UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Universidad Distrital Francisco José de Caldas

Computer Engineering Program
School of Engineering

Workshop No. 2 – 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 for Workshop No. 2 in Systems Analysis & Design
Semester 2025-III

October 2025, Bogotá D.C.

Contents

List of Figures

Chapter 1

System Design

1.1 Review of Workshop No. 1 Findings

1.1.1 Summary of Systems Analysis

In Workshop 1, we conducted a comprehensive systems analysis of the CIBMTR Kaggle competition focused on equity in post-HCT survival predictions. Our analysis revealed the complex nature of the medical system surrounding hematopoietic cell transplantation procedures, characterized by multiple interconnected clinical, genetic, and demographic factors that influence patient outcomes.

Key findings from our analysis include:

- **System Components:** The system comprises diverse data sources including disease characteristics, transplant-specific inputs, demographic factors, and temporal variables, all contributing to survival predictions.
- **High Complexity:** Post-HCT survival prediction involves nonlinear interactions between variables, forming a high-dimensional system where small parameter changes can lead to significant outcome differences.
- **Sensitivity Analysis:** We identified critical sensitivity parameters including patient age, disease risk indices, genetic compatibility scores, and comorbidities, where minor variations can dramatically alter survival predictions.
- **Chaos Theory:** Medical outcomes following HCT exhibit characteristics of chaos theory and inherent randomness, where small inaccuracies in measurements can lead to unpredictable outcomes.
- **Equity Considerations:** The stratified C-index metric ensures model performance is consistent across ethnic subgroups, highlighting the need for fairness-aware modeling approaches.

Our analysis emphasized the importance of addressing both the technical challenges of accurate prediction and the ethical imperatives of ensuring equitable performance across diverse patient populations.

1.1.2 Critical Constraints

Our analysis from Workshop 1 identified several critical constraints that must be addressed in our system design:

- **Data Limitations:** The competition restricts participants to use only the provided CIBMTR datasets, without external data sources. This constraint limits our ability to supplement missing information or incorporate additional variables that might improve predictive performance.

- **Fairness Requirements:** The system must meet stringent equity criteria, ensuring similar performance across different demographic groups as measured by the stratified C-index. This requires explicit fairness considerations throughout the modeling pipeline.

- **Clinical Validity:** Predictions must maintain medical relevance and interpretability for real-world application. Solutions that achieve high mathematical accuracy without clinical meaningfulness would fail to meet the competition's underlying healthcare objectives.

- **Missing Data Patterns:** The dataset contains significant missing values with patterns that may vary systematically across demographic groups. The system must handle missing data in ways that don't amplify existing disparities.

- **Computational Feasibility:** While complex modeling approaches are needed to capture the system's nonlinear nature, solutions must remain computationally feasible for practical implementation and evaluation within the competition framework.

- **Temporal Validity:** Medical practices evolve over time, and models must account for potential shifts in treatment protocols and outcomes across the time span represented in the dataset.

These constraints directly inform our architectural decisions and implementation strategies, creating boundaries within which we must develop our solution.

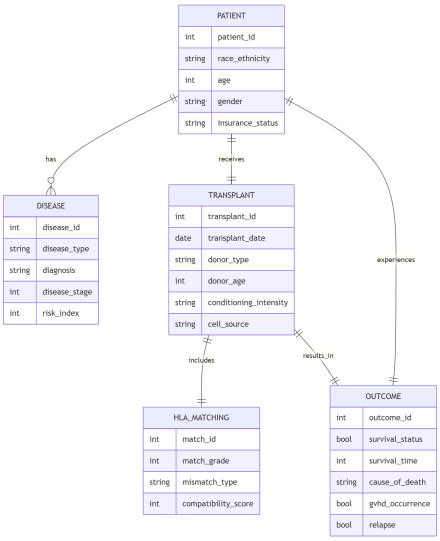

Figure 1.1: Entity-relationship diagram of the CIBMTR dataset structure showing the relationships between key data entities: patients, diseases, transplant procedures, HLA matching, and outcomes. This visualization maps how different data components are interconnected, highlighting the complexity of the dataset and providing context for the data constraints faced in prediction modeling.

1.2 System Requirements

1.2.1 Measurable Design Requirements

- **Accuracy:** Production of results with the highest possible quality, enabling the fulfillment of user-centered needs.
- **Resilience:** Ability to maintain acceptable performance despite variations in the quantity or quality of input variables established for the system. This allows the system to properly handle generated entropy.
- **Efficiency:** Reduction in the amount of time and system resources without losing the established homeostasis in the system.
- **Scalability:** Simplicity when increasing the amount of data entered into the system.

1.2.2 User-Centric Needs

- **Accuracy in post-HCT survival prediction**
- **Equity:** The model should not discriminate based on demographic or socioeconomic characteristics.
- **System resilience:** To handle missing data, outliers, and biological variability.

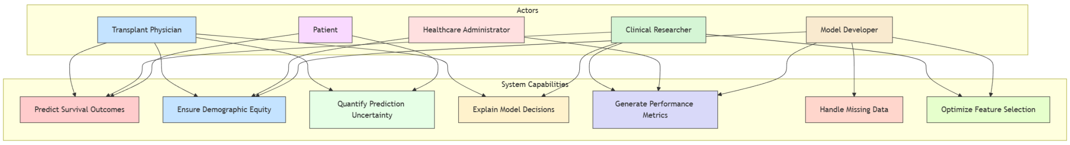

Figure 1.2: Use case diagram showing how different stakeholders interact with the prediction system. This visualization connects the technical capabilities of the system with the specific needs of transplant physicians, clinical researchers, healthcare administrators, model developers, and patients, illustrating how each user group utilizes different aspects of the system's functionality.

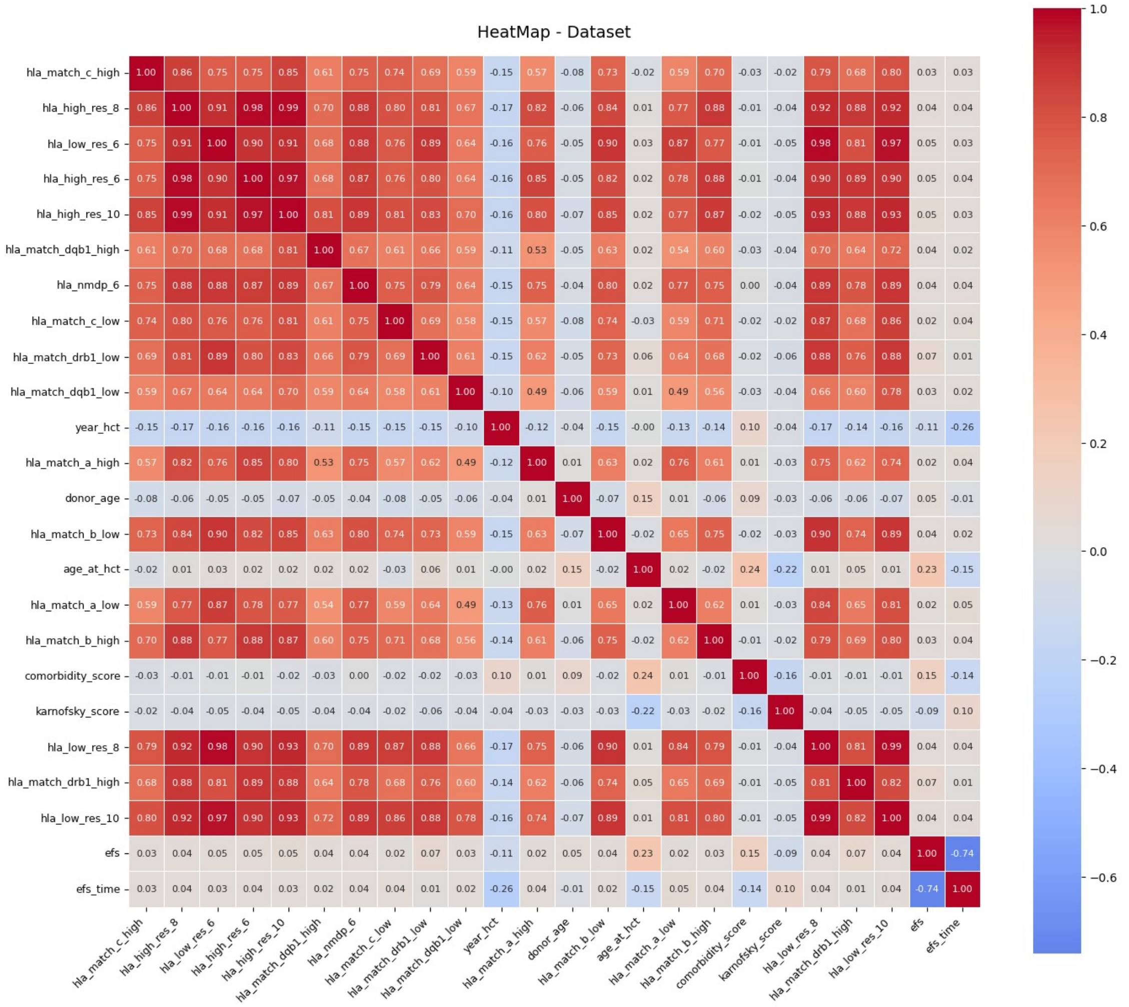

Figure 1.3: Heatmap visualization showing correlations between numerical variables in the dataset. This visualization facilitates understanding of the system's numerical input data to identify potential redundancies. By revealing which variables are highly correlated with each other, the heatmap provides insight into variable relationships that may influence system performance. This understanding allows us to prioritize relationships that contribute most effectively to the established system requirements, particularly with respect to equity considerations across demographic groups.

1.2.3 Evaluation Metric and Fairness Assessment

The stratified C-index serves as the primary evaluation metric for this system, designed specifically to ensure equity across demographic groups. This metric is central to the competition objectives and directly informs our architectural decisions.

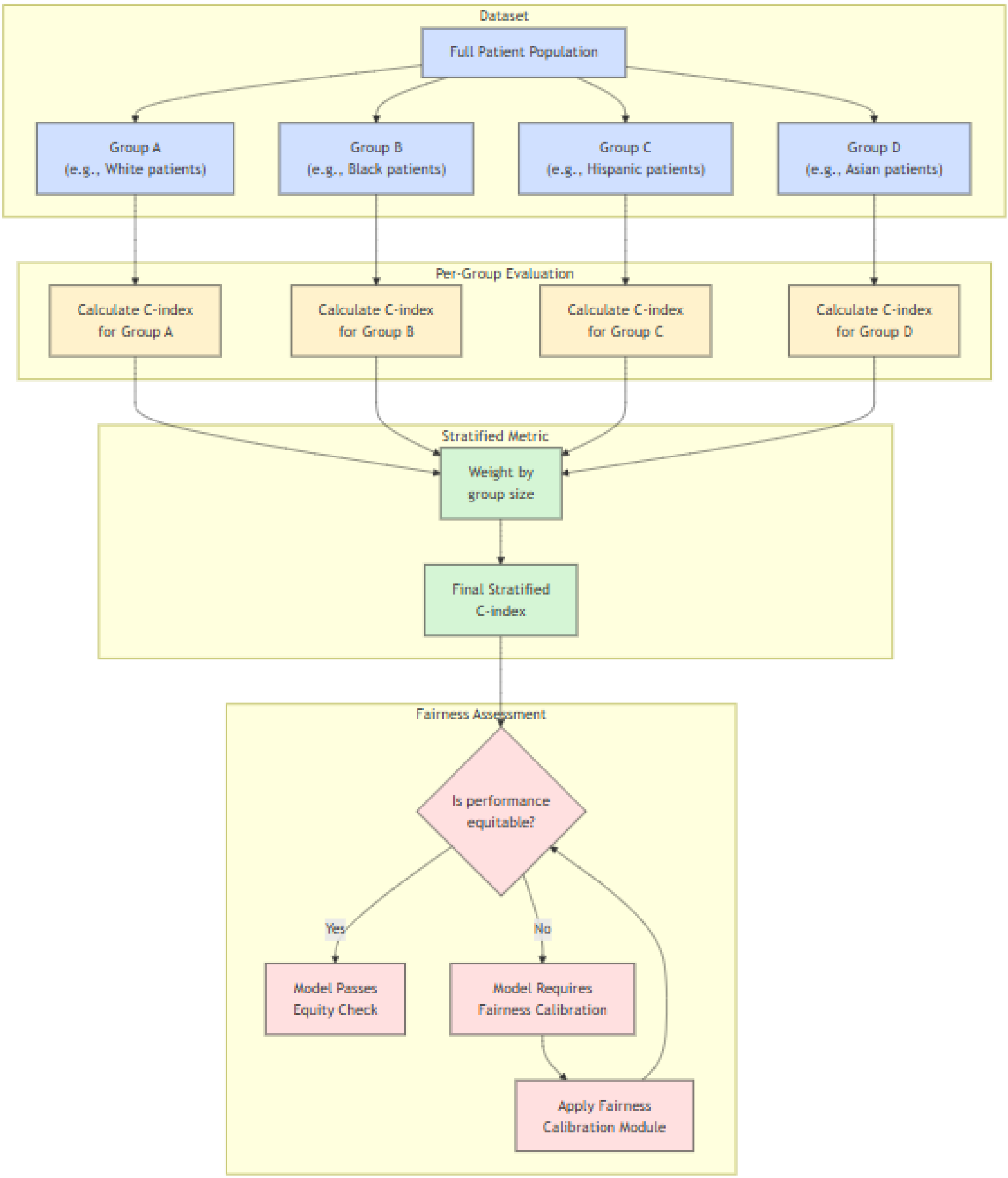
Dataset
Full Patient Population
Group A (e.g., White patients)
Group B (e.g., Black patients)
Group C (e.g., Hispanic patients)
Group D (e.g., Asian patients)
Per-Group Evaluation
Calculate C-index for Group A
Calculate C-index for Group B
Calculate C-index for Group C
Calculate C-index for Group D
Stratified Metric
Weight by group size
Final Stratified C-index
Fairness Assessment
Is performance equitable?
Yes
No
Model Passes Equity Check
Model Requires Fairness Calibration
Apply Fairness Calibration Module

Figure 1.4: Workflow for calculating the stratified C-index and performing fairness assessment. The process begins with stratification of the patient population by demographic groups, followed by separate C-index calculation within each group. These group-specific metrics are then weighted according to group size to produce the final stratified C-index. The fairness assessment loop ensures models deliver equitable performance across all patient populations, triggering fairness calibration when necessary.

This evaluation approach addresses a critical challenge in healthcare prediction: models may perform well on average while performing poorly for specific demographic groups. By requiring consistent performance across all patient populations, the stratified C-index incentivizes fair prediction systems that do not disadvantage underrepresented groups. Our architecture explicitly addresses this requirement through dedicated modules for equity analysis and fairness calibration.

1.2.4 User Stories and Requirements Prioritization

User Stories

We identified the following user stories to guide our system design:

- **As a transplant physician**, I need accurate survival predictions stratified by demographic groups so that I can provide equitable prognostic information to all my patients regardless of their background.
- **As a clinical researcher**, I need to understand which factors most influence survival predictions so that I can identify potential interventions that might improve outcomes.
- **As a healthcare administrator**, I need performance metrics that demonstrate equitable model performance across populations so that I can ensure our institution delivers fair care.
- **As a model developer**, I need robust evaluation frameworks so that I can verify my algorithms maintain accuracy across different demographic subgroups.
- **As a patient**, I need trustworthy survival estimates with appropriate uncertainty bounds so that I can make informed decisions about my treatment options.

Requirements Prioritization (MoSCoW)

- **Must Have:**
  - Equity across demographic groups (measured by stratified C-index)
  - Accurate survival predictions (minimum C-index of 0.70)
  - Clinical validity of identified risk factors
  - Appropriate uncertainty quantification
- **Should Have:**
  - Interpretability of model predictions
  - Efficient computational performance
  - Handling of complex missing data patterns
  - Identification of potential biases in input data
- **Could Have:**
  - Integration capabilities with clinical workflows
  - Advanced visualization of prediction uncertainty
  - Personalization of risk thresholds

  - Continuous learning capabilities
- **Won't Have (This Version):**
  - Integration with electronic health records
  - Real-time prediction updates
  - Patient-facing interfaces
  - External data augmentation

Requirements Traceability

To ensure all requirements are properly addressed by our architecture, we map key requirements to specific system components:

Requirement	**Preproc.**	**Equity Analysis**	**Feature Selection**	**Modeling**	**Fairness**	**Outputs**
Equity across groups	✓	✓	✓	✓	✓	✓
Clinical interpretability	✓		✓			✓
Handling missing data	✓	✓				
Uncertainty estimation				✓	✓	✓
Fairness across demographics	✓	✓	✓		✓	✓

Table 1.1: Requirements traceability matrix showing how each system component addresses specific user and technical requirements. This visualization ensures all requirements are properly addressed within the system architecture.

1.3 High-Level Architecture

1.3.1 Architectural Overview

The system architecture is founded on a modular and sequential pipeline that aims to optimize clinical accuracy and equity, responding to the complexity and high sensitivity identified in Workshop No. 1.

The architecture of the system is organized into modules representing the main stages, from the ingestion of raw data (clinical, genetic, and demographic data) to the generation of predictions and equity metrics for clinical decision support.

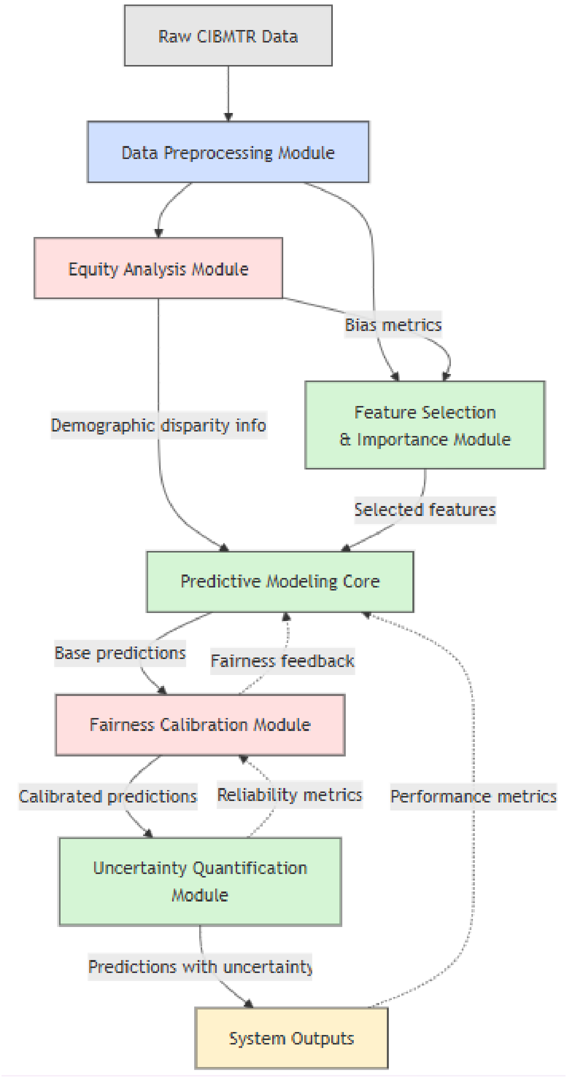

Figure 1.5: System architecture flowchart showing the seven core modules, data flow between components, and feedback loops. The arrows indicate the primary information flow, while dotted lines represent feedback mechanisms that enable continuous optimization. This design explicitly incorporates equity considerations throughout the prediction pipeline, with color-coding to distinguish different functional areas: blue for preprocessing, red for equity-focused components, green for modeling, and yellow for system outputs.

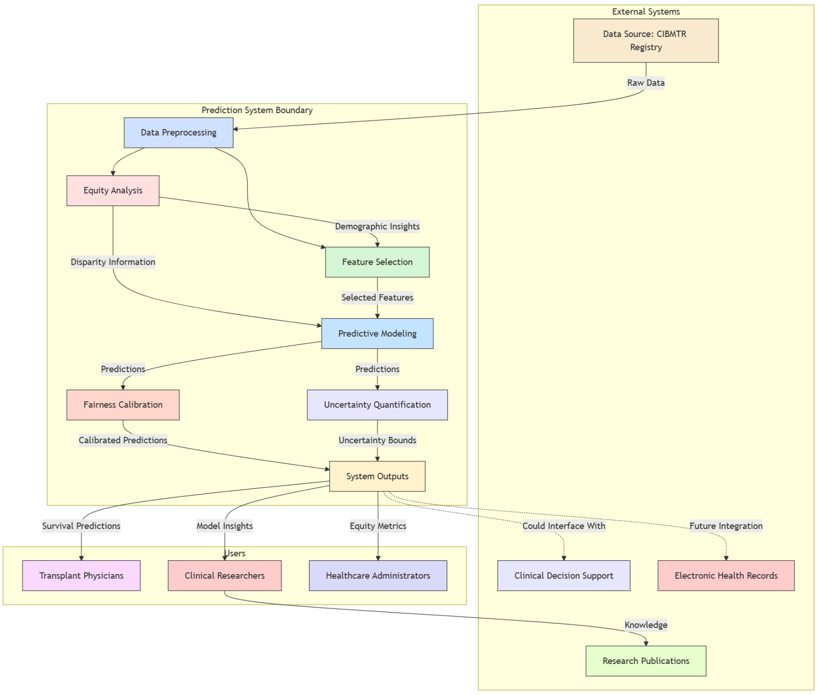

Figure 1.6: System context diagram showing the prediction system's boundaries and its interactions with external entities. The central area represents components within our system boundary, while external systems (data sources, clinical decision support) and users (physicians, researchers, administrators) are positioned outside. This visualization clarifies what is within the scope of our system versus what it interfaces with externally.

1.3.2 Modules and Responsibilities

The architecture consists of seven interconnected modules, each with a specific responsibility:

1. Data Preprocessing Module

Internal Component	**Specific Operations**
Ingestion and Validation	Verifies the integrity and format of raw CIBMTR data.
Equity-Aware Imputation	Handles missing and atypical data. Uses methods that consider demographic variations.
Standardization and Normalization	Scales and transforms continuous variables (e.g., Age, laboratory values) for ML models.
Basic Feature Engineering	Creation of interaction features and logarithmic transformation for skewed variables.

Table 1.2: Data Preprocessing Module Components and Operations

2. Equity Analysis Module

Internal Component	Specific Operations
Stratified Analysis	Performs comprehensive examination of data in demographic subgroups (e.g., Race/Ethnicity) to identify disparities in baseline outcome rates.
Bias Detection Algorithms	Applies algorithms to measure bias in the quality or availability of input features.
Balancing Techniques	Uses fairness-aware preprocessing techniques such as reweighting to improve balanced representation.

Table 1.3: Equity Analysis Module Components and Operations

3. Feature Selection and Importance Module

Internal Component	Specific Operations
Clinical Domain Integration	Prioritizes sensitive variables identified (Age, Disease Risk Index, Genetic Compatibility).
ML/Statistical Rankings	Ranks feature importance using methods such as Recursive Feature Elimination or statistical significance.
Equitable Availability Verification	Ensures that critical predictive features are consistently available across all populations.

Table 1.4: Feature Selection and Importance Module Components and Operations

4. Predictive Modeling Core

Internal Component	Specific Operations
Proportional Hazards Models (Cox)	Base models for time-to-event survival predictions.
Ensemble Algorithms	Uses Gradient Boosting Machines (GBMs) and Random Forests to capture non-linear patterns.
Deep Learning Architectures	Implements neural models for complex non-linear relationships, when necessary.
Cross-Validation Strategy	Executes demographically stratified cross-validation adapted for survival data.

Table 1.5: Predictive Modeling Core Components and Operations

5. Fairness Calibration Module

Internal Component	**Specific Operations**
Probability Calibration	Adjusts survival curves to maintain comparable accuracy across different patient populations.
Risk Threshold Optimization	Modifies risk classification thresholds (low, intermediate, high) to minimize disparity in risk assessment.
Disparity Impact	Quantifies the effect of adjustments to ensure minimization of prediction inequities.

Table 1.6: Fairness Calibration Module Components and Operations

6. Uncertainty Quantification Module

Internal Component	**Specific Operations**
Prediction Interval Calculation	Generates confidence intervals using methods appropriate for survival analysis.
Risk and Reliability	Associates uncertainty bounds with risk stratification.
Low Reliability Identification	Flags cases where prediction is less reliable due to incomplete or atypical data.

Table 1.7: Uncertainty Quantification Module Components and Operations

7. System Outputs

Internal Component	**Specific Operations**
Prediction Generator	Produces survival probabilities and risk stratification categories.
Equity Metrics Dashboard	Displays equity assessment and prediction consistency across groups.
Interpretation Engine (SHAP)	Generates detailed explanations for individual predictions (SHAP values, Feature Importance).
Quality and QA Reports	Delivers performance metrics, calibration plots, and data quality reports.

Table 1.8: System Outputs Components and Operations

1.3.3 System Information Flow

To complement the modular architecture description, we visualize the temporal progression of information through the system with a sequence diagram:

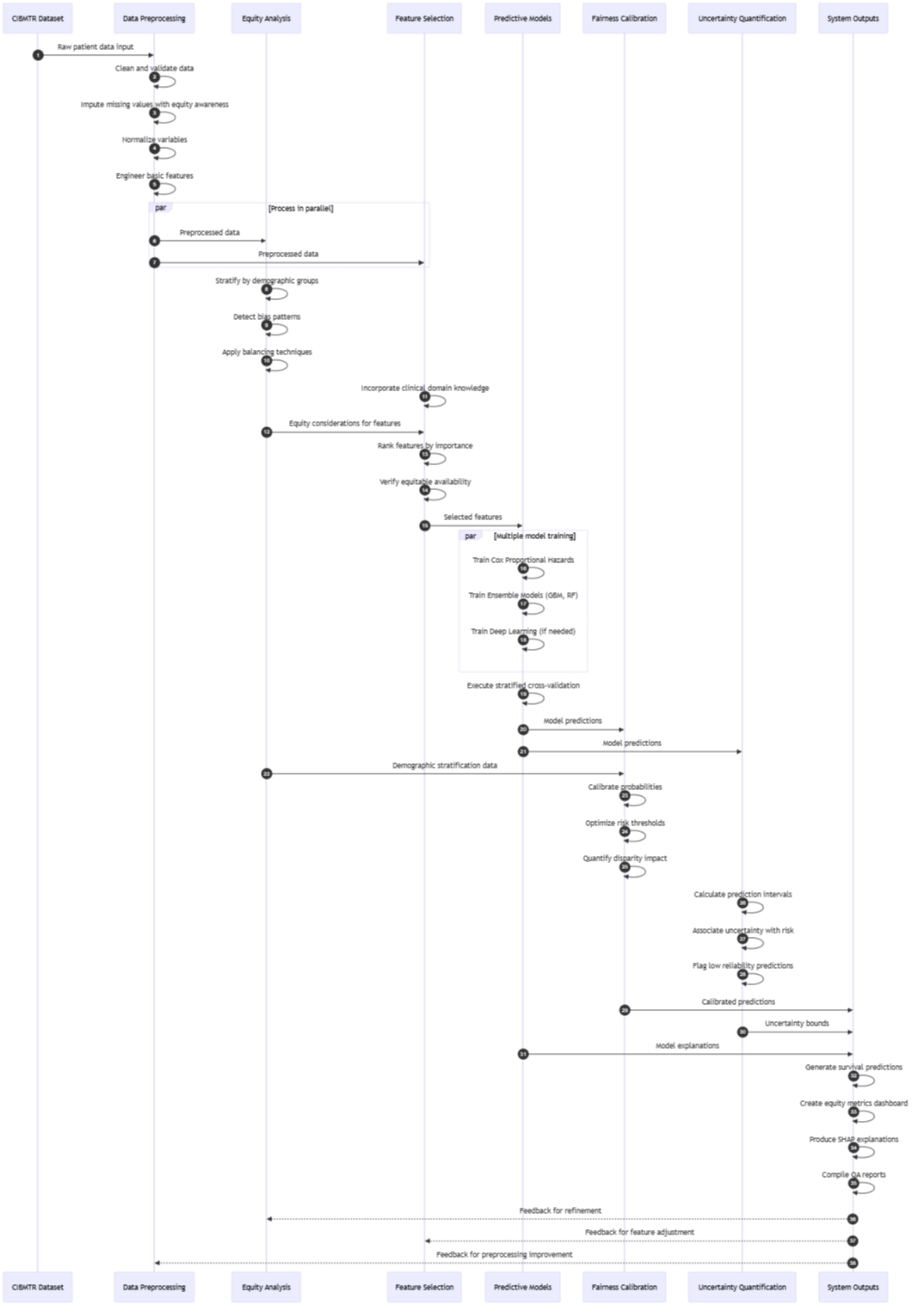
CIBMTR Dataset
Data Preprocessing
Equity Analysis
Feature Selection
Predictive Models
Fairness Calibration
Uncertainty Quantification
System Outputs
1
Raw patient data input
2
Clean and validate data
3
Impute missing values with equity awareness
4
Normalize variables
5
Engineer basic features
par
[Process in parallel]
6
Preprocessed data
7
Preprocessed data
8
Stratify by demographic groups
9
Detect bias patterns
10
Apply balancing techniques
11
Incorporate clinical domain knowledge
12
Equity considerations for features
13
Rank features by importance
14
Verify equitable availability
15
Selected features
par
[Multiple model training]
16
Train Cox Proportional Hazards
17
Train Ensemble Models (GBM, RF)
18
Train Deep Learning (if needed)
19
Execute stratified cross-validation
20
Model predictions
21
Model predictions
22
Demographic stratification data
23
Calibrate probabilities
24
Optimize risk thresholds
25
Quantify disparity impact
26
Calculate prediction intervals
27
Associate uncertainty with risk
28
Flag low reliability predictions
29
Calibrated predictions
30
Uncertainty bounds
31
Model explanations
32
Generate survival predictions
33
Create equity metrics dashboard
34
Produce SHAP explanations
35
Compile QA reports
36
Feedback for refinement
37
Feedback for feature adjustment
38
Feedback for preprocessing improvement
CIBMTR Dataset
Data Preprocessing
Equity Analysis
Feature Selection
Predictive Models
Fairness Calibration
Uncertainty Quantification
System Outputs

Figure 1.7: Detailed sequence diagram illustrating the flow of information through the system pipeline. The diagram shows how data progresses from the CIBMTR dataset through each processing stage, with parallel operations where appropriate and feedback loops enabling continuous improvement. Each numbered step represents a specific operation in the prediction workflow, highlighting the dynamic interactions between modules.

This sequence view demonstrates how information flows through our modules over time, including:

- Parallel processing paths for efficiency
- Key decision points in the pipeline
- Feedback mechanisms that enable system learning and adaptation
- Sequential dependencies between operations

The diagram reinforces how equity considerations permeate the entire workflow, from initial data preprocessing through final output generation, ensuring fairness is maintained throughout the prediction process.

1.3.4 Systems Engineering Principles

The systems engineering principles applied to shape these structural decisions are:

- **Modularity:** The architecture is divided into discrete modules, allowing independent development, testing, and component replacement without affecting overall system integrity.
- **Scalability:** The pipeline design supports complex methods such as ensemble modeling and stratified analysis. This ensures the system can handle both the high volume of data from CIBMTR and the computational complexity required.
- **Maintainability:** Clear documentation and separation of tasks are integrated. This separation facilitates audits and updates in response to new medical advances or equity requirements.
- **Robustness and Complexity Management:**
  - The inclusion of the Uncertainty Quantification Module and the use of ensemble models mitigate the effects of chaotic behavior and high system sensitivity (where small changes in input generate large changes in output).
  - The Fairness Calibration Module reinforces system robustness against latent biases and ensures that the critical stratified C-Index metric is consistently met across all populations.

1.3.5 Complexity and Sensitivity Analysis

Complexity in Post-HCT Survival Modeling

The challenge of predicting survival outcomes after hematopoietic cell transplantation (HCT) centers fundamentally on the system's inherent complexity Auletta, Kou, Chen, and Shaw (2020). Multiple interconnected clinical, genetic, and demographic factors influence patient survival outcomes, creating a nonlinear, high-dimensional system where minor parameter variations can produce significant outcome differences Harrington et al. (2025).

As identified in our analysis, post-HCT survival prediction involves:

- Multifactorial disease characteristics that vary by hematologic malignancy type Salit and Deeg (2024)

- Complex donor-recipient genetic compatibility factors affecting engraftment Shike and Zhang (2024)
- Demographic variables with potential impact on healthcare access and outcomes Doherty, Char, Goodman, Shah, and Oberst (2024)
- Temporal aspects reflecting evolving medical practices over the dataset's time span Zubarovskaya et al. (2023)

Feedback mechanisms such as immune responses, graft-versus-host disease, and intervention-triggered complications further complicate prediction by introducing additional nonlinearity Williams, Chien, Gladwin, and Pavletic (2021).

Sensitivity Analysis

Our analysis identified several parameters where small variations generate disproportionate impact on survival predictions:

- **Age:** Even a single-year difference can shift risk categorization significantly for patients near clinical thresholds Kucab et al. (2024)
- **Disease risk indices:** Minor score changes can alter treatment eligibility and prognosis UpToDate (2025)
- **HLA matching:** Subtle genetic compatibility differences substantially affect transplant rejection rates Krummey and Gareau (2022)
- **Comorbidities:** Presence or absence of specific conditions can dramatically alter survival projections Zubarovskaya et al. (2023)

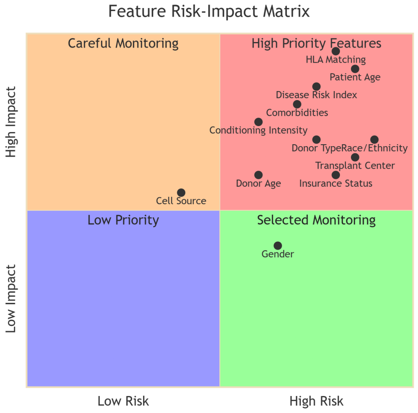

Figure 1.8: Risk-impact matrix for key prediction features, positioning variables according to their predictive power (impact) and potential for introducing bias (risk). High-impact, high-risk features like patient age and HLA matching require particular attention in our fairness calibration process. This visualization guides our feature selection strategy by highlighting which variables need careful handling to balance predictive power against equity considerations.

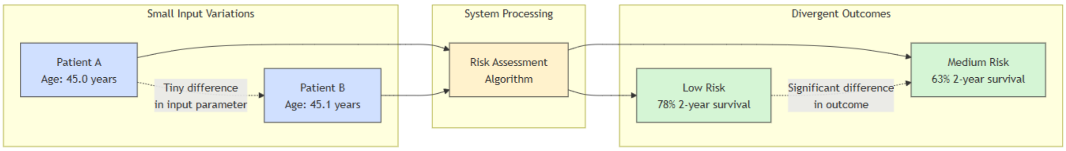

Figure 1.9: Visual representation of sensitivity in post-HCT survival prediction. This diagram illustrates how a minimal difference in patient age (just 0.1 years) can lead to significantly different risk categorization and survival probability estimates. Such high sensitivity to input parameters necessitates specialized modeling approaches that can account for these threshold effects, particularly around clinical decision boundaries.

Chaos and Randomness in Post-HCT Survival Prediction

Medical outcomes following allogeneic HCT exhibit characteristics of chaotic systems, where deterministic processes nonetheless produce seemingly random and unpredictable results Harrington et al. (2025). This chaos manifests in several specific ways:

- **Immune reconstitution dynamics:** Small variations in initial T-cell populations can lead to dramatically different immune recovery trajectories and corresponding survival outcomes Kucab et al. (2024). Studies show that even genetically identical grafts can produce divergent immune reconstitution patterns due to sensitivity to microenvironmental conditions at transplantation.

- **Graft-versus-host disease emergence:** The development and progression of GVHD follows chaotic patterns, with similar patients on identical prophylaxis regimens experiencing vastly different disease trajectories Williams et al. (2021). Minor differences in tissue damage during conditioning or subtle variations in gut microbiota can trigger significantly different inflammatory cascades.

- **Infection susceptibility:** Infection outcomes post-transplant display hallmarks of chaos theory, where minimal differences in pathogen exposure or antimicrobial timing can result in either rapid clearance or life-threatening sepsis UpToDate (2025).

- **Relapse dynamics:** For malignant conditions, disease recurrence follows complex nonlinear patterns influenced by minimal residual disease, graft-versus-leukemia effects, and immune escape mechanisms that exhibit extreme sensitivity to initial conditions Zubarovskaya et al. (2023).

The "butterfly effect" in these systems means that conventional deterministic models struggle to capture the true range of possible outcomes. Our analysis revealed that this unpredictability is further complicated by demographic factors, as certain chaos-inducing variables (like access to prompt care for infections or monitoring for early GVHD signs) may be systematically distributed unequally across patient populations Doherty et al. (2024).

The competition's stratified C-index evaluation metric Kaggle and CIBMTR (2025) presents a particular challenge in this context, as it requires models to maintain consistent performance across demographic subgroups despite these chaotic elements. Achieving both accuracy and equity demands sophisticated approaches that can:

- Quantify uncertainty ranges rather than single-point predictions

- Identify regions of the feature space where chaotic behavior is most likely
- Apply ensemble methods that capture various possible trajectory families
- Maintain calibration across demographic groups even when faced with inherent unpredictability

These considerations directly inform our system design, which employs multiple modeling approaches and explicit uncertainty quantification to address the fundamental chaotic nature of post-HCT outcomes while ensuring equitable performance across patient populations.

1.4 Technical Stack and Implementation Sketch

We have identified a range of widely-used and accessible tools designed to handle everything from data cleaning to results presentation, ensuring that the work can be replicated without difficulties.

1.4.1 Recommended Technology Stack

Data Processing and Management Tools

- **Pandas:** Enables efficient organization and cleaning of information, such as sorting patient tables or transforming medical data GeeksforGeeks (2023b).
- **NumPy:** Facilitates rapid calculations and matrix operations, supporting all statistical processing needs Dataquest (2023).
- **Scikit-learn:** Provides functions for data preparation, filling missing values, and splitting information for model training and validation Pushkarna (2023).

Survival Analysis Libraries

- **Lifelines:** This library assists in analyzing patient survival time. It includes models such as Cox Proportional Hazards and methods to calculate expected time to an event using available information Davidson-Pilon (2023).
- **Scikit-survival:** Enables the use of more advanced survival models and is compatible with other machine learning tools, making the process simpler and more integrated Pölsterl (2023).

Machine Learning Models

- **XGBoost:** Very fast and works well with large data volumes. Additionally, it detects missing values and adjusts calculations if there are fewer cases in one group than another Neptune.ai (2023).
- **LightGBM:** Consumes less memory than other options and has faster training times, ideal for working with many records Ciencia de Datos (2023).
- **CatBoost:** Allows direct use of categorical data without prior conversion, saving time and reducing errors GeeksforGeeks (2023a).

Fairness and Bias Reduction Tools

- **AIF360 (IBM):** Allows fairness measurement using various metrics and applies techniques to reduce bias either before, during, or after model training Trusted AI Team (2023).
- **Fairlearn (Microsoft):** Offers options to adjust models and ensure results are fair across all groups Microsoft (2023).

Model Interpretation Tools

- **SHAP:** Used to explain, through graphics and values, why the model makes certain decisions SHAP (2023).

Data Visualization

- **Matplotlib:** Creates basic or advanced charts to clearly display analysis results Matplotlib (2023).
- **Seaborn:** Enables creation of more attractive and easier-to-interpret statistical graphics, quickly comparing groups or trends Seaborn (2023).

Technical Management and Deployment

- **Docker:** Ensures the project works identically on any computer, avoiding configuration-related errors KDnuggets (2023).
- **PostgreSQL:** Stores all data and results to keep history secure and organized.
- **MLflow:** Records experiments and results of each model version for easy comparison and selection of the best option DataCamp (2023).

1.4.2 Implementation Approach

Development Methodology

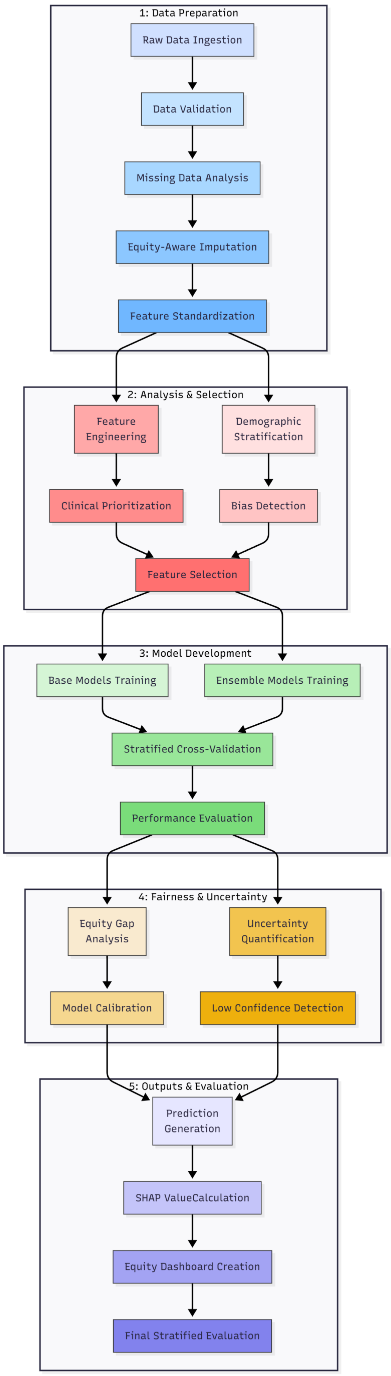

Figure 1.10: Development workflow diagram showing the five implementation phases: Data Preparation, Analysis & Selection, Model Development, Fairness & Uncertainty, and Outputs & Evaluation. Each phase contains specific tasks that must be completed before progressing to subsequent phases, though feedback loops exist between phases for iterative refinement. Color-coding distinguishes different phases of implementation, highlighting both the sequential nature of development and cross-phase dependencies.

Our implementation will follow a structured workflow that addresses both the technical challenges of accurate prediction and the ethical requirements for equity:

1. **Data Preparation:** First, missing data is cleaned and completed, ensuring fair treatment across different groups such as race or ethnicity. Then, values are normalized so all patients are on the same scale.
2. **Fairness Analysis:** Using AIF360, information is converted to a special format and metrics like "Statistical Parity Difference" are calculated to determine if the system treats all groups fairly.
3. **Feature Selection:** The most important variables are chosen using statistical methods and ensemble models that detect which factors most affect survival.
4. **Predictive Modeling:** Several models are trained, including "Cox Proportional Hazards", "Gradient Boosting Survival Analysis", and XGBoost, adjusting parameters when there are fewer cases in some groups.
5. **Fairness Calibration:** Techniques are applied to correct bias after training the model and adjust thresholds to ensure predictions are fair across groups.
6. **Uncertainty Quantification:** Confidence intervals are calculated to determine how certain the model is in its predictions, using repeated simulations (bootstrap).
7. **System Output Generation:** A visual explanation of each prediction is generated with SHAP, showing why the model decides in a particular way and which features have the most influence.

Design Patterns

The implementation incorporates several design patterns to ensure maintainability and scalability:

- **Pipeline Pattern:** All process steps are connected to automate and avoid errors when processing data DataCamp (2023).
- **Strategy Pattern:** The cleaning method, prediction model, or equity techniques can be changed depending on what the study seeks to achieve.
- **Observer Pattern:** With MLflow, every change and result is automatically recorded to monitor the process.

Implementation Considerations

- **Cross-Validation:** To ensure consistent results, data is divided into several groups, mixing the event of interest (such as survival) and demographic characteristics. The model is trained multiple times, and performance in each group is compared using the C-index Scikit-learn (2023).
- **Parallelization:** Computationally intensive processes like cross-validation and ensemble training are parallelized to improve efficiency.
- **Memory Management:** For large datasets, efficient memory management techniques are employed, particularly when working with ensemble models.

- **Data Security:** Although using de-identified data, appropriate security controls are maintained throughout the implementation.
- **Reproducibility:** All random processes use fixed seeds to ensure reproducibility of results.

1.5 Conclusion

In this report, we have presented a comprehensive system design that addresses the dual challenges of accuracy and equity in post-HCT survival predictions identified in Workshop No. 1. Our architecture directly responds to the complex, sensitive, and chaotic nature of the medical domain while maintaining a focus on fair outcomes across demographic groups.

The proposed modular pipeline architecture, consisting of seven interconnected components, systematically addresses each aspect of the prediction challenge:

- The Data Preprocessing and Equity Analysis modules establish a foundation for fairness from the earliest stages of data handling, ensuring demographic disparities are not amplified through the prediction pipeline.
- The Feature Selection and Importance Module prioritizes variables based on both clinical relevance and stability, with special attention to the high-sensitivity parameters identified in our analysis.
- The Predictive Modeling Core employs multiple complementary algorithms that collectively capture the nonlinear patterns essential for accurate survival prediction while mitigating the chaotic elements inherent in post-HCT outcomes.
- The Fairness Calibration Module provides explicit mechanisms to optimize the critical stratified C-index metric by ensuring consistent performance across demographic subgroups.
- The Uncertainty Quantification Module acknowledges and communicates the inherent unpredictability in medical outcomes, enhancing clinical trust and decision-making.
- The System Outputs module delivers not only predictions but explanations and equity metrics that support transparent and fair clinical application.

Our carefully selected technical stack balances state-of-the-art capabilities with practical implementation concerns, ensuring that the sophisticated modeling approaches required for this domain remain computationally feasible and reproducible.

1.5.1 Limitations and Considerations

Despite the comprehensive nature of our design, several important limitations and considerations must be acknowledged:

- **Trade-offs between accuracy and interpretability:** While ensemble methods provide superior predictive performance, they can reduce interpretability compared to simpler models. Our design attempts to balance these concerns through SHAP-based explanations, but this tension remains fundamental to the problem domain.

- **Data constraints:** Working exclusively with the competition dataset limits our ability to incorporate potentially valuable external information. The system design includes robust handling of missing data, but cannot entirely overcome fundamental data limitations.

- **Computational complexity:** The sophisticated modeling approaches necessary for addressing the complex and chaotic nature of post-HCT outcomes require significant computational resources, potentially limiting real-time application in resource-constrained clinical settings.

- **Evolving clinical knowledge:** Medical understanding and practices in HCT continue to evolve, requiring regular updates to model parameters and potentially architectural changes to incorporate new prognostic factors.

- **Ethical dimensions beyond technical fairness:** While our design emphasizes statistical equity across demographic groups, broader ethical considerations in healthcare prediction extend beyond what can be addressed through technical means alone.

1.5.2 Future Enhancements

Looking beyond the current design, several promising directions for future enhancement emerge:

- **Continuous learning systems:** Implementing a feedback loop that incorporates new patient outcomes to continuously refine and recalibrate predictions over time, adapting to evolving medical practices.

- **Expanded demographic considerations:** Extending equity analyses beyond the demographic factors in the current competition to address additional dimensions of potential healthcare disparities.

- **Federated learning approaches:** Developing techniques that allow model training across multiple transplant centers while preserving data privacy, significantly expanding available training data.

- **Temporal modeling improvements:** Incorporating more sophisticated approaches to capture the time-dependent nature of post-transplant complications and interventions.

- **Integration with electronic health records:** Creating interfaces between the prediction system and clinical workflows to facilitate seamless integration into transplant decision-making processes.

- **Patient-specific visualizations:** Developing personalized risk communication tools that effectively convey prediction uncertainty to support shared decision-making between clinicians and patients.

In summary, our system design represents a robust framework for addressing the complex challenges of equitable survival prediction following HCT. By explicitly addressing the sensitivity, chaos, and equity considerations identified in Workshop No. 1, this architecture provides a foundation for models that can deliver both accurate and fair predictions across diverse patient populations.

References

Auletta, J. J., Kou, J., Chen, M., & Shaw, B. E. (2020). Indications for hematopoietic cell transplantation and immune effector cell therapy: Guidelines from the american society for transplantation and cellular therapy. *Biology of Blood and Marrow Transplantation*, *26*(7), 1247–1261. Retrieved from https://www.astctjournal.org/article/S1083-8791(20)30114-2/fulltext doi: 10.1016/j.bbmt.2020.03.002

Ciencia de Datos. (2023). *Forecasting de series temporales con skforecast, xgboost, lightgbm y catboost.* Retrieved 2025-10-18, from https://cienciadedatos.net/documentos/py39-forecasting-series-temporales-con-skforecast-xgboost-lightgbm-catboost

DataCamp. (2023). *Tutorial: Machine learning pipelines, mlops & deployment.* Retrieved 2025-10-18, from https://www.datacamp.com/tutorial/tutorial-machine-learning-pipelines-mlops-deployment

Dataquest. (2023). *Numpy and pandas for data analysis.* Retrieved 2025-10-18, from https://www.dataquest.io/tutorial/numpy-and-pandas-for-data-analysis/

Davidson-Pilon, C. (2023). *Lifelines: Survival analysis in python.* Retrieved 2025-10-18, from https://lifelines.readthedocs.io/

Doherty, T. S., Char, D. S., Goodman, S. N., Shah, N. H., & Oberst, M. (2024). Addressing ai algorithmic bias in health care. *JAMA*. Retrieved from https://jamanetwork.com/journals/jama/fullarticle/2823006 doi: 10.1001/jama.2024.16564

GeeksforGeeks. (2023a). *Gradientboosting vs adaboost vs xgboost vs catboost vs lightgbm.* Retrieved 2025-10-18, from https://www.geeksforgeeks.org/machine-learning/gradientboosting-vs-adaboost-vs-xgboost-vs-catboost-vs-lightgbm/

GeeksforGeeks. (2023b). *Python/data processing with pandas.* Retrieved 2025-10-18, from https://www.geeksforgeeks.org/python/data-processing-with-pandas/

Harrington, P., de Lima Zuchner, C., McGowan, R., Gormley, N., Hill, Q. A., & Snowden, J. A. (2025). Editorial: Improving stem cell transplantation delivery using computational modelling. *Frontiers in Immunology*, *16*. Retrieved from https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2025.1579353/full doi: 10.3389/fimmu.2025.1579353

Kaggle and CIBMTR. (2025). *Cibmtr - equity in post-hct survival predictions.* Retrieved from https://www.kaggle.com/competitions/equity-post-HCT-survival-predictions (Accessed: September 2025)

KDnuggets. (2023). *Build your own simple data pipeline with python and docker.* Retrieved 2025-10-18, from https://www.kdnuggets.com/build-your-own-simple-data-pipeline-with-python-and-docker

Krummey, S. M., & Gareau, A. J. (2022). Donor specific hla antibody in hematopoietic stem cell transplantation: Implications for donor selection. *Frontiers in Immunol-*

ogy, *13*. Retrieved from https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2022.916200/full doi: 10.3389/fimmu.2022.916200

Kucab, M., Mitchell, T. J., Ariffin, H., Campbell, P. J., Rahman, N. A., Scelo, G., ... Prokunina-Olsson, L. (2024). Characterization of clonal dynamics using duplex sequencing in diagnoses and relapses of acute myeloid leukemias. *Science Translational Medicine*, *16*(770), eado5108. Retrieved from https://www.science.org/doi/10.1126/scitranslmed.ado5108 doi: 10.1126/scitranslmed.ado5108

Matplotlib. (2023). *Matplotlib: Visualization with python.* Retrieved 2025-10-18, from https://matplotlib.org/

Microsoft. (2023). *Fairlearn.* Retrieved 2025-10-18, from https://fairlearn.org/

Neptune.ai. (2023). *When to choose catboost over xgboost or lightgbm.* Retrieved 2025-10-18, from https://neptune.ai/blog/when-to-choose-catboost-over-xgboost-or-lightgbm

Pushkarna, C. (2023). *Data preprocessing with pandas, numpy, scikit-learn.* Retrieved 2025-10-18, from https://www.kaggle.com/code/chetalipushkarna/data-preprocessing-with-pandas-numpy-scikit-learn

Pölsterl, S. (2023). *Introduction to scikit-survival.* Retrieved 2025-10-18, from https://scikit-survival.readthedocs.io/en/stable/user_guide/00-introduction.html

Salit, R. B., & Deeg, H. J. (2024). Transplant in all: who, when, and how? *Hematology*, *2024*(1), 93–100. Retrieved from https://ashpublications.org/hematology/article/2024/1/93/526246/Transplant-in-ALL-who-when-and-how doi: 10.1182/hematology.2024000503

Scikit-learn. (2023). *Cross-validation: evaluating estimator performance.* Retrieved 2025-10-18, from https://scikit-learn.org/stable/modules/cross_validation.html

Seaborn. (2023). *Seaborn: Statistical data visualization.* Retrieved 2025-10-18, from https://seaborn.pydata.org/

SHAP. (2023). *An introduction to explainable ai with shapley values.* Retrieved 2025-10-18, from https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

Shike, H., & Zhang, A. (2024). Hla and non-hla factors for donor selection in hematopoietic stem cell transplantation with post-transplant cyclophosphamide gvhd prophylaxis. *Cells*, *13*(24), 2067. Retrieved from https://www.mdpi.com/2073-4409/13/24/2067 doi: 10.3390/cells13242067

Trusted AI Team. (2023). *AI Fairness 360 (AIF360).* Retrieved 2025-10-18, from https://github.com/Trusted-AI/AIF360

UpToDate. (2025). *Allogeneic hematopoietic cell transplantation: Indications, eligibility, and prognosis.* UpToDate. Retrieved from https://www.uptodate.com/contents/allogeneic-hematopoietic-cell-transplantation-indications-eligibility-and-prognosis (Accessed: September 2025)

Williams, K. M., Chien, J. W., Gladwin, M. T., & Pavletic, S. Z. (2021). Updates in chronic graft-versus-host disease. *Hematology*, *2021*(1), 648–654. Retrieved from https://ashpublications.org/hematology/article/2021/1/648/482938/Updates-in-chronic-graft-versus-host-disease doi: 10.1182/hematology.2021000301

Zubarovskaya, L. S., Moiseev, I. S., Vladovskaya, M. D., Mikhailova, N. B., Morozova, E. V., Bykova, T. A., ... Afanasyev, B. V. (2023). Trends in outcome of hematopoietic stem cell transplantation: 5000 transplantations and 30 years of single-center experience. *Cancers*, *15*(19), 4758. Retrieved from https://www.mdpi.com/2072-6694/15/19/4758 doi: 10.3390/cancers15194758