

심부전 데이터에 대한 생존 분석 비교 연구

정보통계학과전공 권소연, 김하나, 박혜인
회계학전공 김소은

요약문

본 연구는 파키스탄 파이살라바드에 입원한 심부전 환자들의 데이터를 기반으로 생존분석 모형을 적합하고, 위험률 존재 여부에 대한 검정을 진행하였다. Weibull 분포, Kaplan-Meier, Nelson-Aalen을 이용해 초기 70일간 높은 사망률을 보이는 경향을 살펴보았다. 로그 순위 검정을 통해 심부전 생존에 유의미한 차이를 보이는 변수를 확인하였고, 이후 단계적 선택법과 유의수준에 따른 변수선택을 통해 Cox 비례위험모형을 적합하였다. 완전모형과 일치성 검정, 비례위험 가정 검정, 다중공선성 점검을 거쳐 단계적 선택법을 적용한 Cox 모형을 최종 모형으로 선정하였다. 분석 결과, 심박출률, 나이, 혈청크레아티닌, 고혈압, CPK, 혈청 나트륨이 심부전 환자의 사망률에 중요한 위험 요인으로 확인되었다.

1. 서론

심부전(Heart Failure)은 심장이 구조적 또는 기능적 이상으로 인해 혈액을 충분히 펌프질하지 못하여 신체 대사 요구를 충족시키지 못하는 만성적인 임상 증후군이다. 좌심실 충만 이상은 폐에 물이 차 호흡곤란을 유발할 수 있으며, 좌심실 박출 장애는 피로와 쇠약감을 동반할 수 있다. 반면, 우심실 충만 장애는 다리 부종, 복수, 흉수로 인한 호흡곤란 등의 증상을 초래 할 수 있다. 심부전은 좌심실의 박출 기능에 따라 박출률 감소 심부전(HFrEF), 박출률 경도 감소 심부전(HFmrEF), 박출률 보존 심부전(HFpEF)으로 분류된다. 심부전의 원인 질환 및 요인으로는 심근경색(Myocardial Infarction), 심장판막 질환(Valvular Heart Disease), 심근병증(Cardiomyopathy), 부정맥(Arrhythmias) 등이 있다(Junho Hyun 외, 2023).

심부전은 전 세계적으로 주요한 사망 원인 중 하나로, 높은 의료 비용과 환자의 삶의 질 저하를 초래하는 심각한 질환이다. 2023년 기준으로 약 26만 5,000명이 심부전으로 진단받았으며, 심부전에 대한 연구도 활발히 진행되고 있다. 생존분석은 이러한 상황에서 심부전 환자 사망률에 영향을 미치는 다양한 요인을 정량적으로 평가할 수 있는 방법을 제공한다.

생존분석은 생존율을 분석하여 치료 효과, 예후 인자 및 기타 관련 사항에 대하여 추론하는 통계 분석 기법이다. 이때 사건의 발생 여부에 대한 불확실한 자료인 ‘중도절단’ 자료가 포함되는데, 사건이 발생하기 전에 데이터가 중단되었거나, 관찰 기간 동안 사건이 발생하지 않았다는 정보를 제공한다. 생존분석은 환자의 예후, 치료 효과, 재입원 위험 및 생존 기간에 대한 정보를 제공하며, 중요한 통계적 분석 방법으로 자리잡고 있다.

본 연구의 목적은 Kaplan-Meier, Nelson-Aalen, Cox 비례위험모형 기법을 활용하여 심부전 환자의 사망률에 영향을 미치는 주요 위험 요인을 규명하고자 한다. 연구 결과를 통해 심부전 환자의 생존율 예측 및 사망 위험 요인을 도출하고, 이를 바탕으로 환자 맞춤형 관리 방안을 제시함으로써 환자 생존율 향상에 기여하고자 한다.

2. 심부전 데이터

2.1 심부전 데이터 출처 및 변수 설명

본 분석에서 사용할 자료는 2015년 4월부터 12월까지 파키스탄 편자브 주 파이살라바드에 위치한 파이살라바드 심장병원(Faisalabad Institute of Cardiology)과 알라이드 병원(Allied Hospital)에서 수집된 심부전 환자 299명의 데이터이다. 데이터는 연속형 8개, 범주형 5개로 구성되어있으며, 분석을 위해 연속형 변수 2개를 범주형 변수로 변환하였다. 다음 내용은 <표 2.1>과 <표 2.2>에 나타내었다.

표 2.1 변수 설명

형태	변수	설명	단위
연속형	time	치료기간	Days (4-285)
	DEATH_EVENT	추적 관찰 기간 동안 사망 여부	사망: 1 생존: 0
	age	환자의 나이	Years (40-95)
	creatinine_phosphokinase	혈액 내 CPK의 수치	mcg/L (23-7861)
	ejection_fraction	박출률: 심장 수축 시 심장에서 나가는 혈액의 비율	Percentage (14-80)
	platelets	혈액 내 혈소판의 수	kiloplatelets/mL (25100-850000)
	serum_creatinine	혈청크레아티닌: 혈액 내 크레아티닌 수치	mg/dL (0.5-9.4)
	serum_sodium	혈청나트륨: 혈액 내 나트륨 수치	mg/dL (113-148)
범주형	anaemia	빈혈: 적혈구 또는 혜모글로빈의 감소	빈혈: 1 (170) 빈혈X: 0
	diabetes	환자의 당뇨병 유무	당뇨병: 1 (125) 당뇨병X: 0 (174)
	high_blood_pressure	환자의 고혈압 유무	고혈압: 1 (105) 고혈압X: 0 (194)
	sex	성별: 남성 또는 여성	남성: 1 (194) 여성: 0 (105)
	smoking	환자의 흡연 여부	흡연: 1 (96) 흡연X: 0 (203)

표 2.2 범주형 변수로 변환된 연속형 변수

변환된 변수	변환기준	범주	범주(번호)
NA_group	serum_sodium(혈청나트륨)<135	P (저나트륨혈증)	P(83)
	serum_sodium(혈청나트륨)>=135	N (정상)	N(216)
EF_Level	ejection_fraction(박출률)≤30	P (심장기능 매우 저하)	P(93)
	ejection_fraction(박출률)>30	N (정상)	N(206)

Ahmad와 동료들은 환자들로부터 사전 동의를 받은 후, 해당 의료 기록 데이터를 공개하였다. 공개된 논문에 따르면, 심부전 환자의 선정 기준은 좌심실 수축 기능 부전이 있고, 뉴욕심장학회(New York Heart Association, NYHA) 기능분류에 따라 3급 및 4급에 해당하는 경우로 정의된다. 해당 심부전 환자의 데이터에 대하여 질병발생부터 사망까지 걸리는 시간을 분석하는 생존분석을 진행하기 위하여 공변량으로 나이, CPK, 박출률, 혈소판, 혈청크레아티닌, 혈청나트륨, 빈혈, 당뇨병, 고혈압, 성별, 흡연 여부를 살펴보고자 한다.

위험 요인에 관련된 정보는 혈액 검사 보고서에서 얻었으며, 빈혈 유무를 판단하기 위하여 해마토크리트 수치가 36% 미만인 환자는 빈혈이 있다고 판단하였다. 그 외의 질병은 심장 초음파 보고서나 의사의 기록을 통해 진단되었으며, 흡연 상태는 의사의 기록에서 가져왔다.

본 연구에서는 환자의 주요 위험 요인을 분석하기 위해 연속형 변수인 박출률(Ejection Fraction, EF)을 범주형 변수로 변환하여 활용하였다. 박출률은 심장의 수축 기능을 나타내는 지표로 심장이 박동할 때마다 심장에서 박출되는 혈액의 비율을 의미하며, 이는 심장이 혈액을 얼마나 잘 공급하고 있는지를 평가하는 중요한 변수이다(Nowell M Fine, 2022). 환자의 심장 기능 상태를 평가하는 중요한 변수이다. 박출률 30을 기준으로 환자의 심장 기능 상태를 ‘정상’과 ‘심장기능 매우 저하’로 구분하였다.

추가로, 혈청나트륨 수치를 135mEq/L를 기준으로 정상과 저나트륨혈증으로 구분하여 분석하였다. 이러한 접근은 임상에서 쉽게 적용할 수 있는 기준을 제공하기 위함이다. 나트륨은 근육과 신경의 정상적인 기능에 필수적인 미네랄로, MSD 매뉴얼에 따르면 혈액 내 나트륨 수치가 비정상적으로 낮을 경우 심부전과 관련된 문제일 수 있다(James L. Lewis III, 2023).

또한, 경상대학교병원 박용휘·정영훈 교수팀의 연구에 따르면 심근경색증 환자에서 치료 후 혈소판이 과도하게 활성화될 경우 심부전 발생의 중요한 지표인 심실 확장의 위험도가 3.15배 증가하며, 혈소판의 과활성화가 염증 반응의 증가와 동반될 경우 이 위험도는 21.49배까지 증가하는 것으로 나타났다. 이는 심부전 발생에 있어 혈소판의 활성도 억제가 중요한 치료법이 될 수 있음을 시사한다(손의식, 2017).

존스홉킨스 루푸스 센터에 따르면, 근육 조직이 손상되면 CPK(크레아틴 포스포카니제)가 혈액으로 유출되어 CPK 수치가 상승하게 된다. 이는 심장이나 다른 근육에 스트레스나 부상이 있음을 나타내는 지표로 활용될 수 있다.

서울아산병원에서는 혈청크레아티닌이 근육에서 생성되는 노폐물로, 대부분 신장을 통해 배출되기 때문에 신장 기능의 중요한 지표가 된다고 설명한다. 혈액 검사에서 크레아티닌 수치가 높으면 신장에서 노폐물이 잘 걸러지지 않고 있다는 것을 의미한다.

이러한 배경 지식을 바탕으로 환자의 나이, 혈청크레아티닌 수치 등 변수가 심부전에 대한 생존에 미치는 영향을 분석하고자 한다.

3. 심부전 데이터 생존 분석

본 연구에서는 치료기간(time) 변수를 생존시간으로 사용하여 심부전 데이터에 대한 분석을 하고자 한다. 본 연구의 분석에서는 통계 시스템 R을 사용하였으며 패키지로는 survival과 MASS을 사용했다.

3.1 모수적 방법

생존분석 방법에는 모수적 방법(parametric methods)과 비모수적 방법(non-parametric methods)이 있다. 모수적 생존분석 방법은 결과 변수가 정규분포, 이분형 분포, 또는 포아송 분포와 같이 기존에 알고 있는 분포를 따른다는 가정하에 회귀분석을 통하여 생존함수를 구축하는 방법이다. Weibull, exponential 등이 흔하게 사용된다(J. In와 D. K. Lee, 2018).

3.1.1 Weibull 분포 가정 생존함수

모수적 생존 함수 추정법으로 Weibull 분포를 가정하고 생존 함수를 추정하고자 한다. 와이블 분포의 평균과 분산은 복잡한 형태를 띠고 있지만 생존함수나 위험함수는 단순한 형태이며 모수의 값에 따라 여러 가지 형태의 함수를 표현할 수 있다. 와이블 분포는 지수분포의 일반화된 형태로 현재 가장 많이 쓰이고 있는 생존분포모형이고, 이 분포의 확률밀도함수와 생존 함수, 그리고 위험함수는 다음과 같이 주어진다.

$$\text{Weibull 확률밀도함수: } f(t) = \alpha \lambda t^{\alpha-1} \exp(-\lambda t^\alpha) \quad \alpha > 0, \lambda > 0, t > 0$$

$$\text{Weibull 생존 함수: } S(t) = \exp(-\lambda t^\alpha)$$

$$\text{Weibull 위험률 함수: } h(t) = \alpha \lambda t^{\alpha-1}$$

$$\text{평균: } E(t) = \frac{\Gamma(1 + \frac{1}{\alpha})}{\lambda^{\frac{1}{\alpha}}} \quad \text{분산: } V(t) = \frac{\Gamma(1 + \frac{2}{\alpha}) - [\Gamma(1 + \frac{1}{\alpha})]^2}{\lambda^{\frac{2}{\alpha}}}$$

생존시간이 Weibull 분포를 따른다고 가정하고 모수 α 와 λ 를 추정하고자 한다. Weibull 분포에 대한 모수 추정 결과 <표 3.1>을 보면 $\widehat{\lambda} = \exp(0.1824) = 1.2$ 이고, $\widehat{\alpha} = 1.623731$ 이다. 그러므로 생존시간의 확률밀도함수는

$$f(t) = 1.94863t^{0.623731} \exp(-1.2t^{1.623731})$$

으로 추정된다. <그림 3.1>을 보면 그래프가 직선과 유사한 형태로 Weibull 분포를 잘 따르는 것처럼 보이지만, <그림 3.2>에서 생존함수를 그린 결과 3일 경과 후 생존 확률이 0으로 나타난다. 이는 Weibull 분포가 중도절단을 고려하지 않기 때문에 실제 사건이 발생하지 않은 시간에 대한 예측이 왜곡된 것으로 보인다. Weibull 분포는 데이터에 적합하지 않다고 판단하여, 비모수적 방법을 사용하여 분석을 진행하였다.

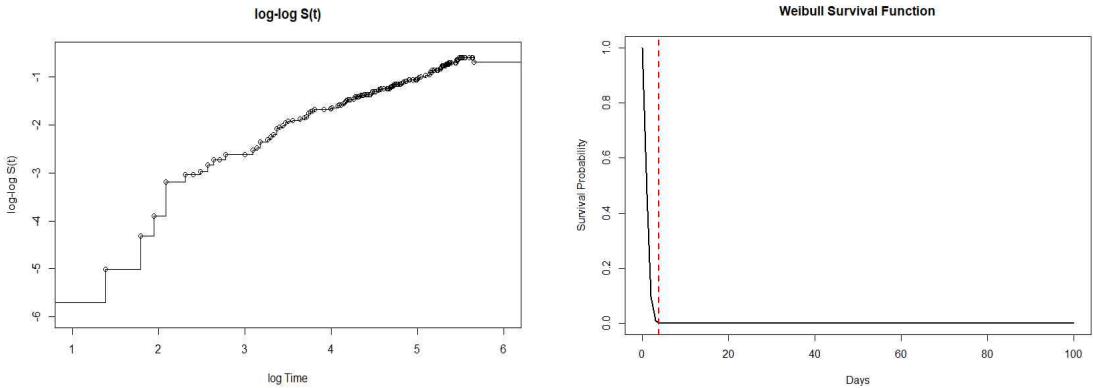


그림 3.1 심부전 데이터 $\log(-\log \hat{S}(t))$

그림 3.2 Weibull 분포 생존함수

3.2 비모수적 방법에 의한 생존분석

비모수적 생존 함수 추정법은 생존시간의 분포에 대한 가정 없이 생존시간과 분산 및 신뢰구간을 추정하여 검정하는 방법이다. 대표적인 방법인 Kaplan-Meier 추정과 Nelson-Aalen 추정을 고려하였다.

3.2.1 Kaplan-Meier Product-Limit 생존함수와 위험률 함수

대표적인 비모수적 방법으로 Kaplan-Meier 추정량을 들 수 있다. Kaplan-Meier 생존 분석 방법은 사건이 발생한 시점마다 구간생존율을 구하여 이들의 누적생존율을 추정하는 방식이며 비모수적 최대우도(nonparametric maximum likelihood)를 활용한 방법으로 누적한계추정량(product-limit estimator)이라고 불린다. Kaplan-Meier Method에서 생존함수와 위험률 함수를 다음과 같이 추정한다.

Kaplan-Meier Product-Limit estimator는

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1, \\ \prod_{t_i \leq t} \left[1 - \frac{d_i}{Y_i} \right], & \text{if } t_1 \leq t. \end{cases}$$

이다. 여기서 d_i 는 t_i 에 사망한 사람의 수를 나타내고, Y_i 는 t_i 바로 직전까지 중도 절단되지 않고 생존해있던 사람의 수를 나타낸다.

Kaplan-Meier estimator의 분산은

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$$

이고 Kaplan-Meier Product-Limit estimator를 이용한 누적 위험률 함수는

$$\hat{H}(t) = -\ln[\hat{S}(t)]$$

으로 추정된다.

<그림 3.3>을 통해 시간이 지남에 따라 생존율은 감소하고, 위험률은 증가하는 경향을 확인할 수 있다. <그림 3.4>에서 70일 간격으로 생존함수를 그린 결과, 생존 시간이 길어짐에 따라 그래프 기울기가 완만해졌으며, 심부전 발생 후 70일 안에 사망 위험이 급격히 증가하는 패턴을 보였다.

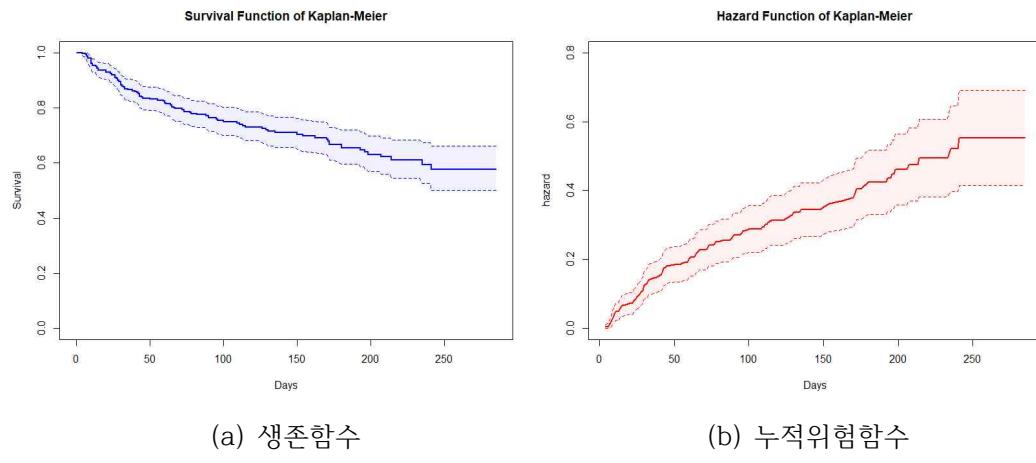


그림 3.3 Kaplan-Meier 방법에 의한 심부전 생존함수와 누적위험함수 그래프

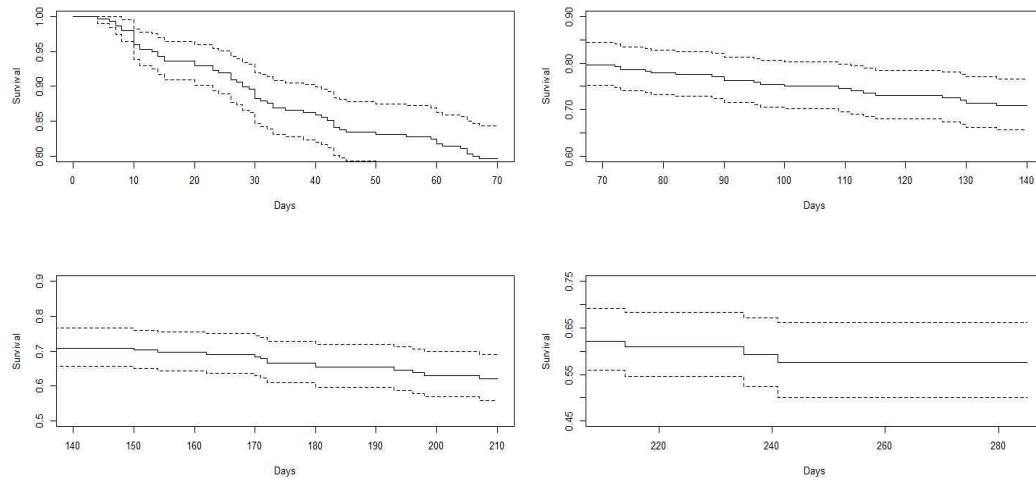


그림 3.4 70일 간격으로 나눈 Kaplan-Meier 생존함수

3.2.2 Nelson-Aalen 누적 위험률함수와 생존함수

Nelson-Aalen 방법은 위험률함수를 추정하는 일반적인 방법이며, 이를 기반으로 생존함수를 추정하는 특수한 방법인 Fleming-Harrington 방법을 이용하고자 한다.

Fleming-Harrington 방법에 의한 추정된 생존함수 $\tilde{S}(t) = \exp[-\tilde{H}(t)]$ 이므로 Nelson-Aalen의 위험률함수는 $\tilde{H}(t) = -\log(\tilde{S}(t))$ 로 구하면 된다. Nelson-Aalen 방법에서 위험률함수를 추정하는 방법과 분산의 추정량은 다음과 같다.

Nelson-Aalen estimator of the cumulative hazards:

$$\tilde{H}(t) = \begin{cases} 0, & \text{if } t \leq t_1, \\ \sum_{t_i \leq t} \frac{d_i}{Y_i}, & \text{if } t_1 \leq t. \end{cases}$$

Nelson-Aalen estimator 의 분산:

$$\sigma_H^2(t) = \sum_{t_i \leq t} \frac{d_i}{Y_i^2}.$$

Nelson-Aalen estimator of the cumulative hazard rate 을 이용한 생존함수:

$$\tilde{S}(t) = \exp\{-\tilde{H}(t)\}.$$

Nelson-Aalen 방법은 위험률을 추정하는 함수로, 생존함수가 시간이 지남에 따라 감소한다면 위험함수는 시간이 지남에 따라 증가한다고 볼 수 있다. <그림 3.5>를 통해 시간이 흐름에 따라 위험률이 증가하고, 생존률이 감소하는 것을 확인할 수 있다.

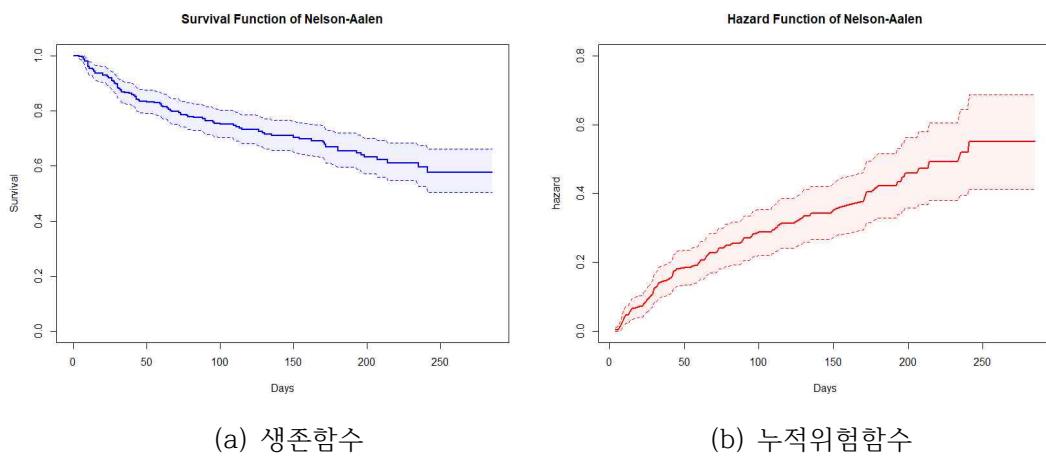


그림 3.5 Nelson-Aalen 방법에 의한 심부전 생존함수와 누적위험함수 그래프

3.3 로그순위검정

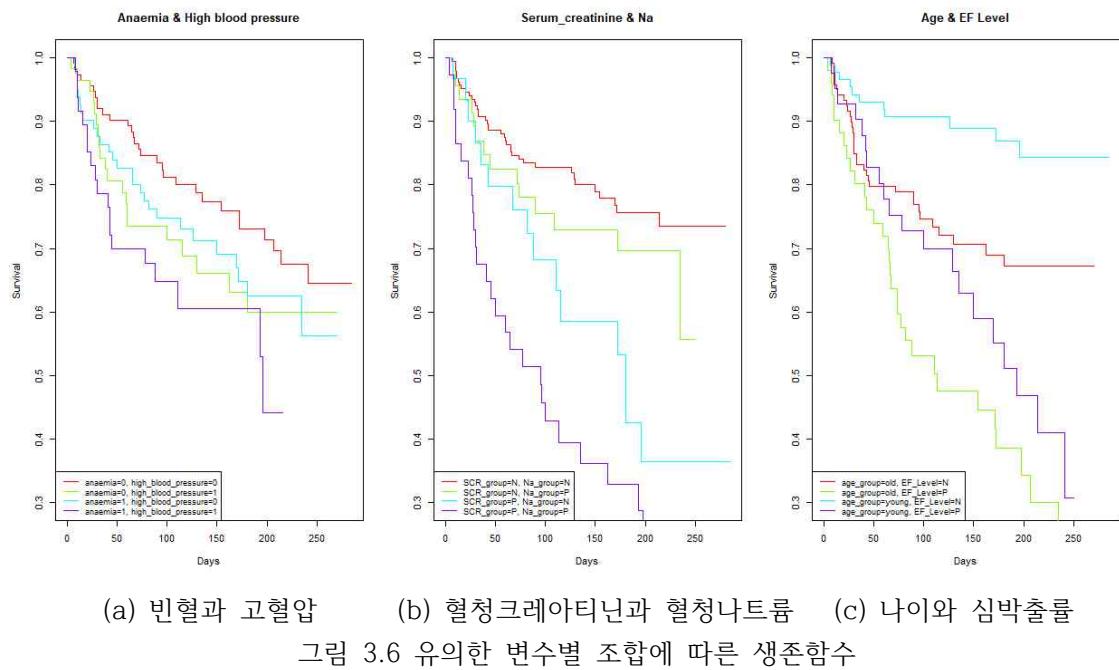
로그순위검정을 위해, 기존의 연속형 변수를 범주형으로 변환하였다. 범주형으로 변환된 변수는 <표 3.1>에 나타나 있으며, 해당 변수들은 로그순위검정에만 사용되었다. 검정 결과, 빈혈, 고혈압, 나이, 심박출률, 혈청크레아티닌, 혈청나트륨 변수가 심부전과 유의한 상관관계를 보였다. <표 3.2>에서 자세한 내용을 확인할 수 있다. 빈혈과 고혈압, 나이와 심박출률, 혈청크레아티닌과 혈청나트륨의 조합에 따른 Kaplan-Meier 생존함수를 그린 결과, 해당 변수들의 조합이 심부전 환자의 생존에 유의한 영향을 미치는 것으로 나타났다. <그림 3.6>에서 변수별 조합에 따른 생존 곡선을 확인할 수 있다.

표 3.1 변수 변환

변수	변환기준	범주
나이	age < 60	young
	age ≥ 60	old
혈청 크레아티닌	serum_creatinine > 1.5	P(신장 기능 장애)
	serum_creatinine ≤ 1.5	N (정상)
CPK	creatinine_phosphokinase < 250	Normal
	250 ≤ creatinine_phosphokinase < 1000	Intermediate
	creatinine_phosphokinase ≥ 1000	Severe
혈소판	platelets ≤ Q1	1 (적음)
	Q1 < platelets < Q3	2 (적당)
	platelets ≥ Q3	3 (많음)

표 3.2 로그순위검정

변수	검정통계량	p-value
혈청크레아티닌	39.612	0.000
심박출률	30.626	0.000
혈청나트륨	15.650	0.000
나이	8.565	0.003
고혈압	4.406	0.036
빈혈	2.727	0.099
혈소판	3.453	0.178
CPK	0.885	0.643
당뇨병	0.041	0.841
성별	0.004	0.950
흡연	0.002	0.964



3.4 Cox 비례 위험률모형

Cox(1972) 비례 위험률모형(Cox Proportional hazard regression model)은 특정 변수들에 대한 상대위험률(relative hazard)을 계산하는 생존 분석법이다. 생존시간에 대한 어떠한 분포도 가정하지 않으므로 비모수적이지만 모형에 근거하여 회귀계수를 추정한다는 점이 모수적 방법과 유사하여 비모수와 모수의 중간형태 모형이라 알려져 있다. Therneau and Grambsch(2000)은 Cox모형을 이용한 생존분석에 대한 설명을 제공한다.

이 절에서는 유방암 관련 공변량을 고려한 Cox 비례 위험률모형을 적합하고 유방암 생존율에 유의한 변수와 적합한 모형을 찾고자 한다.

$h(t | Z)$ 는 t시점에서 공변량 Z 를 갖는 위험함수로 정의하고, $h_0(t)$ 는 모든 공변량 값이 0인 경우의 기저 위험률 함수(baseline hazard function)라고 하며 비례위험 회귀모형은 다음과 같다:

$$h(t | Z) = h_0(t) \exp[\beta Z].$$

위험 함수 중 log 선형 관계를 가지는 모형이고, $h_0(t)$ 는 기저 위험률 함수로 식에 시간이 포함되어 있으므로 시간에 의존한다고 하며 $\exp[\beta Z]$ 는 일정하다고 할 수 있다.

심부전 생존에 영향을 미치는 공변량을 고려하여 Cox 비례위험 모형을 적합하고자 한다. 본 연구에서는 나이, 심박출률, 혈청크레아티닌 등 총 11개의 변수들이 심부전 환자의 생존에 미치는 영향을 분석하였다. <표 3.3>은 전체 변수를 포함한 완전모형(full model)을 적합한 결과이다. 분석 결과, 나이, 심박출률, 혈청크레아티닌, CPK, 고혈압이 유의수준 5%에서 유의한 결과를 보였으며, 나머지 변수들은 유의하지 않았다. 이후, 유의한 변수들만을 포함한 모형과 단계적 선택법을 적용하여 AIC가 가장 작은 모형을 최종적으로 적합하였다.

완전모형에서 유의했던 나이, 심박출률, 혈청크레아티닌, CPK, 고혈압 변수 5개를 공변량으로 하여 비례위험모형을 적합한 결과는 <표 3.4>에 확인할 수 있다. 단계적 선택법을 적용하여 나이, 심박출률, 혈청크레아티닌, CPK, 고혈압, 혈청나트륨 변수를 공변량으로 하여 비례위험모형을 적합한 결과는 <표 3.5>와 같다.

표 3.3 Full 비례 위험률모형

변수	회귀계수	exp(회귀계수)	se(회귀계수)	z	p-값	신뢰구간 하한	신뢰구간 상한
심박출률_P	1.15	3.16	0.2173	5.294	0.00000012	2.0637	4.838
나이	0.04934	1.051	0.009746	5.063	0.000000414	1.0307	1.071
혈청크레아티닌	0.2679	1.307	0.06765	3.959	0.0000752	1.1448	1.492
CPK	0.000242	1	0.00009728	2.488	0.0129	1.0001	1
고혈압	0.5068	1.66	0.2163	2.343	0.0191	1.0863	2.537
혈청나트륨_P	0.4273	1.533	0.228	1.875	0.0609	0.9807	2.397
빈혈	0.3118	1.366	0.2177	1.432	0.152	0.8915	2.093
당뇨병	0.2295	1.258	0.2209	1.039	0.299	0.8158	1.94
성별	-0.1935	0.8241	0.2515	-0.769	0.4417	0.5034	1.349
흡연	0.1786	1.196	0.2567	0.696	0.4865	0.7229	1.977
혈소판	-6.26×10^{-7}	1	0.000001117	-0.56	0.5753	1	1

표 3.4 나이, 심박출률, 혈청크레아티닌, CPK, 고혈압을 공변량으로 하는 축소 비례위험모형

변수	회귀계수	exp(회귀계수)	se(회귀계수)	z	p-값	신뢰구간 하한	신뢰구간 상한
심박출률_P	1.216	3.373	0.2092	5.812	6.18×10^{-9}	2.239	5.083
나이	0.05019	1.051	0.009038	5.553	0.000000028	1.033	1.07
혈청크레아티닌	0.2928	1.34	0.06357	4.605	0.00000412	1.183	1.518
고혈압	0.4997	1.648	0.2127	2.349	1.88×10^{-2}	1.086	2.501
CPK	0.000194	1	0.00009548	2.03	0.0423	1	1

표 3.5 나이, 심박출률, 혈청크레아티닌, CPK, 고혈압, 혈청나트륨을 공변량으로 하는 단계적 선택 축소 비례 위험 모형

변수	회귀계수	exp(회귀계수)	se(회귀계수)	z	p-값	신뢰구간 하한	신뢰구간 상한
심박출률_P	1.151	3.16	0.213	5.41	6.30×10^{-8}	2.0838	4.799
나이	0.04677	1.048	0.00914	5.117	3.10×10^{-7}	1.0293	1.067
혈청크레아티닌	0.2724	1.313	0.06661	4.09	0.0000432	1.1524	1.496
고혈압	0.5412	1.718	0.214	2.527	0.0115	1.1291	2.614
CPK	0.0002035	1	0.00009495	2.143	0.0321	1	1
혈청나트륨_P	0.3686	1.446	0.2201	1.675	0.094	0.9391	2.226

<그림 3.7>은 완전모형과 유의한 변수들만 선택한 축소모형, 그리고 단계적 선택법을 적용한 축소모형의 생존함수 그래프이다. 단계적 선택법을 적용한 모형의 생존 그래프가 완전모형과 유사한 형태를 보였다. 6개의 공변량으로도 예측이 잘 된다고 판단하여, 축소모형인 단계적 선택법 비례위험모형을 더 적절한 모형이라고 평가하였다. <표 3.5> 결과를 바탕으로 단계적 선택법 비례위험모형에 대한 추정식은

$$h(t) = h_0(t) \exp [1.151 \times \text{심박출률} P + 0.04677 \times \text{나이} + 0.2724 \times \text{혈청크레아티닌} + 0.5412 \times \text{고혈압} + 0.0002035 \times \text{CPK} + 0.3686 \times \text{혈청나트륨} P]$$

과 같다. 심박출률이 비정상 상태일수록, 나이가 많을수록, 혈청크레아티닌과 CPK 수치가 높을수록, 고혈압이 있을수록, 혈청나트륨 수치가 좋지 않을수록 위험률이 높아진다.

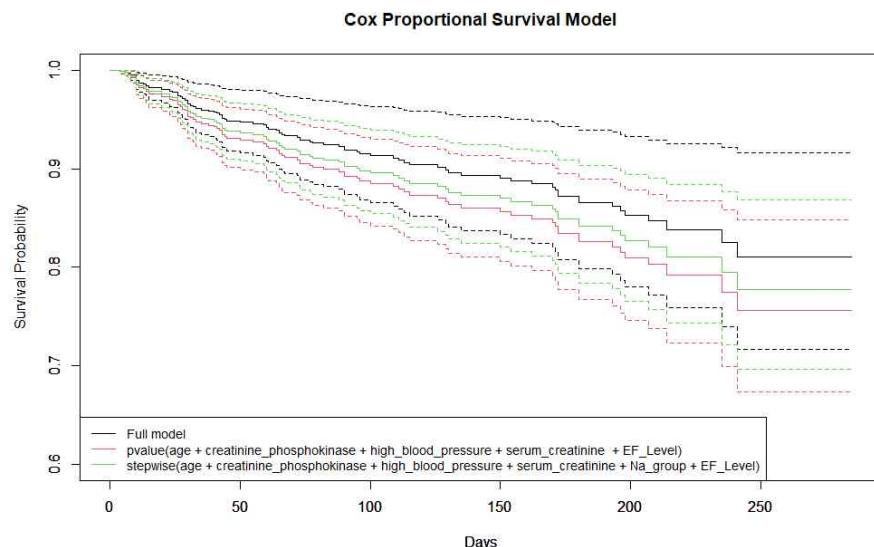


그림 3.7 완전모형과 축소모형에서 비례 위험률모형 생존함수 그래프

EF(ejection_fraction)은 심장에서 박출되는 혈액의 비율로서, 심장이 혈액을 잘 공급하고 있는지 평가하는 중요한 척도가 된다. 심박출률이 정상범위와 비정상적으로 저하된 경우에 따라 생존율에 대한 차이를 알아보고자 나이, 혈청크레아티닌, CPK, 고혈압, 혈청나트륨을 고려한 Cox 비례위험률 모형을 각각 구하였다.

<표 3.6>은 EF가 비정상적으로 저하된 경우, <표 3.7>은 EF가 정상인 경우의 Cox 비례위험률 추정 결과를 보여준다. <그림 3.8>을 보면 심박출률이 비정상적으로 저하된 경우 생존률이 현저하게 낮음을 알 수 있다.

표 3.6 EF가 비정상적으로 저하(P)인 경우 비례 위험률모형

변수	회귀계수	exp(회귀계수)	se(회귀계수)	z	p-값	신뢰구간 하한	신뢰구간 상한
나이	0.044944	1.04597	0.014528	3.094	0.00198	1.0166	1.076
혈청크레아티닌	0.028314	1.028719	0.1749	0.162	0.87139	0.7302	1.449
고혈압	0.637456	1.891662	0.303543	2.1	0.03572	1.0434	3.429
CPK	0.000137	1.000137	0.000155	0.885	0.37633	0.9998	1
혈청나트륨_P	0.146512	1.157789	0.316425	0.463	0.64335	0.6227	2.153

표 3.7 EF가 정상(N)인 경우 비례 위험률모형

변수	회귀계수	exp(회귀계수)	se(회귀계수)	z	p-값	신뢰구간 하한	신뢰구간 상한
나이	0.055597	1.057171	0.013293	4.183	2.88×10^{-5}	1.03	1.085
혈청크레아티닌	0.350923	1.420378	0.069004	5.086	3.67×10^{-7}	1.2407	1.626
고혈압	0.368217	1.445156	0.318431	1.156	0.248	0.7742	2.698
CPK	0.000153	1.000153	0.000132	1.166	0.244	0.9999	1
혈청나트륨_P	0.694001	2.001707	0.31589	2.197	0.028	1.0777	3.718

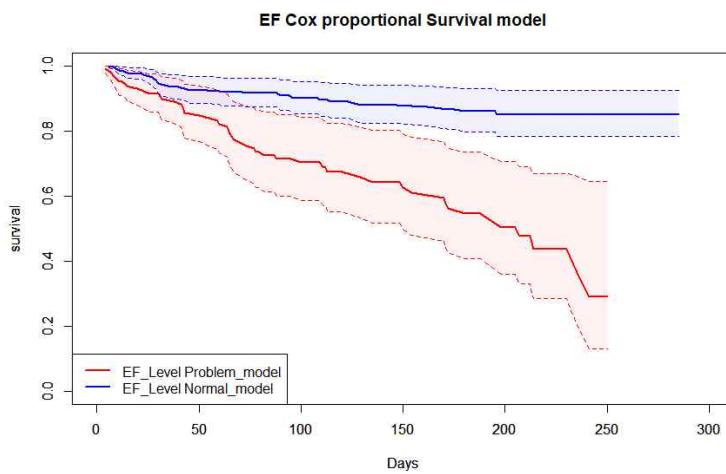


그림 3.8 심박출률에 따른 비례 위험률함수 생존함수 그래프

3.5 생존함수와 위험률 모형 비교

Kaplan-Meier 모형, Nelson-Aalen 모형과 Cox 모형에 의한 생존함수를 비교해보고자 한다. <그림 3.9> 그래프를 보면 공변량을 포함한 비례위험률 모형에서 생존율이 높게 나타나며 위험률은 작게 나타난다. 생존 시간에 대해 공변량으로 설명되는 부분 때문에 그렇지 않은 경우에 비해 생존율이 높아진 것으로 여겨진다.

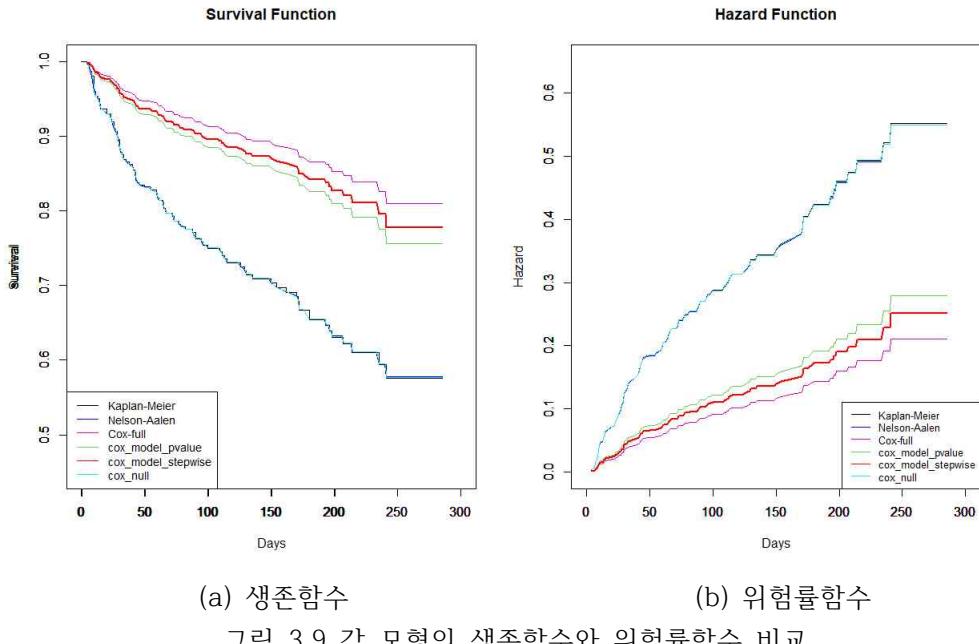


그림 3.9 각 모형의 생존함수와 위험률함수 비교

3.6 모형 평가

3.6.1 모형 적합도 검정

<표 3.8>은 모형 적합도 검정 결과이다. Likelihood test 결과 공변량을 포함한 모델이 공변량이 없는 기본 모델에 비해 유의미한 영향을 미친다. Wald test와 Score test 결과, 심부전 생존에 유의미하게 영향을 미치는 변수가 있다는 것을 통해, 단계적 선택 Cox 모형이 공변량이 없는 Cox 모형보다 더 우수한 성능을 지닌 것으로 판단된다.

표 3.8 단계적 선택 축소모형에 대한 적합도 검정

통계량	통계량 값	자유도	p-값
Likelihood ratio test	81.34	6	2.00×10^{-15}
Wald test	82.19	6	1.00×10^{-15}
Score (logrank) test	89.50	6	$<2 \times 10^{-16}$

3.6.2 모형 비례성 검정

범주형 변수에 대한 로그누적기저위험함수 결과, 그래프의 선이 교차하지 않으면 비례성 가정이 만족된다. <그림 3.10>을 보면, 범주형 변수들의 로그누적위험함수 그래프에서 선이 교차하는 부분이 거의 보이지 않으므로, 비례성 가정을 대체로 만족한다고 판단하였다.

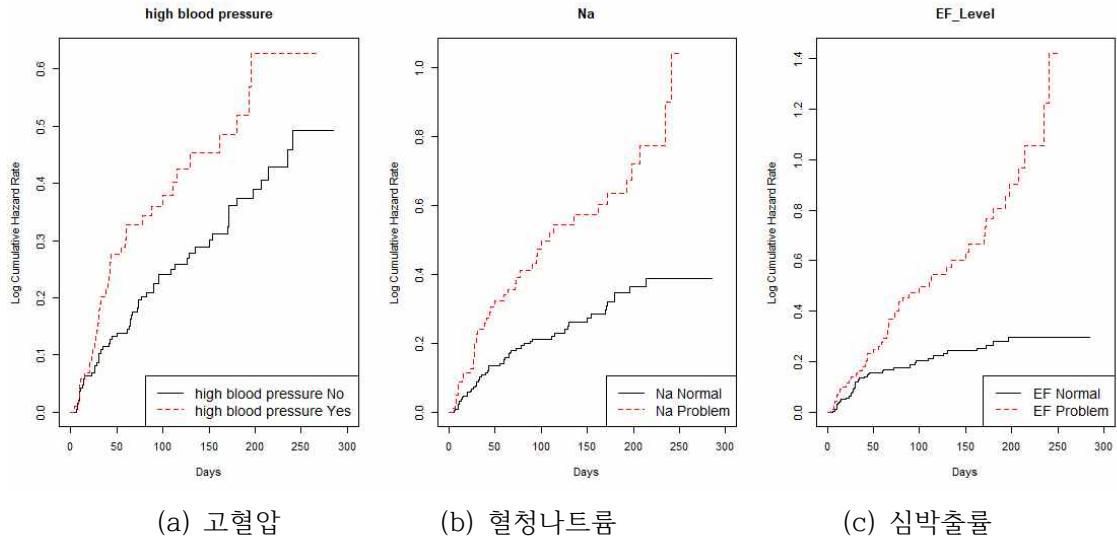


그림 3.10 범주형 변수에 대한 로그누적기저위험함수

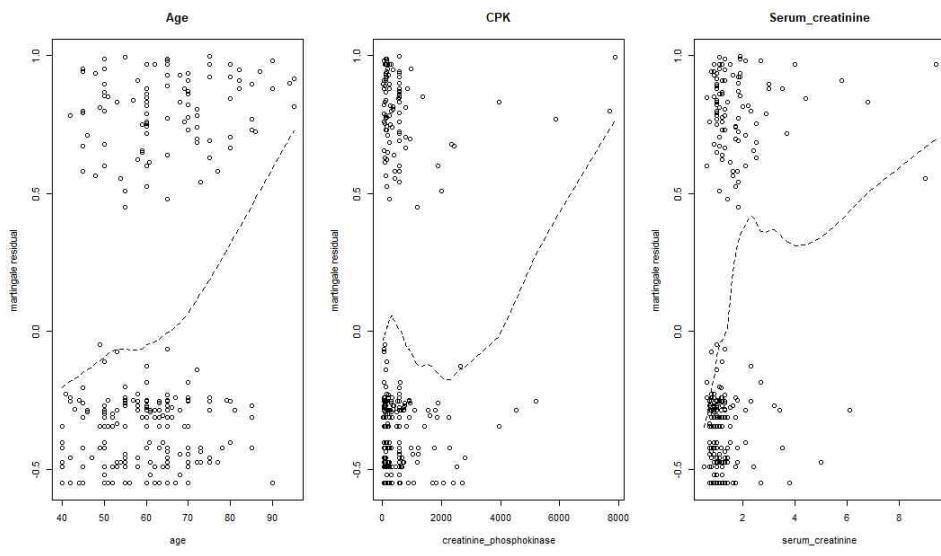
<표 3.9>는 Schoenfeld 잔차를 사용하여, 모형에 대한 비례위험 가정에 대한 검정을 진행한 결과이다. 유의수준 5%에서 심박출률은 비례성 가정을 만족하지 않았으나, 전체 모형과 다른 변수들은 비례성 가정을 모두 만족하였다. 따라서, 모형은 비례위험 가정을 만족하는 편이라고 판단하였다.

표 3.9 Schoenfeld 잔차 비례 위험 가정

변수	나이	고혈압	혈청크레아티닌	혈청나트륨	CPK효소	심박출률	전체모형
검정통계량	0.0423	0.1218	0.2475	0.2513	1.4827	7.8771	10.4974
p-값	0.837	0.727	0.619	0.616	0.223	0.005	0.105

3.6.3 잔차분석

연속형 변수에 대한 잔차 분석을 진행하였다. 마팅게일 잔차 분석에서 공변량이 없을 때 나아이는 선형 관계를 보였고, CPK와 혈청크레아티닌은 비선형 관계를 보였으나 최종 모형에서는 선형성이 사라진 형태로 나타났다. CPK와 혈청크레아티닌이 비선형적인 관계를 띠고 있으며, 최종 모델에서 연속형 변수들이 생존 시간과의 관계를 더 적합하게 반영했음을 알 수 있다. Schoenfeld 잔차 결과, 모든 변수가 특정 패턴을 뚜렷하게 보이지 않아 비례성 가정도 만족하는 것으로 보인다. 공변량이 없을 때 마팅게일 잔차는 <그림 3.11>에, 최종 모형의 잔차는 그림 <그림 3.12>, <그림 3.13>, <그림 3.14>에 나타나 있다.



(a) 나이

(b) CPK

(c) 혈청크레아티닌

그림 3.11 공변량 없을 경우 잔차분석

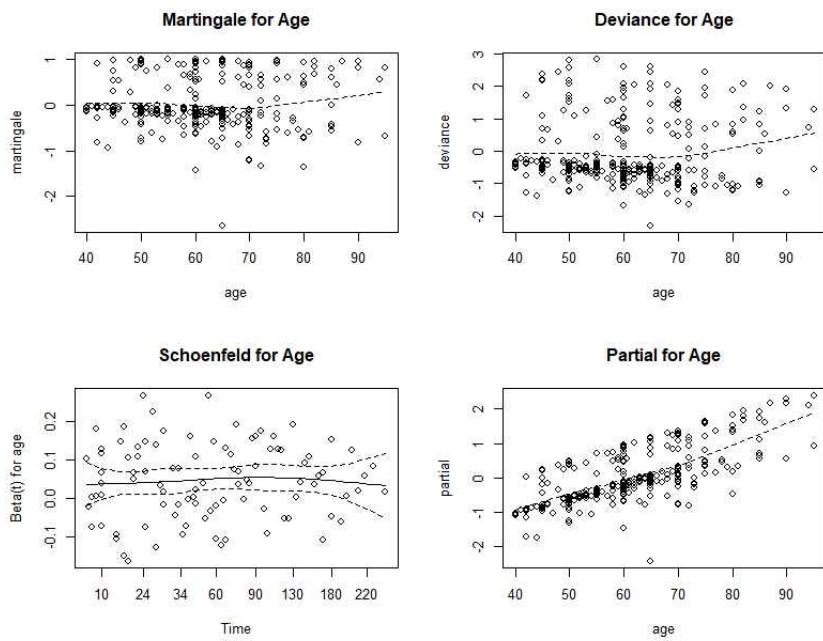


그림 3.12 최종 모형의 나이 잔차분석

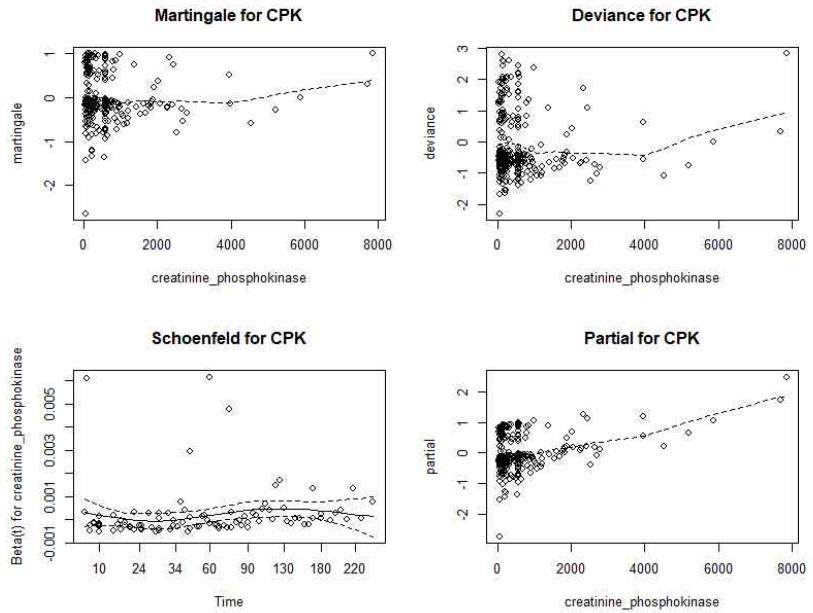


그림 3.13 최종 모형의 CPK 잔차분석

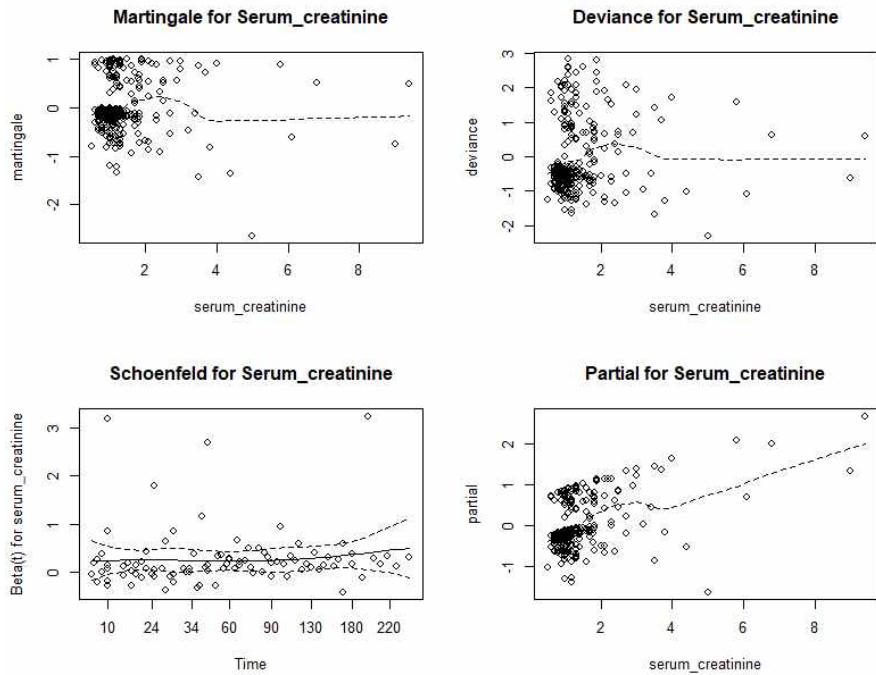


그림 3.14 최종 모형의 혈청크레이티닌 잔차분석

3.6.4 다중공선성

다중공선성 검토 결과를 바탕으로 최종 모델을 선택하였다. 카이제곱 검정, T검정, 피어슨 상관분석을 통해 변수 간의 다중공선성 문제를 분석한 결과, 혈청나트륨과 나이, 혈청나트륨과 혈청크레아티닌 사이에 다중공선성이 의심되었으나, 다른 변수들 간의 관계에서는 다중공선성이 영향을 미치지 않았다. 일부 변수들 간에만 다중공선성이 존재했으므로, 이들이 최종 모델에 미치는 영향은 제한적이라고 판단하였다. 검정 결과는 <표 3.10>, <표 3.11>, <그림 3.15>을 통해 확인할 수 있다.

표 3.10 범주형 변수 간 카이제곱 검정

변수		검정통계량	p-값
심박출률	혈청나트륨	2.5144	0.1128
심박출률	고혈압	0.048463	0.8258
혈청나트륨	고혈압	0.0091128	0.9239

표 3.11 범주형 변수와 연속형 변수의 T검정

p-값	나이	혈청크레아티닌	CPK 효소
심박출률	0.2177	0.4344	0.9651
혈청나트륨	0.001137	0.0004147	0.1964
고혈압	0.1031	0.9387	0.2023

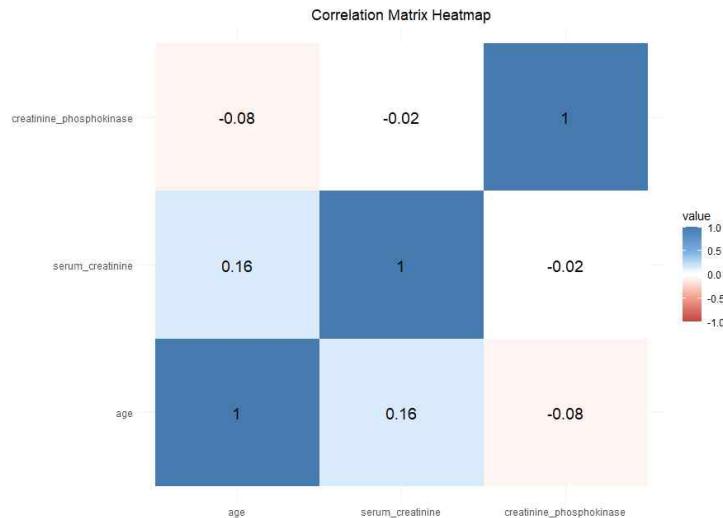


그림 3.15 연속형 변수 간 피어슨 상관분석

해당 모형이 모형 적합도와 비례위험 가정을 잘 만족하는 편이고, 다중공선성이 모델에 미치는 영향이 크지 않다고 판단하였다. 따라서 단계적 선택법을 적용한 Cox 비례위험 모델을 심부전 데이터에 적합한 최종 모형으로 선정하였다.

4. 결론

본 분석은 심부전 발생부터 사망까지의 시간에 영향을 주는 변수를 파악하고자 하였다. Weibull 분포를 가정한 생존 함수 추정에서 중도절단을 반영하지 않아 왜곡이 발생할 수 있음을 확인하고, Kaplan-Meier와 Nelson-Aalen 방법을 통해 생존 함수와 위험 함수의 경향을 분석하였다. 심부전 발생 후 70일 이내에 사망 위험이 급격히 증가하고, 250일 이후에는 약 60% 확률로 안정화되는 패턴이 나타났다. 로그 순위 검정 결과, 혈청크레아티닌, 심박출률, 혈청나트륨, 나이, 고혈압, 빈혈이 생존에 유의한 영향을 미쳤으며, Kaplan-Meier 생존함수를 통해 변수 간의 상호작용이 생존 확률에 영향을 미침을 확인했다.

이후 Cox 비례 위험 모형 분석결과, 유의수준 5%에서 나이, 심박출률, 혈청크레아티닌, CPK, 고혈압의 5개 변수가 확인되었다. 단계적 선택법을 적용한 결과에서는 혈청나트륨을 포함한 6개 변수가 최종 모형에 포함되었다. 모형 비교 결과, 단계적 선택법이 적용된 Cox 모형이 데이터에 더 적합하였으며, Likelihood test, Wald test, Score test 결과에서도 유의미한 변수가 있었다. 로그누적기저위험함수와 잔차분석을 통해 범주형 변수와 연속형 변수 모두 비례성을 만족하는 것으로 나타났고, Schoenfeld 잔차를 사용한 비례 위험 가정 검정 결과 심박출률이 비례성을 충족하지 않았으나 전체 모형은 유의수준 5%에서 비례성을 만족하는 것처럼 보였다. 다중공선성 검토 결과, 혈청나트륨과 나이, 혈청크레아티닌 간의 다중공선성이 의심되었으나, 최종 모델에 미치는 영향은 제한적이었다. 따라서 단계적 선택법을 적용한 Cox 비례위험 모형을 최종 모델로 선정하였다.

분석 결과, 혈청나트륨 수치가 낮을 경우 저나트륨혈증이 발생하고, 심박출률 저하는 심장 기능 저하로 이어져 사망 위험이 증가하는 경향이 있었다. 나이, 고혈압, 빈혈, 혈청크레아티닌, CPK 수치가 증가할수록 사망 확률도 높아진다. 저나트륨혈증은 뇌세포로 수분이 이동하여 뇌부종을 유발하고, 이로 인해 신경학적 증상이 나타나며 심한 경우 정신 이상이나 간질 발작의 위험이 증가한다(서울대학교 의과대학 국민건강지식센터, 2015). 이러한 증상은 심부전 환자의 상태를 악화시킨다. 심장 기능 저하는 혈액 펌프 기능을 저하시켜 사망 위험을 높인다. 나이는 생물학적 노화와 관련이 있으며, 고혈압은 심장에 부담을 주고 혈관을 손상시켜 심부전의 주요 원인이 된다(질병관리청 국가건강정보포털). 빈혈은 혈액의 산소 운반 능력을 감소시켜 여러 장기에 손상을 일으키며, 혈청크레아티닌 수치는 신장 기능을 나타내는 지표로 신장 기능 저하가 심혈관계에 부정적인 영향을 미친다. 마지막으로, CPK 수치는 심장 손상의 지표로, 수치가 높아지면 심장 질환 발생 가능성이 증가한다. 이러한 변수들을 고려할 때, 심부전 환자에 대한 지속적인 관리와 조기 개입이 중요하다. 이를 통해 심부전 발생부터 사망까지의 시간을 지연시키고 생존율을 높이기 위한 노력이 필요하다.

이번 연구는 파키스탄 병원에서 수집된 데이터로, 특정 지역과 의료 환경에 제한이 있으며 표본 크기가 작아 결과의 일반화에 어려움이 있다. 특성들의 차이로 연구 결과를 국내 상황에 직접 적용하는 데 한계가 있다. 더불어, 해당 데이터는 약 9개월간 40세 이상의 심부전 환자를 대상으로 하였으므로, 장기적인 상태 변화나 젊은 연령대 환자의 상태를 반영하지 못하였다. 마지막으로, 본 연구의 Cox 모형은 비례위험 가정에 기반한 분석법이므로 시간이 지남에 따라 각 변수의 효과가 달라질 수 있다. 따라서 통계적 한계와 Cox 모형의 선택에 따른 결함이 존재할 수 있다. 더 큰 규모의 국내 데이터 수집과 장기적 추적 조사를 통해 국내 환자에 최적화된 치료 방안을 고려해야 하며, 기계 학습 모델을 활용한 생존 분석을 통해 심도 있는 연구로 이어져야 한다.

참고문헌

- 서울아산병원 (n.d.) 검사/시술/수술 정보, m.amc.seoul.kr (Accessed: 2024-12-15).
- 서울대학교 의과대학 국민건강지식센터 (2015) hqcenter.snu.ac.kr
- 손의식(2017-03-02). 국내 연구진, 혈소판-심부전 발생 인과관계 세계 최초 규명.
Medical Times, medicaltimes.com
- 이시은, 심병용, 김재희 (2015). 직장암 데이터에 대한 위험률 함수 추정 및 위험률 변화점 추정, 한국데이터정보과학회지, 26(6), 1225-1238.
- 질병관리청 국가건강정보포털, health.kdca.go.kr
- Ahmad, T., Aftab, M., Bhatti, S. H.*, Munir, A., Raza, M. A. (2017), Survival Analysis of Heart Failure Patients: A Case Study, Department of Statistics, Government College University, Faisalabad, Pakistan.
- Chicco, D., Jurman, G. (2020), Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone, BMC Medical Informatics and Decision Making, 20(1), 16. DOI: 10.1186/s12911-020-1023-5.
- In, J., Lee, D. K. (2018), Survival Analysis: Part I – Analysis of Time-to-Event, Korean Journal of Anesthesiology, DOI: 10.4097/kja.d.18.00067.
- Hyun, Junho, et al. (2023). "Korean Society of Heart Failure Guidelines for the Management of Heart Failure: Advanced and Acute Heart Failure." Korean Circulation Journal, 53(7), 452-471. <https://doi.org/10.4070/kcj.2023.0115>
- James L. Lewis III, MD, Brookwood Baptist Health and Saint Vincent's Ascension Health, Birmingham
- Johns Hopkins Lupus Center.(n.d.).Other clinical tests. Retrieved from hopkinslupus.
- Klein, J. P., Moeschberger, M. L. (2003), Survival Analysis Techniques for Censored and Truncated Data, Springer, New York.
- Nowell M Fine, MD, SM, Libin Cardiovascular Institute, Cumming School of Medicine, University of Calgary, 2022)
- Therneau, T. M., Grambsch, P. M. (2000), Modeling Survival Data Extending the Cox Model, Springer, New York.