Institute for Clinical and Translational Research
AT EINSTEIN AND MONTEFIORE
Building Bridges in the Bronx and Beyond

Data Science 101: Electronic Health Record (EHR) Data

Pavel Goriacko, PharmD, MPH

Co-Director, Health Research Informatics

EINSTEIN
Albert Einstein College of Medicine

Montefiore

Today's Presentation

- Logistics of Working with EHR Data
- Challenges Inherent in EHR Data
- Methodological Considerations
- Building EHR Research Team

What Are EHR Data?

Institute for Clinical and Translational Research
AT EINSTEIN AND MONTEFIORE
Building Bridges in the Bronx and Beyond

Digital record of care delivery generated during clinical encounters

Demographics	Encounter Information	Diagnoses & Problems	Procedures & Orders
Medications	Laboratory, Test, & Vital Signs Measurements	Clinical Narratives and Reports	Special Tests and Reports (ECG, EEG, genomics)

Why EHR Data Matter

Nearly every patient encounter generates digital data

Huge scale, real-world context, inexpensive and quick to obtain

Enables questions impractical or unethical for trials

Case Study: Tocilizumab in COVID-19

Original Investigation

FREE

October 20, 2020

Association Between Early Treatment With Tocilizumab and Mortality Among Critically Ill Patients With COVID-19

Shruti Gupta, MD, MPH[1]; Wei Wang, PhD[2]; Salim S. Hayek, MD[3]; et al

» Author Affiliations | Article Information

JAMA Intern Med. 2021;181(1):41-51. doi:10.1001/jamainternmed.2020.6252

Mortality OR 0.68 (0.46-0.99)

Study Design and Oversight

We emulated a hypothetical target trial in which critically ill adults with COVID-19 received or did not receive tocilizumab in the first 2 days of intensive care unit (ICU) admission. We used data from the Study of the Treatment and Outcomes in Critically Ill Patients With COVID-19 (STOP-COVID), a multicenter cohort study that enrolled consecutive adults with laboratory-confirmed COVID-19 (detected by nasopharyngeal or oropharyngeal swab) admitted to participating ICUs at 68 hospitals across the United States (eTable 1 in Supplement 1). Study personnel at each site collected data by detailed medical

Case Study: Tocilizumab in COVID-19

ORIGINAL ARTICLE

Interleukin-6 Receptor Antagonists in Critically Ill Patients with Covid-19

Author: The REMAP-CAP Investigators* Author Info & Affiliations

Published February 25, 2021 | N Engl J Med 2021;384:1491-1502 | DOI: 10.1056/NEJMoa2100433

VOL. 384 NO. 16 |

Mortality OR: 0.65 (0.43-0.99)

Case Study: Tocilizumab in COVID-19

Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial

RECOVERY Collaborative Group †

Affiliations & Notes   Article Info   Linked Articles (7)

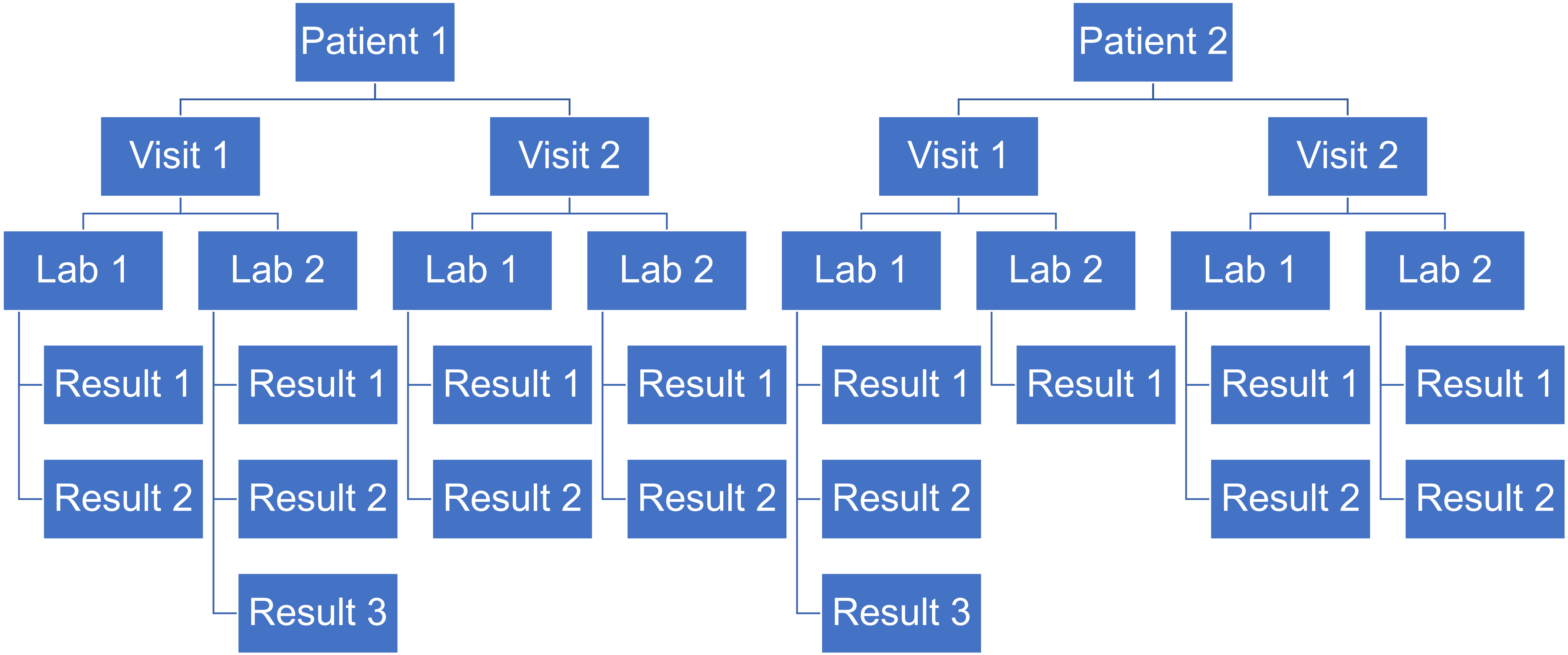
Publication History: Published May 1, 2021

DOI: 10.1016/S0140-6736(21)00676-0   Also available on ScienceDirect

Mortality OR: 0.76 (0.66-0.88)

Working with EHR Data

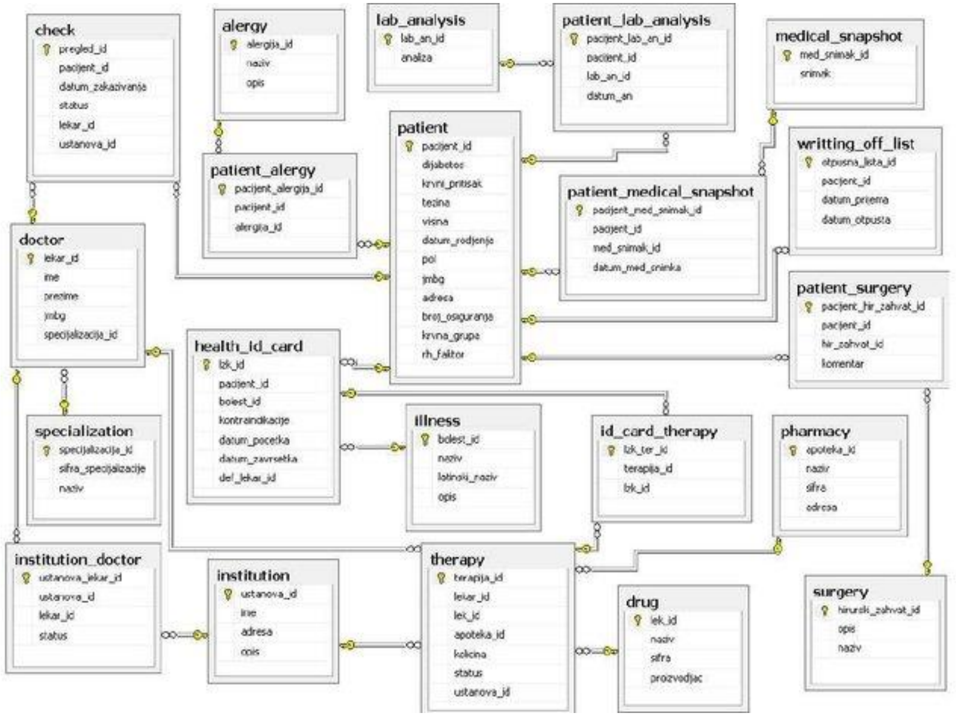
How Are EHR Data Stored?



EHR Data Structure

Institute for Clinical and Translational Research
AT EINSTEIN AND MONTEFIORE
Building Bridges in the Bronx and Beyond

- Stored in relational tables (patients, visits, meds, labs, notes…)
- Requires joins, cleaning, reshaping
    - ▻ SQL and data wrangling skills
- Different vendors & data models
    - ▻ Epic alone has 3 data models!



M. Vucetic, A. Uzelac and N. Gligoric, "E-Health Transformation Model in Serbia: Design, Architecture and Developing," *2011 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, Beijing, China, 2011, pp. 566-573, doi: 10.1109/CyberC.2011.96.

Epic Clarity Data Dictionary

Institute for Clinical and Translational Research AT EINSTEIN AND MONTEFIORE
Building Bridges in the Bronx and Beyond

A B C D E F G H I K L M N O P Q R S T U V W

Tables Starting with A

Top of page

- ABN_DOCUMENT_ID
- ABN_FOLLOW_UP
- ABN_MEDICATIONS
- ABN_NOTES
- ABN_NOTE_COMMENTS
- ABN_NOTE_CONTACT_SERVICE
- ABN_NOTE_PROC
- ABN_ORDERS
- ABN_ORDER_INFO
- ABN_PLAN_INFO
- ABN_SERVICE_INFO
- ACCCATH3_MED_DISCH
- ACCESSIBLE_DOCUMENTS_PREF
- ACCESSIBLE_SERVICES
- ACCOUNT
- ACCOUNT_2
- ACCOUNT_3
- ACCOUNT_CONS_SP_SA_BILL
- ACCOUNT_CONTACT
- ACCOUNT_CONTACT_2
- ACCOUNT_CREATION
- ACCOUNT_DISCON_CVG
- ACCOUNT_FPL_INFO
- ACCOUNT_MYC_GP_FAILED_LOG
- ACCOUNT_RQG_GROUPERS

Over 18,000 tables!

ORDER_PROC

Description: The ORDER_PROC table enables you to report on the procedures ordered in the clinical system. We have also included patient and contact identification information for each record.

Primary Key

Column Name	Ordinal Position
ORDER_PROC_ID	1

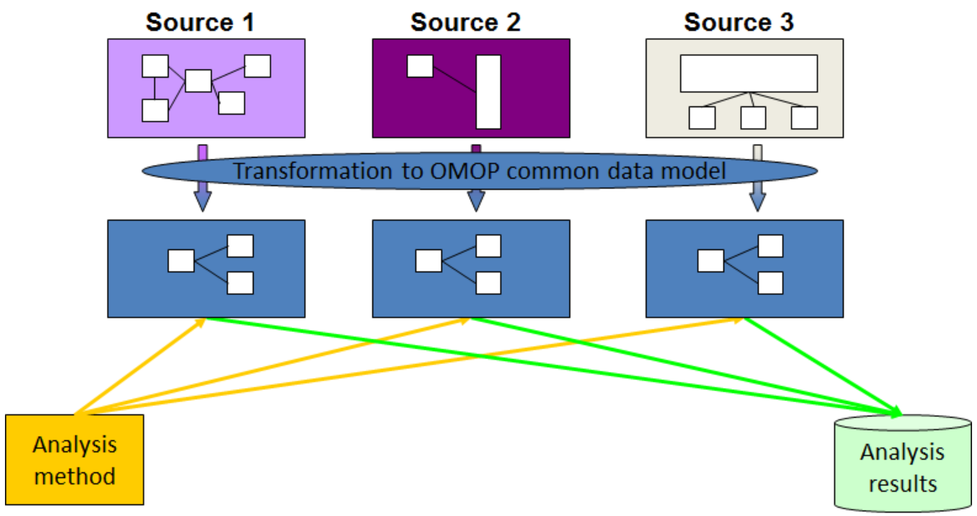
Column Information

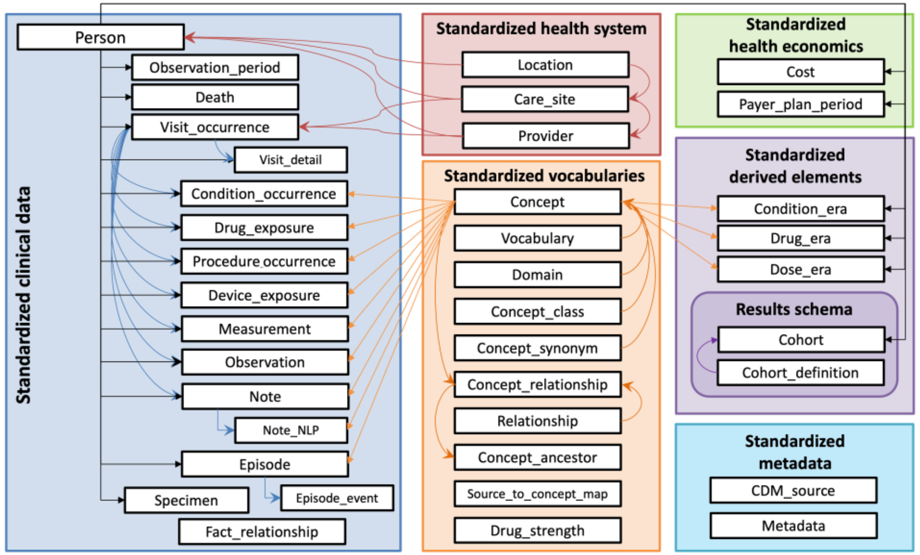
	Name	Type	Discontinued?
1	**ORDER_PROC_ID**	**NUMERIC**	No
	The unique ID of the order record associated with this procedure order.		
2	**PAT_ID**	**VARCHAR**	No
	The unique ID of the patient record for this order. This column is frequently used to link to the PATIENT table.		
3	**PAT_ENC_DATE_REAL**	**FLOAT**	No
	A unique, internal contact date in decimal format. The integer portion of the number indicates the date of the contact. The digits after the decimal distinguish different contacts on the same date and are unique for each contact on that date. For example, .00 is the first/only contact, .01 is the second contact, etc.		
4	**PAT_ENC_CSN_ID**	**NUMERIC**	No
	The unique contact serial number for this contact. This number is unique across patients and encounters in your system. If you use IntraConnect this is the Unique Contact Identifier (UCI).		
5	**RESULT_LAB_ID**	**NUMERIC**	No
	The unique ID of the lab or other resulting agency, such as radiology, that provided the order results.		
6	**RESULT_LAB_ID_LLB_NAME**	**VARCHAR**	No
	Interface laboratory name.		
7	**ORDERING_DATE**	**DATETIME**	No
	The date when the procedure order was placed.		
8	**ORDER_TYPE_C_NAME**	**VARCHAR**	No
	The order type category number for the procedure order. May contain organization-specific values: Yes Category Entries: Medications		
9	**PROC_ID_PROC_NAME**	**VARCHAR**	No

https://open.epic.com/EHITables/GetTable/_index.htm

Common Data Models

- OMOP, PCORnet standardize structure & vocabularies for research
- Enable multicenter studies, repositories, shared tools





How Are EHR Data Accessed?

- EHR data contain protected health information (PHI)
- Use governed by IRB, HIPAA, and institutional data use policies
- Often accessed through secure enclaves or virtual data centers
    - Access restricted to IRB-approved subset of data
    - Structured Query Language (SQL) or specialized user interface tools (ATLAS, SlicerDicer, etc)
- Data rarely exported; analyses happen inside the firewall/enclave

EHR Data Analyses

Institute for Clinical and Translational Research
AT EINSTEIN AND MONTEFIORE
Building Bridges in the Bronx and Beyond

Description	• What is currently happening? • Example: what proportion of patients with heart failure with reduced ejection fraction get readmitted within 30 days?
Prediction	• What will likely happen? • Which patients will likely have adverse outcomes after a procedure?
Causal Inference	• What would happen under different courses of action? • How would outcomes differ if we used terlipressin versus current standard of care as first-line treatment for HRS?

Adopted from Hernan et al," A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks"

Phenotyping

- Unlike in prospective studies (including clinical trials), EHR data does not have pre-defined cohorts or study-specific variables
- Phenotyping = process of turning available data into study-specific features
  - This is a key step for working with EHR data
- Using data as a surrogate of true disease or health event
  - Example: COPD exacerbation
- Can be done using code (SQL) or specialized tools (ATLAS, TriNetX)

Cohort Building

- Descriptive and feasibility analyses can often be conducted using simple queries
  - E.g. how many COPD admissions do we get per year?
- Most clinical research questions are *person-centric*
  - E.g. in patients admitted with COPD not on triple therapy, how would outcomes differ if we initiated triple therapy in the hospital versus leaving it to the outpatient provider?
- Clinical questions require cohort framework and complex queries
  - Foundation for causal inference and real-world evidence generation

Cohort Definition Framework

Cohort entry event – what triggers the patient to be evaluated for your cohort?

↓

Inclusion criteria – patients who meet them at entry event enter your cohort

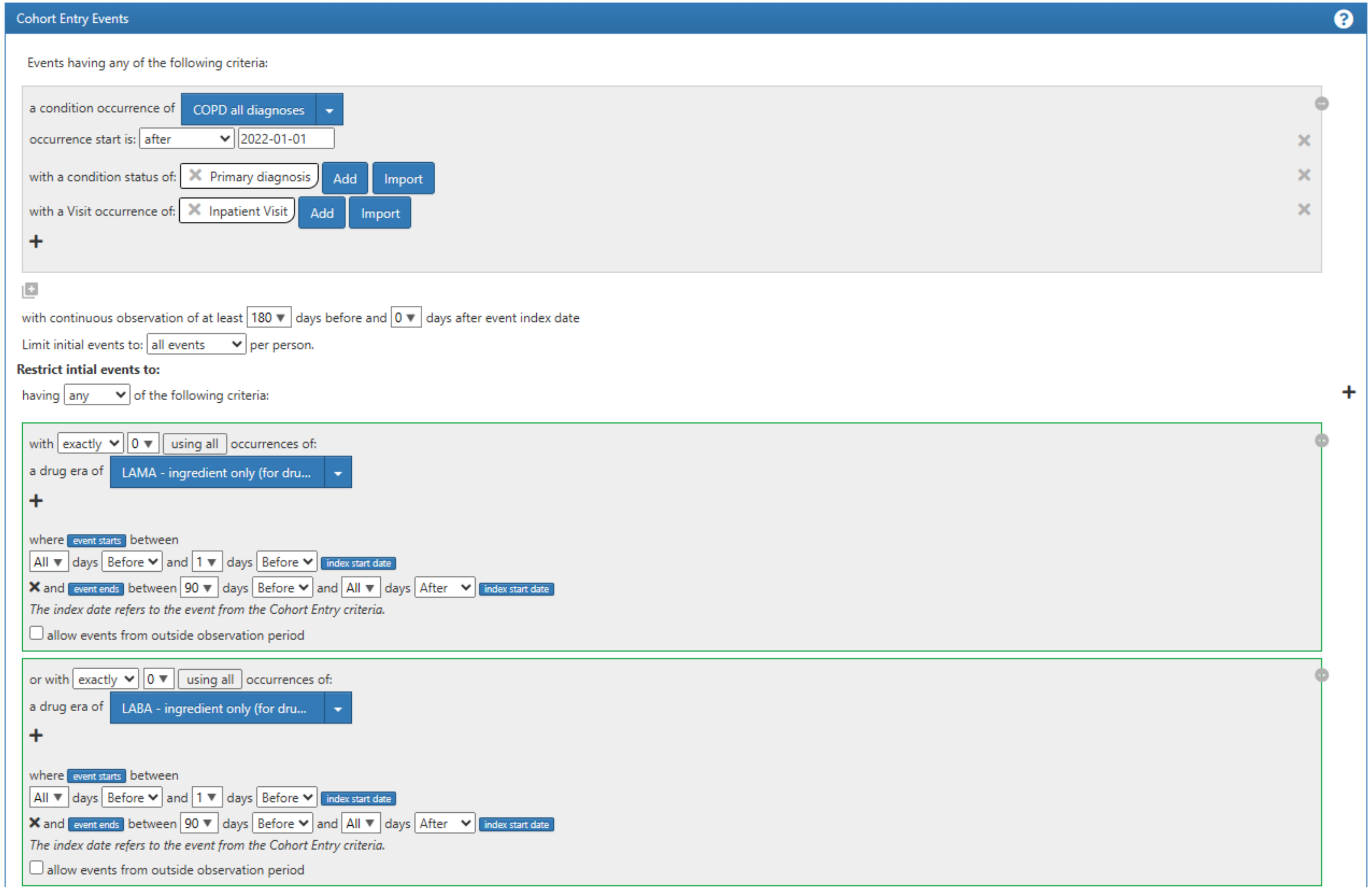
↓

Exit criteria – what triggers the patient to exit cohort?

Project Example: COPD

- Study question: in patients admitted with COPD not on triple therapy, how would outcomes differ if we initiated triple therapy in the hospital versus leaving it to the outpatient provider
- Phenotyping tasks:
  - Eligible patients
    - Admitted with COPD exacerbation
    - Not on triple therapy at the time of admission
  - Prescribed triple therapy at discharge
- Phenotypes combined into complex study cohort

Project Examples: COPD



LAMA Concept Sets

ICTR Institute for Clinical and Translational Research
AT EINSTEIN AND MONTEFIORE
Building Bridges in the Bronx and Beyond

LAMA

Concept Set Expression | Included Concepts 1284 | Included Source Codes | Recommend | Explore Evidence | Export | Import | Compare | Versions | Message

Enter the concept set description here

Show 50 entries

Filt

Showing 1 to 4 of 4 entries

Concept Id	Concept Code	Concept Name	Domain	Standard Con
44785907	1487514	umeclidinium	Drug	Standard
1106776	69120	tiotropium	Drug	Standard
35201615	2102775	revefenacin	Drug	Standard
42873639	1303098	aclidinium	Drug	Standard

Show columns▼ Copy CSV Show 50 entries

Filter: Search...

howing 1 to 50 of 1,284 entries

Previous 1 2 3

Id	Code	Name	Class	RC	DRC	PC	DPC	Domain	Vocabulary
44785907	1487514	umeclidinium	Ingredient	21,489	103,151	13,174	33,851	Drug	RxNorm
1106776	69120	tiotropium	Ingredient	19,770	108,874	11,880	30,745	Drug	RxNorm
19112761	485032	tiotropium 0.018 MG Inhalation Powder	Clinical Drug	16,916	53,931	6,382	11,415	Drug	RxNorm
19121580	580261	tiotropium 0.018 MG Inhalation Powder [Spiriva]	Branded Drug	37,015	37,015	5,033	5,033	Drug	RxNorm
37003605	2395775	30 ACTUAT fluticasone furoate 0.2 MG/ACTUAT / umeclidinium 0.0625 MG/ACTUAT / vilanterol 0.025 MG/ACTUAT Dry Powder Inhaler [Trelegy]	Quant Branded Drug	10,599	10,599	3,675	3,675	Drug	RxNorm
44785545	1487524	7 ACTUAT umeclidinium 0.0625 MG/ACTUAT / vilanterol 0.025 MG/ACTUAT Dry Powder Inhaler [Anoro]	Quant Branded Drug	23,596	23,596	3,364	3,364	Drug	RxNorm
792494	1945048	30 ACTUAT fluticasone furoate 0.1 MG/ACTUAT / umeclidinium 0.0625 MG/ACTUAT / vilanterol 0.025 MG/ACTUAT Dry Powder Inhaler [Trelegy]	Quant Branded Drug	7,175	7,175	3,198	3,198	Drug	RxNorm
45775241	1539885	30 ACTUAT umeclidinium 0.0625 MG/ACTUAT Dry Powder Inhaler [Incruse]	Quant Branded Drug	7,478	7,478	2,821	2,821	Drug	RxNorm
45892418	1596445	7 ACTUAT umeclidinium 0.0625 MG/ACTUAT Dry Powder Inhaler [Incruse]	Quant Branded Drug	8,475	8,475	2,483	2,483	Drug	RxNorm
1511446	2166204	10 ACTUAT tiotropium 0.0025 MG/ACTUAT Inhalation Spray [Spiriva]	Quant Branded Drug	20,037	20,037	2,423	2,423	Drug	RxNorm
44785540	1487527	30 ACTUAT umeclidinium 0.0625 MG/ACTUAT / vilanterol 0.025 MG/ACTUAT Dry Powder Inhaler	Quant Clinical Drug	4,718	4,813	2,294	2,332	Drug	RxNorm
42873639	1303098	aclidinium	Ingredient	3,301	35,523	2,215	5,248	Drug	RxNorm
			Quant						

SQL Code

Institute for Clinical and Translational Research
AT EINSTEIN AND MONTEFIORE
Building Bridges in the Bronx and Beyond

```
select event_id, person_id, start_date, end_date, op_start_date, op_end_date
into #included_events
FROM (
  SELECT event_id, person_id, start_date, end_date, op_start_date, op_end_date, row_number() over (partition by person_id order by start_date ASC) as ordinal
  from
  (
    select Q.event_id, Q.person_id, Q.start_date, Q.end_date, Q.op_start_date, Q.op_end_date, SUM(coalesce(POWER(cast(2 as bigint), I.inclusion_rule_id), 0)) as inclusion_rule_mask
    from #qualified_events Q
    LEFT JOIN #inclusion_events I on I.person_id = Q.person_id and I.event_id = Q.event_id
    GROUP BY Q.event_id, Q.person_id, Q.start_date, Q.end_date, Q.op_start_date, Q.op_end_date
  ) MG -- matching groups

  -- the matching group with all bits set ( POWER(2,# of inclusion rules) - 1 = inclusion_rule_mask
  WHERE (MG.inclusion_rule_mask = POWER(cast(2 as bigint),4)-1)

) Results

;

-- date offset strategy

select event_id, person_id,
  case when DATEADD(day,1,end_date) > op_end_date then op_end_date else DATEADD(day,1,end_date) end as end_date
INTO #strategy_ends
from #included_events;


-- generate cohort periods into #final_cohort
select person_id, start_date, end_date
INTO #cohort_rows
from ( -- first_ends
    select F.person_id, F.start_date, F.end_date
    FROM (
      select I.event_id, I.person_id, I.start_date, CE.end_date, row_number() over (partition by I.person_id, I.event_id order by CE.end_date) as ordinal
      from #included_events I
      join ( -- cohort_ends
-- cohort exit dates
-- End Date Strategy
```

445 lines

EINSTEIN
Albert Einstein College of Medicine
Montefiore

Case Study: Local EHR Data Access at Montefiore Einstein

- Epic EHR + specialized data warehouse mapped to OMOP common data model
- Accessible via ATLAS for cohort building and analytics + Epic tools for basic reporting (SlicerDicer)
  - SQL access only for informatics analysis or deidentified data
- Extraction possible to a Montefiore-managed device for local analysis
  - Built-in IRB check to ensure extraction approved
- Larger extractions and ECG/DICOM require work on HPC
  - Has SQL/R/python available

Case Study: Epic Cosmos

- 208 institutions across U.S. contribute data
- Access via participating institution (institutional data use agreements)
- SlicerDicer-like interface for basic summary-level queries
- For patient level data, must undergo Epic certification as Data Scientist or Data Architect
- Project review and approval by Epic + local governance
- Data access fees
- Remote repository and virtual analytic workspace (SQL, R, python)
    - All data and code stay on remote workspace and are audited

Data-Specific Challenges

EHR: Not Designed for Research

- Built for care delivery, billing, regulatory compliance
- Documentation priorities do not align with research priorities
- Results: incomplete, inconsistent, biased

Data Completeness & Bias

- Data present when patients seek care at the institution
- Health status influences follow-up & testing
    - Unlike in trials where regular follow up is built-in
- Coding practices largely driven by reimbursement incentives
- Insurance, access, and location drive representativeness

Data Quality & Structure

- Not all data structured (notes, images, PDFs)
- Structured fields inconsistently filled
- Standard codes (LOINC, RxNorm) incomplete + local quirks
- Solutions:
  - Mapping concepts to standardized vocabularies
  - Extraction of structured concepts from unstructured data

Example: CAM-ICU

- Used to indicate delirium
- Theoretically two options: delirium present/absent
- Actual values

	MEAS_VALUE	TOTAL_OCCURRENCES
1	Negative - not Delirious	943862
2	Unable to Assess	722439
3	Delirious	44327
4	Incomplete	23170
5	(INCOMPLETE)	20278
6	Unable to assess	19549
7	Not delirious	11515
8	(null)	3681
9	Negative	2907
10	Positive	1781

Irregular/Nonrandom Measurement

- Labs/vitals not done on fixed schedule
- Healthy patients less likely to return → nonrandom missing data
    - Can bias finding if inappropriately handled

Irregular Measurements: Example

You're comparing two cohorts – those prescribed SPS versus patiromer (two potassium binding agents)

You're interested in checking their potassium lowering trends over next 6 months

Challenges

Some never return at 6 months	Some will only return at month 1 or 6	Some will only return at months 2, 3, and 5

What are some potential solutions?

Irregular/Nonrandom Measurement

- Labs/vitals not done on fixed schedule
- Healthy patients less likely to return → nonrandom missing data
  - Can bias finding if inappropriately handled
- Possible solution: require evidence of regular care pre-baseline

Nonstandard & Complex Data Types

- Imaging, ECG waveforms, pathology slides
- Require specialized methods (NLP, AI/ML)
- Integration with structured data is difficult
- Potential solutions:
  - Signals/images into structured features (e.g. QT interval, digital pathology quantified features)
  - Natural language processing to extract distinct findings

Natural Language Processing

- FLT-ITD mutated AML with monocytic differentiation

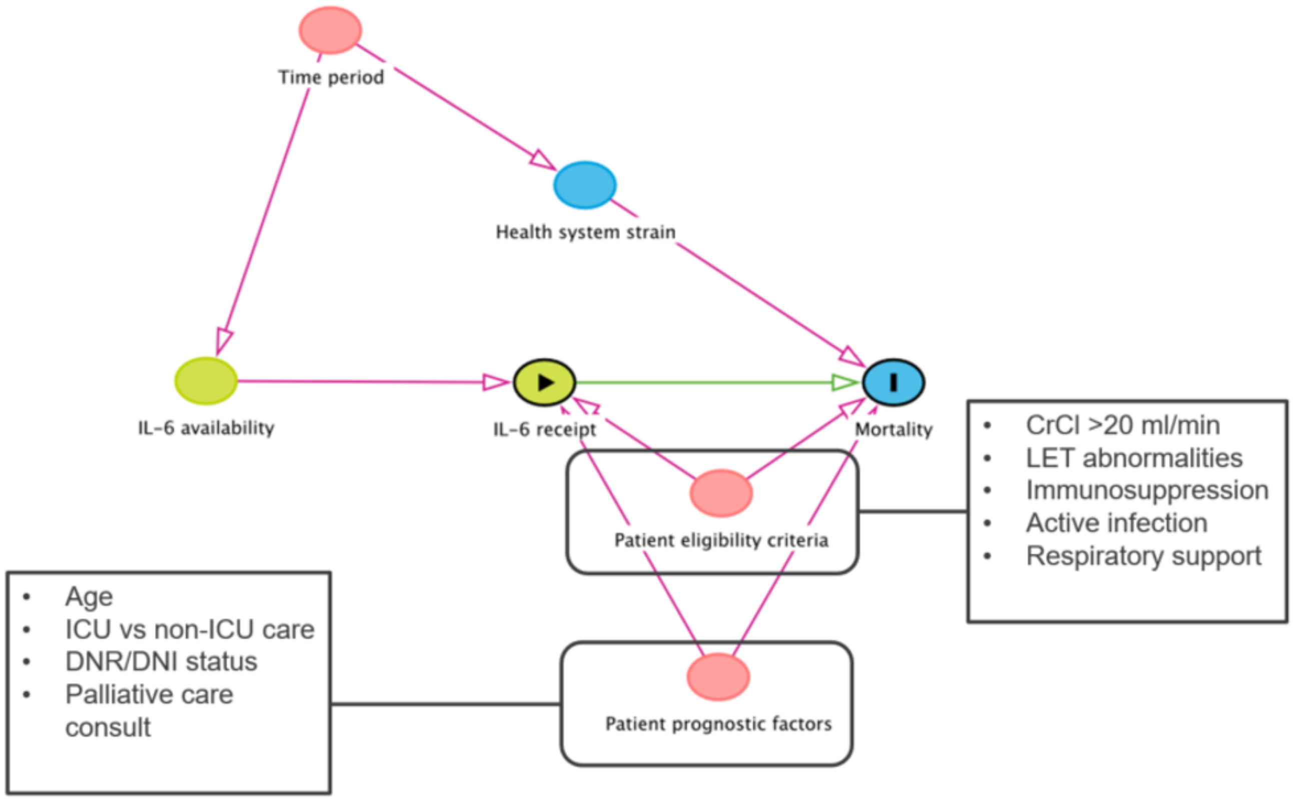
```
l information: The source of the history was the patientAnd medical record Reliability:  The patient was judged to be
hief Complaint: Elective admission for HiDAC #1HPI[redacted]diagnosis of FLT3-ITD mutated AML diagnosed ********
, pt also had pharyngitis during same admission and was treated w Amoxicillin, pt also received multiple supportive
er on **** without symptoms. Started on cefepime/vanco for neutropenic fever, seen by ID.  Patient developed morbilliform
man syndrome, over several days rash continued to spread throughout body. Seen by derm and allergy, likely related to amox,
llopurinol which were all stopped. Biopsy taken, labs were monitored for DRESS but remained stable.  Pt returns to hospital
 has mostly resolved, except for feet to mid-calf w peeling skin now, but no longer itchy, pt otherwise says feels well, NA
l not be started tonight.Denies HA/malaise/fever/vision changes/CP/SOB/abdm
n/dysuria/hematuria/BRBPR/leg pain or swelling.SARS-COV-2 PCR/SWAB (COVID-19) - STANDARD [**********] (Normal) Collected:
al result Specimen: Swab from Nasopharyngeal   SARS-CoV-2 PCR/Swab (COVID-19) Not Detected Bone marrow [redacted]o
 aspirate: - Acute myeloid leukemia with monocytic differentiation. - Hypercellular marrow (>90%) with approximately 80-90%
```

Methodological Challenges

Confounding & Treatment Selection

- Real-world treatment selection is not random
- Indication, severity, provider preference drive choice
- Need to measure & adjust for these factors

Treatment Selection
IL-6 Inhibitors for COVID-19



Defining Time Zero

- No natural randomization/enrollment date
- Must define cohort entry & follow-up start carefully
- Mis-specification of time zero leads to bias

Time Zero Bias in Practice

ICTR Institute for Clinical and Translational Research
AT EINSTEIN AND MONTEFIORE
Building Bridges in the Bronx and Beyond

ELSEVIER
Current Problems in Cardiology
Volume 44, Issue 10, October 2019, 100407

Review

Heart Failure Postdischarge Clinic: A Pharmacist-led Approach to Reduce Readmissions

Rasha Al-Bawardy MD, Angela Cheng-Lai PharmD, Lendita Prlesi PharmD, Manaf Assafin MD, Shuo Xu ScM, Kiana Chen, Samvit Tandan MD, PhD, Chino S. Aneke MD, PhD, Sandhya Murthy MD, Ileana L. Piña MD, MPH ✉

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.cpcardiol.2018.12.004 ↗ Get rights and content ↗

Introduction

Heart failure (HF) is a major cardiovascular syndrome that affects approximately 5.7 million adults in the United States.[1] Although the survival rate after HF diagnosis has improved over time, the death rate from this disease remains high. It is estimated that 50% of those diagnosed with HF will die within 5 years after the first diagnosis.[1] The cost

ELSEVIER
Current Problems in Cardiology
Volume 48, Issue 2, February 2023, 101507

Evaluating Pharmacist-Led Heart Failure Transitions of Care Clinic: Impact of Analytic Approach on Readmission Rate Endpoints

Angela Cheng-Lai [a b c] ✉, Lendita Prlesi [a], Sandhya Murthy [b c], Eran Y. Bellin [b c d], Mark J. Sinnett [a], Pavel Goriacko [a c d]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.cpcardiol.2022.101507 ↗ Get rights and content ↗

Abstract

Studies evaluating pharmacist-led transitions of care (TOC) services for heart failure patients reported profound decreases in hospital readmissions. Most studies restricted their analysis to clinic attendees (as-treated analysis), which can introduce selection and immortal time bias. In this study, we evaluated the impact of including only clinic attendees vs all clinic referrals in assessing the effectiveness of a pharmacist-led heart failure transitions of care (PharmD HF TOC) clinic program on 30-day readmissions. This

EINSTEIN Albert Einstein College of Medicine Montefiore

Loss of Longitudinally

- Detailed encounter data but fragmented over time
- Patients move between systems (especially in NYC)
- Leads to misclassification and missing data bias
- Possible solutions: repositories + linkage
    - Claims, registries, mortality → fill longitudinal gaps
- Analytic solution: restrict cohort to those “engaged with health system”
    - Based on regular previous visits

Key Takeaways

- EHR data is not research ready, needs cleaning & careful design
- Bias, missingness, and access hurdles are real
- Multidisciplinary expertise is essential (esp. BERD and Informatics)
- When done well: faster, cheaper, ethically feasible studies

Building Your EHR Research Team

Clinical Expertise

Frame meaningful questions, interpret clinical context, validate findings

Informatics and Data Science

Extract, transform, and manage complex EHR data with institutional and regulatory compliance

Epidemiology and Biostatistics

Design studies, address bias, and analyze results appropriately

Institute for Clinical and Translational Research
AT EINSTEIN AND MONTEFIORE
Building Bridges in the Bronx and Beyond

Electronic Health Record (EHR) Data in Clinical Research

Pavel Goriacko, PharmD, MPH
Co-Director, Health Research Informatics

Albert Einstein College of Medicine