated due to other cond; autopsy only (1973-2002)
##                               121124
##      Recommended but not performed, patient refused
##                               84228
##      Recommended but not performed, unknown reason
##                               200104
##      Recommended, unknown if performed
##                               31741
```

```
##
##
##
##
```

	Surgery performed
	4546974
Unknown; death certificate; or autopsy only (2003+)	
	147067

```
unique(seer$`Primary Site`)
```

```
## [1] 629 209 381 508 220 749 445 19 210 670 343 659 778 649 619 250 252 49
## [19] 188 509 187 22 673 421 180 99 541 505 504 446 447 569 341 444 679 443
## [37] 672 340 502 30 678 770 609 420 155 31 503 493 449 809 739 182 480 186
## [55] 539 441 442 621 674 79 179 185 259 529 349 501 713 411 183 676 419 1
## [73] 542 379 519 482 211 711 384 779 160 712 669 716 320 154 163 184 189 675
## [91] 23 168 170 171 251 218 380 221 690 422 239 718 671 109 158 119 29 40
## [109] 153 719 269 9 253 90 165 241 771 169 0 140 494 760 518 240 342 772
## [127] 258 181 773 696 602 348 499 51 249 680 321 162 511 570 161 329 491 413
## [145] 69 199 490 60 495 129 695 112 693 559 311 496 694 754 151 159 500 715
## [163] 62 39 481 300 310 91 414 319 510 530 689 172 257 150 139 448 412 717
## [181] 506 164 80 774 50 166 28 402 543 301 58 440 700 512 762 108 470 474
## [199] 132 710 492 579 763 601 103 322 328 4 21 765 52 714 41 152 89 131
## [217] 3 608 400 383 688 775 312 571 699 138 268 61 753 677 59 403 720 600
## [235] 323 538 260 2 110 173 578 130 100 318 549 212 631 632 48 410 740 339
## [253] 6 692 248 111 81 409 148 473 498 68 475 102 98 620 750 142 729 20
## [271] 531 577 401 8 254 751 424 113 24 178 723 5 313 589 698 741 721 479
## [289] 709 382 548 118 637 471 725 728 101 764 759 691 701 752 488 104 755 761
## [307] 476 88 472 574 630 423 540 418 573 767 388 408 724 638 639 768 399 398
## [325] 572 390 681 478 722 758
```

Frankie: Cols 23-32

Going through variables 23-32 from our original data set to verify quality in data. Making sure to remove columns that have too many missing variables. Removed columns 23-26 because they contained too many blanks.

Saul: Cols 33-47

I checked variables for suspicious values, changed variables from strings to numeric (when appropriate), and renamed variables for ease of use. I then dropped a few variables that did not add any useful information.

```
a <- seer[, 33:47] # Subsetting data
b <- a # For reference
rm(seer)

colnames(a) <- gsub(" ", ".", colnames(a)) # replace spaces with periods
colnames(b) <- gsub(" ", ".", colnames(b))

# Three age variables... which do we want
table(a$`Age.recode.with.<1.year-olds.and.90+`) # looks clean
```

```
##
## 00 years 01-04 years 05-09 years 10-14 years 15-19 years 20-24 years
## 5654 20254 14755 17137 27444 44140
```



```
## 25-29 years 30-34 years 35-39 years 40-44 years 45-49 years 50-54 years
##      70581      109493      165149      266308      423858      643732
## 55-59 years 60-64 years 65-69 years 70-74 years 75-79 years 80-84 years
##      869581      1041522      1147748      1070188      930759      707451
## 85-89 years 90+ years
##      428942      211473
```

```
table(a$`Age.recode.with.single.ages.and.85+`) # if we decide we want this remove " years" and convert
```

```
##
## 00 years 01 years 02 years 03 years 04 years 05 years 06 years 07 years
##      5654      5229      5570      5114      4341      3417      3133      2831
## 08 years 09 years 10 years 11 years 12 years 13 years 14 years 15 years
##      2654      2720      2847      3072      3344      3719      4155      4765
## 16 years 17 years 18 years 19 years 20 years 21 years 22 years 23 years
##      4955      5646      5780      6298      7221      7899      8867      9557
## 24 years 25 years 26 years 27 years 28 years 29 years 30 years 31 years
##      10596      11784      12644      14068      15249      16836      18613      20140
## 32 years 33 years 34 years 35 years 36 years 37 years 38 years 39 years
##      21869      23563      25308      27845      30004      32739      35415      39146
## 40 years 41 years 42 years 43 years 44 years 45 years 46 years 47 years
##      45096      48138      52609      56945      63520      70593      76882      84428
## 48 years 49 years 50 years 51 years 52 years 53 years 54 years 55 years
##      91745      100210      114538      119922      126394      136600      146278      155772
## 56 years 57 years 58 years 59 years 60 years 61 years 62 years 63 years
##      164815      174238      183245      191511      198087      204390      210326      214111
## 64 years 65 years 66 years 67 years 68 years 69 years 70 years 71 years
##      214608      237351      230015      228172      227433      224777      220617      217912
## 72 years 73 years 74 years 75 years 76 years 77 years 78 years 79 years
##      215276      211297      205086      199244      194695      187163      179204      170453
## 80 years 81 years 82 years 83 years 84 years 85+ years
##      158613      152006      142938      132634      121260      640415
```

```
table(a$`Age.recode.with.single.ages.and.90+`) # same as above
```

```
##
## 00 years 01 years 02 years 03 years 04 years 05 years 06 years 07 years
##      5654      5229      5570      5114      4341      3417      3133      2831
## 08 years 09 years 10 years 11 years 12 years 13 years 14 years 15 years
##      2654      2720      2847      3072      3344      3719      4155      4765
## 16 years 17 years 18 years 19 years 20 years 21 years 22 years 23 years
##      4955      5646      5780      6298      7221      7899      8867      9557
## 24 years 25 years 26 years 27 years 28 years 29 years 30 years 31 years
##      10596      11784      12644      14068      15249      16836      18613      20140
## 32 years 33 years 34 years 35 years 36 years 37 years 38 years 39 years
##      21869      23563      25308      27845      30004      32739      35415      39146
## 40 years 41 years 42 years 43 years 44 years 45 years 46 years 47 years
##      45096      48138      52609      56945      63520      70593      76882      84428
## 48 years 49 years 50 years 51 years 52 years 53 years 54 years 55 years
##      91745      100210      114538      119922      126394      136600      146278      155772
## 56 years 57 years 58 years 59 years 60 years 61 years 62 years 63 years
##      164815      174238      183245      191511      198087      204390      210326      214111
## 64 years 65 years 66 years 67 years 68 years 69 years 70 years 71 years
```

```
##      214608      237351      230015      228172      227433      224777      220617      217912
## 72 years 73 years 74 years 75 years 76 years 77 years 78 years 79 years
##      215276      211297      205086      199244      194695      187163      179204      170453
## 80 years 81 years 82 years 83 years 84 years 85 years 86 years 87 years
##      158613      152006      142938      132634      121260      108644      97349      85394
## 88 years 89 years 90+ years
##      74633      62922      211473
```

```
a$`Age.recode.with.single.ages.and.90+` <- gsub(" years", "", a$`Age.recode.with.single.ages.and.90+`)
table(a$`Age.recode.with.single.ages.and.90+`)
```

```
##
##      00      01      02      03      04      05      06      07      08      09      10
## 5654  5229  5570  5114  4341  3417  3133  2831  2654  2720  2847
##      11      12      13      14      15      16      17      18      19      20      21
## 3072  3344  3719  4155  4765  4955  5646  5780  6298  7221  7899
##      22      23      24      25      26      27      28      29      30      31      32
## 8867  9557 10596 11784 12644 14068 15249 16836 18613 20140 21869
##      33      34      35      36      37      38      39      40      41      42      43
## 23563 25308 27845 30004 32739 35415 39146 45096 48138 52609 56945
##      44      45      46      47      48      49      50      51      52      53      54
## 63520 70593 76882 84428 91745 100210 114538 119922 126394 136600 146278
##      55      56      57      58      59      60      61      62      63      64      65
## 155772 164815 174238 183245 191511 198087 204390 210326 214111 214608 237351
##      66      67      68      69      70      71      72      73      74      75      76
## 230015 228172 227433 224777 220617 217912 215276 211297 205086 199244 194695
##      77      78      79      80      81      82      83      84      85      86      87
## 187163 179204 170453 158613 152006 142938 132634 121260 108644 97349 85394
##      88      89      90+
##      74633      62922      211473
```

```
a <- a %>%
  mutate(
    `CS.Reg.Node.Eval.(2004-2015)` = as.numeric(`CS.Reg.Node.Eval.(2004-2015)`),
    `CS.Tumor.Size/Ext.Eval.(2004-2015)` = as.numeric(`CS.Tumor.Size/Ext.Eval.(2004-2015)`))
```

```
## Warning: There were 2 warnings in 'mutate()'.
## The first warning was:
## i In argument: 'CS.Reg.Node.Eval.(2004-2015)' =
##   as.numeric('CS.Reg.Node.Eval.(2004-2015)')'.
## Caused by warning:
## ! NAs introduced by coercion
## i Run 'dplyr::last_dplyr_warnings()' to see the 1 remaining warning.
```

```
table(a$`CS.Reg.Node.Eval.(2004-2015)`)
```

```
##
##      0      1      2      3      5      6      8      9
## 2046842 131378      2575 1347191 42616 50367 2350 910124
```

```
table(b$`CS.Reg.Node.Eval.(2004-2015)` ) # Check recode
```

```
##
##      0      1      2      3      5      6      8      9
## 2046842 131378 2575 1347191 42616 50367 2350 910124
## Blank(s)
## 3682726
```

```
table(a$`CS.Tumor.Size/Ext.Eval.(2004-2015)` )
```

```
##
##      0      1      2      3      4      5      6      8      9
## 755967 981882 3616 1833007 192513 71948 40181 2877 651486
```

```
table(b$`CS.Tumor.Size/Ext.Eval.(2004-2015)` ) # Check recode
```

```
##
##      0      1      2      3      4      5      6      8
## 755967 981882 3616 1833007 192513 71948 40181 2877
##      9 Blank(s)
## 651486 3682692
```

```
table(a$Marital.status.at.diagnosis) # looks clean... Were non-answers replaced with 'Unknown'?
```

```
##
##              Divorced Married (including common law)
##              723496                                4397646
##      Separated              Single (never married)
##              78147                                1221427
##              Unknown  Unmarried or Domestic Partner
##              661424                                17921
##              Widowed
##              1116108
```

```
table(a$`Race.recode.(W,.B,.AI,.API)` ) # Looks clean
```

```
##
## American Indian/Alaska Native      Asian or Pacific Islander
##              48935                                551271
##              Black                                Unknown
##              812554                                89742
##              White
##              6713667
```

```
table(a$`Race/ethnicity` ) #Looks clean
```

```
##
##      American Indian/Alaska Native      Asian Indian (2010+)
##              48935                                29034
```

```
## Asian Indian or Pakistani, NOS (1988+) Black
## 15070 812554
## Chamorran (1991+) Chinese
## 242 107637
## Fiji Islander (1991+) Filipino
## 1060 120697
## Guamanian, NOS (1991+) Hawaiian
## 1218 28673
## Hmong (1988+) Japanese
## 1284 69481
## Kampuchean (1988+) Korean (1988+)
## 4512 40716
## Laotian (1988+) Melanesian, NOS (1991+)
## 3813 111
## Micronesian, NOS (1991+) New Guinean (1991+)
## 1412 40
## Other Other Asian (1991+)
## 21444 63766
## Pacific Islander, NOS (1991+) Pakistani (2010+)
## 5507 2687
## Polynesian, NOS (1991+) Samoan (1991+)
## 316 4887
## Tahitian (1991+) Thai (1994+)
## 72 4074
## Tongan (1991+) Unknown
## 1556 68311
## Vietnamese (1988+) White
## 43393 6713667
```

```
a$Survival.months <- as.numeric(a$Survival.months)
```

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

```
table(a$Survival.months)
```

```
##
##      0      1      2      3      4      5      6      7      8      9     10
## 461194 304059 239520 190704 165833 147122 141527 118420 108220 114559 108569
##      11      12      13      14      15      16      17      18      19      20      21
## 109751  98476  94920  90602  86176  84728  78400  80844  72886  74639  68783
##      22      23      24      25      26      27      28      29      30      31      32
##  68551  64520  60159  61701  60327  57666  56907  57574  52915  52576  54626
##      33      34      35      36      37      38      39      40      41      42      43
##  49156  51891  48053  47166  46777  47914  44109  45110  45627  43553  44639
##      44      45      46      47      48      49      50      51      52      53      54
##  41833  43531  41923  40145  39917  38672  38845  37782  40139  35180  40562
##      55      56      57      58      59      60      61      62      63      64      65
##  36480  35311  37846  36179  35146  33523  34848  33508  33534  34823  32070
##      66      67      68      69      70      71      72      73      74      75      76
##  35963  31802  33452  32246  31450  30784  29227  29953  30131  30911  28215
##      77      78      79      80      81      82      83      84      85      86      87
##  31225  29517  27388  30067  28863  28358  27389  26241  25726  27211  26966
##      88      89      90      91      92      93      94      95      96      97      98
```

```
## 26271 28483 26200 26571 26407 25751 26075 25344 23490 24128 25682
## 99 100 101 102 103 104 105 106 107 108 109
## 22818 25135 23150 24281 25033 24030 23790 24142 23688 21206 23251
## 110 111 112 113 114 115 116 117 118 119 120
## 22207 21507 23814 20996 24245 22147 21990 22535 21967 21620 20006
## 121 122 123 124 125 126 127 128 129 130 131
## 21955 19909 21550 20903 20868 20730 20034 20015 20371 19974 19154
## 132 133 134 135 136 137 138 139 140 141 142
## 19150 18512 19107 19766 18345 19790 19042 18340 18356 19451 17706
## 143 144 145 146 147 148 149 150 151 152 153
## 18118 17500 16192 18249 16647 16787 16876 18637 16437 17133 16556
## 154 155 156 157 158 159 160 161 162 163 164
## 15847 17165 14300 14997 16712 15060 15918 14965 16554 15044 16194
## 165 166 167 168 169 170 171 172 173 174 175
## 14636 15272 15412 13071 14890 14122 13562 13937 12934 14533 13509
## 176 177 178 179 180 181 182 183 184 185 186
## 13476 13097 13700 12537 12094 12174 11744 12318 12105 12483 12127
## 187 188 189 190 191 192 193 194 195 196 197
## 12437 11337 11739 12307 11461 11018 11035 11066 10878 11113 10116
## 198 199 200 201 202 203 204 205 206 207 208
## 11730 10624 10417 11249 9994 10156 8975 9164 9925 9920 9076
## 209 210 211 212 213 214 215 216 217 218 219
## 9222 10534 8655 9712 9382 8518 9594 8148 8665 9000 8858
## 220 221 222 223 224 225 226 227 228 229 230
## 8227 8375 9248 8161 8902 8122 8385 8166 7067 7871 7869
## 231 232 233 234 235 236 237 238 239 240 241
## 7407 7417 8235 7059 7673 7159 6743 7485 7100 6047 6373
## 242 243 244 245 246 247 248 249 250 251
## 6486 6157 6510 5875 6585 6357 5828 6104 5817 5714
```

```
table(b$Survival.months) # Check recode
```

```
##
## 0000 0001 0002 0003 0004 0005 0006 0007 0008 0009
## 461194 304059 239520 190704 165833 147122 141527 118420 108220 114559
## 0010 0011 0012 0013 0014 0015 0016 0017 0018 0019
## 108569 109751 98476 94920 90602 86176 84728 78400 80844 72886
## 0020 0021 0022 0023 0024 0025 0026 0027 0028 0029
## 74639 68783 68551 64520 60159 61701 60327 57666 56907 57574
## 0030 0031 0032 0033 0034 0035 0036 0037 0038 0039
## 52915 52576 54626 49156 51891 48053 47166 46777 47914 44109
## 0040 0041 0042 0043 0044 0045 0046 0047 0048 0049
## 45110 45627 43553 44639 41833 43531 41923 40145 39917 38672
## 0050 0051 0052 0053 0054 0055 0056 0057 0058 0059
## 38845 37782 40139 35180 40562 36480 35311 37846 36179 35146
## 0060 0061 0062 0063 0064 0065 0066 0067 0068 0069
## 33523 34848 33508 33534 34823 32070 35963 31802 33452 32246
## 0070 0071 0072 0073 0074 0075 0076 0077 0078 0079
## 31450 30784 29227 29953 30131 30911 28215 31225 29517 27388
## 0080 0081 0082 0083 0084 0085 0086 0087 0088 0089
## 30067 28863 28358 27389 26241 25726 27211 26966 26271 28483
## 0090 0091 0092 0093 0094 0095 0096 0097 0098 0099
## 26200 26571 26407 25751 26075 25344 23490 24128 25682 22818
## 0100 0101 0102 0103 0104 0105 0106 0107 0108 0109
```

```
## 25135 23150 24281 25033 24030 23790 24142 23688 21206 23251
## 0110 0111 0112 0113 0114 0115 0116 0117 0118 0119
## 22207 21507 23814 20996 24245 22147 21990 22535 21967 21620
## 0120 0121 0122 0123 0124 0125 0126 0127 0128 0129
## 20006 21955 19909 21550 20903 20868 20730 20034 20015 20371
## 0130 0131 0132 0133 0134 0135 0136 0137 0138 0139
## 19974 19154 19150 18512 19107 19766 18345 19790 19042 18340
## 0140 0141 0142 0143 0144 0145 0146 0147 0148 0149
## 18356 19451 17706 18118 17500 16192 18249 16647 16787 16876
## 0150 0151 0152 0153 0154 0155 0156 0157 0158 0159
## 18637 16437 17133 16556 15847 17165 14300 14997 16712 15060
## 0160 0161 0162 0163 0164 0165 0166 0167 0168 0169
## 15918 14965 16554 15044 16194 14636 15272 15412 13071 14890
## 0170 0171 0172 0173 0174 0175 0176 0177 0178 0179
## 14122 13562 13937 12934 14533 13509 13476 13097 13700 12537
## 0180 0181 0182 0183 0184 0185 0186 0187 0188 0189
## 12094 12174 11744 12318 12105 12483 12127 12437 11337 11739
## 0190 0191 0192 0193 0194 0195 0196 0197 0198 0199
## 12307 11461 11018 11035 11066 10878 11113 10116 11730 10624
## 0200 0201 0202 0203 0204 0205 0206 0207 0208 0209
## 10417 11249 9994 10156 8975 9164 9925 9920 9076 9222
## 0210 0211 0212 0213 0214 0215 0216 0217 0218 0219
## 10534 8655 9712 9382 8518 9594 8148 8665 9000 8858
## 0220 0221 0222 0223 0224 0225 0226 0227 0228 0229
## 8227 8375 9248 8161 8902 8122 8385 8166 7067 7871
## 0230 0231 0232 0233 0234 0235 0236 0237 0238 0239
## 7869 7407 7417 8235 7059 7673 7159 6743 7485 7100
## 0240 0241 0242 0243 0244 0245 0246 0247 0248 0249
## 6047 6373 6486 6157 6510 5875 6585 6357 5828 6104
## 0250 0251 Unknown
## 5817 5714 96680
```

```
table(a$Survival.months.flag) # Looks clean
```

```
##
## Complete dates are available and there are 0 days of survival
## 34407
## Complete dates are available and there are more than 0 days of survival
## 7651896
## Incomplete dates are available and there cannot be zero days of follow-up
## 415516
## Incomplete dates are available and there could be zero days of follow-up
## 17670
## Not calculated because a Death Certificate Only or Autopsy Only case
## 96680
```

```
table(a$`Total.number.of.benign/borderline.tumors.for.patient`) # Looks clean, not sure if this variabl
```

```
##
## 00 01 02 03 04 05 06 07 09 Unknown
## 8165335 49026 1613 146 16 19 8 1 1 4
```

```
table(a$`Total.number.of.in.situ/malignant.tumors.for.patient`)
```

```
##
##      01      02      03      04      05      06      07      08      09      10
## 5973827 1691725 413002 98212  24397  8065  2883  1427  742  466
##      11      12      13      14      15      16      17      18      19      20
##      311      144      130      115      53      19      21      45      6      36
##      21      22      23      24      25      26      27      28      30      32
##      12      28      18      13      3      3      13      10      4      18
##      34      36      43      50      56      58      59 Unknown
##      1      1      1      1      1      24      4      388
```

```
a$`Total.number.of.in.situ/malignant.tumors.for.patient` <- as.numeric(a$`Total.number.of.in.situ/malignant.tumors.for.patient`)
```

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

```
table(a$`Total.number.of.in.situ/malignant.tumors.for.patient`)
```

```
##
##      1      2      3      4      5      6      7      8      9      10
## 5973827 1691725 413002 98212  24397  8065  2883  1427  742  466
##      11      12      13      14      15      16      17      18      19      20
##      311      144      130      115      53      19      21      45      6      36
##      21      22      23      24      25      26      27      28      30      32
##      12      28      18      13      3      3      13      10      4      18
##      34      36      43      50      56      58      59
##      1      1      1      1      1      24      4
```

```
table(b$`Total.number.of.in.situ/malignant.tumors.for.patient`) # Checking recode
```

```
##
##      01      02      03      04      05      06      07      08      09      10
## 5973827 1691725 413002 98212  24397  8065  2883  1427  742  466
##      11      12      13      14      15      16      17      18      19      20
##      311      144      130      115      53      19      21      45      6      36
##      21      22      23      24      25      26      27      28      30      32
##      12      28      18      13      3      3      13      10      4      18
##      34      36      43      50      56      58      59 Unknown
##      1      1      1      1      1      24      4      388
```

```
table(a$`Tumor.marker.1.(1990-2003)`) # Not sure how to treat these tumor marker variables, need a better way
```

```
##
##      0      1      2      3      4      5      6      8
##      59022 132492 39666  475  1332  292  135  4357
##      9 Blank(s)
##      574923 7403475
```

```
table(a$`Tumor.marker.2.(1990-2003)`)
```

```
##
##      0      1      2      3      4      5      6      8
##  9574 105855  55359  1097  1453   242   124  4212
##      9 Blank(s)
## 424308 7613945
```

```
table(a$`Tumor.marker.3.(1998-2003)`)
```

```
##
##      0      2      4      5      6      8      9 Blank(s)
##  1101   740   499   199   141   191 568777 7644521
```

```
## Changing names of variables we want to keep
```

```
a <- a %>%
```

```
  mutate(
    age = `Age.recode.with.single.ages.and.90+`,
    cs.reg.node.eval.04.15 = `CS.Reg.Node.Eval.(2004-2015)`,
    cs.tumor.size.ext.eval.04.15 = `CS.Tumor.Size/Ext.Eval.(2004-2015)`,
    race.recode = `Race.recode.(W,.B,.AI,.API)`,
    race.ethnicity = `Race/ethnicity`,
    survival.months = Survival.months,
    survival.months.flag = Survival.months.flag,
    marital.status.at.diagnosis = Marital.status.at.diagnosis,
    benign.borderline.tumors = `Total.number.of.benign/borderline.tumors.for.patient`,
    situ.malignant.tumors = `Total.number.of.in.situ/malignant.tumors.for.patient`,
    tm.1.1990.2003 = `Tumor.marker.1.(1990-2003)`,
    tm.2.1990.2003 = `Tumor.marker.2.(1990-2003)`,
    tm.3.1998.2003 = `Tumor.marker.3.(1998-2003)`)
```

```
saul_semiclean_vars <- a %>% select(age, cs.reg.node.eval.04.15, cs.tumor.size.ext.eval.04.15,
  race.recode, race.ethnicity, survival.months,
  survival.months.flag, Marital.status.at.diagnosis,
  benign.borderline.tumors, situ.malignant.tumors,
  tm.1.1990.2003, tm.2.1990.2003, tm.3.1998.2003)
```