ETL from InCor ePR to OMOP CDM

Daniel Mário de Lima

2019-05-02

Summary

- Introduction
  - Motivation
  - Objectives
- Concepts
  - Health Information Systems
  - Clinical Research
  - KDD
  - OMOP CDM
- Method
  - Evaluation and Results
- Conclusion

Introduction

- 2010 -- 2020
- Web 2.0 ---> 3.0
- Parallel/Distributed
- Cloud Computing
- Social Networks
- Big Data
- Large-Scale KDD
- Machine Learning

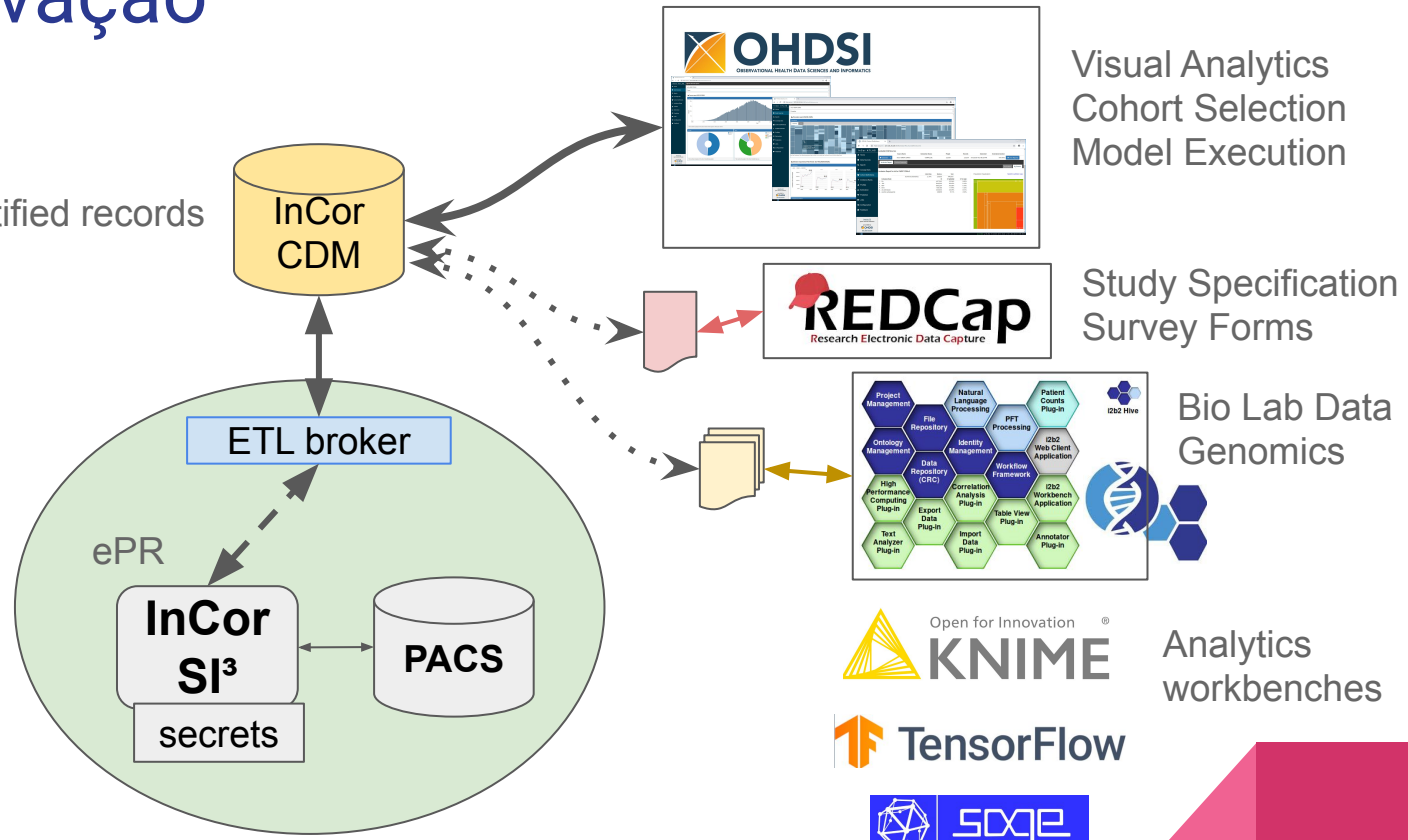
Introduction

- 2010 -- 2020
- Web 2.0 ---> 3.0
- Parallel/Distributed
- Cloud Computing
- Social Networks
- **Big Data**
- **Large-Scale KDD**
- Machine Learning

- Volume
- Velocity
- Variety
- Veracity
- Value

(Gudivada, Baeza-Yates, Raghavan, 2015)

Motivação



Open for Innovation KNIME

Analytics
workbenches

TensorFlow

sage

jupyter

Objectives

- Prepare a new ETL (extract-transform-and-load) layer for InCor's ePR;
- Curate a anonymized database following an international standardized data model for clinical research (OMOP CDM);
- Evaluate the new database quality at recreating patient cohorts of a previous reference study.

Concepts

Healthcare Information Systems

Hospital-centric

- HIS (Hospital Information System)
- Registers all hospital activities
- Patients, Materials, Nursery, Administrative, Billing, Pharmacy...

Patient-centric

- ePR (electronic patient records)
- Registers interactions between patients and providers
- Visits, Hospitalization, Medication, Tests, Procedures...

Clinical Research

Evidence-based diagnosis (Cruz e Pimenta, 2005)

V. expert opinions, case reports, descriptive studies

IV. non-experimental studies from several sources

III. non-randomized trials, cohorts, time series, case-control studies

II. randomized controlled trials (RCT)

I. systematic reviews of RCTs

Clinical Research

Evidence-based diagnosis (Cruz e Pimenta, 2005)

I. systematic reviews of RCTs

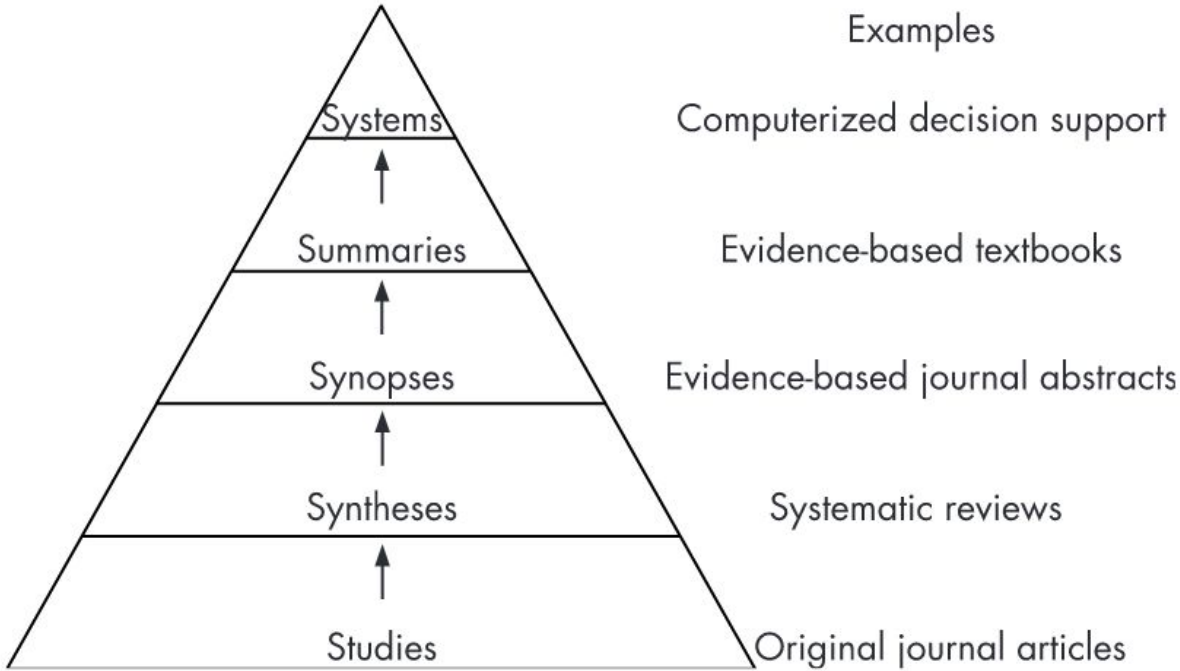
II. randomized controlled trials (RCT)

III. non-randomized trials, cohorts, time series, case-control studies

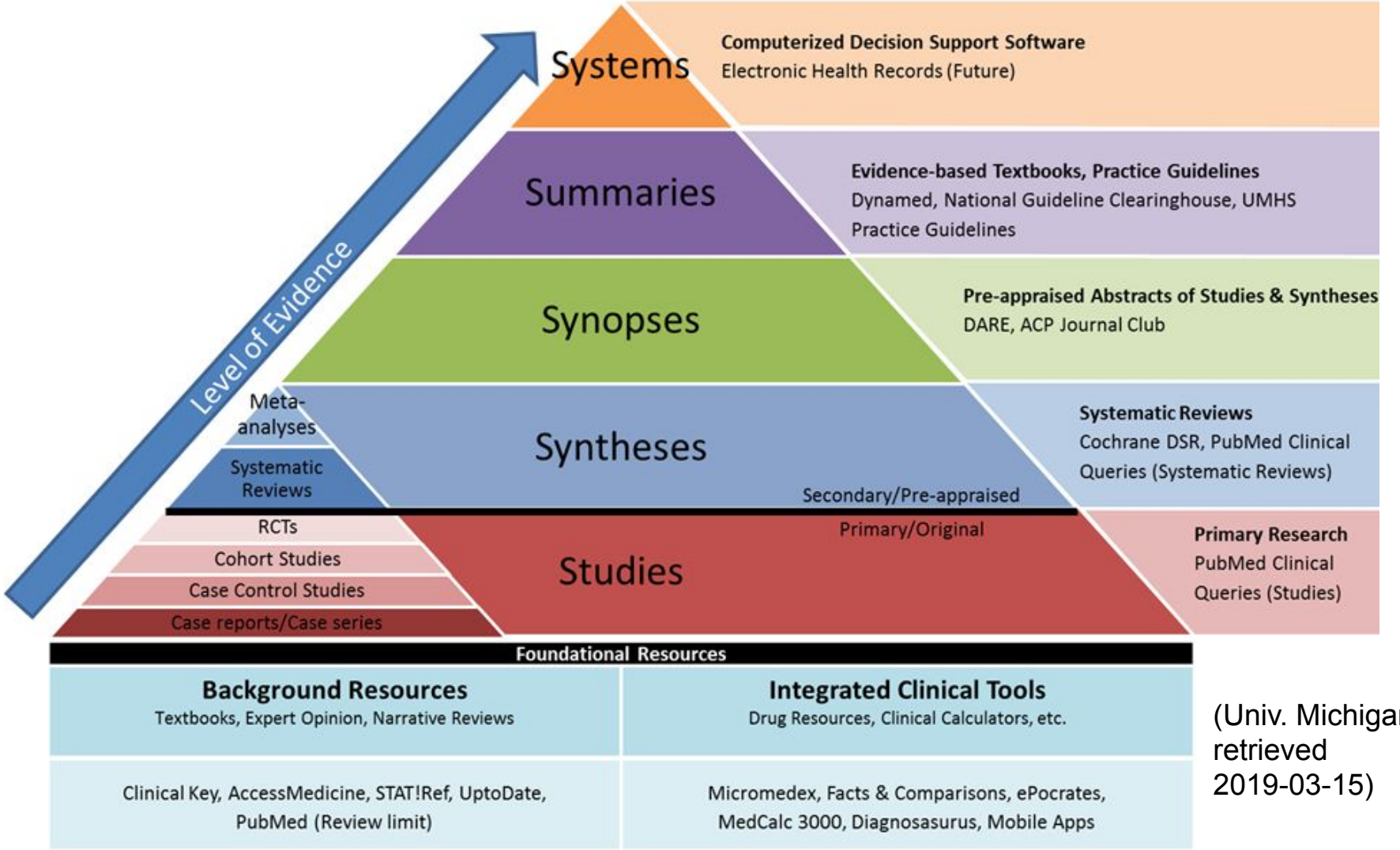
IV. non-experimental studies from several sources

V. expert opinions, case reports, descriptive studies

“5S” Model



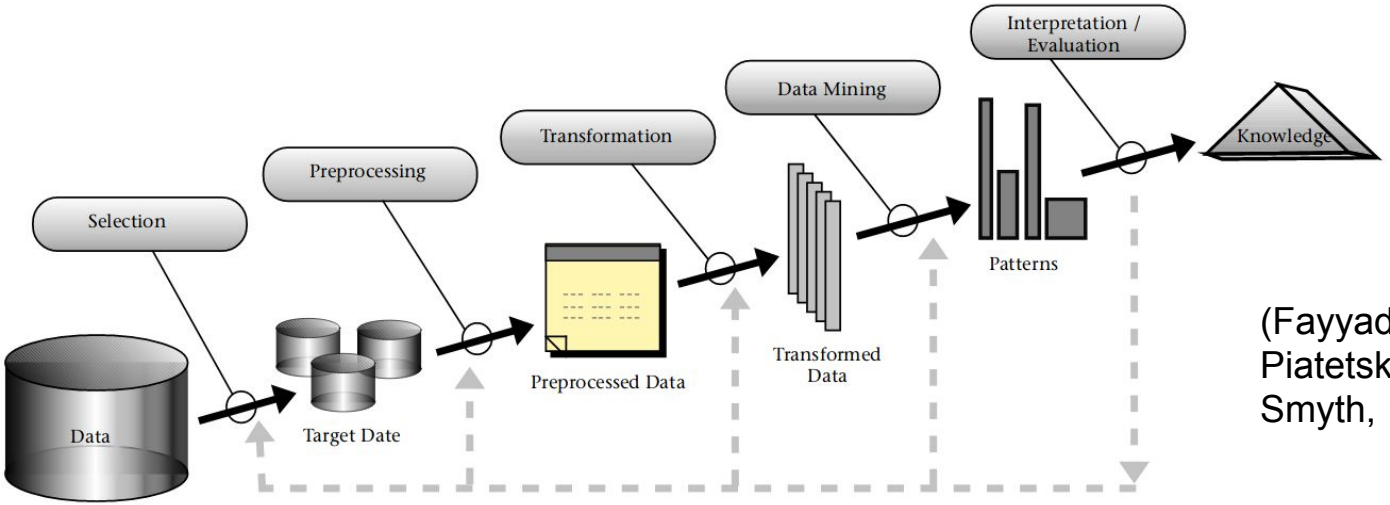
(Haynes, 2007)



(Univ. Michigan, retrieved 2019-03-15)

KDD

Knowledge Discovery in Databases



(Fayyad, Piatetsky-Shapiro, Smyth, 1996)

Clinical Data Acquisition

Ethics

- Morals, legal codes, *Ethos*, Hippocrates, Spinoza
- Nuremberg Trials (1945-49)
- Belmont Report (1979)
  - Respect for Persons, Beneficence, Justice
- Health Insurance Portability and Accountability Act (HIPAA)
  - Protected Health Information (PHI)
    - name, address, birth date, Social Security Number, etc.
  - De-Identified Health Information
  - Research clause

OLTP vs OLAP

On-Line Transaction Processing

- Stores an Information System’s data
- ACID protocol (atomicity, consistency, isolation, durability)
- Performance e scalability
- Relational Model

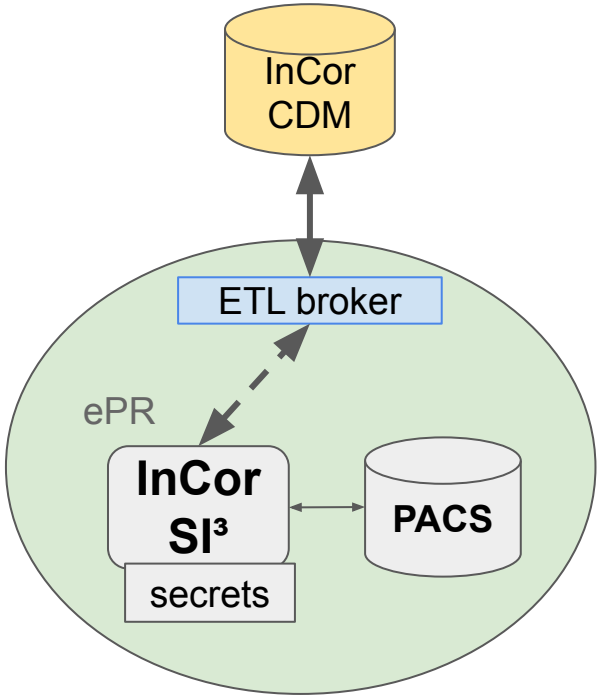
On-Line Analytical Processing

- Retrieval and interpretation of data in a DB
- Organization, aggregation and summarization of values
- Dimensional Modelling (via OLAP cubes)
- Execution over Relational DBs (ROLAP)

ETL

Extract, Transform & Load

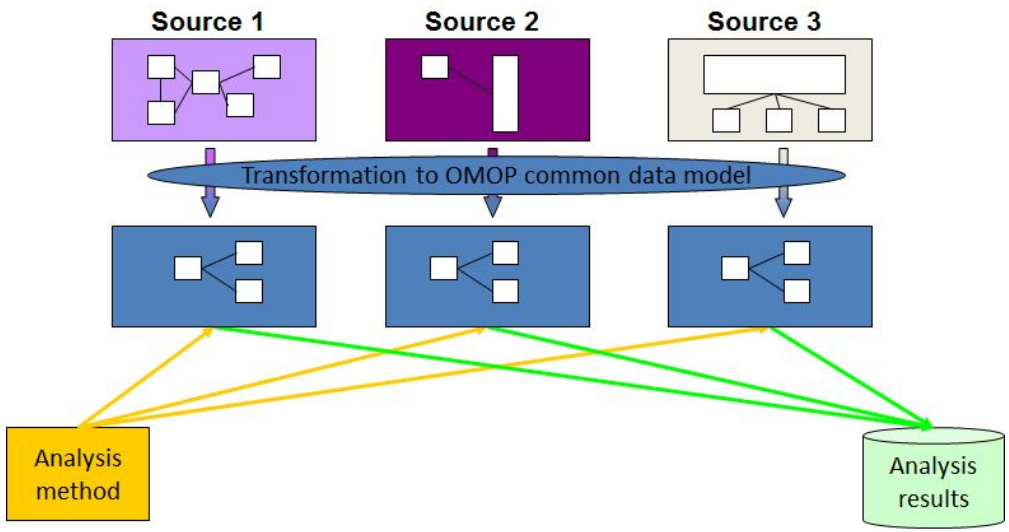
Usage of data definition and manipulation languages (DDL/DML) to transport data acquired from several DBs to a **data mart** for analysis.



OMOP ---> OHDSI

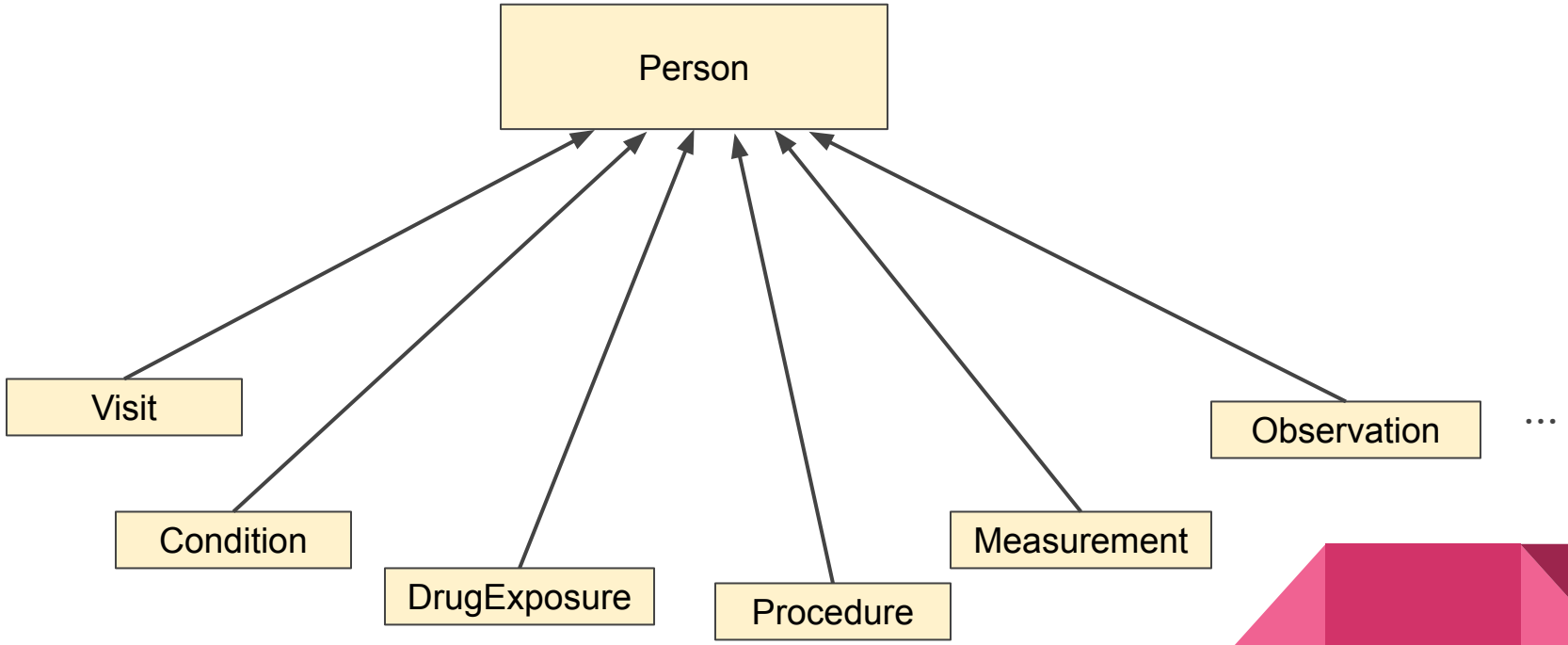
Observational Medical Outcomes Partnership

Observational Health Data Sciences and Informatics

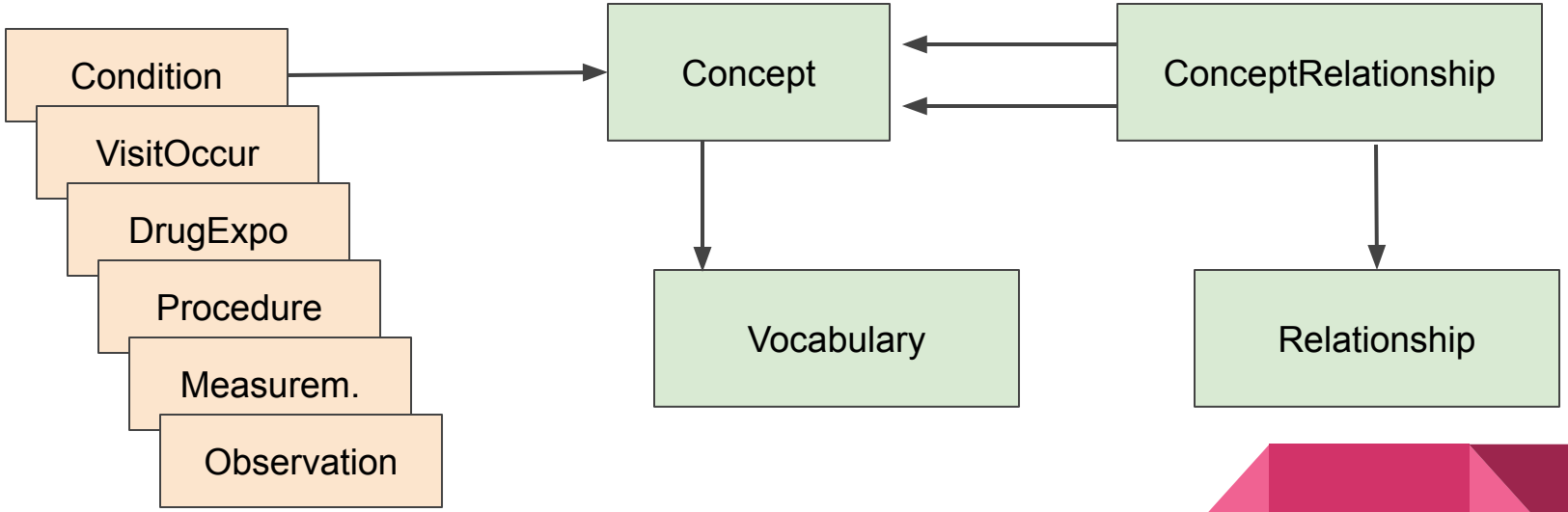


(ohdsi.org, acessado em 2019-03-15)

Common Data Model (CDM)



CDM Metadada



Example

Person

person_id	year_of_birth	gender_concept_id
128172	1985	8205

Condition_Occurrence

condition_occurrence_id	person_id	condition_concept_id	start_date	end_date
8127	128172	812739	2015-01-02	2017-01-01

Concept

concept_id	concept_name	vocabulary_id
812739	PNEUMONIA	InCor

Example

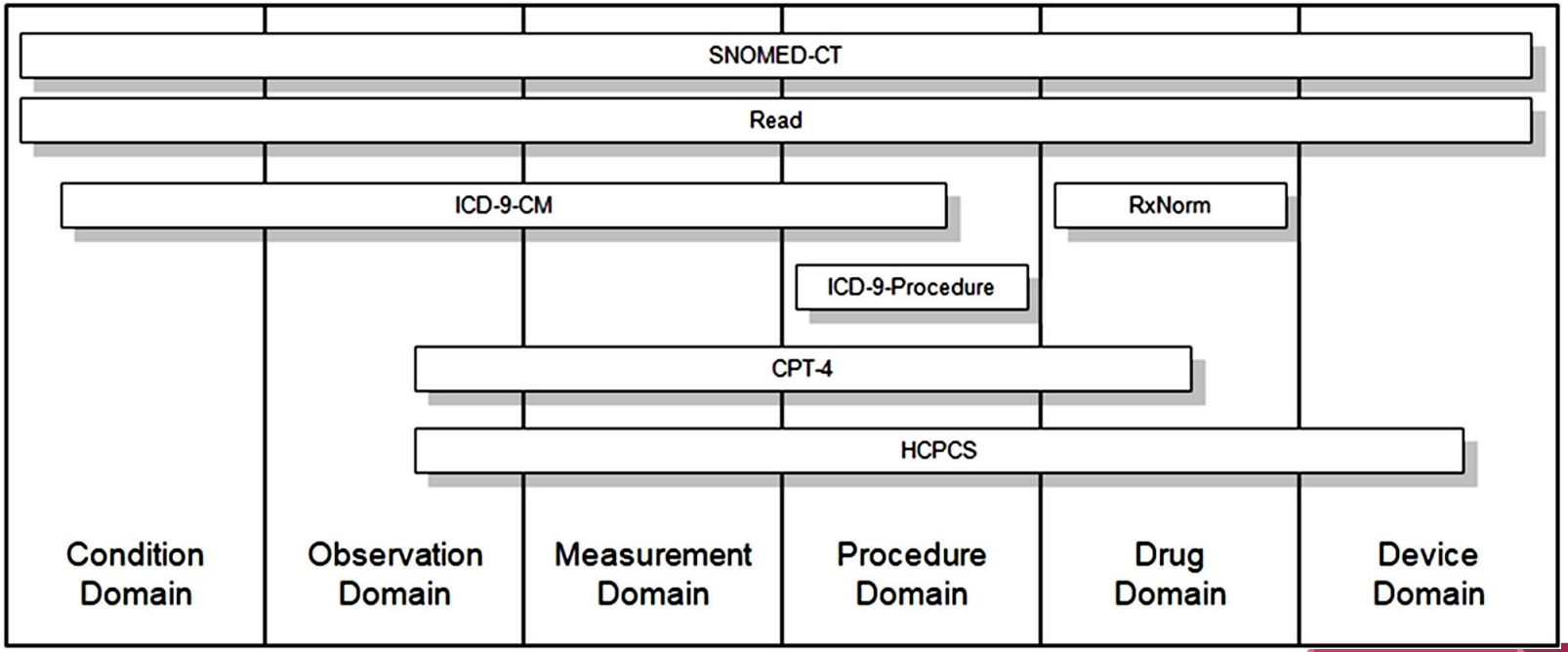
- Concept

concept_id	concept_name	vocabulary_id
812739	PNEUMONIA	InCor
53084003	Bacterial pneumonia	SNOMED-CT
8783836492	J15 - Pneumonia bacteriana não classificada em outra parte	CID-10

- Concept_Relationship

concept_id_1	concept_id_2	relationship_id
812739	53084003	Maps to
812739	8783836492	Subsumes
53084003	8783836492	Subsumes

Standard CDM Vocabularies



Data Mining

Exploratory Data Analysis, using computable [mathematical and statistical] properties of objects under study.

Fayyad et al, 1996:

- Regression
- Classification
- Cluster analysis
- Summarization, dimensionality reduction
- Dependency modelling
- Anomaly, change and deviation detection

Method

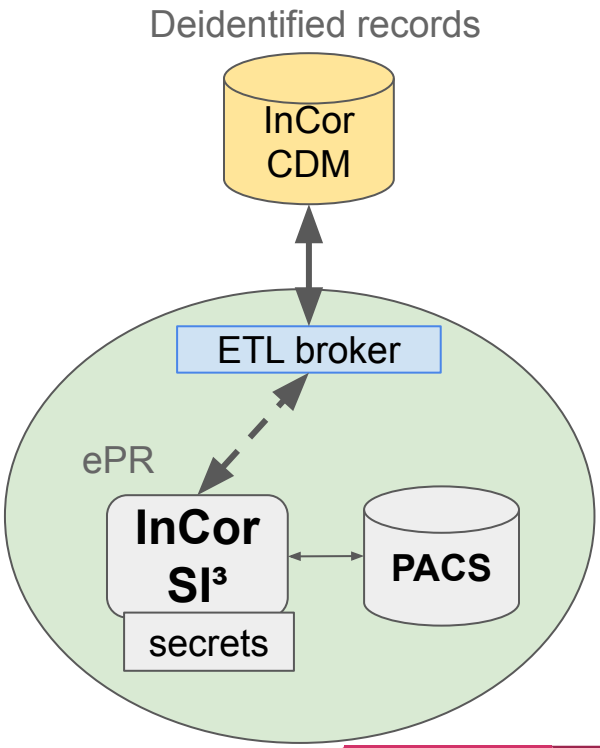
InCor SI³ ---> InCor-CDM

Objective: Prepare a CDM DB for clinical research (InCor-CDM)

Dataset: InCor SI³ (ePR / EHR)

Domínio	SI3-2016	Pauá	SI3-2018	InCor-CDM
Person	1.116	323	1.346	946
Visit Occurrence	6.427	5.686	7.499	7.305
Condition Occurrence	1.205	1.007	1.361	1.324
Procedure Occurrence	45.024	144	53.945	51.479
Drug Exposure	83.283	2.775	100.052	38.962
Measurement	22.025	20.528	31.095	30.177
Death	17	21	18	18

×1000



ETL

1. Pseudonymize Patients
   a. All PKs are reassigned to random new keys
   b. Patient and Visit PKs are stored in a private table, for medical use if needed
2. For each CDM table:
   a. Find source tables in SI³
   b. Join patient id and pseudonym PK
   c. Truncate PHI values -- k-anonimização
   d. Join CDM standard concept
   e. Project to CDM schema

- AG_TRANSFUSIONAL
- AMBULATORIO
- APEX_030200
- APEX_040100
- APEX_040200
- CAIXA
- CALLCENTER
- CIR_CIRURGIA
- CMP
- COCI
- COM
- COP_SEG
- CTXSYS
- DBA_FABIO
- DBA_VALDEMIR
- EASY
- ENFERMAGEM
- FABIANO
- FONETICA
- GENEHY
- GREA
- GSMADMIN_INTERNAL
- GUICHELATTES
- HELPDESK
- HISTORICO
- INFOSAUDE
- INFOSAUDE_SI3
- INF_GER
- INTEGRACAO
- JAVA_ADM
- JOB_OPER
- KIT_INCOR
- LAUDO
- LAUDO_SI3
- MAGIC
- MAILING
- MANUT_HIST
- MASS
- MAT
- MOBILEADMIN
- MYSQL_DRIVER
- NETL_HC_SI3

```sql
INSERT INTO person
--CREATE OR REPLACE VIEW v_person AS
SELECT
        person.id                       person_id,
        gender.id                       gender_concept_id,
        EXTRACT(YEAR FROM paci_dt_nasc) year_of_birth,
        NULL                            month_of_birth,
        NULL                            day_of_birth,
        NULL                            birth_datetime,
        coalesce(color.id, 0)           race_concept_id,        -- unknown (0) se pardo ou cns:99
        0                               ethnicity_concept_id,   -- unknown
        NULL                            location_id,
        NULL                            provider_id,
        NULL                            care_site_id,
        NULL                            person_source_value,
        paciente.paci_tp_sexo           gender_source_value,
        NULL                            gender_source_concept_id,
        paciente.paci_sg_cor            race_source_value,
        NULL                            race_source_concept_id,
        NULL                            ethnicity_source_value,
        NULL                            ethnicity_source_concept_id
FROM PACIENTE.PAC_PACIENTE paciente
JOIN keys person
        ON  person.rel = 1302
        AND person.src = paci_id
LEFT JOIN gender
        ON gender.src = paciente.paci_tp_sexo
LEFT JOIN color
        ON paci_sg_cor = color.src
```

keys

table
src_id
new_id

```sql
INSERT INTO visit_occurrence
--CREATE OR REPLACE VIEW v_visit_occurrence AS
SELECT  item.id                         visit_occurrence_id,
        person.id                       person_id,
        visit.id                        visit_concept_id,
        trunc(adm_dt_ingresso,'DD')
                AS                      visit_start_date,
        NULL                            visit_start_datetime,
        trunc(coalesce(said_dt_hr, adm_dt_fim_am_hosp_dia, adm_dt_ingresso), 'DD') -- REVIEW
                AS                      visit_end_date,
        NULL                            visit_end_datetime,
        32035                           visit_type_concept_id,          -- EHR encounter record
        NULL                            provider_id,
        NULL                            care_site_id,                   -- dependência
        adm_tp                          visit_source_value,
        NULL                            visit_source_concept_id,
        NULL                            admitting_source_concept_id,
        NULL                            admitting_source_value,
        NULL                            discharge_to_concept_id,
        NULL                            discharge_to_source_value,
        NULL                            preceding_visit_occurrence_id
FROM KEYS item
JOIN PACIENTE.ADM_ADMISSAO
        ON  item.rel = 2478 -- (SELECT id FROM tables WHERE schem = 'PACIENTE' AND name = 'ADM_ADMISSAO')
        AND item.src = adm_nr||';'||adm_ano||';'||adm_inst_cd
JOIN KEYS person
        ON  person.rel = 1302 -- (SELECT id FROM tables WHERE schem = 'PACIENTE' AND name = 'PAC_PACIENTE')
        AND person.src = to_char(adm_paci_id)
JOIN visit
        ON  visit.src = adm_tp
LEFT JOIN PACIENTE.SAI_SAIDA
        ON  said_adm_nr  = adm_nr
        AND said_adm_ano = adm_ano
        AND said_inst_cd = adm_inst_cd
        AND said_dt_hr_canc IS NULL
```

person

person_id
year_of_birth
month_of_birth
gender_concept_id
race_concept_id
gender_source_val
race_source_value
...

visit

visit_id
person_id
concept_id
start_date
end_date
...

ETL

1. Pseudonymize Patients
   a. All PKs are reassigned to random new keys
   b. Patient and Visit PKs are stored in a private table, for medical use if needed
2. For each CDM table:
   a. Find source tables in SI³
   b. Join patient id and pseudonym PK
   c. Truncate PHI values -- k-anonimização
   d. Join CDM standard concept
   e. Project to CDM schema

ETL

```
INSERT INTO omop.person
SELECT K.new_id                          AS person_id,
       EXTRACT(YEAR FROM P.dt_nasc)      AS year_of_birth,
       COALESCE(G.id, 0)                 AS gender_concept_id
FROM si3.pac_paciente P
JOIN omop.keys K ON  K.table='pac_paciente'
                 AND K.src_id=P.paci_id
LEFT JOIN omop.gender_map G ON P.tp_sexo=G.src;
```

Evaluation

1. OHDSI tools (Achilles Heel):

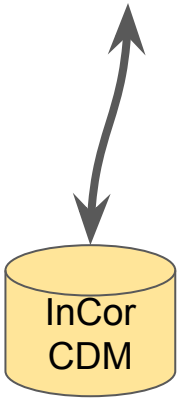
   Referential integrity (FKs)

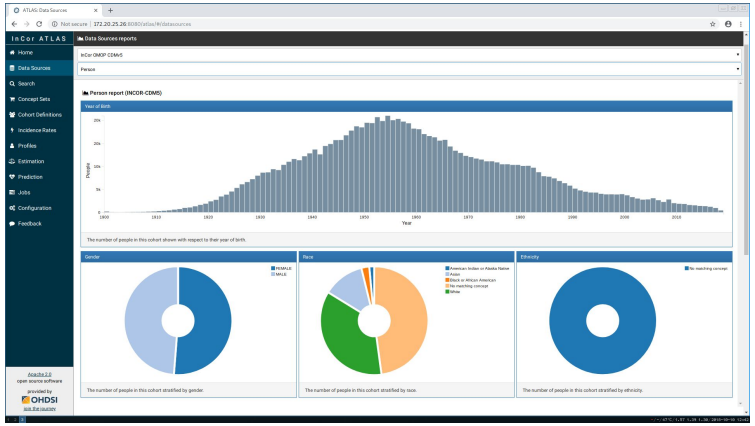
   Consistence

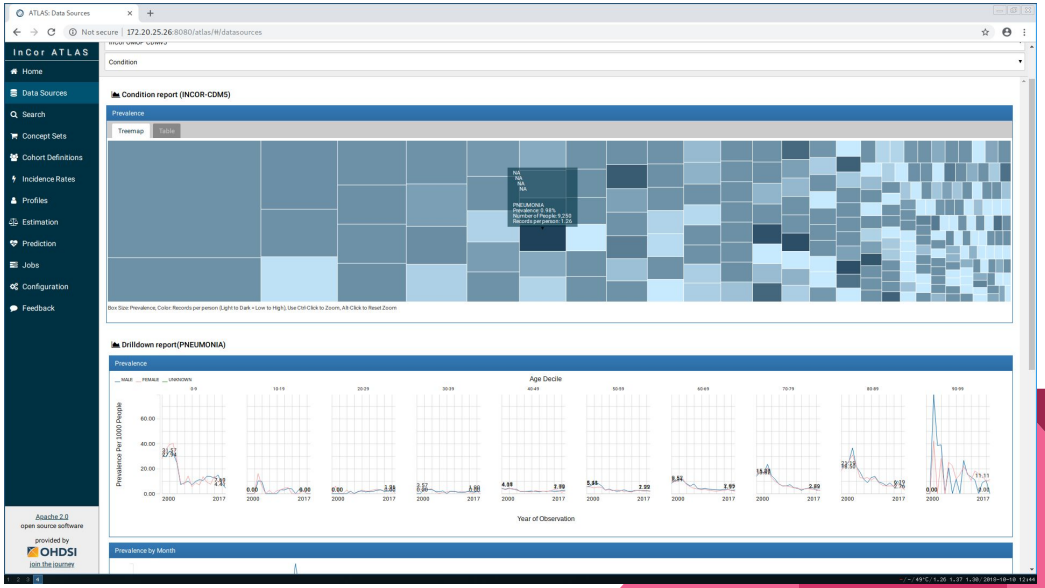
   Missing data

   Veracity

Visual Analytics







ATLAS: Cohort Definitions

Not secure | 172.20.25.26:8080/atlas/#/cohortdefinition/6

InCor ATLAS

Atenção: projeto em desenvolvimento – Pesquisa / Informática / InCor / HCFMUSP

- Home
- Data Sources
- Search
- Concept Sets
- Cohort Definitions
- Incidence Rates
- Profiles
- Estimation
- Prediction
- Jobs
- Configuration
- Feedback

Cohort #6

DCV Abrahao 2016

Definition | Concept Sets | Generation | Reporting | Export | Warnings 13

enter a cohort definition description here

Initial Event Cohort

People having any of the following:

+ Add Initial Event

a visit occurrence of Qualquer Admissão

+ Add criteria attribute...

Delete Criteria

with continuous observation of at least 0 days before and 0 days after event index date

Limit initial events to: all events per person.

Initial event inclusion criteria: From among the initial events, include:

having all of the following criteria:

+ Add criteria to group...

Limit cohort of initial events to: all events per person.

Remove initial event inclusion criteria

Additional Qualifying Inclusion Criteria

New qualifying inclusion criteria

1. 18+
2. M/F
3. Dx
4. DCV
5. 2a admissao
6. evento subsequente

18+ | Copy | Delete

enter an inclusion rule description

having all of the following criteria: + Add criteria to group...

with the following event criteria:

+ Add criteria attribute...

Delete Criteria

with age Between 18 and 80

Limit qualifying cohort to: all events per person.

Cohort Exit Criteria

Add a cohort exit criteria:

- Based on a fixed time period relative to initial event start or end date
- Based on the end of an era of persistent exposure to any drug within a defined concept set

The minimum date from amongst the selected cohort exit criteria occurring after the cohort entry date will be selected as the end date for the person's episode.

Apache 2.0 open source software

provided by OHDSI

join the journey

1 2 3 4

-/-/47°C/1.21 1.35 1.31/2018-10-10 12:46

ATLAS: Cohort Definitions

Not secure | 172.20.25.26:8080/atlas/#/cohortdefinition/6

InCor ATLAS

Atenção: projeto em desenvolvimento -- Pesquisa / Informática / InCor / HCFMUSP

- Home
- Data Sources
- Search
- Concept Sets
- Cohort Definitions
- Incidence Rates
- Profiles
- Estimation
- Prediction
- Jobs
- Configuration
- Feedback

Apache 2.0
open source software
provided by
OHDSI
join the journey

Cohort #6

DCV Abrahao 2016

Definition | Concept Sets | Generation | Reporting | Export | Warnings 13

Available CDM Sources

	Source Name	Generation Status	People	Records	Generated	Generation Duration	
Generate	InCor OMOP CDMv5	COMPLETE	23,339	23,339	10/26/2018 2:48:20 PM	470.431s	View Reports

Inclusion Report | Cohort Features

By Events | By Person

Inclusion Report for InCor OMOP CDMv5

	Match Rate	Matches	Total
Summary Statistics:	2.76%	26,065	946,031

Inclusion Rule	N	% Remain	% Diff
1. Dx	497,263	52.56%	47.44%
2. 18+	444,307	46.97%	5.60%
3. M/F	417,555	44.14%	2.83%
4. DCV	95,115	10.05%	34.08%
5. 2a admissao	72,718	7.69%	2.37%
6. evento subsequente	26,065	2.76%	4.93%

Switch to intersect view

Attrition Visualization

1 2 3 gbdi100%/▲ 65.64% 11:42:12/46°C/0.67 1.04 1.03/2018-11-09 09:13

Evaluation

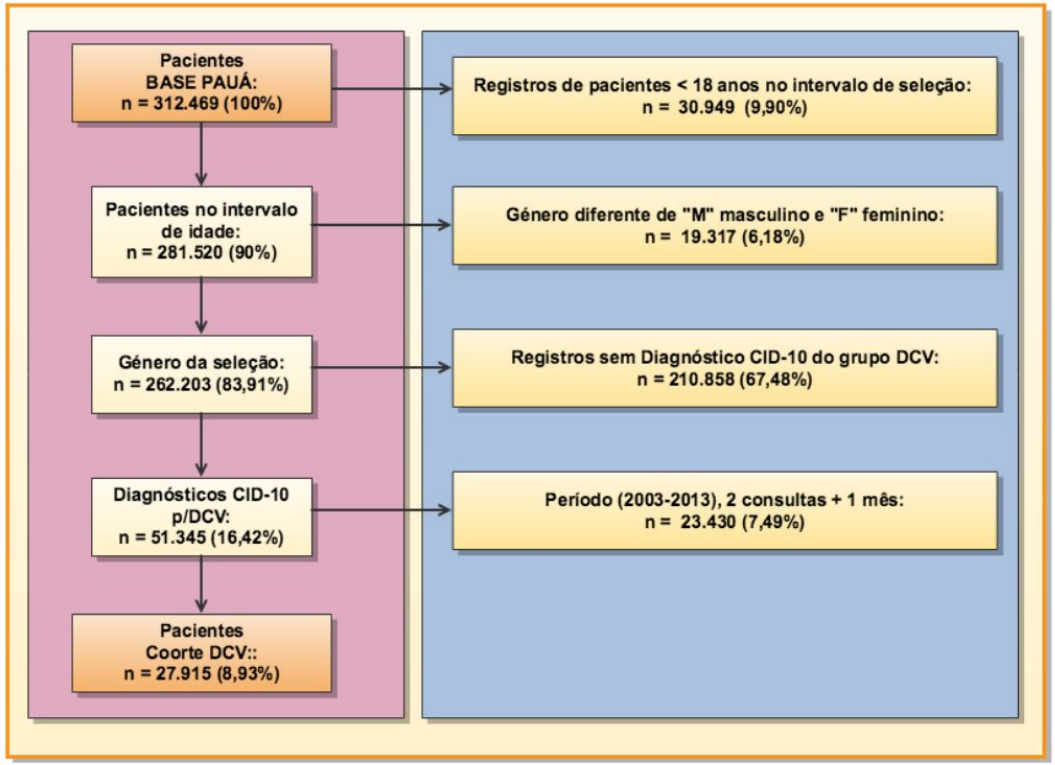
1. OHDSI tools (Achilles Heel):

   Referential integrity (FKs), Consistence, Missing data, Veracity

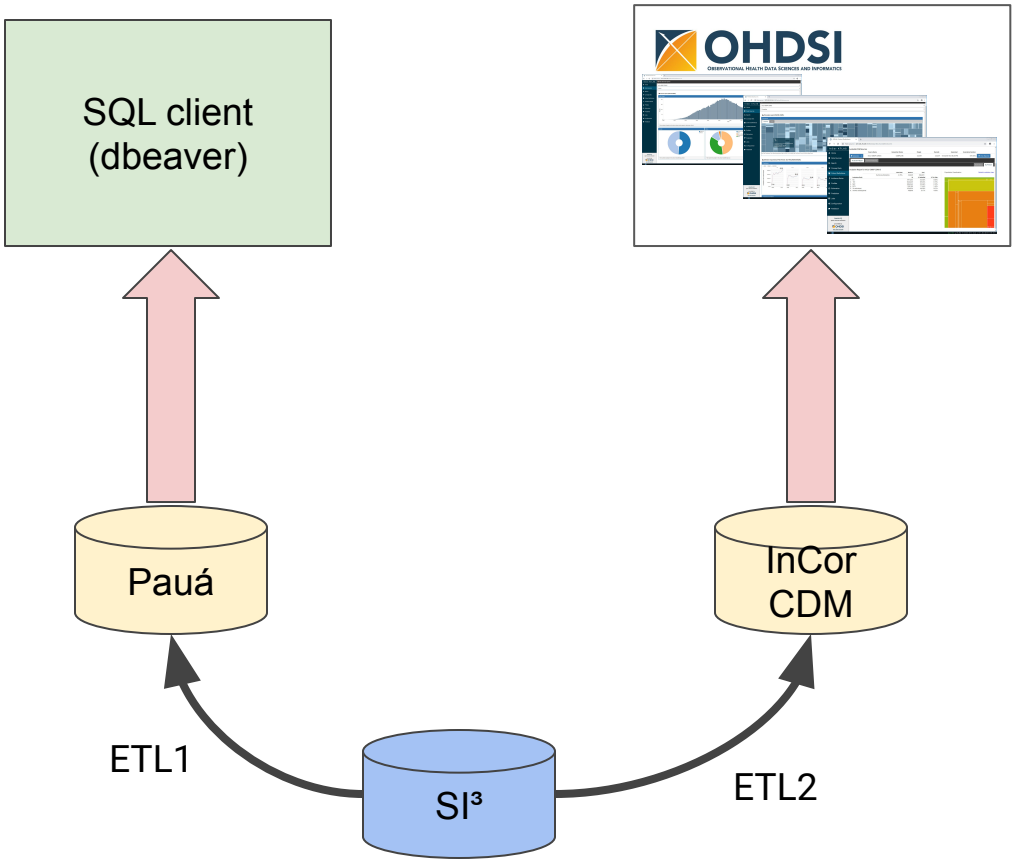
2. Reselect Abrahao et al (2010) CVD cohort

   Patients diagnosed with CardioVascular Disease and under treatment with statins.

CVD cohort



(Abrahao et al, 2010)


SQL client
(dbeaver)
OHDSI
Pauá
InCor
CDM
ETL1
SI³
ETL2

Varying parameters / thresholds

Table 5 – Varying condition start periods.

(years) Criteria	2003-2013	2000-2013	2000-2016
Initial		778,015	
Dx, 18+, M/F		303,847	
CVD	45,710	49,942	63,656
People	39,498	43,293	54,126

Table 6 – Varying 2nd visit event start after index.

(days) Criteria	All	365	180	90
Initial		778,015		
Dx, 18+, M/F		303,847		
CVD	45,710	44,228	43,950	43,667
People	39,498	35,457	32,767	29,414

Information retrieval statistics

Table 5 – Varying condition start periods.

Criteria (years)	2003-2013	2000-2013	2000-2016
Initial		778,015	
Dx, 18+, M/F		303,847	
CVD	45,710	49,942	63,656
People	39,498	43,293	54,126

Table 6 – Varying 2nd visit event start after index.

Criteria (days)	All	365	180	90
Initial		778,015		
Dx, 18+, M/F		303,847		
CVD	45,710	44,228	43,950	43,667
People	39,498	35,457	32,767	29,414

#	TPR	FPR	PPV	NPV	ACC	F1
1	.905	.041	.674	.990	.953	.772
2	.901	.040	.678	.990	.954	.774
3	.903	.040	.677	.990	.954	.774
4	.904	.041	.676	.990	.954	.773
5	.905	.041	.674	.990	.953	.772
6	.907	.052	.623	.990	.944	.738
7	.907	.052	.622	.990	.944	.738
8	.907	.052	.620	.990	.943	.736
9	.889	.040	.680	.990	.954	.775
10	.877	.031	.727	.988	.960	.795
11	.829	.027	.743	.983	.960	.784
12	.754	.023	.752	.976	.957	.753

ROC curve from previous table (AUC=0.938)

#	TPR	FPR	PPV	NPV	ACC	F1
1	.905	.041	.674	.990	.953	.772
2	.901	.040	.678	.990	.954	.774
3	.903	.040	.677	.990	.954	.774
4	.904	.041	.676	.990	.954	.773
5	.905	.041	.674	.990	.953	.772
6	.907	.052	.623	.990	.944	.738
7	.907	.052	.622	.990	.944	.738
8	.907	.052	.620	.990	.943	.736
9	.889	.040	.680	.990	.954	.775
10	.877	.031	.727	.988	.960	.795
11	.829	.027	.743	.983	.960	.784
12	.754	.023	.752	.976	.957	.753

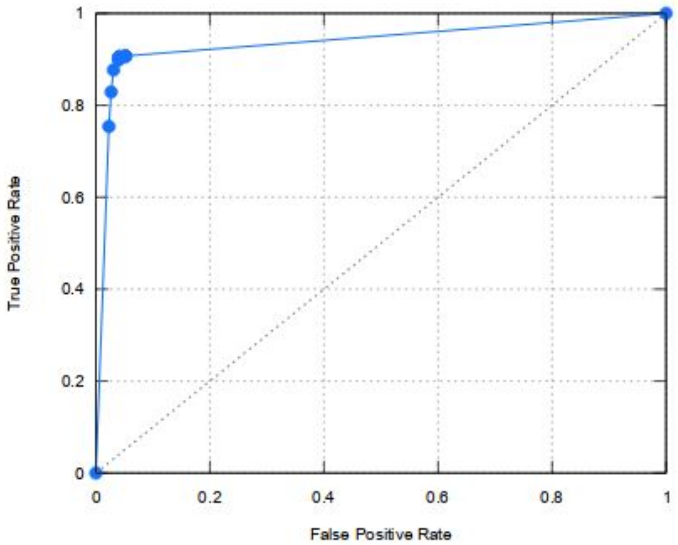


Figure 1 – Empirical ROC curve for Table 8 (AUC = 0.938).

Conclusion

Conclusion

- InCor-CDM allows cohort selection for clinical research
- Data quality can (and should) improve
  - Agreement with Abrahao et al 2012:
    - Precision = 62~75%
    - Recall = 75~91%
    - F1 = 74~80%
    - AUC = 0.938

Future work

Add complex features (e.g. PACS) to InCor-CDM

Evaluate quality at estimating population-level effects

Evaluate quality between different data mining techniques

Image: Poti River canyons

This project is supported CAPES, CNPq and FAPESP (grant #2018/11424-0).