OHDSI
Observational Health Data Sciences and Informatics

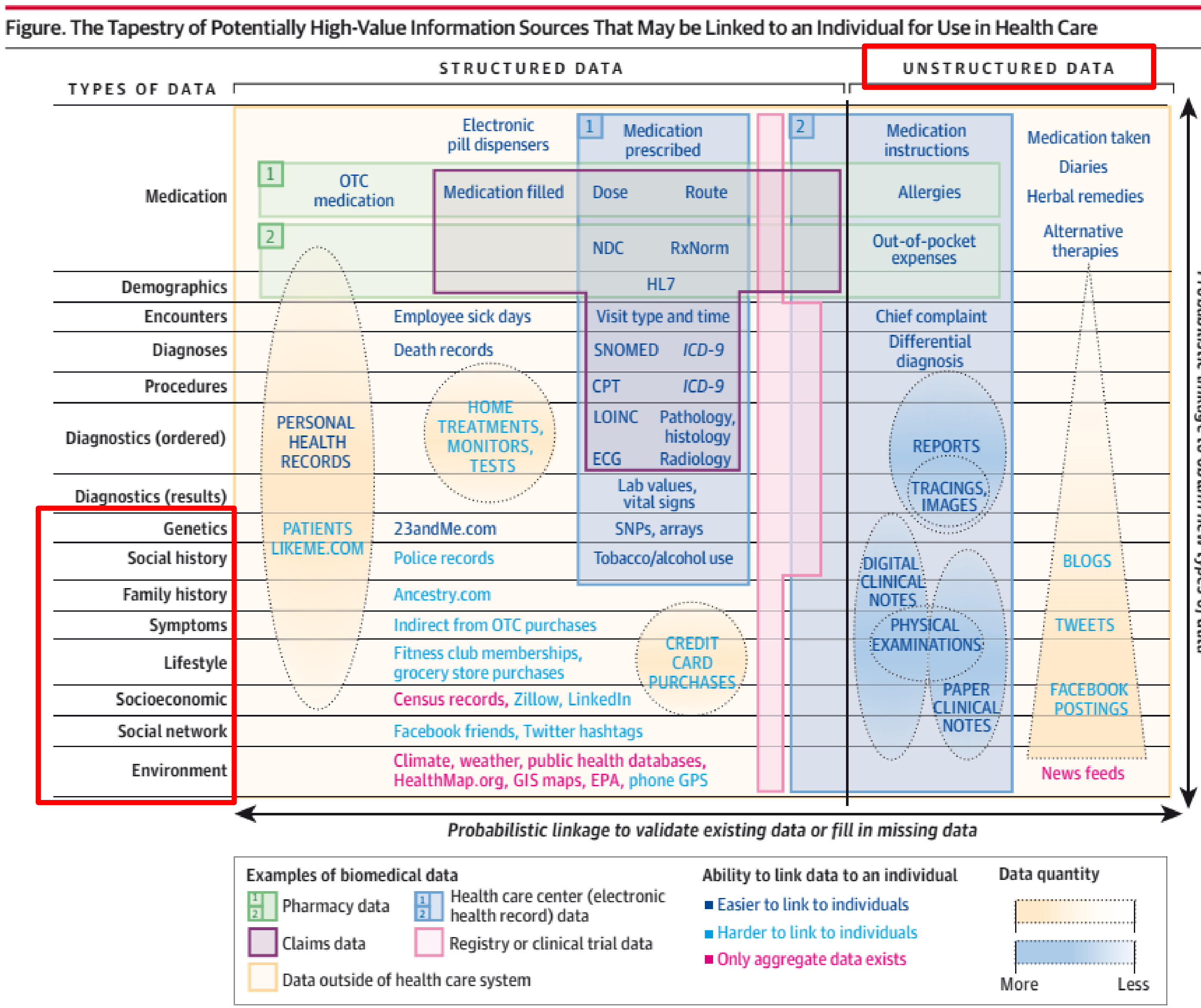
AJOU UNIVERSITY

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

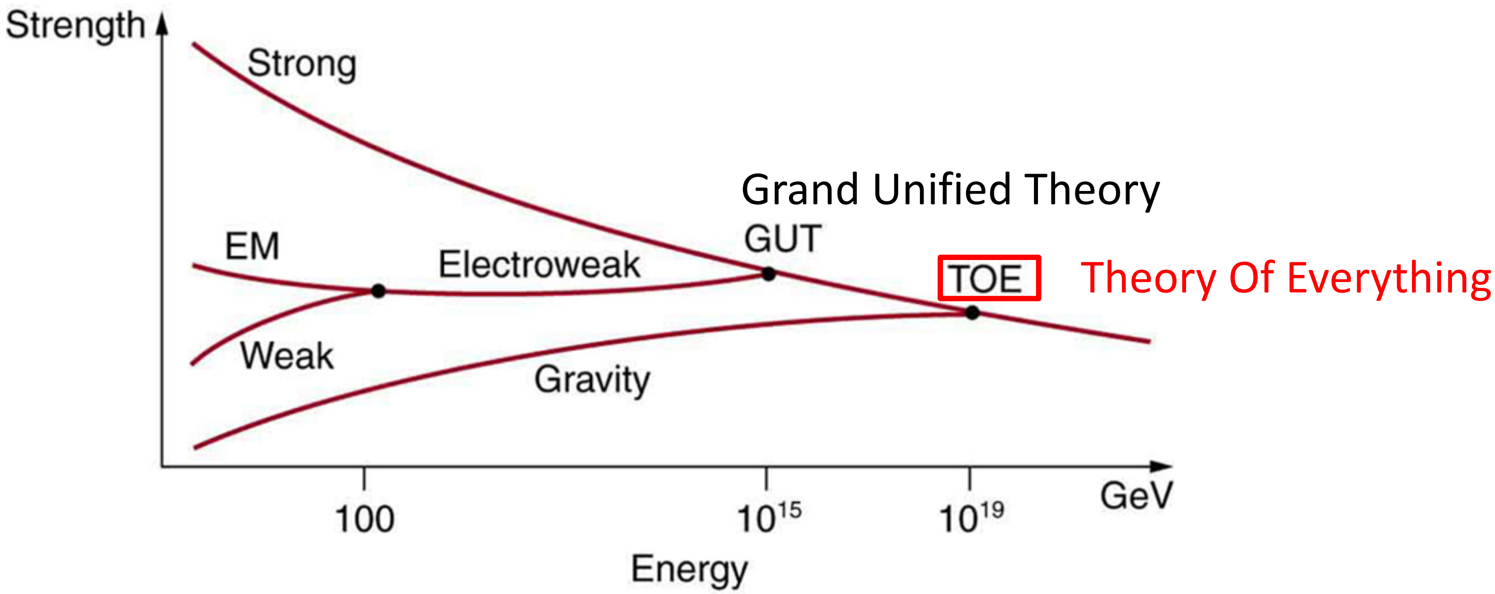


Weber et al., "Finding the Missing Link for Big Biomedical Data." JAMA (June 25, 2014)

Physics: A search for Simplicity, Beauty and Symmetry

- 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 modem 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

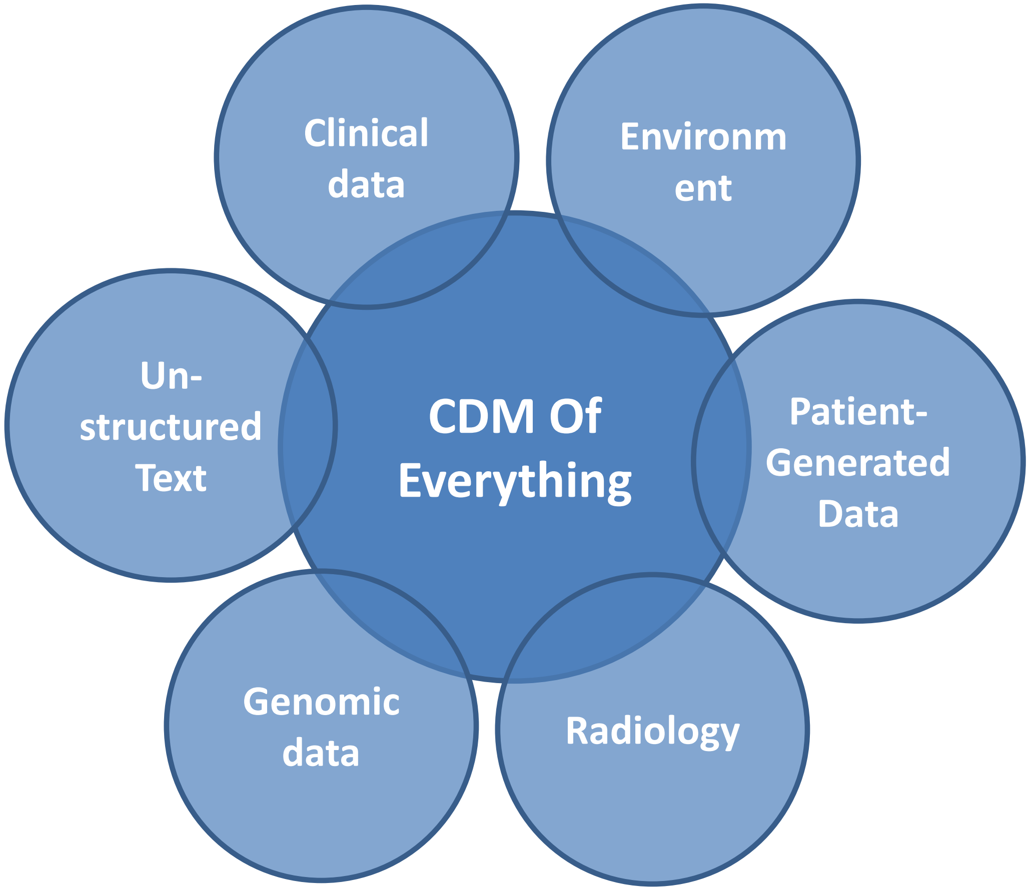
Physics: A search for Simplicity, Beauty and Symmetry



- 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."

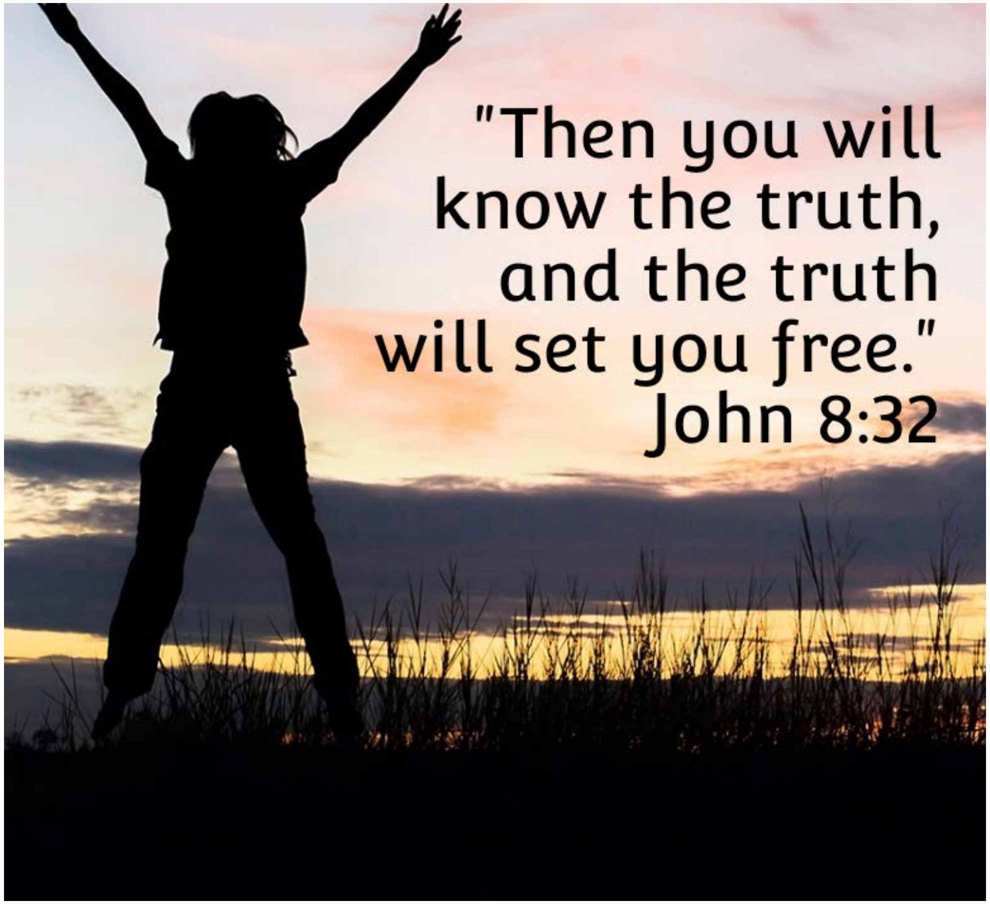
OHDSI: A Journey for Simplicity, Beauty and Symmetry in Medical Data

- 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

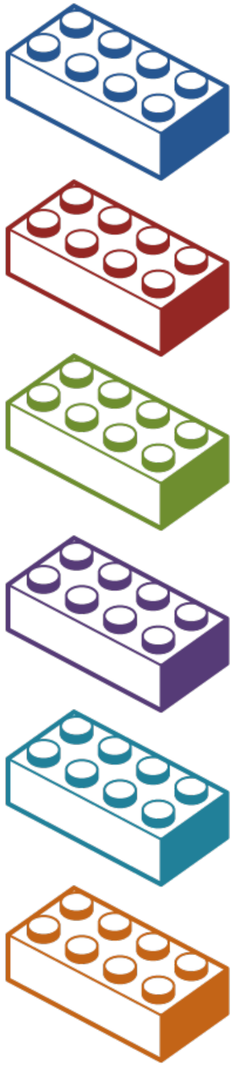


OHDSI: A Journey for Simplicity, Beauty and Symmetry in Medical Data

- 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

Estimation methods

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).

Prediction methods

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.

Method characterization

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.

ATLAS

Supporting packages

Database Connector
Connect directly to a wide range of database platforms, including SQL Server, Oracle, and PostgreSQL.

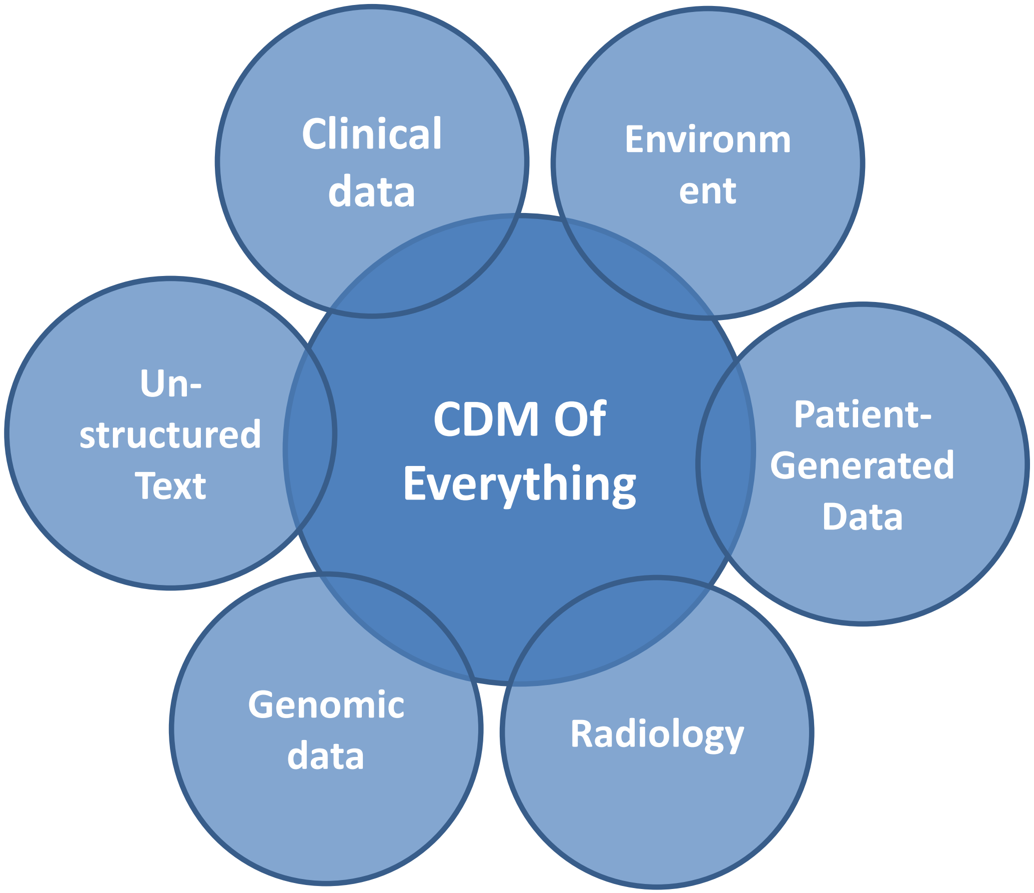
Sql Render
Generate SQL on the fly for the various SQL dialects.

Cyclops
Highly efficient implementation of regularized logistic, Poisson and Cox regression.

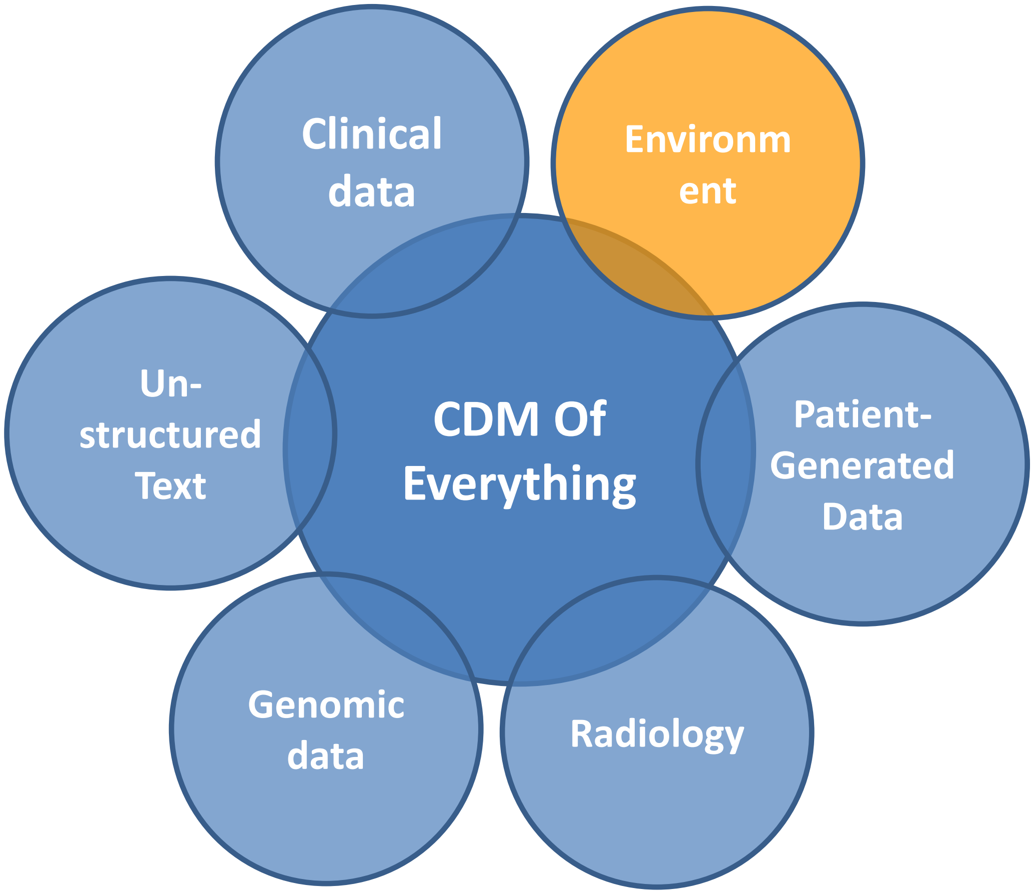
Ohdsi R Tools
Support tools that didn't fit other categories, including tools for maintaining R libraries.

Under construction

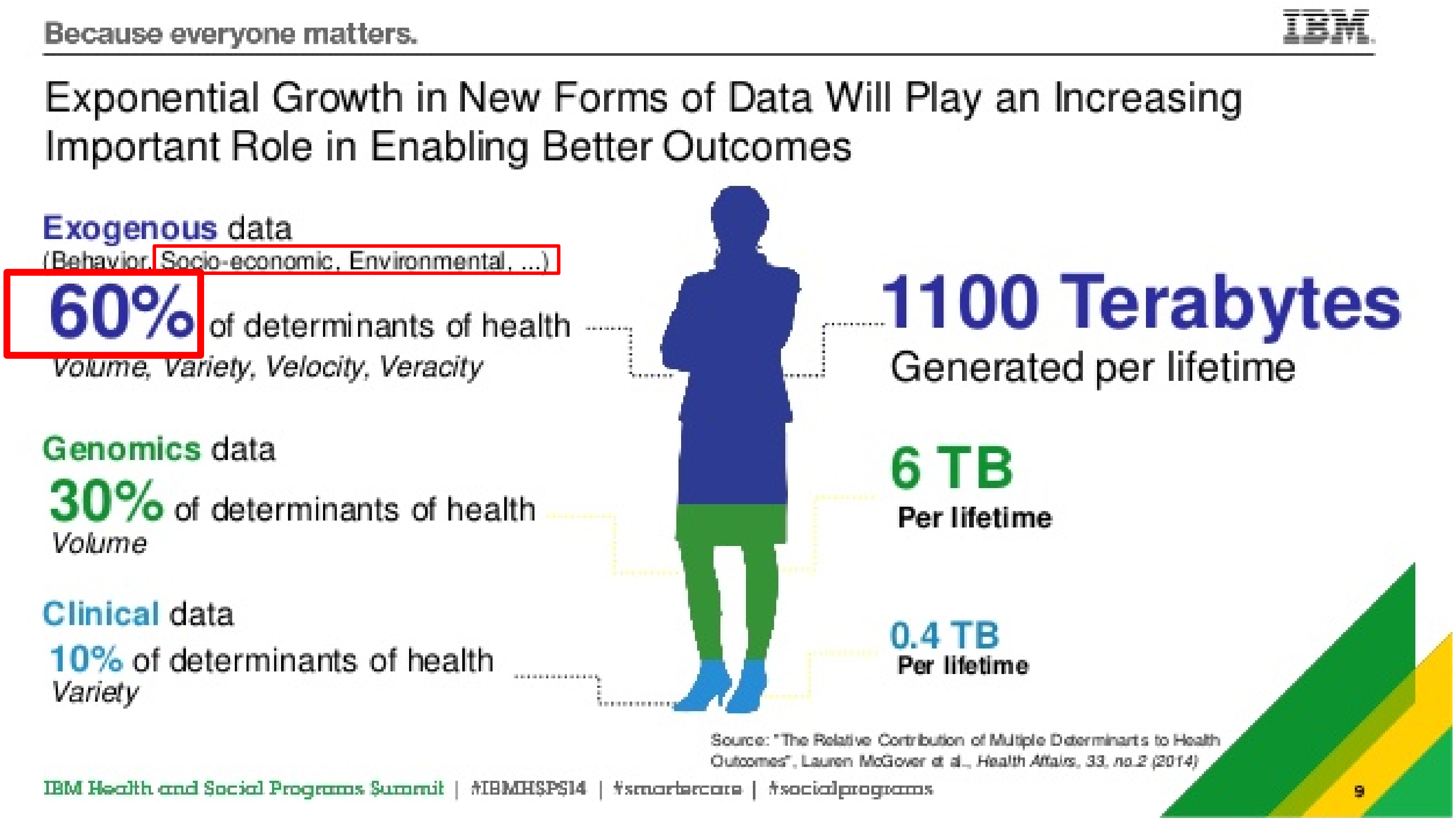
Common Data Model of Everything in Medicine



Common Data Model of Everything in Medicine



Environment in Health



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"*

Carlsten et al., EMBO Rep, 2014

The definition of environment in medicine

- **Environment** is everything that is around us

https://simple.wikipedia.org/wiki/Environment

- ***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://en.wikipedia.org/wiki/Environmental_medicine

What is the environment in medicine?

- **Environment** is everything that is around us

https://simple.wikipedia.org/wiki/Environment

- ***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

https://en.wikipedia.org/wiki/Environmental_medicine

What is the environment in medicine?

- **Environment** is everything that is around us

https://simple.wikipedia.org/wiki/Environment

- ***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

https://en.wikipedia.org/wiki/Environmental_medicine

- All above are based on **G**eographic 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

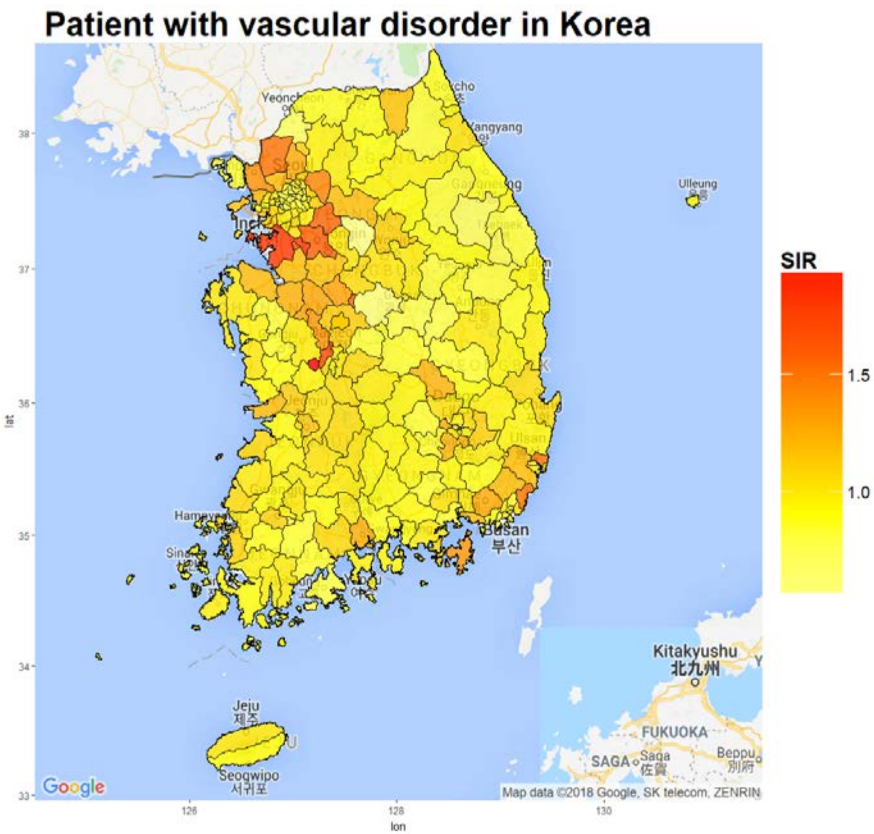
Dept. of Biomedical Informatics, Ajou University School of Medicine, Yeongtong-gu, Suwon

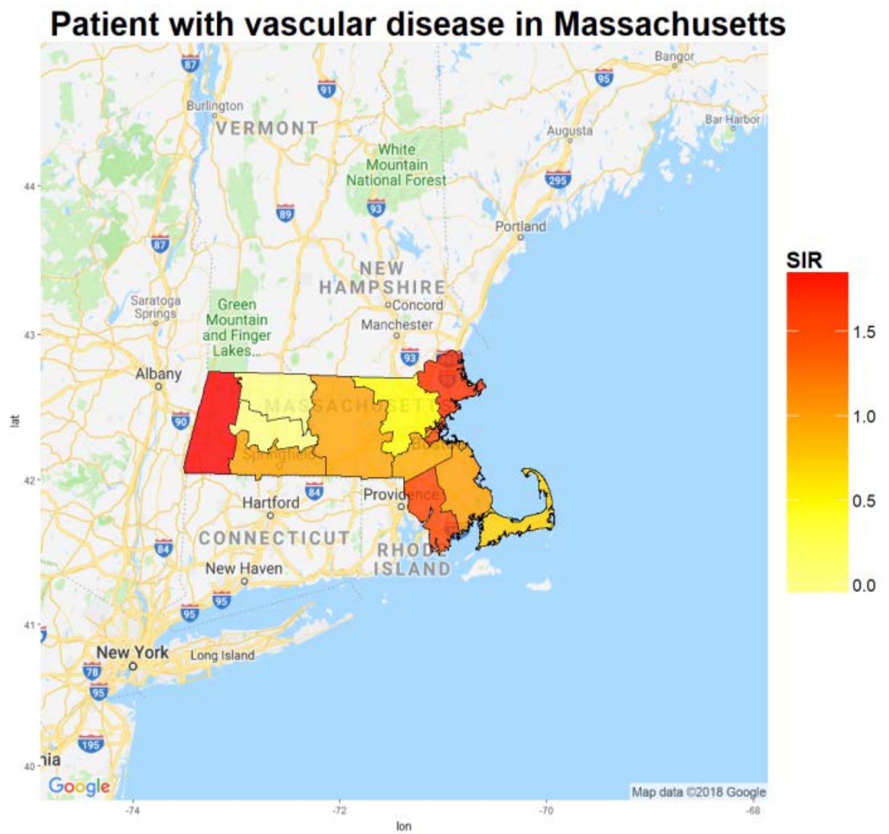
Yeungnam University Graduate school of Medicine, Nam-gu, Daegu

Aim

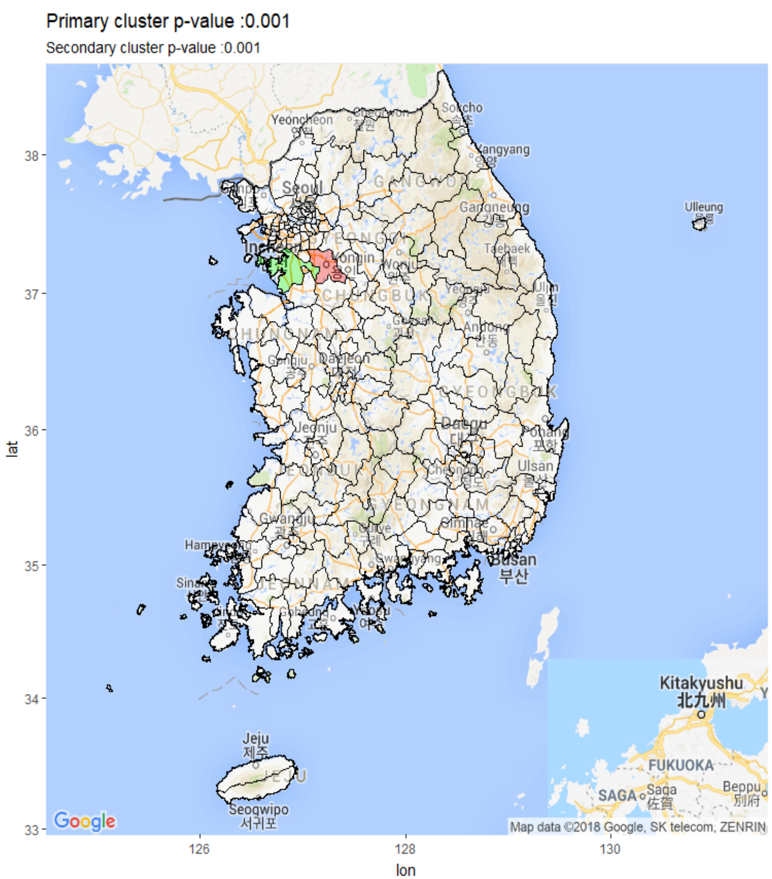
- AEGIS development
  - AEGIS : Application for Epidemiological Geographic Information System
  - A tool to conduct disease mapping and cluster analysis considering age and gender-adjustment and spatial autocorrelation using GIS database based on CDM
  - AEGIS is open-source software, which is harmonized within OHDSI eco-system

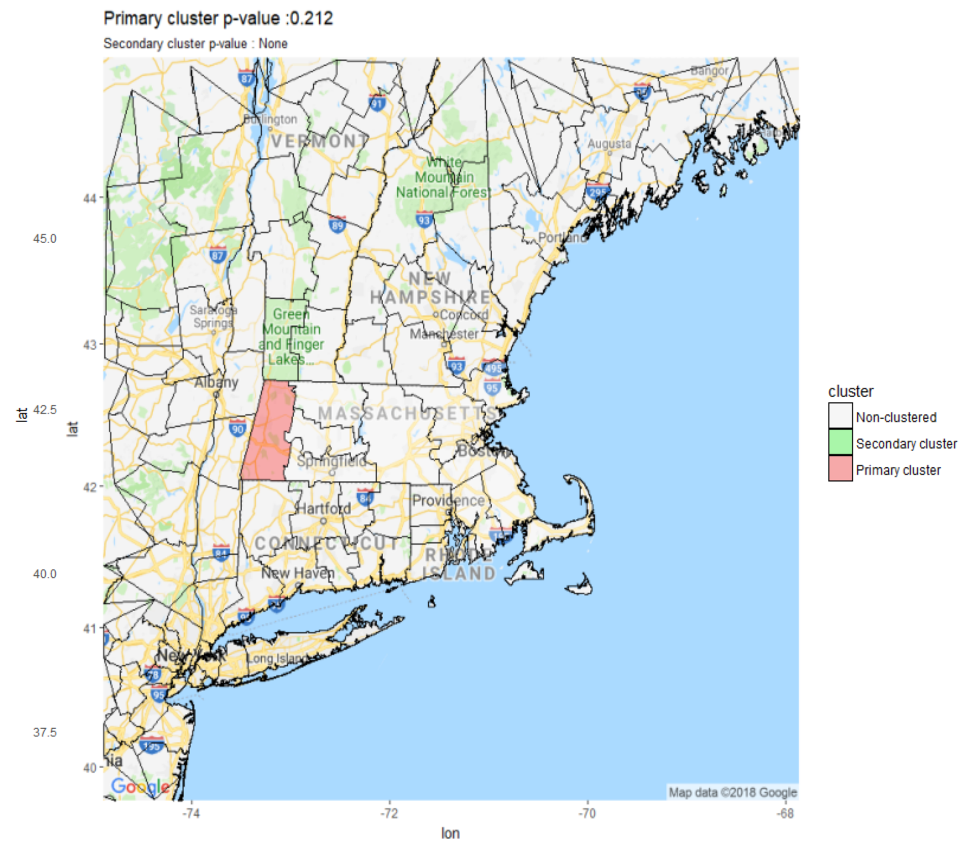
- Based on Global Administrative Database (GADM), AEGIS can depicts cohorts on the map according to the country's own administrative district.





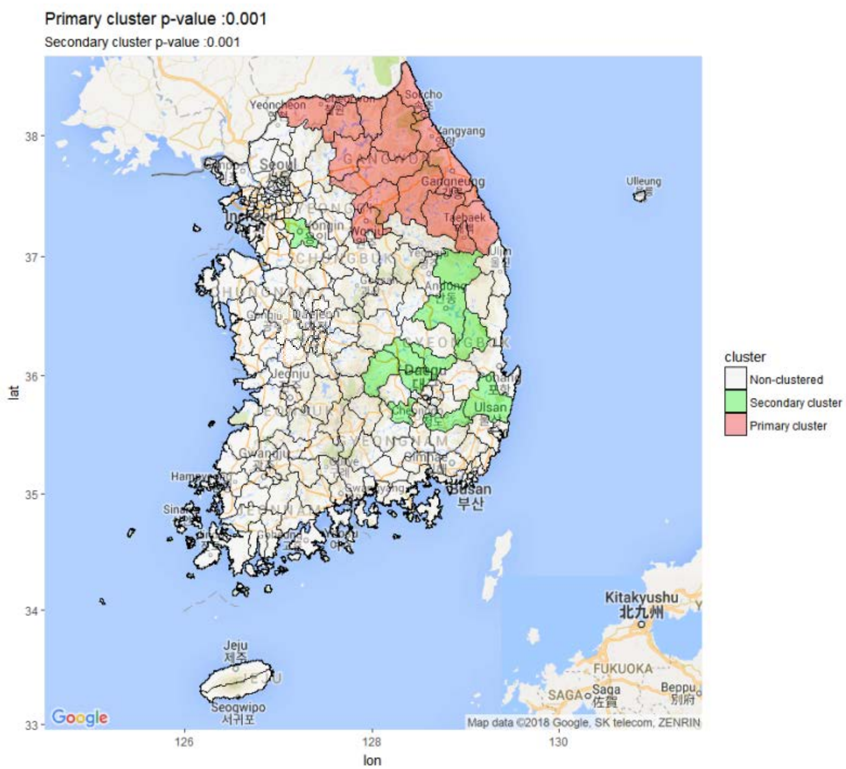
- Based on Global Administrative Database (GADM), AEGIS can depicts cohorts on the map according to the country's own administrative district.





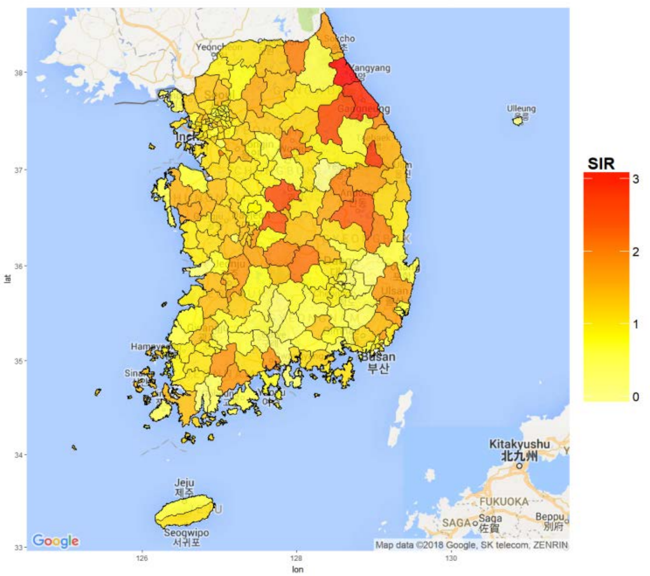
4 Identification of Disease Cluster → AEGIS

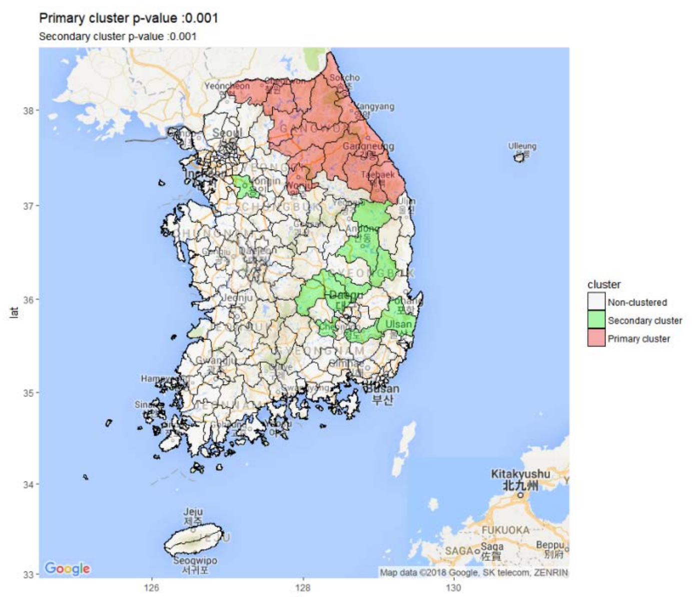
- Clustering of emergency department visit due to asthma among patients with asthma

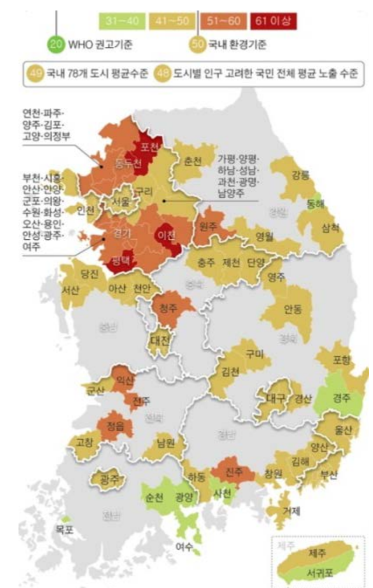


4 Identification of Disease Cluster → AEGIS

- Association of Asthma Exacerbation and Air pollution







Air pollution map in Korea (PM-10)

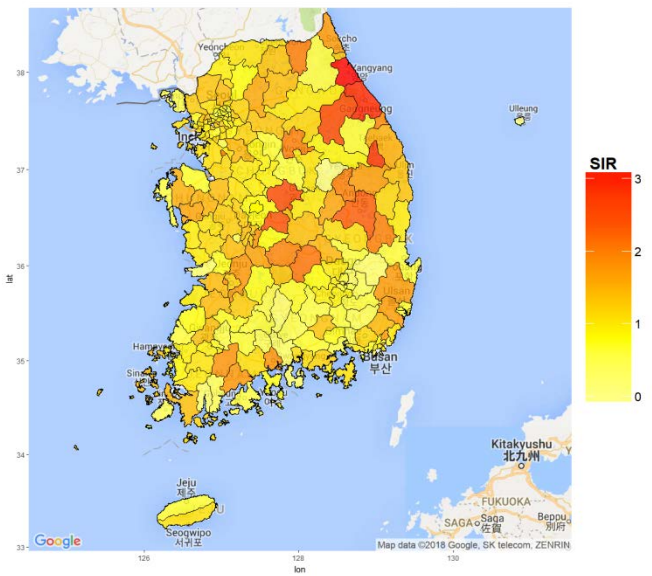
NATIONAL INSTITUTE OF ENVIRONMENTAL RESEARCH, 2018

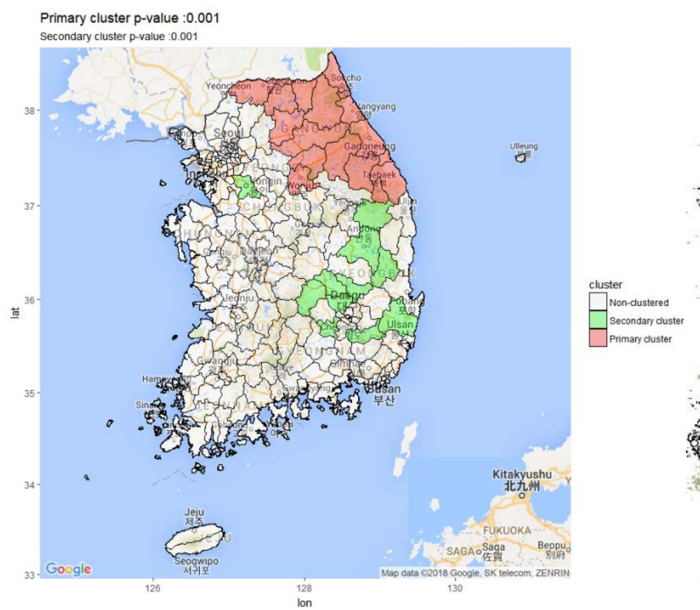
아주대학교

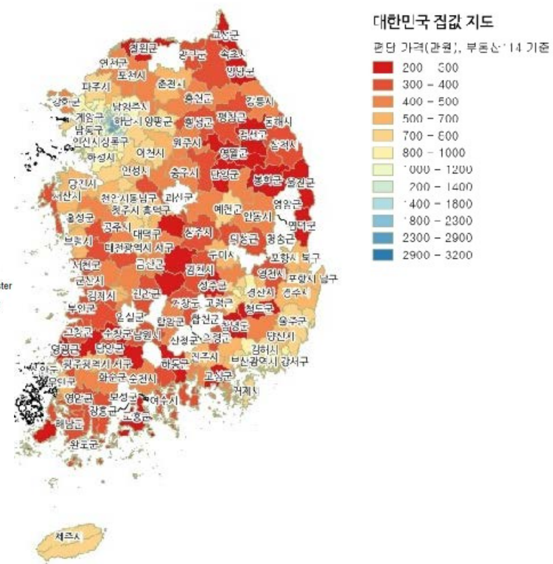
They don't seem to be correlated

4 Identification of Disease Cluster → AEGIS

- Association of Asthma Exacerbation and House price



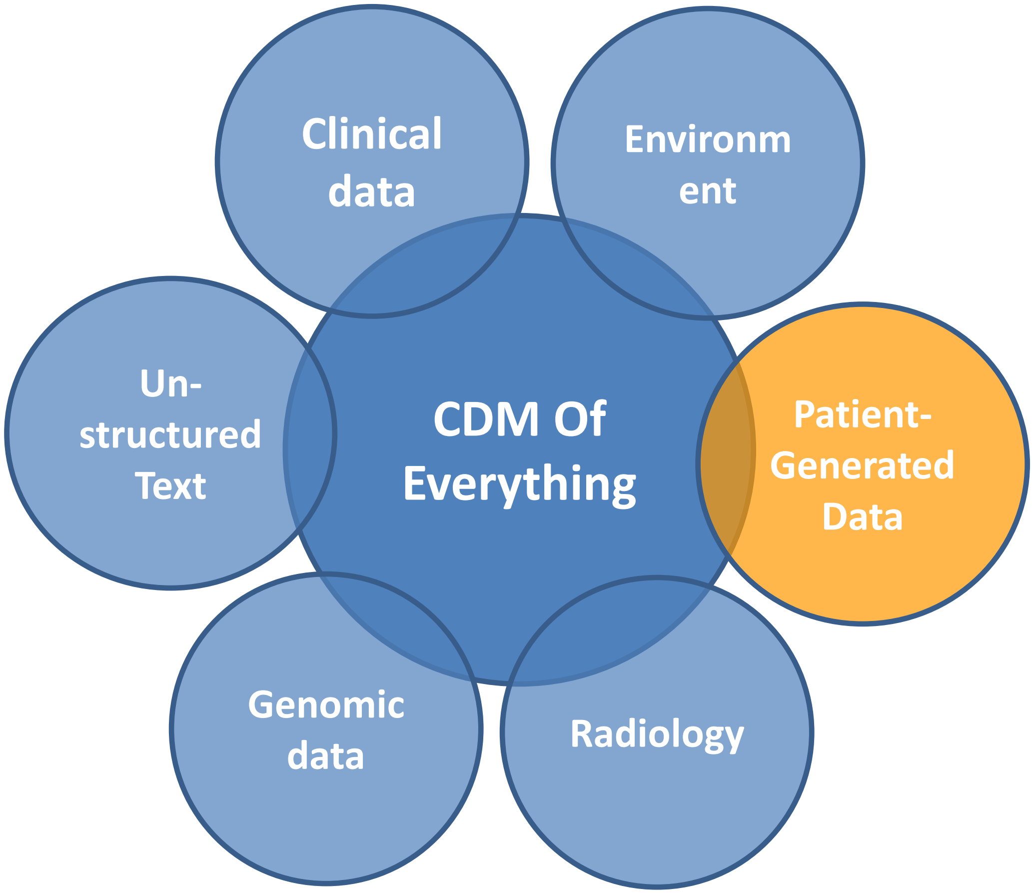




House prices map in Korea

They seem to be correlated!

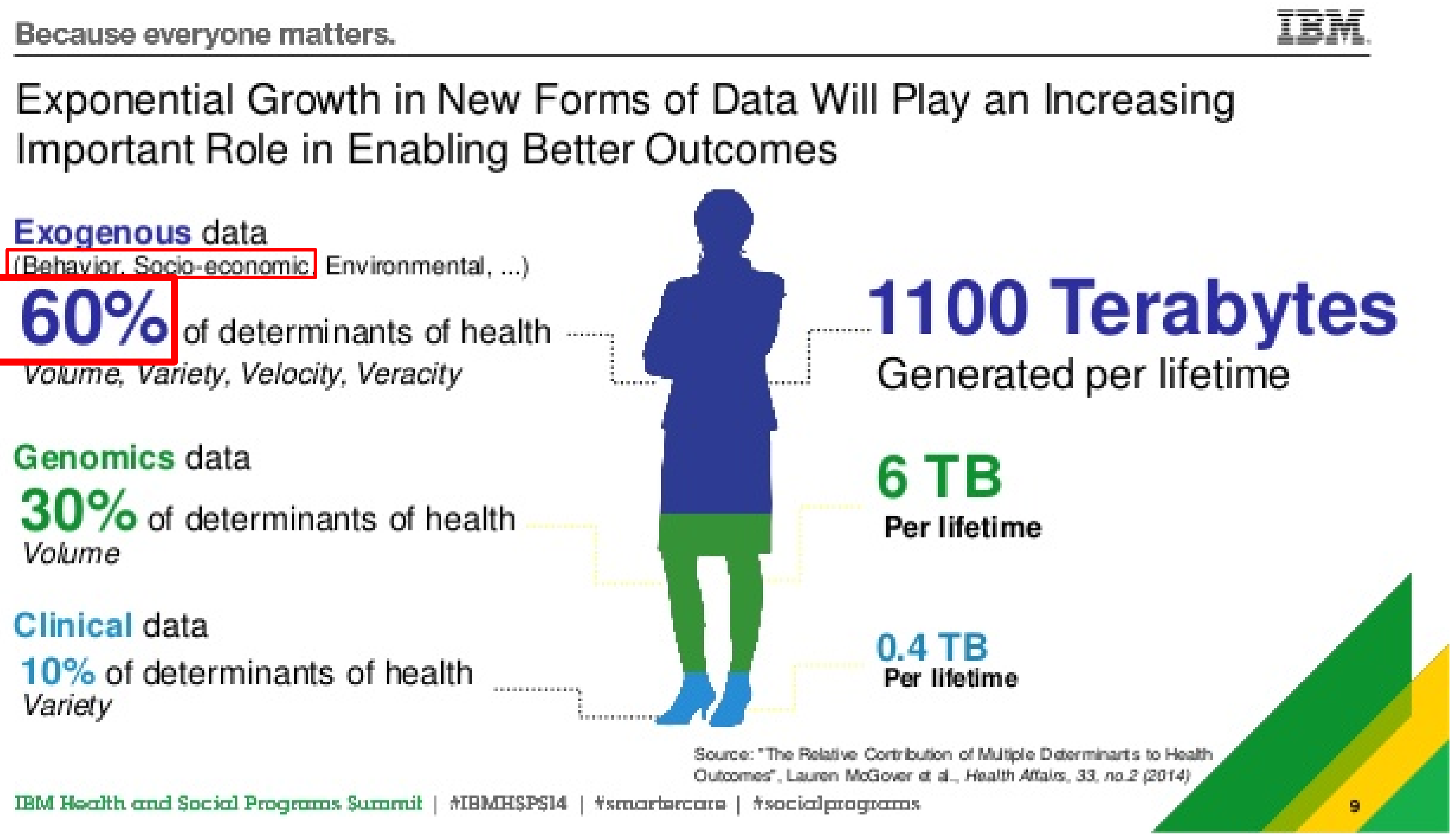
Common Data Model of Everything in Medicine



Seng Chan You, MD[1], Youngin Kim, MD[2], Jaehyung Cho[1], Rae Woong Park, MD, PhD[1,3]

[1]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea;
[2]Medicine, Noom, Inc, Seoul, Korea
[3]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea

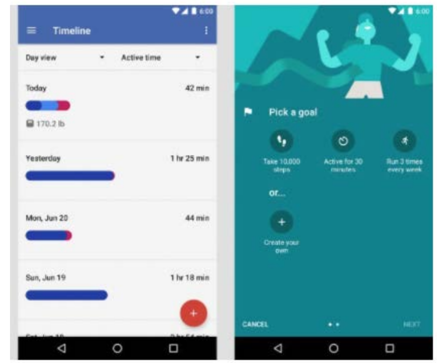
Patient-Generated Health Data



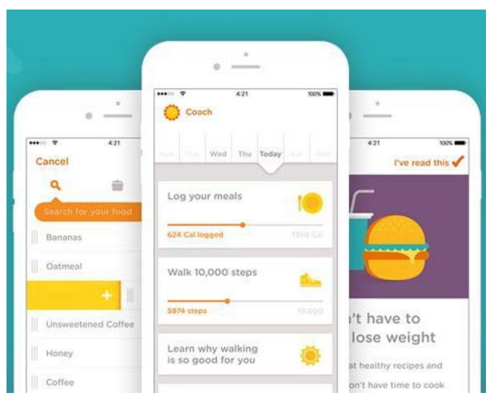
Applications in smartphone collecting health data

Apple Health

Google Fit

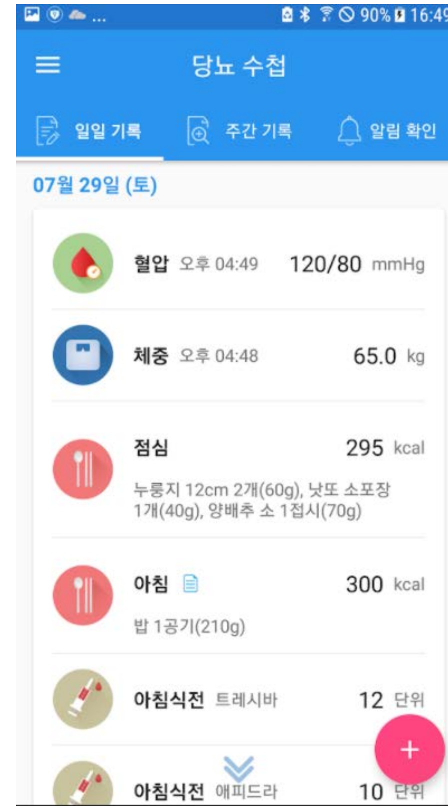


NOOM



Efil

Samsung Medical Center
Diabetes Note



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

General

SCYou Seng Chan You — 1 3d

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 7)
The primitive ETL convention for this data is released, too (PGHD_ETL_convention 7)

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

7 Likes — Reply

created	last reply	13	133	11	25	8
2 days	2 days	replies	views	users	likes	links

Wonchul Wonchul Cha — 3d

Great work Seungchan! Let's make some progress!

2 Likes — Reply

Rijnbeek Peter Rijnbeek — 3d

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 1 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

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

https://1drv.ms/x/s!Aseo02iy8GQ-jfopHLrrp_p9NoD8XQ

NOOM converted their data into CDM

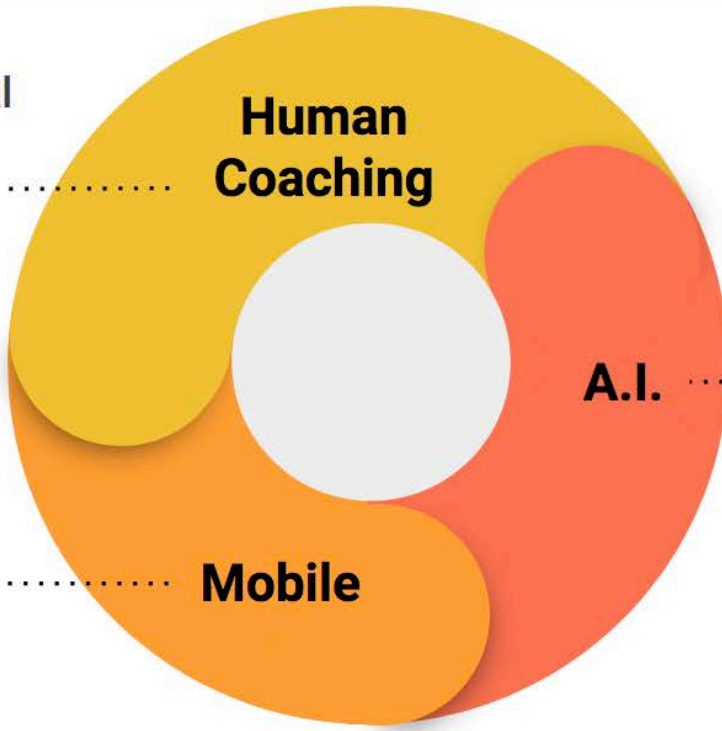
Noom is a behavior change company that uses **A.I., Human Coaching and Mobile Technology** to create the world's most effective solutions for lifestyle & chronic conditions

Use latest psychological approaches (CBT)

Patented scalable coaching using A.I.

Mobile-first company

47 million users

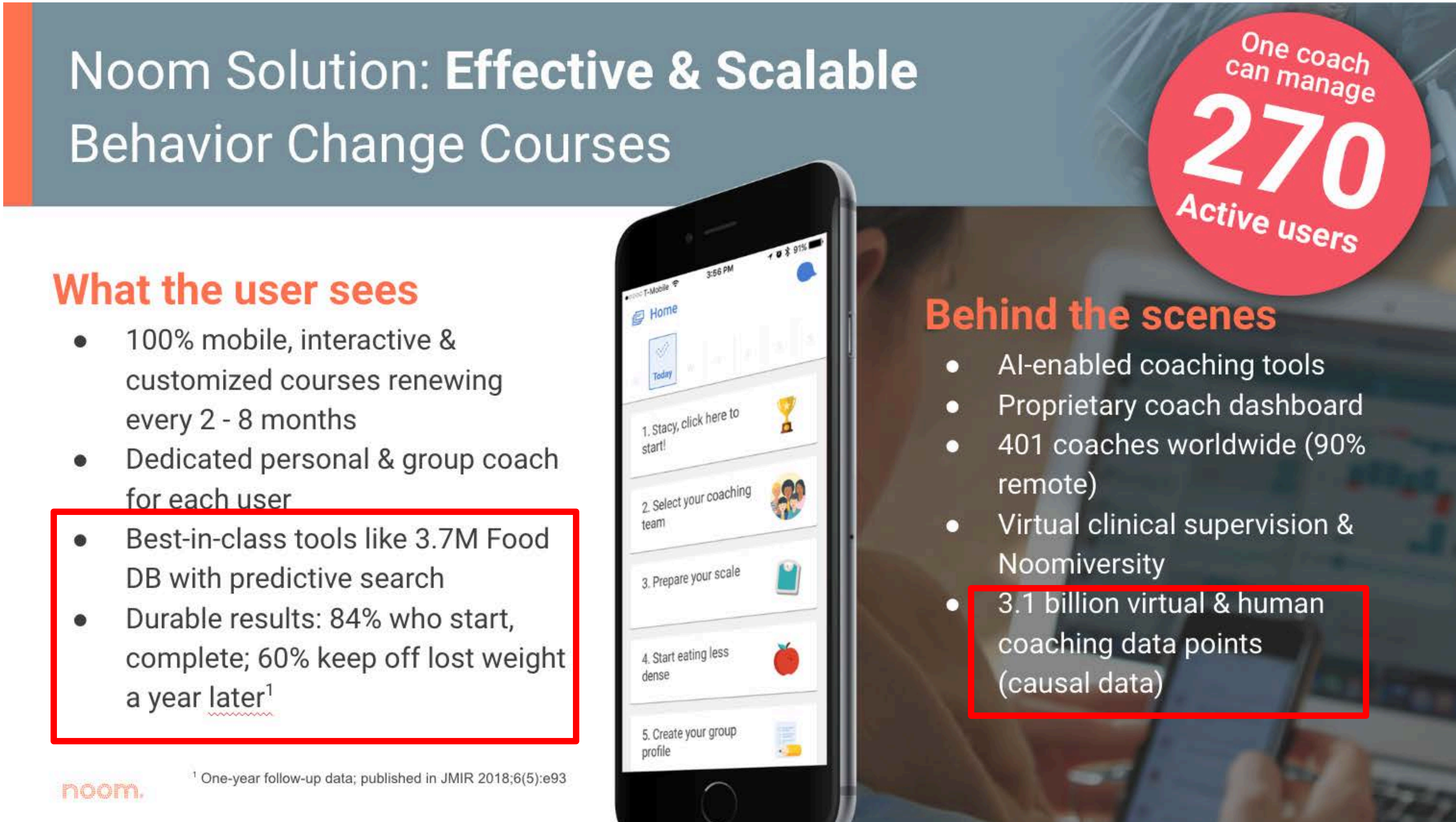


3.1 billion coaching data points

Self-learning A.I.

noom.

NOOM converted their data into CDM



ETL result of sample data from NOOM

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

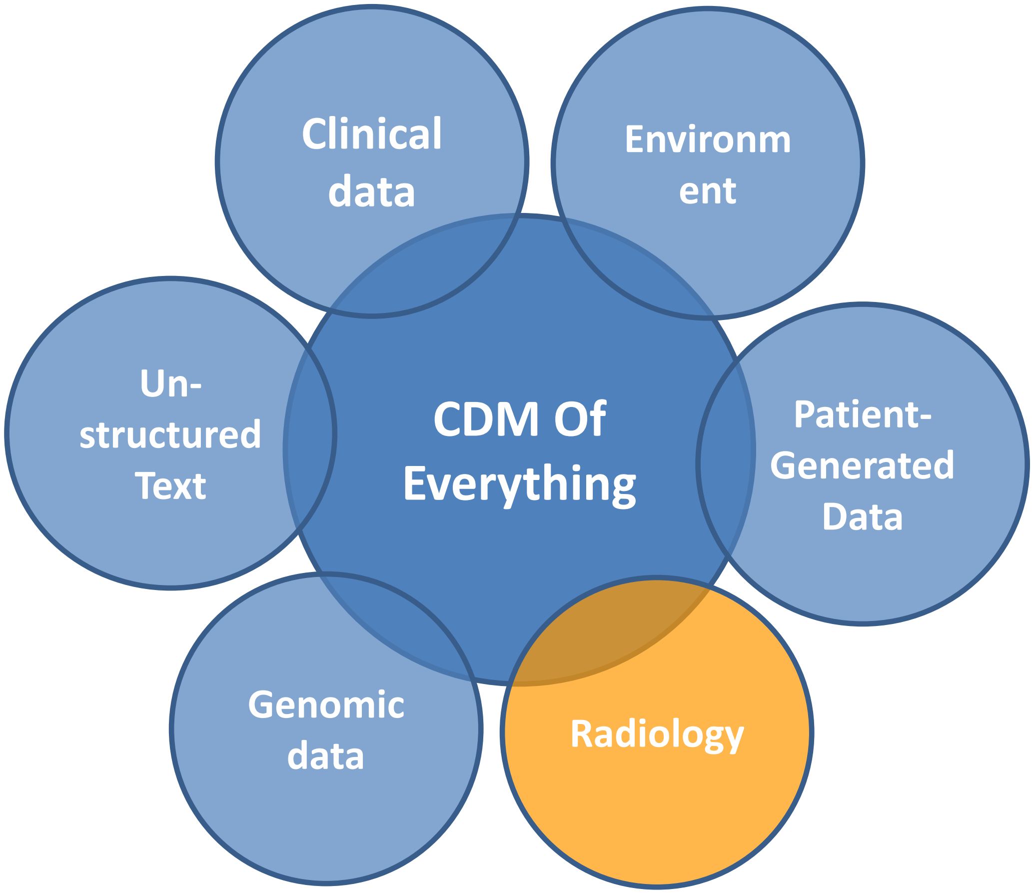
measurem	person_id	measurem	value_sour	unit_sourc	measurem	concept_name	measurement_date	measurement_datetime	value_as_n	unit_conce	unit_conce	measurem
1	1	Weight	103.4	kg	3025315	Body weight	2017-05-08	2017-05-08 22:56	103.4	4122383	kg	44818704
2	1	Weight	108	kg	3025315	Body weight	2017-03-22	2017-03-23 10:27	105	4122383	kg	44818704
3	1	Weight	109	kg	3025315	Body weight	2017-03-04	2017-03-04 9:46	106.7	4122383	kg	44818704
31	2	Weight	69.9	kg	3025315	Body weight	2017-07-11	2017-07-11 9:30	69.9	4122383	kg	44818704
32	2	Weight	70	kg	3025315	Body weight	2018-04-26	2018-04-26 9:39	65.8	4122383	kg	44818704
33	2	Weight	69.8	kg	3025315	Body weight	2018-02-28	2018-02-28 9:24	69.8	4122383	kg	44818704

observatio	person_id	observation_source_value	value_sour	unit_sourc	observatio	concept_name	observation_date	value_as_number	unit_conce	unit_conce	observatio	observatio	Observation_type_concept_name
1	1	Steps	9097	count	3034985	Number of steps in 24 hour Measured	2017-07-04	9348	44777556	per 24 hours		44814721	App generated
2	1	Steps	1600	count	3034985	Number of steps in 24 hour Measured	2017-04-24	1519	44777556	per 24 hours		44814721	App generated
3	1	Steps	7200	count	3034985	Number of steps in 24 hour Measured	2017-05-15	7269	44777556	per 24 hours		44814721	App generated
170	2	Steps	4944	count	3034985	Number of steps in 24 hour Measured	2018-04-28	4944	44777556	per 24 hours		44814721	App generated
171	2	Steps	1800	count	3034985	Number of steps in 24 hour Measured	2017-08-09	1687	44777556	per 24 hours		44814721	App generated
172	2	Steps	4381	count	3034985	Number of steps in 24 hour Measured	2018-02-14	4943	44777556	per 24 hours		44814721	App generated
173	2	Steps	8735	count	3034985	Number of steps in 24 hour Measured	2017-09-15	3626	44777556	per 24 hours		44814721	App generated
9147	19	Dietary Calories	1598000	calorie	4037128	Dietary calorie intake	2018-04-03	1498000	9472	calorie		44814721	Patient reported
9148	19	Dietary Calories	1186000	calorie	4037128	Dietary calorie intake	2018-04-04	1176000	9472	calorie		44814721	Patient reported
9149	19	Dietary Calories	1772000	calorie	4037128	Dietary calorie intake	2018-04-05	1672000	9472	calorie		44814721	Patient reported
9150	19	Dietary Calories	1329000	calorie	4037128	Dietary calorie intake	2018-04-06	1309000	9472	calorie		44814721	Patient reported

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, MS[1], Kwang Soo Jeong[1], Si Hyung No[2], Kwon-Ha Yoon, MD, PhD[3], Chang-Won Jeong, PhD[2], Rae Woong Park, MD, PhD[1,4]

[1]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea;
[2]Imaging Science based Lung and Bone Disease Research Center, Wonkwang University, Iksan, Korea;
[3]Department of Radiology, Wonkwang University Colledge of Medicine
[4]Department 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

CDM Builders

Patrick_Ryan

Dec '16

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?

Reply

created Dec 14, '16 | last reply 54 mins | 22 replies | 1.6k views | 13 users | 1 like | 11 links | 5 | 3 | 3

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**.

Langer, et al., "Collaborative and Reproducible Research: Goals, Challenges, and Strategies." *Journal of Digital Imaging* (2018)

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

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

Kohli et al., "Medical Image Data and Datasets in the Era of Machine Learning—Whitepaper from the 2016 C-MIMI Meeting Dataset Session." *Journal of Digital Imaging* (2017)

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

AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

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 Langlotz[8]

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

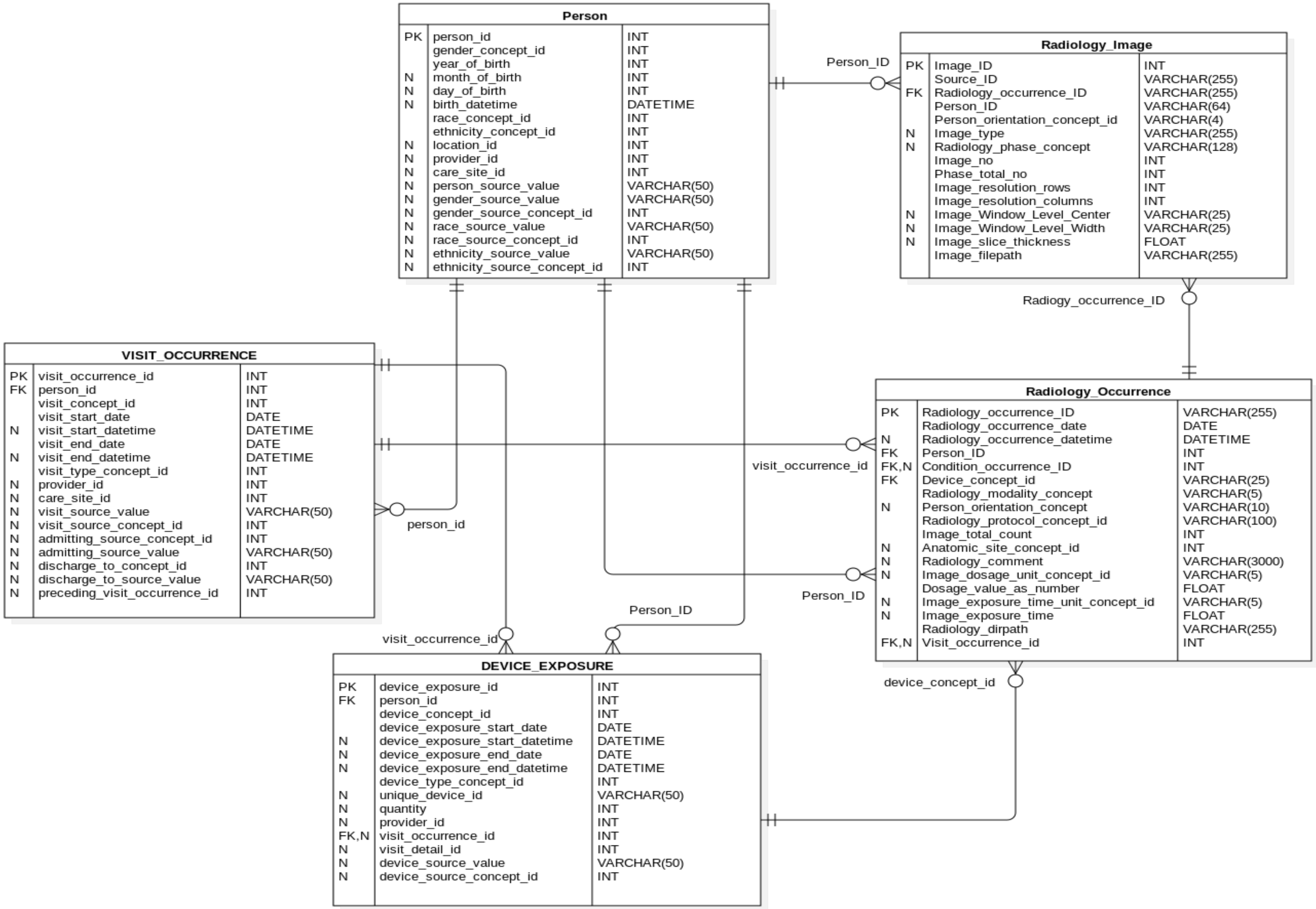
PK	Radiology_occurrence_ID	VARCHAR(255)
	Radiology_occurrence_date	DATE
N	Radiology_occurrence_datetime	DATETIME
	Person_ID	VARCHAR(64)
FK,N	Condition_occurrence_ID	INT
FK	Device_concept_id	VARCHAR(25)
	Radiology_modality_concept_id	VARCHAR(5)
N	Person_orientation_concept_id	VARCHAR(10)
	Radiology_protocol_concept_id	VARCHAR(100)
	Image_total_count	INT
N	Anatomic_site_concept_id	INT
N	Radiology_comment	VARCHAR(3000)
N	Image_dosage_unit_concept_id	VARCHAR(5)
	Dosage_value_as_number	FLOAT
N	Image_exposure_time_unit_concept_id	VARCHAR(5)
N	Image_exposure_time	FLOAT
	Radiology_dirpath	VARCHAR(255)
N	Visit_occurrence_id	INT

Radiology_Image

PK	Image_ID	INT
	Source_ID	VARCHAR(255)
FK	Radiology_occurrence_ID	VARCHAR(255)
	Person_ID	VARCHAR(64)
	Person_orientation_concept_id	VARCHAR(4)
N	Image_type	VARCHAR(255)
N	Radiology_phase_concept_id	VARCHAR(128)
	Image_no	INT
	Phase_total_no	INT
	Image_resolution_rows	INT
	Image_resolution_columns	INT
N	Image_Window_Level_Center	VARCHAR(25)
N	Image_Window_Level_Width	VARCHAR(25)
N	Image_slice_thickness	FLOAT
	Image_filepath	VARCHAR(255)

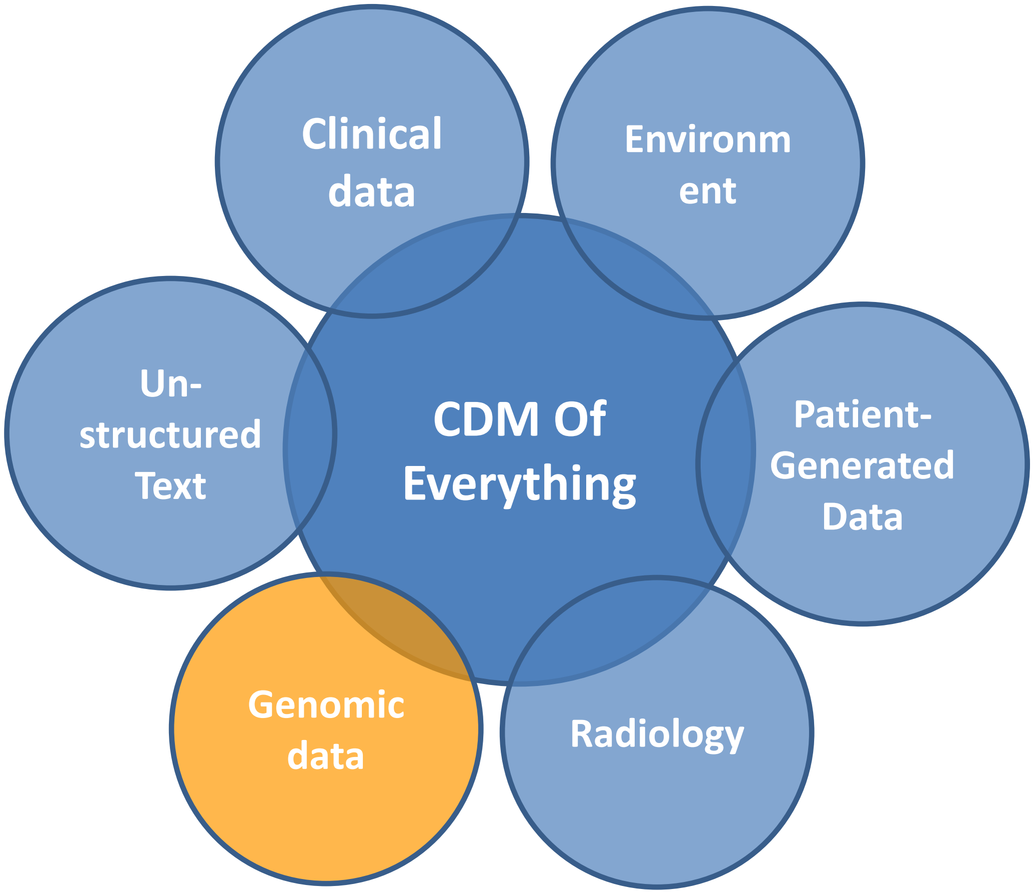
http://forums.ohdsi.org/t/oncology-radiology-imaging-integration-into-cdm/2018/23?u=scyou

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



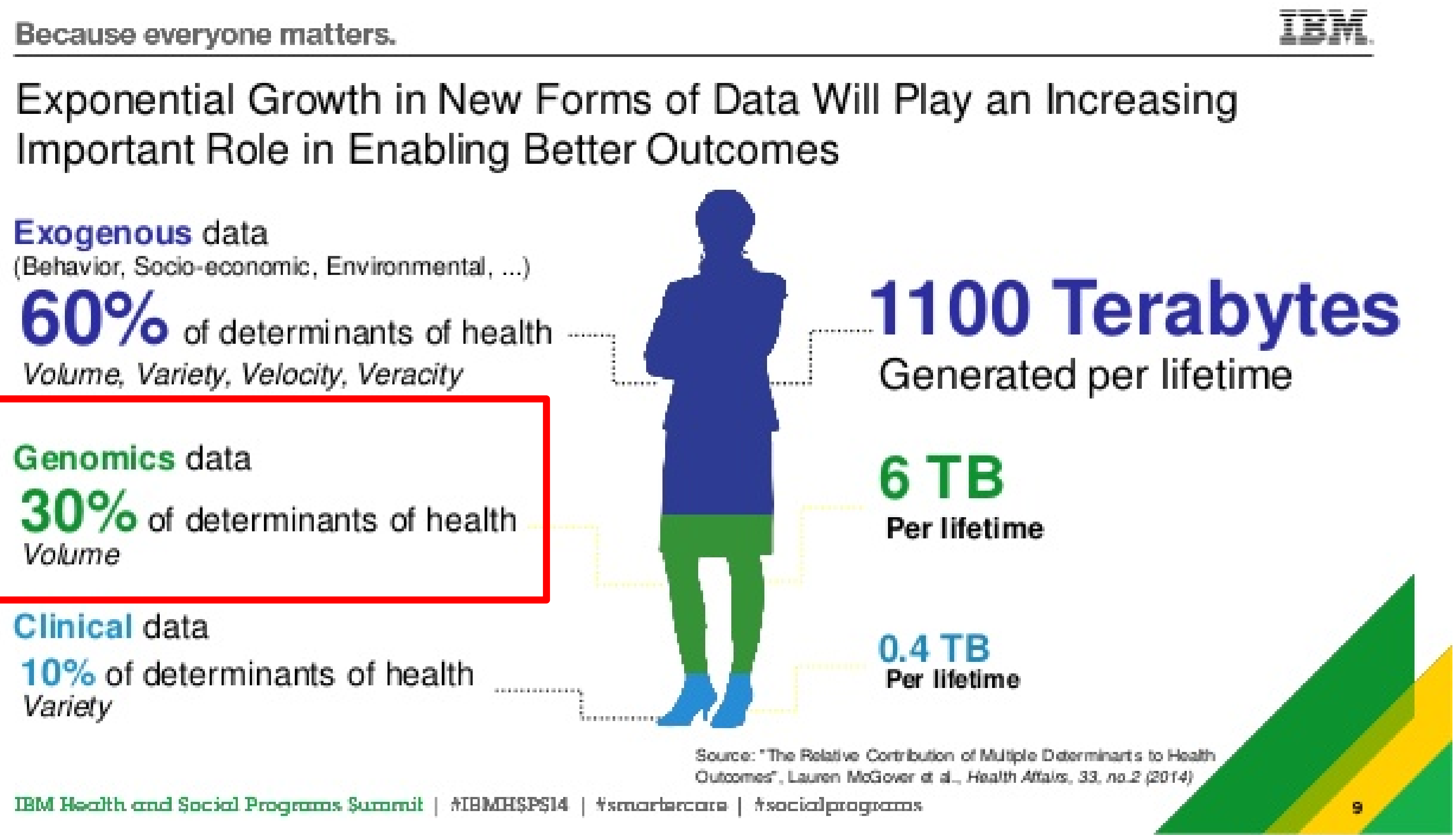
http://forums.ohdsi.org/t/oncology-radiology-imaging-integration-into-cdm/2018/23?u=scyou

Common Data Model of Everything in Medicine



Seo Jeong Shin, MS[1], Seng Chan You, MD, MS[1], Jin Roh, MD, PhD[2], Rae Woong Park, MD, PhD[1, 3]

[1]Dept. of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea; [2]Dept. of Pathology, Ajou University Hospital, Suwon, South Korea; [3]Dept. of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, South Korea


Because everyone matters.
IBM
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
6 TB
Per lifetime
Clinical data
10% of determinants of health
Variety
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)
IBM Health and Social Programs Summit | #IBMHSPS14 | #smartercare | #socialprograms
9

Background: Surge of genomic data

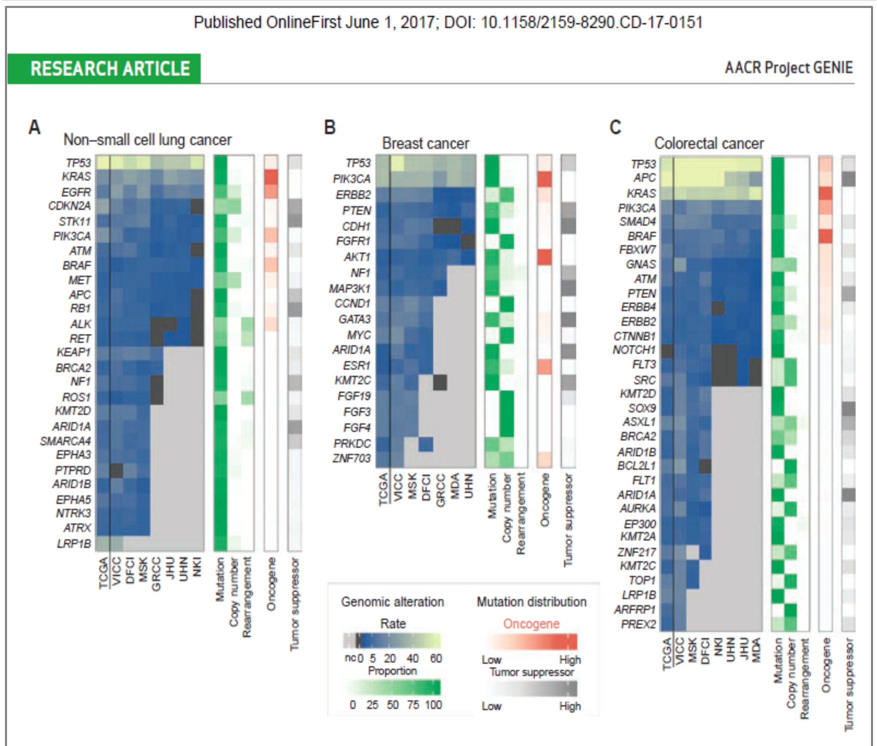
- Global waves of 'precision medicine'
  - Precision medicine initiative in US: Population of 1M, $215M
  - Precision medicine initiative in China
- Insurance coverage of NGS in Korea
  - Since March 2017, national insurance coverage for targeted NGS in cancer patients has started in Korea.
  - No. of 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**

AACR
American Association for Cancer Research
FINDING CURES TOGETHER

PROJECT GENIE
Genomics Evidence Neoplasia Information Exchange



Development of G-CDM based **ISO standard**

TECHNICAL SPECIFICATION

ISO/TS 20428

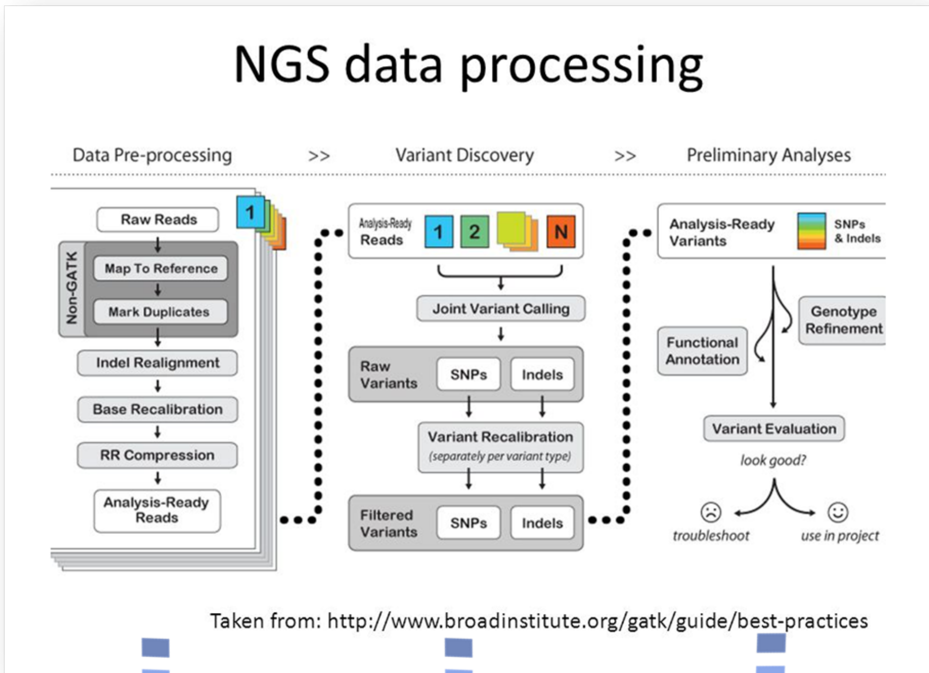
First edition
2017-05

Health informatics — Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records

Informatique de santé — Éléments de données et leurs métadonnées pour décrire l'information structurée de la séquence génomique clinique dans les dossiers de santé électroniques

- 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 | **2. Variant_occurrence** | **3. Variant_annotation**

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**

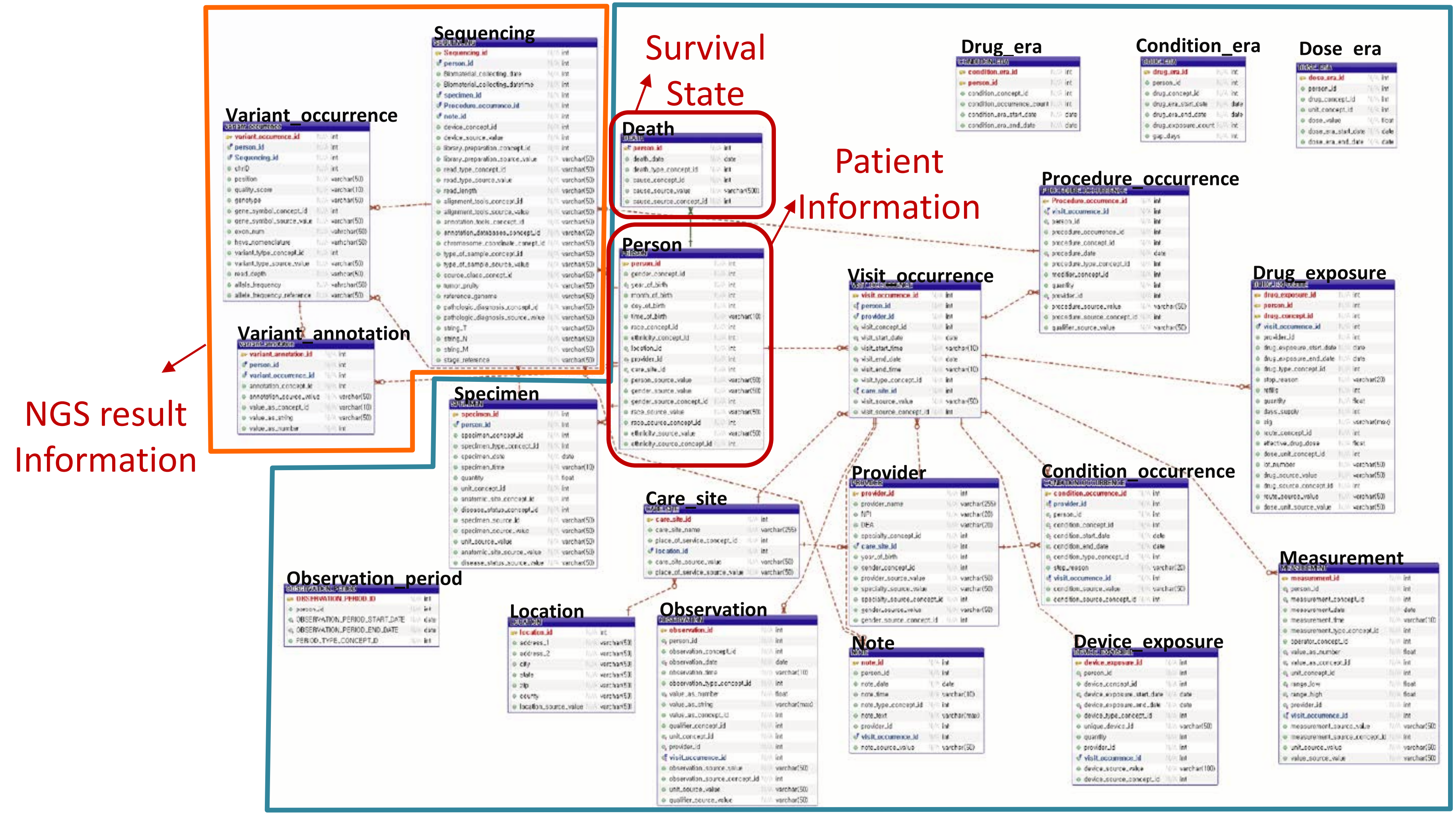
3. Variant_annotation

- Each row represents each **annotation**
- **Flexibility** for any annotation tool

Brief review: G-CDM

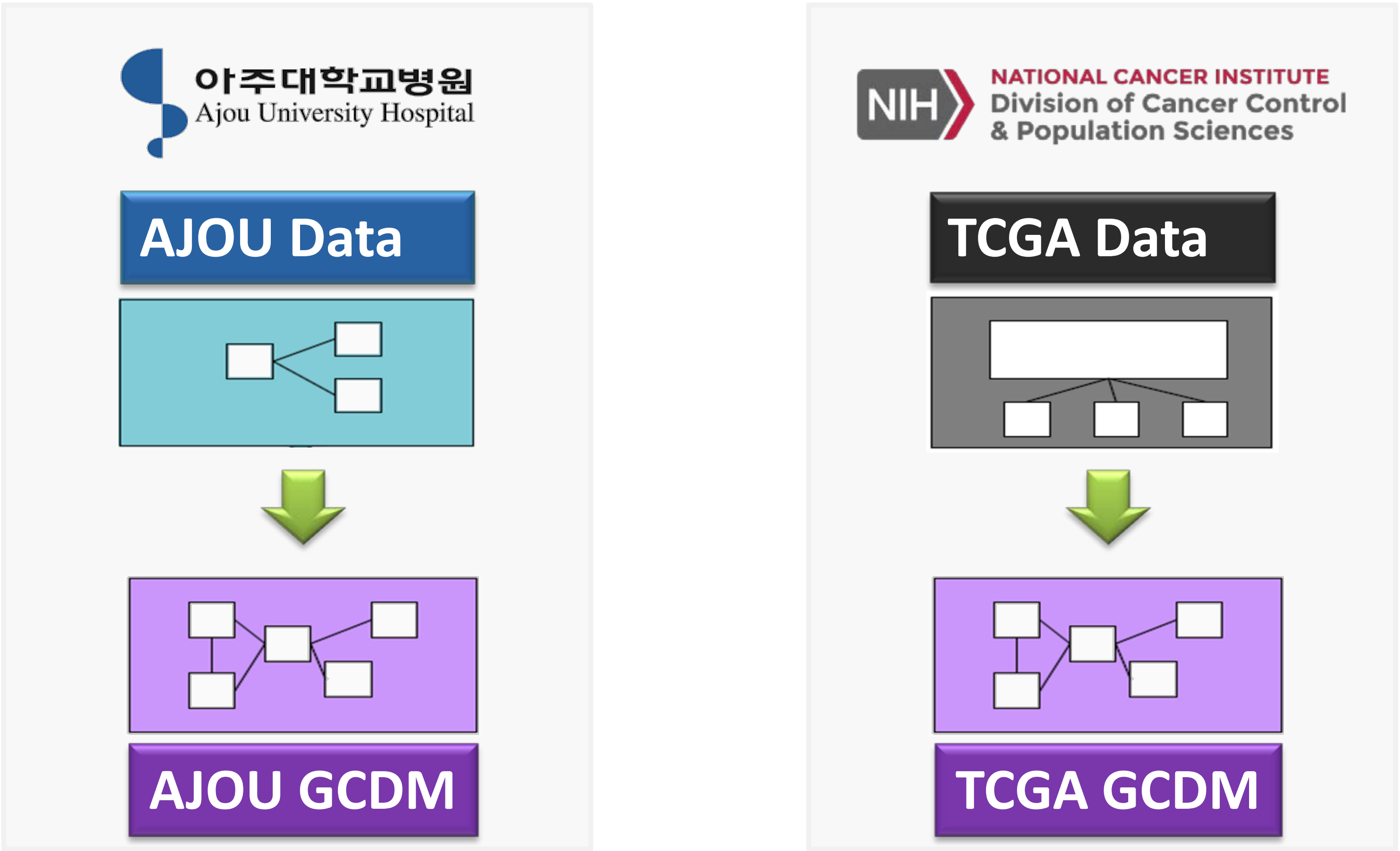
- Overall, three tables are added
- Priority: compatibility with existing OMOP-CDM and OHDSI tools (eg Feature Extraction / Patient Level Prediction package)
- Sequencing table
  - Each row represents **each sequencing** (multiple sequencing is possible for same specimen of same patients)
  - Foreign 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



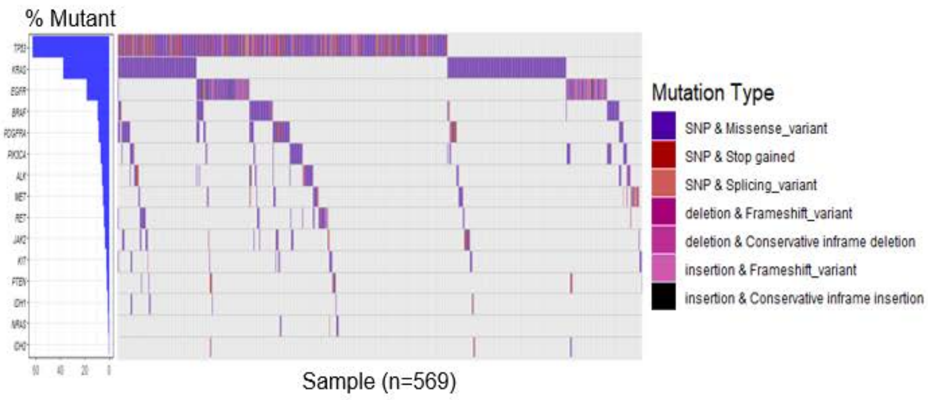
Conversion of G-CDM

- The data structures of the two institutes were unified.

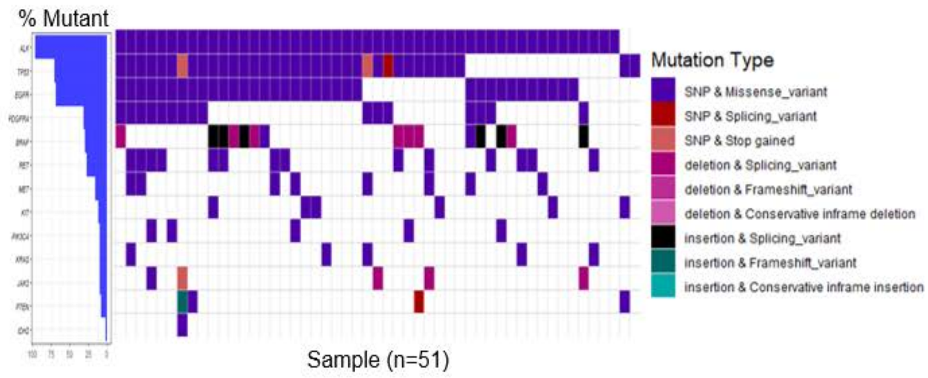


Study Results:
Waterfall plot of adenocarcinoma and squamous cell carcinoma of lung

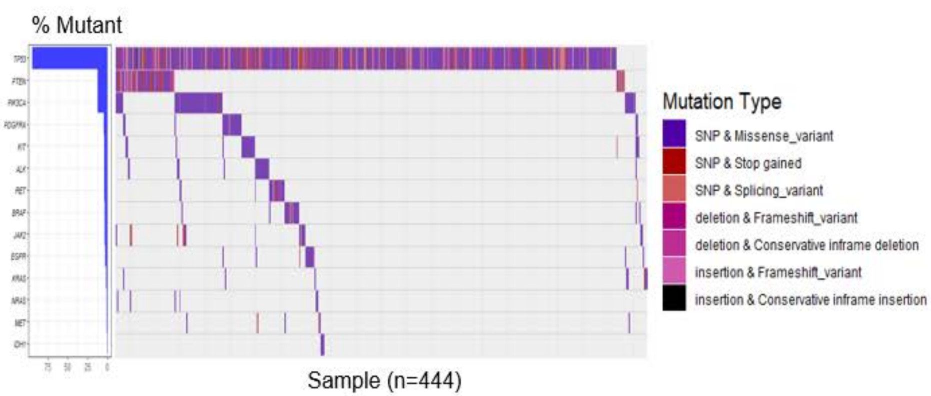
TCGA LUAD



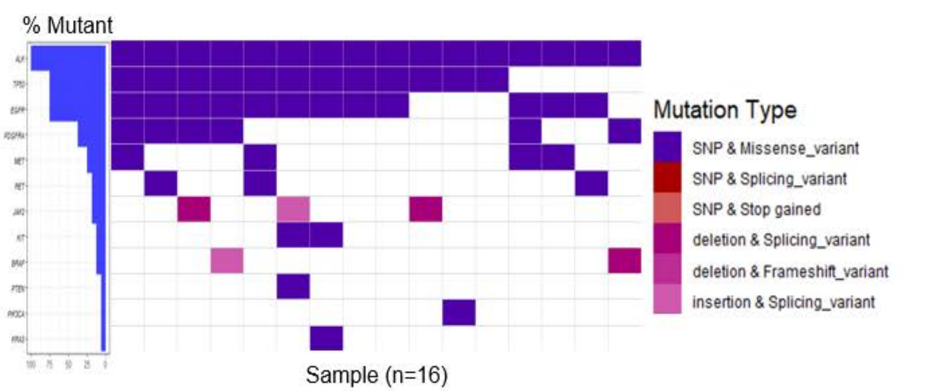
AJOU LUAD



TCGA LUSC



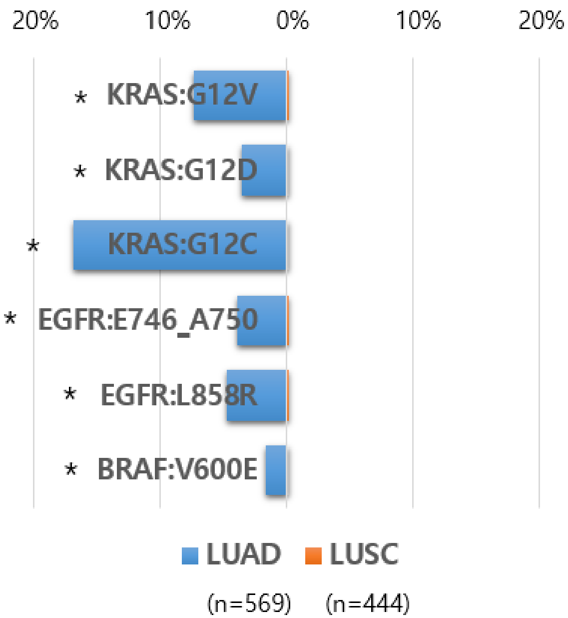
AJOU LUSC



Study Results:
Waterfall plot of adenocarcinoma and squamous cell carcinoma of lung

TCGA

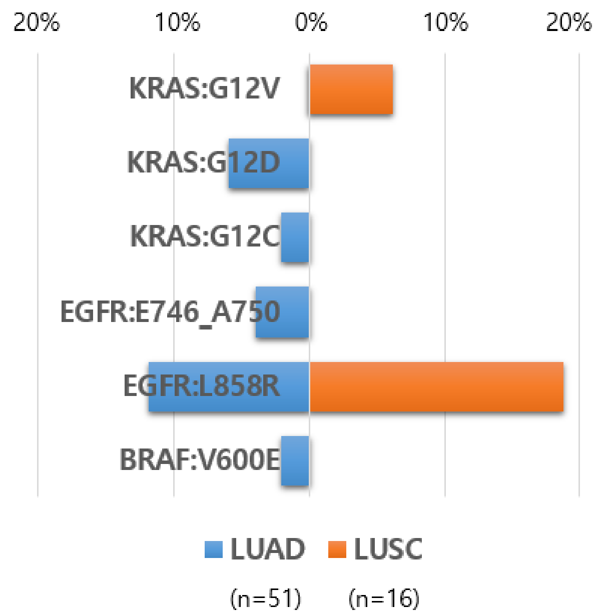
Actionable Variant Proportion Comparison (TCGA)



Gene:HGVSp	*P*-value
KRAS:G12V	0.000
KRAS:G12D	0.000
KRAS:G12C	0.000
EGFR:E746_A750	0.000
EGFR:L858R	0.000
BRAF:V600E	0.020

AJOU

Actionable Variant Proportion Comparison (AJOU)



Gene:HGVSp	*P*-value
KRAS:G12V	0.537
KRAS:G12D	0.764
KRAS:G12C	1.000
EGFR:E746_A750	1.000
EGFR:L858R	0.768
BRAF:V600E	1.000

[Onco-Achilles]
Future plans for Oncology

- Converting whole cancer patients data from National Insurance Claim data

Cancer statistics across OHDSI networks: ONCO-ACHILLES

Researchers

SCYou Seng Chan You 1 17d

Dear colleagues,

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

> SCYou:
>
> 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 @rchen '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 …

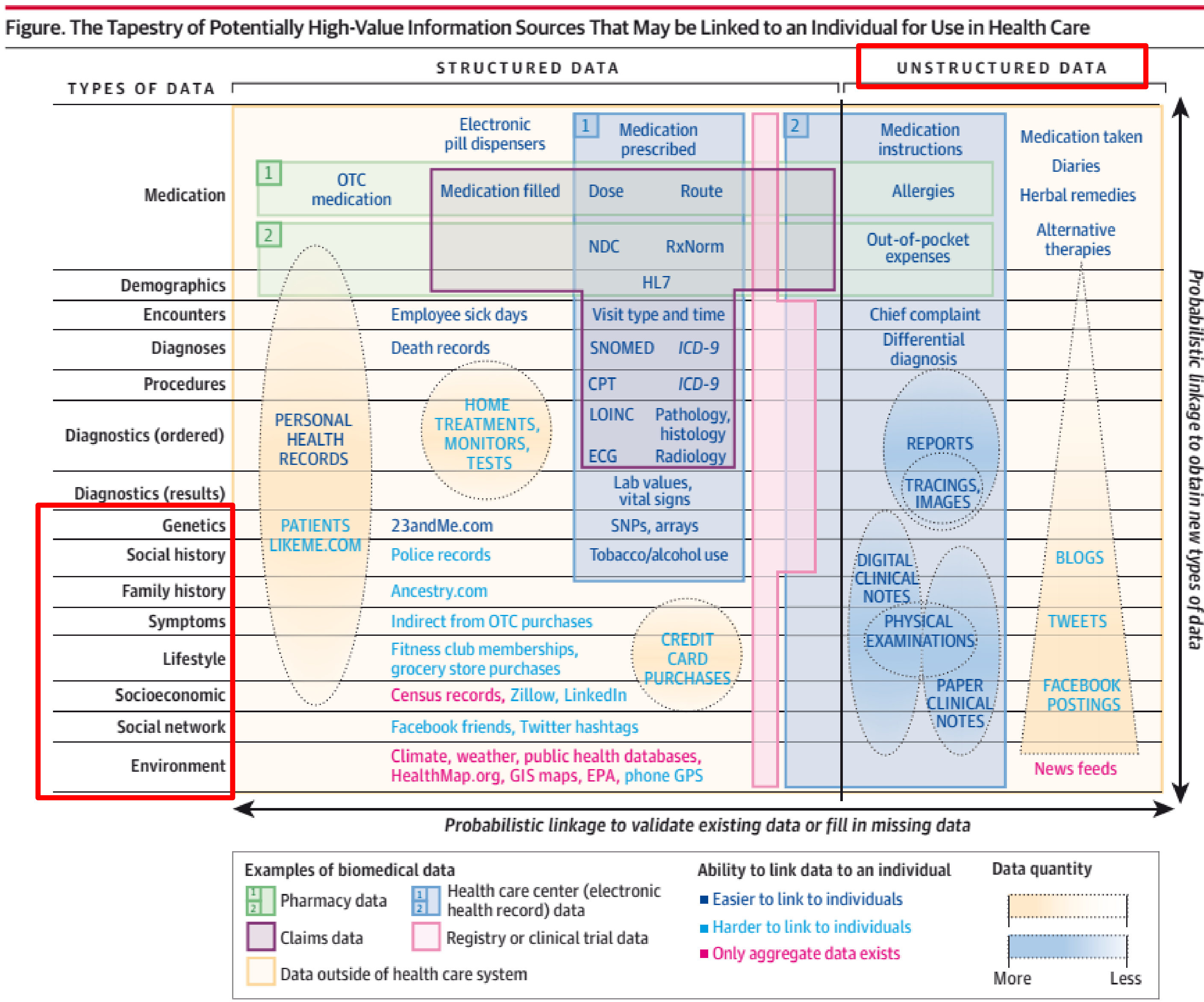
I will extract three components of information from this as the first research:

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]
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



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

Estimation methods

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).

Prediction methods

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.

Method characterization

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.

ATLAS

Supporting packages

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.

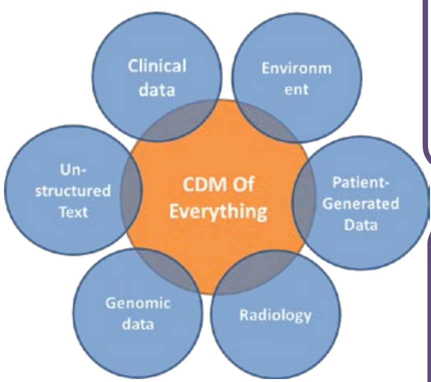
Cyclops
Highly efficient implementation of regularized logistic, Poisson and Cox regression.

Ohdsi R Tools
Support tools that didn't fit other categories, including tools for maintaining R libraries.

Under construction

OHDSI Tools Ecosystem with CDM of Everything

CDM of Everything → DATABASE CONNECTION → Phenotyping & Cohort Generation → Feature Extraction → Prediction & Estimation



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.

ATLAS

Cohort generation by Phenotyping

Feature Extraction

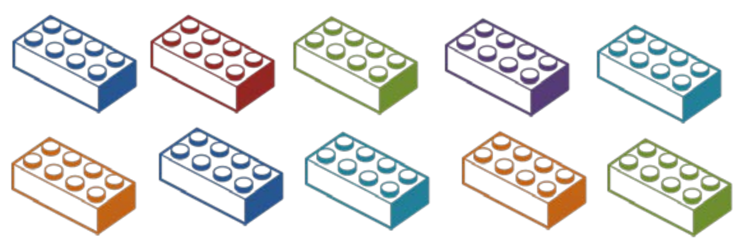
Automatically extract large sets of features for user-specified cohorts using data in the CDM.

Patient Level Prediction

Build and evaluate predictive models for user-specified outcomes, using a wide array of machine learning algorithms.

Cohort Method

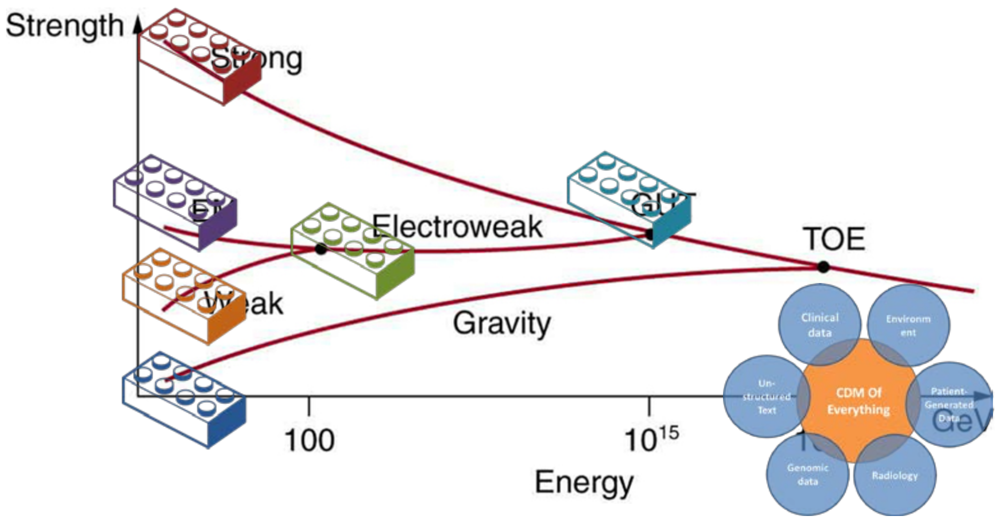
New-user cohort studies using large-scale regression for propensity and outcome models



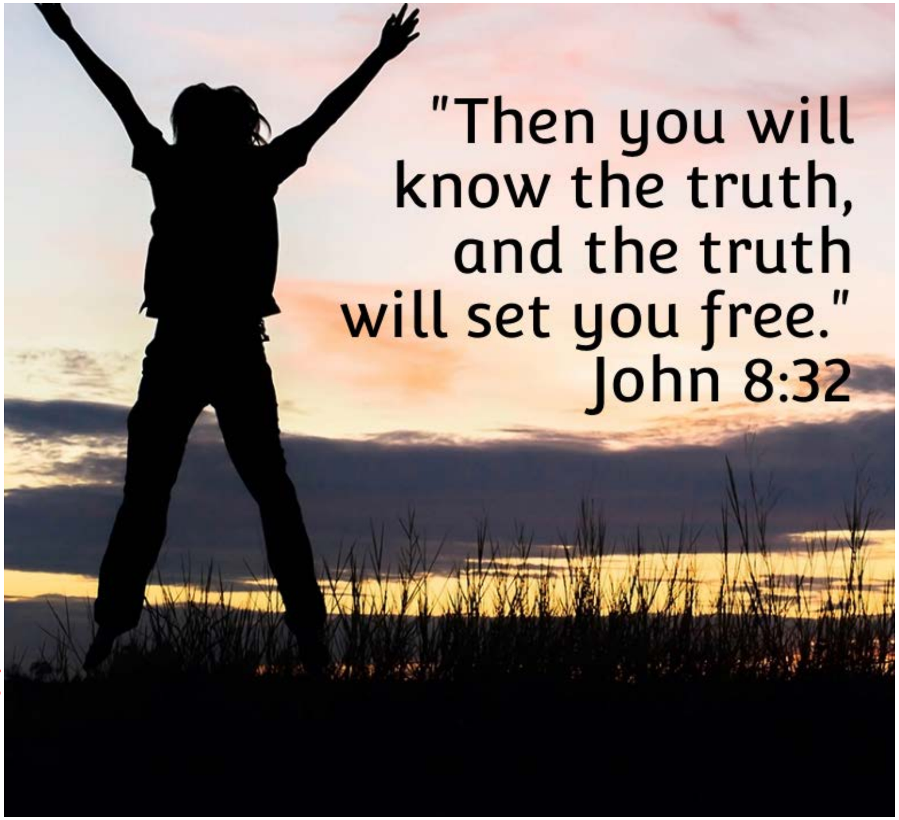
Symmetry in medical data

- 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

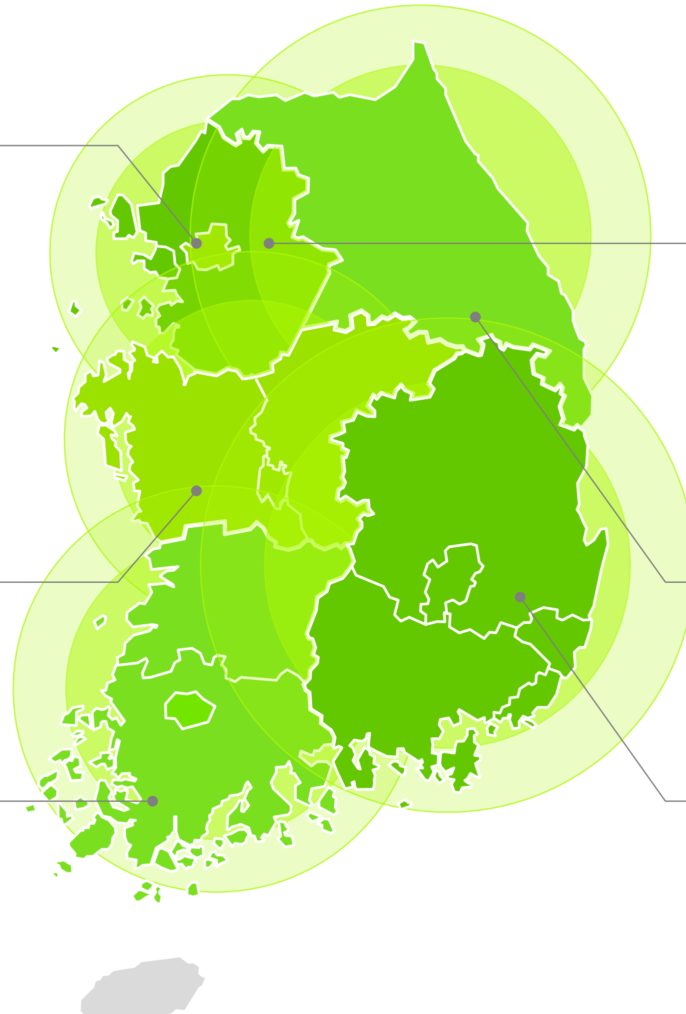


CDM Of Everything



Status of Korean OHDSI Network

Data Network of 41 Hospitals, 55M Patients



Seoul
- 세브란스병원
- 고려대학교안암병원
- 고려대학교구로병원
- 삼성서울병원
- 강동경희대학교병원
- 경희대학교병원
- 건국대학교병원
- 원자력병원
- 한양대학교의료원
- 중앙보훈병원
- 이화여자대학교 의과대학 부속 목동병원
- 강북삼성병원
- 한림대학교강동성심병원
- 순천향대학교 서울병원
- 연세대학교의과대학 강남세브란스병원

Chungcheong
- 순천향대학교 천안병원
- 건양대학교병원

Jeolla
- 원광대학교병원
- 전북대학교병원
- 전라북도 군산의료원
- 진안군의료원
- 전라북도 남원의료원
- 예수병원

Incheon / Gyeonggi
- 순천향대학교 부천병원
- 인하대병원
- 아주대학교병원
- 고려대학교안산병원
- 가천대길병원
- 용인세브란스병원
- 동국대학교일산병원
- CHA 의과학대학교 차병원

Gangwon
- 원주세브란스기독병원
- 강원대학교병원

Gyeongsang
- 순천향대학교 구미병원
- KNUH 경북대학교병원
- 부산대학교병원
- 동국대학교경주병원
- 칠곡경북대학교병원
- 대구가톨릭대학교병원

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).
  과학 혁명은 지식혁명이 아니었다. 무엇보다 무지의 혁명이었다. 과학혁명을 출발시킨 위대한 발견은 **인류는 가장 중요한 질문들에 대한 해답을 모른다**는 발견이었다.
- 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)
  농업 혁명 이후 수천 년에 이르는 인간의 역사를 이해하려는 시도는 단 하나의 질문으로 귀결된다. 인류는 어떻게 자신들을 **대규모 협력망**으로 엮었는가? 그런 망을 지탱할 생물할적 본능이 결핍된 상태에서 말이다.


Thank You
for your time