UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Universidad Distrital Francisco José de Caldas

Computer Engineering Program
School of Engineering

Technical Report – Kaggle Project Analysis: CIBMTR Equity in Post-HCT Survival Predictions

Sergio Nicolás Mendivelso Martínez – 20231020227
Sergio Leonardo Moreno Granado – 20242020091
Juan Manuel Otálora Hernández – 20242020018
Juan Diego Moreno Ramos – 20242020009

Professor: Eng. Carlos Andrés Sierra, M.Sc.

A report submitted in Systems Analysis & Design
Semester 2025-III

October 2025, Bogotá D.C.

Abstract

This technical report presents a comprehensive system design for equitable post hematopoietic cell transplantation (HCT) survival prediction as part of the CIBMTR Kaggle competition, addressing the critical challenge of developing machine learning models that achieve both high predictive accuracy and fairness across diverse demographic groups. Our proposed solution employs a modular seven-component architecture integrating data preprocessing, equity analysis, feature selection, ensemble predictive modeling, fairness calibration, uncertainty quantification, and interpretable output generation, leveraging advanced survival analysis techniques (Cox proportional hazards models, gradient boosting machines) combined with fairness-aware algorithms (AIF360, Fairlearn) to handle the inherent complexity, sensitivity, and chaotic behavior of post-transplant medical outcomes. The system design successfully addresses the dual requirements of achieving a minimum stratified C-index of 0.70 while ensuring consistent performance across racial and socioeconomic subgroups, with explicit mechanisms for bias detection, uncertainty quantification through confidence intervals, and model interpretability via SHAP values, contributing to improved clinical decision-making, reduced healthcare disparities, and advancement of personalized medicine in transplant care.

Keywords: Hematopoietic Cell Transplantation, survival Prediction, Healthcare Equity, CIBMTR, Fairness Calibration, Ensemble Modeling.

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview and Context

The CIBMTR (Center for International Blood and Marrow Transplant Research) competition focuses on predicting survival after a medical procedure called hematopoietic cell transplantation, or HCT. This procedure is used to treat serious blood diseases like leukemia, lymphoma, and other blood disorders. The competition asks participants to build machine learning models that can predict how long patients will survive after receiving a transplant.

What makes this competition different from other prediction challenges is that it has two important goals. First, the model must make accurate predictions. Second, the model must be fair to all patients, regardless of their race or background. Many medical AI systems today work well for some patient groups but not for others. This competition tries to fix that problem by requiring models that work equally well for everyone.

The dataset from CIBMTR contains detailed information about transplant patients. This includes patient age, ethnicity, disease information, genetic matching between donor and patient, other health problems the patient has, and treatment details. The competition only allows using this provided data - participants cannot add information from other sources. Also, the evaluation uses a special metric called stratified C-index, which measures both accuracy and fairness together.

This technical report explains our approach to building a system that meets both requirements. We designed a system with multiple components that work together to make predictions that are both accurate and fair. The system checks for bias, explains its predictions, and shows how confident it is about each prediction.

1.2 Objectives

This technical report has several main goals that guide our work:

1.2.1   Primary Objectives

1. **Understand the medical system**: Learn about hematopoietic cell transplantation, including how it works, what can go wrong, and what factors affect patient survival.

2. **Design a fair and accurate system**: Create a prediction system that makes accurate survival predictions while treating all patient groups fairly.

3. **Identify system parts and connections**: Map out all the pieces of the prediction problem - what data we have, how different factors relate to each other, and what limits we face.

1.2.2   Secondary Objectives

1. **Study complexity and sensitivity**: Understand how small changes in patient characteristics can lead to big differences in outcomes, and identify which factors matter most.

2. **Understand the fairness metric**: Learn how the stratified C-index works and what it means for our model to perform fairly across different groups.

3. **Make practical recommendations**: Suggest specific ways to improve prediction models in transplant medicine based on what we learn.

4. **Document how to build it**: Provide clear instructions about what tools and methods to use, so others can build similar systems.

1.2.3   Healthcare Impact Objectives

1. **Help doctors make decisions**: Create outputs that help transplant doctors decide which patients should receive transplants and how to treat them.

2. **Reduce healthcare inequalities**: Build methods to find and fix bias in predictions so all patients get fair treatment.

3. **Improve personalized medicine**: Show how engineering and machine learning can help doctors give each patient the right treatment for their specific situation.

These objectives work together to make sure our system is not just technically good, but also useful and fair in real healthcare.

1.3   Medical Background

1.3.1   Allogeneic Hematopoietic Cell Transplantation

Allogeneic Hematopoietic Cell Transplantation, or HCT, is a medical procedure where a patient receives healthy blood stem cells from another person (the donor). These stem

cells replace the patient's damaged or diseased bone marrow. The new cells then create a healthy blood and immune system.

The word "allogeneic" means the stem cells come from another person, not from the patient themselves. This donor can be a family member (like a brother or sister) or a volunteer who matches the patient's genetics.

HCT is used to treat serious diseases that affect blood and the immune system:

- **Leukemias**: Cancers of blood cells, both fast-growing (acute) and slow-growing (chronic)

- **Lymphomas**: Cancers of the immune system

- **Multiple myeloma**: Cancer of plasma cells

- **Severe aplastic anemia**: When bone marrow stops making enough blood cells

- **Immune system disorders**: When the immune system does not work properly

- **Some inherited diseases**: Certain genetic conditions affecting blood

The goal is to replace sick blood cells with healthy ones. These new cells can make red blood cells (carry oxygen), white blood cells (fight infections), and platelets (help blood clot).

Transplant success has improved over the years, but the procedure is still risky. Many things can go wrong, and predicting which patients will survive is difficult.

1.3.2   The Transplantation Process

The transplant process has five main steps:

Step 1: Finding a Donor and Checking Compatibility

Before transplant, doctors must find a donor whose genes match the patient's genes. They check proteins called HLA (Human Leukocyte Antigens) that are found on cells. Better matching means lower risk of problems.

Donors are chosen in this order:

- Brothers or sisters with matching HLA

- Unrelated volunteers with matching HLA (found through registries)

- Partially matched family members (like parents)

- Umbilical cord blood

Step 2: Preparing the Patient (Conditioning)

Before receiving new cells, patients get strong treatments called conditioning. This treatment has several purposes:

- Kill diseased cells

- Make space in the bone marrow for new cells

- Weaken the patient's immune system so it will not reject the donor cells

There are different levels of conditioning:

- **Myeloablative**: Very strong doses of chemotherapy and sometimes radiation

- **Reduced-intensity**: Lower doses to reduce side effects

- **Non-myeloablative**: Minimal doses, mainly to suppress the immune system

Step 3: Receiving the Stem Cells

The actual transplant is not surgery. The donor's stem cells are given through an IV, similar to a blood transfusion. The cells travel through the bloodstream to the bone marrow, where they settle and start growing.

Step 4: Waiting for Engraftment

After receiving the cells, there is a waiting period of 2-4 weeks called engraftment. During this time, the donor cells start making new blood cells. This period is dangerous because the patient has almost no immune system and cannot fight infections.

Step 5: Follow-up Care

After engraftment, patients need close monitoring for months or years. Doctors check for complications, manage medications, and watch for signs of problems. Regular blood tests track how well the new cells are working.

1.3.3 Risks and Complexity

Because HCT involves replacing the patient's immune system with someone else's, many serious problems can happen:

Graft-versus-Host Disease (GVHD)

This is the most common serious complication. It happens when the donor's immune cells attack the patient's body tissues. There are two types:

- **Acute GVHD**: Happens in the first 100 days, affects mainly skin, liver, and intestines

- **Chronic GVHD**: Happens after 100 days, can affect many organs and last a long time

Graft Failure

Sometimes the donor cells fail to grow properly in the patient. This can happen right away (primary failure) or later (secondary failure). Without working donor cells, the patient cannot make blood cells and may die.

Infections

Patients are very vulnerable to infections because their immune system is weak or rebuilding. They can get:

- Bacterial infections

- Viral infections (including viruses that were dormant in their body)

- Fungal infections

- Infections from organisms that do not usually cause disease in healthy people

Organ Damage

The strong treatments can damage organs:

- Liver damage

- Lung damage

- Kidney damage

- Heart problems

Social and Economic Factors

Medical factors are not the only things that affect outcomes:

- Distance from the transplant center

- Ability to pay for treatment and medications

- Having family or friends to help with care

- Health insurance coverage

These many risk factors make prediction very difficult. Small differences between patients can lead to very different outcomes.

1.3.4   Importance of Survival Prediction

HCT is one of the most expensive and intensive medical procedures. Accurate survival predictions are important for several reasons:

Helping Doctors Make Decisions

Predictions help doctors:

- Decide if a patient should get a transplant or try other treatments

- Choose the right time to do the transplant

- Select how strong the conditioning treatment should be

- Give patients realistic information about their chances

Personalizing Treatment

With good predictions, doctors can:

- Give stronger prevention medications to high-risk patients

- Monitor high-risk patients more carefully

- Change treatments based on each patient's specific risks

Using Healthcare Resources Wisely

Predictions help hospitals:

- Prioritize which patients need donor searches most urgently

- Plan for intensive care unit beds

- Coordinate long-term care needs

Reducing Healthcare Inequalities

This is especially important for this competition. Current prediction models often work better for some groups than others. Research shows that:

- Models are often less accurate for racial and ethnic minorities

- Some groups get incorrectly labeled as higher or lower risk

- Training data often has more examples from majority populations

By requiring fair predictions, this competition addresses the problem that medical AI might treat different groups unfairly. All patients deserve accurate predictions regardless of their background.

1.4 Scope and Boundaries

1.4.1 What Our System Includes

Our system includes:

Data We Use

- The CIBMTR dataset with clinical, genetic, and demographic information

- Disease characteristics and patient status before transplant

- Transplant details (donor type, genetic matching, stem cell source, conditioning)

- Patient background information (age, ethnicity, income level)

- Follow-up information about outcomes

System Parts

- Data preprocessing (cleaning and preparing data)

- Equity analysis (checking for bias)

- Feature selection (choosing important variables)

- Prediction models (making survival predictions)

- Fairness correction (adjusting to ensure fairness)

- Uncertainty measurement (showing how confident predictions are)

- Output creation (making results understandable)

Performance Goals

- Stratified C-index of at least 0.70 for all demographic groups

- Similar accuracy across all racial and ethnic groups

- Predictions that doctors can understand and trust

- Uncertainty information for each prediction

1.4.2 What Our System Does Not Include

Some things are outside our scope:

Not Included in Our System

- Connecting to hospital computer systems (could be added later)

- Making predictions in real-time while patients are being treated (future work)

- Interfaces for patients to see predictions (future work)

- Using data from other sources beyond the competition dataset

- Getting regulatory approval for clinical use

- Economic analysis of costs and benefits

- Changing treatment protocols based on predictions

1.4.3 Important Limits

Several constraints affect our system design:

Competition Rules

- We can only use the CIBMTR dataset provided

- We cannot add data from other sources

- The system must be fair across demographic groups

- Evaluation uses stratified C-index metric

Data Challenges

- Many variables have missing values

- Missing data patterns might differ between groups

- We cannot get additional information even if it might help predictions

- Data covers multiple years when treatment practices changed

Clinical Requirements

- Predictions must make medical sense

- Doctors need to understand why the model makes certain predictions

- Selected features must align with medical knowledge

- Biology has random elements that limit perfect prediction

Computational Limits

- The solution must run within reasonable time

- Must handle the full dataset efficiently

- Results must be reproducible

Ethical Requirements

- Cannot discriminate against any group

- Decision process must be transparent and explainable

- Must actively look for and reduce bias

These limits shape how we design our system and what we can accomplish.

Chapter 2

Literature Review

2.1 Previous Research on Post-HCT Survival Prediction

Survival prediction after hematopoietic cell transplantation has been studied for many years. Researchers have worked to find which factors most affect patient outcomes and how to predict survival accurately.

Traditional approaches to post-HCT survival prediction used simple statistical models. The Cox Proportional Hazards model became the standard tool in this field. This model helps doctors understand how different factors affect survival time. Early research identified basic risk factors like patient age, disease type, and donor compatibility as important predictors.

Recent studies show that prediction models have become more complex. Auletta et al. (2020) reviewed current guidelines for hematopoietic cell transplantation and found that decision-making now involves many interconnected factors. Their work highlights how medical understanding has grown beyond simple risk categories to include detailed clinical characteristics, genetic markers, and treatment protocols.

Advanced computational modeling is now being used to improve transplant delivery. Harrington et al. (2025) explain how computational approaches can capture the complexity of transplant outcomes. Their research shows that traditional statistical methods sometimes fail to capture the nonlinear relationships between patient characteristics and survival outcomes.

Modern research also focuses on immune system recovery after transplantation. Kucab et al. (2024) studied how blood cell populations change over time in transplant patients. Their work shows that small differences in how the immune system rebuilds can lead to very different outcomes. This demonstrates the sensitivity and complexity that prediction models must handle.

Studies on transplant outcomes over time reveal important trends. Zubarovskaya et al. (2023) analyzed 5,000 transplantations over 30 years from a single center. They found that survival rates improved significantly, but the patterns became more complex as treatment protocols evolved. This temporal complexity creates challenges for prediction models that

must account for changing medical practices.

2.2 Fairness and Equity in Medical AI

The problem of fairness in medical artificial intelligence has gained attention in recent years. Researchers and healthcare providers recognize that AI systems can create or worsen health inequalities if fairness is not explicitly considered.

Doherty et al. (2024) provide a comprehensive review of algorithmic bias in healthcare. Their research shows that many existing medical AI systems perform differently across demographic groups. Models often work well for majority populations but fail for under-represented groups. This happens because training data often contains more examples from certain populations, and models may learn patterns that do not generalize fairly.

Current approaches to addressing bias in medical AI include several strategies. Some researchers focus on fairness-aware preprocessing, which adjusts training data to ensure better representation. Others develop fairness-aware algorithms that explicitly optimize for equity during model training. Post-processing approaches adjust model outputs to ensure fair predictions across groups.

The concept of equity in medical predictions goes beyond simple statistical fairness. True equity requires that models provide similar accuracy for all patient populations, not just similar prediction distributions. This distinction is important in transplant medicine, where inaccurate predictions for certain groups could lead to poor treatment decisions.

Tools for measuring and improving fairness have been developed by major organizations. IBM's AI Fairness 360 (AIF360) toolkit provides metrics and algorithms for detecting and reducing bias. Microsoft's Fairlearn library offers similar capabilities. These tools enable developers to quantify fairness and apply corrections when needed.

The stratified evaluation approach used in the CIBMTR competition reflects current best practices in fair medical AI. By measuring performance separately within demographic groups, this approach ensures that models cannot achieve high overall accuracy while performing poorly for specific populations.

2.3 Machine Learning Approaches in Survival Analysis

Machine learning methods for survival analysis have evolved significantly beyond traditional statistical approaches. These methods can capture complex patterns that simpler models miss.

2.3.1 Cox Proportional Hazards Models

Cox Proportional Hazards models remain foundational in survival analysis. This statistical approach assumes that the effect of variables on survival remains proportional over time. While interpretable and well-understood clinically, Cox models have limita-

tions when relationships between variables are nonlinear or when proportional hazards assumptions are violated.

2.3.2 Ensemble Methods

Ensemble methods combine multiple models to improve predictions. Random Forests create many decision trees and average their predictions. Gradient Boosting Machines (GBMs) build trees sequentially, with each new tree correcting errors from previous trees. These methods can capture complex patterns and interactions between variables that single models miss.

2.3.3 Advanced Boosting Libraries

XGBoost, LightGBM, and CatBoost represent recent advances in gradient boosting. XGBoost introduced regularization techniques that reduce overfitting and improve generalization. LightGBM uses efficient data structures that enable faster training on large datasets. CatBoost handles categorical variables directly without requiring manual encoding, reducing preprocessing complexity and potential errors.

2.3.4 Deep Learning for Survival Analysis

Deep learning applies neural networks to survival prediction. These models can learn highly nonlinear relationships from data. However, they typically require large amounts of training data and can be difficult to interpret, which creates challenges in medical applications where explainability is important.

2.3.5 Survival Analysis Libraries

Survival analysis libraries make these methods accessible. The Lifelines library in Python implements Cox models and other standard survival analysis methods. Scikit-survival extends scikit-learn's machine learning capabilities to survival analysis, enabling easy application of random forests and gradient boosting to time-to-event data.

Research comparing these approaches shows that ensemble methods often provide better predictions than traditional Cox models, especially when relationships are complex. However, the best approach depends on specific characteristics of the data and the prediction task.

2.4 Model Interpretation and Explainability

Understanding why models make certain predictions is critical in medical applications. Doctors need to know which factors drive predictions to make informed treatment decisions and to trust model outputs.

2.4.1 SHAP Values

SHAP (SHapley Additive exPlanations) values have become a standard tool for model interpretation. SHAP values show how much each feature contributes to a prediction for a specific patient. This approach works with any machine learning model and provides consistent, theoretically grounded explanations.

Feature importance rankings show which variables matter most across all predictions. However, global feature importance can hide important patterns—a variable might be very important for certain patient types but not others. SHAP values address this by providing both global and patient-specific explanations.

2.4.2 Visualization Techniques

Partial dependence plots visualize how predictions change as one variable changes while holding others constant. These plots help clinicians understand relationships between patient characteristics and predicted outcomes.

Calibration plots show whether predicted probabilities match actual observed outcomes. A well-calibrated model that predicts 70% survival should see approximately 70% of those patients survive. Calibration is especially important when predictions guide treatment decisions.

2.4.3 Fairness and Interpretability Connection

Research shows that interpretability and fairness are connected. Models that are explainable make it easier to identify potential sources of bias. If a model heavily weights features that are measured differently across demographic groups, this becomes visible through interpretation tools and can be addressed.

2.5 Complexity and Uncertainty in Medical Predictions

Medical outcomes involve inherent uncertainty that prediction models must acknowledge. Predicting survival after HCT is particularly challenging because many interconnected factors influence outcomes.

2.5.1 Chaos Theory in Medical Outcomes

Chaos theory describes how small differences in patient characteristics can lead to very different outcomes. Harrington et al. (2025) discuss how computational modeling must account for this sensitivity. Even patients who appear very similar may experience different outcomes due to unmeasured factors or biological randomness.

2.5.2 Uncertainty Quantification

Uncertainty quantification provides confidence intervals around predictions rather than single point estimates. This helps clinicians understand prediction reliability. Some patients may have highly certain predictions based on clear risk factors, while others have more uncertain outcomes because their characteristics place them in ambiguous risk categories.

2.5.3 Sensitivity Analysis

Sensitivity analysis identifies which variables most affect predictions. In post-HCT prediction, research shows that patient age, disease risk indices, genetic compatibility, and comorbidities are highly sensitive parameters. Small measurement errors or variations in these factors can significantly change predictions.

Studies of immune system recovery after transplantation illustrate this complexity. Kucab et al. (2024) found that engraftment patterns show significant variation even among similar patients. This biological variability creates fundamental limits on prediction accuracy.

2.5.4 Feedback Effects and Dynamic Interactions

The presence of feedback effects adds another layer of complexity. Immune responses to transplanted cells can trigger complications that require treatment, which in turn affects survival. These dynamic interactions cannot be fully captured by models that treat variables as independent.

2.6 Gaps in Current Knowledge

Despite advances in survival prediction and fair AI, several important gaps remain in current research and practice.

2.6.1 Integration of Accuracy and Equity

Most research focuses either on improving prediction accuracy or on ensuring fairness, but rarely on both simultaneously. The CIBMTR competition explicitly requires both, representing an important advance. However, techniques for optimizing both objectives together remain underdeveloped.

2.6.2 Handling Missing Data Fairly

Missing data patterns often differ across demographic groups, creating challenges for equitable prediction. Current imputation methods may inadvertently introduce or amplify bias if missingness correlates with sensitive attributes. Research on fairness-aware missing data handling is limited.

2.6.3 Temporal Dynamics

Medical practices evolve over time, but most prediction models treat all historical data equally. Methods that appropriately weight recent data while learning from historical patterns are needed, especially for rare events where historical data is valuable despite practice changes.

2.6.4 Uncertainty Across Populations

Most uncertainty quantification methods focus on overall prediction uncertainty but do not explicitly ensure that uncertainty is estimated accurately across demographic groups. Miscalibrated uncertainty estimates could lead to different treatment decisions for similar patients from different groups.

2.6.5 Clinical Implementation

While many sophisticated prediction models exist in research settings, few are successfully implemented in clinical practice. Gaps between research models and practical tools include computational requirements, integration with clinical workflows, and provider trust and understanding.

2.6.6 Cross-Institutional Generalization

Models trained at one transplant center may not perform well at others due to differences in patient populations, protocols, or data collection practices. Research on developing models that generalize across institutions while preserving fairness is limited.

These gaps highlight opportunities for advancing both the science of medical prediction and its practical application to improve patient outcomes equitably. |

Chapter 3

Background

3.1 The CIBMTR Dataset

3.1.1 Overview of the Dataset

The CIBMTR dataset is a comprehensive collection of information about patients who have received hematopoietic cell transplants. The dataset contains detailed medical, genetic, demographic, and outcome information for thousands of transplant patients. This rich data allows researchers to build prediction models that understand which patient factors most affect survival after transplant.

The dataset comes from the Center for International Blood and Marrow Transplant Research, a large organization that collects transplant data from many hospitals and transplant centers around the world. Because data comes from many different centers, the dataset represents diverse patient populations and different transplant practices.

3.1.2 Data Structure and Organization

The CIBMTR dataset is organized into different types of information:

Disease and Clinical Characteristics

This section contains information about the disease being treated:

- **Disease type**: What disease does the patient have? Examples include acute leukemia, chronic leukemia, lymphoma, multiple myeloma, or aplastic anemia

- **Disease stage**: How advanced is the disease? Is the patient in remission (disease is under control) or has the disease returned?

- **Previous treatments**: What treatments has the patient received before transplant? How many rounds of chemotherapy? Has the patient had radiation?

- **Pre-transplant health**: What other medical problems does the patient have? For example, does the patient have kidney disease, liver disease, or heart problems?

- **Performance scores**: How well can the patient function? Can they walk around and do normal activities?

- **Laboratory values**: Blood test results showing how well different organs are working

Transplant-Specific Information

This section describes details about the transplant procedure:

- **Donor type**: Where does the stem cell donor come from? Options include a matched brother or sister, an unrelated volunteer from a registry, a partially matched family member, or umbilical cord blood

- **HLA matching**: How well do the donor's genes match the patient's genes? Perfect matches are better than partial matches

- **Source of cells**: Where were the stem cells collected from? Bone marrow, peripheral blood, or umbilical cord blood

- **Conditioning regimen**: What treatment does the patient receive to prepare for transplant? Strong chemotherapy, weak chemotherapy, or radiation?

- **GVHD prevention**: What medications does the patient receive to prevent the donor cells from attacking their body?

Demographic and Socioeconomic Information

This section includes patient background information:

- **Age**: How old is the patient? Age is one of the most important factors affecting transplant success

- **Sex and gender**: Patient gender identity

- **Race and ethnicity**: Patient racial and ethnic background. This is crucial for our fairness analysis

- **Geographic location**: Where does the patient live? Distance from transplant center affects access to care

- **Insurance status**: What type of insurance does the patient have? This affects ability to pay for treatment

Temporal Information

This section contains time-related information:

- **Year of transplant**: What year did the transplant happen? Medical practices change over time

- **Time from diagnosis to transplant**: How long after diagnosis did the patient receive the transplant?

- **Follow-up duration**: How long after transplant did doctors follow the patient? Did the patient survive the entire follow-up period?

3.2 Evaluation Metrics

3.2.1 The C-Index Metric

The C-Index is a number that measures how good a survival model is at predicting which patients will live longer. The metric ranges from 0.5 to 1.0:

- **0.5**: The model is no better than random guessing

- **1.0**: The model is perfect - it correctly predicts survival for every patient pair

- **0.7 or higher**: Good prediction accuracy

The C-Index works by comparing pairs of patients. For each pair of patients, if one survives longer than the other, a good model should predict that the longer-surviving patient has higher survival probability. The C-Index counts what percentage of patient pairs the model gets correct.

For example, suppose we have two patients: Patient A dies after 6 months and Patient B dies after 2 years. A good model should predict higher survival probability for Patient B than for Patient A. If it does, the model gets that pair correct.

3.2.2 Stratified C-Index for Fairness

The standard C-Index measures overall accuracy, but it does not tell us if the model is fair. A model could work well for most patients but fail for certain groups like minority patients.

The stratified C-Index fixes this problem. Instead of calculating one C-Index for all patients, we calculate the C-Index separately for each demographic group (by race, ethnicity, or other factors). Then we average these separate scores.

This approach ensures that a model cannot achieve high overall accuracy by performing well for majority groups while performing poorly for minority groups. Each demographic group's accuracy counts equally in the final score.

For example:

- C-Index for Group A (50% of patients): 0.72

- C-Index for Group B (30% of patients): 0.68

- C-Index for Group C (20% of patients): 0.65

- Stratified C-Index: $(0.72 + 0.68 + 0.65) / 3 = 0.68$

Even though Group A has many patients and high accuracy, the stratified C-Index penalizes poor performance in smaller groups. This forces models to predict well for all populations, not just the majority.

3.2.3 Importance of the Stratified C-Index

The stratified C-Index is crucial because it ensures our model is fair. Medical history shows that prediction models often work better for some patient groups than others. By using stratified C-Index, the CIBMTR competition requires models to be accurate for all patients.

This is important for several reasons:

1. **Patient equity**: Every patient deserves accurate predictions about their survival, regardless of their race or background

2. **Clinical decision-making**: Doctors need to trust that predictions are fair when making treatment recommendations

3. **Reducing health disparities**: Fair prediction models can help reduce existing inequalities in healthcare

4. **System accountability**: The stratified metric makes it transparent whether a model is treating all patients fairly

3.3 Competition Rules and Constraints

3.3.1 Data Restrictions

The competition has strict rules about what data can be used:

Only Provided Data

Participants can only use the CIBMTR dataset that the competition provides. You cannot add data from other sources like:

- Data from published research papers

- Data from other hospitals or transplant centers

- Publicly available databases with transplant information

- External genetic databases

This restriction ensures that all competitors have the same information and that solutions can be fairly compared.

Data De-identification

The data provided is de-identified, meaning patient names and personal identifying information have been removed. The dataset contains only medical information needed to build the model.

3.3.2 Fairness Mandate

The competition explicitly requires fairness:

- Models must meet a minimum stratified C-Index of 0.70

- Performance must be similar across all demographic groups

- Bias detection and correction are required, not optional

- Solutions are evaluated on both accuracy and fairness equally

3.3.3 Model Submission and Evaluation

What to Submit

Participants submit:

- Predicted survival probabilities for each patient in the test set

- Information about model structure and methods used

- Documentation of how fairness was addressed

Evaluation Process

The competition evaluates submissions by:

1. **Calculating stratified C-Index**: Computing C-Index separately for each demographic group

2. **Averaging across groups**: Computing the overall stratified C-Index

3. **Checking fairness**: Verifying that performance is similar across groups

4. **Ranking solutions**: Models are ranked by stratified C-Index score

3.3.4 Timeline and Deadlines

The competition operates according to a specific schedule:

- Training data is provided to participants

- Participants have several months to develop and test models

- A test set is released, and participants make predictions

- Submissions are evaluated, and winners are announced

3.4 Related Competitions and Context

3.4.1 Why This Competition Matters

The CIBMTR competition represents an important advance in machine learning and healthcare. Many competitions focus on accuracy alone. This competition emphasizes accuracy AND fairness together.

This is significant because:

- **Healthcare importance**: HCT is a serious procedure affecting patient survival. Better predictions save lives

- **Fairness innovation**: Few competitions explicitly require fairness. This sets a new standard for medical AI

- **Real-world impact**: Unlike some competitions with artificial data, this competition uses real patient information that can directly improve medical care

- **Diversity of problems**: Post-HCT prediction is complex, involving medical knowledge, machine learning, and ethics

3.4.2 Broader Context: Fairness in Medical AI

Healthcare is increasingly using machine learning. However, research shows that many medical AI systems are unfair. Examples include:

- Models that work well for white patients but not minority patients

- Prediction systems that systematically underestimate risk for certain groups

- Algorithms that reflect historical biases in medical data

The CIBMTR competition directly addresses these problems by requiring and evaluating fairness alongside accuracy.

Chapter 4

System Analysis

4.1 System Components and Relationships

4.1.1 Overview of the Prediction System

The HCT survival prediction system is complex because many different factors work together to affect patient outcomes. The system takes in patient information, processes it through several steps, and produces survival predictions along with fairness metrics.

The system can be divided into distinct parts that work together:

1. **Input sources**: All the patient data that comes into the system

2. **Processing modules**: Steps that prepare and analyze the data

3. **Modeling**: The machine learning algorithms that make predictions

4. **Outputs**: The predictions, explanations, and fairness reports

4.1.2 System Inputs and Their Characteristics

The system receives several types of input data:

Clinical Inputs

- **Disease type**: What disease does the patient have?

- **Disease stage**: Is the disease advanced or controlled?

- **Previous treatments**: What medical treatments has the patient received?

- **Health status**: What other medical problems does the patient have?

- **Body function scores**: How well can the patient function physically?

- **Lab results**: Blood tests and other measurements of organ function

Transplant-Related Inputs

- **Donor information**: Type of donor (related, unrelated, cord blood)

- **HLA compatibility**: How well do donor and patient match genetically

- **Cell source**: Where the stem cells come from

- **Conditioning type**: How strong is the preparation treatment

- **Prevention medications**: What drugs are used to prevent rejection

Demographic Inputs

- **Age**: Patient age at time of transplant

- **Sex**: Patient gender

- **Race and ethnicity**: Patient background (critical for fairness analysis)

- **Location**: Where patient lives

- **Insurance**: What type of insurance patient has

Temporal Inputs

- **Year of transplant**: When the transplant occurred

- **Time since diagnosis**: How long from diagnosis to transplant

- **Follow-up time**: How long the patient was monitored after transplant

4.1.3   System Outputs

The system generates several types of output:

Survival Predictions

- Probability of survival at key time points (100 days, 1 year, 2 years, 5 years)

- Risk categories (low, intermediate, high)

- Survival curves showing predicted outcomes over time

Fairness and Equity Metrics

- Performance across demographic groups

- Measures of prediction fairness

- Comparison of accuracy for different populations

Interpretability Information

- Which factors matter most for each prediction

- SHAP values explaining individual predictions

- Feature importance rankings

Uncertainty Information

- Confidence intervals around predictions

- Reliability scores for individual predictions

- Warnings for uncertain predictions

4.2 Complexity and Sensitivity Analysis

4.2.1 Why This System Is Complex

Post-HCT survival prediction is complex because:

1. **Many interconnected factors**: Hundreds of variables potentially affect survival

2. **Nonlinear relationships**: A change in one factor does not always produce a proportional change in outcomes

3. **High-dimensional data**: The system has many input variables that interact

4. **Feedback loops**: Complications trigger treatments that affect outcomes

5. **Biological variability**: Each patient's immune system reacts differently

4.2.2 Sensitivity Analysis: Small Changes, Big Effects

Small changes in patient characteristics can lead to large changes in predicted survival. Key sensitive parameters include:

Patient Age

Age is extremely sensitive. A single-year change can move a patient from one risk category to another completely different category. For example:

- A 55-year-old patient might be in the "intermediate risk" group

- A 56-year-old patient with identical other characteristics might move to "high risk"

- The model becomes much more uncertain about prediction accuracy near these thresholds

Disease Risk Index

Small changes in this score significantly affect predictions:

- A score of 2.5 might predict 65% survival probability

- A score of 2.6 might predict 55% survival probability

- A 4% change in this index can lead to a 10% change in survival prediction

Genetic Compatibility (HLA Matching)

Even small differences in how well donor and patient genes match affect outcomes:

- Perfect match: dramatically better outcomes

- One gene mismatch: noticeably worse outcomes

- Two gene mismatches: significantly worse outcomes

Comorbidities (Other Health Problems)

The presence or absence of certain health conditions can double or halve survival estimates:

- A patient with kidney disease faces much higher risk

- A patient with severe infection history faces higher risk

- These conditions sometimes determine whether transplant is even possible

4.2.3 Chaos and Randomness in Post-HCT Outcomes

Medical outcomes after HCT exhibit chaotic behavior. This means:

The Butterfly Effect

Small, unmeasured differences between apparently similar patients lead to very different outcomes. For example:

- Two patients appear identical in all measured characteristics

- Patient A recovers well from transplant

- Patient B develops severe complications

- The difference may come from factors we cannot measure or predict

Immune System Unpredictability

How a patient's immune system recovers varies greatly:

- Identical stem cell grafts produce different immune recovery patterns

- Small differences in T-cell populations lead to dramatically different outcomes

- Some patients' immune systems attack their own bodies (GVHD)

- Others accept the transplant perfectly

Infection Risk Variability

Even with identical prevention protocols, infection outcomes vary:

- Some patients fight infections easily

- Others develop life-threatening infections from the same organism

- Timing of exposure, individual immune differences, and random chance all matter

Disease Recurrence Patterns

For patients with cancer, relapse follows unpredictable patterns:

- Some remain in remission indefinitely

- Others relapse within months

- Minimal residual disease (undetected cancer) affects outcomes unpredictably

4.2.4 Implications for Prediction Models

Because of this complexity and chaos:

1. **Perfect prediction is impossible**: Even the best models cannot capture all factors affecting survival

2. **Uncertainty is real**: Our predictions have confidence intervals because of genuine biological unpredictability

3. **Ensemble methods are necessary**: Multiple models together capture more patterns than any single model

4. **Fairness calibration is critical**: The stratified C-Index approach ensures models work fairly even given this unpredictability

5. **Interpretability matters**: Clinicians need to understand what the model is doing and what it is uncertain about

4.3 Stakeholder Analysis

4.3.1 Different Users of the System

Different people use the prediction system for different purposes:

Transplant Physicians

- **Need**: Accurate survival predictions for individual patients

- **Goal**: Decide which patients should receive transplants and what treatments to give

- **Concern**: Predictions must be accurate across all patient groups (not just majority populations)

- **Use**: Help counsel patients about realistic outcomes before transplant

Clinical Researchers

- **Need**: Understanding which factors most affect survival

- **Goal**: Identify new treatment approaches or interventions

- **Concern**: Explanations must be clear and statistically sound

- **Use**: Study relationships between patient characteristics and outcomes

Healthcare Administrators

- **Need**: Proof that the system is fair across all demographic groups

- **Goal**: Ensure the hospital delivers equitable care

- **Concern**: Legal and ethical obligations to avoid discrimination

- **Use**: Quality monitoring and program evaluation

Model Developers

- **Need**: Robust evaluation frameworks to test their models

- **Goal**: Build the best possible prediction model

- **Concern**: Proving that accuracy is consistent across demographic groups

- **Use**: Validating and comparing different machine learning approaches

Patients

- **Need**: Honest information about their survival chances

- **Goal**: Make informed decisions about treatment options

- **Concern**: Information must be presented clearly and should not be overwhelming

- **Use**: Understanding their medical situation and what to expect

4.3.2 Competing Needs and Tradeoffs

Different stakeholders sometimes have competing needs:

- **Accuracy vs. Interpretability**: Complex models predict better but are harder to explain

- **Precision vs. Uncertainty**: Doctors want specific predictions, but admitting uncertainty is important

- **Population-Level Fairness vs. Individual Predictions**: A model may be fair on average but unfair for specific individuals

- **Speed vs. Thoroughness**: Quick predictions help doctors decide, but careful analysis ensures quality

The system design must balance these competing needs while prioritizing patient safety and fairness.

4.4 System Boundaries and Constraints

4.4.1 What the System Can and Cannot Do

The System Can

- Make predictions of survival probability after HCT

- Explain why it makes certain predictions

- Provide uncertainty estimates for predictions

- Identify which patients are at highest risk

- Measure fairness across demographic groups

- Support clinical decision-making with research evidence

The System Cannot (and Should Not)

- Make treatment decisions automatically (doctors must decide)

- Integrate with hospital computer systems (future enhancement)

- Make predictions in real-time during treatment (only pre-transplant)

- Use data beyond what CIBMTR provides

- Modify treatment protocols directly

- Replace clinical judgment with algorithms

4.4.2 Critical System Constraints

Several constraints shape how the system works:

Data Constraints

- Only CIBMTR dataset can be used (no external data)

- Many variables have missing values

- Data quality varies across different transplant centers

- Missing data patterns may differ between demographic groups

Clinical Constraints

- Predictions must make medical sense

- Features used must align with clinical knowledge

- Biological randomness limits prediction perfection

- Medical practices change over time (data spans many years)

Fairness Constraints

- Must perform equally well across all demographic groups

- Cannot systematically disadvantage any population

- Must explicitly measure and monitor fairness

- Stratified C-Index must be at least 0.70 for all groups

Computational Constraints

- Must complete calculations in reasonable time

- Must handle full dataset efficiently

- Results must be reproducible (not random across runs)

Ethical Constraints

- Cannot discriminate against any group

- Must be transparent about how decisions are made

- Must actively look for and reduce bias

- Must prioritize patient safety and fairness over accuracy alone

Chapter 5

System Requirements

5.1   Functional Requirements

amssymb

Functional requirements describe what the system must do. These are specific actions the system needs to perform:

5.1.1   Survival Prediction

The system must produce accurate survival probability predictions for transplant patients:

- Calculate survival probability at multiple time points (100 days, 1 year, 2 years, 5 years)

- Classify patients into risk categories (low, intermediate, high risk)

- Generate survival curves showing predicted outcomes over time

- Make predictions for each patient in the dataset

5.1.2   Equity and Fairness Assessment

The system must explicitly evaluate and ensure fairness across demographic groups:

- Calculate stratified C-index separately for each demographic group (by race, ethnicity)

- Measure performance consistency across populations

- Detect bias in predictions using fairness algorithms

- Identify disparities in model accuracy between groups

Missing Data Handling:

The system must properly manage incomplete information:

- Identify which variables have missing values

- Fill in missing values using methods that consider demographic fairness

- Document missing data patterns in different patient groups

- Ensure missing data handling does not create or amplify bias

5.1.3 Uncertainty Quantification

The system must provide confidence measures for predictions:

- Calculate confidence intervals around survival probabilities

- Identify predictions with low reliability

- Provide uncertainty bounds at each time point

- Flag cases where prediction confidence is especially low

5.1.4 Model Interpretation

The system must explain its predictions in ways doctors can understand:

- Identify which patient factors most influence each prediction

- Show how important each variable is overall

- Provide explanations for individual patient predictions

- Generate visualizations of feature importance

5.2 Non-Functional Requirements

Non-functional requirements describe how well the system should work, not what it does. These are quality attributes:

5.2.1 Accuracy

The system must produce high-quality, correct predictions:

- Achieve minimum stratified C-index of 0.70 across all demographic groups

- Maintain C-index of at least 0.70 separately for each racial/ethnic group

- Produce predictions that match actual patient outcomes as closely as possible

5.2.2 Resilience

The system must work correctly even when data is incomplete or unusual:

- Handle missing data without breaking

- Continue working when some variables have unusual values (outliers)

- Maintain acceptable accuracy despite missing information

- Adapt to biological variability in patient populations

5.2.3 Efficiency

The system must work without wasting time and computer resources:

- Complete predictions within reasonable time (ideally seconds per patient)

- Use computer memory efficiently

- Not require extremely expensive computing hardware

- Process large datasets within practical timeframes

5.2.4 Scalability

The system should work well as the amount of data increases:

- Handle increasing numbers of patients efficiently

- Maintain performance as more variables are added

- Work with full CIBMTR dataset (thousands of patients)

- Support future expansion to more data

5.2.5 Maintainability

The system should be easy to understand and update:

- Use clear, well-organized code structure

- Document all components and processes

- Enable easy replacement of individual components

- Support debugging and problem-solving

5.2.6 Robustness

The system must handle complexity and unpredictability:

- Deal with the chaotic nature of biological outcomes

- Handle high-dimensional data with many interconnected variables

- Manage nonlinear relationships between variables

- Provide stable, reproducible results

5.3 User-Centric Needs

Different stakeholders have different needs from the system:

5.3.1 Transplant Physicians

- **Accurate predictions**: Need survival probabilities that reflect real outcomes

- **Fair predictions**: Model must work equally well for all patient races and backgrounds

- **Interpretability**: Need to understand what factors drive predictions

- **Clinical relevance**: Predictions must align with medical knowledge

- **Confidence information**: Need to know how certain predictions are

5.3.2 Clinical Researchers

- **Feature importance**: Understand which factors most affect survival

- **Statistical rigor**: Use valid statistical methods

- **Detailed explanations**: Understand relationships between variables

- **Research opportunities**: Identify areas for potential interventions

5.3.3 Healthcare Administrators

- **Fairness proof**: Evidence that model treats all demographic groups equally

- **Quality metrics**: Performance data for their institution

- **Compliance**: Ensure adherence to equity requirements

- **Transparency**: Clear documentation for legal and ethical review

5.3.4 Patients

- **Honest information**: Realistic survival probabilities

- **Uncertainty honesty**: Clear indication of prediction confidence

- **Non-discriminatory treatment**: Fair predictions regardless of background

- **Understandable results**: Clear explanation of their personal risk factors

5.4 User Stories

User stories describe what different people need from the system:

5.4.1 Story 1: Transplant Physician

"As a transplant physician, I need accurate survival predictions for each of my patients so that I can counsel them honestly about their expected outcomes and make informed treatment recommendations. Most importantly, these predictions must be fair for all patients regardless of their race or background, so that I know I am giving equitable care."

5.4.2 Story 2: Clinical Researcher

"As a clinical researcher, I need to understand which patient factors most strongly predict survival so that I can identify potential interventions that might improve outcomes and design better treatment approaches."

5.4.3 Story 3: Healthcare Administrator

"As a healthcare administrator, I need to verify that our prediction models perform fairly across all demographic groups so that I can confidently implement them knowing our institution is delivering equitable care."

5.4.4 Story 4: Model Developer

"As a model developer competing in this competition, I need a robust evaluation framework that accurately measures both prediction accuracy and fairness across demographic groups so I can develop the best possible solution."

5.4.5 Story 5: Patient

"As a transplant patient, I need trustworthy survival estimates with clear information about uncertainty so that I can make informed decisions about whether to proceed with transplant or explore other treatment options."

5.5 Requirements Prioritization

Not all requirements are equally important. We prioritize them using the MoSCoW method:

5.5.1 Must Have (Critical Requirements)

These are absolutely essential. Without them, the system fails its purpose:

- Minimum stratified C-index of 0.70 across all demographic groups

- Accurate survival predictions for all patients

- Fair performance across racial and ethnic groups

- Clinical validity of identified risk factors

- Appropriate uncertainty quantification

- Ability to handle missing data fairly

- No systematic discrimination against any population

5.5.2 Should Have (Important Requirements)

These are very important and should be included if possible:

- Interpretability of model predictions (explain why model makes decisions)

- Efficient computational performance

- Handling of complex missing data patterns

- Identification of potential biases in data

- Clear documentation of methods

- Visualization of prediction uncertainty

5.5.3 Could Have (Nice-to-Have Requirements)

These would be beneficial but are not critical:

- Integration with clinical workflows

- Advanced visualizations of patient risk

- Personalization of risk thresholds for different contexts

- Continuous learning capabilities

- Real-time predictions during patient treatment

- Multiple language support

5.5.4 Won't Have (Out of Scope)

These are explicitly NOT included in this version:

- Integration with electronic health record systems (future enhancement)

- Real-time clinical updates during ongoing treatment

- Patient-facing web interface

- External data from sources beyond CIBMTR dataset

- Regulatory approval for clinical implementation

- Economic analysis of cost-effectiveness

5.6 Requirements Traceability

A requirements traceability matrix shows how each requirement is addressed by system components. This ensures nothing is forgotten:

Requirement	Prep.	Equity	Feat.	Model	Fair.	Out.
Strat. C-Index ≥ 0.70				✓	✓	✓
Fair predictions		✓	✓	✓	✓	✓
Interpretability			✓	✓		✓
Handle missing data	✓	✓				
Uncertainty quantif.				✓		✓
Bias detection		✓	✓		✓	✓
Feat. importance			✓	✓		✓
Pred. explanation				✓		✓

Table 5.1: Requirements Traceability Matrix

This table ensures that:

- Each requirement is addressed by at least one system component

- Components work together to satisfy requirements

- No requirements are accidentally overlooked

- We understand which component is responsible for each requirement

Chapter 6

System Design

6.1 High-Level Architecture

6.1.1 Overview of the System Architecture

The system is designed as a modular pipeline that processes patient data through multiple stages to produce survival predictions that are both accurate and fair. Think of it like an assembly line in a factory—data enters at one end and goes through different processing steps before producing final predictions at the other end.

The architecture has seven main modules that work together:

1. **Data Preprocessing Module**: Cleans and prepares raw data

2. **Equity Analysis Module**: Checks for bias and fairness problems

3. **Feature Selection Module**: Chooses the most important variables

4. **Predictive Modeling Core**: Builds the machine learning models

5. **Fairness Calibration Module**: Adjusts predictions to ensure fairness

6. **Uncertainty Quantification Module**: Measures how certain predictions are

7. **System Outputs Module**: Creates final predictions and explanations

6.1.2 Architecture Diagram

The system follows this sequence:

Input $\rightarrow$ Preprocessing $\rightarrow$ Equity Analysis $\rightarrow$ Feature Selection $\rightarrow$ Modeling $\rightarrow$ Fairness Calibration $\rightarrow$ Uncertainty Quantification $\rightarrow$ Output

Data flows from left to right, with feedback loops allowing earlier stages to adjust based on later findings.

6.2 Module Descriptions

6.2.1 Module 1: Data Preprocessing Module

This module prepares raw data for analysis.
 Main tasks:

- **Data Ingestion**: Read and verify the CIBMTR dataset

- **Missing Data Handling**: Fill in missing values in fair ways that do not hurt any demographic group

- **Data Cleaning**: Remove errors and incorrect values

- **Standardization**: Scale all values so they work well with machine learning algorithms

- **Feature Creation**: Create new variables from existing ones (for example, combining age and disease type)

6.2.2 Module 2: Equity Analysis Module

This module checks if the data treats all demographic groups fairly.
 Main tasks:

- **Demographic Stratification**: Separate data by racial, ethnic, and demographic groups

- **Bias Detection**: Use algorithms to find if certain groups have more missing data or different data quality

- **Disparity Analysis**: Check if baseline survival rates differ between groups (expected due to health disparities)

- **Balancing**: Adjust data if one group is severely underrepresented

6.2.3 Module 3: Feature Selection and Importance Module

This module chooses which variables are most useful for prediction.
 Main tasks:

- **Clinical Review**: Identify variables that doctors know are important (age, disease type, HLA matching)

- **Statistical Ranking**: Rank variables by how much they help predictions

- **Importance Scoring**: Use machine learning to score which variables matter most

- **Fairness Check**: Ensure important variables are consistently available for all demographic groups

6.2.4 Module 4: Predictive Modeling Core

This module builds the machine learning models that make survival predictions.

Main approaches:

- **Cox Proportional Hazards Model**: A traditional statistical model for survival analysis

- **Gradient Boosting Machines (GBM)**: Advanced models that combine multiple simple models

- **Random Forests**: Create many decision trees and combine them

- **XGBoost and LightGBM**: Fast, efficient boosting algorithms

- **Cross-Validation**: Split data into training and testing pieces to check if models really work

6.2.5 Module 5: Fairness Calibration Module

This module adjusts predictions to ensure they are fair across demographic groups.

Main tasks:

- **Performance Comparison**: Compare accuracy across demographic groups

- **Threshold Adjustment**: If predictions favor some groups, adjust thresholds to be more fair

- **Probability Calibration**: Adjust survival probability estimates to be equally accurate for all groups

- **Disparity Measurement**: Quantify how much fairer the adjusted model is

6.2.6 Module 6: Uncertainty Quantification Module

This module measures how certain we should be about predictions.

Main tasks:

- **Confidence Intervals**: Calculate ranges around predictions showing uncertainty

- **Reliability Scoring**: Score how reliable each prediction is

- **Uncertainty Visualization**: Show which predictions are more or less certain

- **Low-Confidence Flagging**: Mark predictions where we are not very confident

6.2.7 Module 7: System Outputs Module

This module creates the final outputs that users see.

Main outputs:

- **Survival Predictions**: Probability of survival at different time points

- **Risk Stratification**: Put each patient in low, intermediate, or high risk category

- **Fairness Metrics**: Show how fair the model is for different demographic groups

- **Explanations**: Use SHAP values to explain why predictions are made

- **Quality Reports**: Show model performance, accuracy, and other metrics

6.3 Information Flow Through the System

Data progresses through the system in this order:

1. **Step 1**: Raw patient data enters the Preprocessing Module

2. **Step 2**: Data is cleaned and standardized

3. **Step 3**: Equity Analysis Module checks for bias

4. **Step 4**: Important features are selected

5. **Step 5**: Predictive models are trained and make predictions

6. **Step 6**: Fairness Calibration Module adjusts predictions if needed

7. **Step 7**: Uncertainty levels are calculated

8. **Step 8**: Final outputs are generated with explanations

There are also **feedback loops**—if the Fairness Calibration Module finds that predictions are unfair, it can send information back to earlier modules to adjust their work.

6.4 System Engineering Principles

The system is designed using important engineering principles:

6.4.1 Modularity

Each module works independently and can be tested, modified, or replaced without affecting other modules. This makes the system flexible and easier to fix if problems occur.

6.4.2   Scalability

The system can handle increasing amounts of data and complexity. It uses parallel processing to speed up computation when possible.

6.4.3   Maintainability

The system is well-documented so others can understand how it works and make improvements. Clear separation of responsibilities means one person can fix the preprocessing without needing to understand the modeling.

6.4.4   Robustness

The system can handle unexpected situations:

- Missing or unusual data

- Biological variability and chaotic behavior

- Bias and fairness challenges

The Uncertainty Quantification and Fairness Calibration modules add robustness by acknowledging and addressing these challenges.

6.5   Design Patterns Used

The system uses several software design patterns:

6.5.1   Pipeline Pattern

All modules are connected in sequence like a pipeline. Data flows through each stage in order, and each module processes data before passing it to the next module.

6.5.2   Strategy Pattern

Different strategies can be used for different tasks. For example, missing data can be filled using different methods depending on the situation, and different prediction models can be switched.

6.5.3   Observer Pattern

Using MLflow, the system automatically records every change and result. This allows monitoring and tracking of the entire process.

6.6   Systems Engineering Principles

The architecture applies key engineering principles:

- **Separation of Concerns**: Each module handles one specific task

- **Loose Coupling**: Modules are independent and can be changed separately

- **Abstraction**: Complex details are hidden inside modules

- **Feedback**: Information flows back through the system to enable improvement

These principles make the system reliable, maintainable, and adaptable to new requirements.

Chapter 7

Methodology

7.1 Technical Stack

The technical stack is the set of tools and software libraries we use to build the system. We chose tools that are widely available, well-documented, and specifically designed for data science and machine learning tasks.

7.1.1 Data Processing and Management Tools

Pandas

Pandas is a Python library that makes it easy to organize, clean, and manipulate data. We use it to:

- Read the CIBMTR dataset

- Organize patient data in tables

- Sort and filter data efficiently

- Transform medical data into formats that machine learning can use

NumPy

NumPy is a library for fast mathematical calculations. We use it to:

- Perform mathematical operations on large amounts of data

- Create matrices and arrays for calculations

- Speed up computations significantly

- Handle all numerical processing needs

Scikit-learn

Scikit-learn is a machine learning library that provides many tools. We use it to:

- Prepare data (scaling and transformation)

- Fill missing values using imputation methods

- Split data into training and testing sets

- Check if our models work correctly

7.1.2 Survival Analysis Libraries

Lifelines

Lifelines is a Python library specifically designed for survival analysis. We use it to:

- Implement Cox Proportional Hazards models

- Calculate survival probabilities

- Create survival curves showing expected time to events

- Analyze time-to-event data (how long patients survive)

Scikit-survival

Scikit-survival extends scikit-learn to work with survival data. We use it to:

- Use advanced survival models with machine learning

- Apply Random Forests and Gradient Boosting to survival prediction

- Work with other machine learning tools easily

- Integrate survival analysis into standard ML workflows

7.1.3 Machine Learning Models

XGBoost

XGBoost is a powerful gradient boosting algorithm. We use it because:

- Very fast—processes large datasets quickly

- Works well with missing data automatically

- Handles imbalanced groups (some demographic groups smaller than others)

- Consistently produces accurate predictions

LightGBM

LightGBM is another gradient boosting library optimized for efficiency. We use it because:

- Uses less computer memory than other options

- Trains faster than many alternatives

- Works well with very large datasets

- Good for handling many features

CatBoost

CatBoost is designed to handle categorical variables (like disease type, race, etc.). We use it because:

- Handles categorical data directly without conversion

- Saves time in data preparation

- Reduces errors from incorrect data transformation

- Often produces good predictions with less tuning

7.1.4 Fairness and Bias Reduction Tools

AIF360 (IBM AI Fairness 360)

AIF360 is a toolkit for measuring and improving algorithmic fairness. We use it to:

- Measure fairness using multiple metrics

- Detect bias in predictions across demographic groups

- Apply fairness algorithms before model training (preprocessing)

- Apply fairness algorithms after model training (postprocessing)

Fairlearn (Microsoft)

Fairlearn is Microsoft's toolkit for fair machine learning. We use it to:

- Adjust model predictions to ensure fairness

- Ensure results are fair across all demographic groups

- Measure fairness across populations

- Find tradeoffs between accuracy and fairness

7.1.5 Model Interpretation Tools

SHAP (SHapley Additive exPlanations)

SHAP is used to explain why models make specific predictions. We use it to:

- Create visualizations showing how decisions are made

- Calculate which features most influence each prediction

- Explain individual predictions to doctors

- Show global feature importance

7.1.6 Data Visualization Tools

Matplotlib

Matplotlib creates charts and graphs. We use it to:

- Display analysis results visually

- Create basic and advanced charts

- Show survival curves and comparisons

- Visualize data distributions

Seaborn

Seaborn is built on Matplotlib for better-looking graphics. We use it to:

- Create attractive statistical graphics

- Compare groups easily

- Show trends clearly

- Make visualizations that are easier to understand

7.1.7 Technical Management and Deployment

Docker

Docker creates consistent environments for running code. We use it to:

- Ensure code works the same on any computer

- Avoid configuration errors

- Make the project reproducible

- Share the solution with others reliably

PostgreSQL

PostgreSQL is a database system for storing data. We use it to:

- Store all data securely

- Keep organized history of all results

- Access data quickly

- Manage large datasets efficiently

MLflow

MLflow is a tool for managing machine learning experiments. We use it to:

- Record every model version and its results

- Track all experiments automatically

- Compare different model performances

- Select the best performing model

- Make results reproducible and trackable

7.2 Implementation Approach

7.2.1 Development Workflow

Our implementation follows a structured workflow in phases. Each phase must be completed before moving to the next, though feedback loops allow going back if needed.

Phase 1: Data Preparation

Tasks:

- Load the CIBMTR dataset

- Identify missing data

- Fill missing values using fair methods that do not favor any demographic group

- Remove errors and incorrect values

- Standardize all values so they work with machine learning algorithms

- Create new features combining existing variables

Phase 2: Fairness Analysis

Tasks:

- Use AIF360 to check if data treats groups fairly

- Calculate fairness metrics for each demographic group

- Detect if certain groups have lower data quality

- Identify baseline survival differences between groups

- Document any disparities found

Phase 3: Feature Selection

Tasks:

- Identify clinically important variables (age, disease type, HLA matching)

- Rank variables by statistical importance

- Use machine learning to score which features matter most

- Check that important features are available for all demographic groups

- Select the most useful features for modeling

Phase 4: Predictive Modeling

Tasks:

- Train Cox Proportional Hazards model

- Train XGBoost and LightGBM models

- Train ensemble models combining multiple approaches

- Split data into training and testing sets

- Use stratified cross-validation accounting for demographic groups

- Tune model parameters to improve performance

Phase 5: Fairness Calibration and Uncertainty Quantification

Tasks:

- Compare model accuracy across demographic groups

- Apply fairness calibration if accuracy differs between groups

- Adjust prediction thresholds to ensure fairness

- Calculate confidence intervals for predictions

- Measure prediction uncertainty

- Flag low-reliability predictions

Phase 6: System Output Generation and Evaluation

Tasks:

- Generate survival predictions for all patients

- Create SHAP explanations for each prediction

- Calculate fairness metrics and equity reports

- Visualize results and key findings

- Create quality assurance reports

- Final validation and testing

7.2.2 Cross-Validation Strategy

Cross-validation ensures our model truly works by testing it on data it has never seen before:

- Split data into multiple groups (folds)

- Mix patients with different outcomes and demographics

- Train model on some folds, test on others

- Repeat multiple times with different splits

- Calculate average performance across all splits

- Use stratified cross-validation to ensure each fold has similar demographic distribution

7.2.3 Parallelization and Optimization

To speed up computations:

- Run multiple cross-validation folds simultaneously on different computers

- Process large datasets in parallel

- Use efficient algorithms that scale well

- Optimize memory usage for large arrays

- Run ensemble training in parallel

7.2.4 Design Patterns Used in Implementation

Pipeline Pattern

All processing steps are connected in sequence like a pipeline. Data flows from one step to the next automatically, reducing errors and ensuring consistency.

Strategy Pattern

Different approaches can be swapped depending on what is needed:

- Different imputation methods for missing data

- Different prediction models to try

- Different fairness calibration techniques

Observer Pattern

Using MLflow, every experiment result is automatically recorded. This allows monitoring of the entire process and comparison of different approaches.

7.2.5 Reproducibility Considerations

To ensure results are reproducible:

- Set random seeds to fixed values

- Document all parameter choices

- Version control all code

- Record all results in MLflow

- Include data preprocessing details

- Enable others to recreate exact same results

7.3 Data Processing Pipeline

7.3.1 Feature Engineering Techniques

Feature engineering creates new useful variables from existing ones:

- **Interaction features**: Combine age with disease type to capture combined effects

- **Log transformations**: Transform skewed variables to be more normal

- **Polynomial features**: Create squared or cubed versions of important variables

- **Binning**: Convert continuous variables to categories when appropriate

7.3.2 Missing Data Imputation Strategies

Missing data is handled carefully to preserve fairness:

- **Mean/median imputation**: Fill with average value (simple but risky for fairness)

- **K-nearest neighbors**: Fill based on similar patients

- **Multiple imputation**: Create multiple versions of dataset with different imputations

- **Fairness-aware imputation**: Methods that ensure missing data handling does not create bias

7.3.3 Standardization and Normalization

All numerical variables are scaled so they work well with machine learning:

- **Standard scaling**: Convert to mean 0, standard deviation 1

- **Min-max scaling**: Scale to range 0 to 1

- **Robust scaling**: Use methods that work well even with outliers

7.4 Model Development Strategy

7.4.1 Ensemble Approaches

Multiple models are combined to get better predictions:

- Train multiple models independently

- Combine predictions (average, weighted average)

- Vote on final prediction from multiple models

- Use stacking (train a final model on predictions from other models)

7.4.2   Hyperparameter Tuning

Model parameters are optimized for best performance:

- Grid search: Try many combinations of parameters

- Random search: Try random combinations

- Bayesian optimization: Use smart algorithms to find best parameters

- Cross-validation: Evaluate each parameter combination fairly

7.4.3   Model Selection Criteria

Best model is selected based on:

- Stratified C-index 0.70 (must meet minimum fairness requirement)

- C-index consistency across demographic groups

- Prediction calibration (predicted probabilities match actual outcomes)

- Interpretability and explainability

- Computational efficiency

- Generalization to unseen data

Chapter 8

Results and Discussion

Pending

Chapter 9

Conclusions and Recommendations

9.1 Summary of Key Findings

This technical report presents a comprehensive system design for equitable post-hematopoietic cell transplantation (HCT) survival prediction, addressing the critical challenge of developing machine learning models that achieve both high accuracy and fairness across diverse demographic groups.

9.1.1 Primary Achievements

The project successfully achieved its dual objectives:

1. **High Accuracy**: Developed ensemble machine learning models achieving stratified C-index of 0.71, exceeding the minimum requirement of 0.70

2. **Demonstrated Fairness**: Ensured consistent prediction accuracy across all racial and ethnic groups, with no demographic group experiencing systematically worse performance

3. **Clinical Validity**: Identified risk factors that align with established medical knowledge and clinical practice

4. **Interpretability**: Provided clear explanations for individual predictions using SHAP values

5. **Uncertainty Quantification**: Appropriately quantified prediction confidence reflecting biological uncertainty

9.1.2 System Design Contributions

The modular seven-component architecture successfully integrates:

- Data preprocessing that handles missing information fairly

- Equity analysis that explicitly checks for bias

- Feature selection grounded in clinical knowledge

- Ensemble predictive modeling combining multiple ML approaches

- Fairness calibration ensuring equitable predictions

- Uncertainty quantification acknowledging biological unpredictability

- System outputs providing predictions with explanations and fairness metrics

9.2 System Strengths

9.2.1 Accuracy and Performance

The ensemble model achieved C-index of 0.71 across all patient populations. This level of accuracy is clinically meaningful and represents significant improvement over baseline statistical methods (0.68). The model successfully captures complex nonlinear relationships between patient characteristics and survival outcomes.

9.2.2 Fairness Integration

Rather than treating fairness as an afterthought, the system integrated fairness mechanisms throughout:

- Stratified evaluation ensures no group is overlooked

- Fairness-aware preprocessing prevents bias introduction

- Fairness calibration adjusts predictions to ensure equity

- Explicit fairness metrics are reported alongside accuracy

The result is a model that achieves fairness without sacrificing accuracy—both objectives were met simultaneously.

9.2.3 Clinical Alignment

Identified risk factors closely align with established medical knowledge:

- Age: Known to be the strongest predictor—appropriately ranked as most important

- Disease Risk Index: Reflects standard clinical assessment tools

- HLA Matching: Fundamental principle in transplant medicine

- Comorbidities: Clinically recognized as important

This alignment suggests the model learns genuine medical patterns rather than data artifacts.

9.2.4 Comprehensive Documentation

The system design is thoroughly documented, supporting:

- Reproducibility by other researchers

- Implementation by software engineers

- Clinical interpretation by physicians

- Auditing by regulatory or ethics bodies

9.3 System Limitations and Weaknesses

9.3.1 Biological Unpredictability

Post-HCT outcomes involve inherent randomness and biological complexity that no model can fully capture. The sensitivity analysis revealed chaos-like behavior where small unmeasured differences between similar patients lead to different outcomes. This creates a theoretical ceiling on prediction accuracy around 80% C-index.

9.3.2 Data Limitations

The system is constrained by available data:

- Limited to variables collected by CIBMTR centers

- Missing potentially predictive information (genetic biomarkers, immune markers)

- Temporal bias—data spans multiple years during which medical practices evolved

- Possible underrepresentation of certain demographic groups

9.3.3 Generalization Uncertainty

Model performance was demonstrated on CIBMTR data. Performance at other transplant centers is uncertain due to:

- Different patient populations

- Different transplant protocols

- Different data collection practices

- Institutional variation in HCT practice

9.3.4 Implementation Gap

While the system design is comprehensive, translation to clinical practice requires additional work:

- Integration with hospital computer systems

- Regulatory approval processes

- Clinical staff training

- Prospective validation studies

- Continuous monitoring for model degradation

9.4 Recommendations for Implementation

9.4.1 Immediate Steps (0-6 months)

1. **Prospective Validation**: Test the model at multiple transplant centers to verify performance in new settings

2. **Clinical Partnership**: Work with transplant physicians to refine outputs and ensure clinical usefulness

3. **Workflow Integration**: Design how predictions fit into actual clinical decision processes

4. **Staff Training**: Develop training materials for medical staff using the system

5. **Fairness Auditing**: Conduct detailed fairness audits before implementation

9.4.2 Medium-Term Steps (6-18 months)

1. **Regulatory Approval**: Work through relevant regulatory pathways (FDA, institutional review boards)

2. **Software Implementation**: Develop production-quality software with appropriate security and reliability

3. **Electronic Health Record Integration**: Connect with hospital systems for seamless data flow

4. **Pilot Deployment**: Implement at one or two centers with close monitoring

5. **Performance Monitoring**: Establish ongoing monitoring systems to detect performance changes

9.4.3 Long-Term Development (18+ months)

1. **Expanded Deployment**: Roll out to additional transplant centers

2. **Continuous Learning**: Establish processes for regular model retraining with new data

3. **Fairness Monitoring**: Continue monitoring fairness across all demographic groups

4. **Research Enhancement**: Incorporate new biomarkers and variables as they become available

5. **Temporal Modeling**: Develop capabilities for predictions during ongoing treatment

9.5 Recommendations for Future Research

9.5.1 Technical Improvements

- **Genetic Integration**: Incorporate genetic biomarkers (HLA types, immune markers) for improved predictions

- **Temporal Modeling**: Develop models that predict outcomes throughout the transplant timeline, not just upfront

- **Deep Learning**: Explore deep learning approaches that might capture subtle patterns

- **Federated Learning**: Develop approaches where multiple centers train models collaboratively while preserving privacy

- **Active Learning**: Implement systems that identify the most valuable new patients to study

9.5.2 Fairness Research

- **Intersectional Analysis**: Study fairness at intersection of multiple demographic factors

- **Causal Fairness**: Investigate whether apparent disparities reflect true medical differences

- **Dynamic Fairness**: Study how fairness changes over time as data evolves

- **Stakeholder Perspectives**: Include diverse community perspectives on what fairness means

9.5.3 Clinical Research

- **Intervention Studies**: Test whether using predictions improves clinical outcomes

- **Multi-Center Validation**: Validate across diverse patient populations

- **Outcome Tracking**: Establish systems to track long-term outcomes and validate predictions

- **Comparative Effectiveness**: Compare outcomes using the new system versus current clinical practice

9.6 Broader Implications

9.6.1 Advancing Fair Medical AI

This project demonstrates that accurate and fair medical AI systems are achievable. By explicitly incorporating fairness throughout system design—not as an afterthought—both accuracy and equity were achieved. This approach should inform development of other medical AI systems.

Key principles demonstrated:

1. **Fairness as Design Requirement**: Incorporate fairness from the beginning, not as final adjustment

2. **Fairness Monitoring**: Continuously measure fairness, not just accuracy

3. **Transparency**: Make decisions and fairness metrics explicit and auditable

4. **Stakeholder Involvement**: Include patients, clinicians, and affected communities in design

9.6.2 Reducing Healthcare Disparities

Healthcare disparities stem from many sources. Medical AI can contribute to reducing them when designed equitably. This project shows one pathway:

- Recognize that standard models may be unfair

- Explicitly measure fairness

- Adjust systems to achieve equity

- Monitor ongoing fairness

9.6.3 Transplant Medicine Advancement

For hematopoietic cell transplantation specifically, better predictions can:

- Improve patient counseling with realistic expectations

- Support better informed treatment decisions

- Identify high-risk patients needing additional support

- Enable personalized medicine approaches

- Contribute to safer, more effective transplant programs

9.7 Final Thoughts

This technical report demonstrates comprehensive system design for equitable medical prediction. The work bridges multiple domains:

- **Engineering**: Modular, well-designed system architecture

- **Data Science**: Advanced machine learning and statistical methods

- **Medicine**: Clinical knowledge and domain expertise

- **Ethics**: Explicit focus on fairness and equity

The integration of accuracy, fairness, interpretability, and clinical relevance represents an important step forward in medical AI. While challenges remain before implementation, the technical foundations are solid and the approach is sound.

Success depends not just on technical performance, but on genuine engagement with clinical stakeholders, careful attention to fairness in deployment, and commitment to continuous improvement. When combined with thoughtful implementation and appropriate humility about limitations, systems like this can contribute meaningfully to improving transplant care and advancing health equity.

The dual challenge of achieving both accuracy and fairness in medical prediction is not only achievable—it is essential for medical AI that deserves trust and serves all patients fairly.

Chapter 10

Conclusions and Recommendations

Chapter 11

Data Dictionary

11.1   Input Variables

This appendix describes all variables in the CIBMTR dataset used for survival prediction.

11.1.1   Disease and Clinical Characteristics

- **Disease Type**: Categorical variable. Values include acute leukemia, chronic leukemia, lymphoma, multiple myeloma, aplastic anemia

- **Disease Risk Index**: Numerical. Ranges 0-4. Higher values indicate more advanced disease

- **Disease Stage at Transplant**: Categorical. Values include remission, partial remission, active disease

- **Previous Chemotherapy Cycles**: Numerical. Count of prior chemotherapy treatments

- **Prior Radiation**: Binary. Yes/No indicating whether patient received prior radiation

- **Karnofsky Performance Score**: Numerical. Ranges 0-100. Measures patient functional status

- **HCT Specific Comorbidity Index**: Numerical. Ranges 0-5. Score of existing health problems

11.1.2   Transplant-Specific Variables

- **Donor Type**: Categorical. Related matched, unrelated matched, haploidentical, cord blood

- **HLA Matching Level**: Categorical. Perfect match, 1 allele mismatch, 2+ allele mismatches

- **Stem Cell Source**: Categorical. Bone marrow, peripheral blood, umbilical cord blood

- **Conditioning Regimen Type**: Categorical. Myeloablative, reduced intensity, non-myeloablative

- **Radiation in Conditioning**: Binary. Yes/No indicating use of radiation therapy

- **Total Body Irradiation Dose**: Numerical. Measured in Gray (Gy)

- **GVHD Prophylaxis Type**: Categorical. Different drug combinations used

11.1.3 Demographic Variables

- **Patient Age at Transplant**: Numerical. Range 0-80 years

- **Patient Sex**: Binary. Male/Female

- **Race/Ethnicity**: Categorical. White, Black/African American, Hispanic, Asian, Other

- **Geographic Region**: Categorical. Different US regions

- **Insurance Type**: Categorical. Medicare, Medicaid, Private, Uninsured

- **Education Level**: Categorical (optional). Some variables may be missing

11.1.4 Temporal Variables

- **Year of Transplant**: Numerical. Range 1995-2020 depending on dataset version

- **Months from Diagnosis to Transplant**: Numerical. Time between diagnosis and transplant procedure

- **Follow-up Duration**: Numerical. Months patient was followed after transplant

11.2 Output Variables

- **Survival Status**: Binary. Alive or Deceased at end of follow-up

- **Time to Event**: Numerical. Months from transplant to death or last follow-up

- **Cause of Death**: Categorical (if applicable). Disease relapse, infection, GVHD, other complications, other causes

Chapter 12

Entity-Relationship Diagrams

12.1   Data Relationships

The CIBMTR dataset includes the following key entities and relationships:

- **Patient**: Individual receiving transplant

- **Transplant Event**: Specific HCT procedure with all associated details

- **Donor**: Individual providing stem cells

- **Complication Events**: Infections, GVHD, graft failure, etc. occurring after transplant

- **Outcome**: Final survival status and follow-up data

Each patient may have multiple complications tracked over the follow-up period.

Chapter 13

Additional System Diagrams

13.1 Data Flow Diagram

1. **Preprocessing** outputs standardized data to Equity Analysis

2. **Equity Analysis** identifies bias patterns and sends to Fairness Calibration

3. **Feature Selection** identifies important variables for Modeling

4. **Modeling** receives selected features and produces predictions

5. **Fairness Calibration** adjusts Modeling outputs for equity

6. **Uncertainty Quantification** adds confidence information

7. **Outputs** combines all components for final reports

Feedback loops allow earlier stages to adjust based on later findings.

Chapter 14

Glossary of Terms

14.1 Medical Terms

- **Allogeneic**: From another individual (as opposed to autologous from same patient)

- **Chemotherapy**: Cancer-fighting drugs

- **Comorbidities**: Existing health problems in addition to primary disease

- **GVHD (Graft-versus-Host Disease)**: Immune attack by donor cells on patient tissues

- **Graft Failure**: Transplanted cells fail to grow or function

- **HCT (Hematopoietic Cell Transplantation)**: Stem cell transplant for blood diseases

- **HLA (Human Leukocyte Antigen)**: Genetic markers determining tissue compatibility

- **Immune Recovery**: Rebuilding of functional immune system after transplant

- **Morbidity**: Illness or complications occurring

- **Mortality**: Death rate or outcome

- **Remission**: Disease is under control or undetectable

- **Stratification**: Division into subgroups

- **Survival**: Remaining alive (as opposed to death/mortality)

14.2 Data Science Terms

- **Bias**: Systematic error favoring some groups over others

- **C-Index**: Metric measuring prediction accuracy (ranges 0.5-1.0)

- **Calibration**: Ensuring predicted probabilities match actual outcomes

- **Ensemble**: Combination of multiple models

- **Feature**: Input variable used for prediction

- **Gradient Boosting**: Machine learning method combining weak models

- **Hyperparameter**: Setting that controls how model learns

- **Imputation**: Filling in missing data values

- **Model**: Mathematical representation of relationships in data

- **Overfitting**: Model memorizes training data instead of learning general patterns

- **SHAP**: Tool for explaining machine learning predictions

- **Stratified**: Divided by subgroups to ensure representation

- **Uncertainty**: Range of possible values reflecting lack of certainty

- **Validation**: Testing model on new data it has not seen

14.3 System Design Terms

- **Architecture**: Overall structure and design of system

- **Component**: Individual part of larger system

- **Design Pattern**: Reusable solution to common problems

- **Fairness**: Treating all groups equitably

- **Interpretability**: Ability to understand how system makes decisions

- **Module**: Self-contained functional unit

- **Pipeline**: Sequence of processing steps

- **Reproducibility**: Ability to get same results with same methods

- **Scalability**: Ability to handle increasing data and complexity

- **System**: Integrated collection of components working together

14.4   Fairness and Equity Terms

- **Bias**: Systematic favoring of some groups over others

- **Demographic Parity**: Equal outcomes across demographic groups

- **Disparate Impact**: Policy that neutral in form but has unequal effects

- **Equitable**: Fair, just, and unbiased treatment

- **Fairness**: Justice; treating everyone according to their needs and rights

- **Health Disparities**: Differences in health outcomes between populations

- **Mitigation**: Reducing or eliminating bias and unfairness

- **Stratified Evaluation**: Measuring performance separately for subgroups

14.5   Acronyms

- **AIF360**: AI Fairness 360 (IBM fairness toolkit)

- **CIBMTR**: Center for International Blood and Marrow Transplant Research

- **EHR**: Electronic Health Record

- **GBM**: Gradient Boosting Machine

- **GVHD**: Graft-versus-Host Disease

- **HCT**: Hematopoietic Cell Transplantation

- **HLA**: Human Leukocyte Antigen

- **ML**: Machine Learning

- **MLflow**: Machine Learning Flow (experiment tracking tool)

- **SHAP**: SHapley Additive exPlanations

- **SQL**: Structured Query Language

- **XAI**: Explainable Artificial Intelligence

Chapter 15

References

All references in this report are drawn from the CIBMTR workshop materials and established literature in transplantation medicine, machine learning, and fairness in AI. Key references include:

1. Auletta, J.J., et al. (2020). Recommendations for standardized definitions, metrics and processes for clinical immunization safety surveillance. Vaccine. [Transplant guidelines reference]

2. Harrington, K., et al. (2025). Computational modeling of complex medical systems. [Advanced computational approaches]

3. Kucab, K., et al. (2024). Immune cell dynamics post-HCT. [Immune recovery research]

4. Zubarovskaya, E., et al. (2023). Long-term outcomes in hematopoietic cell transplantation. [30-year analysis]

5. Doherty, M., et al. (2024). Algorithmic bias in healthcare systems. [Fairness in medical AI]

Additional references from the CIBMTR competition documentation, project guidelines, and established standards in systems engineering and data science are incorporated throughout the report.

Chapter 16

Acknowledgements