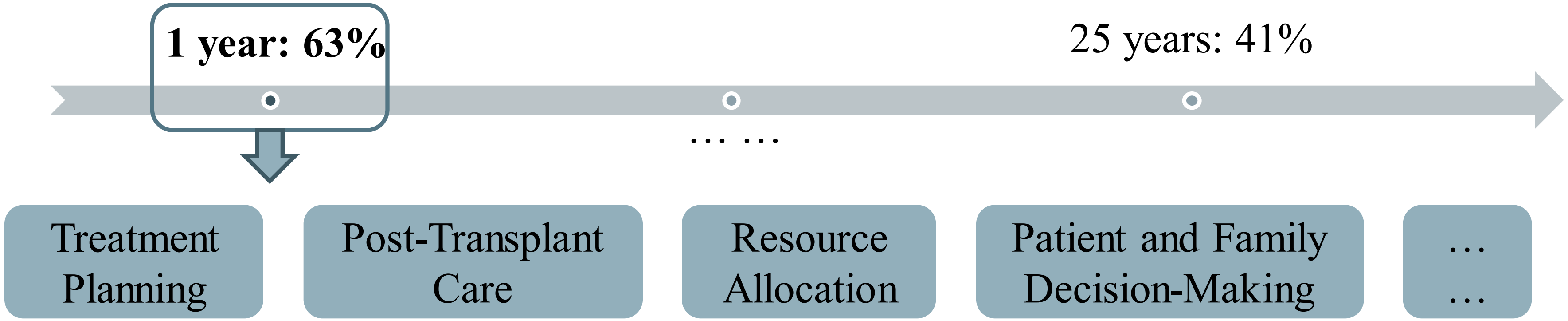
Pediatric Bone Marrow Transplant Survival Prediction

Zoey Yu Brown University Oct 24, 2024

https://github.com/Zoey-Yu98/data1030-midterm.git

Bone Marrow Transplant (BMT)

- To replace unhealthy bone marrow with healthy cells.
- One of the most effective treatments for blood related cancer and diseases (leukemia, lymphoma, and multiple myeloma)
- **BMTs have serious risks and low survival rate**



1 year: 63%

25 years: 41%

… …

Treatment Planning

Post-Transplant Care

Resource Allocation

Patient and Family Decision-Making

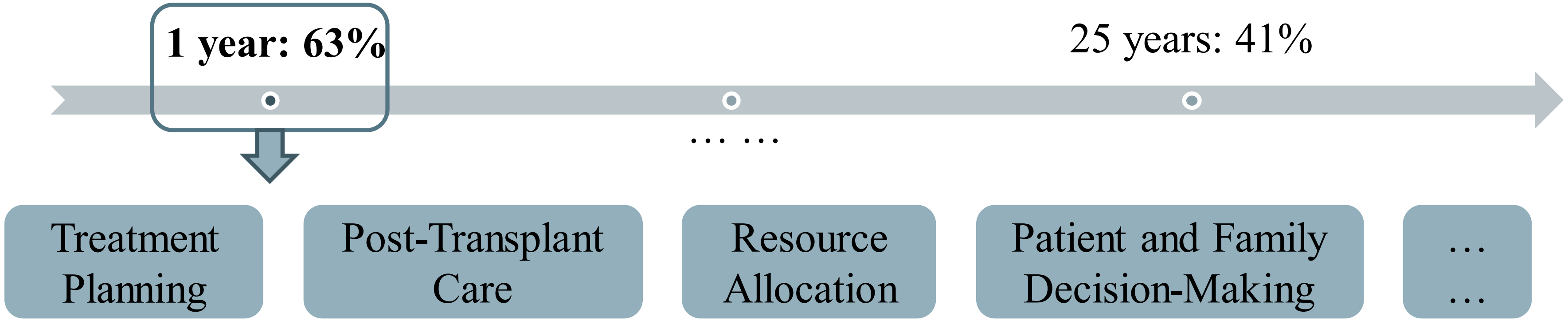
… …

Biology of Blood and Marrow Transplantation, Volume 12, Issue 4, 422 - 429

Bone Marrow Transport (BMT)

A Post-op 1-yr Survival Status Prediction ➔ Decisions Making

- **BMTs have serious risks and low survival rate**



1 year: 63%

25 years: 41%

… …

Treatment Planning

Post-Transplant Care

Resource Allocation

Patient and Family Decision-Making

… …

Biology of Blood and Marrow Transplantation, Volume 12, Issue 4, 422 - 429

Dataset Brief

- Hospital in Poland, from 2000 to 2010
- **187 Pediatric Patients** who underwent BMTs
- **37 Attributes**
  - **Pre-op Parameters** (e.g. Age, Gender, Blood Type, Disease, Antigen)
  - **Post-op Assessments** (e.g. Graft versus Host Disease (GvHD))
- **Classification** Problem (1 Year after Surgery, **Alive** or **Dead**)

Source: Sikora, M., Wróbel, Ł., & Gudyś, A. (2020). Bone marrow transplant: children [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5NP6Z.

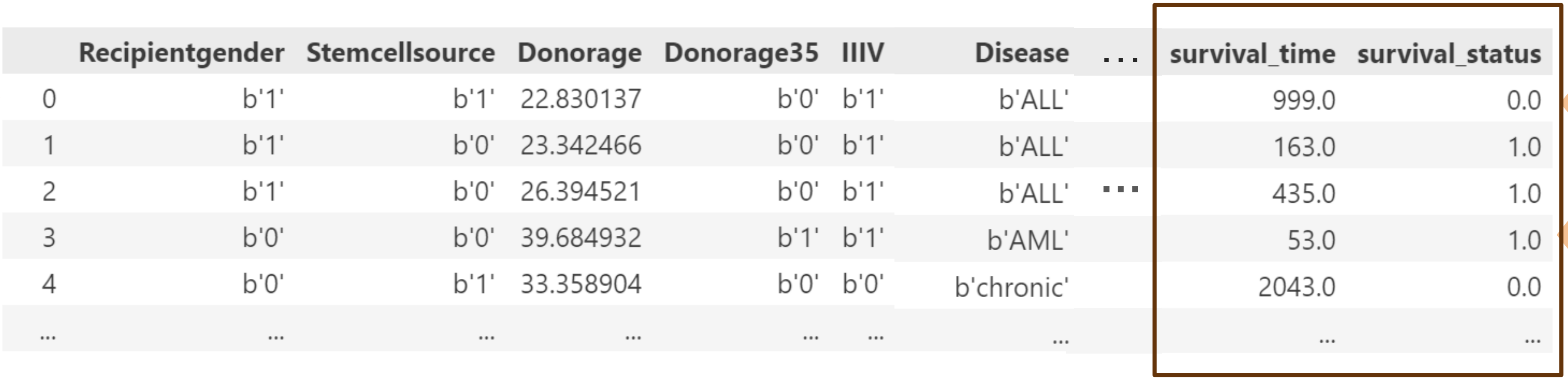
Basic Understanding Of The Data

1. **Data is iid within each feature**

Target Variable
1-yr survival status

One patient

187 rows
×
37 columns

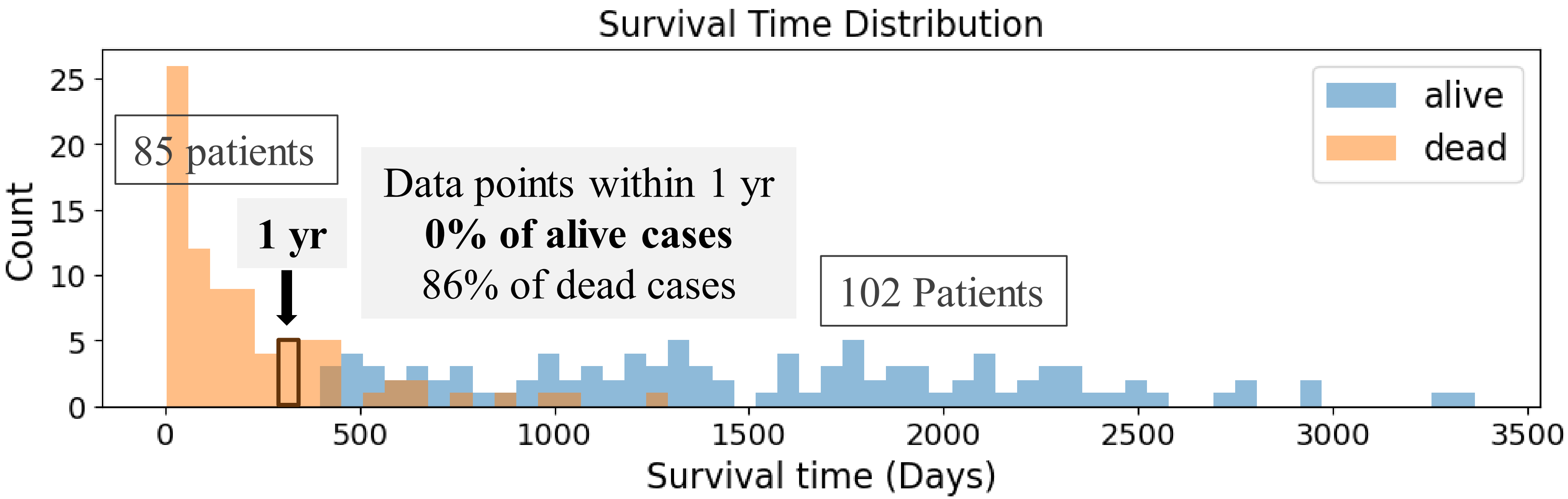


	Recipientgender	Stemcellsource	Donorage	Donorage35	IIIV	Disease	. . .	survival_time	survival_status
0	b'1'	b'1'	22.830137	b'0'	b'1'	b'ALL'		999.0	0.0
1	b'1'	b'0'	23.342466	b'0'	b'1'	b'ALL'		163.0	1.0
2	b'1'	b'0'	26.394521	b'0'	b'1'	b'ALL'	...	435.0	1.0
3	b'0'	b'0'	39.684932	b'1'	b'1'	b'AML'		53.0	1.0
4	b'0'	b'1'	33.358904	b'0'	b'0'	b'chronic'		2043.0	0.0
...	...	...	...	...	...	...		...	...

alive

dead

Survival Time: Time of observation (if alive ‘0’) or time of death (if dead ‘1’) in days



85 patients

1 yr

Data points within 1 yr
0% of alive cases
86% of dead cases

102 Patients

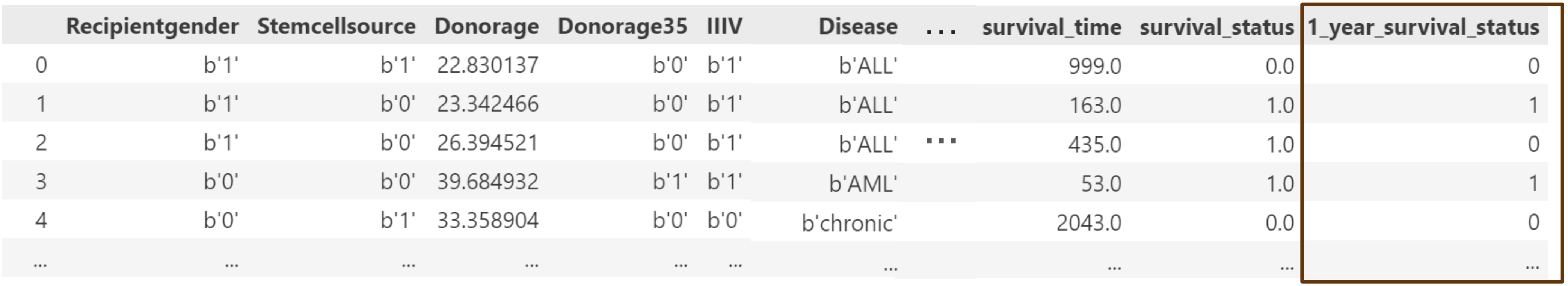
Basic Understanding Of The Data

1. **Data is iid within each feature**

Target Variable

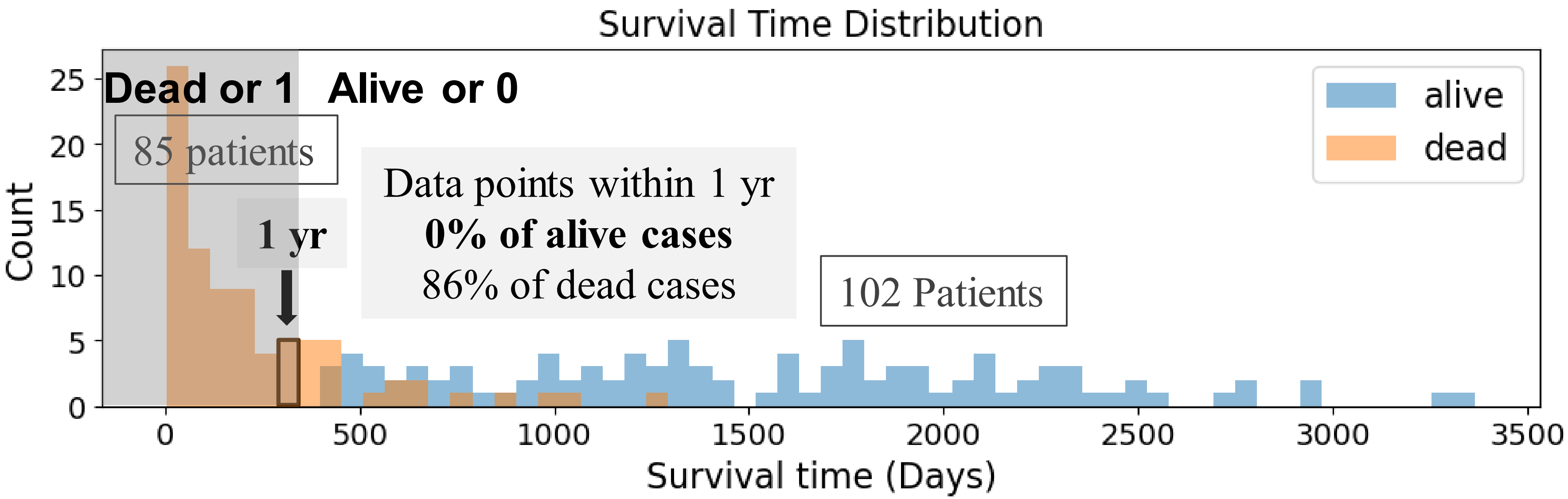
One patient

187 rows × 37 columns



	Recipientgender	Stemcellsource	Donorage	Donorage35	IIIV	Disease	...	survival_time	survival_status	1_year_survival_status
0	b'1'	b'1'	22.830137	b'0'	b'1'	b'ALL'		999.0	0.0	0
1	b'1'	b'0'	23.342466	b'0'	b'1'	b'ALL'		163.0	1.0	1
2	b'1'	b'0'	26.394521	b'0'	b'1'	b'ALL'	...	435.0	1.0	0
3	b'0'	b'0'	39.684932	b'1'	b'1'	b'AML'		53.0	1.0	1
4	b'0'	b'1'	33.358904	b'0'	b'0'	b'chronic'		2043.0	0.0	0
...	...	...	...	...	...	...		...	...	...

Goal: Postoperatively, to predict 1-year survival status.



Dead or 1 **Alive or 0**

85 patients

1 yr

Data points within 1 yr
0% of alive cases
86% of dead cases

102 Patients

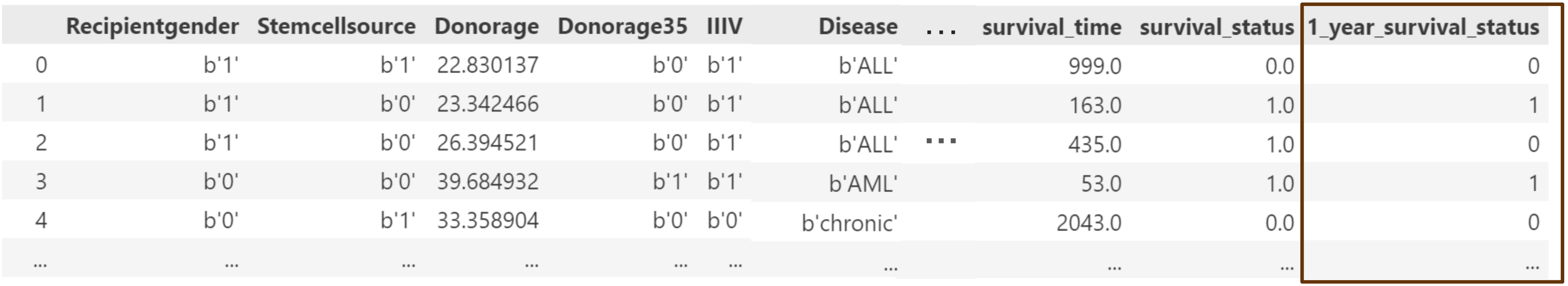
Basic Understanding Of The Data

1. **Data is iid within each feature**

One patient

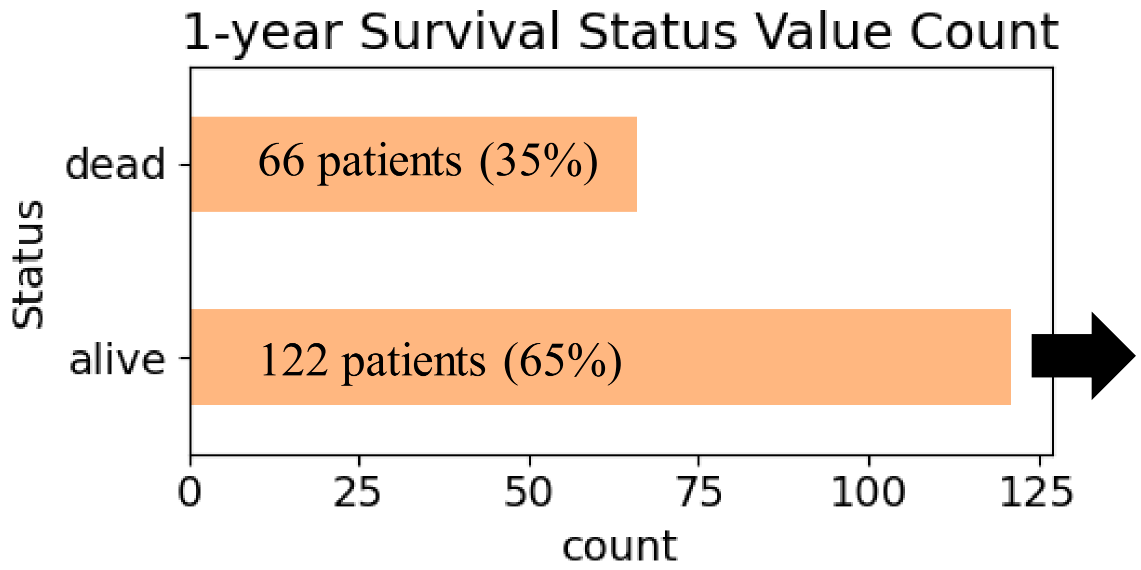
187 rows × 37 columns

Target Variable



	Recipientgender	Stemcellsource	Donorage	Donorage35	IIIV	Disease	...	survival_time	survival_status	1_year_survival_status
0	b'1'	b'1'	22.830137	b'0'	b'1'	b'ALL'		999.0	0.0	0
1	b'1'	b'0'	23.342466	b'0'	b'1'	b'ALL'		163.0	1.0	1
2	b'1'	b'0'	26.394521	b'0'	b'1'	b'ALL'	···	435.0	1.0	0
3	b'0'	b'0'	39.684932	b'1'	b'1'	b'AML'		53.0	1.0	1
4	b'0'	b'1'	33.358904	b'0'	b'0'	b'chronic'		2043.0	0.0	0
...	...	...	...	...	...	...		...	...	...

Goal: Postoperatively, to predict 1-year survival status.



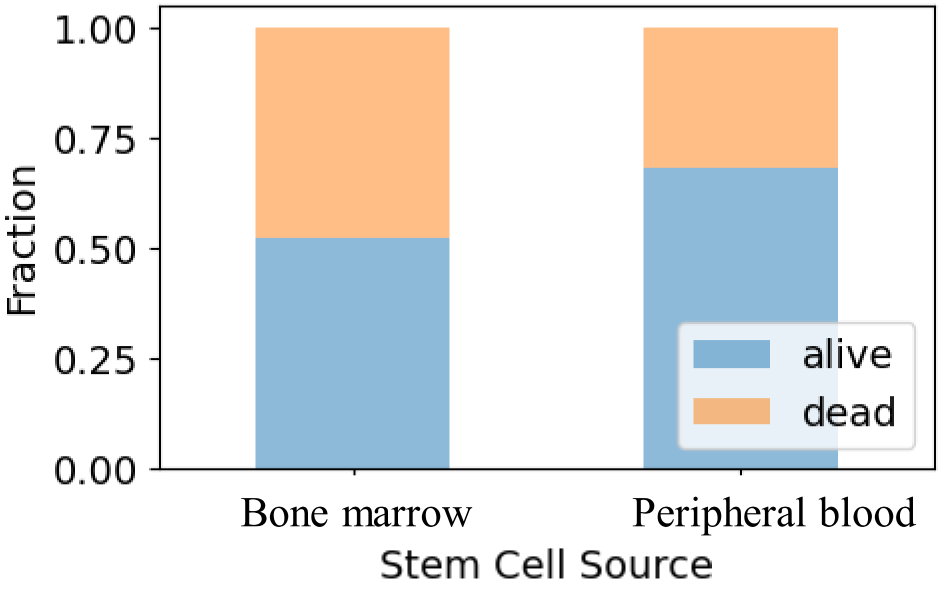
baseline accuracy of our classification model

Basic Understanding Of The Data

2. **Some important correlations between features and the target variable**

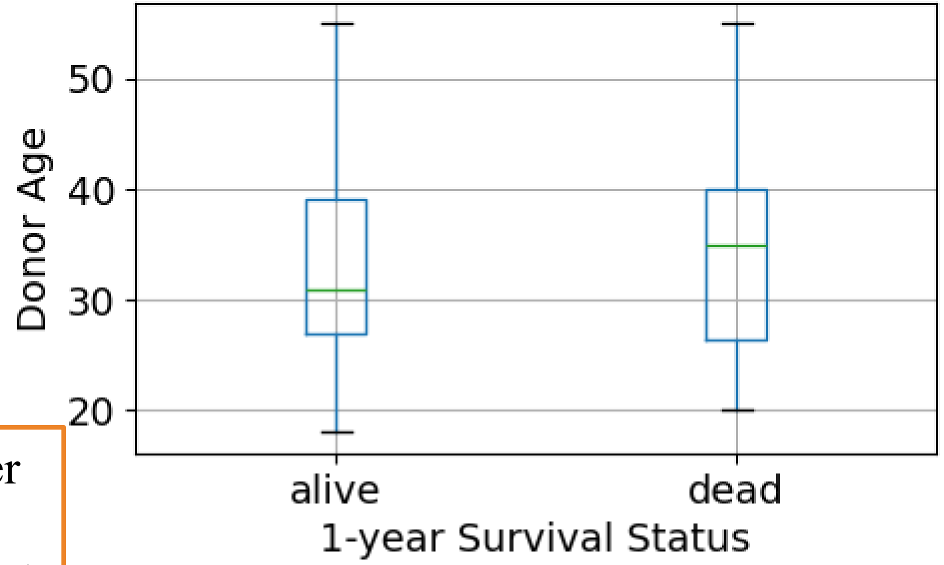
Pre-op

Source of Hematopoietic Stem Cells



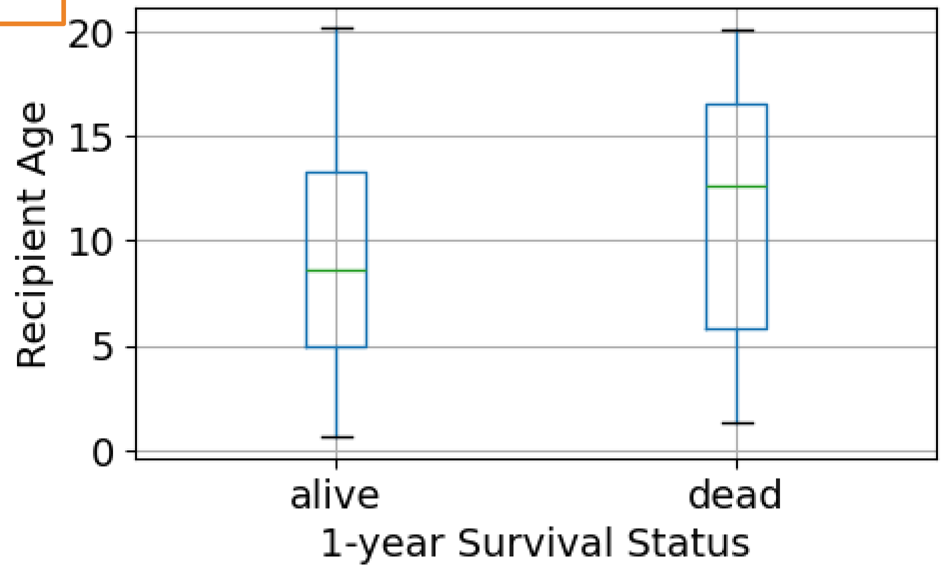
BMT sourced from peripheral blood worked better than from bone marrow.

Donor Age



Lower Age Worked Better

Recipient Age

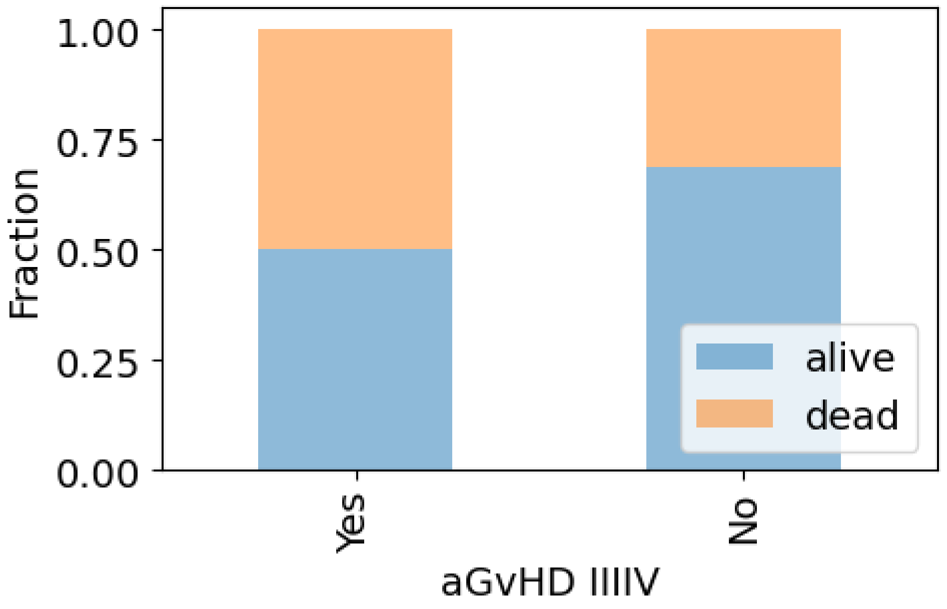


Basic Understanding Of The Data

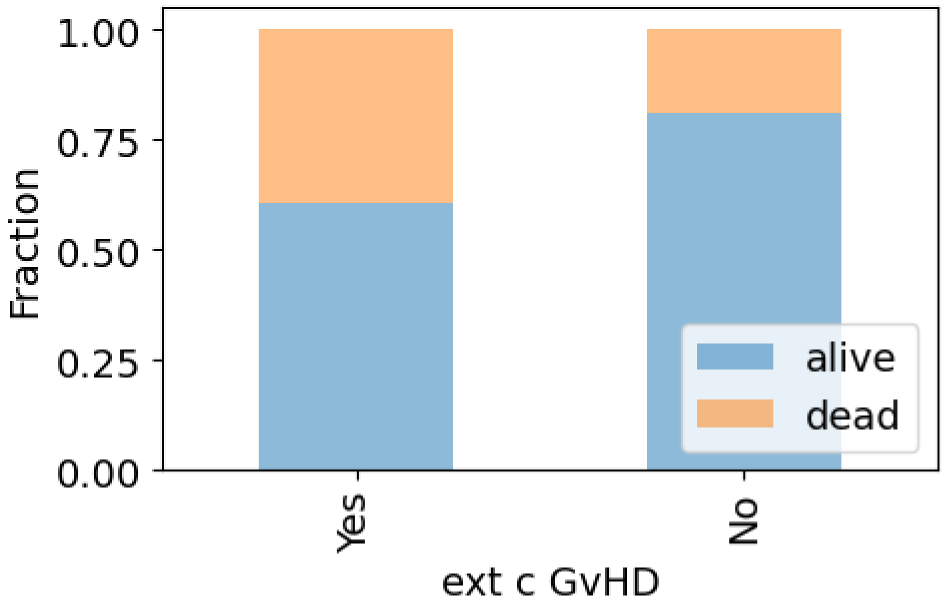
2. Some important correlations between features and the target variable

Post-op

Development of acute GvHD stage III or IV



Development of extensive chronic GvHD



Acute or extensive GvHD increased the risk of death, though it was not always fatal.

Splitting Of The Data

iid data splitting: `from sklearn.model_selection import train_test_split`

	Training Set	Validation Set	Test Set
% of patients	60 %	20 %	20 %
# of patients	112	37	38

Missing Data

Features	Fraction of NA
R blood type	0.005348
R blood cell Rh	0.010695
Blood type match	0.005348
Serological compatibility	0.085561

Features	Fraction of NA
D CMV infection	0.010695
R CMV infection	0.074866
Antigen difference	0.005348
Allele difference	0.005348

Features	Fraction of NA
extensive chronic GvHB	0.165775
CD3+/CD34+	0.026738
CD3+ dose	0.026738
R HSC before surgery	0.010695

Fraction of points with missing values (NA) : 0.24064171122994651

R = recipient ***D*** = donor HSC = hematopoietic stem cells (CD34+ cells)

Dealing with Missing Data --- Clarify The Data Type

Features	Type
R blood type	Categorical
R blood cell Rh	Categorical
Blood type match	Categorical
Serological compatibility	Ordinal

Features	Type
D CMV infection	Categorical
R CMV infection	Categorical
Antigen difference	Ordinal
Allele difference	Ordinal

Features	Type
extensive chronic GvHB	Categorical
CD3+/CD34+	Continuous
CD3+ dose	Continuous
R HSC before surgery	Continuous

Most of them are **categorical** (non ordinal), some are **ordinal**, and some are **continuous**

R = recipient ***D*** = donor HSC = hematopoietic stem cells (CD34+ cells)

Dealing with Missing Data --- **Categorical** (Not Ordinal) Data

Features	Type
R blood type	**Categorical**
R blood cell Rh	**Categorical**
Blood type match	**Categorical**
Serological compatibility	**Ordinal**
D CMV infection	**Categorical**
R CMV infection	**Categorical**
Antigen difference	**Ordinal**
Allele difference	**Ordinal**
extensive chronic GvHB	**Categorical**
CD3+/CD34+	Continuous
CD3+ dose	Continuous
R HSC before surgery	Continuous

Learned in class: the BEST thing we can do is to **treat missing values as another category**

E.g. **Blood type**:

- Unlikely the hospital did not have blood type tested
- Over 30 blood group systems in addition to ABO and Rh
  - E.g. Duffy, K antigen (or Kell), Lutheran, and Kidd blood groups

SimpleImputer (strategy='constant’)

R = recipient ***D*** = donor HSC = hematopoietic stem cells (CD34+ cells)

Dealing with Missing Data --- **Ordinal** Data

Features	Type
R blood type	**Categorical**
R blood cell Rh	**Categorical**
Blood type match	**Categorical**
Serological compatibility	**Ordinal**

Features	Type
D CMV infection	**Categorical**
R CMV infection	**Categorical**
Antigen difference	**Ordinal**
Allele difference	**Ordinal**

Features	Type
extensive chronic GvHB	**Categorical**
CD3+/CD34+	Continuous
CD3+ dose	Continuous
R HSC before surgery	Continuous

Learned in class: the BEST thing we can do is to **treat missing values as another category**

How should they be ordered?

To treat missing values **as the highest category**

R = recipient ***D*** = donor HSC = hematopoietic stem cells (CD34+ cells)

Dealing with Missing Data --- **Ordinal** Data

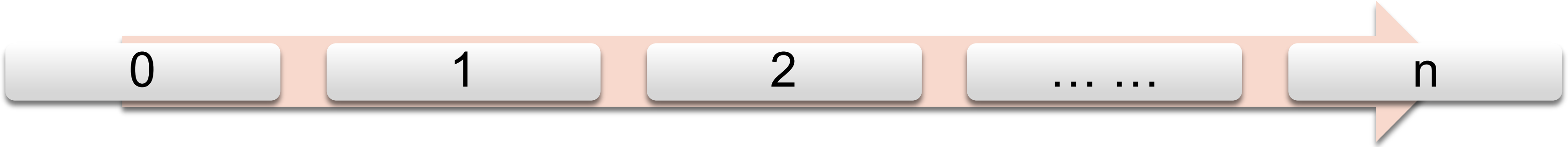
Features	Type
R blood type	**Categorical**
R blood cell Rh	**Categorical**
Blood type match	**Categorical**
Serological compatibility	**Ordinal**

Features	Type
D CMV infection	**Categorical**
R CMV infection	**Categorical**
Antigen difference	**Ordinal**
Allele difference	**Ordinal**

Features	Type
extensive chronic GvHB	**Categorical**
CD3+/CD34+	Continuous
CD3+ dose	Continuous
R HSC before surgery	Continuous

Learned in class: the BEST thing we can do is to **treat missing values as another category**

How should they be ordered? To treat missing values **as the highest category**



Difference	No Difference	n-1 Differences
Compatibility	Most Compatible	Lest Compatible

SimpleImputer (strategy='constant’)

R = recipient ***D*** = donor HSC = hematopoietic stem cells (CD34+ cells)

Dealing with Missing Data --- Continuous Data

Features	Type
R blood type	Categorical
R blood cell Rh	Categorical
Blood type match	Categorical
Serological compatibility	Ordinal

Features	Type
D CMV infection	Categorical
R CMV infection	Categorical
Antigen difference	Ordinal
Allele difference	Ordinal

Features	Type
extensive chronic GvHB	Categorical
CD3+/CD34+	Continuous
CD3+ dose	Continuous
R HSC before surgery	Continuous

For now, based on our current knowledge: SimpleImputer (strategy='mean')

Will change to better method when we learn more in the class.

R = recipient ***D*** = donor HSC = hematopoietic stem cells (CD34+ cells)

High Correlation (Redundant) Data Reduction

Donor age, Donor age 35; **Recipient age**, Recipient age 10, Recipient age int

Fit_Transformation X_Train

SimpleImputer + OneHotEncoder

Features =

- gender, age, matching, Disease, judging-related features

SimpleImputer + Ordinal Encoder

Features =

- leveled status, matching grade, age groups-related features

SimpleImputer + StandardScaler

Features =

- Dosed for CD34+, CD3+
- Body mass
- Recovery time for some side effects

MinMaxScaler

Features =

- donor age, recipient age

R = recipient ***D*** = donor HSC = hematopoietic stem cells (CD34+ cells)

Transformation X_val and X_test

Data Ready for Machine Learning!

```
X_train.shape
(112, 36)
```

```
X_train_prep.shape
(112, 68)
```

THANK YOU

Zoey “Ziyan” Yu

Ziyan_yu@brown.edu

PhD Candidate in Chemistry

Brown University

Oct 24, 2024