의료와 데이터사이언스 4주차

- 다양한 종류의 의료데이터
- EMR 정형 데이터

Kwangsoo Kim, PhD

Professor

AI Healthcare Institute
Seoul National University Hospital

SNUH

서울대학교
SEOUL NATIONAL UNIVERSITY

CONTENTS

◆ 의료데이터의 종류와 특징

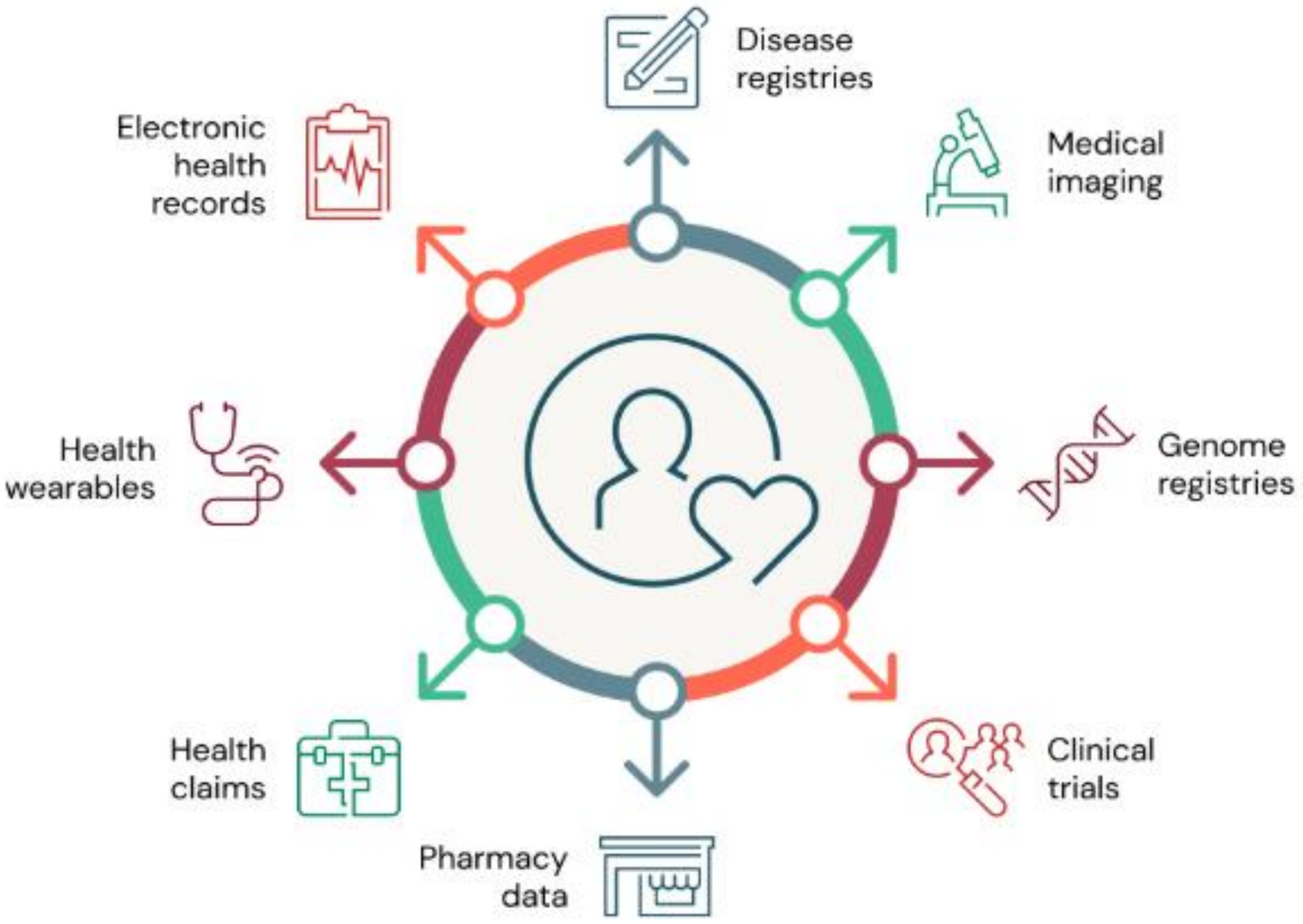
◆ EMR 정형데이터와 MIMIC-IV 사례

- EMR 정형데이터 개요
- MIMIC-IV 데이터 구조와 연구 활용

Tyes of Medical Data

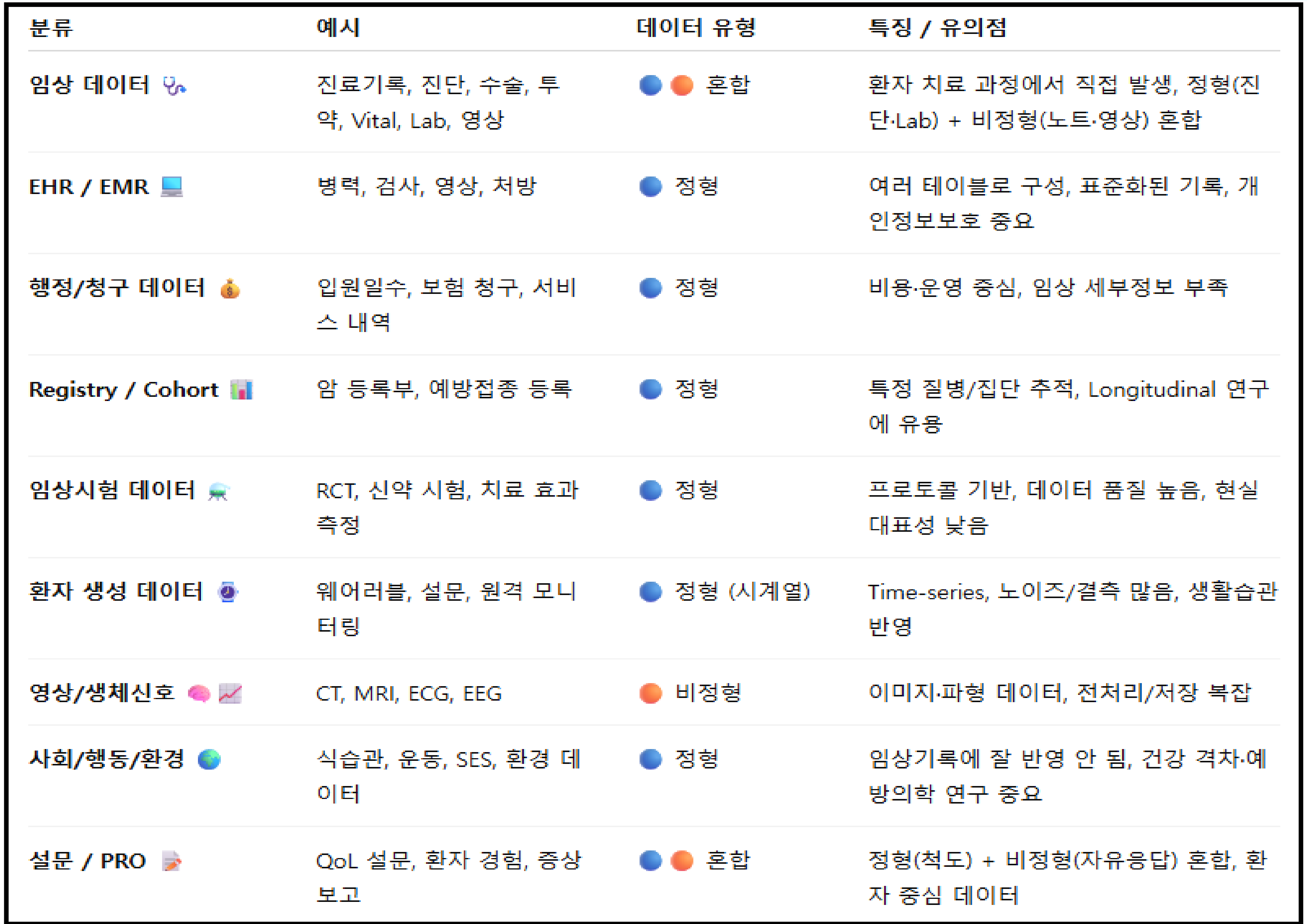
SNUH

A single patient produces 80+ megabytes of medical data every year



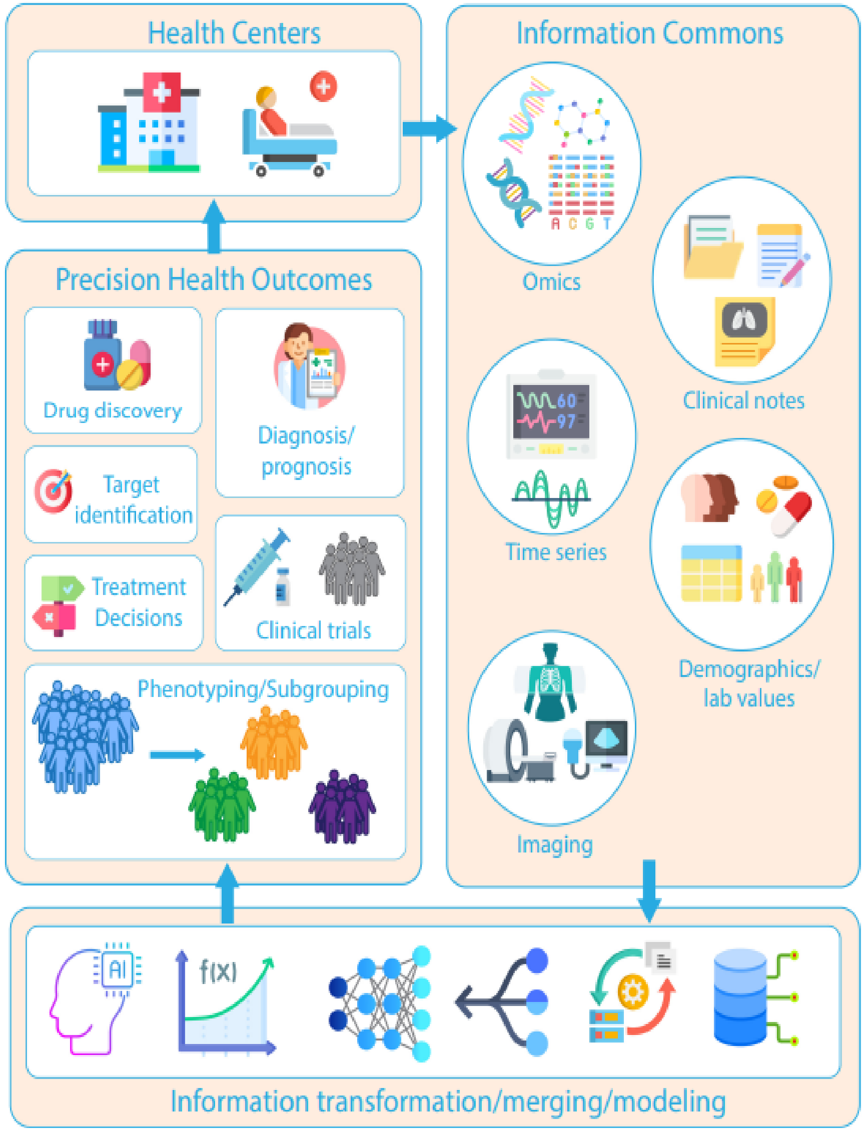
Databricks - Unlocking the power of health data with a modern data lakehouse

Tyes of Medical Data

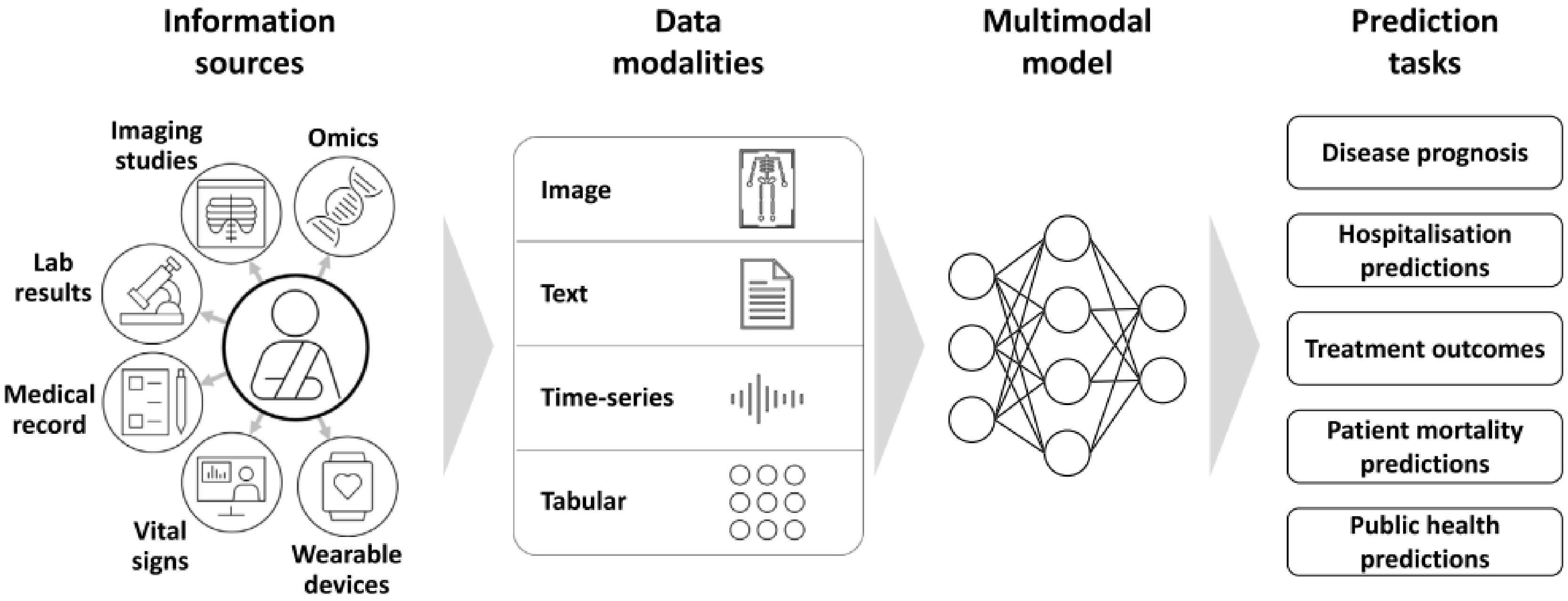


분류	예시	데이터 유형	특징 / 유의점
임상 데이터 🩺	진료기록, 진단, 수술, 투약, Vital, Lab, 영상	🔵🟠 혼합	환자 치료 과정에서 직접 발생, 정형(진단·Lab) + 비정형(노트·영상) 혼합
EHR / EMR 💻	병력, 검사, 영상, 처방	🔵 정형	여러 테이블로 구성, 표준화된 기록, 개인정보보호 중요
행정/청구 데이터 💰	입원일수, 보험 청구, 서비스 내역	🔵 정형	비용·운영 중심, 임상 세부정보 부족
Registry / Cohort 📊	암 등록부, 예방접종 등록	🔵 정형	특정 질병/집단 추적, Longitudinal 연구에 유용
임상시험 데이터 🧪	RCT, 신약 시험, 치료 효과 측정	🔵 정형	프로토콜 기반, 데이터 품질 높음, 현실 대표성 낮음
환자 생성 데이터 ⌚	웨어러블, 설문, 원격 모니터링	🔵 정형 (시계열)	Time-series, 노이즈/결측 많음, 생활습관 반영
영상/생체신호 🧠 📈	CT, MRI, ECG, EEG	🟠 비정형	이미지·파형 데이터, 전처리/저장 복잡
사회/행동/환경 🌍	식습관, 운동, SES, 환경 데이터	🔵 정형	임상기록에 잘 반영 안 됨, 건강 격차·예방의학 연구 중요
설문 / PRO 📝	QoL 설문, 환자 경험, 증상 보고	🔵🟠 혼합	정형(척도) + 비정형(자유응답) 혼합, 환자 중심 데이터

Tyes of Medical Data



Kline et al., npj Digital Medicine, 2022 (https://doi.org/10.1038/s41746-022-00712-8)



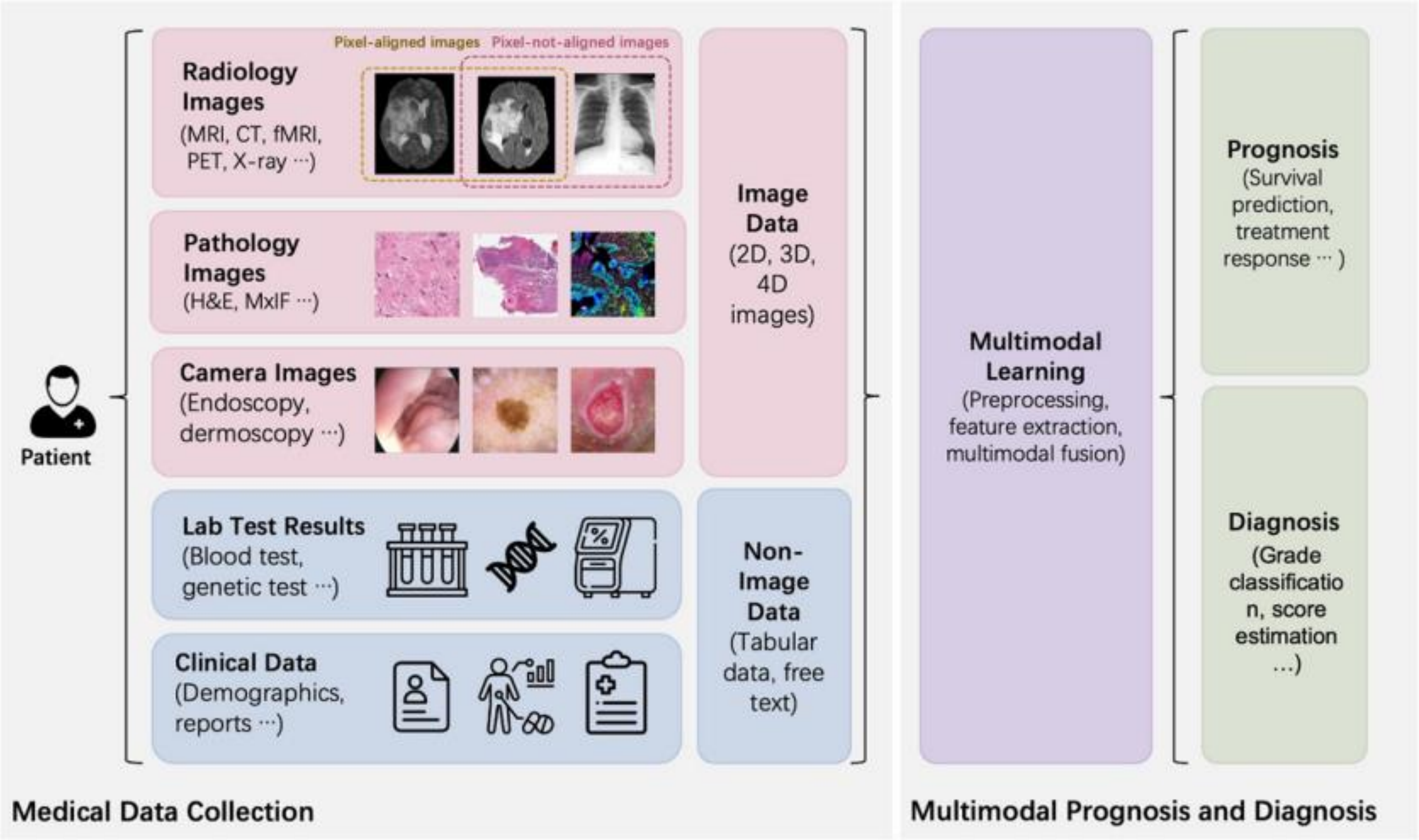
Krones et al., Information Fusion, 2025

다양한 의료데이터의 순환적 활용

- 임상 현장에서 생성되는 다양한 데이터 (EMR, 영상, 오믹스 등)는 정보 저장소에서 통합,분석되어 알고리즘 모델링에 활용됩니다.
- 이를 통해 얻은 인사이트는 다시 의료 현장으로 환류되어 정밀의료를 가능하게 합니다.

Tyes of Medical Data

SNUH

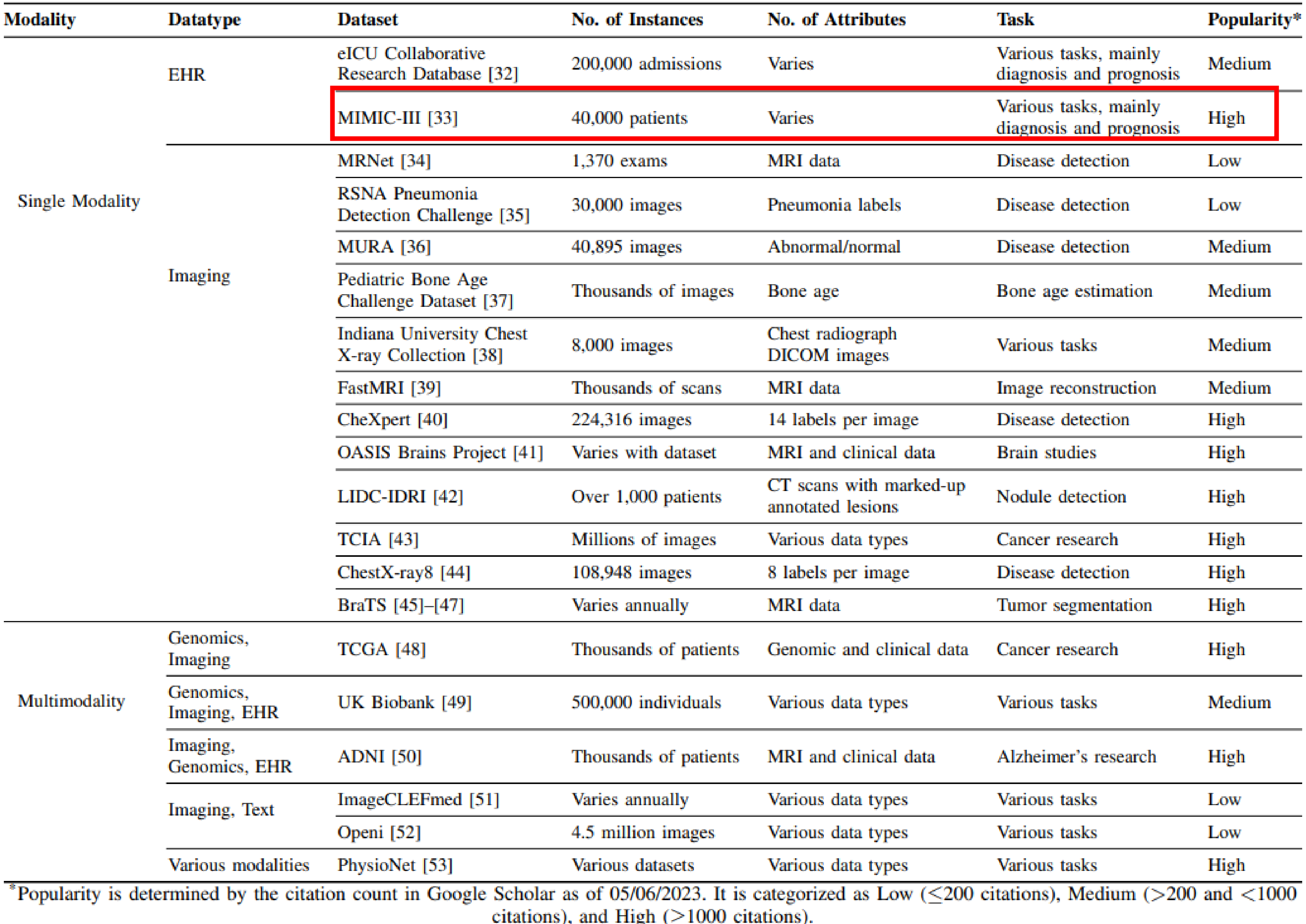

Pixel-aligned images
Pixel-not-aligned images
Radiology Images
(MRI, CT, fMRI, PET, X-ray ···)
Pathology Images
(H&E, MxIF ···)
Camera Images
(Endoscopy, dermoscopy ···)
Lab Test Results
(Blood test, genetic test ···)
Clinical Data
(Demographics, reports ···)
Patient
Image Data
(2D, 3D, 4D images)
Non-Image Data
(Tabular data, free text)
Medical Data Collection
Multimodal Learning
(Preprocessing, feature extraction, multimodal fusion)
Prognosis
(Survival prediction, treatment response ···)
Diagnosis
(Grade classification, score estimation …)
Multimodal Prognosis and Diagnosis

다양한 의료데이터의와 그 융합

- 의료데이터는 크게 영상 데이터와 비영상 데이터로 구분할 수 있습니다.
- 서로 다른 데이터들은 각각 고유한 특성과 장점을 가지며, 멀티모달 학습 기법으로도 분석될 수 있습니다.
- 이를 통해 단일 데이터로는 얻기 어려운 통찰을 확보하고, 질병의 진단과 예후예측에 활용될 수 있습니다.

Cui et al., Prog Biomed Eng, 2023

Tyes of Medical Data



Modality	Datatype	Dataset	No. of Instances	No. of Attributes	Task	Popularity*
Single Modality	EHR	eICU Collaborative Research Database [32]	200,000 admissions	Varies	Various tasks, mainly diagnosis and prognosis	Medium
		MIMIC-III [33]	40,000 patients	Varies	Various tasks, mainly diagnosis and prognosis	High
	Imaging	MRNet [34]	1,370 exams	MRI data	Disease detection	Low
		RSNA Pneumonia Detection Challenge [35]	30,000 images	Pneumonia labels	Disease detection	Low
		MURA [36]	40,895 images	Abnormal/normal	Disease detection	Medium
		Pediatric Bone Age Challenge Dataset [37]	Thousands of images	Bone age	Bone age estimation	Medium
		Indiana University Chest X-ray Collection [38]	8,000 images	Chest radiograph DICOM images	Various tasks	Medium
		FastMRI [39]	Thousands of scans	MRI data	Image reconstruction	Medium
		CheXpert [40]	224,316 images	14 labels per image	Disease detection	High
		OASIS Brains Project [41]	Varies with dataset	MRI and clinical data	Brain studies	High
		LIDC-IDRI [42]	Over 1,000 patients	CT scans with marked-up annotated lesions	Nodule detection	High
		TCIA [43]	Millions of images	Various data types	Cancer research	High
		ChestX-ray8 [44]	108,948 images	8 labels per image	Disease detection	High
		BraTS [45]–[47]	Varies annually	MRI data	Tumor segmentation	High
Multimodality	Genomics, Imaging	TCGA [48]	Thousands of patients	Genomic and clinical data	Cancer research	High
	Genomics, Imaging, EHR	UK Biobank [49]	500,000 individuals	Various data types	Various tasks	Medium
	Imaging, Genomics, EHR	ADNI [50]	Thousands of patients	MRI and clinical data	Alzheimer's research	High
	Imaging, Text	ImageCLEFmed [51]	Varies annually	Various data types	Various tasks	Low
		Openi [52]	4.5 million images	Various data types	Various tasks	Low
	Various modalities	PhysioNet [53]	Various datasets	Various data types	Various tasks	High

*Popularity is determined by the citation count in Google Scholar as of 05/06/2023. It is categorized as Low ($\leq$200 citations), Medium ($>$200 and $<$1000 citations), and High ($>$1000 citations).

Shaik et al., arXiv, 2023)

CONTENTS

MIMIC-IV

MIMIC-CXR

MIMIC-NOTE

MIMIC-ECG / Waveform

Trends in Medical Data Science

SNUH

Open Dataset

- Publicly Available Datasets
- Open Source, Collaborations
- ICU Datasets
  - **MIMIC (MIT, US)**
  - K-MIMIC (SNUH, Korea)
  - eICU-CRD (MIT, US)
  - AmsterdamUMCdb (Netherland)
  - HiRID (Bern Univ, Switzerland)
  - SICdb (Salzburg Univ, Austria)
- Perioperative Datasets
  - VitalDB, INSPIRE (SNUH, Korea)
  - MOVER (UC Irvine, US)

데이터는 각 기관에 남겨둔 채, 분석/모델 학습을 네트워크로 연결해서 공동으로 하는 방식

Federated Network

- Distributed, Restricted Data
- Moving Models and Queries
- Common Data Model: OMOP-CDM
- Research Networks
  - OHDSI Network
  - Research Border Free Zone
- Platforms
  - Mayo Clinic Platform (US)
  - FeederNet (47 hospitals, 57M pts, Korea)

➡ **의료 데이터에서 환자 프라이버시를 지키면서도 대규모 연구 AI 학습이 가능하게 해주는 핵심 인프라**

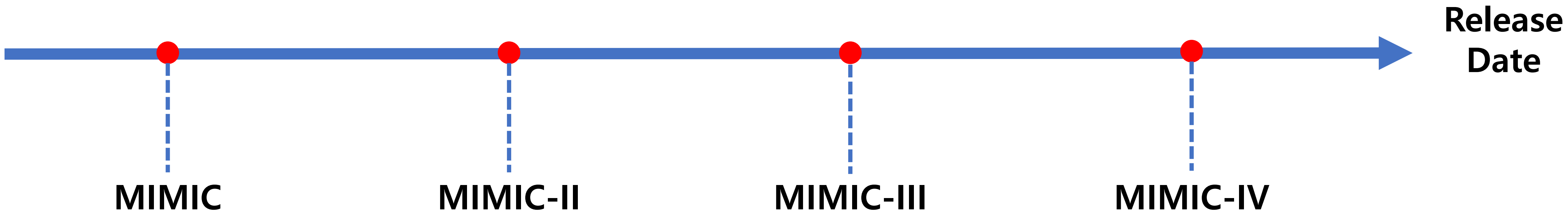
H.C. Lee. (2023.10) Perioperative Data Science, Global trends and Future Directions

MIMIC Dataset

MIMIC (Medical Information Mart for Intensive Care)는 **MIT**의 **Laboratory for Computational Physiology (LCP)**가 개발한, 자유롭게 접근할 수 있는 (승인 필요) **대규모 중환자실 환자 전자의무기록지** (EHR,EMR) 데이터 베이스

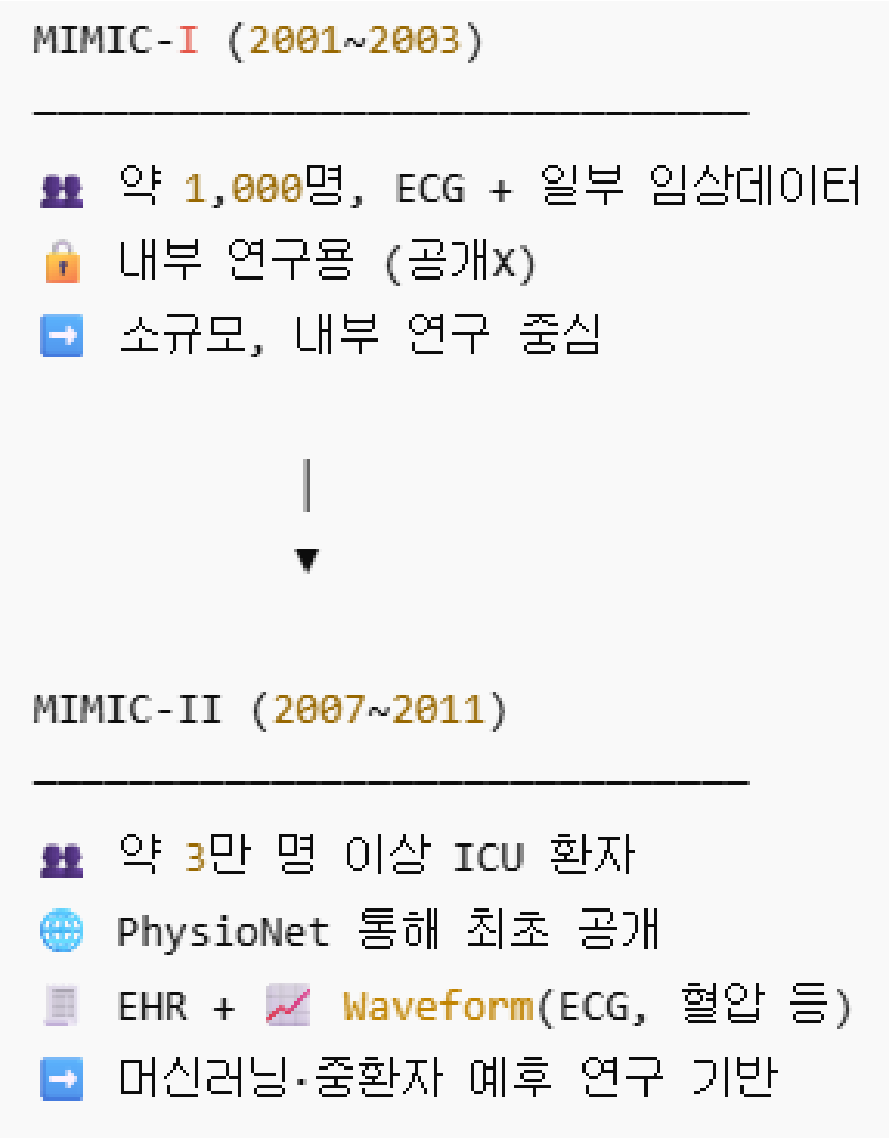
- 최신버전: MIMIC-IV (v2.2 기준, 2008-2019년 데이터 포함)
- 환자 수: 약 38만 명 이상의 환자, 50만 건 이상의 입원 기록
- 수집 기관: Beth Israel Deaconess Medical Center (BIDMC, Boston, MA)
- 목적: 인공지능, 임상연구, 생체신호 분석, 전산병리 등 다양한 학문 분야에서 재현성 있는 연구 데이터 셋 제공

- **M**edical **I**nformation **M**art for **I**ntensive **C**are
- A large, single-center database comprising information relating to patients admitted to critical care units at a large tertiary care hospital



MIMIC Dataset

SNUH



MIMIC-I (2001~2003)

────────────────────────────────

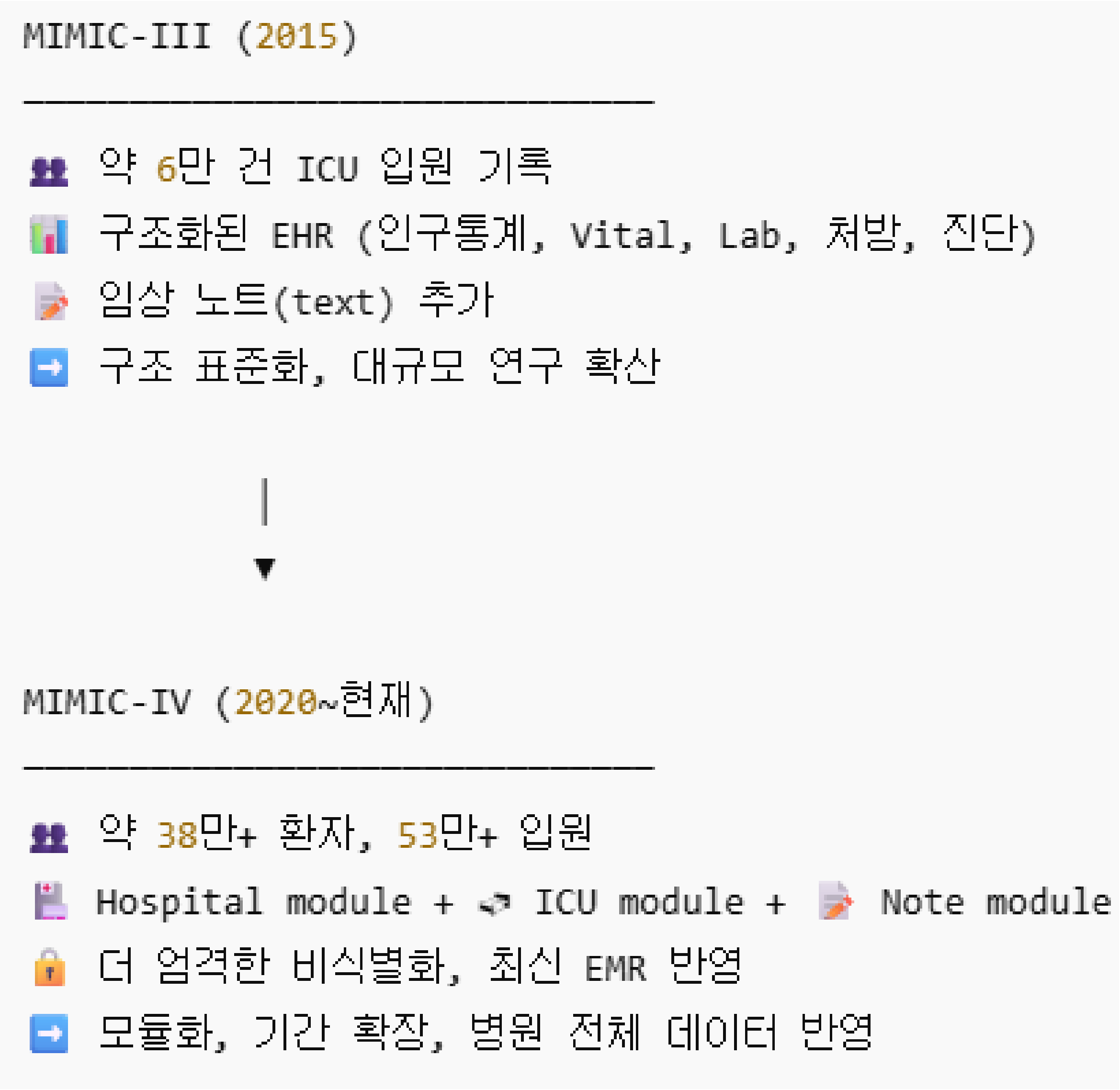
👥 약 1,000명, ECG + 일부 임상데이터
🔒 내부 연구용 (공개X)
➡️ 소규모, 내부 연구 중심

│
▼

MIMIC-II (2007~2011)

────────────────────────────────

👥 약 3만 명 이상 ICU 환자
🌐 PhysioNet 통해 최초 공개
🧾 EHR + 📈 Waveform(ECG, 혈압 등)
➡️ 머신러닝·중환자 예후 연구 기반



MIMIC-III (2015)

──────────────────────────────

👥 약 6만 건 ICU 입원 기록
📊 구조화된 EHR (인구통계, Vital, Lab, 처방, 진단)
📝 임상 노트(text) 추가
➡️ 구조 표준화, 대규모 연구 확산

│
▼

MIMIC-IV (2020~현재)

──────────────────────────────

👥 약 38만+ 환자, 53만+ 입원
🏥 Hospital module + 🛏️ ICU module + 📝 Note module
🔒 더 엄격한 비식별화, 최신 EMR 반영
➡️ 모듈화, 기간 확장, 병원 전체 데이터 반영

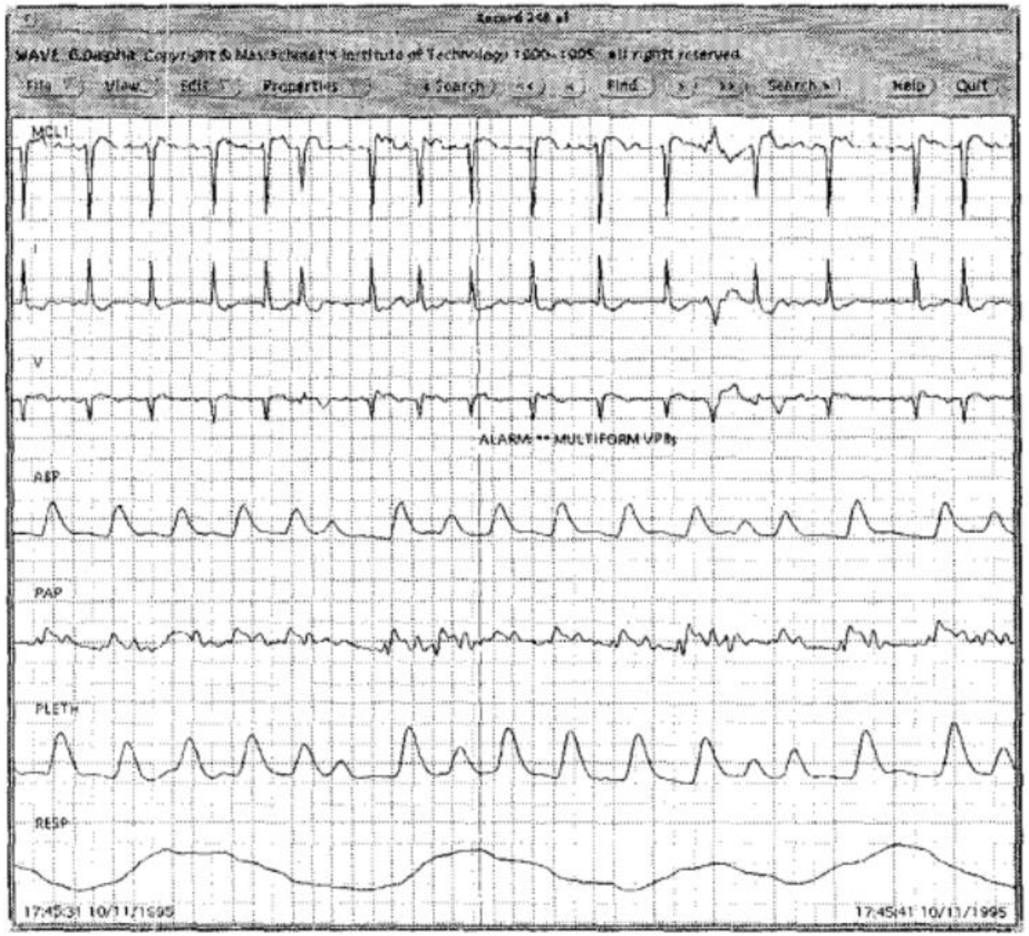
MIMIC I

Multiparameter Intelligent Monitoring for Intensive Care

- In 1996, 90 ICU patients
- 20 hours of ECG, ABP, PAP, PLETH
- First attempt to build a collection of multi-parameter recordings of ICU patients

Roger Mark

George Moody



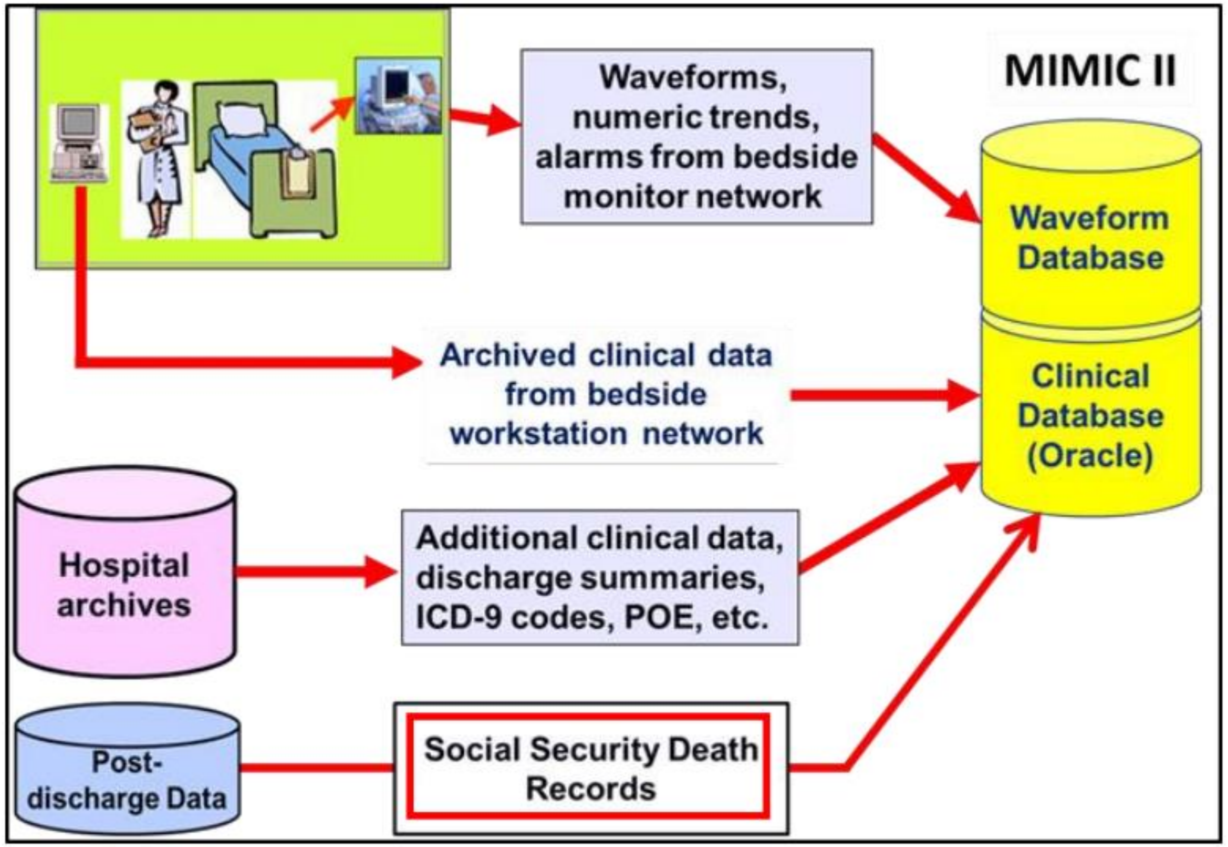
Record 248 a1

WAVE 6.Dsphit Copyright © Massachusetts Institute of Technology 1990-1995 all rights reserved.

Moody GB, Mark RG, A Database to Support Development and Evaluation of Intelligent Intensive Care Monitoring, Computers in Cardiology 1996

MIMIC II

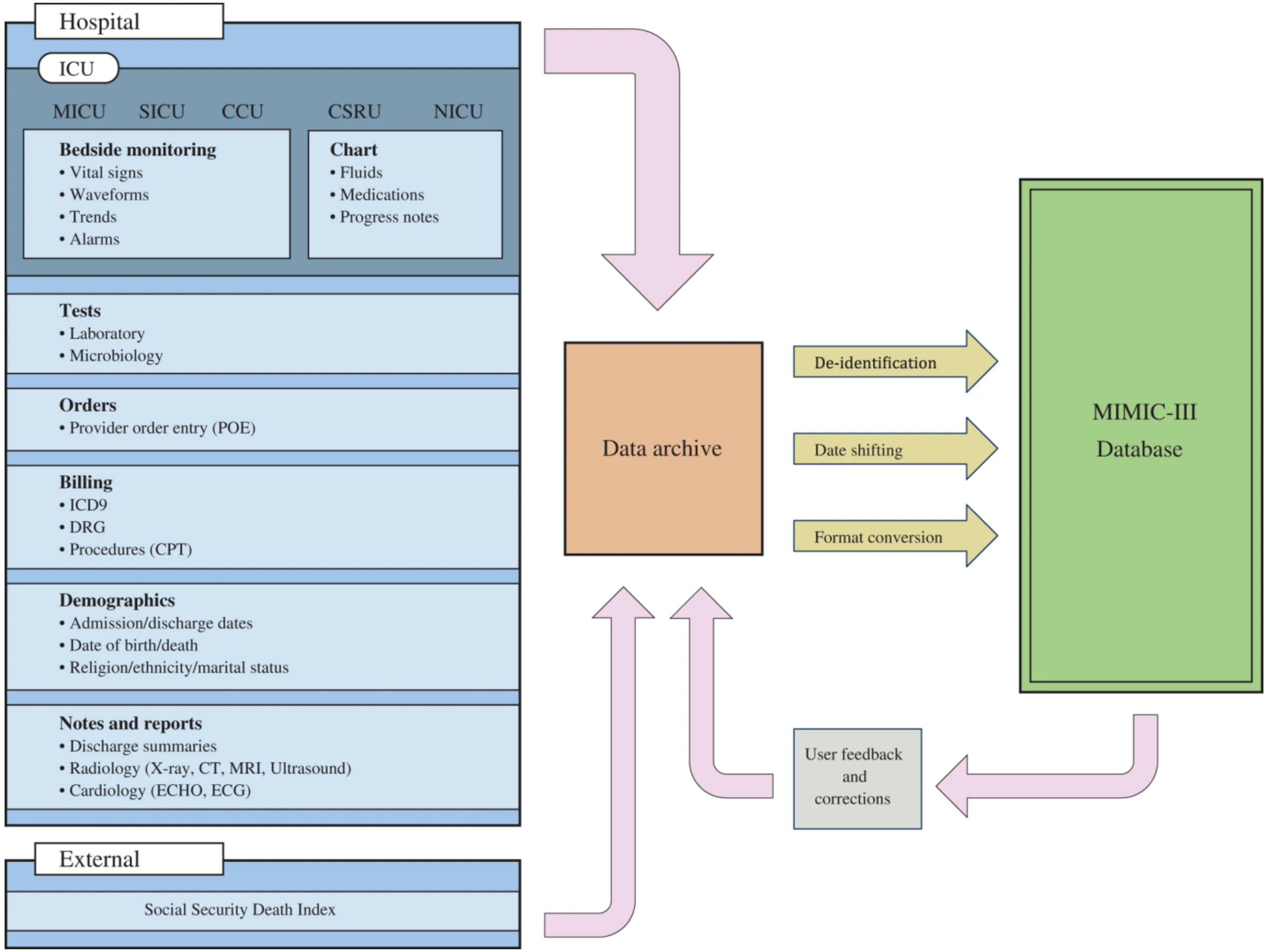
- In 2003, MIT, Philips, BIH received NIH fund
- Clinical data were collected between 2001 and 2007 from a variety of ICUs (MICU, SICU, CCU, NICU)
- Publicly released in 2010
- 25,328 ICU stays, 22,870 admissions


MIMIC II
Waveforms, numeric trends, alarms from bedside monitor network
Waveform Database
Archived clinical data from bedside workstation network
Clinical Database (Oracle)
Hospital archives
Additional clinical data, discharge summaries, ICD-9 codes, POE, etc.
Post-discharge Data
Social Security Death Records

Joon Lee, et al. Open-access MIMIC-II database for intensive care research. Annu Int Conf IEEE Eng Med Biol Soc, 2011

MIMIC III

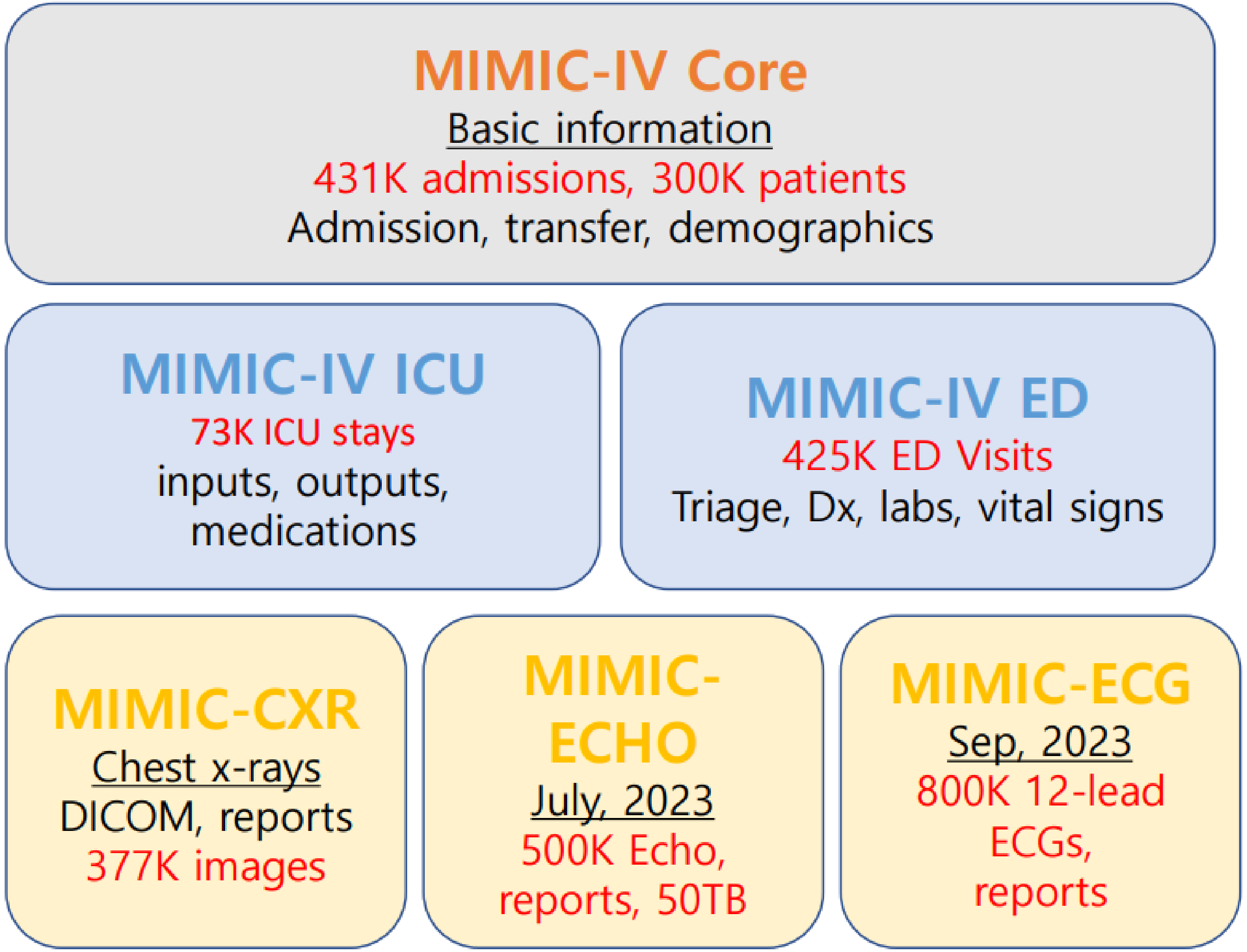
- In 2015, major update: includes **clinical notes**
- Renamed to the "Medical Information Mart for the Intensive Care"
- >40,000 patients, >50,000 ICU stays, between 2001-2012
- Cited by >**7000 publications (24.07.31)**



Johnson, A., Pollard, T., Shen, L. *et al.* MIMIC-III, a freely accessible critical care database. *Scientific Data*, 2016.

MIMIC IV

- In March **2021**
- between 2008 – 2019 (overlapped patients btwn 2008-2012)
- Admissions: 431,231
  Patients: 299,712
  ICU stays: 73,181 (v2.2)
- **Modular approach**



MIMIC-IV Core
Basic information
431K admissions, 300K patients
Admission, transfer, demographics

MIMIC-IV ICU
73K ICU stays
inputs, outputs, medications

MIMIC-IV ED
425K ED Visits
Triage, Dx, labs, vital signs

MIMIC-CXR
Chest x-rays
DICOM, reports
377K images

MIMIC-ECHO
July, 2023
500K Echo, reports, 50TB

MIMIC-ECG
Sep, 2023
800K 12-lead ECGs, reports

Alistair Johnson et al. MIMIC-IV, a freely accessible electronic health records dataset. Scientific Data 2023.

Patient composition

- **MIMIC-IV (3.0)** : patients admitted to the ED and the ICU
  - **364,627 unique patients**
    - **223,452 patients** : at least one hospitalization
    - **141,175 patients** : only seen in the ED

Data Structure

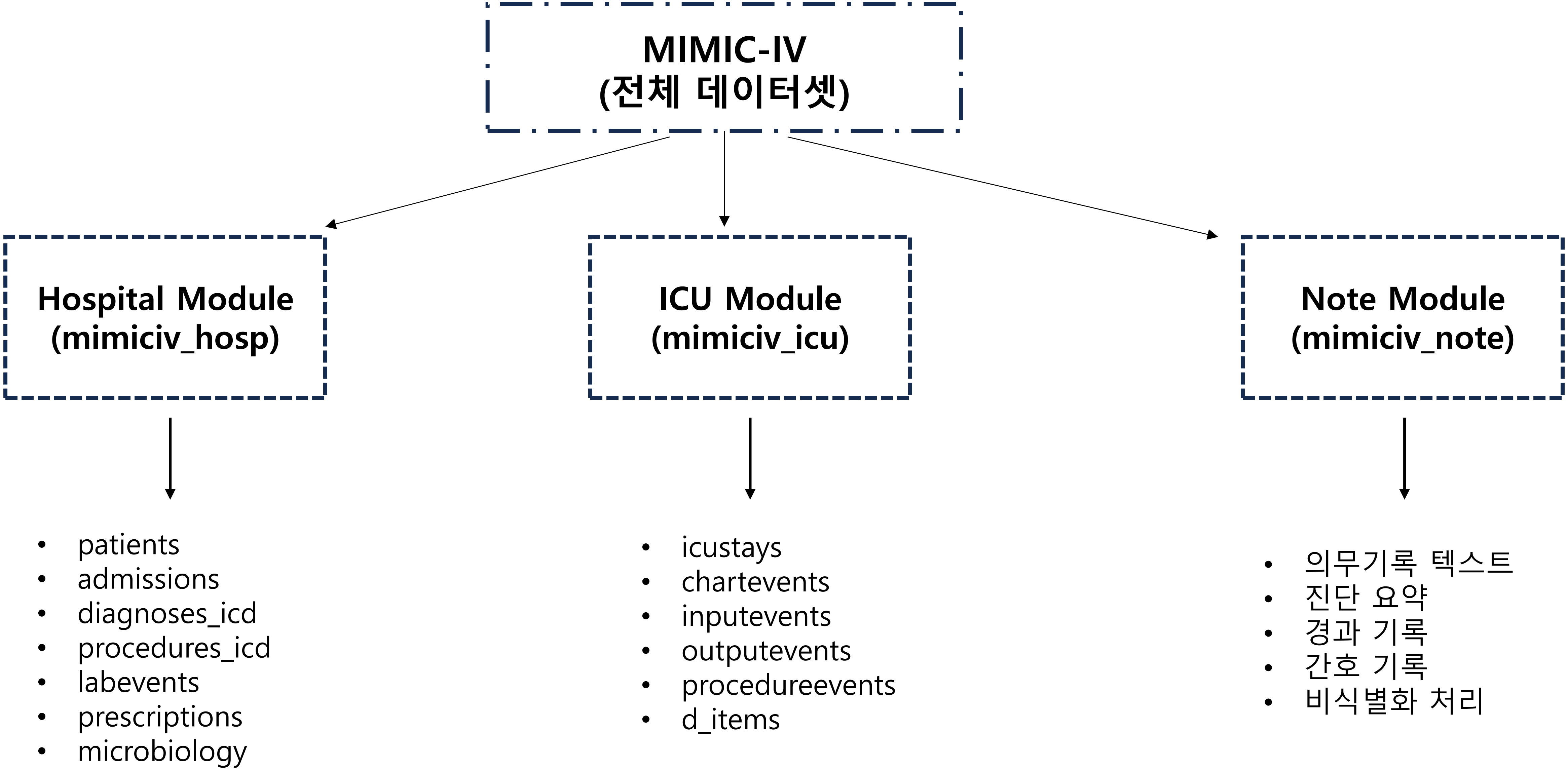
MIMIC-IV 는 크게 세 가지 모듈(스키마)로 나뉠 수 있음

1. **Hospital Module (mimiciv_hosp)**
   - 전체 병원 진료 기록 (ICU 포함, 외래/일반병동 포함)
   - 주요 테이블:
     - **patients**: 인구통계(성별, 출생연도, 사망 여부 등)
     - **admissions**: 입원 정보 (입원/퇴원, 날짜, 병동, 사망 여부)
     - **diagnoses_icd**: 진단 코드 (ICD-9. ICD-10)
     - **procedures_icd**: 수술 및 시술 코드
     - **labevents**: 실험실 검사 결과
     - **pharmacy/prescriptions**: 약물 정보
     - **microbiologyevents**: 미생물 배양 검사, 항생재 감수성 결과

2. **ICU Module (mimiciv_icu)**
   - ICU에서 수집된 고빈도 데이터
   - 주요 테이블:
     - **icustays**: ICU 체류 기록
     - **chartevents**: 환자 모니터링 이벤트 (vital signs, GCS, 투약량 등)
     - **inputevents/outputevents**: 수액, 약물 주입, 체액 배출량
     - **procedureevnets**: ICU 내 시술
     - **d_items**: 측정 항목 사전(itemid -> label)

3. **Note Module (mimiciv_note)**
   - 의무기록 텍스트 데이터
   - 진단 요약, 경과 기록, 간호 기록 등
   - 비식별화, 개인정보 제거됨

Data Structure



MIMIC-IV
(전체 데이터셋)

Hospital Module
(mimiciv_hosp)

- patients
- admissions
- diagnoses_icd
- procedures_icd
- labevents
- prescriptions
- microbiology

ICU Module
(mimiciv_icu)

- icustays
- chartevents
- inputevents
- outputevents
- procedureevents
- d_items

Note Module
(mimiciv_note)

- 의무기록 텍스트
- 진단 요약
- 경과 기록
- 간호 기록
- 비식별화 처리

Data Structure

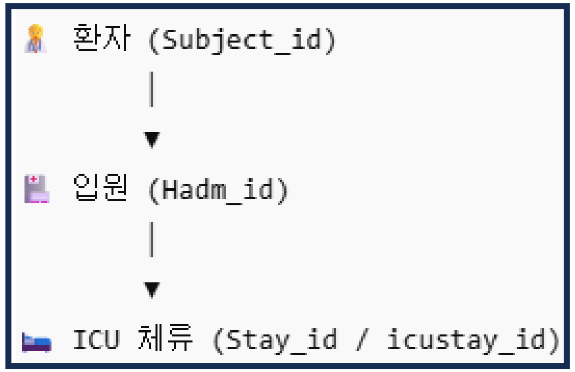
주요 키 (Identifiers)

데이터를 연결할 때 중요한 3단계 키 구조:

- **Subject_id**: 환자(익명화된 개인):
- **Hadm_id**: 한 번의 병원 입원 (admission)
- **Stay_id (icustay_id)**: ICU 체류

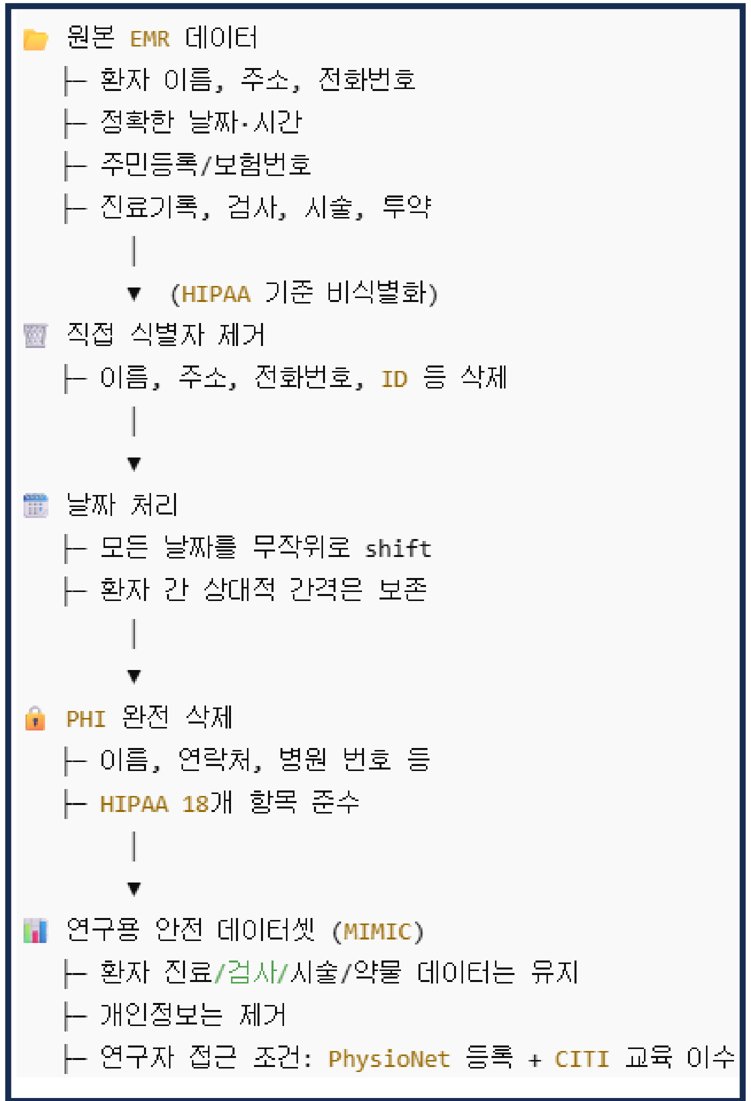
⬇

이 계층 구조 덕분에 환자 → 입원 → ICU stay 순서로 추적가능



```
🧑 환자 (Subject_id)
        │
        ▼
🏥 입원 (Hadm_id)
        │
        ▼
🛏 ICU 체류 (Stay_id / icustay_id)
```

개인정보 보호 (De-identification)



```
📁 원본 EMR 데이터
 ├─ 환자 이름, 주소, 전화번호
 ├─ 정확한 날짜·시간
 ├─ 주민등록/보험번호
 ├─ 진료기록, 검사, 시술, 투약
        │
        ▼  (HIPAA 기준 비식별화)
🗑 직접 식별자 제거
 ├─ 이름, 주소, 전화번호, ID 등 삭제
        │
        ▼
📅 날짜 처리
 ├─ 모든 날짜를 무작위로 shift
 ├─ 환자 간 상대적 간격은 보존
        │
        ▼
🔒 PHI 완전 삭제
 ├─ 이름, 연락처, 병원 번호 등
 ├─ HIPAA 18개 항목 준수
        │
        ▼
📊 연구용 안전 데이터셋 (MIMIC)
 ├─ 환자 진료/검사/시술/약물 데이터는 유지
 ├─ 개인정보는 제거
 ├─ 연구자 접근 조건: PhysioNet 등록 + CITI 교육 이수
```

Application

활용예시

1. **예측 모델링**
   - 첫 24시간 데이터로 병원 사망률 예측
   - 재입원 예측, 장기 예후 분석

2. **자연어 처리(NLP)**
   - 임상 노트 요약, Phenotyping, adverse event 탐지

3. **시계열 분석**
   - Vital sign 시계열로 상태 변동 분석
   - 치료 반응 패턴 학습

4. **실제 임상 연구**
   - 특정 치료 (예: 항생제, 수액요법) 효과 관찰
   - Real-world evidence (RWE) 생성

연구 시 주의사항

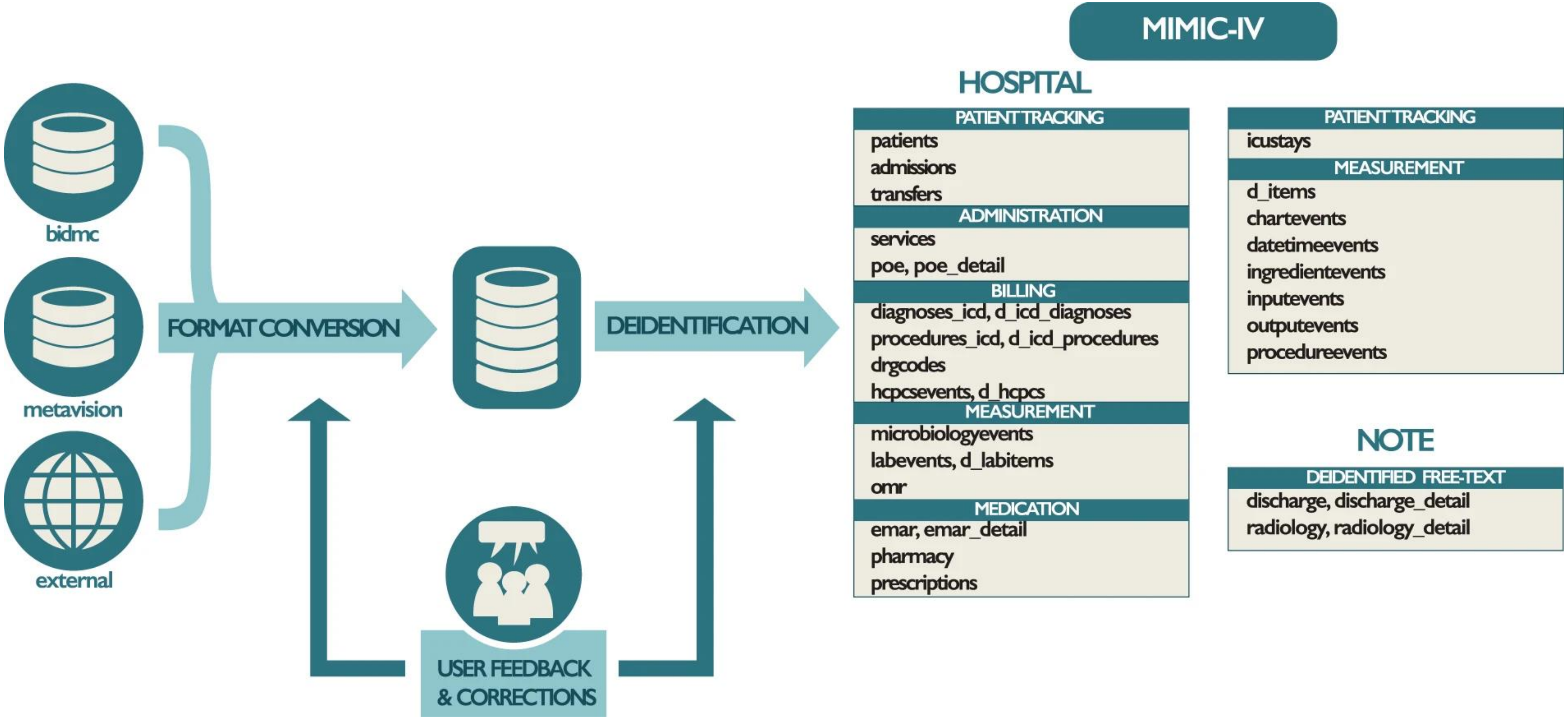
✓ **Index event 정의**
- 같은 환자 여러 입원/ICU stay가 있을 수 있으므로,
  연구에서는 보통 첫 입원이나 첫 ICU stay를 index로 삼음

✓ **결측치 / 이상치**
- 센서 오류, 기록 누락이 많아 전 처리 필수

✓ **대표성 한계**
- 단일 병원 데이터라 전체 환자를 대표하지 않을 수 있음

✓ **시계열 불규칙성**
- Irregular sampling → re-sampling, imputation 필요

Dataset collection process



bidmc

metavision

external

FORMAT CONVERSION

DEIDENTIFICATION

USER FEEDBACK & CORRECTIONS

MIMIC-IV

HOSPITAL

PATIENT TRACKING
- patients
- admissions
- transfers

ADMINISTRATION
- services
- poe, poe_detail

BILLING
- diagnoses_icd, d_icd_diagnoses
- procedures_icd, d_icd_procedures
- drgcodes
- hcpcsevents, d_hcpcs

MEASUREMENT
- microbiologyevents
- labevents, d_labitems
- omr

MEDICATION
- emar, emar_detail
- pharmacy
- prescriptions

PATIENT TRACKING
- icustays

MEASUREMENT
- d_items
- chartevents
- datetimeevents
- ingredientevents
- inputevents
- outputevents
- procedureevents

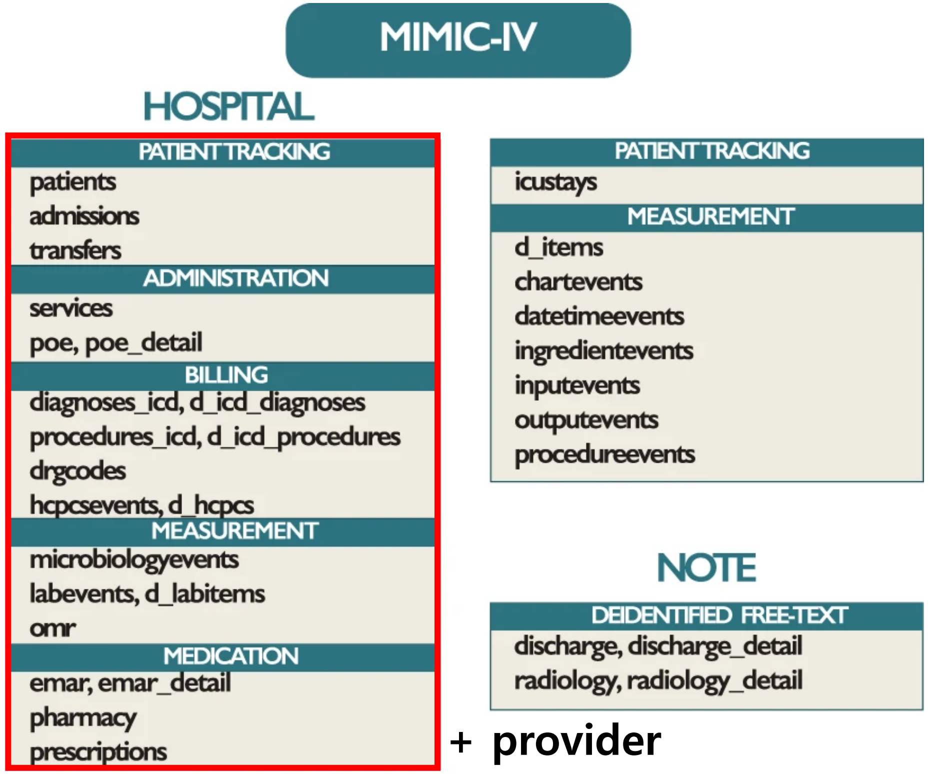
NOTE

DEIDENTIFIED FREE-TEXT
- discharge, discharge_detail
- radiology, radiology_detail

BIDMC: Beth Israel Deaconess Medical Center

Alistair Johnson et al. MIMIC-IV, a freely accessible electronic health records dataset. Scientific Data 2023.

Data description - *hosp* module



MIMIC-IV

HOSPITAL

PATIENT TRACKING
patients
admissions
transfers

ADMINISTRATION
services
poe, poe_detail

BILLING
diagnoses_icd, d_icd_diagnoses
procedures_icd, d_icd_procedures
drgcodes
hcpcsevents, d_hcpcs

MEASUREMENT
microbiologyevents
labevents, d_labitems
omr

MEDICATION
emar, emar_detail
pharmacy
prescriptions

PATIENT TRACKING
icustays

MEASUREMENT
d_items
chartevents
datetimeevents
ingredientevents
inputevents
outputevents
procedureevents

NOTE

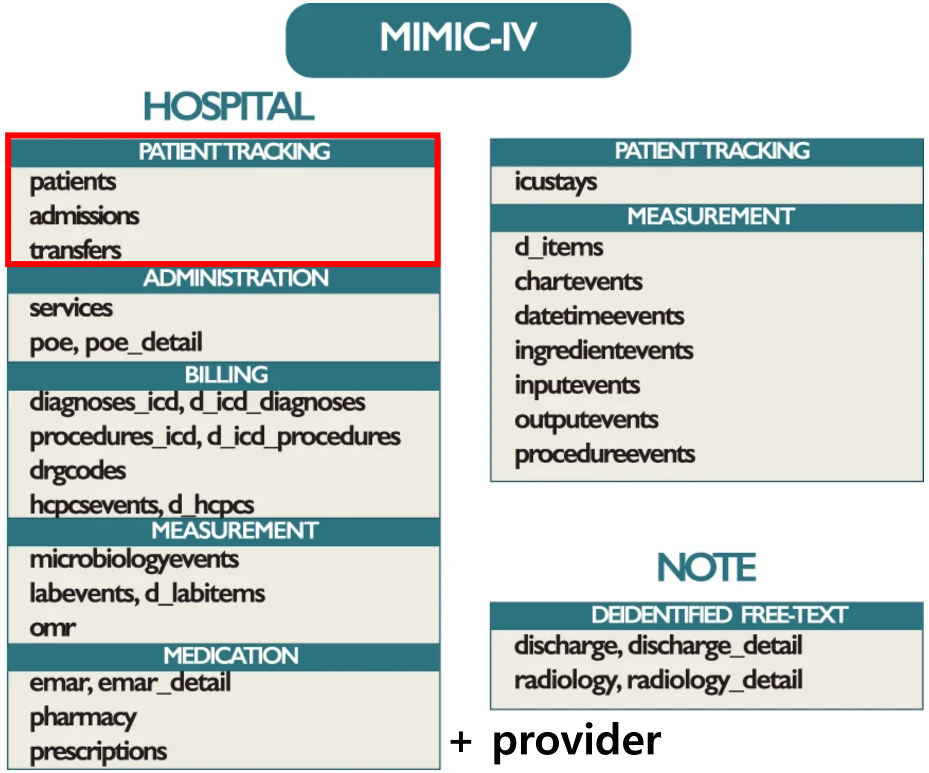
DEIDENTIFIED FREE-TEXT
discharge, discharge_detail
radiology, radiology_detail

+ **provider**

hosp

- **546,028** unique hospitalizations, **223,452** unique patients
- **Sources**: BIDMC EHR
- **Identifiers**
  - subject_id
  - hadm_id : a single hospitalization
    - NaN value – outside of an inpatient encouter
  - item_ids : linkage to d_{table} (description)

Data description - *hosp* module



MIMIC-IV

HOSPITAL

PATIENT TRACKING
patients
admissions
transfers

ADMINISTRATION
services
poe, poe_detail

BILLING
diagnoses_icd, d_icd_diagnoses
procedures_icd, d_icd_procedures
drgcodes
hcpcsevents, d_hcpcs

MEASUREMENT
microbiologyevents
labevents, d_labitems
omr

MEDICATION
emar, emar_detail
pharmacy
prescriptions

PATIENT TRACKING
icustays

MEASUREMENT
d_items
chartevents
datetimeevents
ingredientevents
inputevents
outputevents
procedureevents

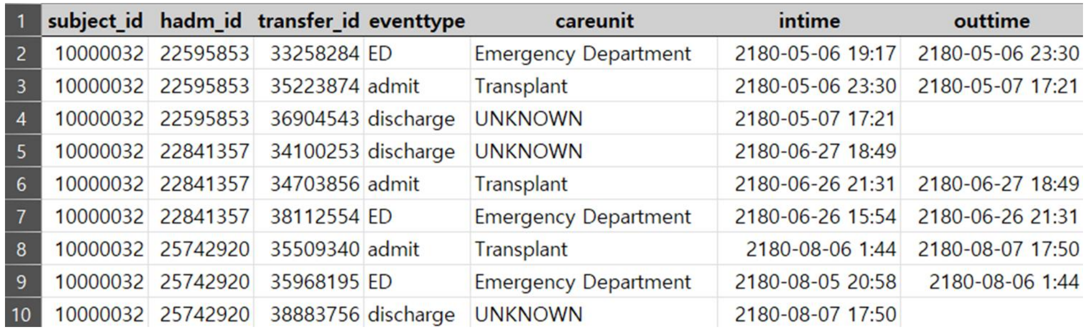
NOTE

DEIDENTIFIED FREE-TEXT
discharge, discharge_detail
radiology, radiology_detail

+ provider

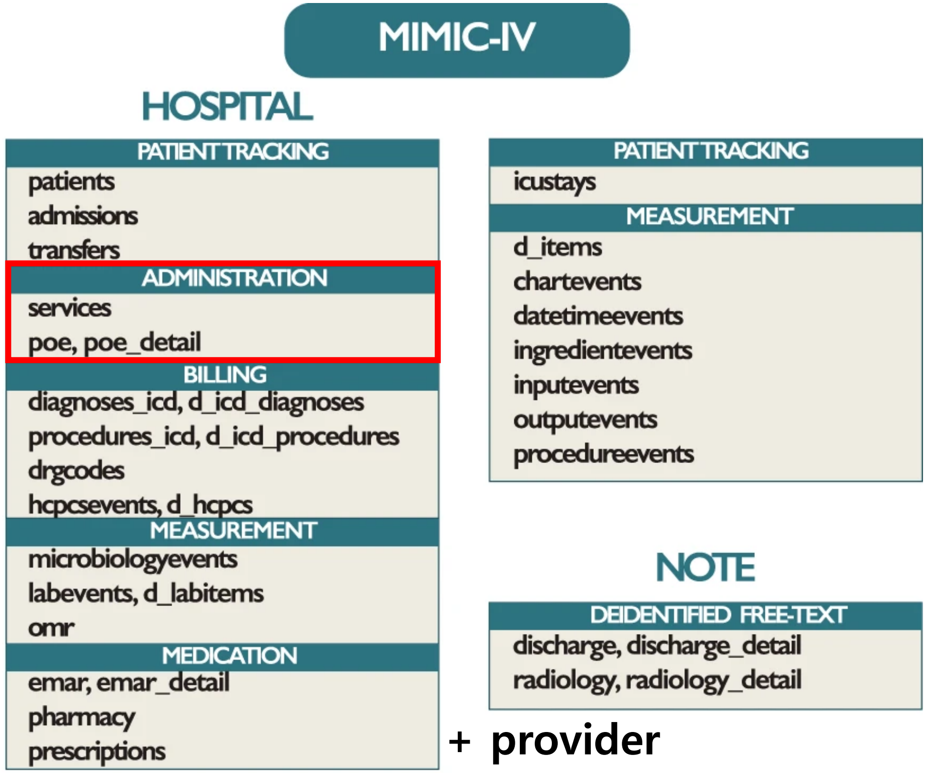
hosp

- Patient tracking
  - ***patients*** : patient demographics
    - Patient's administrative gender, age, date of death
  - ***admissions*** : hospitalizations
  - ***transfers*** : intra-hospital transfers



1	subject_id	hadm_id	transfer_id	eventtype	careunit	intime	outtime
2	10000032	22595853	33258284	ED	Emergency Department	2180-05-06 19:17	2180-05-06 23:30
3	10000032	22595853	35223874	admit	Transplant	2180-05-06 23:30	2180-05-07 17:21
4	10000032	22595853	36904543	discharge	UNKNOWN	2180-05-07 17:21	
5	10000032	22841357	34100253	discharge	UNKNOWN	2180-06-27 18:49	
6	10000032	22841357	34703856	admit	Transplant	2180-06-26 21:31	2180-06-27 18:49
7	10000032	22841357	38112554	ED	Emergency Department	2180-06-26 15:54	2180-06-26 21:31
8	10000032	25742920	35509340	admit	Transplant	2180-08-06 1:44	2180-08-07 17:50
9	10000032	25742920	35968195	ED	Emergency Department	2180-08-05 20:58	2180-08-06 1:44
10	10000032	25742920	38883756	discharge	UNKNOWN	2180-08-07 17:50	

Data description - *hosp* module



MIMIC-IV

HOSPITAL

PATIENT TRACKING
- patients
- admissions
- transfers

ADMINISTRATION
- services
- poe, poe_detail

BILLING
- diagnoses_icd, d_icd_diagnoses
- procedures_icd, d_icd_procedures
- drgcodes
- hcpcsevents, d_hcpcs

MEASUREMENT
- microbiologyevents
- labevents, d_labitems
- omr

MEDICATION
- emar, emar_detail
- pharmacy
- prescriptions

PATIENT TRACKING
- icustays

MEASUREMENT
- d_items
- chartevents
- datetimeevents
- ingredientevents
- inputevents
- outputevents
- procedureevents

NOTE

DEIDENTIFIED FREE-TEXT
- discharge, discharge_detail
- radiology, radiology_detail

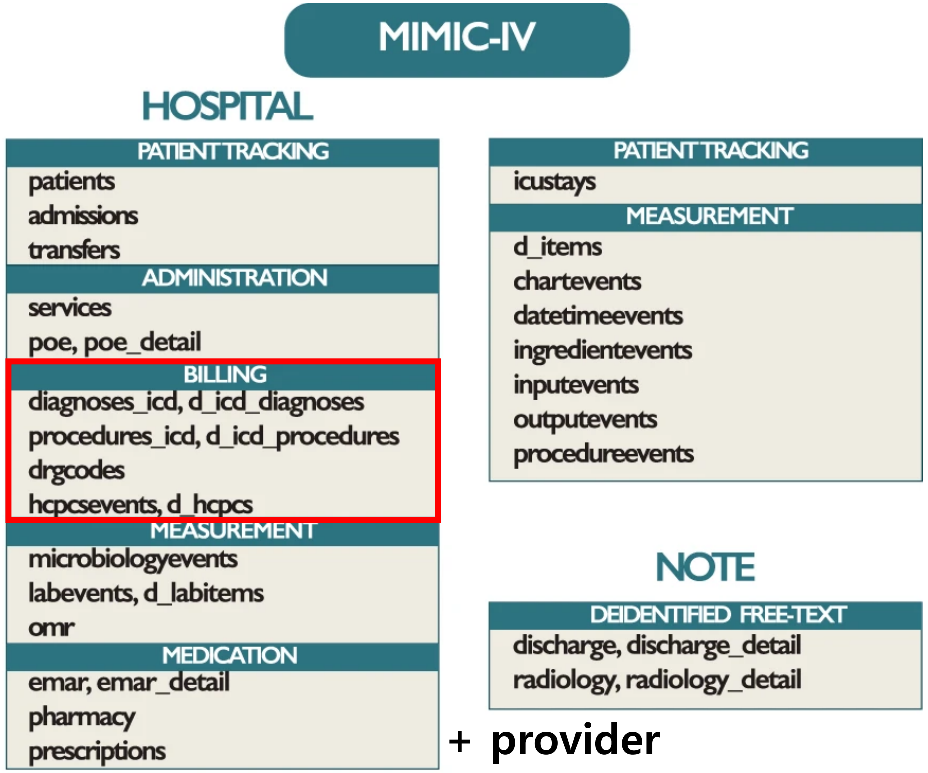
+ **provider**

hosp

- **Administration**
  - ***services*** : hospital-related services related information
  - ***poe, poe_detail*** : orders made in the provider order entry (POE) system *
    - Provide the date and time of an order

* POE system : used within the hospital to make orders related to diagnoses, imaging, consultation, treatment

Data description - *hosp* module



MIMIC-IV

HOSPITAL

PATIENT TRACKING
patients
admissions
transfers

ADMINISTRATION
services
poe, poe_detail

BILLING
diagnoses_icd, d_icd_diagnoses
procedures_icd, d_icd_procedures
drgcodes
hcpcsevents, d_hcpcs

MEASUREMENT
microbiologyevents
labevents, d_labitems
omr

MEDICATION
emar, emar_detail
pharmacy
prescriptions

PATIENT TRACKING
icustays

MEASUREMENT
d_items
chartevents
datetimeevents
ingredientevents
inputevents
outputevents
procedureevents

NOTE

DEIDENTIFIED FREE-TEXT
discharge, discharge_detail
radiology, radiology_detail

+ **provider**

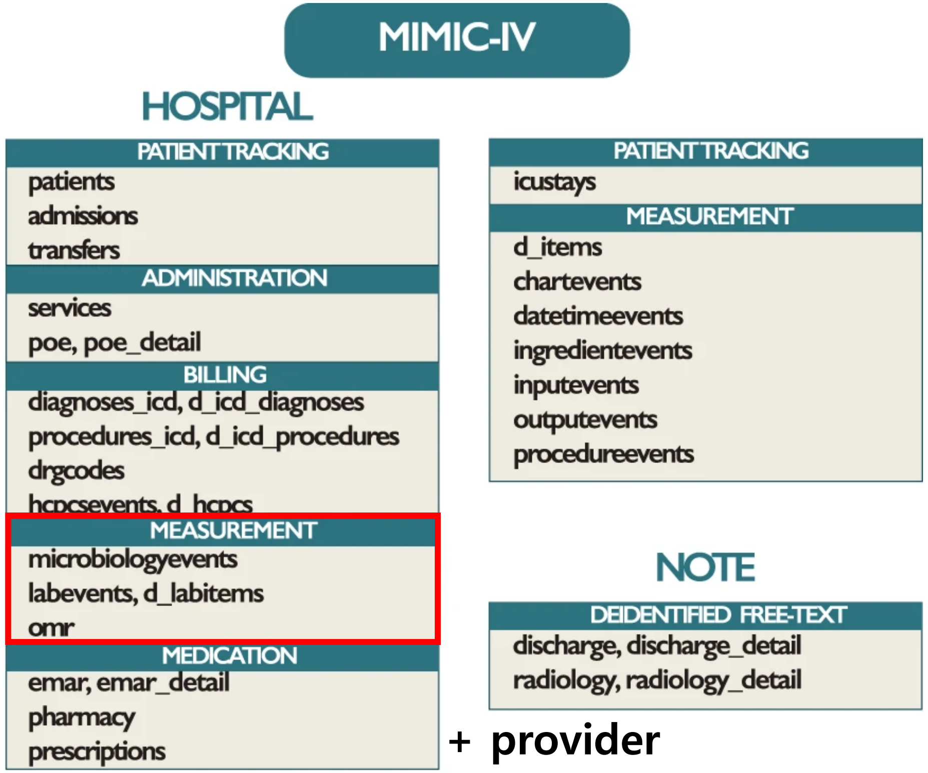
hosp

- **Billing**
  - ***diagnoses_icd*** : coded diagnoses representing the hospitalization
    - Ontology: ICD-9-CM, ICD-10-CM
    - *d_icd_diagnoses* : definitions for ICD codes
  - ***procedures_icd*** *:* coded procedures
    - Ontology: ICD-9-PCS, ICD-10-PCS
  - ***drgcodes*** : Diagnosis Related Groups codes *
  - ***hcpcevents*** : billing by the hospital for provided services (ex. mechanical ventilation)

* DRG: billable codes used to assign an overall cost to a hospitalization

Data description - *hosp* module

SNUH



MIMIC-IV

HOSPITAL

PATIENT TRACKING
patients
admissions
transfers

ADMINISTRATION
services
poe, poe_detail

BILLING
diagnoses_icd, d_icd_diagnoses
procedures_icd, d_icd_procedures
drgcodes
hcpcsevents, d_hcpcs

MEASUREMENT
microbiologyevents
labevents, d_labitems
omr

MEDICATION
emar, emar_detail
pharmacy
prescriptions

PATIENT TRACKING
icustays

MEASUREMENT
d_items
chartevents
datetimeevents
ingredientevents
inputevents
outputevents
procedureevents

NOTE

DEIDENTIFIED FREE-TEXT
discharge, discharge_detail
radiology, radiology_detail

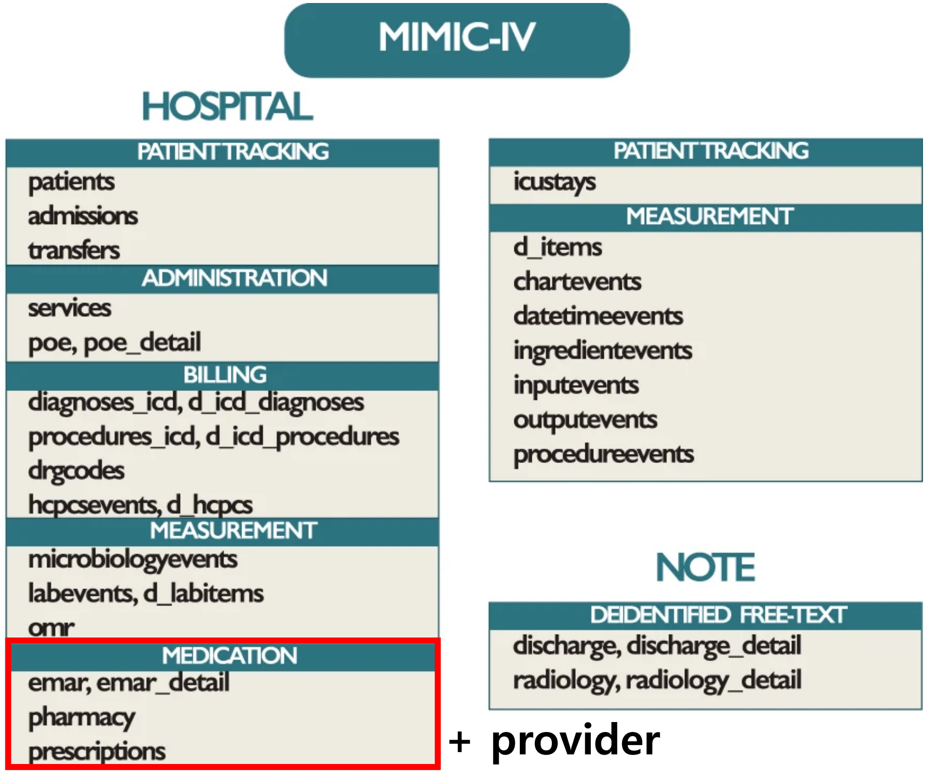
+ **provider**

hosp

- **Measurement**
  - ***microbiologyevents*** : microbiology measurements
  - ***labevents*** *:* laboratory measurements
    - *d_labitems*: definitions for concepts in *labevents*
  - ***omr*** : information from the Online Medical Record (OMR) *
    - Five measurements: blood pressure, height, weight, body mass index, eGFR
    - Both inpatient, outpatient visits
      - Including ‘baseline value’ before hospitalization

* OMR: a general system used for documenting patient information from visits at BIDMC affiliated institutes

Data description - *hosp* module



MIMIC-IV

HOSPITAL

PATIENT TRACKING
patients
admissions
transfers

ADMINISTRATION
services
poe, poe_detail

BILLING
diagnoses_icd, d_icd_diagnoses
procedures_icd, d_icd_procedures
drgcodes
hcpcsevents, d_hcpcs

MEASUREMENT
microbiologyevents
labevents, d_labitems
omr

MEDICATION
emar, emar_detail
pharmacy
prescriptions

PATIENT TRACKING
icustays

MEASUREMENT
d_items
chartevents
datetimeevents
ingredientevents
inputevents
outputevents
procedureevents

NOTE

DEIDENTIFIED FREE-TEXT
discharge, discharge_detail
radiology, radiology_detail

+ provider

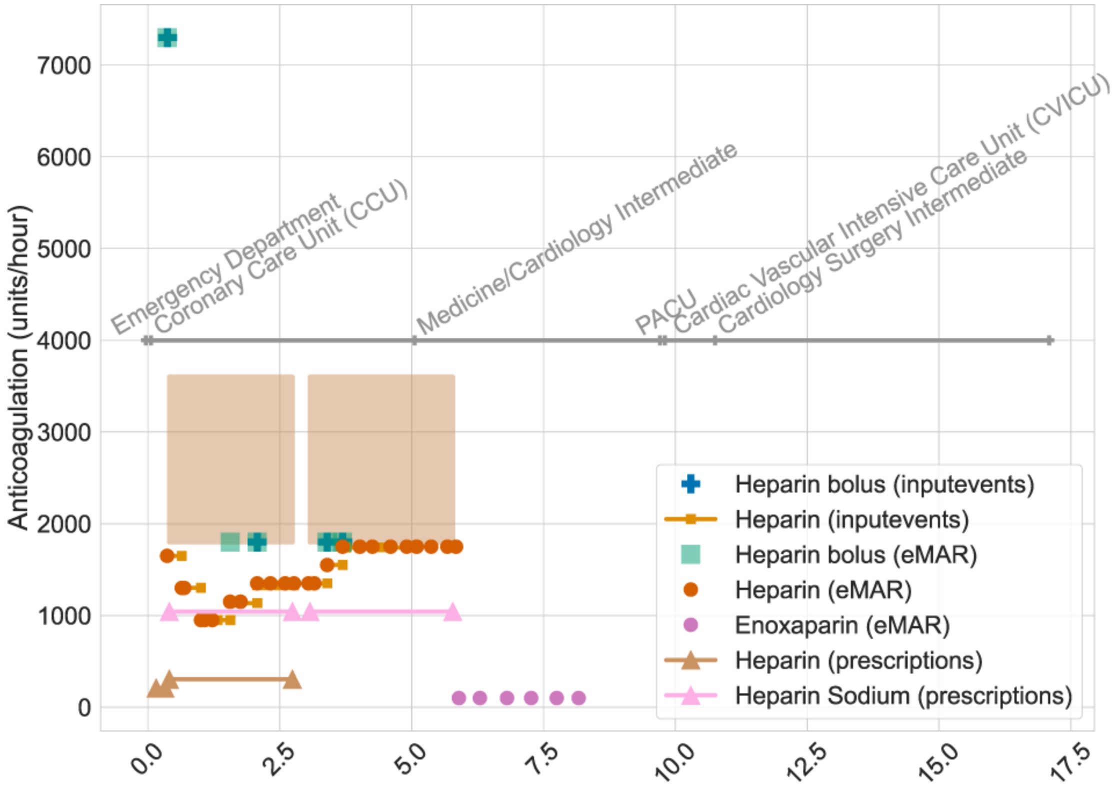
hosp

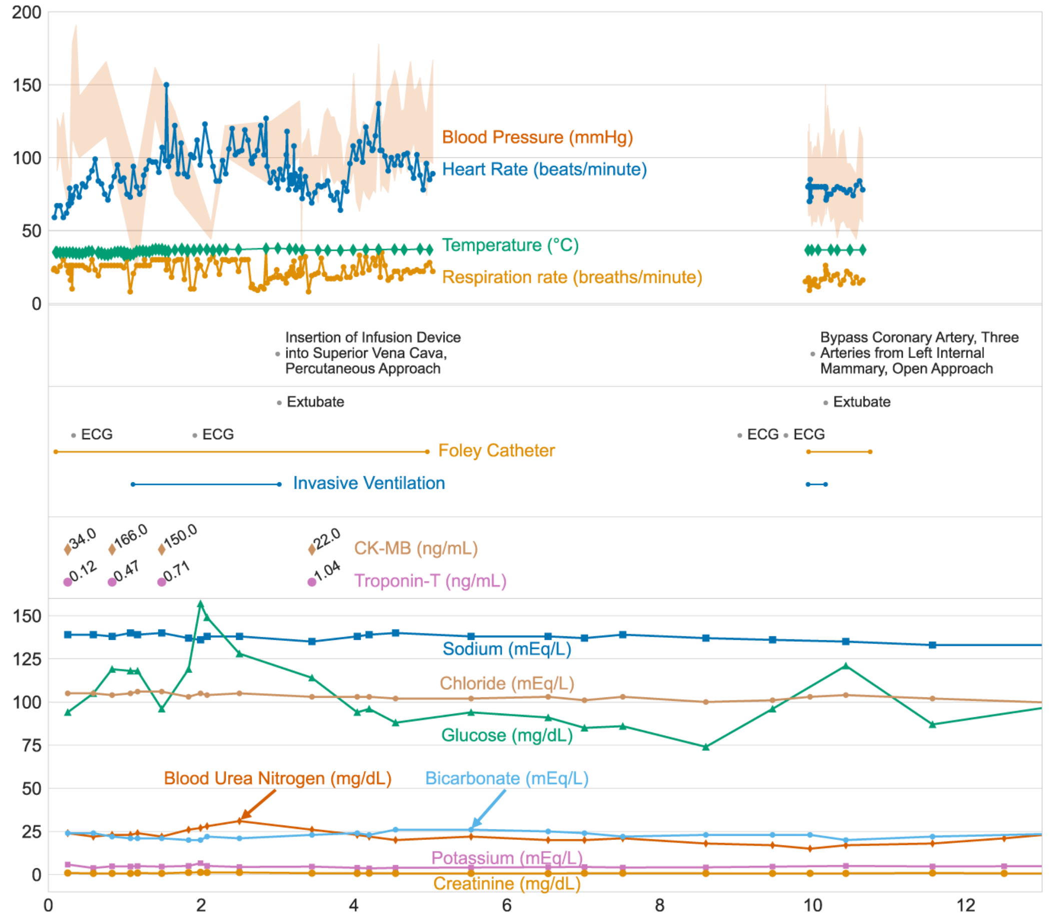
- **Medication**
  - ***prescriptions*** : ‘**order**’ made by a provider
  - ***pharmacy*** *:* detailed information regarding the compoud prescribed
  - ***emar*** : ‘**administration**’ records from the electronic Medicine Administration Record (eMAR)
    - 2014-2016 : first deployed => By 2016, all units
    - Link with *poe* (poe_id), *pharmacy* (pharmacy_id) table

Data description - *hosp* module

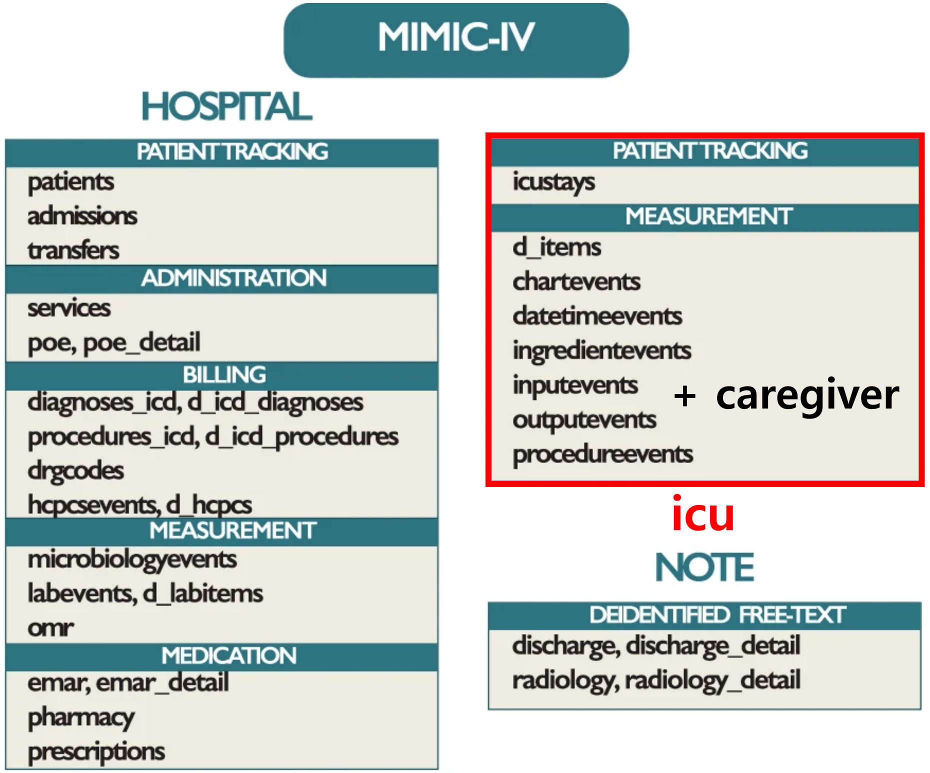
SNUH

A single patient's hospitalization (hadm_id 28503629)





Data description - *icu* module



MIMIC-IV

HOSPITAL

PATIENT TRACKING
- patients
- admissions
- transfers

ADMINISTRATION
- services
- poe, poe_detail

BILLING
- diagnoses_icd, d_icd_diagnoses
- procedures_icd, d_icd_procedures
- drgcodes
- hcpcsevents, d_hcpcs

MEASUREMENT
- microbiologyevents
- labevents, d_labitems
- omr

MEDICATION
- emar, emar_detail
- pharmacy
- prescriptions

icu

PATIENT TRACKING
- icustays

MEASUREMENT
- d_items
- chartevents
- datetimeevents
- ingredientevents
- inputevents
- outputevents
- procedureevents

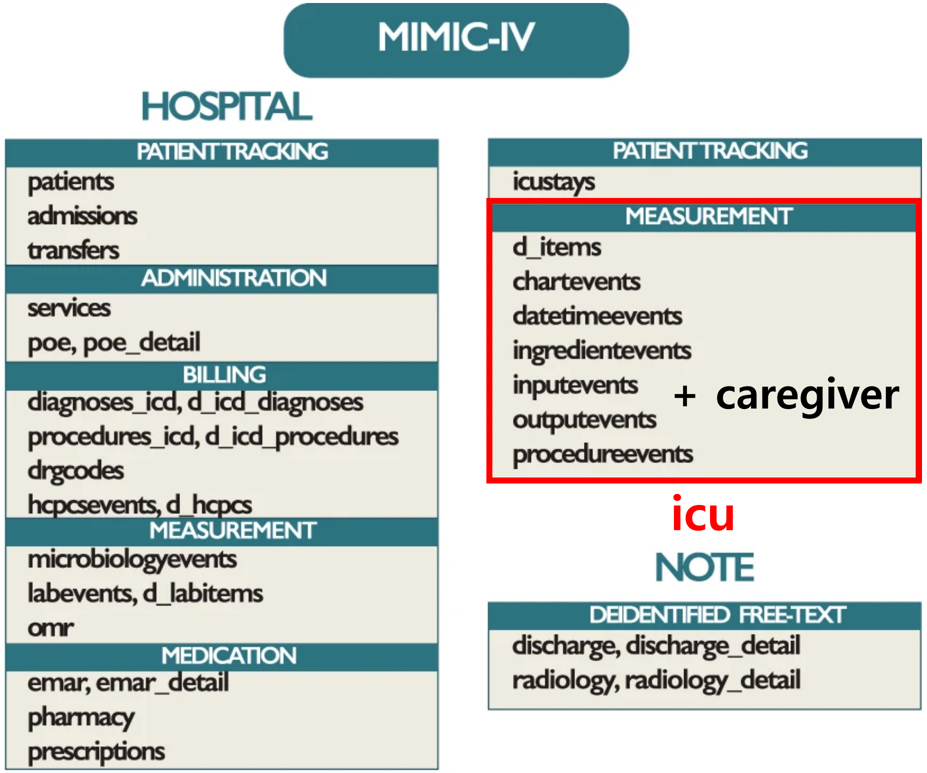
+ caregiver

NOTE

DEIDENTIFIED FREE-TEXT
- discharge, discharge_detail
- radiology, radiology_detail

- 94,458 ICU stays
- 65,366 unique patients
- **Sources**: MetaVision
- **Identifiers**
  - subject_id
  - stay_id : ICU stay
    - Consecutive transfers => a single stay_id
    - Transfer to a non-ICU ward between two ICU stays -> unique stay_id for each stay
  - itemid : identification of the concept in d_items
  - ***icustays*** : records of ICU stays
    - Derived from the ***transfers*** table in the *hosp* module

Data description - *icu* module



MIMIC-IV

HOSPITAL

PATIENT TRACKING
patients
admissions
transfers

ADMINISTRATION
services
poe, poe_detail

BILLING
diagnoses_icd, d_icd_diagnoses
procedures_icd, d_icd_procedures
drgcodes
hcpcsevents, d_hcpcs

MEASUREMENT
microbiologyevents
labevents, d_labitems
omr

MEDICATION
emar, emar_detail
pharmacy
prescriptions

PATIENT TRACKING
icustays

MEASUREMENT
d_items
chartevents
datetimeevents
ingredientevents
inputevents
outputevents
procedureevents

+ caregiver

icu

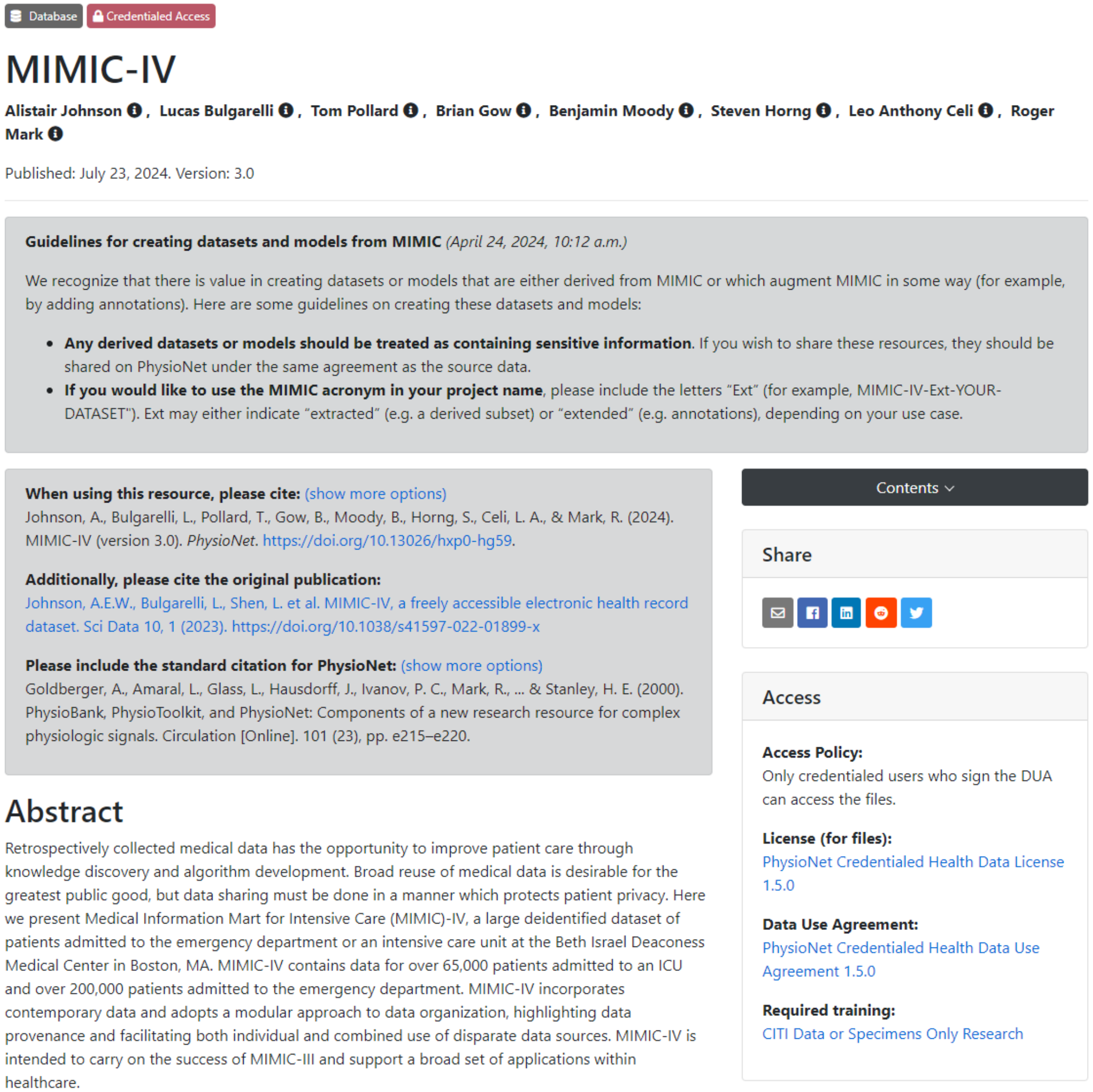
NOTE

DEIDENTIFIED FREE-TEXT
discharge, discharge_detail
radiology, radiology_detail

- **Measurement**
  - "events" based on the data type
  - ***inputevents*** : intravenous and fluid inputs
  - ***ingredientevents*** : ingredients for the inputs
  - ***outputevents*** : patient outputs
  - ***procedureevents*** : procedures including organ support treatments
  - ***datetimeevents*** : information documented as a date or time
  - ***chartevents*** : other charted information at the bedside

MIMIC-IV Usage note

- Available in PhysioNet



Database Credentialed Access

MIMIC-IV

Alistair Johnson , Lucas Bulgarelli , Tom Pollard , Brian Gow , Benjamin Moody , Steven Horng , Leo Anthony Celi , Roger Mark

Published: July 23, 2024. Version: 3.0

Guidelines for creating datasets and models from MIMIC *(April 24, 2024, 10:12 a.m.)*

We recognize that there is value in creating datasets or models that are either derived from MIMIC or which augment MIMIC in some way (for example, by adding annotations). Here are some guidelines on creating these datasets and models:

- **Any derived datasets or models should be treated as containing sensitive information**. If you wish to share these resources, they should be shared on PhysioNet under the same agreement as the source data.
- **If you would like to use the MIMIC acronym in your project name**, please include the letters "Ext" (for example, MIMIC-IV-Ext-YOUR-DATASET"). Ext may either indicate "extracted" (e.g. a derived subset) or "extended" (e.g. annotations), depending on your use case.

When using this resource, please cite: (show more options)
Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., & Mark, R. (2024). MIMIC-IV (version 3.0). *PhysioNet*. https://doi.org/10.13026/hxp0-hg59.

Additionally, please cite the original publication:
Johnson, A.E.W., Bulgarelli, L., Shen, L. et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 10, 1 (2023). https://doi.org/10.1038/s41597-022-01899-x

Please include the standard citation for PhysioNet: (show more options)
Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

Abstract

Retrospectively collected medical data has the opportunity to improve patient care through knowledge discovery and algorithm development. Broad reuse of medical data is desirable for the greatest public good, but data sharing must be done in a manner which protects patient privacy. Here we present Medical Information Mart for Intensive Care (MIMIC)-IV, a large deidentified dataset of patients admitted to the emergency department or an intensive care unit at the Beth Israel Deaconess Medical Center in Boston, MA. MIMIC-IV contains data for over 65,000 patients admitted to an ICU and over 200,000 patients admitted to the emergency department. MIMIC-IV incorporates contemporary data and adopts a modular approach to data organization, highlighting data provenance and facilitating both individual and combined use of disparate data sources. MIMIC-IV is intended to carry on the success of MIMIC-III and support a broad set of applications within healthcare.

Access

Access Policy:
Only credentialed users who sign the DUA can access the files.

License (for files):
PhysioNet Credentialed Health Data License 1.5.0

Data Use Agreement:
PhysioNet Credentialed Health Data Use Agreement 1.5.0

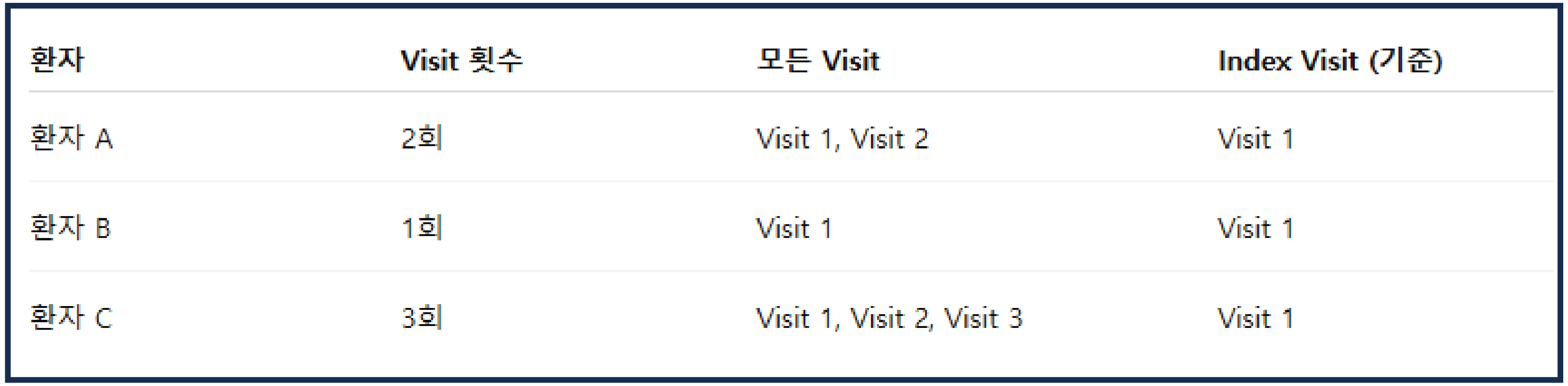
Required training:
CITI Data or Specimens Only Research

Tips for Data analysis

SNUH

MIMIC 같은 EMR/EHR 데이터에서 "visit"이 여러 번 일 수 있다는 개념, 그리고 index data의 필요성

1. Visit이 여러 번인 이유
   - 환자 단위(subject_id): 한 사람 (익명화된 환자)
   - 입원 단위(hadm_id): 한 번의 병원 입원 (Hospital Admission). 한 환자가 여러 번 입원할 수 있음
   - ICU 체류 단위(icustay_id, stay_id): 한 번의 ICU 체류 (ICU stay). 한 입원 안에서도 여러 번 ICU에 들어갔다 나올 수 있음



환자	Visit 횟수	모든 Visit	Index Visit (기준)
환자 A	2회	Visit 1, Visit 2	Visit 1
환자 B	1회	Visit 1	Visit 1
환자 C	3회	Visit 1, Visit 2, Visit 3	Visit 1

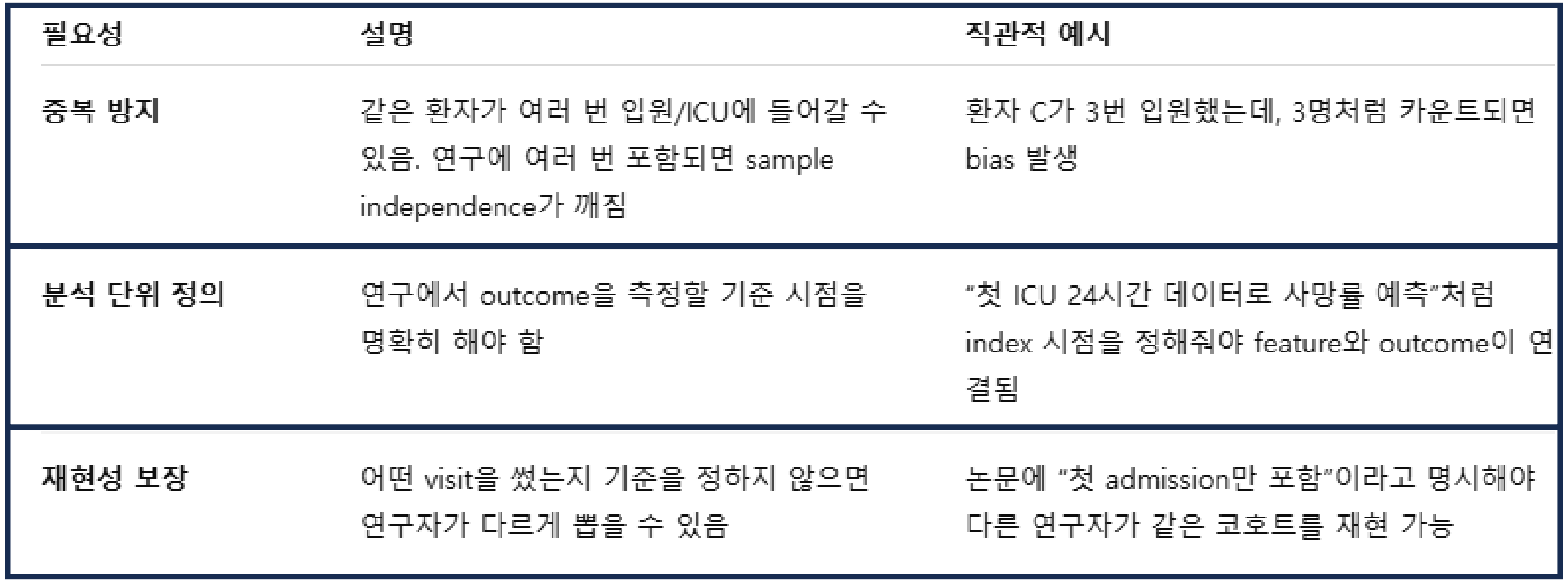
2. Index 데이터(Index admission/index event)란?
   - 연구를 설계할 때 "어떤 방문/사건을 기준(index)으로 할지"를 정해야 함
   - 예)
     - 병원 전체 입원(admission) 데이터가 있으면, 한 환자가 5번 입원했을 수 있음
       → **첫 입원**만 선택해서 **index admission**으로 정할 수 있음

   - ICU 연구라면, 여러 번 ICU에 들어간 사람 중 **첫 번째 ICU stay**를 index로 삼고, 나머지는 제외할 수 있음
   - 어떤 약물 효과를 보는 연구라면, 그 약물 **첫 처방일**을 index date로 삼을 수 있음

Tips for Data analysis

MIMIC 같은 EMR/HER 데이터에서 “visit”이 여러 번 일 수 있다는 개념, 그리고 index data의 필요성

3. 왜 index가 필요할까?



필요성	설명	직관적 예시
중복 방지	같은 환자가 여러 번 입원/ICU에 들어갈 수 있음. 연구에 여러 번 포함되면 sample independence가 깨짐	환자 C가 3번 입원했는데, 3명처럼 카운트되면 bias 발생
분석 단위 정의	연구에서 outcome을 측정할 기준 시점을 명확히 해야 함	“첫 ICU 24시간 데이터로 사망률 예측”처럼 index 시점을 정해줘야 feature와 outcome이 연결됨
재현성 보장	어떤 visit을 썼는지 기준을 정하지 않으면 연구자가 다르게 뽑을 수 있음	논문에 “첫 admission만 포함”이라고 명시해야 다른 연구자가 같은 코호트를 재현 가능

환자 단위로 중복되지 않고 cohort가 명확해짐

Q&A