Common Data Model Of Everything in Medicine:
Journey for integration of Environmental, Genomic data, Radiology, and Patient-Generated Health Data with clinical data in OMOP-CDM

Seng Chan You
Finding the missing link for big biomedical data
The identification of the degree of symmetry of an object or idea with the degree of perfection of that object or idea is both as old as the ancient Greeks and as new as the current ideas of modern physics.

From its beginnings in ancient astronomy, the goal of the science of physics has always been to find ‘the simple Theory Of Everything’

Symmetry in Mathematics

- A symmetry operation is a mathematical operation which leaves the final state indistinguishable from the initial state

Richard C. Morrison, PHYSICS: A Search for Simplicity, Beauty and Symmetry
Symmetry in Physics

- At the ultimate extreme of contraction - the instant of the "big bang," all particles and all forces would be indistinguishable.
- Only as the universe cools and expands do particles separate into quarks then into protons and neutrons, and the primordial single force splits into distinct gravitational, electromagnetic and nuclear forces.
- Modern physicists would like nothing better than to prove that the universe really does behave according to this model of "perfect symmetry."
Symmetry in medical data

- By grand unification across all aspects of health data, various types of medical data, such as clinical, genomic, radiologic, and patient-generated data, would be **indistinguishably accessible** in the single database.
- OHDSI tools ecosystem can work across various types of medical data.
Symmetry in medical data

- By grand unification across all aspects of health data, various types of medical data, such as clinical, genomic, radiologic, and patient-generated data, would be indistinguishably accessible in the single database.
- OHDSI tools ecosystem can work across various types of medical data.
Data are Like Lego Bricks for Phenotyping in CDM

- Conditions
- Drugs
- Procedures
- Measurements
- Observations
- Visits
Data are Like Lego Bricks for Phenotyping in CDM

- Conditions
- Drugs
- Procedures
- Measurements
- Observations
- Visits
- Genomic variants
- Radiology
- Topics from Free-Text
- Patient-Generated Health Data
- Environment
## OHDSI Tools Ecosystem

| **Cohort Method** | New-user cohort studies using large-scale regression for propensity and outcome models. |
| **Self-Controlled Case Series** | Self-Controlled Case Series analysis using few or many predictors, includes splines for age and seasonality. |
| **Self-Controlled Cohort** | A self-controlled cohort design, where time preceding exposure is used as control. |
| **IC Temporal Pattern Disc.** | A self-controlled design, but using temporal patterns around other exposures and outcomes to correct for time-varying confounding. |

| **Case-control** | Case-control studies, matching controls on age, gender, provider, and visit date. Allows nesting of the study in another cohort. |
| **Case-crossover** | Case-crossover design including the option to adjust for time-trends in exposures (so-called case-time-control). |

| **Patient Level Prediction** | Build and evaluate predictive models for user-specified outcomes, using a wide array of machine learning algorithms. |
| **Feature Extraction** | Automatically extract large sets of features for user-specified cohorts using data in the CDM. |

| **Empirical Calibration** | Use negative control exposure-outcome pairs to profile and calibrate a particular analysis design. |
| **Method Evaluation** | Use real data and established reference sets as well as simulations injected in real data to evaluate the performance of methods. |

| **Database Connector** | Connect directly to a wide range of database platforms, including SQL Server, Oracle, and PostgreSQL. |
| **Sql Render** | Generate SQL on the fly for the various SQL dialects. |
| **Cyclops** | Highly efficient implementation of regularized logistic, Poisson and Cox regression. |
| **Ohdsi R Tools** | Support tools that didn’t fit other categories, including tools for maintaining R libraries. |

---

**ATLAS**

*Under construction*
Common Data Model of Everything in Medicine

CDM Of Everything

- Clinical data
- Environment
- Patient-Generated Data
- Genomic data
- Radiology
- Unstructured Text
Common Data Model of Everything in Medicine

CDM Of Everything

- Clinical data
- Genomic data
- Radiology
- Unstructured Text
- Patient-Generated Data
- Environment
Exponential Growth in New Forms of Data Will Play an Increasing Important Role in Enabling Better Outcomes

Exogenous data
(Behavior, Socio-economic, Environmental, ...)

60% of determinants of health
Volume, Variety, Velocity, Veracity

Genomics data
30% of determinants of health
Volume

Clinical data
10% of determinants of health
Variety

1100 Terabytes
Generated per lifetime

6 TB
Per lifetime

0.4 TB
Per lifetime

Source: “The Relative Contribution of Multiple Determinants to Health Outcomes”, Lauren McGovern et al., Health Affairs, 33, no.2 (2014)
Environmental information and precision medicine

• We need to harness all of environmental, genetic, and clinical data to maximize personal and population health
  – “... the prevailing focus on an individual’s genes and biology insufficiently incorporates the important role of environmental factors in disease etiology and health”
  – “... a better understanding of the relationship between environmental exposure and the epigenome might lead to more efficient preventive measures”
  – “... embracing the impact of the environment on health will require a new framework to guide both research and its application, and to steer public investment and research efforts”
The definition of environment in medicine

• **Environment** is everything that is around us

• **Environmental medicine** is a multidisciplinary fields... Environmental factors can be classified into:
  – Physical
  – Chemical
  – Biological
  – Social (including Psychological and Culture variables)
  – Ergonomic
  – Safety
  – Any combination of the above

https://simple.wikipedia.org/wiki/Environment
https://en.wikipedia.org/wiki/Environmental_medicine
What is the environment in medicine?

- **Environment** is everything that is around us
  
- **Environmental medicine** is a multidisciplinary fields... Environmental factors can be classified into:
  
  - **Physical**: e.g. Weather
  - **Chemical**: e.g. Pollution
  - **Biological**: e.g. Zoonotic source (Lyme disease)
  - **Social**: e.g. Culture, Economic status
What is the environment in medicine?

• **Environment** is everything that is around us

• **Environmental medicine** is a multidisciplinary fields... Environmental factors can be classified into:
  
  – **Physical**: e.g. Weather
  – **Chemical**: e.g. Pollution
  – **Biological**: e.g. Zoonotic source (Lyme disease)
  – **Social**: e.g. Culture, Economic status

• All above are based on **Geographic Information System**
AEGIS- An open source spatial analysis tool based on CDM

Jaehyeong Cho, B.S.¹, Seng Chan You, M.D. M.S.², Kyehwon Kim, B.E.³, Doyeop Kim, B.E.², Rae Woong Park, M.D., Ph.D.¹,²

Dept. of Biomedical Sciences, Ajou University Graduate School of Medicine, Yeongtong-gu, Suwon

Dept. of Biomedical Informatics, Ajou University School of Medicine, Yeongtong-gu, Suwon

Yeungnam University Graduate school of Medicine, Nam-gu, Daegu
• AEGIS development

  • AEGIS: Application for Epidemiological Geographic Information System

  • A tool to conduct disease mapping and cluster analysis considering age and gender-adjustment and spatial autocorrelation using GIS database based on CDM

  • AEGIS is open-source software, which is harmonized within OHDSI eco-system
Based on the Global Administrative Database (GADM), AEGIS can depict cohorts on the map according to the country’s own administrative district.

**Experiment:** Disease map of Vascular disease in Korea and Massachusetts, US

**Result**

- **Patient with vascular disorder in Korea**
- **Patient with vascular disease in Massachusetts**
Based on Global Administrative Database (GADM), AEGIS can depict cohorts on the map according to the country’s own administrative district.
Experiment 2: Among asthma patients, regional variations in patients who visited the emergency department due to asthma

- Clustering of emergency department visit due to asthma among patients with asthma
Experiment 2: Among asthma patients, regional variations in patients who visited the emergency department due to asthma

Result

4 Identification of Disease Cluster

- Association of **Asthma Exacerbation and Air pollution**

They don’t seem to be correlated

Air pollution map in Korea (PM-10)

NATIONAL INSTITUTE OF ENVIRONMENTAL RESEARCH, 2018
Experiment 2: Among asthma patients, regional variations in patients who visited the emergency department due to asthma

4 Identification of Disease Cluster

- Association of Asthma Exacerbation and House price

They seem to be correlated!
Common Data Model of Everything in Medicine

- Clinical data
- Genomic data
- Radiology
- Unstructured Text
- Environment
- Patient-Generated Data

Seng Chan You, MD¹, Youngin Kim, MD², Jaehyung Cho¹, Rae Woong Park, MD, PhD¹,³

¹Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea; ²Medicine, Noom, Inc, Seoul, Korea; ³Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea
Patient-Generated Health Data

Exponential Growth in New Forms of Data Will Play an Increasing Important Role in Enabling Better Outcomes

Exogenous data
(Behavior, Socio-economic, Environmental, ...)

60% of determinants of health
Volume, Variety, Velocity, Veracity

Genomics data
30% of determinants of health
Volume

Clinical data
10% of determinants of health
 Variety

1100 Terabytes
Generated per lifetime

6 TB
Per lifetime

0.4 TB
Per lifetime

Source: “The Relative Contribution of Multiple Determinants to Health Outcomes”, Lauren McGover et al., Health Affairs, 33, no.2 (2014)
Applications in smartphone collecting health data

Apple Health

Google Fit

Samsung Medical Center

Diabetes Note

NOOM

Efil
Basic concept for standardization of patient generated health data

- **Data Sources**
  - **Measuring**
    - Phone / Wearable / medical device / Report
  - **SmartPhone**
    - iOS: AppleHealth
    - Android: GoogleFit, S-Health
  - **Third-party Applications**
    - Samsung Medical Center: Diabetes Note
    - NOOM
    - Life Semantics: Efil

- **CDM Database Schema**
  - OMOP-CDM
Basic concept for standardization of patient generated health data

- Data Sources
  - Measuring
    - Phone / Wearable / medical device / Report
  - SmartPhone
    - iOS: AppleHealth
    - Android: GoogleFit, S-Health
  - Third-party Applications
    - Samsung Medical Center: Diabetes Note
    - NOOM
    - Life Semantics: Efil
- CDM Database Schema
  - OMOP-CDM
Start PGHD Working Group in OHDSI

Patient Generated Health Data (PGHD) Working Group

1 3d

SCYou Seng Chan You

Dear colleagues,

I would like to propose to start Patient Generated Health Data (PGHD) Working Group. The goal of this WG would be developing ETL conventions, integration process with clinical data, and analytic process for PGHD, which is generated through Smart Phone/App/Wearable devices.

I've released the sample for PGHD, which was generated by QS app of iPhone (sample_data)

The primitive ETL convention for this data is released, too (PGHD_ETL_convention)

Please join if you're interested in this topic.
@yipaulkim @Wonchul

Wonchul Wonchul Cha

Great work Seungchan! Let's make some progress! 😊

2 Likes

Rijnbeek Peter Rijnbeek

Hi Chan,

Interesting. Within our upcoming European project EHDEN there is some work planned in this direction. Also in Europe the Radar project https://www.imi.europa.eu/projects-results/project-factsheets/radar-cns has a focus on collecting data from wearables and they are looking into OMOP-CDM to host it.

http://forums.ohdsi.org/t/patient-generated-health-data-pghd-working-group/4612
Data types in PGHD

1. Activity
   - Steps, Flight climbed, Distance
2. Nutrition
   - Calorie intake (24hr / breakfast, lunch, dinner)
   - Nutrients
3. Sleep
   - Total minutes / Minutes asleep, Time to fall sleep, Number of sleep periods
4. Body measurements
   - Height, Weight, BMI, Lean body, Body fat
5. Vital signs
   - HR, BP, ...
6. Self-medication
   - Insulin
7. Laboratory measurement
   - Glucose
8. Self-report
9. Mindfulness
# Granularities of Data in PGHD

## Macro-level

1. **Activity**  
   - Steps, Flight climbed, Distance

2. **Nutrition**  
   - Calorie intake (24hr / breakfast, lunch, dinner)  
   - Nutrients

3. **Sleep**  
   - Total minutes / Minutes asleep, Time to fall sleep, Number of sleep periods

4. **Body measurements**  
   - Height, Weight, BMI, Lean body, Body fat

5. **Vital signs**  
   - HR, BP, ...

6. **Self-medication**  
   - Insulin

7. **Laboratory measurement**  
   - Glucose

8. **Self-report**

9. **Mindfulness**

## Micro-level

1. **Activity**  
   - Acceleration, Angular velocity unit value (GyroMeter)

2. **Nutrition**  
   - Temporal relationship to meal

3. **Sleep**  
   - Temporal relationship to sleep, REM/non-REM sleep

4. **Body measurements**  
   - Body location, Body posture, Ventilation cycle time

5. **Self-report**  
   - Ambient temperature, Geoposition, Magnetic force
## ETL convention for macro-level PGHD

<table>
<thead>
<tr>
<th>PGHD Types</th>
<th>Source Value</th>
<th>Domain</th>
<th>Event_ID</th>
<th>Concept_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Activity</strong></td>
<td>Steps 1</td>
<td>OBSERVATION</td>
<td>1</td>
<td>3034985</td>
</tr>
<tr>
<td></td>
<td>Flight climbed 2</td>
<td>OBSERVATION</td>
<td>2</td>
<td>4121036</td>
</tr>
<tr>
<td></td>
<td>Distance 3</td>
<td>OBSERVATION</td>
<td>3</td>
<td>3031111</td>
</tr>
<tr>
<td></td>
<td>Active Calories 4</td>
<td>OBSERVATION</td>
<td>4</td>
<td>3032128</td>
</tr>
<tr>
<td><strong>Nutrition</strong></td>
<td>Dietary Calories 5</td>
<td>OBSERVATION</td>
<td>5</td>
<td>4037128</td>
</tr>
<tr>
<td></td>
<td>Nutrients</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sleep</strong></td>
<td>Sleep start 1</td>
<td>CONDITION_OCCURRENCE</td>
<td>1</td>
<td>4086839</td>
</tr>
<tr>
<td></td>
<td>Sleep end 1</td>
<td>CONDITION_OCCURRENCE</td>
<td>1</td>
<td>4086839</td>
</tr>
<tr>
<td></td>
<td>Minutes asleep</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time to fall sleep</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of sleep periods</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total sleep minutes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Body measurement</strong></td>
<td>Weight 1</td>
<td>MEASUREMENT</td>
<td>1</td>
<td>3025315</td>
</tr>
<tr>
<td></td>
<td>BMI 2</td>
<td>MEASUREMENT</td>
<td>2</td>
<td>3032843</td>
</tr>
<tr>
<td></td>
<td>Lean Body Mass 3</td>
<td>MEASUREMENT</td>
<td>3</td>
<td>3010914</td>
</tr>
<tr>
<td></td>
<td>Body Fat Percentage</td>
<td>MEASUREMENT</td>
<td>4</td>
<td>3012888</td>
</tr>
<tr>
<td></td>
<td>Body Temperature</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Vital signs</strong></td>
<td>Heart Rate 5</td>
<td>MEASUREMENT</td>
<td>5</td>
<td>3028737</td>
</tr>
<tr>
<td></td>
<td>Blood Pressure (Systolic) 6</td>
<td>MEASUREMENT</td>
<td>6</td>
<td>3038553</td>
</tr>
<tr>
<td></td>
<td>Blood Pressure (Diastolic) 7</td>
<td>MEASUREMENT</td>
<td>7</td>
<td>4239408</td>
</tr>
<tr>
<td></td>
<td>Respiratory Rate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Self-medication</strong></td>
<td>Insulin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inhaler Usage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Laboratory measurement</strong></td>
<td>Blood Glucose</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
NOOM converted their data into CDM

Noom is a behavior change company that uses A.I., Human Coaching and Mobile Technology to create the world's most effective solutions for lifestyle & chronic conditions.

- Use latest psychological approaches (CBT)
- Patented scalable coaching using A.I.
- Mobile-first company
- 47 million users

A.I. 3.1 billion coaching data points
Self-learning A.I.

Mobile

Human Coaching
NOOM converted their data into CDM

Noom Solution: Effective & Scalable Behavior Change Courses

What the user sees
- 100% mobile, interactive & customized courses renewing every 2 - 8 months
- Dedicated personal & group coach for each user
- Best-in-class tools like 3.7M Food DB with predictive search
- Durable results: 84% who start, complete; 60% keep off lost weight a year later

Behind the scenes
- AI-enabled coaching tools
- Proprietary coach dashboard
- 401 coaches worldwide (90% remote)
- Virtual clinical supervision & Noomiversity
- 3.1 billion virtual & human coaching data points (causal data)

1 One-year follow-up data; published in JMIR 2018;8(5):e93
ETL result of sample data from NOOM

- NOOM converted their sample data (n=100) into CDM
  - weight, daily step count, and daily dietary calories

<table>
<thead>
<tr>
<th>measurement_id</th>
<th>person_id</th>
<th>measurem</th>
<th>value_source</th>
<th>unit_source</th>
<th>measurement_concept_name</th>
<th>measurement_datetime</th>
<th>measurement_date</th>
<th>value_as_number</th>
<th>unit_concept_value</th>
<th>unit_concept_datetime</th>
<th>observation_id</th>
<th>observation_source_value</th>
<th>value_source</th>
<th>unit_source</th>
<th>observation_concept_name</th>
<th>observation_date</th>
<th>value_as_number</th>
<th>unit_concept_value</th>
<th>unit_concept_datetime</th>
<th>observation_type_concept_name</th>
<th>observation_data_type</th>
<th>observation_data_type_code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Weight</td>
<td>103.4</td>
<td>kg</td>
<td>3025315 Body weight</td>
<td>2017-05-08</td>
<td>2017-05-08 22:56</td>
<td>103.4</td>
<td>4122383 kg</td>
<td>44818704</td>
<td>1</td>
<td>Steps</td>
<td>9097</td>
<td>count</td>
<td>Number of steps in 24 hour Measured</td>
<td>2017-07-04</td>
<td>9348</td>
<td>44777556 per 24 hours</td>
<td>App generated</td>
<td>44814721/App generated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Weight</td>
<td>108</td>
<td>kg</td>
<td>3025315 Body weight</td>
<td>2017-03-22</td>
<td>2017-03-23 10:27</td>
<td>105</td>
<td>4122383 kg</td>
<td>44818704</td>
<td>2</td>
<td>Steps</td>
<td>1600</td>
<td>count</td>
<td>Number of steps in 24 hour Measured</td>
<td>2017-04-24</td>
<td>1519</td>
<td>44777556 per 24 hours</td>
<td>App generated</td>
<td>44814721/App generated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Weight</td>
<td>109</td>
<td>kg</td>
<td>3025315 Body weight</td>
<td>2017-03-04</td>
<td>2017-03-04 9:46</td>
<td>106.7</td>
<td>4122383 kg</td>
<td>44818704</td>
<td>3</td>
<td>Steps</td>
<td>7200</td>
<td>count</td>
<td>Number of steps in 24 hour Measured</td>
<td>2017-05-15</td>
<td>7269</td>
<td>44777556 per 24 hours</td>
<td>App generated</td>
<td>44814721/App generated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>2</td>
<td>Weight</td>
<td>69.9</td>
<td>kg</td>
<td>3025315 Body weight</td>
<td>2017-07-11</td>
<td>2017-07-11 9:30</td>
<td>69.9</td>
<td>4122383 kg</td>
<td>44818704</td>
<td>32</td>
<td>Steps</td>
<td>70</td>
<td>count</td>
<td>Number of steps in 24 hour Measured</td>
<td>2018-04-26</td>
<td>65.8</td>
<td>44777556 per 24 hours</td>
<td>App generated</td>
<td>44814721/App generated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>2</td>
<td>Weight</td>
<td>69.8</td>
<td>kg</td>
<td>3025315 Body weight</td>
<td>2018-02-28</td>
<td>2018-02-28 9:24</td>
<td>69.8</td>
<td>4122383 kg</td>
<td>44818704</td>
<td>33</td>
<td>Steps</td>
<td>69.8</td>
<td>count</td>
<td>Number of steps in 24 hour Measured</td>
<td>2018-02-28</td>
<td>9472</td>
<td>44777556 per 24 hours</td>
<td>App generated</td>
<td>44814721/App generated</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>observation_id</th>
<th>observation_source_value</th>
<th>value_source</th>
<th>unit_source</th>
<th>observation_concept_name</th>
<th>observation_date</th>
<th>value_as_number</th>
<th>unit_concept_value</th>
<th>unit_concept_datetime</th>
<th>observation_type_concept_name</th>
<th>observation_data_type</th>
<th>observation_data_type_code</th>
</tr>
</thead>
<tbody>
<tr>
<td>9147</td>
<td>19 Dietary Calories</td>
<td>1598000 calorie</td>
<td>4037128 Dietary calorie intake</td>
<td>2018-04-03</td>
<td>1498000</td>
<td>9472 calorie</td>
<td>44777556 per 24 hours</td>
<td>Patient reported</td>
<td>44814721</td>
<td>Patient reported</td>
<td></td>
</tr>
<tr>
<td>9148</td>
<td>19 Dietary Calories</td>
<td>1186000 calorie</td>
<td>4037128 Dietary calorie intake</td>
<td>2018-04-04</td>
<td>1176000</td>
<td>9472 calorie</td>
<td>44777556 per 24 hours</td>
<td>Patient reported</td>
<td>44814721</td>
<td>Patient reported</td>
<td></td>
</tr>
<tr>
<td>9149</td>
<td>19 Dietary Calories</td>
<td>1772000 calorie</td>
<td>4037128 Dietary calorie intake</td>
<td>2018-04-05</td>
<td>1672000</td>
<td>9472 calorie</td>
<td>44777556 per 24 hours</td>
<td>Patient reported</td>
<td>44814721</td>
<td>Patient reported</td>
<td></td>
</tr>
<tr>
<td>9150</td>
<td>19 Dietary Calories</td>
<td>1329000 calorie</td>
<td>4037128 Dietary calorie intake</td>
<td>2018-04-06</td>
<td>1309000</td>
<td>9472 calorie</td>
<td>44777556 per 24 hours</td>
<td>Patient reported</td>
<td>44814721</td>
<td>Patient reported</td>
<td></td>
</tr>
</tbody>
</table>
Basic concept for standardization of patient generated health data

• Development of PGHD ETL convention
  – Macro-level Data: Convert PGHD of each data source into conventional OMOP-CDM by the ETL guidance
  – Micro-level Data: Add new extension model (tables) to OMOP-CDM
  – Extract converted PGHD from 3rd-party apps

• Integration of PGHD from and EHR
  – Send PGHD data (CDM) from IT company to the hospital when patients approves it
  – PGHD will be integrated with EHR data in the format of CDM

• Analytic Tool
  – Development of Visualization tool for Time-Series data
  – Development of Standardized Time-Series Analysis Tool

• Ultimate goal
  – Clinicians can utilize integrated PGHD data in their practice
Common Data Model of Everything in Medicine

Seng Chan You, MD, MS¹, Kwang Soo Jeong¹, Si Hyung No², Kwon-Ha Yoon, MD, PhD³, Chang-Won Jeong, PhD², Rae Woong Park, MD, PhD¹,⁴

¹Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea;
²Imaging Science based Lung and Bone Disease Research Center, Wonkwang University, Iksan, Korea;
³Department of Radiology, Wonkwang University College of Medicine
⁴Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea
Why do we need CDM extension for Radiology (R-CDM)?

Oncology radiology imaging integration into CDM

Patrick Ryan

Team: I'm in Sweden right now, they've got some exciting research going on that involves linking various national registries (including prescription, hospitalization, and cancer) with a new dataset that pulls out radiology images of tumor sites, that can then be used for predictive modeling via deep learning and other algorithms. The team at Karolinska Institute have already demonstrated successful ETL for most of the registers, but as a community, we don't yet have a common solution for storing the imaging files and whatever associated records to link to them. Has anyone in the community worked on this problem, whether it be for oncology or for other areas? @Rijnbeek, does the work you've led in EKG imaging have some applicability here?
Collaborative and Reproducible Research using Radiology data

- Combining imaging biomarkers with genomic and clinical phenotype data is the foundation of precision medicine research efforts.

- **Current image studies are scattered** across numerous archives, hindering collaborative and reproducible research using radiology data.

- By definition, **reproducible science** requires being able to reproduce results. *Without access to another researcher’s code and data, there is no way a third party can duplicate that researcher’s results. Github and Docker vastly lower the learning curve required to share code and runtime environments-for those who want to.* What they do no address is the **commonality of dataset.**
Basic concept for standardization of radiology data (R-CDM)

- Most of radiologic images are stored in **DICOM** (Digital Imaging and Communications in Medicine) format
  - DICOM provides a standard for medical image storage and a set of network operations for transmission and retrieval
  - DICOM file contains required and optional **metadata** fields: patient ID, row, columns (pixel), modality, manufacturer, phase, etc.

### Table 1: Examples of commonly available metadata

<table>
<thead>
<tr>
<th>Element</th>
<th>Source</th>
<th>Example</th>
<th>Storage location</th>
</tr>
</thead>
<tbody>
<tr>
<td>PatientsName</td>
<td>EHR/ADT</td>
<td>MARY^JONES^B</td>
<td>DICOM header</td>
</tr>
<tr>
<td>PatientID</td>
<td>EHR/ADT</td>
<td>1232391-3</td>
<td>DICOM header</td>
</tr>
<tr>
<td>StudyDescription</td>
<td>RIS</td>
<td>CT BRAIN W/O</td>
<td>DICOM header</td>
</tr>
<tr>
<td>Rows</td>
<td>Imaging modality</td>
<td>512</td>
<td>DICOM header</td>
</tr>
<tr>
<td>Columns</td>
<td>Imaging modality</td>
<td>512</td>
<td>DICOM header</td>
</tr>
<tr>
<td>BitsStored</td>
<td>Imaging modality</td>
<td>12</td>
<td>DICOM header</td>
</tr>
</tbody>
</table>

Basic concept for standardization of radiology data (R-CDM)

• Why do we need R-CDM if we have DICOM?
  – In practice, data fields in DICOM are often filled incorrectly or left blank
  – Study description heterogeneity between institutions (eg, ‘brain CT’, ‘CT brain’, ’CT brain non-contrast’, etc.)
    • We need standard vocabulary and map local study description to the standard vocabulary for radiology.
  – De-identified datasets of DICOM may result in the removal of metadata that is required for advanced processing
Ontology for R-CDM

• **LOINC RSNA radiology playbook:** Unified terminology of RadLex and LOINC
  - **RadLex** is a comprehensive lexicon of radiology terms for indexing and retrieval of radiology information resources, specifically aimed at representing clinical content associated with radiology reports
  - RadLex has been incorporated into LOINC, and OMOP vocabulary!

*Journal of the American Medical Informatics Association, 25(7), 2018, 885–893*

doi: 10.1093/jamia/ocy053
Advance Access Publication Date: 29 May 2018
Research and Applications

Research and Applications

**The LOINC RSNA radiology playbook - a unified terminology for radiology procedures**

Daniel J Vreeman,1,2 Swapna Abhyankar,1 Kenneth C Wang,3,4 Christopher Carr,5 Beverly Collins,6 Daniel L Rubin,7,8 Curtis P Langlotz8
Basic concept for standardization of radiology data (R-CDM)

- **MetaData** and Path of images are stored in two tables
  - Radiology_Occurrence: each row represents single radiologic procedure
    - Device, Modality(CT/MRI,...), Total image counts, Radiology dosages, path, and etc.
  - Radiology_Image: each row represents single image from radiologic procedure
    - Phase (Non-contrast/contrast), Image number, pixel data, path, and etc.

---

### Radiology_Occurrence

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK</td>
<td>Radiology_occurrence_ID</td>
</tr>
<tr>
<td>FK, N, FK</td>
<td>Device_concept_id</td>
</tr>
<tr>
<td>FK</td>
<td>Radiology_modality_concept_id</td>
</tr>
<tr>
<td>FK</td>
<td>Radiology_protocol_concept_id</td>
</tr>
<tr>
<td>FK</td>
<td>Image_total_count</td>
</tr>
<tr>
<td>FK</td>
<td>Anatomic_site_concept_id</td>
</tr>
<tr>
<td>FK</td>
<td>Radiology_comment</td>
</tr>
<tr>
<td>FK</td>
<td>Dosage_value_number</td>
</tr>
<tr>
<td>FK</td>
<td>Image_exposure_time_unit_concept_id</td>
</tr>
<tr>
<td>FK</td>
<td>Image_exposure_time</td>
</tr>
<tr>
<td>FK</td>
<td>Radiology_dipath</td>
</tr>
<tr>
<td>FK</td>
<td>Visit_occurrence_id</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Radiology_occurrence_date</td>
</tr>
<tr>
<td>N</td>
<td>Radiology_occurrence_datetime</td>
</tr>
<tr>
<td>N</td>
<td>Person_ID</td>
</tr>
<tr>
<td>N</td>
<td>Condition_occurrence_ID</td>
</tr>
<tr>
<td>N</td>
<td>Image_value_number</td>
</tr>
</tbody>
</table>

### Radiology_Image

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK</td>
<td>Image_ID</td>
</tr>
<tr>
<td>FK</td>
<td>Source_ID</td>
</tr>
<tr>
<td>FK</td>
<td>Radiology_occurrence_ID</td>
</tr>
<tr>
<td>FK</td>
<td>Person_ID</td>
</tr>
<tr>
<td>FK</td>
<td>Person_orientation_concept_id</td>
</tr>
<tr>
<td>FK</td>
<td>Image_type</td>
</tr>
<tr>
<td>FK</td>
<td>Radiology_phase_concept_id</td>
</tr>
<tr>
<td>FK</td>
<td>Image_no</td>
</tr>
<tr>
<td>FK</td>
<td>Phase_total_no</td>
</tr>
<tr>
<td>FK</td>
<td>Image_resolution_rows</td>
</tr>
<tr>
<td>FK</td>
<td>Image_resolution_columns</td>
</tr>
<tr>
<td>FK</td>
<td>Image_Window_Level_Center</td>
</tr>
<tr>
<td>FK</td>
<td>Image_Window_Level_Width</td>
</tr>
<tr>
<td>FK</td>
<td>Image_slice_thickness</td>
</tr>
<tr>
<td>FK</td>
<td>Image_filepath</td>
</tr>
</tbody>
</table>
Basic concept for standardization of radiology data (R-CDM)

Common Data Model of Everything in Medicine

Seo Jeong Shin, MS¹, Seng Chan You, MD, MS¹, Jin Roh, MD, PhD², Rae Woong Park, MD, PhD¹, ³

¹Dept. of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea; ²Dept. of Pathology, Ajou University Hospital, Suwon, South Korea; ³Dept. of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, South Korea
Exponential Growth in New Forms of Data Will Play an Increasing Important Role in Enabling Better Outcomes

*Exogenous* data
(Behavior, Socio-economic, Environmental, ...)

60% of determinants of health
*Volume, Variety, Velocity, Veracity*

1100 Terabytes
Generated per lifetime

**Genomics** data
30% of determinants of health
*Volume*

**Clinical** data
10% of determinants of health
*Variety*

Source: "The Relative Contribution of Multiple Determinants to Health Outcomes", Lauren McGover et al., Health Affairs, 33, no.2 (2014)

IBM Health and Social Programs Summit | #IBMHSPS14 | #smartercare | #socialprograms
Background: Surge of genomic data

- Global waves of ‘precision medicine’
  - Precision medicine initiative in US: Population of 1M, $215M
  - Precision medicine initiative in China

- Insurance coverage of NGS in Korea
  - Since March 2017, national insurance coverage for targeted
    NGS in cancer patients has started in Korea.
  - No. of target genes
    - Level 1: 5~50 (cost paid by the patient: $450)
    - Level 2: 51~ (cost paid by the patient: $640)

- Despite much progress, genomic and clinical data are
  still generally collected and studies in silos, in individual
  institutions, or individual nations
Background: Surge of genomic data

- Collaborative research platform for genomic data in Oncology
Development of G-CDM based ISO standard

• ISO (International Organization for Standardization): a worldwide federation of national standards bodies

• Scope of this document (ISO/TS 20428)
  – Genetic variation from **human sample**
  – Whole genome sequencing, whole exome sequencing, targeted sequencing with **NGS** (not including Sanger)
  – **Clinical** application (eg, clinical trial, translational; not including basic or other area research)
Brief review: G-CDM

1. Sequencing
   - Each row represents each sequencing
   - Linking Clinical Information
   - Sequencing Process
     (Patient, Pathologic Diagnosis, Tumor Stage, Somatic/Germ-line, Sequencer, Reference Genome, Alignment Library, Quality Score etc.)

2. Variant_occurrence
   - Each row represents each variant
   - Structural / Functional variant classification
   - HGVS Nomenclature
   - Quality Score

3. Variant_annotation
   - Each row represents each annotation
   - Flexibility for any annotation tool

Taken from: http://www.broadinstitute.org/gatk/guide/best-practices
Brief review: G-CDM

- Overall, three tables are added
- Priority: compatibility with existing OMOP-CDM and OHDSI tools (eg Feature Extraction / Patient Level Prediction package)
- Sequencing table
  - Each row represents each sequencing (multiple sequencing is possible for same specimen of same patients)
  - Foreign keys (person, specimen, procedure, note, device)
  - Sequencing process (sequencer, reference genome, library for alignment, QC, …)
- Variant_occurrence table
  - Each row represents each variant (SNP, insertion, deletion, translocation, CNV)
  - Chromosome / Position (1st and 2nd for translocation/CNV)
  - HGVS nomenclature (according to the ISO)
  - Quality
- Variant_annotation table
  - Each row represent each secondary information resulted from variable annotation library for variant on variants (eg, clinical implication / eg, gnomAD, ClinVar, COSMIC)
  - Flexibility for any annotation tool (like Measurement table)
Relationship between G-CDM and OMOP-CDM
The data structures of the two institutes were unified.
Study Results:
Waterfall plot of adenocarcinoma and squamous cell carcinoma of lung
Study Results:
Waterfall plot of adenocarcinoma and squamous cell carcinoma of lung

**TCGA**

**Actionable Variant Proportion**

<table>
<thead>
<tr>
<th>Comparison (TCGA)</th>
<th>Gene:HGVS</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>KRAS:G12V</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>KRAS:G12D</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>KRAS:G12C</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>EGFR:E746_A750</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>EGFR:L858R</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>BRAF:V600E</td>
<td>0.020</td>
<td></td>
</tr>
</tbody>
</table>

**AJOU**

**Actionable Variant Proportion**

<table>
<thead>
<tr>
<th>Comparison (AJOU)</th>
<th>Gene:HGVS</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>KRAS:G12V</td>
<td>0.537</td>
<td></td>
</tr>
<tr>
<td>KRAS:G12D</td>
<td>0.764</td>
<td></td>
</tr>
<tr>
<td>KRAS:G12C</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>EGFR:E746_A750</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>EGFR:L858R</td>
<td>0.768</td>
<td></td>
</tr>
<tr>
<td>BRAF:V600E</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

LUAD (n=569)  LUSC (n=444)
LUAD (n=51)   LUSC (n=16)
Future plans for Oncology

• Converting whole cancer patients data from National Insurance Claim data

**Cancer statistics across OHDSI networks: ONCO-ACHILLES**

Dear colleagues,

As I mentioned earlier, we decided to convert whole Korean cancer patients data into CDM from National Insurance data (2007-2017).

Hi everyone, We’re planning to convert whole Korean cancer patients data into CDM from National Insurance data of HIRA (Korean national insurance data covers almost 99% population of Korea. This insurance covers 95% of cancer-related claim (If the patients should pay 100$ for the treatment, it covers 95$)). Then, we can run @rchens’s treatment pattern in cancer patient on much bigger data. We’ll perform descriptive analysis about incidence, overall survival and the whole cost within 1, 3 and 5 …

I will extract three components of information from this as the first research:

1. Quarterly incidence of each cancer from 2008-2017 according to the birth year (5-year base) and sex (and hopefully ethnic groups)
2. All-cause mortality within 1-year, 3-year and 5-year after cancer diagnosis from 2008-2017 in these quarterly cohorts according to the birth year and sex (and ethnic group)
3. Whole medical expenditure, cost amount paid by insurer, cost amount paid by the patients within 1-month, 6-months, 1-year, 3-year and 5-year after cancer diagnosis from 2008-2017 in these quarterly cohorts
Onco-Achilles

• Converting whole cancer patients data from National Insurance Claim data
  – Quarterly incidence of each cancer from 2008-2017 according to the birth year (5-year base) and sex (and hopefully ethnic groups)
  – All-cause mortality within 1-year, 3-year and 5-year after cancer diagnosis from 2008-2017 in these quarterly cohorts according to the birth year and sex (and ethnic group)
  – Whole medical expenditure, cost amount paid by insurer, cost amount paid by the patients within 1-month, 6-months, 1-year, 3-year and 5-year after cancer diagnosis from 2008-2017 in these quarterly cohorts according to birth year and sex.
Finding the missing link for big biomedical data

Figure. The Tapestry of Potentially High-Value Information Sources That May be Linked to an Individual for Use in Health Care

Probabilistic linkage to validate existing data or fill in missing data

Examples of biomedical data
- Pharmacy data
- Health care center (electronic health record) data
- Claims data
- Registry or clinical trial data
- Data outside of health care system

Ability to link data to an individual
- Easier to link to individuals
- Harder to link to individuals
- Only aggregate data exists

Data quantity
- More
- Less

Weber et al., “Finding the Missing Link for Big Biomedical Data.” JAMA (June 25, 2014)
Data are Like Lego Bricks for Phenotyping

- Conditions
- Drugs
- Procedures
- Measurements
- Observations
- Visits

- Genomic variants
- Radiology
- Topics from Free-Text
- Patient-Generated Health Data
- Environment
OHDSI Tools Ecosystem

- **Cohort Method**: New-user cohort studies using large-scale regression for propensity and outcome models.
- **Self-Controlled Case Series**: Self-Controlled Case Series analysis using few or many predictors, includes splines for age and seasonality.
- **Self-Controlled Cohort**: A self-controlled cohort design, where time preceding exposure is used as control.
- **IC Temporal Pattern Disc.**: A self-controlled design, but using temporal patterns around other exposures and outcomes to correct for time-varying confounding.
- **Case-control**: Case-control studies, matching controls on age, gender, provider, and visit date. Allows nesting of the study in another cohort.
- **Case-crossover**: Case-crossover design including the option to adjust for time-trends in exposures (so-called case-time-control).
- **Patient Level Prediction**: Build and evaluate predictive models for user-specified outcomes, using a wide array of machine learning algorithms.
- **Feature Extraction**: Automatically extract large sets of features for user-specified cohorts using data in the CDM.
- **Empirical Calibration**: Use negative control exposure-outcome pairs to profile and calibrate a particular analysis design.
- **Method Evaluation**: Use real data and established reference sets as well as simulations injected in real data to evaluate the performance of methods.
- **Database Connector**: Connect directly to a wide range of database platforms, including SQL Server, Oracle, and PostgreSQL.
- **Sql Render**: Generate SQL on the fly for the various SQL dialects.
- **Cyclops**: Highly efficient implementation of regularized logistic, Poisson and Cox regression.
- **Ohdsi R Tools**: Support tools that didn’t fit other categories, including tools for maintaining R libraries.
**Symmetry** in medical data

- By grand unification across all aspects of health data, various types of medical data would be **indistinguishably accessible** in the single database
- OHDSI tools ecosystem can work across various types of medical data
OHDSI: A Journey for Simplicity, Beauty and **Symmetry** in Medical Data

"Then you will know the truth, and the truth will set you free."  
John 8:32
Status of Korean OHDSI Network

Data Network of 41 Hospitals, 55M Patients

Seoul

Incheon / Gyeonggi

Chungcheong

Gangwon

Jeolla

Gyeongsang
I need your help!

- The **Scientific Revolution** has not been a revolution of knowledge. It has been above all a **revolution of ignorance**. The great discovery that launched the Scientific Revolution was the discovery that **humans do not know the answers to their most important questions** (Yuval Harari, A Brief history of Humankind, Ch14. Ignoramus).

- Understanding human history in the millennia following the **Agricultural Revolution** boils down to a single question: **how did humans organise themselves in mass-cooperation networks**, when they lacked the biological instincts necessary to sustain such networks? (Yuval Harari, A Brief history of Humankind, Ch8. There is No Justice in History)
Thank You for your time