# R로 진행하는 의료 데이터 분석 - 기초

차라투 주식회사

데이터 분석가 조은서

# 목차

# 자기소개

### 의료 데이터 분석 전문가

차라투에서 2022년 7월부터 근무 중

### 의학 논문 게재

JAMA, CHEST 등 의학 학술지에 공저자 논문 보유

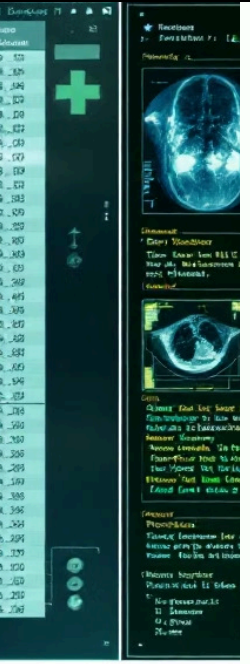
### 자격 및 교육

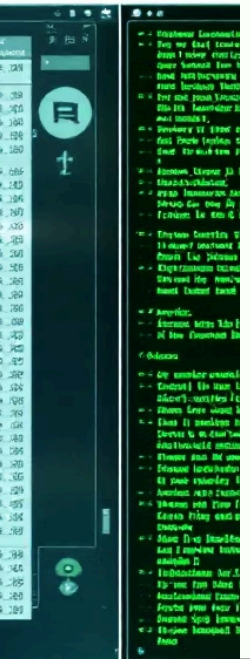
국민대학교 경영학/빅데이터경영통계 전공

ADP, SQLD 등 자격증 보유

# 1부: 의료 데이터 특성과 R 기초 실습

# 의료 데이터 기본 이해와 분류





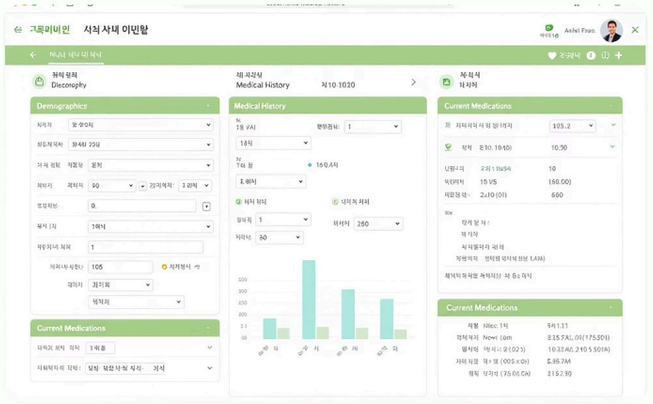
# 1. 의료 데이터란?

| 구분 | 정의 | 예시 |
| --- | --- | --- |
| 정형 데이터 (Structured) | 표 형태로 저장된 수치, 코드 등 | 나이, 혈압, 진단코드, 약물명 등 |
| 비정형 데이터 (Unstructured) | 일정한 구조 없이 저장된 데이터 | 진료 기록 텍스트, CT/MRI 이미지, 음성 녹취 등 |

의료 데이터는 환자의 진료 과정에서 생성되는 모든 정보를 포괄하는 개념으로, 크게 정형 데이터와 비정형 데이터로 구분할 수 있습니다. 정형 데이터는 명확한 구조와 형식을 가진 데이터로, 데이터베이스에 쉽게 저장되고 분석될 수 있습니다.

반면 비정형 데이터는 구조화되지 않은 형태로 존재하며, 텍스트, 이미지, 음성 등 다양한 형태로 나타납니다. 이러한 데이터는 주로 인공지능이나 머신러닝과 같은 고급 분석 기법을 통해 의미 있는 정보를 추출합니다.

# 2. 정형화 데이터

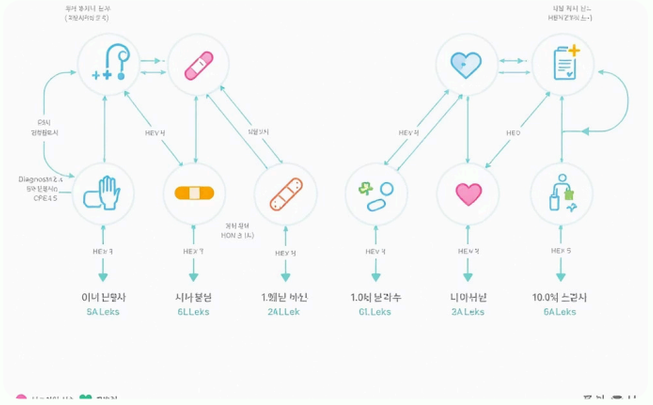


전자의무기록(EMR)

병원에서 일상적으로 수집되는 진료 정보로, 환자의 진단, 검사 결과, 처방 내역 등이 포함됩니다. 병원별로 시스템과 구조에 차이가 있어 데이터 통합에 주의가 필요합니다.

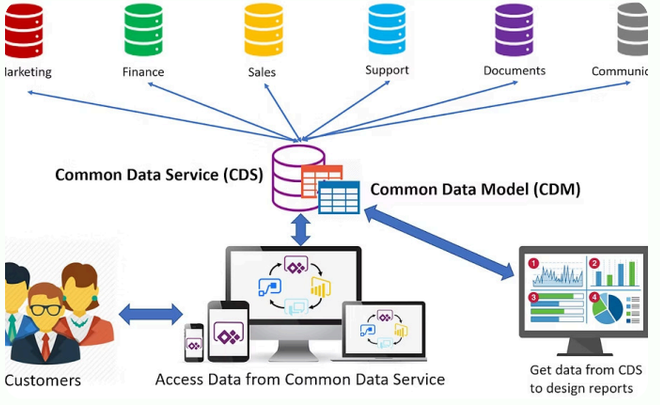

임상시험 데이터

계획된 연구 프로토콜에 따라 수집되는 고품질 데이터입니다. 일반적으로 규모는 작지만 표준화가 잘 되어 있어 분석의 정확도가 높은 편입니다.



청구 데이터

보험 심사를 위해 의료기관에서 제출하는 데이터로, 전국 단위의 정보를 담고 있어 대규모 분석에 유용합니다. 다만 임상적 상세 정보는 제한적입니다.




CDM (Common Data Model)

다양한 진료 데이터를 표준화하여 구조와 용어를 일관성 있게 정리합니다. condition_occurrence, drug_exposure 등의 개념으로 데이터를 정의하여 연구 공유가 용이해집니다. 최근 병원과 연구소를 중심으로 CDM 도입이 증가하고 있습니다.

# 2. 정형화 데이터

**시간성**
반복 측정, 입원기간 등 시간 흐름 중요

**표준 코드 기반**
ICD, ATC, KCD 등 다양한 코드 체계 사용

**결측값 존재**
검사 누락, 진단 누락 등 불완전한 데이터

**비균형 데이터**
질환별/약물별 환자 수 차이가 큼

**다기관 데이터의 구조 차이**
EMR, CDM, 청구 데이터의 구조 상이

정형화된 의료 데이터는 고유한 특성을 가지고 있어 분석 시 특별한 주의가 필요합니다. 의료 데이터는 시간의 흐름에 따른 변화가 중요하며, 다양한 국제 표준 코드 체계를 사용합니다. 이러한 코드 체계는 분석 전 적절한 매핑 작업이 필요합니다.

또한 실제 임상 환경에서 수집된 데이터는 결측값이 많고, 질환별 환자 수에 큰 차이가 있는 비균형 데이터인 경우가 많습니다. 여러 의료기관에서 수집된 데이터는 구조적 차이가 있어 통합 분석 시 이를 고려한 전처리가 필수적입니다.

# 3. 비정형 데이터

## 의료 영상 데이터

CT, MRI, X-ray 이미지

컴퓨터 비전과 딥러닝(CNN 등) 기술을 활용하여 분석합니다.

폐렴이나 종양의 자동 진단 모델 개발에 활용되고 있습니다.

## 의료 텍스트 데이터

진료 기록, 소견서 같은 자유 서술식 텍스트

자연어처리(NLP)와 개체명 인식 기술로 분석합니다. 부작용 보고 탐지나 질환명 추출에 활용됩니다.

## 음성 데이터

의사-환자 대화, 수술 중 발화

음성 인식(STT)과 감정 분석 기술로 처리합니다. 자동 문서화나 감정 상태 모니터링에 활용되고 있습니다.

**비정형 의료 데이터는 구조화되지 않은 형태로 존재하여 분석에 특별한 기술이 필요합니다.**

**비정형 의료 데이터는 민감한 개인정보를 포함할 가능성이 높아 보안과 윤리적 관리가 매우 중요합니다.**

**또한 병변 표시, 텍스트 정제 등 별도의 라벨링 및 전처리 작업이 필요하며  파일 용량이 매우 커서 저장과 분석 환경에도 제약이 따를 수 있습니다.**

# 건강보험공단 빅데이터 소개

# 2부: 건강보험 공단 데이터 실습 1

# R 기초 실습

# 왜 의학 통계 실습에 R을 선택했나?

## 의학 연구에 특화된 통계 분석 기능

- 복잡한 통계 모형(생존분석, 로지스틱 회귀, 다변량 분석 등)에 강점
- `survival`, `tableone`, `cmprsk` 등 임상연구 맞춤 패키지 풍부
- 논문·학회에서 재현 가능한 코드 기반 분석 가능

## 고품질 시각화로 결과 전달력 향상

- `ggplot2`, `survminer` 등을 활용해 논문급 그래프 제작
- 데이터 패턴·변화를 직관적이고 설득력 있게 표현
- 색상·레이아웃 커스터마이징이 자유로워 학술발표에 최적

## 오픈소스의 장점과 접근성

- 무료, 설치와 사용이 간단하며 OS(Windows, Mac, Linux) 제약 없음
- 전 세계 연구자 커뮤니티와 패키지 공유·지속 업데이트
- 다양한 예제와 문서로 학습 곡선 완화

# data.table 이란?

R의 고성능 데이터 처리 패키지

- 대규모 데이터셋을 빠르고 효율적으로 다루기 위한 data.frame의 확장판
- 메모리 사용 최소화 + 속도 극대화

## 주요 특징

### 속도

- 수백만~수천만 행의 데이터도 초고속 처리
- 병렬 처리와 효율적인 메모리 관리 지원

### 간결한 문법

- DT[i, j, by] 형태로 필터링, 계산, 그룹화를 한 번에 작성
- SQL·pandas의 기능을 R 스타일로 구현

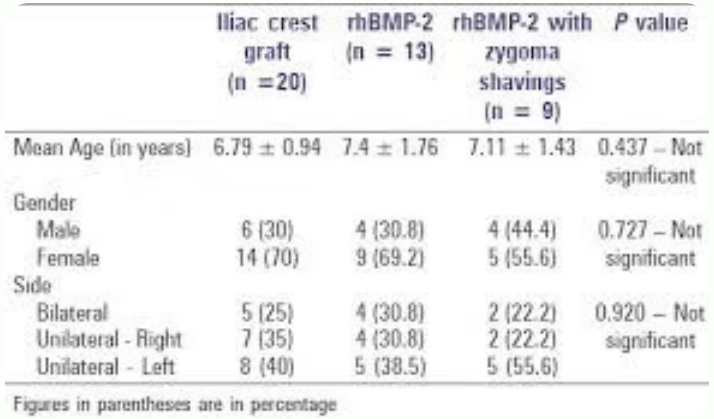
### 데이터 분석 최적화

- 대규모 의료·통계 데이터 전처리에 적합
- CSV, TSV 등 외부 파일도 빠르게 읽기(fread)

# 실습 !

# 기술 통계와 탐색적 데이터 분석

# tableone을 이용한 기술통계



| | Iliac crest graft (n =20) | rhBMP-2 (n = 13) | rhBMP-2 with zygoma shavings (n = 9) | P value |
|---|---|---|---|---|
| Mean Age (in years) | 6.79 ± 0.94 | 7.4 ± 1.76 | 7.11 ± 1.43 | 0.437 – Not significant |
| Gender | | | | |
| Male | 6 (30) | 4 (30.8) | 4 (44.4) | 0.727 – Not significant |
| Female | 14 (70) | 9 (69.2) | 5 (55.6) | |
| Side | | | | |
| Bilateral | 5 (25) | 4 (30.8) | 2 (22.2) | 0.920 – Not significant |
| Unilateral - Right | 7 (35) | 4 (30.8) | 2 (22.2) | |
| Unilateral - Left | 8 (40) | 5 (38.5) | 5 (55.6) | |

Figures in parentheses are in percentage

연구 대상자 기본 특성 표 (Table 1) 예시

## 기술통계 (Descriptive Statistics)

데이터의 **전반적 특성을 한눈에 파악**하는 것이 목적입니다.

- 평균, 표준편차, 빈도, 비율 등 **기본 지표 산출**
- 의학 데이터에서의 역할: 연구 집단의 특성 비교, 이상치 확인, 분석 설계 기초

## 왜 중요한가?

기술통계는 데이터 분석의 첫 단추이자 핵심입니다.

- 연구 대상 집단 간 **동질성 판단**
- 변수 분포 확인 → **적절한 통계 분석 방법 결정**
- 최종 분석 **결과 해석의 기반 자료** 제공

## tableone 패키지 소개

- **연속형·범주형 변수** 동시 처리 가능
- 그룹 간 비교 및 **p-value 자동 생성**
- 논문 형식의 **'Table 1' 제작에 최적화**

**실습:** `tableone`으로 연구 대상자의 기본 특성 표 만들기

# 실습 !

# 시각화를 통한 탐색적 데이터 분석 (EDA)

데이터의 전반적인 패턴과 관계를 시각적으로 이해하는 것을 목적으로 합니다.

이를 통해 데이터의 분포, 경향, 이상치, 그리고 그룹 간의 차이를 효과적으로 파악할 수 있습니다. 의료 데이터 분석에서 시각화는 데이터의 중요성을 강조하고 통찰력을 제공하는 핵심적인 단계입니다.

## 의료 데이터 분석에서의 시각화 특징

- 군별 비교, 추세선, 생존곡선 등 논문 표준형식에 맞는 시각화 요구
- 정확하고 신뢰성 있는 그래프를 통해 연구 결과의 타당성 확보
- 복잡한 의료 데이터를 직관적으로 이해하고 전달하는 데 필수적

## ggpubr 패키지 장점

- **ggplot2** 기반으로, 논문용 고품질 그래프를 손쉽게 제작
- 평균±표준편차, boxplot, barplot, scatterplot 등 의료 논문에서 자주 사용되는 형식 지원
- p-value 및 통계적 비교 결과를 그래프에 자동으로 표기하여 분석 시간 단축

**실습:** ggpubr 패키지를 활용하여 논문급 그래프 2~3종 만들기

# 실습 !

# 3부: 건강보험 공단 데이터 실습 2

# 로지스틱 회귀(Logistic Regression) 소개

로지스틱 회귀는 통계학에서 널리 사용되는 분류 모델 중 하나로, 특히 의학 연구에서 중요한 역할을 합니다.

이 모델은 특정 사건이 발생할 확률을 예측하고, 다양한 설명변수들이 그 사건에 미치는 영향을 분석하는 데 활용됩니다.

- **목적:** 이진형 결과(예/아니오, 생존/사망 등)를 설명변수로 예측
- **의학 연구 예:** 약물 복용 여부가 질병 발생 확률에 미치는 영향
- **특징:** 결과를 **확률(0~1)**로 추정하여 해석이 직관적

의료 데이터 분석에서 로지스틱 회귀는 질병 발생 위험 예측, 치료 효과 분석, 예후 예측 등 다양한 분야에서 강력한 도구로 활용됩니다.

이는 복잡한 의학 데이터를 통계적으로 명확하게 해석하고, 임상적 의사결정을 지원하는 데 기여합니다.

# 로지스틱 회귀의 수학적 원리

## 1 로지스틱 회귀 기본식

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

- **p = P(Y=1)** : 사건 발생 확률
- **p/(1-p)** : 사건 발생 오즈(odds)
- **$\beta_j$** : 변수 $X_j$의 회귀계수

## 2 오즈(odds) 형태로 변환

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}$$

- 로그함수의 역함수는 지수함수(exp)
- 따라서 logit(p) = $\beta_0 + \beta_1 X_1$ → 오즈 = $\exp(\beta_0 + \beta_1 X_1)$

## 3 변수 1단위 증가 시 오즈비 계산

$$OR = \frac{\text{odds when } X_1 + 1}{\text{odds when } X_1} = e^{\beta_1}$$

- $X_1$이 1 증가하면 오즈는 e^$\beta_1$배로 증가
- 따라서 $\beta_1$의 지수(exp)를 취하면 오즈비(OR)가 됨

## 4 핵심 포인트

- 로지스틱 회귀는 로그 오즈(logit)를 선형결합으로 모델링
- $\exp(\beta_j)$ = 변수 1단위 증가 시 오즈비
- 회귀계수 $\beta_j$가 양수면 오즈 증가, 음수면 감소

# 생활습관·인구학적 요인과 당뇨병 발생 연관성: 로지스틱 회귀분석

**연구 배경 및 목적**

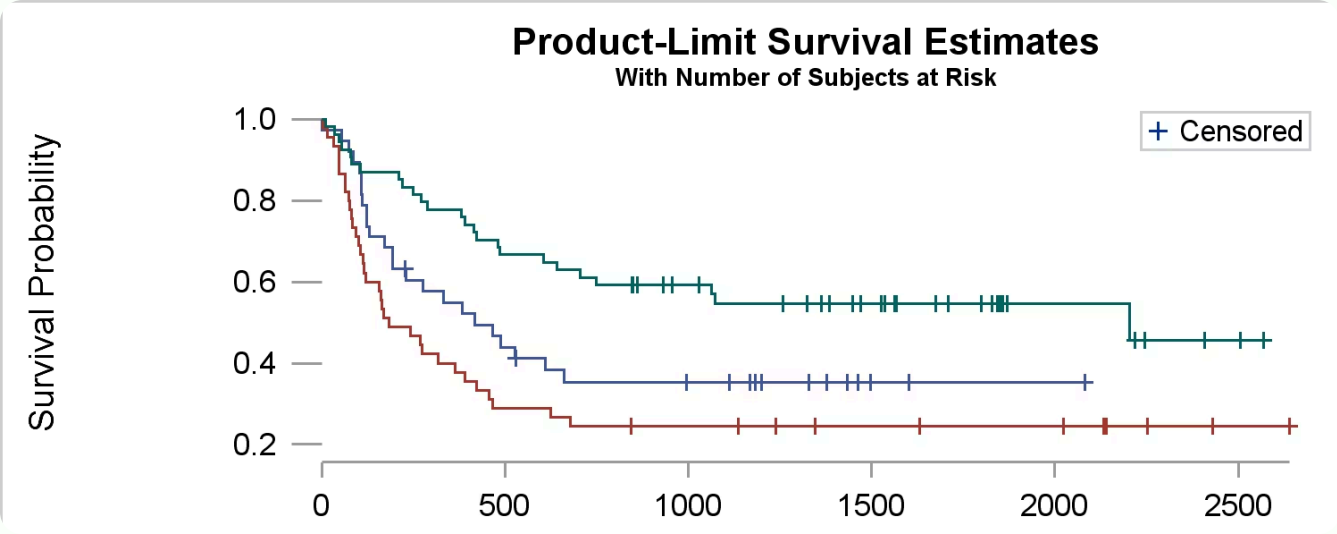
당뇨병은 전 세계적으로 유병률이 증가하고 있는 주요 만성질환으로,
흡연, 비만, 나이 등 다양한 생활습관 및 인구학적 요인과 밀접한 관련이 있습니다.
이러한 요인들의 영향을 정량적으로 파악하는 것은 **고위험군 선별과 예방 전략 수립**에 필수적입니다.

본 연구는 **흡연 여부, 비만, 나이**가 당뇨병 발생에 미치는 영향을 파악하고,
Logistic Regression을 통해 각 요인의 **오즈비(OR)**를 산출하여
위험 요인으로서의 기여도를 비교·평가하고자 합니다.

**데이터 및 방법론**

- **데이터 출처:** 건강보험공단 제공 1000명의 샘플 데이터 (2002년부터 2015년까지의 데이터).
- **당뇨 환자 정의:** E10~E14 에 해당되는 ICD 코드 보유자
- **대조군 정의:** 전체 기간 동안 당뇨 진단이 없는 자.

# 실습 !




# Kaplan-Meier Plot 소개

### 시간에 따른 생존 확률 곡선

- 그룹 간 생존 패턴 비교
- 비모수적 방법 → 분포 가정 불필요
- 생존 분석에서의 table1, 직관적으로 표현 됨

Kaplan-Meier Plot은 시간에 따른 **생존 확률**을 시각화하여 치료군과 대조군의 생존 패턴 차이를 직관적으로 보여줍니다.
차이의 통계적 유의성은 **log-rank test**로 확인하며, 이러한 시각화는 **임상적 의사결정과 연구 해석**에 중요한 기반을 제공합니다.

# Log-rank Test

**목적:** 그룹 간 생존곡선 차이 통계적 검정

## 원리

- 각 사건 발생 시점마다 risk set(위험군) 확인
- **관찰된 사건 수(Observed)**와 기대 사건 수(Expected) 비교
- 구간별 차이를 모두 합산 → Chi-square 검정

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{V_i}$$

- $O_i$ : i번째 사건에서 관찰된 사건 수
- $E_i$ : i번째 사건에서 기대 사건 수
- $V_i$ : 분산

## 비례위험 가정과의 관계

Log-rank Test는 구간별 기대 사건 수 계산 시 HR(Hazard Ratio)이 시간에 따라 일정하다고 가정합니다.

즉, **"각 시점에서 그룹 간 상대 위험이 비례한다"**는 가정이 암묵적으로 포함되어 있습니다.

만약 비례위험 가정이 깨지면 검정 결과의 신뢰도가 낮아질 수 있습니다.

# CCI 점수에 따른 사망률 분석

## 연구 배경 및 목적

CCI는 환자의 다양한 만성 질환을 종합적으로 평가하여 예후를 파악하는 데 사용되는 지표입니다.

이 연구를 통해 CCI 점수가 높을수록 사망률도 높아지는 지를 확인하고자 합니다.

## 데이터 및 방법론

- **데이터 출처**: 건강보험공단 제공 1000명의 샘플 데이터 (2002년부터 2015년까지의 데이터).
- **기준 날짜**: 건강검진 받은 가장 첫 날짜를 기준으로 설정.
- **CCI 계산**: 기준 날짜로부터 과거 1년 이내의 질병을 평가하여 CCI 점수 산출.
- **그룹 분류**: CCI 3점 미만/ CCI 3점 이상.
- **분석 방법**: Cox 비례 위험 회귀 분석 및 Kaplan-Meier 생존 곡선 플로팅.

| | | | | |
|---|---|---|---|---|
| **심근경색증: 1점** | **울혈성 심부전: 1점** | **만성 폐 질환: 1점** | **류마티스 질환: 1점** | **소화성 궤양 질환: 1점** |
| **뇌혈관 질환: 1점** | **말초 혈관 질환: 1점** | **치매: 1점** | **경증 간 질환: 1점** | **당뇨병(합병증 없음): 1점** |
| **당뇨병(합병증 있음): 2점** | **신장 질환: 2점** | **편마비 및 마비: 2점** | **악성종양의 백혈병/림프종: 2점** | **중등도/심각한 간 질환: 3점** |
| **전이성 고형 종양: 6점** | **AIDS/HIV: 6점** | | | |

# Cox 비례위험모형 소개

## 1. 목적

- **Time-to-event 데이터 분석:** 생존 시간, 재발까지 걸린 시간, 치료 후 합병증 발생 시간 등 특정 사건이 발생하기까지의 시간을 분석합니다.

## 2. 의학적 예시

- **치료 효과 비교:** 치료 시작 후 사망까지 걸린 시간 비교 (예: 항암제 단독 투여 vs 병용 투여 그룹의 생존 위험 비교).
- **위험 요인 분석:** 특정 요인(나이, 성별, 기저 질환)이 환자의 생존 또는 사건 발생에 미치는 영향을 평가합니다.

## 3. 주요 특징

- **비례위험 가정(Proportional Hazards) 기반:**
  - 변수에 따른 상대 위험(Hazard Ratio, HR)은 시간에 따라 일정하다고 가정합니다.
  - 절대 위험(Baseline Hazard, $h_0(t)$)를 명시적으로 지정하지 않고도 계수($\beta$) 추정을 통해 요인별 상대 위험(HR)을 계산할 수 있어 모형의 유연성이 높습니다.

# Cox 비례위험 모형의 수학적 구조

$$h(t|X) = h_0(t) \cdot e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}$$

- **h(t|X)** : 시간 t에서 사건 발생 위험(hazard)
- **$h_0(t)$** : 기준위험(baseline hazard), X=0일 때 위험
- **$\beta_j$** : 변수 $X_j$의 회귀계수, 로그 위험비(log hazard ratio)

## 2. 기준위험(Baseline hazard)

정의: 모든 설명변수 X가 0일 때의 시간 t에서 사건 발생 위험

### 특징:

- 시간에 따라 변할 수 있음 → Cox 모형은 $h_0(t)$의 형태를 특정하지 않음
- HR 계산 시 상대비율이므로 기준위험은 직접적으로 계산되지 않아도 됨
- 기준위험을 통해 개별 변수의 **상대적 위험(HR)**을 평가 가능 즉, **"모든 변수 영향이 없을 때 사건 발생 위험"**이 기준 위험

## 3. 비례위험 가정 (Proportional Hazards Assumption)

Cox 모형의 핵심 가정: 변수에 따른 위험비(HR)가 시간에 따라 변하지 않는다

$$h(t|X_i)/h(t|X_j) = e^{\beta(X_i - X_j)}$$

- 시간 t에 관계없이 HR이 일정
- 위 가정이 깨지면 Cox 모형 결과의 해석이 어려워짐
- 검증 방법: Schoenfeld residual, 시각적 log(-log) 생존곡선 확인 등
- HR = e^$\beta_j$ (변수 1단위 증가 시 위험비)

# 실습 !

# Q&A

문의사항이 있으시면 아래 연락처로 연락해주세요


- 이메일: joes@zarathu.com
- github: **https://github.com/joeeerere**
- linked in: **https://www.linkedin.com/in/%EC%9D%80%EC%84%9C-%EC%A1%B0-485720303/**

감사합니다.