Breast Exploratory Analysis

Jonathan Ma

2025-12-01

Table of Contents

1 Introduction

This exploratory data analysis (EDA) is the first step in our project pipeline:

1. Bayesian Hierarchical Modeling
2. Posterior Inference & Diagnostics
3. Clustering of Individual Risk Profiles

We use data from the SEER Program focused on breast cancer cases, structured across three levels:

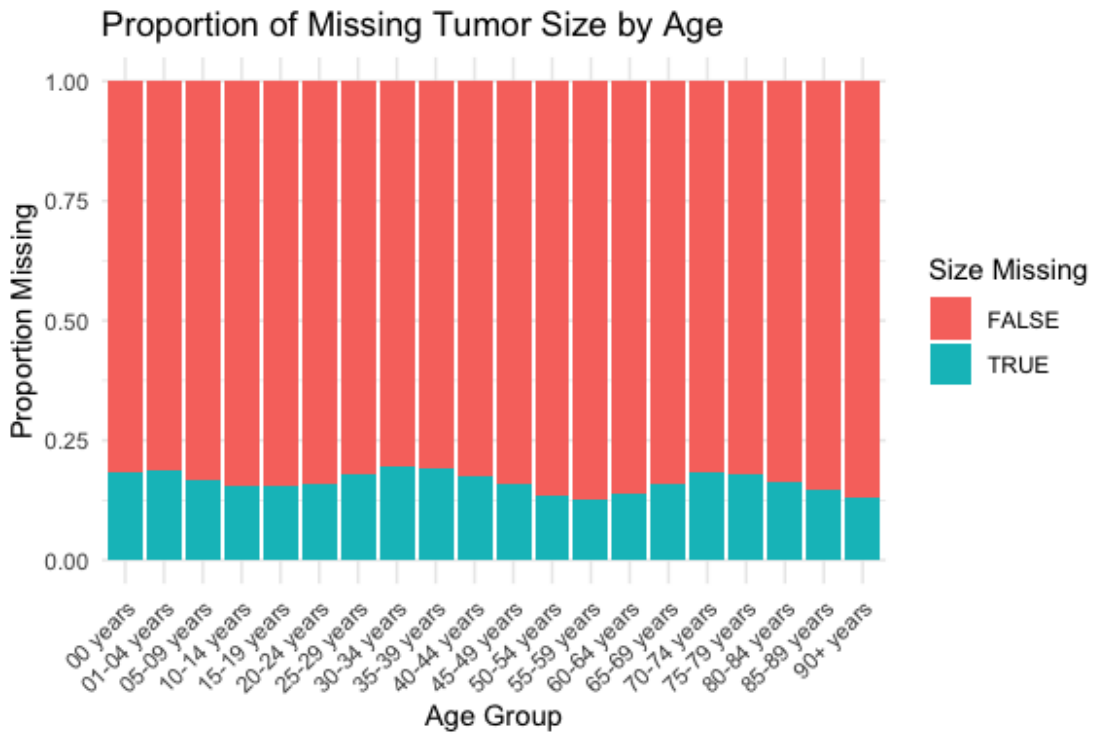
- **Level 1 (Tumor):** Tumor-specific details (e.g., stage, size, grade)
- **Level 2 (Patient):** Demographics and marital status
- **Level 3 (Region):** County-level metro/rural designations

2 Data Structure and Levels

Level	Variable	Description
Tumor	stage	SEER stage: Localized, Regional, Distant
Tumor	grade	Tumor grade: 1–4 or unknown
Tumor	size	EOD Tumor Size (mm)
Tumor	site	Tumor site (ICD code)
Patient	id	Patient identifier
Patient	age	Age group at diagnosis
Patient	sex	Biological sex
Patient	raceth	Race/ethnicity
Patient	year	Year of diagnosis
Patient	marry	Marital status
Region	region	County-level metro/rural code

We replace “Unknown” and “Blank(s)” with `NA`, and then analyse missing data. We want to see if missing data is specific to a group so we can determine if the missingness is MCAR or not.

Is missing data more common in older patients?



Is missing data associated with Race/Ethnicity?

```
##
##                                    FALSE    TRUE
##   American Indian/Alaska Native    40048    4775
##   Asian or Pacific Islander       453003   58740
##   Black                           369571   62192
##   White                          3427119  685078
```

Is stage more likely to be missing in nonmetro rural areas?

```
##
##                                                                    FALSE
##   Counties in metropolitan areas ge 1 million pop                0.5234788
##   Counties in metropolitan areas of 250,000 to 1 million pop     0.5292829
##   Counties in metropolitan areas of lt 250 thousand pop          0.5255872
##   Nonmetropolitan counties adjacent to a metropolitan area       0.5250781
##   Nonmetropolitan counties not adjacent to a metropolitan area   0.5329171
##   Unknown/missing/no match (Alaska or Hawaii - Entire State)     0.4969591
##   Unknown/missing/no match/Not 1990-2023                         0.6902941
##
##                                                                     TRUE
##   Counties in metropolitan areas ge 1 million pop                0.4765212
##   Counties in metropolitan areas of 250,000 to 1 million pop     0.4707171
##   Counties in metropolitan areas of lt 250 thousand pop          0.4744128
##   Nonmetropolitan counties adjacent to a metropolitan area       0.4749219
##   Nonmetropolitan counties not adjacent to a metropolitan area   0.4670829
```

```
##   Unknown/missing/no match (Alaska or Hawaii - Entire State)   0.5030409
##   Unknown/missing/no match/Not 1990-2023                       0.3097059
```

Although we did not do a statistical test for MCAR, visual and tabular inspection clearly show patterns in missingness. Tumor size by age is non-uniform, with younger and older ends having higher missing rates. Missing rates also vary by racial group, with whites having notably higher missingness.

Logistic regression is a standard method for exploring missing data mechanisms. By modeling the missingness indicator (e.g., stage_miss) as a binary outcome and regressing it on observed variables, we can test whether missingness is related to any known covariates.

If none of the predictors are significant, this supports the MCAR assumption. But if missingness is significantly associated with observed variables (as seen here), MCAR is rejected, and MAR becomes a more reasonable assumption for downstream modeling.

The missingness in stage is likely not MCAR, because missingness depends significantly on several observed variables (e.g., sex, raceth, region, age, marry, grade, size). Therefore, it's reasonable to assume MAR (Missing at Random) for modeling purposes. We find we have 209123 cases to work with (4.06% of the whole thing) after removing "Unknown" and "Blank(s)" Values.

```
## Complete cases: 209123 out of 5149008 (4.06% complete)
```

3 Variable-Specific Treatment

Further cleaning is done via tabling all variables.

We want to filter for top ten cancer sites (breast, lung, etc). We exclude 998 and 999 because they are unknown.

```
##
##  010  012  015  018  020  025  030  035  040  050
## 4269 3940 5738 3284 5440 4994 6046 3768 4953 4415
```

Collapse and recode into ordered, interpretable regional factors, removes the Alaska/Hawaii "unknown" entries.

```
##
##    Large    Small Nonmetro
##    39835     2410     4497
```

Restrict to tumors that are invasive (drop In situ, Unknown/unstaged).

```
##
## Localized  Regional   Distant
##     26446     13382      3918
```

Extract numeric grade 1–4; drop cell-line entries (“B-cell”, “T-cell” etc.). Convert them to Start (1/2) and End (3/4) since our model is logistic.

```
##
## Start   End
## 42877  3650
```

Making this numeric and then filtering implausible values.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.00   23.00   30.00   34.66   44.00   99.00    6305
```

Making this binary, only married or unmarried.

```
##
##   Married Unmarried
##     28032     18815
```

Recode RACETH to be three levels, Black, White, and Other.

```
##
##     W     B     O
## 36288  3978  6581
```

3.1 Complete Summary of available data

Seeing how many cases are available now:

```
## Remaining complete cases: 37530
```

4 Hierarchal Encoding

To do hierarchal models in R, identifiers must be encoded as numeric integers (e.g., 1, 2, ..., N) for levels like REGION and ID. We complete that, then set factor variables, and check resulting data structure.

```
## 'data.frame':    37530 obs. of  13 variables:
##  $ id       : int  1028 1034 1152 2172 4489 5518 9181 9314 9415 10092 ...
##  $ region   : Factor w/ 3 levels "Large","Small",..: 1 1 1 1 1 1 1 1 1 1
 ...
##  $ stage    : Factor w/ 3 levels "Localized","Regional",..: 1 1 1 1 1 1 1
2 1 2 ...
##  $ age      : Factor w/ 20 levels "00 years","01-04 years",..: 14 16 17 19
 15 13 17 18 18 17 ...
##  $ sex      : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 1 1 1 1 ...
##  $ raceth   : Factor w/ 3 levels "W","B","O": 1 2 1 3 1 2 3 1 3 1 ...
##  $ grade    : Factor w/ 2 levels "Start","End": 1 1 1 2 1 1 1 1 2 1 ...
##  $ size     : num  30 30 38 27 22 49 22 50 27 27 ...
##  $ year     : int  2017 2016 2016 2017 2017 2016 2017 2017 2017 2016 ...
##  $ marry    : Factor w/ 2 levels "Married","Unmarried": 1 2 1 2 2 2 1 1 2
```

```
2 ...
##  $ site     : chr  "020" "025" "035" "020" ...
##  $ regionid : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ patientid: int  1 2 3 4 5 6 7 8 9 10 ...
```