조혈모세포이식수술의 생존 분석을 위한 머신러닝 및 타겟 변환 기법 비교 분석

류한비[1], 김나현[2], 김지민[2], 조석헌*
[1]연세대학교, [2]숙명여자대학교, *University of California, San Diego (UCSD)

hanbie.ryu@yonsei.ac.kr, nh0126@sookmyung.ac.kr, moonlight@sookmyung.ac.kr,
*justinshcho@gmail.com

Benchmarking Machine Learning and Target Transformation Methods for Survival Analysis in Post-HCT Risk Prediction

Hanbie Ryu[1], Nahyun Kim[2], Jimin Kim[2], Seokheon Cho*
[1]Yonsei University, [2]Sookmyung Women's University,
*University of California San Diego (UCSD)

Abstract

Machine learning (ML) methods, such as Random Forest and Gradient Boosting, have long been the standard for analyzing tabular data, including survival analysis. In this study, we aim to predict relative risk measurements in post-hematopoietic cell transplantation (Post-HCT) patient data by benchmarking established ML methods. Our experimental results demonstrate that a Random Forest model using a Rank-based transformation scored the highest on prediction metrics (Race-Stratified C-Index ≈ 0.64). Further feature importance analysis using scikit-learn emphasized the clinical significance of disease severity, donor/patient age, and comorbidities in Post-HCT survival. Notably, restricting training to the most influential features showed comparable performance to standard prediction models, demonstrating the potential for a more efficient yet effective approach for survival analysis.

Ⅰ. 서 론

조혈모세포는 정상적인 피를 만들어내는 어미 세포로서, 각종 혈구의 중간 과정인 전구세포를 거쳐 혈액 내의 적혈구, 혈소판, 백혈구 등으로 분화할 수 있는 능력을 갖추고 있다. 조혈모세포는 골수, 혈액, 제대혈에 존재하며, 이를 이식하는 것을 조혈모세포이식(造血母細胞移植, Hematopoietic Stem Cell Transplantation)이라고 한다 [1]. 조혈모세포이식 이후 발생하는 적응증은 꾸준히 증가하는 추세를 보이고 있으며, 처음에는 백혈병과 재생불량성빈혈 및 면역결핍성 질환에 적용되었으나 점차 림프종, 고형종양, 유전질환으로 확장되고 있다. 조혈모세포 이식의 적응은 병의 종류, 전신상태, 공여자의 유무 등을 보고 판단해야 하기에, 이에 대한 지속적인 연구가 요구되는 상황이다 [2]. 본 연구에서는 조혈모세포이식 후 환자 정보를 담은 '우측 검열 사건 발생 시간 데이터 (right-censored time-to-event data)' 에서 샘플별 상대적인 위험 정도를 예측하기 위해 생존 분석 이론을 적용하여 위험 척도를 설정하고, 다양한 머신러닝 (ML) 모델을 적용하여 성능을 비교·분석하였다. 더 나아가, 개별 환자의 생체 데이터를 활용하여 이식 후의 적응 능력을 정밀하게 평가하는 가능성을 탐색하며, 머신러닝 기반 예측 모델이 향후 이식 환자 관리 및 맞춤형 치료 전략 수립에 기여할 수 있음을 제시하고자 한다.

본 연구는 Centre for International Blood and Marrow Transplant Research (CIBMTR) 에서 주최하는 대회 (Equity in post-HCT Survival Predictions) 에서 제공된 데이터셋을 기반으로 얻은 실험 데이터를 정리한 것이다 [3]. 규정을 위반하지 않도록 대회의 목적으로 데이터셋을 이용하였음을 밝힌다.

Ⅱ. 데이터세트 구성 및 데이터 전처리

본 연구는 SurvivalGAN으로 생성된 28,800 sample의 인공 데이터셋으로 진행되었다. CIBMTR에서 의료 데이터의 민감성을 감안하여 실제 의료 기록을 기반으로 생성한 공식 데이터이기 때문에 연구 목적에 적합하다고 판단되었다.

본 연구에서 활용된 데이터셋은 총 59개의 고유한 변수로 구성되어 있으며, 데이터 타입은 27개의 float64형, 16개의 int64형, 16개의 categorical 변수로 이루어져 있다. 각 변수는 환자의 기본 정보, 임상 점수 및 건강 상태, 질환 및 건강 문제, HLA 및 면역 관련 변수, 환자 및 공여자 정보, 치료 및 이식 관련 변수, 환경 및 기타 정보, 치료 약물 및 면역 관련 변수 등을 포함하고 있다. 각 환자의 생존 여부와 시간에 대한 정보는 '우측 검열 사건 발생 시간 데이터 (right-censored time-to-event data)'에서의 EFS 및 EFS Time 형태로 주어지며, 이에 대한 설명은 **3.1**에서 다루도록 한다.

최대한 정제된 훈련 데이터셋을 확보하기 위해 전처리 과정을 세 단계로 구분하여 수행하였다. 이러한 전처리 절차를 통해 두 개의 데이터셋을 생성하였으며, 각각의 데이터셋을 활용하여 머신러닝 실험을 진행하였다.

2.1. Feature Selection

이 과정은 EDA에서 무의미하다고 판단되거나, 결측값이 과하게 존재하여 모델의 예측 성능을 저하시킬 것으로 예상되는 feature를 제거하는 과정이다. 또한, 'hla_nmdp_6' 등 다른 feature 간의 연산으로 계산할 수 있는 feature 또한 차원 축소를 위해 제거 대상에 포함시켰다.

Feature	Type	Missing Value (%)
hla_match_a_low	Numerical	8.3%
hla_match_b_low	Numerical	8.91%
hla_match_c_low	Numerical	9.72%
hla_match_drb1_low	Numerical	9.18%
hla_match_dqb1_low	Numerical	14.56%
hla_high_res_6	Numerical	18.35%
hla_low_res_6	Numerical	11.35%
hla_nmdp_6	Numerical	14.57%
hla_high_res_8	Numerical	20.24%
hla_low_res_8	Numerical	12.68%
hla_high_res_10	Numerical	24.87%
hla_low_res_10	Numerical	17.58%
psych_disturb	Categorical	7.67%
tce_imm_match	Categorical	38.66%
cyto_score_detail	Categorical	45.52%
mrd_hct	Categorical	57.63%
tce_match	Categorical	65.96%
tce_div_match	Categorical	39.57%

표 1. Feature Selection 과정에서 제거된 feature의 목록

2.2. N/A Unification

이는 데이터 차원을 축소시켜 모델의 효율을 개선시키기 위해 정보가 없는 불필요한 Feature value를 N/A로 통합시키는 과정이다.

데이터 전처리 과정에서 결측값 처리를 일관되게 수행하였다. 범주형 변수에서 해당 feature에 대해 정보를 제공하지 않는 값은 모두 'N/A'로 통일하였으며, 연속형 변수의 결측값(missing value)은 null 값으로 유지하였다. 이를 통해 데이터의 정합성을 확보하고 모델 학습 과정에서 결측값 처리를 일관되게 적용할 수 있도록 하였다.

2.3 Sample Selection

위 두 과정을 거친 데이터셋을 대상으로 N/A 또는 Null value가 포함된 sample을 제거하여 전처리 과정을 마쳤다. 이로서 (8061, 42)의 형태를 갖는 정제된 데이터셋(CFD, Constructed Feature Dataset)을 완성하였으며, 이를 활용하여 이후 연구를 진행하였다.

Ⅲ. 머신러닝 알고리즘 및 성능 평가 지표

본 연구에서는 조혈모세포이식(HCT) 적합성을 예측하기 위해 회귀(regression) 기반의 모델들을 활용하였다. 회귀모델이 EFS와 EFS time으로부터 샘플의 위험 정도를 예측하기 위해서는 두 변수를 정량화된 위험 지표로 통일하는 과정이 필요하다. Random Survival Forest (RSF) 모델 (Ishwaran et al., 2008)에서는 “right-censored data”에 대한 예후(prognosis) 성능을 평가하기 위해 Harrell's Concordance Index (C-index)를 활용하며, 또한 Nelson-Aalen Estimator를 이용하여 누적 위험 함수(Cumulative Hazard Function, CHF)를 계산하고 이를 기반으로 Risk Score를 산출한다. 이러한 접근 방식은 생존 확률과 누적 위험도를 효과적으로 활용하여 개별 샘플의 상대적 위험도를 평가할 수 있도록 한다.

본 연구에서는 이러한 개념을 기반으로, 생존 분석 이론을 활용하여 Risk Score를 추정하는 방법을 모사하였다. 이를 통해 기존 RSF 접근법과 유사한 효과를 얻을 수 있도록 하였으며, Kaplan-Meier 생존 곡선 및 누적 위험 함수 변환을 적용하여 샘플별 위험도를 정량화하는 방법을 탐색하였다. 이러한 방법론적 모사는 생존 분석 기반의 예측 모델이 censored data를 다룰 때 가지는 장점을 유지하면서도, 비교적 단순한 방식으로 위험도를 해석할 수 있는 대안을 제공하는 데에 목적이 있다.

최종 평가 지표는 모델이 Risk Score를 얼마나 잘 예측했는지가 아닌, 예측된 샘플별 위험 척도가 실제 생존 데이터와의 관계를 얼마나 잘 대변하는지를 평가하기에, 모델이 예측할 척도인 Risk Score 를 구하는 과정을 적절하게 설정하는 것도 과제의 일부이다. 최종 평가 지표에 관해서는 **3.2**에서 다루도록 한다.

3.1 Target Transformation

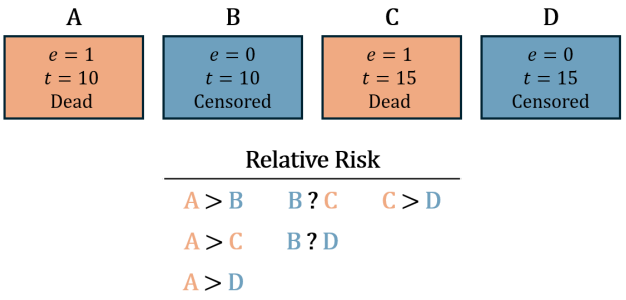

그림 1. 생존 분석에서 사건 시간 데이터의 샘플별 상대적 위험 정도. e: EFS(사건 발생 여부), t: EFS time(관측 시간)

샘플 간의 상대적인 위험 정도는 EFS와 EFS time에 의해 결정된다. EFS는 Event Free Survival을 의미하며, 0 또는 1의 값을 가지고, EFS time은 해당 샘플의 관측 시간을 의미한다.[1] EFS=1 (event)은 조혈모세포이식수술 이후 관측 시간 EFS time 시점에 심각한 부작용 또는 사망 사건으로 이어졌음을 의미하며, EFS=0 (censored)은 이러한 사건이 관측 시간 EFS time까지 발생하지 않았음을 의미한다. 이때, EFS가 0이라고 해서 해당 샘플이 안전한 것은 아니며, 관측 시간 이후에 사건이 발생했을 수 있다. 샘플별 관측 시간은 임의로 정해지기에, 그림 1의 B와 C, B와 D와 같은 경우 위험 정도에 대한 샘플간의 비교가 불가능할 수 있다. 비교 가능한 샘플에 대해서는 더 빨리 사건이 발생한 샘플의 상대적 위험이 더 크다고 정의한다. 이와 같이 일부 샘플의 사건 발생 여부가 불확실한 형식의 데이터를 ‘검열된 사건 발생 시간 데이터 (censored time-to-event data)’ 라고 하며, 이를 분석하기 위해 생존 분석 이론(Survival Analysis Theory)을 적용할 수 있다.

3.1.1. Kaplan-Meier Target Transformation

Kaplan-Meier (KM)추정법은 개별 환자의 사건 발생 시점을 고려하여 생존 확률을 추정하는 비모수적(non-parametric) 방법이다. 이 방법은 특정 시점에서 사건이 발생하지 않고 생존할 확률을 누적으로 계산하며, 생존 곡선(Survival Curve)을 생성하는 데 사용된다. 생존 곡선에 대한 KM 예측 함수 $\hat{S}_{KM}(t)$는 식 (1)과 같이 정의될 수 있다. n_i는 시점 t_i에서 아직 사건이 발생하지 않은 관찰 대상의 수를 의미하며, d_i는 해당 시점에서 사건(EFS=1)이 발생한 대상의 수를 나타낸다.

Kaplan-Meier 추정법은 검출된 사건의 발생 시점과 생존 확률을 고려하여 시간 경과에 따른 생존 패턴을 추정할 수 있다.

$$\hat{S}_{KM}(t) = \prod_{i:t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$$

- t : 특정 시점 (time)
- t_i : 사건이 발생한 시간 중 i 번째 시점
- n_i : 시점 t_i까지 사건이 발생하지 않은 개체의 수
- d_i : 시점 t_i 에서 사건이 발생한 개체 수

이때 누적생존율은 시간에 따라 감소하는 형태를 가지기 때문에, harrell’s C-index의 정의에 의하면 사건이 발생한 샘플에 대해 생존 시간이 짧은 샘플이 더 위험하다고 해석할 수 있다. 각 샘플의 관측 시간 EFS time 에서의 $\hat{S}(t)$를 해당 샘플의 Risk Score η로 간주할 수 있다.따라서, 본 연구에서는 각 샘플의 EFS와 EFS time을 하나의 위험 척도 η로 정량화하는 변환 과정(target transformation) H_{KM}을 아래와 같이 정의한다.

[1] EFS는 Event Free Survival의 축약이지만, 제공된 데이터셋에서 사망/부작용 사건이 발생한 샘플을 EFS=1로 정의하기에 본 연구에서도 EFS=1을 사건 발생 샘플로 정의한다.

$$\eta_{KM} = H_{KM}(e,\ t,\ x_F) \equiv \hat{S}_{KM}(t)$$

e는 EFS, t는 EFS time을 의미하며, x_F는 해당 샘플의 39가지 feature를 의미한다. 이렇게 KM 변환으로 Risk Score η를 구하는 경우, 각 η는 해당 샘플의 EFS time으로부터 결정된다.

그림 2는 Kaplan-Meier (KM) 기반으로 산출한 Risk Score 값의 히스토그램을 나타낸 것이다. 분포를 살펴보면, 검열된 샘플(EFS=0)은 특정 구간(약 0.3 부근)에 집중적으로 분포하는 경향을 보인다. 반면, 사건이 발생한 샘플(EFS=1)의 Risk Score는 전반적으로 고르게 분포하는 모습을 보인다.

히스토그램을 통해 KM 기반 Risk Score가 사건 발생 여부에 따라 차별적인 분포를 가지며, 검열된 샘플은 비교적 낮은 Risk Score를 갖는 반면, 사건이 발생한 샘플은 더 높은 Risk Score를 가지는 경향이 있음을 보여준다.

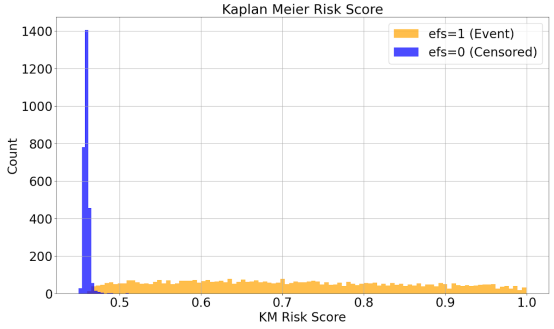

그림 2. Kaplan-Meier Risk Score H_{KM} histogram

3.1.2. Nelson-Aalen Target Transformation

생존 분석 이론에서 Nelson-Aalen Estimator는 특정 시점까지의 누적생존율 대신 누적 위험 지수(Cumulative Hazard Function)을 추정하는 비모수적 방법이다. 이 위험지수로부터 생존률을 계산할 수도 있다.

$$\hat{H}_{NA}(t) = \sum_{t_i \le t} \frac{d_i}{n_i}, \quad \hat{S}_{NA}(t) = e^{-\hat{H}_{NA}(t)}$$

$\hat{H}_{NA}(t)$은 시간 t에 따라 증가하는 함수이기에, Harrell's C-index의 가정에 부합하는 상대적 지표로 변환하기 위해서는 부호를 바꿔야 한다. NA 변환으로 Risk Score를 계산하기 위해 아래와 같은 변환 H_{NA}를 정의하였다.

$$\eta_{NA} = H_{NA}(e,\ t,\ x_F) \equiv -\hat{H}_{NA}(t)$$

그림 3은 Nelson-Aalen(NA) 기반으로 산출한 Risk Score 값의 히스토그램을 나타낸 것이다. NA의 계산 방식에 따라, $\hat{H}_{NA}(t)$의 값은 0에서 시작하여 위험 지수를 누적으로 합하며 수치가 증가하고, 이를 음수로 변환하여 Risk Score로 사용한다. 따라서 X축이 가질 수 있는 최댓값은 0이며, 최솟값은 계산된 누적합 값의 최대치를 음수로 변환한 값이 된다.

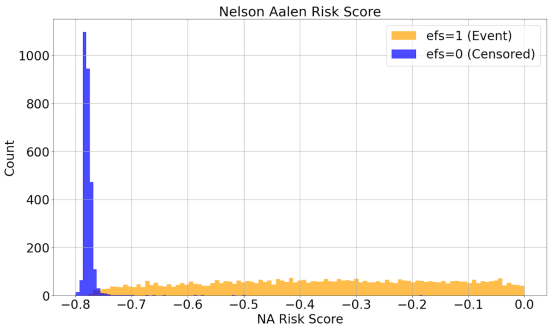

그림 3. Nelson-Aalen Risk Score H_{NA} histogram

3.1.3. Rank-Based Target Transformation

Risk Score는 샘플별 위험 정도를 나타내는 상대적인 척도인 점을 감안하여, Risk Score를 임의로 정의할 수도 있다. 본 연구에서는 Deotte의 MLP Baseline 코드에서 사용한 EFS 여부와 EFS time의 순위를 기준으로 정의된 transformation H_R을 차용하였다 (Deotte, 2025) [5].

Algorithm 1 Chris Deotte's Transform, MLP Baseline

Input:
  EFS_time: List of event times
  EFS: List of event indicators (1 = event, 0 = censored)
Output: $risk\_scores$: Transformed risk scores
1. Shift censored EFS_time
for each i **where** $EFS[i] = 0$ **do**
  $EFS\_time[i] \leftarrow EFS\_time[i] + \max(EFS\_time|EFS = 1)$
    $- \min(EFS\_time|EFS = 0)$
end for
2. Rank EFS_time
$ranks \leftarrow RANK(EFS\_time)$
3. Offset censored ranks
for each i **where** $EFS[i] = 0$ **do**
  $ranks[i] \leftarrow ranks[i] + 2 \times LENGTH(EFS\_time)$
end for
4. Transform ranks
$risk\_scores \leftarrow -\log(ranks/\max(ranks))$
    $+MEAN(\log(ranks/\max(ranks)))$
Return $risk\_scores$

그림 4. Rank-Based Target Transformation의 Pseudocode

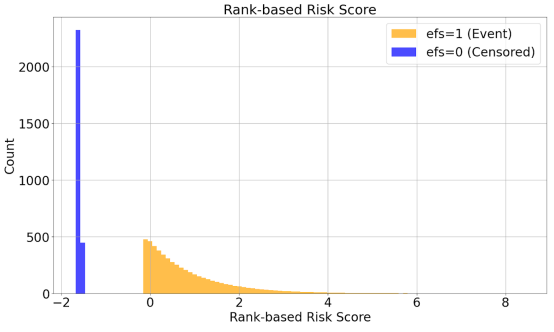

그림 5. Deotte's Rank-Based Risk Score H_R histogram

그림 5는 Rank 기반으로 산출한 Risk Score 값의 히스토그램을 나타낸 것이다. NA의 계산 방식에 따라, 샘플을 EFS를 기준으로 나누고, EFS_time을 기준으로 순위를 매겨 배열하는 형식의 Risk Score를 설계하였을 때, 그림 5와 같이 EFS=0과 EFS=1의 분포가 극명하게 나뉘게 된다. 다만, 그림 1의 샘플 B, C와 같이 판별이 불가능한 두 샘플에 대해서도 극명히 갈리는 타깃을 설정하게 된다는 점에서 이론적인 단점이 존재한다. 이에 대해서는 4.3에서 다루도록 한다.

이 외에도 딥러닝 기법을 활용한 DeepSurv, 환자의 특성(나이, 성별, 임상 변수 등)을 회귀 계수와 내적하여 하나의 값으로 변환한 Cox 비례위험 모형(Cox Proportional Hazards, CoxPH) 을

통해 Risk Score를 계산하는 방식이 존재한다. CoxPH의 경우, 생존 분석에서 일반적으로 사용하는 방법이지만, 특성에 대해 완벽히 선형적인 비례관계를 가지는 데이터가 아닌 경우, 이를 Risk Score로 설정하였을 때 이론적인 성능의 한계가 존재한다.

3.2 성능 평가 지표

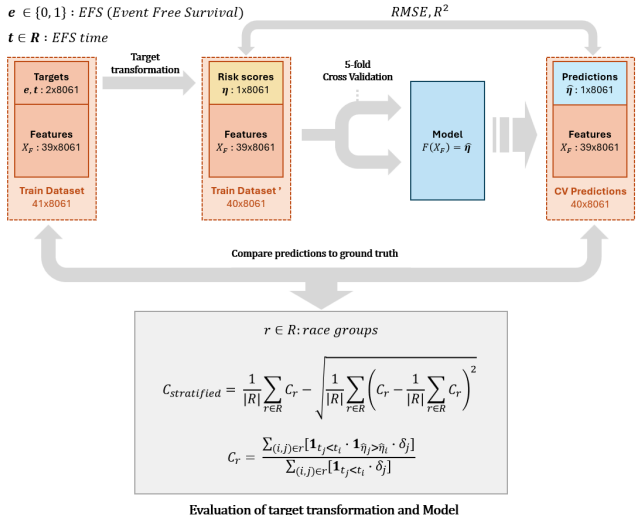

그림 6. Model pipeline

본 연구에서는 모델의 예측 성능을 평가하기 위해 harrell's Concordance Index (C-Index) (harrell, et al., 1982)를 주요 지표로 사용한다 [6]. C-Index는 생존 분석에서 널리 활용되는 평가 척도로, 모델이 환자의 위험도를 정확히 순서화할 수 있는지를 측정한다. 값이 높을수록 모델의 판별력이 우수함을 의미한다. Ground Truth로 간주될 Risk Score를 모델이 얼마나 정확하게 예측했는지 평가하기 위한 지표로는 RMSE 및 R^2를 사용하였다.

Root Mean Squared Error (RMSE)는 예측값과 실제값 간의 평균적인 차이를 측정하는 대표적인 오류 측정 지표이다. RMSE는 각 예측 오차의 제곱을 평균한 후, 그 값의 제곱근을 취하는 방식으로 계산되며, 다음과 같은 수식으로 정의된다. RMSE 값이 작을수록 모델의 Risk Score 예측값이 실제값과 가까움을 의미하며, 보다 높은 예측 정확도를 나타낸다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

C-index는 샘플별 상대적인 위험 수준을 예측한 Risk Score (η)를 기준으로 모델의 성능을 평가하며, 이때 Risk Score의 샘플별 대소관계만이 주요 평가 요소로 작용한다. C-index는 생존 분석에서 모델의 예측 순위가 실제 결과와 얼마나 일치하는지를 평가하는 지표로, 0.5는 무작위 예측, 1은 완벽한 예측을 의미한다.

$$C - index = \frac{\Sigma_{i,j}[1_{t_j<t_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j]}{\Sigma_{i,j}[1_{t_j<t_i} \cdot \delta_j]}$$

정리하자면, C-Index는 상대적인 위험 정도의 비교가 가능한 모든 쌍에 대해 올바르게 Risk Score의 대소 관계를 예측한 샘플의 수를 의미한다.

이때, H_{KM}와 H_{NA}의 경우, Risk Score가 EFS time에 비례하는 관계를 가지기 때문에, 모델이 Risk Score를 완벽히 예측할 경우 이론적인 C-index의 최대치는 1이다.

기존의 C-Index는 전체 데이터에서 모델의 성능을 평가하는 데 유용하지만, 특정 인종 그룹에서 발생할 수 있는 예측 편향을 반영하지 못하는 한계를 가진다. 따라서 본 대회에서는 Race-Stratified C-Index ($C_{stratified}$)를 활용 하여 모델의 인종별 공정성을 평가하며, 본 연구에서도 해당 지표를 동일하게 사용하였다.

$C_{stratified}$를 계산하기 위해 먼저 각 인종 그룹별 C-Index인 C_r을 구한다. 인종 그룹에 대해 C-Index는 다음과 같이 정의된다.

$$C_r = \frac{\Sigma_{(i,j)\in r}[1_{T_j<T_i} \cdot 1_{\hat{\eta}_j>\hat{\eta}_i} \cdot \delta_j]}{\Sigma_{(i,j)\in r}[1_{T_j<T_i} \cdot \delta_j]}$$

이때, 분자는 순응(concordant) 쌍의 개수를 의미하며, 분모는 비교 가능한 전체 쌍의 개수를 나타낸다. 즉, 인종 그룹별로 모델이 얼마나 정확하게 위험도를 순서화하는지를 나타낸다. 여기서 모든 계산은 특정 인종 그룹 r 내에서만 수행되며, 이는 모델이 동일한 인종 그룹 내에서 얼마나 일관된 성능을 보이는지를 평가하기 위함이다.

이후, 각 인종 그룹별 C-Index 값을 조합하여 최종적인 $C_{stratified}$를 계산한다. 이때 단순한 평균을 취하는 것이 아니라, 인종 그룹 간의 성능 편차를 반영하는 보정 항을 포함하여 계산한다.

$$C_{stratified} = \frac{1}{|R|}\sum_{r\in R} C_r - \sqrt{\frac{1}{|R|}\sum_{r\in R}\left(C_r - \frac{1}{|R|}\sum_{r\in R} C_r\right)^2}$$

여기서 $|R|$은 전체 인종 그룹의 개수를 나타내며, 첫 번째 항은 각 인종 그룹에서의 평균 C-Index를 의미한다. 두 번째 항은 인종 그룹 간의 성능 차이를 측정하는 표준 편차를 포함하여, 특정 인종 그룹에서 성능 편차가 클 경우 $C_{stratified}$ 값을 낮추는 역할을 한다.

$C_{stratified}$값이 높을수록, 모델이 모든 인종 그룹에서 균형 잡힌 성능을 보이며, 공정한 예측을 수행하고 있음을 의미한다. $C_{stratified}$ 값이 낮을 경우, 특정 인종 그룹에서 예측 성능이 저하되었음을 나타내며, 모델이 편향되어 있을 가능성이 크다.

기존 C-Index가 모델의 전반적인 성능만을 측정하는 데 반해, $C_{stratified}$는 예측 성능의 공정성까지 포함하여 평가할 수 있도록 한다.

$C_{stratified}$은 초기에 설정한 Risk Score를 모델이 얼마나 잘 예측했는지가 아닌, 모델이 예측한 Risk Score가 실제 샘플간 상대적인 위험 정도를 얼마나 잘 대변하는지를 평가하기 때문에, 적절한 Risk Score를 설계하는 것 또한 과제의 일부이다.

3.3 Random Forest (RF)

본 연구에서는 Risk Score 예측을 위해 Random Forest (RF) 모델을 활용하였다. RF는 여러 개의 결정 트리(Decision Trees)를 결합하여 예측 성능을 향상시키는 배깅(Bagging) 기반의 앙상블 학습 기법이다. 개별 결정 트리는 주어진 데이터의 일부를 무작위로 선택하여 학습하며, 최종 예측은 다수결 투표(분류) 또는 평균 계산(회귀) 방식을 통해 결정된다. 이 방식은 단일 결정 트리보다 높은 예측 성능을 보이나, 과적합(overfitting)에 대한 우려는 여전히 존재한다. 이를 완화시키기 위해 교차 검증(cross-validation)을 활용하여 모델의 일반화 능력을 평가하고자 하였다. Parameter로는 최적 트리 깊이(tree depth), 최소 잎 노드 개수(minimum node size), 트리 개수(number of trees) 를 고려하였다.

Ⅳ. Post-HCT 생존 예측 모델 성능 평가

4.1 모델 학습 및 하이퍼파라미터 설정

본 연구에서는 Random Forest 알고리즘을 활용하여 분석을 수행하였다. 모델의 신뢰성을 확보하고 데이터의 편중을 방지하기 위해 K-Fold 교차 검증(K-Fold Cross Validation)을 적용하였으며, K 값은 5로 설정하였다. 또한, 학습 데이터와 테스트 데이터는 80:20의 비율로 분할하여 모델 성능을 평가하였다.

본 연구에서는 Grid search 기법을 적용하여 최적의 조합을 도출하였다. maxDepth와 minChildSize의 탐색 범위를 설정하고 실험을 진행하였다. Random Forest (RF)에서 KM, NA, Rank 기반의 타깃에 따라 minimum child size (minChildSize), maximum tree depths (maxDepth) 두 파라미터를 대상으로

Grid search를 진행하였다.

Model + Target	minChildSize	maxDepth	Number of Models
KM-based Target	11	29	500
NA-based Target	10	28	500
RANK-based Target	9	31	500

표 2. 세 가지 target 값으로 변형한 모델 (RF_{KM}, RF_{NA}, RF_{RANK})의 하이퍼 파라미터 튜닝 결과

Hyperparameter grid search를 진행하여 표 2와 같은 결과를 얻었다. KM 기반 target 설정 데이터에서 minChildSize = 11, maxDepth = 29일 때 모델의 성능이 가장 우수하였다. NA 기반 target 설정 데이터의 경우 minChildSize = 10, maxDepth = 28에서 최적의 성능을 보였으며, Rank 기반 target 설정 데이터에서는 minChildSize = 9, maxDepth = 31에서 성능이 가장 뛰어났다. 또한, 세 모델의 Number of Models 값은 500으로 통일하여 실험을 진행하였다.

4.2 post-HCT 적합성 예측 모델 결과

본 연구에서는 Random Forest (RF) 기반의 다양한 target 변환 방법을 적용한 모델을 비교하였다. 평가 지표로는 평균제곱근오차(RMSE, Root Mean Square Error) 및 Harrell's C-index를 변환한 $C_{stratified}$ 를 사용하였으며, 세 가지 변형 모델(RF_{KM}, RF_{NA}, RF_{RANK})의 결과를 표 3에 제시하였다.

Model + Target	RF_{KM}	RF_{NA}	RF_{RANK}
Number of Models	500	500	500
Minimum Node Size	11	10	9
Maximum Tree Depth	29	28	31
RMSE	0.1582	0.2447	1.3061
$C_{stratified}$	0.6205	0.6187	**0.6383**

표 3. 세 가지 target 값으로 변형한 모델 (RF_{KM}, RF_{NA}, RF_{RANK})의 RMSE, $C_{stratified}$ 평가 결과

RF_{KM} 모델은 RMSE가 0.1582로 가장 낮아 예측 값과 실제 값 간의 오차가 가장 적었다. 반면, RF_{RANK} 모델은 RMSE가 1.3061로 가장 높았다. 이는 RF_{RANK} 모델의 target 변환 방식이 상대적으로 넓은 값의 분포를 가지는 데 기인한 것으로 보인다. 반면, $C_{stratified}$ 를 기준으로 평가한 모델의 순위는 RMSE와 다소 차이가 있다. RF_{RANK} 모델이 0.6383으로 가장 높은 값을 기록하였다. 이는 RF_{RANK} 방식이 개별 예측값의 정량적 차이보다는 샘플 간 상대적 순위 정보를 보다 효과적으로 보존함을 시사한다.

4.3 한계점

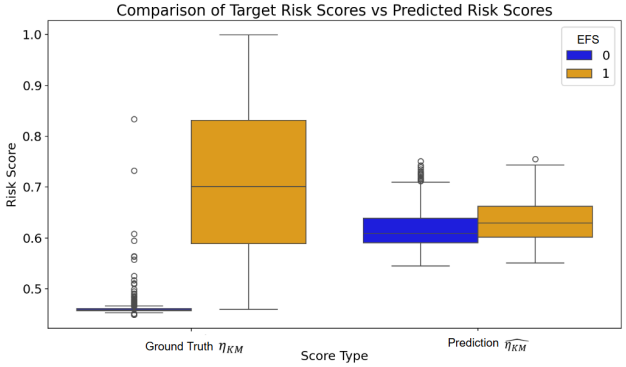

그림 7. H_{KM} target에 대한 EFS 별 ground truth risk score와 prediction risk score 분포

모델이 예측한 Kaplan-Meier risk score η_{KM}는 EFS에 대해 그림 6과 같은 분포를 가졌다. Ground truth와 비교하면 실제로는 EFS=0, 1에 대해 극명하게 나뉘는 분포를 가지지만, 예측값은 하나의 연속된 분포 형태를 가지는 것으로 확인된다. 이는 H_{KM}이 샘플 별로 임의로 주어질 수 있는 EFS Time에 종속적이며, 예측이 불가능한 부분이 있기 때문에 모델이 이를 충분히 재현하지 못한 것으로 파악된다.

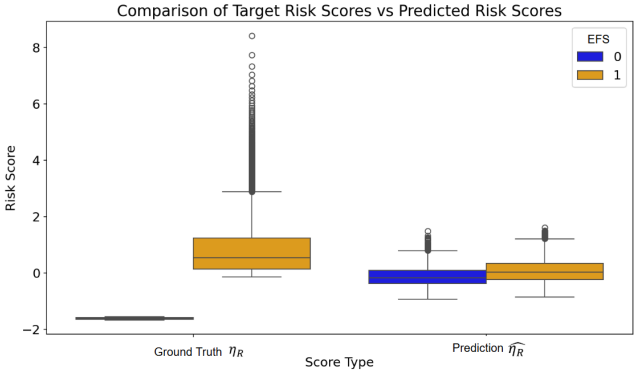

그림 8. H_R target에 대한 EFS 별 ground truth risk score와 prediction risk score 분포

이러한 실제값과 예측값의 분포 차이는 Rank-based transform H_R에서도 확인되었다. H_R의 경우, EFS=0, 1에 대해 분포를 완전히 분리시키기에, 관측이 일찍 끝나 상대적 위험 수준의 비교가 불가능한 샘플의 target 값이 벌어지게 된다. 따라서, 실제로는 샘플들의 위험 수준이 하나의 연속된 분포를 가지나, EFS 여부에 따라 임의의 기준으로 분포가 갈라지기 때문에 모델이 을 완벽히 재현하지 못한 것으로 해석된다.

4.4 Feature Importance 분석 및 SFD 모델 평가

Feature	RF_{KM}	RF_{NA}	RF_{Rank}
dri_score	2.07×10^{-1}	1.92×10^{-1}	1.50×10^{-1}
donor_age	1.15×10^{-1}	1.17×10^{-1}	1.25×10^{-1}
age_at_hct	1.10×10^{-1}	1.20×10^{-1}	1.20×10^{-1}
comorbidity_score	1.01×10^{-1}	1.01×10^{-1}	9.79×10^{-2}
year_hct	7.95×10^{-2}	7.72×10^{-2}	8.69×10^{-2}
karnofsky_score	5.03×10^{-2}	4.43×10^{-2}	5.70×10^{-2}
cyto_score	4.06×10^{-2}	4.04×10^{-2}	3.31×10^{-2}
race_group	3.48×10^{-2}	3.67×10^{-2}	4.00×10^{-2}
sex_match	3.22×10^{-2}	3.60×10^{-2}	3.98×10^{-2}
prim_disease_hct	2.53×10^{-2}	2.69×10^{-2}	—
cmv_status	—	—	2.82×10^{-2}

표 4. 세 가지 target 값으로 변형한 모델(RF_{KM}, RF_{NA}, RF_{RANK})로부터 추출한 상위 Feature 10개

데이터셋의 Feature 수에 따른 모델의 성능 차이를 비교하기 위해 사이킷런(scikit-learn)을 활용하여 feature importance를 추출했다. dri_score, donor_age, age_at_hct, comorbidity_score 등의 feature는 세 모델에서 모두 높은 중요도를 기록하였으며, 이는 조혈모세포이식 환자의 생존 분석에서 질병 중증도, 공여자 및 환자의 나이, 기저질환 등의 요소가 중요한 역할을 함을 시사한다.

한편, 조혈모세포이식의 원인 질환 정보인 prim_disease_hct는 KM과 NA 기반의 모델에서는 공통적으로 발견되었으나 Rank 기반 모델에서는 중요도가 상대적으로 낮아 순위권에서 제외되었다. 대신 공여자와 수혜자 간의 사이토메갈로바이러스(cmv) 상태를 나타내는 cmv_status가 주요 feature로 선정되었다. 이러한 차이는 타겟 변환 방식에 따라 특정 변수의 상대적 중요도가 달라질 수 있음을 보여주며, 조혈모세포이식 환자의 예후 분석 시 변수 선택이 모델링 방식에 따라 영향을 받을 가능성을 시사한다.

Parameter	RF_{KM}^{SFD}	RF_{NA}^{SFD}	RF_{Rank}^{SFD}
C-index	0.6168	0.6212	0.6283
Number of Models	500	500	500
Minimum Node Size	4	7	5
Maximum Tree Depth	46	38	37

표 5. SFD RF optimal hyperparameters (grid search)

Model	Metric	Standard	SFD	Δ
RF_{KM}	RMSE	0.1582	0.1546	-0.0036
	C-index	0.6205	0.6168	-0.0037
RF_{NA}	RMSE	0.2447	0.2403	-0.0044
	C-index	0.6205	0.6212	+0.0007
RF_{Rank}	RMSE	1.3061	1.3028	-0.0033
	C-index	0.6383	0.6283	-0.0100

표 6. 일반 Standard RF와 SFD RF 지표 비교

모델과 타깃별로 39개의 feature 중 importance 상위 10개 feature만을 담은 SFD (Selected Feature Dataset)로 4.1과 동일한 방법으로 hyperparameter grid search를 진행하여 표 6과 같은 결과를 얻었다. 타깃별 RMSE 기준으로는 비슷하거나 향상된 예측 성능을 보여주었으나, $C_{stratified}$기준으로는 다소 감소한 경향을 보였다. 다만, 지표의 변화가 오차 범위 내로 해석될 수 있으며, ¼ 의 feature 만으로 거의 동일한 성능을 낼 수 있다는 점에서 효율이 유의미하게 향상된 것으로 파악된다.

V. 결론

본 연구에서는 조혈모세포이식(HCT) 후 발생하는 사건 발생 시간 데이터(time-to-event data)에 대해, 생존 분석 이론을 활용한 위험 척도(Risk Score)를 정의하고 머신러닝 모델의 예측 성능을 비교하였다. 이를 위해 Kaplan-Meier, Nelson-Aalen, Rank 기반의 세 가지 타깃 변환 기법을 제안하였으며, Random Forest 모델을 활용하여 위험 척도를 예측하고, 그 결과를 RMSE와 Race-Stratified C-Index($C_{stratified}$)를 통해 평가하였다.

실험 결과, Kaplan-Meier 기반 모델(RF_{KM})은 RMSE가 가장 낮아(0.1582) 실제 위험 척도와의 정량적 오차가 가장 작았다. 반면, Rank 기반 모델(RF_{RANK})은 RMSE는 상대적으로 컸으나($C_{stratified}$=0.6383) 샘플 간 상대적 순위를 보다 효과적으로 보존하여 전반적인 분류 성능 지표(공정성 측면의 $C_{stratified}$)에서 우수한 결과를 보였다. 이는 위험 척도를 예측하는 목적에 따라 모델 및 타깃 변환 방식의 선택이 달라질 수 있음을 시사한다.

이후 사이킷런(scikit-learn)을 통해 세 가지 타겟 변환 기법에 따른 feature 중요도를 추출하였고, 그 결과 dri_score, donor_age, age_at_hct, comorbidity_score 등의 변수는 모든 모델에서 일관되게 높은 중요도를 보였다. 반면 prim_disease_hct와 cmv_status의 중요도는 타겟 변환 방식에 따라 차이를 나타냈다. 또한, feature selection을 통해 상위 10개의 주요 변수만을 포함하는 SFD를 구성하여 모델을 학습시킨 결과, 전체 39개 변수로 학습한 모델과 비교하여 유사한 예측 성능을 유지하면서도 보다 효율적인 모델을 구축할 수 있음을 확인하였다. 특히, RMSE 기준에서는 기존 모델과 비슷하거나 향상된 성능을 보였으며, C-index 기준에서는 다소 감소하는 경향이 있었으나, 해당 변화는 오차 범위 내에서 해석 가능하였다. 이러한 결과는 생존 분석에서 모델의 해석 가능성과 예측 성능을 균형 있게 유지하면서도, 불필요한 변수를 제거하여 계산 효율성을 향상시킬 수 있는 가능성을 제시한다.

또한, 의료 데이터는 검열(censored) 특성이 강하므로, 생존 분석 기법을 적극적으로 활용해야 하나, 적절한 target transformation을 설계하는 것이 좋은 효과를 보일 수 있음이 확인되었다. 본 연구에서 제안한 세 가지 타깃 변환 방법은 각각 장단점이 존재하기 때문에, 추후 실제 임상 데이터를 활용할 때에는 데이터 특성, 관심 지표, 공정성 이슈 등을 종합적으로 고려하여 최적의 접근법을 선택할 필요가 있다.

향후 연구에서, 다양한 모델과 target을 설정하여 앙상블(ensemble)모델을 설계하여 성능을 끌어올리는 방법을 실험해보고자 한다. 또한, TabTransformer, TabM등 정형화 데이터를 처리하기에 최적화된 모델을 XGBoost 등전통적인 알고리즘과 성능을 비교함으로써, 모델 간 시너지 효과와 일반화 능력을 함께 평가할 계획이다. 이러한 후속 연구를 통해 조혈모세포이식 환자의 예후 예측에 대한 한층 정교하고 실효성 높은 해법을 제시할 수 있을 것으로 기대한다.

ACKNOWLEDGMENT

The following are results of a study on the “Leaders in INdustry-university Cooperation 3.0” Project, supported by the Ministry of Education and National Research Foundation of Korea.

참 고 문 헌

[1] Seoul National University Hospital. (n.d.). Hematopoietic stem cell transplantation indications and applications. Retrieved February 8, 2024, from http://www.snuh.org/m/board/B019/view.do?bbs_no=6835

[2] National Cancer Center Korea. (n.d.). Hematopoietic stem cell transplantation. National Cancer Center Korea. Retrieved February 8, 2024, from https://www.cancer.go.kr/lay1/S1T295C296/contents.do

[3] CIBMTR. (n.d.). Equity in post-HCT survival predictions [Data set]. Kaggle. Retrieved from https://www.kaggle.com/competitions/equity-post-HCT-survival-predictions

[4] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (n.d.). Random survival forests.

[5] Deotte, C. (2021). NN MLP Baseline CV 0.670 LB 0.676. Kaggle. Retrieved February 11, 2025, from https://www.kaggle.com/code/cdeotte/nn-mlp-baseline-cv-670-lb-676

[6] Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982 May 14;247(18):2543-6.

VI. Appendix

Index	Variable	Description	Type
Disease Characteristics			
1	dri_score	Refined disease risk index	Categorical
2	cyto_score	Cytogenetic score	Categorical
3	prim_disease_hct	Primary disease for HCT	Categorical
Patient Demographics & Comorbidities			
4	diabetes	Diabetes	Categorical
5	arrhythmia	Arrhythmia	Categorical
6	vent_hist	History of mechanical ventilation	Categorical
7	renal_issue	Renal, moderate / severe	Categorical
8	pulm_severe	Pulmonary, severe	Categorical
9	ethnicity	Ethnicity	Categorical
10	year_hct	Year of HCT	Numerical
11	obesity	Obesity	Categorical
12	hepatic_severe	Hepatic, moderate / severe	Categorical
13	prior_tumor	Solid tumor, prior	Categorical
14	peptic_ulcer	Peptic ulcer	Categorical
15	age_at_hct	Age at HCT	Numerical
16	rheum_issue	Rheumatologic	Categorical
17	sex_match	Donor/recipient sex match	Categorical
18	race_group	Race	Categorical
19	comorbidity_score	Sorror comorbidity score	Numerical
20	karnofsky_score	KPS at HCT	Numerical
21	hepatic_mild	Hepatic, mild	Categorical
22	cardiac	Cardiac	Categorical
23	pulm_moderate	Pulmonary, moderate	Categorical
Transplant & Donor Characteristics			
24	tbi_status	TBI	Categorical
25	graft_type	Graft type	Categorical
26	rituximab	Rituximab given in conditioning	Categorical
27	prod_type	Product type	Categorical
28	conditioning_intensity	Computed planned conditioning intensity	Categorical
29	in_vivo_tcd	In-vivo T-cell depletion (ATG/alemtuzumab)	Categorical
30	donor_age	Donor age	Numerical
31	gvhd_proph	Planned GVHD prophylaxis	Categorical
32	donor_related	Related vs. unrelated donor	Categorical
33	melphalan_dose	Melphalan dose (mg^2)	Categorical
HLA Matching & Serostatus			
34	hla_match_c_high	Recipient/1st donor allele level (high resolution) matching at HLA-C	Numerical
35	cmv_status	Donor/recipient CMV serostatus	Categorical
36	hla_match_dqb1_high	Recipient/1st donor allele level (high resolution) matching at HLA-DQB1	Numerical
37	hla_match_a_high	Recipient/1st donor allele level (high resolution) matching at HLA-A	Numerical
38	hla_match_b_high	Recipient/1st donor allele level (high resolution) matching at HLA-B	Numerical
39	hla_match_drb1_high	Recipient/1st donor allele level (high resolution) matching at HLA-DRB1	Numerical

표 7. 최종 선택된 39개 Feature dictionary