

MIMIC-III Information_upload

MIMIC 데이터 기초 정보

- [MIMIC data 가지고 놀아보기](#)
- [\[논문 리뷰\] MIMIC-IV 데이터 구조 이해하기](#)
- [미국 중환자실 데이터 MIMIC-III 정리](#)
- [\[ML\] MIMIC -III 중환자실 빅데이터 DEMO](#)

퍼실님 제공 자료

-

추가 참고 자료

- Mimic-IV-ICD(미믹 4 하위 데이터셋)
 - <https://github.com/thomasnguyen92/MIMIC-IV-ICD-data-processing>
- 파이프라인
 - <https://github.com/healthylaife/MIMIC-IV-Data-Pipeline>
- 의학적 지식
 - **MIMIC-III (Medical Information Mart for Intensive Care)**
 - 내용 하단 부분 데이터 정보 + 레퍼런스 참고(논문)

★ MIMIC-III 스키마 전체 내용 파악

SchemaSpy - mimic.mimiciii

<https://mit-lcp.github.io/mimic-schema-spy/>

Table	Children	Parents	Columns	Rows	Comments
admissions	18	1	19	58,976	Hospital admissions associated with an ICU stay.
callout		2	24	34,499	Record of when patients were ready for discharge (called out), and the actual time of their discharge (or more generally, their outcome).
caregivers	7		4	7,567	List of caregivers associated with an ICU stay.
chartevents		5	15	330,712,483	Events occurring on a patient chart.
chartevents_1			15	38,033,561	Partition of chartevents. Should not be directly queried.
chartevents_10			15	9,584,888	Partition of chartevents. Should not be directly queried.
chartevents_11			15	470,141	Partition of chartevents. Should not be directly queried.
chartevents_12			15	265,413	Partition of chartevents. Should not be directly queried.
chartevents_13			15	39,066,570	Partition of chartevents. Should not be directly queried.
chartevents_14			15	100,075,138	Partition of chartevents. Should not be directly queried.
chartevents_2			15	13,116,197	Partition of chartevents. Should not be directly queried.
chartevents_3			15	38,657,533	Partition of chartevents. Should not be directly queried.
chartevents_4			15	9,374,587	Partition of chartevents. Should not be directly queried.
chartevents_5			15	18,201,026	Partition of chartevents. Should not be directly queried.
chartevents_6			15	28,014,688	Partition of chartevents. Should not be directly queried.
chartevents_7			15	255,967	Partition of chartevents. Should not be directly queried.
chartevents_8			15	34,322,082	Partition of chartevents. Should not be directly queried.
chartevents_9			15	1,274,692	Partition of chartevents. Should not be directly queried.
cpevents		2	12	573,146	Events recorded in Current Procedural Terminology.
d_cpt			9	134	High-level dictionary of the Current Procedural Terminology.
d_icd_diagnoses	1		4	14,710	Dictionary of the International Classification of Diseases, 9th Revision (Diagnoses).
d_icd_procedures	1		4	3,898	Dictionary of the International Classification of Diseases, 9th Revision (Procedures).
d_items	8		10	12,487	Dictionary of non-laboratory-related charted items.
d_labitems	1		6	753	Dictionary of laboratory-related items.
datetimeevents		5	14	4,485,937	Events relating to a datetime.
diagnoses_icd		3	5	651,047	Diagnoses relating to a hospital admission coded

- 사이트에 접속하여 테이블명 클릭 시 세부 내용 확인 가능

CSV 데이터셋 종류

ADMISSIONS.csv
CALLOUT.csv
CAREGIVERS.csv
CHARTEVENTS.csv
CPTEVENTS.csv
DATETIMEEVENTS.csv
DIAGNOSES_ICD.csv
DRGCODES.csv
D_CPT.csv
D_ICD_DIAGNOSES.csv
D_ICD_PROCEDURES.csv

D_ITEMS.csv
D_LABITEMS.csv
ICUSTAYS.csv
INPUTEVENTS_CV.csv
INPUTEVENTS_MV.csv
LABEVENTS.csv
LICENSE.txt
MICROBIOLOGYEVENTS.csv
NOTEVENTS.csv
OUTPUTEVENTS.csv
PATIENTS.csv

PRESCRIPTIONS.csv
PROCEDUREEVENTS_MV.csv
PROCEDURES_ICD.csv
README.md
SERVICES.csv
SHA256SUMS.txt
TRANSFERS.csv
Untitled.ipynb
checksum_md5_unzipped.txt
checksum_md5_zipped.txt

공통점

- ROW_ID 컬럼 : 유니크 ID 컬럼으로 보임
 - 컬럼 설명에서 이 컬럼은 제외하고 작성
- 주요 컬럼만 작성

💡 주요 테이블

1. ADMISSIONS.csv

- 환자 입원 정보
- 컬럼 개수 : 19개

Column Name	Data type	설명
HADM_ID	int	입원 ID
SUBJECT_ID	int	환자 ID
ADMITTIME	datetime	입원 시간
DISCHTIME	datetime	퇴원 시간
DEATHTIME	datetime(null 허용)	사망 시간
ADMISSION_TYPE	text	입원 유형(응급, 계획 등)
DIAGNOSIS	text	주요 진단 정보

2. PATIENTS.csv

- 환자 기본 정보(각 환자에 대한 인구통계학적 데이터, HIPPAA(미국 의료 정보 보호법)에 의해 특정 정보는 비식별화, 성별, 입원-퇴원 일시 등 정보 기록)
- 컬럼 개수 : 8개

Column Name	Data type	설명
SUBJECT_ID	int	환자 ID
GENDER	text	성별(M, F)
DOB	datetime	출생일
DOD	datetime(null 허용)	사망일
DOD_HOSP	datetime(null 허용)	병원 내 사망일
DOD_SSN	datetime(null 허용)	<ul style="list-style-type: none">• SSA(Social Security Administration) 기록에서 가져온 사망 일• 병원을 벗어난 후 사망한 환자에 대한 정보
EXPIRE_FLAG	int	사망 여부 플래그 컬럼(0=생존, 1=사망)

3. ICUSTAYS.csv

- 중환자실 체류 정보
- 컬럼 개수 : 12개

Column Name	Data type	설명
ICUSTAY_ID	int	ICU 체류 ID
HADM_ID	int	입원 ID
INTIME	datetime	ICU 입원 시간
OUTTIME	datetime	ICU 퇴원 시간
LOS	float	ICU 체류 기간(일 단위)

4. LABEVENTS.csv

- 실험실 검사 결과
- 컬럼 개수 : 9개

Column Name	Data type	설명
SUBJECT_ID	int	환자 ID
HADM_ID	int	입원 ID
ITEMID	int	검사 항목 ID
CHARTTIME	datetime	기록 시간
VALUE	text	검사 결과 (문자 형태 가능)
VALUENUM	float	검사 결과 (수치값)
VALUEUOM	text	검사 단위
FLAG	text	검사 결과의 정상 범주 여부 파악(normal: 범위 내, abnormal : 정상 범위를 벗어남)

5. CHARTEVENTS.csv

- 해당 테이블은 구글 드라이브에는 존재하지 않음(test.ipynb에 목록만 존재)
- 환자 관찰 기록
- 컬럼 개수 : 알 수 없음

Column Name	Data type	설명
SUBJECT_ID	int	환자 ID
ITEMID	int	검사 항목 ID
CHARTTIME	datetime	기록 시간
VALUE	text	관찰 값 (문자 형태 가능)
VALUENUM	float	관찰 값 (수치값)
VALUEUOM	text	단위

6. PRESCRIPTIONS.csv

- 약물 사전
- 컬럼 개수 : 19개

Column Name	Data type	설명
SUBJECT_ID	int	환자 ID
HADM_ID	int	입원 ID
STARTDATE	datetime	처방 시작 시간
ENDDATE	datetime	처방 종료 시간
DRUG	text	약물 이름
DRUG_NAME_POE	text	경구 약물 이름
DRUG_NAME_GENERIC	text	약물 일반 이름(주요 성분으로 기록, e.g. 타이레놀을 아세트아미노펜으로 작성하는 것)
FORMULARY_DRUG_CD	int	고유 약물 코드(처방 가능한 약물 목록)
GSN		약물 일반 코드(상표명 약물 대체가 가능한 제네릭 약물 식별 고유 코드)
NDC	int	<ul style="list-style-type: none"> • FDA에서 부여하는 약물 고유 식별 코드(National Drug Code) • 3가지 단위로 구성됨 <ul style="list-style-type: none"> ◦ Labeler code: 약물을 제조하거나 유통하는 회사의 식별 번호 ◦ Product code: 특정 약물과 그 약물의 강도, 제형, 투여 방법 등 ◦ Package code: 약물의 포장 크기와 유형 식별
PROD_STRENGTH	text	처방 약물 용량(단, 활성 성분 양 기준)
DOSE_VAL_RX	float	복용량
DOSE_UNIT_RX	text	복용량 단위
FORM_VAL_DISP	text	약물의 용량 혹은 투여 날짜 등을 작성
FORM_UNIT_DISP	text	처방 약물의 강도 및 제형(CAP: 캡슐, TAB: 정제, BAG: 주사제 혹은 액체 약물, ml: 액체 약물 양, dose: 특정 용량, PKT: 약물 포장 단위(?), BTL: 병 형태 약물, VIAL: 액체 혹은 분말 형태의 약물을 담은 병, SYR: 주사 형태, SUPP: 좌약...)
ROUTE	text	<ul style="list-style-type: none"> • 약물 투여 방식(약어로 작성되어 있음) <ul style="list-style-type: none"> ◦ Oral: 경구 투여 ◦ Intravenous (IV): 정맥 주사나 IV 투여 ◦ Intramuscular (IM): 근육 주사 ◦ Subcutaneous (SC): 피하 주사 ◦ Topical: 피부에 국소 적용 (크림, 연고 등) ◦ Inhalation: 흡입 (네불라이저 등)

7. D_ICD_DIAGNOSES.csv

- 진단 관련 질병 및 관련 건강 문제 ICD-9 기반 코드 사전
- 컬럼 개수 : 4개

Column Name	Data type	설명
ICD9_CODE	int	특정 질병 또는 진단을 나타내는 코드
SHORT_TITLE	text	ICD9-CODE와 연결된 간략 설명
LONG_TITLE	text	ICD9-CODE와 연결된 상세 진단명

8. NOTEEVENTS.csv

- 의료진 메모 및 진단 노트 기록
- 컬럼 개수 : 11개

Column Name	Data type	설명
SUBJECT_ID	int	환자 ID
HADM_ID	int	입원 ID
CHARTDATE	datetime	차팅한 날짜
CHARTTIME	datetime	차팅 날짜 + 시간까지 기록
STORETIME	datetime	EMR(전자 의료 기록) 시스템에 저장한 시간
CATEGORY	text(or varchar)	<ul style="list-style-type: none">• 임상 노트(Clinical Note)의 유형 기록<ul style="list-style-type: none">◦ 환자 건강 상태, 치료 과정, 진단 결과, 처방 내용 등 기술◦ EMR의 중요 요소◦ 환자 관리 및 치료 핵심 데이터
DESCRIPTION	text(or varchar)	임상 노트 세부 내용
CGID	int	치료 담당자가 누군지에 대한 고유 ID
ISERROR	int	임상노트 오류 여부 플래그(0: 오류없음, 0이 아닌 값: 오류 발생)
TEXT	text	의료진 메모 및 진단 노트 전문

▼ 1. CATEGORY 데이터 의미 요약(종류는 정확하지 않음, 이 테이블에 이 종류가 포함된다는 정도)

종류	설명
Discharge Summary(퇴원 요약)	퇴원 시 작성되는 요약 보고서 (환자의 전체 입원 요약)
Radiology(방사선)	방사선학 결과 노트 (X-ray, CT 등 영상 판독 결과)
Nursing/other(간호기록)	간호사나 기타 의료진이 작성한 노트
ECG	심전도(Electrocardiogram) 결과 노트
Physician(진료 노트)	의사가 작성한 일반 메모
Respiratory	호흡 치료와 관련된 노트
Nutrition	환자의 영양 상태나 식이 요법에 대한 노트
General	기타 일반 노트

▼ 2. CATEGORY & DESCRIPTION 관계

- Discharge summary
 - Report(환자의 퇴원 요약 기록)
- Radiology
 - CT Abdomen W Contrast (복부 CT)
 - Portable Chest X-ray (휴대용 흉부 엑스레이)
- Nursing/other
 - Nursing Progress Note (간호 진행 상황 기록)
 - Respiratory Care Shift Note (호흡 치료 관련 노트)
 - ICU Note - CVI(중환자실 기록, 심혈관계 ICU동)
- Physician
 - Addendum (임상 노트 수정(추가) 사항 기록)
 - Physician Resident Progress Note (의사가 기록한 진단 혹은 환자 상태)

9. CHARTEVENTS.csv

- 해당 테이블은 구글 드라이브에는 존재하지 않음(test.ipynb에 목록만 존재)
- 환자 생체 신호 데이터(심박수, 혈압, 호흡 빈도 등) == ICU 환자 차트 데이터
- 컬럼 개수 : 15개

Column Name	Data type	설명
SUBJECT_ID	int	환자 ID
HADM_ID	int	입원 ID
ICUSTAY_ID	int	ICU 체류 ID
ITEMID	int	검사 항목 ID
CHARTTIME	datetime	차팅 날짜 + 시간까지 기록
CGID	int	치료 담당자가 누군지에 대한 고유 ID
VALUE	float or double precision	CHARTTIME(컬럼)에서 기록된 측정 값
VALUENUM	float or double precision	value 컬럼에서 숫자만 저장한 것
VALUEUOM	varchar	값 단위
WARNING	text	경고 또는 이슈 발생 내용 기록

Column Name	Data type	설명
ERROR	text	에러 내용 기록
RESULTSTATUS	varchar	최종 결과 상태(경고, 에러, 중단, 성공 등 모두)
STOPPED	varchar	측정 값의 중단 여부 표기(커넥션 리퓨스같은 거)

10. DIAGNOSES_ICD.csv

- ICD-9 진단 코드 데이터 테이블(환자 진단 정보 기록)
- 컬럼 개수 : 5개

Column Name	Data type	설명
SUBJECT_ID	int	환자 ID
HADM_ID	int	입원 ID
SEQ_NUM	int	진단 순서 번호 부여
ICD9_CODE	varchar	환자 고유 ICD-9 코드(특정 질병 또는 상태를 의미)

*IDC10_CODE가 있기도 함(ICD-9 대체 코드 체계)

11. INPUTEVENTS_CV.csv

- 환자의 입원 후 투약한 수액 종류 및 투여량 확인
- 컬럼 개수 : 22개

Column Name	Data type	설명
SUBJECT_ID	int	환자 ID
HADM_ID	int	입원 ID
ICUSTAY_ID	int	ICU 체류 ID
CHARTTIME	datetime	차팅 날짜 + 시간까지 기록
ITEMID	int	검사 항목 ID
AMOUNT		
AMOUNTUOM		
RATE		
RATEUOM		
STORETIME	datetime	EMR(전자 의료 기록) 시스템에 저장한 시간
CGID	int	치료 담당자가 누군지에 대한 고유 ID
ORDERID	int	
LINKORDERID		
STOPPED	varchar	측정 값의 중단 여부 표기(커넥션 리퓨스같은 거)
ORIGINALAMOUNT		
ORIGINALAMOUNTUOM	varchar	ORIGINALAMOUNT 단위
ORIGINALRATE		
ORIGINALRATEUOM	varchar	ORIGINALRATE 단위
ORIGINALSITE		

MIMIC III vs MIMIC IV



출처

- <https://physionet.org/content/mimiciv/0.4/>
- <https://ieeexplore.ieee.org/document/10386585>
- <https://alistairewj.github.io/talk/2020-mimic-iv-data-tutorial/>
- https://lcp.mit.edu/news_2

1. 카테고리 분리

- III : 모든 데이터가 하나의 테이블에 들어가 있는 형태
- IV : `core` , `hosp` , `icu` 로 분리

2. 새로운 데이터 추가

- chest x-ray images(흉부 x선 이미지)
- electronic medicine administration record(전자 약물 투여 기록)
- *microbiologyevents()*
 - 컬럼 : `spec_type_desc` , `test_name` , `org_name` , `ab_name`
 - 생물체, 항생제, 검사, 검체 text name 포함
 - 검사 정보 및 테스트 결과 일부 포함(ex. viral load tests)

3. 날짜 시프트

- III : 연도와 날짜가 복합적으로 구성
- IV : 연도만 기록

4. 테이블 제거

- d_micro → 테이블 삭제됨

MIMIC-IV로 만들어낼만한 주제(참고, 3으로도 가능)

- **최다 처방약물 10가지 영양소 및 효능 분석**
- 출처 : MIMIC-IV 데이터를 활용한 약물 분석
 - 학교 수업으로 진행한 내용 → velog 시리즈 살펴보면 진행 과정이 모두 있어서 참고하기 좋음
- **MIMIC-IV데이터셋을 이용한 급성신손상 예측 모델 개발**
 - 출처 : 분당 서울대학교병원 급성신손상 환자 데이터톤
 - 해당 대회 예선으로 나온 주제
 - 레퍼런스는 조금 더 찾아봐야하는 문제가 있음
- **패혈증 예측**
 - 퍼실님께서 주신 레퍼런스 존재
 -

아래 3가지는 논문 기반, 실제 구현까지는 시간이 다소 소요될 수 있음.

- ICU 병상 효율성 최적화
 - 출처 : Evaluating the Fairness of the MIMIC-IV Dataset and a Baseline Algorithm: Application to the ICU Length of Stay Prediction
 - LOS 예측(Length of Stay, 입원 환자의 병상 이용 기간)
 - XGBoost 모델 사용
- ICU 환자의 재입원 예측
 - 출처 : https://pods4h.com/wp-content/uploads/2021/10/ICPM2021_PODS4H_paper_ext_abst_174.pdf
- 실시간 사망 예측
 - 출처 : Real-time Mortality Prediction Using MIMIC-IV ICU Data Via Boosted Nonparametric Hazards
 - 사망 위험의 실시간 예측을 위해 BoXHED 모델
 - 시간 변화에 다른 Cox 모델
 - AUC-PRC로 성능 평가
- 참고) MIMIC IV 튜토리얼 노트북
<https://colab.research.google.com/drive/1REu-ofzNzqsTT1cxLHlegPB0nGmwKaM0?usp=sharing>

MIMIC III - 기타 레퍼런스


데이터 인포메이션

- 의학적 소개 + 데이터셋 정보 요약 잘되어 있음
- <https://blueorbit.tistory.com/398>

코드 참고할만한 저장소

- 1. ICU 사망률 예측
 - 사용할 테이블 : ADMISSIONS, ICUSTAYS, PATIENTS, LABEVENTS
- 2. LOS(체류기간) 예측 관련 → CHARTEVENTS 테이블 있어야 가능
 - 사용할 테이블 : ADMISSIONS, ICUSTAYS, (CHARTEVENTS), DIAGNOSES_ICD
 - 레퍼런스
 - https://github.com/MLforHealth/MIMIC_Extract
 - 코드 사용시 인용문 첨부해야하는 이슈 있음
 - **LOS 예측 관련 노트북**

MIMIC_Extract/notebooks/Baselines for Mortality and LOS prediction - Sklearn.ipynb at master · MLforHealth/MIMIC_Extract
MIMIC-Extract:A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III - MLforHealth/MIMIC_Extract

 https://github.com/MLforHealth/MIMIC_Extract/blob/master/notebooks/Baselines%20for%20Mortality%20and%20LOS%20prediction%20-%20Sklearn.ipynb

MLforHealth/
MIMIC_Extract

MIMIC-Extract:A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III

6

29

415

122

- 3. 재입원 예측
 - 사용할 테이블 : ADMISSIONS, ICUSTAYS, DIAGNOSES_ICD, PATIENTS
 - https://github.com/YaronBlinder/MIMIC-III_readmission
 - 노트북, 데이터 가공 SQL문 모두 있으며, 프로젝트 최종 PDF도 있음

https://github.com/YaronBlinder/MIMIC-III_readmission/blob/master/Report.pdf

- 4. 약물 처방 패턴 분석
 - 사용할 테이블 : PRESCRIPTIONS, ADMISSIONS, INPUTEVENTS_MV + INPUTEVENTS_CV
 - 환자 그룹별 약물 처방 패턴을 분석을 통해 → 특정 질병에 대한 약물 효과를 파악하는 정도로 진행
 - K-means나 DBSCAN으로 클러스터링 가능할 것 같음
 - 시각화도 간단히 진행하기 편할듯

- 5. 부가) COVID-19 분석 → CHARTEVENTS 테이블이 있어야 가능
 - 사용할 테이블 : ADMISSIONS, LABEVENTS, (CHARTEVENTS)
 - 시일이 많이 지나기는 했으나, 전염병으로 인해 의료 체계가 많이 바뀐만큼 여전히 중요한 지표일 것으로 판단

- K-MIMIC
 - <https://sites.google.com/view/k-mimic>
- 벤치마킹 모델
 - <https://github.com/YerevaNN/mimic3-benchmarks>
 - 조기 입원 데이터에서 사망률 예측(분류)
 - 실시간 보상 장애 감지(시계열 분류)
 - 입원 기간 예측(회귀)
 - 표현형 분류(다중 라벨 시퀀스 분류)
 - 레포지터리 사용 방법 : <https://kagus2.tistory.com/44>
- 큰 데이터 위주로 정보 서치
 - 자료 거의 ..없음..
- 기타 참고
 - DL : <https://github.com/DanielSola/mimic-iii-project>
 - NLP : ~~논문리뷰~~ | **[NLP ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission \(2020\) Summary](#)**