Prepared by Group 8

Predicting AIDS Patient Survival

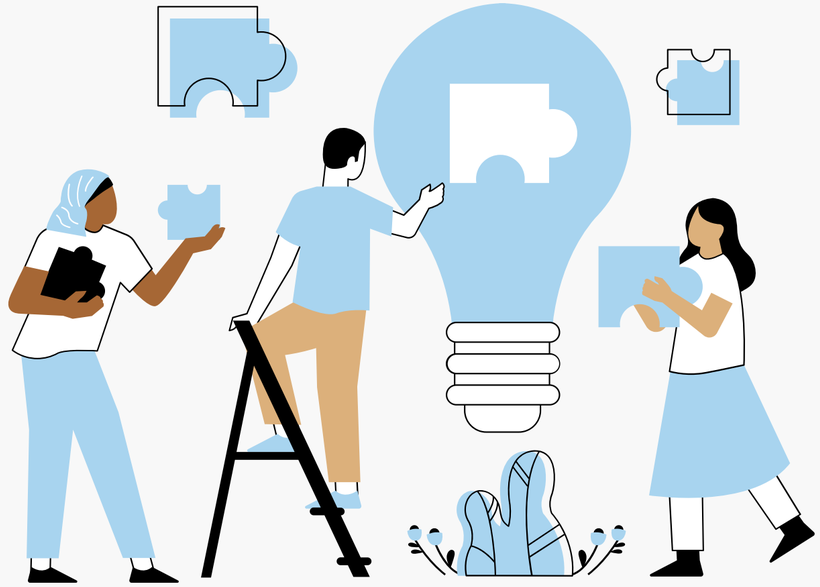
Hang Do Minh, Alex Nirvinen, Jani Oasmaa, Lana St. James-Kautny,
Ekaterina Ustiukhina, Anette Vehniäinen

2 April, 2025

Motivation and Problem Statement

Our goal was to develop predictive models to estimate whether a patient survived during the study period, using clinical and demographic features from ACTG Study 175. The dataset was collected in the United States during the study, and was initially published in 1996.

AIDS remains a critical global health challenge. Timely and data-informed clinical decisions can significantly improve patient outcomes.By identifying high-risk individuals early, healthcare professionals can adapt interventions and drug manufacturers can assess treatment efficacy.



We frame this as a **binary classification problem**: did the patient die during the trial or not?

Dataset Background

Sourced from Kaggle:
AIDS Clinical Trials Group Study 175
2,139 patients with advanced HIV infection

Patients received one of four treatments:

- 0 = Zidovudine (ZDV) only
- 1 = ZDV + Didanosine (ddI)
- 2 = ZDV + Zalcitabine (Zal)
- 3 = Didanosine (ddI) only

Key variable: CD4 lymphocyte count at 20 weeks – an indicator of immune system strength

#	Column	Non-Null Count	Dtype
0	time	2139 non-null	int64
1	trt	2139 non-null	int64
2	age	2139 non-null	int64
3	wtkg	2139 non-null	float64
4	hemo	2139 non-null	int64
5	homo	2139 non-null	int64
6	drugs	2139 non-null	int64
7	karnof	2139 non-null	int64
8	oprior	2139 non-null	int64
9	z30	2139 non-null	int64
10	zprior	2139 non-null	int64
11	preanti	2139 non-null	int64
12	race	2139 non-null	int64
13	gender	2139 non-null	int64
14	str2	2139 non-null	int64
15	strat	2139 non-null	int64
16	symptom	2139 non-null	int64
17	treat	2139 non-null	int64
18	offtrt	2139 non-null	int64
19	cd40	2139 non-null	int64
20	cd420	2139 non-null	int64
21	cd80	2139 non-null	int64
22	cd820	2139 non-null	int64
23	label	2139 non-null	int64

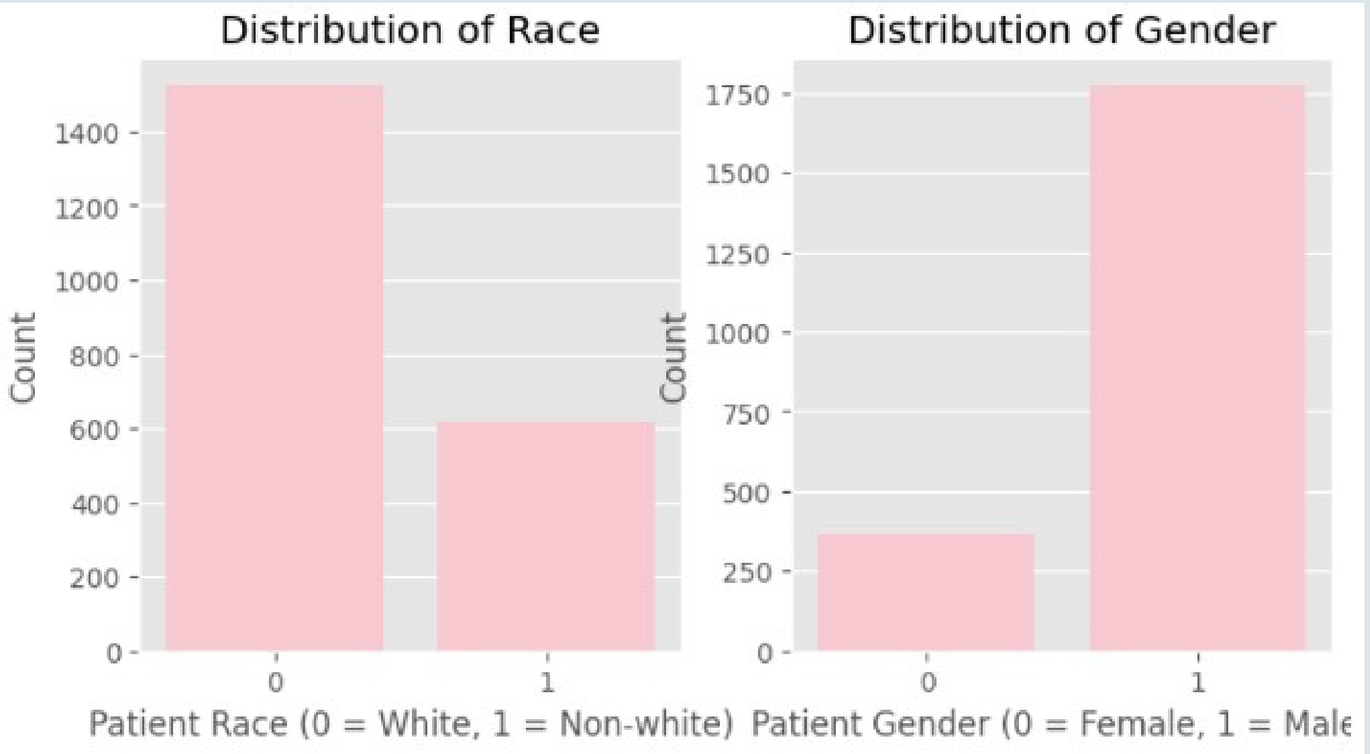
The dataset includes:

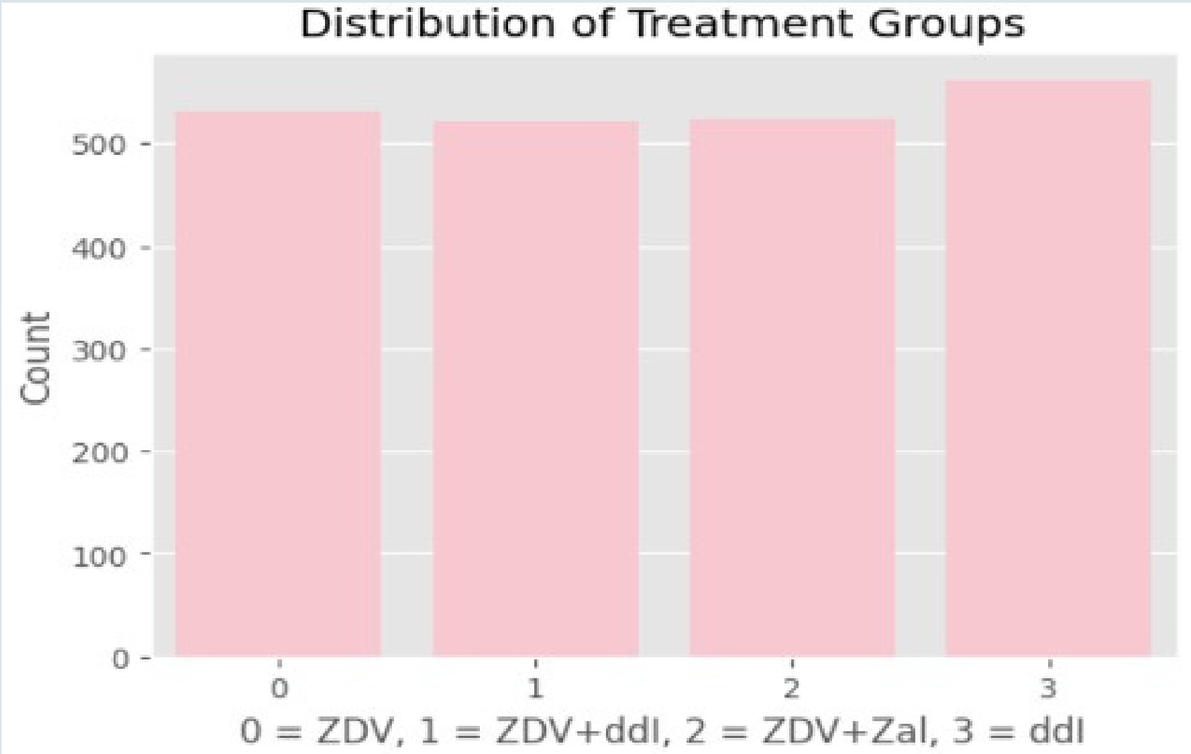
- **Demographics**: Age, weight, gender, race, sexual orientation
- **Clinical measures**: CD4 count, CD8 count, Karnofsky score
- **Medical history**: Hemophilia status, IV drug use
- **Treatment information**: Antiretroviral drug type
- **Outcome**: Survival status within the study period (binary label: censored = survived; failure = died or progressed)

There are no missing values, making the dataset suitable for predictive modeling. Checking the feature distributions also did not reveal any outliers.

Demographics and Treatment

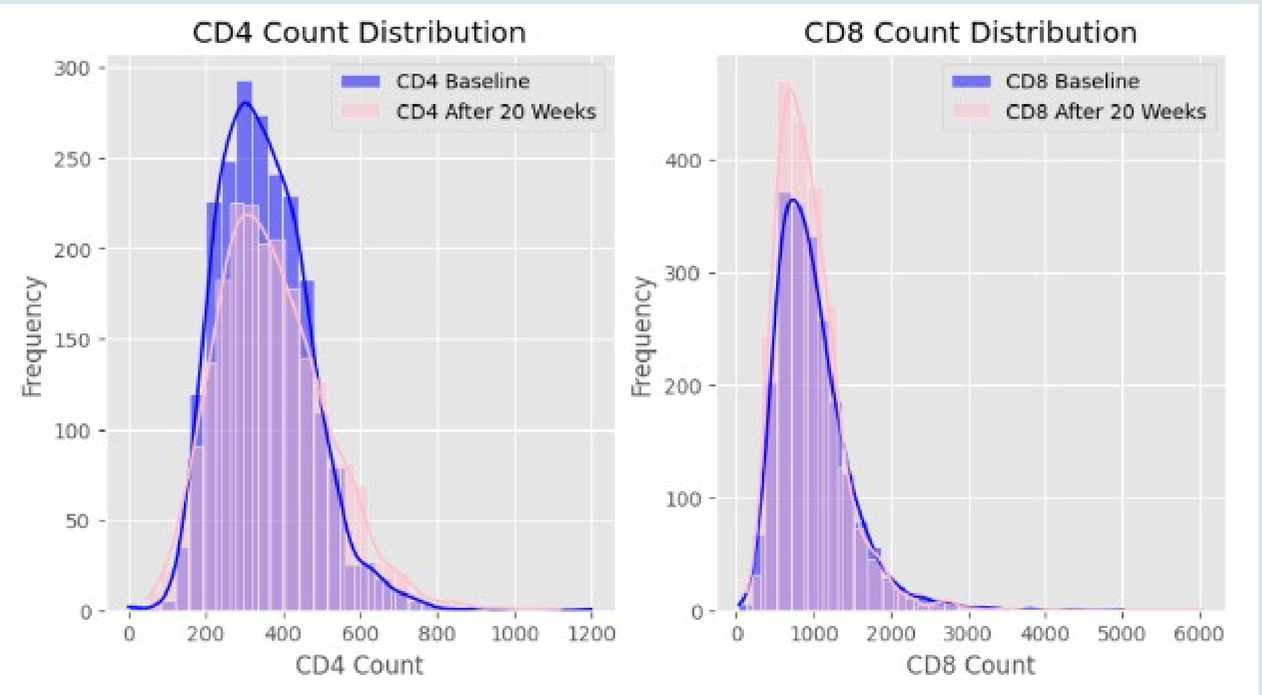
- Majority of patients are **white males**, which may limit the generalizability of results to underrepresented populations.
- Patients are **fairly evenly split** across treatment groups (~500–600 per group), which improves robustness.

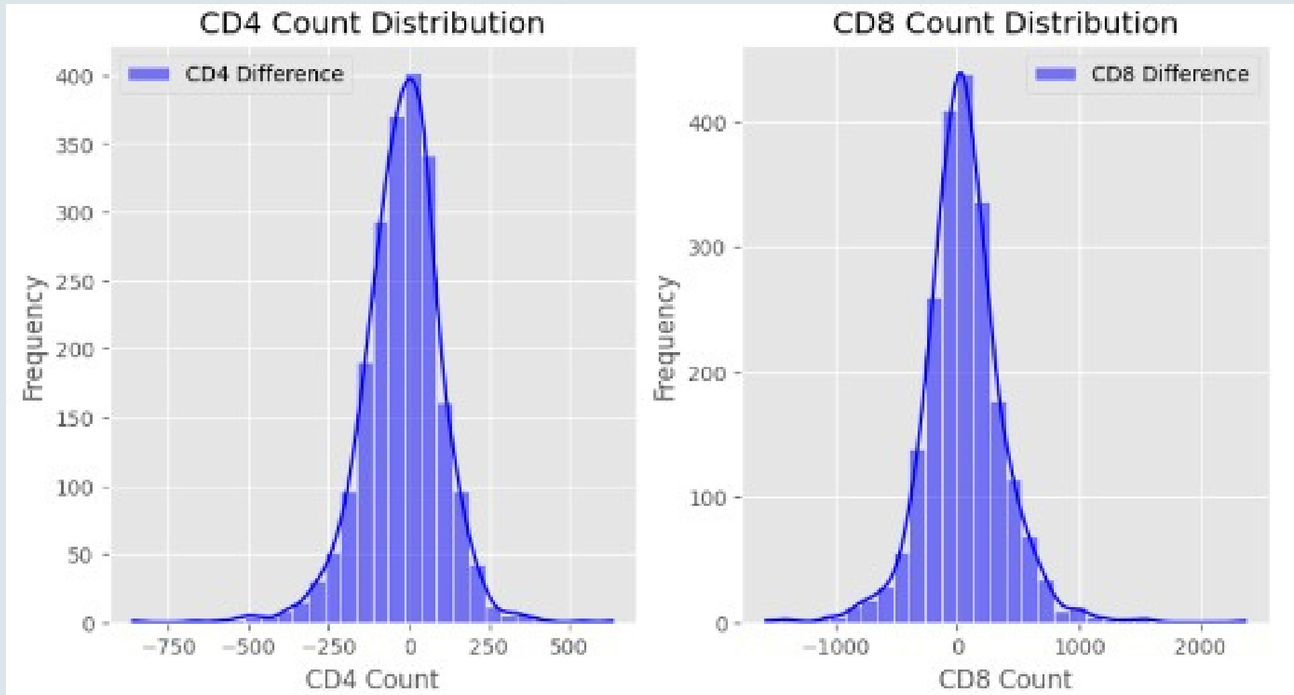


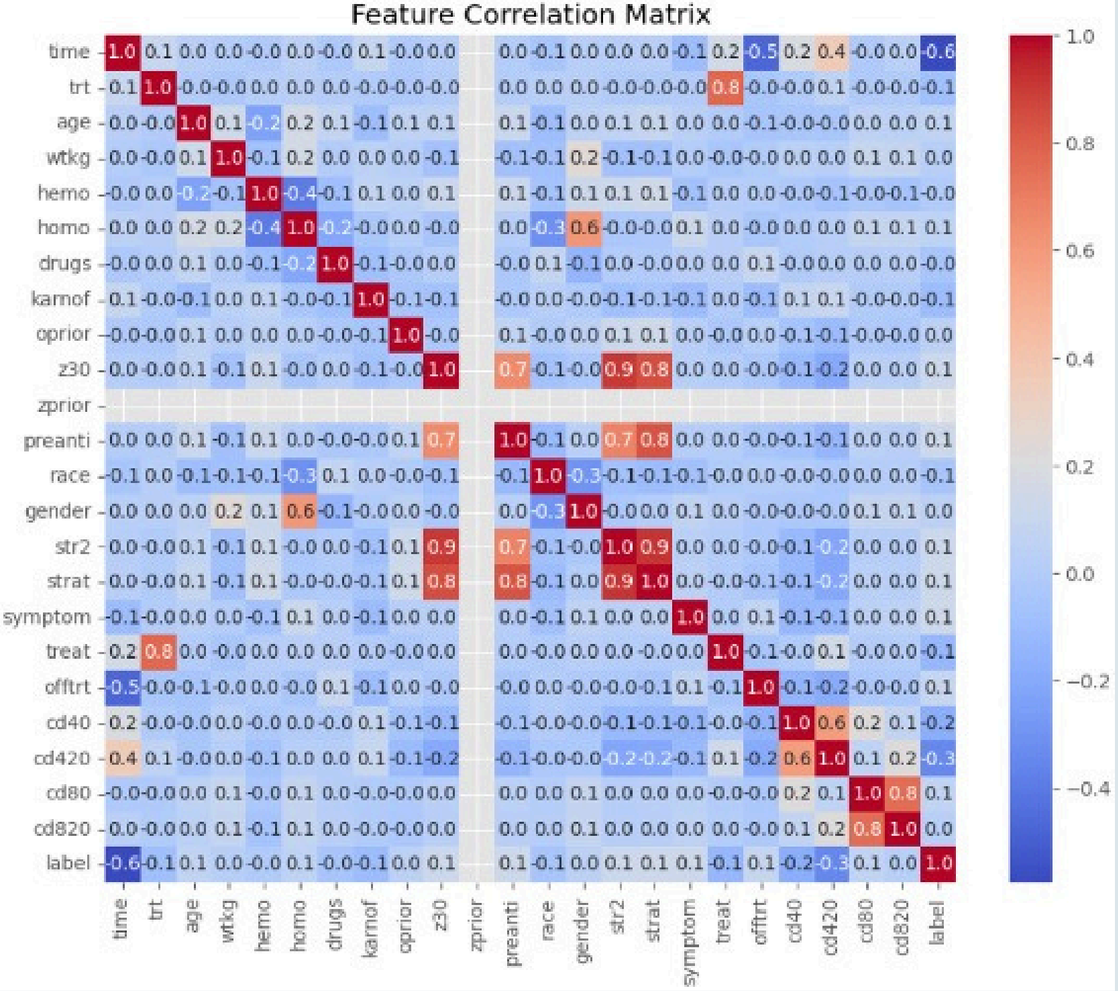


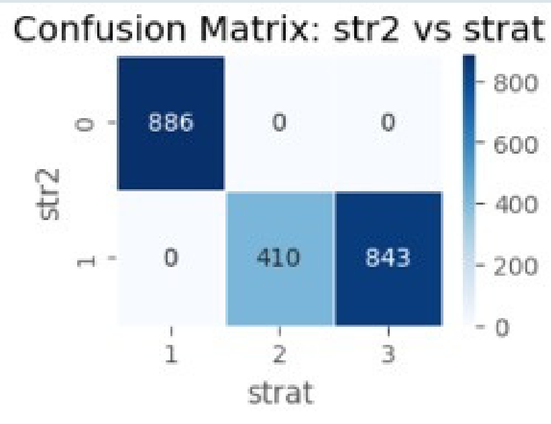
CD4 and CD8 Cell Analysis

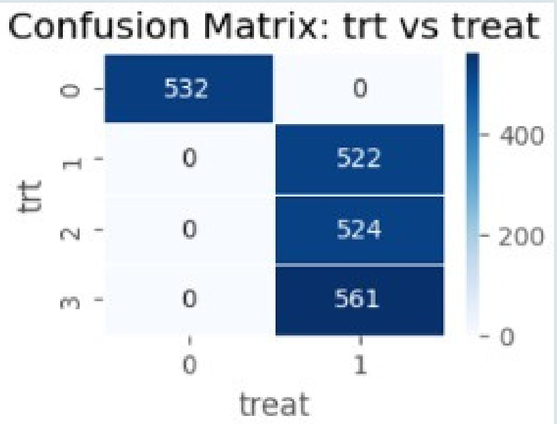
- **CD4 cells**: Infected and destroyed by HIV; lower counts indicate weaker immunity
- **CD8 cells**: Attack infected cells; higher counts generally reflect stronger response
- On average, **CD4 counts declined**, but some patients CD4 counts increased
- **CD8 counts increased slightly**, though gains were modest
- These immune metrics are likely to be essential predictors of survival











Data Preprocessing

To prepare data for modeling, we:

- **Dropped columns** leaking target info (time)
- Removed features that were duplicates of other features (str2, treat), see confusion matrices
- Removed constant-value columns (zprior)
- Encoded treatment groups with **one-hot encoding**
- Retained strat (antiretroviral history) as ordinal
- Applied **standardization** to continuous variables to suit models like logistic regression

We used a classic 70% / 30% train and test dataset split, making sure that the distribution of the label was the same in both train and test sets.

Predictive Task and Model Choice

>> Our objective is to model a binary outcome: did the patient survive during the study or not? We chose three diverse but complementary models:

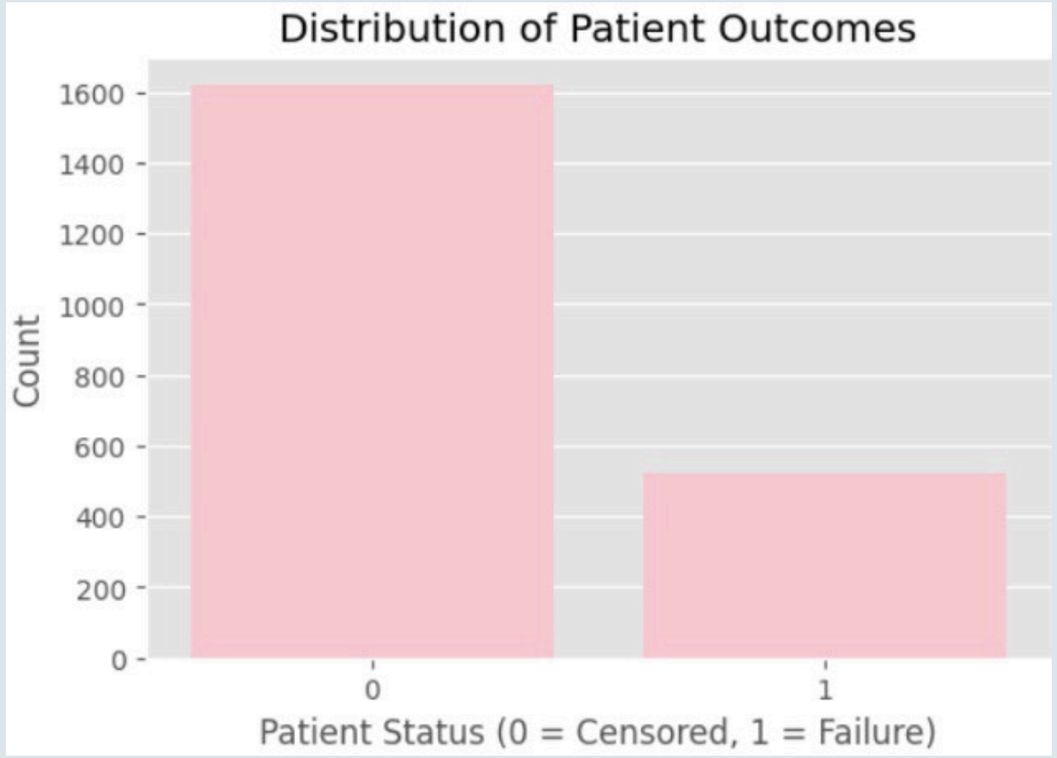
Logistic Regression – for simplicity and clinical interpretability

Random Forest – for handling non-linear interactions and feature importance

Support Vector Machine (SVM) – for robust, high-dimensional classification

Handling Class Imbalance

- **Only ~24% of the dataset represents patient deaths. This could skew predictive models.**
- **Class weighting**: Used for all models to give more importance to underrepresented outcome
- **SMOTE**: Tested with Random Forest but did not outperform class weighting
- Balanced approaches helped improve recall on the minority class without sacrificing overall accuracy



Comparison to unweighted case: We also tried each of our three models without using any method to handle class imbalance. In these cases, the models generally predicted the majority class more accurately, but the minority class much less accurately than when using class weighting.

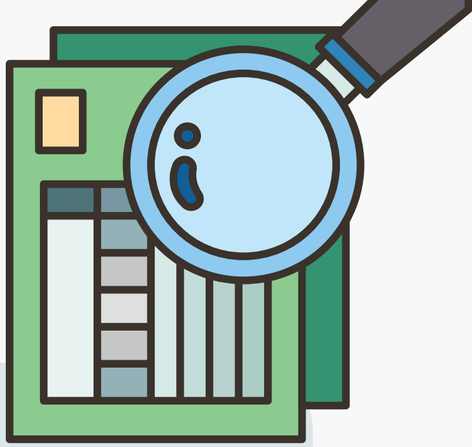
Evaluation Metrics

We selected these evaluation metrics:

- **Balanced Accuracy** as the main metric – addresses class imbalance by averaging true positive and true negative rates
- **Recall (Sensitivity)** – especially important in healthcare where **false negatives** (missing at-risk patients) can be dangerous
- **Precision** and **F1 Score** supported further evaluation

As a **baseline to compare our models to**, we used a dummy classifier that always predicts survival (the majority class). This is a reasonable baseline for an imbalanced classification problem.

Feature Selection and Model Tuning



We implemented:

- **Forward stepwise feature selection** to add predictors that improved model performance → 7-8 features remaining in each model, see next slide
- **Grid search** for hyperparameter tuning, see final chosen values in parentheses:
  - Logistic Regression: C i.e. inverse of regularization strength (100)
  - Random Forest: min samples to split (10), number of estimators (300), min samples to be at a leaf node (4)
  - SVM: kernel type (rbf), C (1), gamma i.e. kernel coefficient (auto)
- We prioritized features and hyperparameters contributing to improved **balanced accuracy**.

Selected Features & Analysis

For logistic regression:

- Gender, Indicator of getting off treatment, ZDV use, ddI use, Zal use, Karnofsky score, CD4 count at 20 weeks, CD8 count at 20 weeks

For Random Forest:

- Race, Indicator of antiretroviral therapy before study, Indicator of ZDV use before study, Karnofsky score, Indicator of symptom status, CD4 count at 20 weeks, CD8 count at baseline

For SVM:

- Indicator of antiretroviral therapy before study, ZDV use, ddI use, Karnofsky score, CD4 count at 20 weeks, CD80 count at 20 weeks

- **CD4 count at 20 weeks selected for all models** – immune strength, biologically central to HIV outcomes
- **Karnofsky score selected for all models** – represents overall patient functioning
- Prior antiretroviral use – reflects drug resistance or effectiveness history
- ZDV, ddI and Zal use – represents the treatment group the patient was in

Results and Model Comparison

Cross validation was used to make sure that the differences are not due to only sampling variance.

Model	Balanced Accuracy	Recall (Death)	Precision (Death)	F1 Score
Dummy Classifier	50%	0%	–	–
Logistic Regression	~66%	**~66%**	~40%	~49%
Random Forest	~65%	~51%	**~47%**	~48%
SVM Classifier	**~68%**	~61%	46%	**~52%**

All models significantly outperformed the baseline, but none of them are able to achieve very high scores. The three models achieve similar balanced accuracy, and it's difficult to say which one is best.

Additional Insights: Cost-Benefit Analysis

We used cost-benefit analysis to calculate the expected values of our models. False positives (failing to identify patients who will die) were heavily penalized, as it is the worst-case scenario. Incorrectly categorizing patients who will survive had a light penalty, as it leads to unnecessary resource allocation, but is not regarded as such a grave problem.

Model	Expected value	Value per patient
Logistic Regression	-14160	-22
Random Forest	-28260	-44
SVM Classifier	-17200	-27

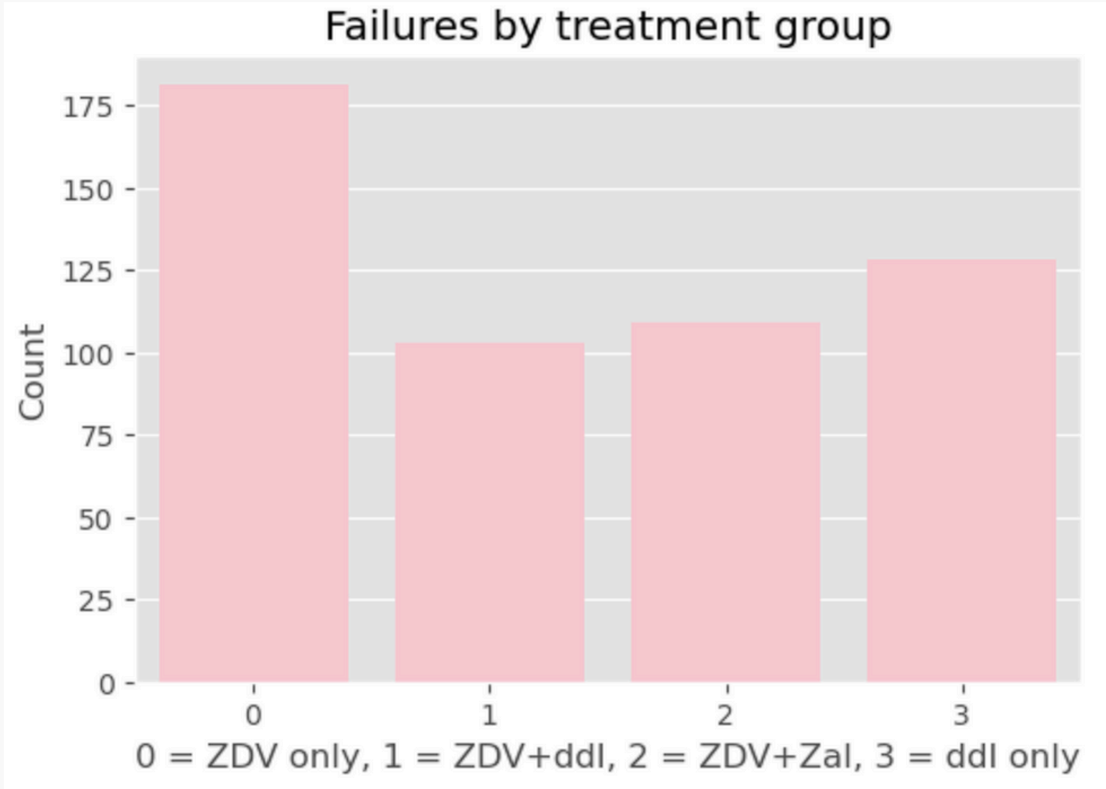
```
benefit_TP = 100
cost_FP = -20
cost_FN = -500
benefit_TN = 0
```

Controlled Setup Context & Interpretation

Our approach mimics a controlled clinical study: randomized treatment arms, consistent baseline measurements, and no missing values. This enables semi-experimental inference — like concluding that low CD4 counts predict higher risk, not just correlate.

Despite using interpretable and robust models, best performance hovered near 68% balanced accuracy. This suggests:

- Predicting survival is inherently difficult with baseline data alone
- There's value in adding dynamic metrics (e.g., time-series lab data) or genetic markers in future models, or perhaps trying neural networks



Use of Generative AI

Use of AI tools in this assignment:

- Generative AI was used to aid in writing the code to generate some of the figures displayed in our presentation.
  - Our group was not very familiar with the Matplotlib and Seaborn visualization libraries before, so AI helped in making the plots good-looking and easy to read.
- Generative AI also aided in some general debugging of our code.
- AI was used for helping with writing (e.g. grammar)