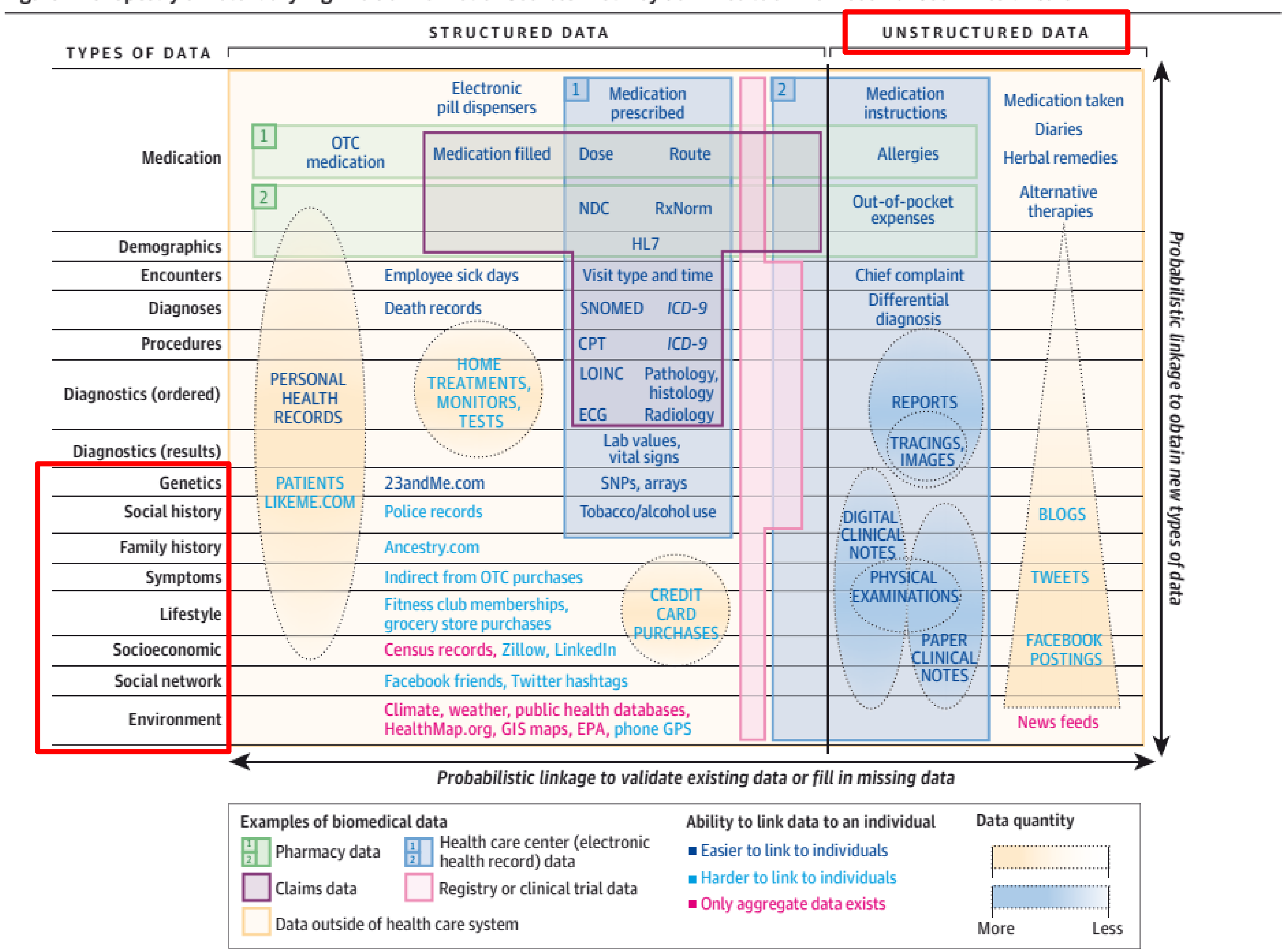
Integration of heterogeneous medical data based on common data model -2nd Part of FEEDER-NET

Seng Chan You

Finding the missing link for big biomedical data

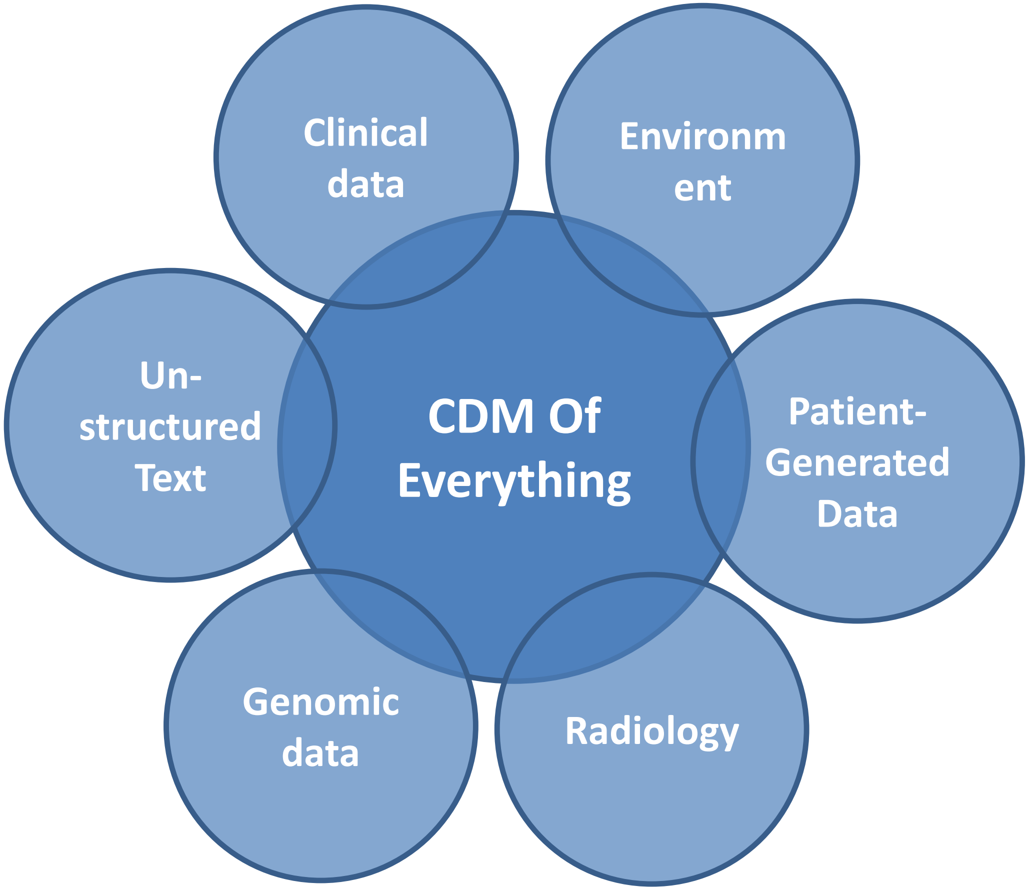
Figure. The Tapestry of Potentially High-Value Information Sources That May be Linked to an Individual for Use in Health Care



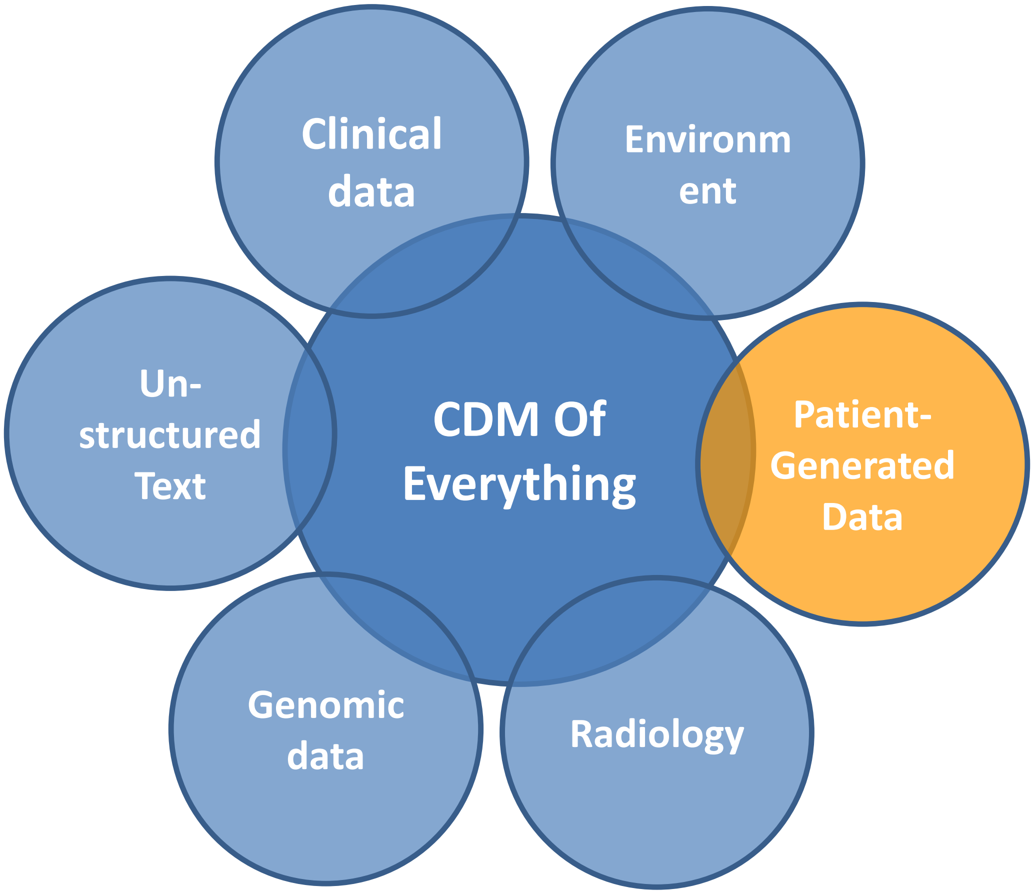
Weber et al., "Finding the Missing Link for Big Biomedical Data." JAMA (June 25, 2014)

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



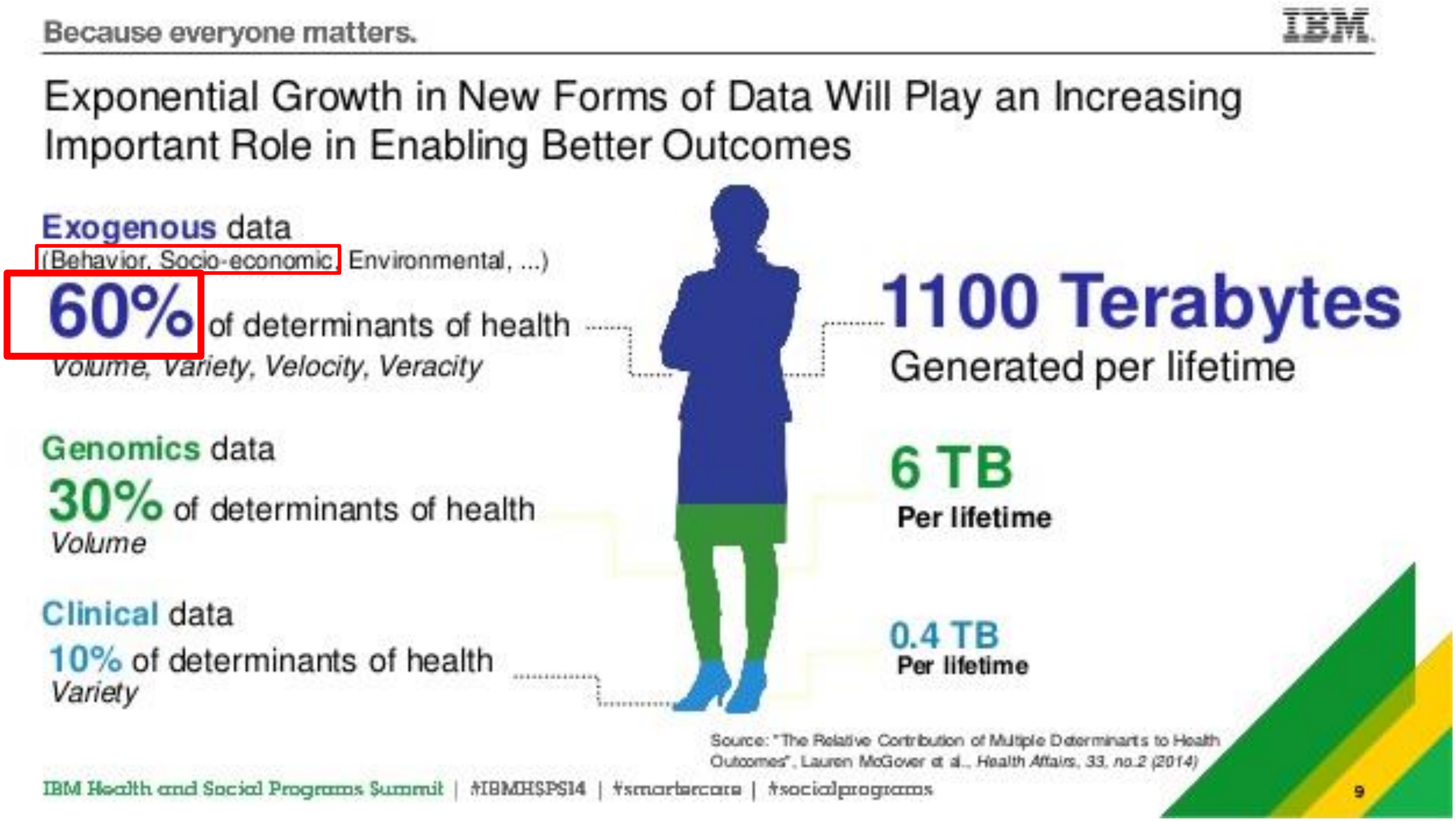
Common Data Model of Everything in Medicine



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

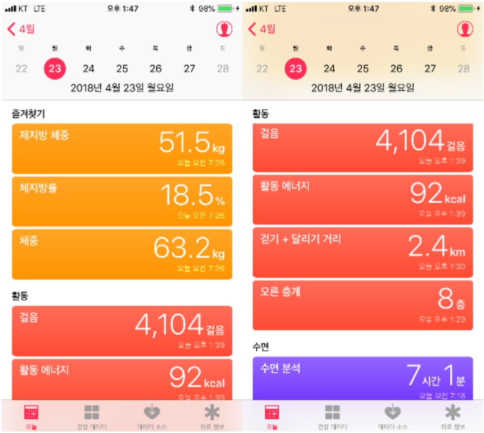
[1]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea;
[2]Medicine, Noom, Inc, Seoul, Korea
[3]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea

Patient-Generated Health Data

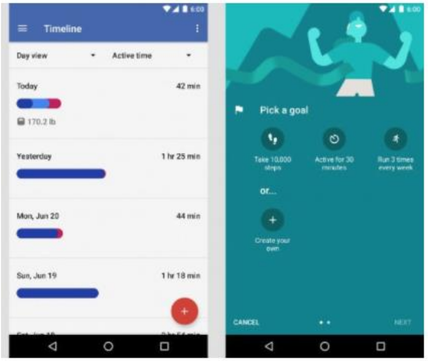


Applications in smartphone collecting health data

Apple Health

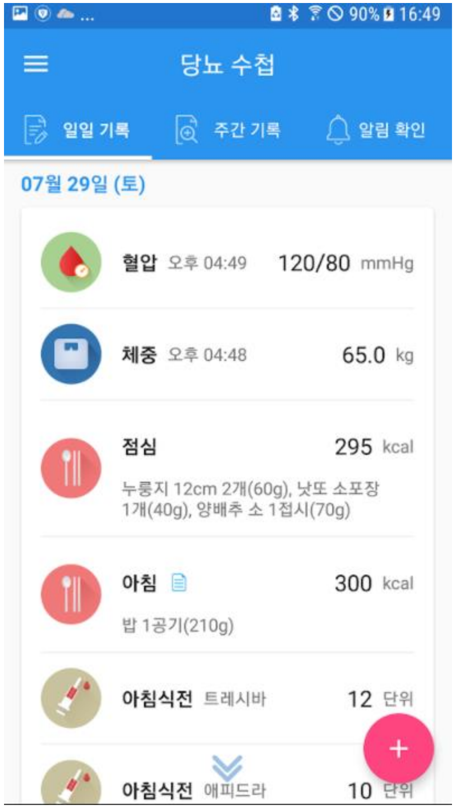


Google Fit

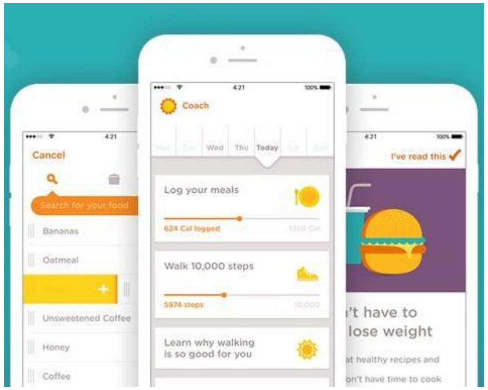


Samsung Medical Center

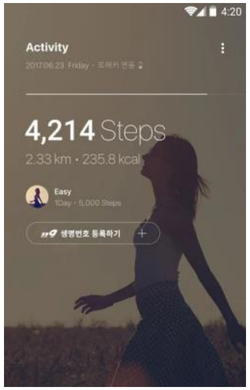
Diabetes Note



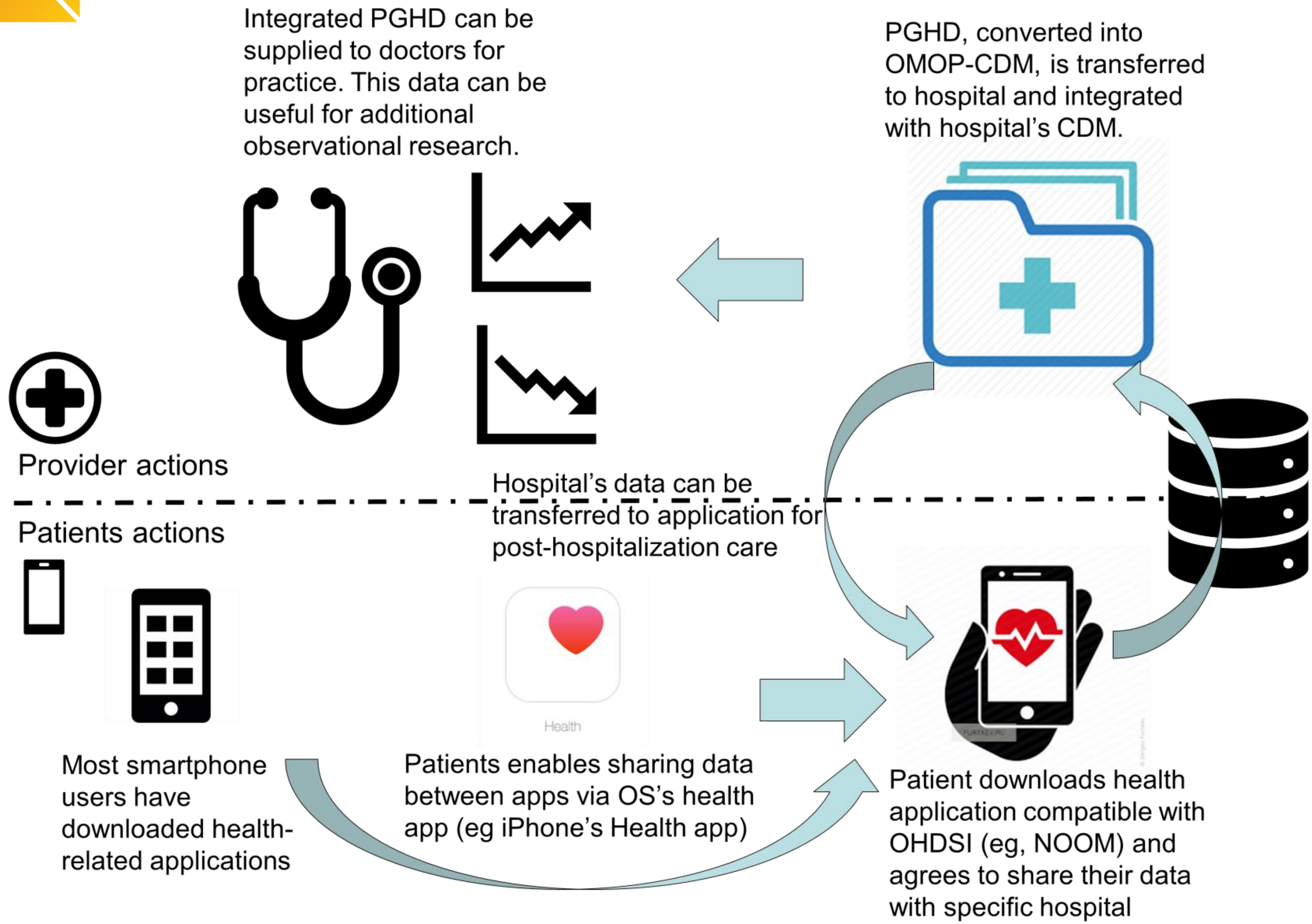
NOOM



Efil



Schematic data flow



NOOM converted their data into CDM

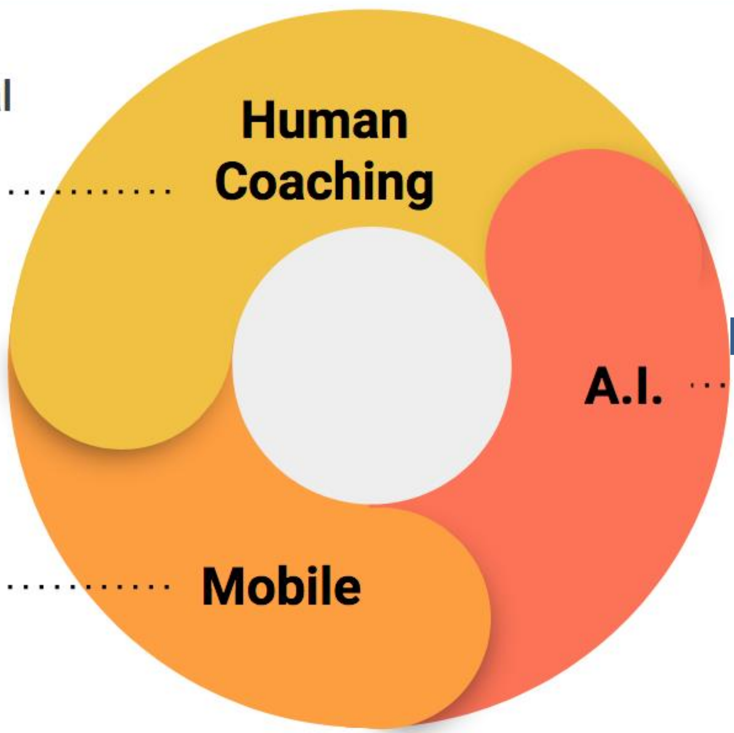
Noom is a behavior change company that uses **A.I., Human Coaching and Mobile Technology** to create the world's most effective solutions for lifestyle & chronic conditions

Use latest psychological approaches (CBT)

Patented scalable coaching using A.I.

Mobile-first company

47 million users

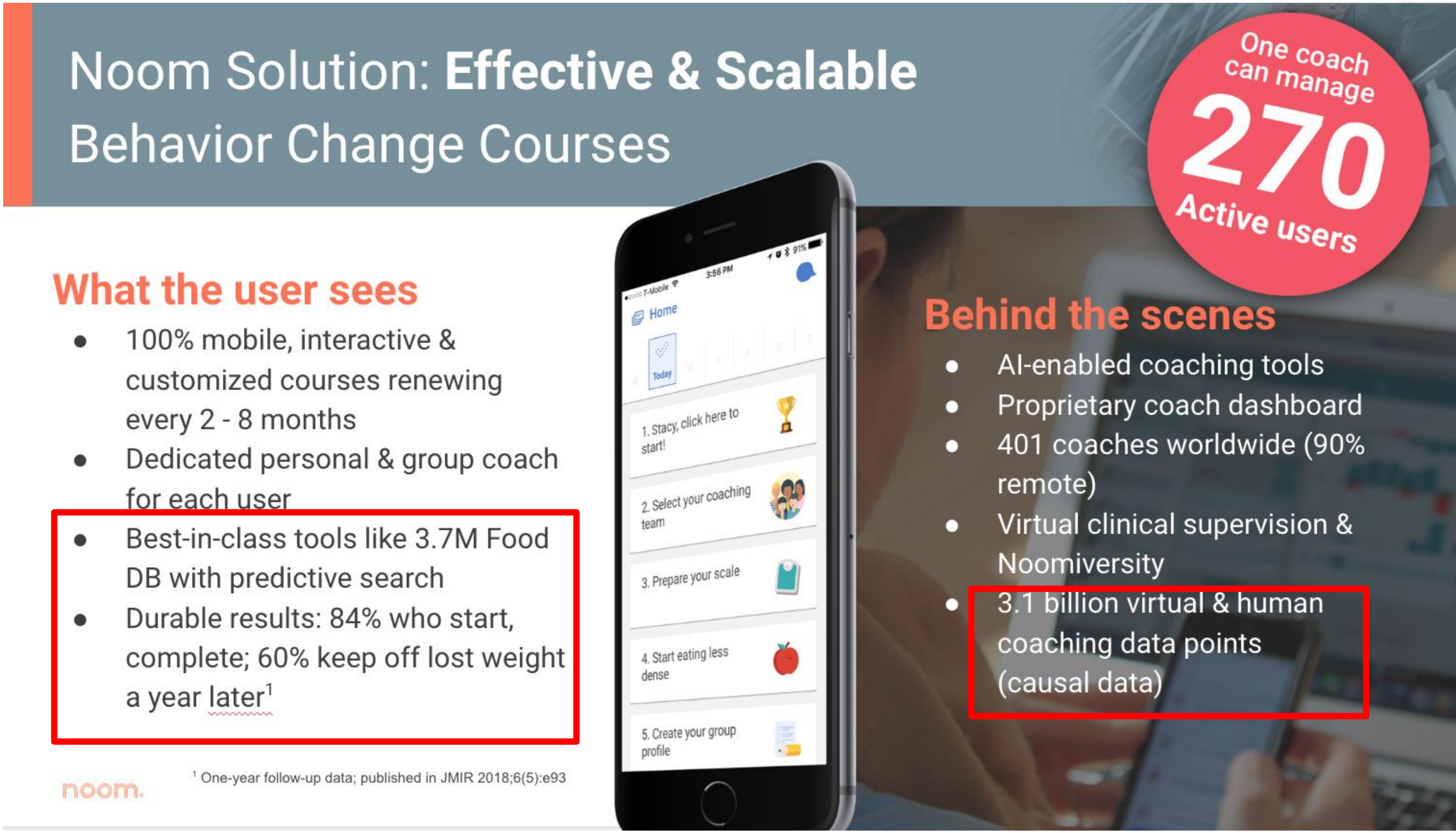


3.1 billion coaching data points

Self-learning A.I.

noom.

NOOM converted their data into CDM



ETL result of sample data from NOOM

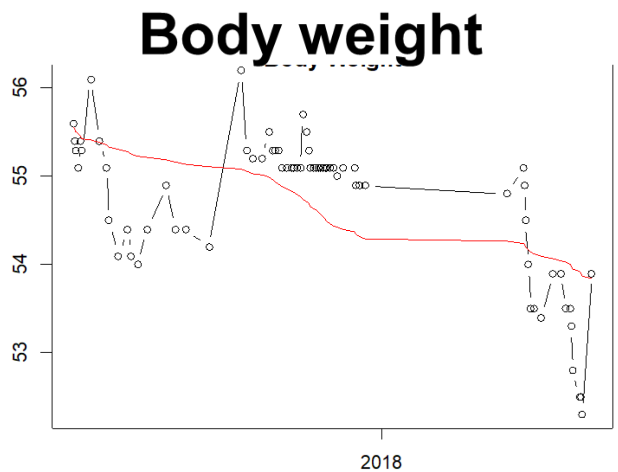
- NOOM converted their sample data (n=100) into CDM
  - weight, daily step count, and daily dietary calories

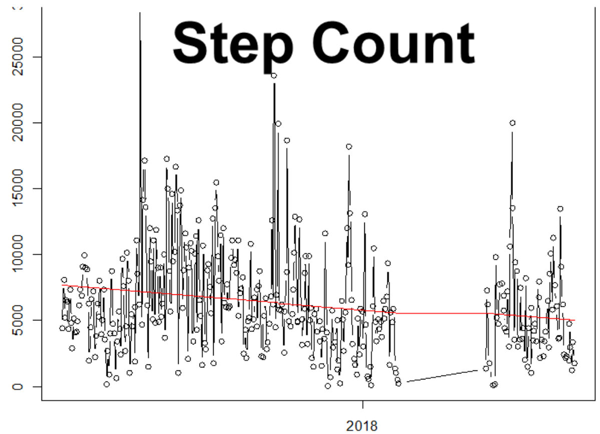
measurem	person_id	measurem	value_sour	unit_sourc	measurem	concept_name	measurement_date	measurement_datetime	value_as_n	unit_conce	unit_conce	measurem
1	1	Weight	103.4	kg	3025315	Body weight	2017-05-08	2017-05-08 22:56	103.4	4122383	kg	44818704
2	1	Weight	108	kg	3025315	Body weight	2017-03-22	2017-03-23 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

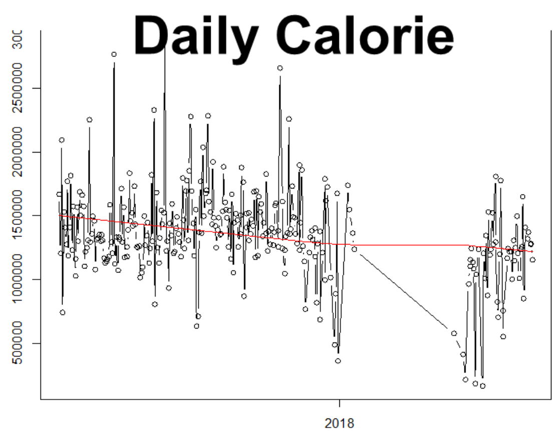
observatio	person_id	observation_source_value	value_sour	unit_sourc	observatio	concept_name	observation_date	value_as_number	unit_conce	unit_conce	observatio	observatio	Observation_type_concept_name
1	1	Steps	9097	count	3034985	Number of steps in 24 hour Measured	2017-07-04	9348	44777556	per 24 hours		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

Visualization of irregular time-series data

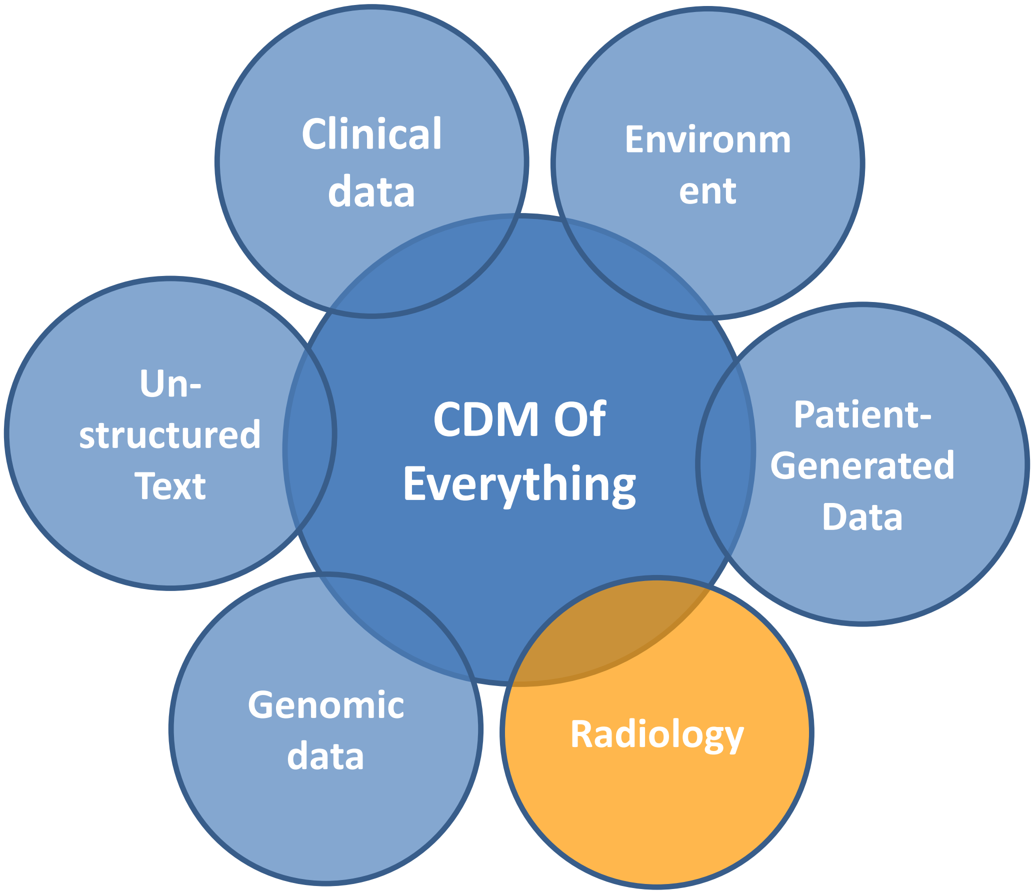
- Identification of **trend** in time-series data might be very difficult, if the patient records it **irregularly**.
- By using **dynamic linear regression**, the trend can be shown in the plot (red line) for better understanding of the data.







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

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

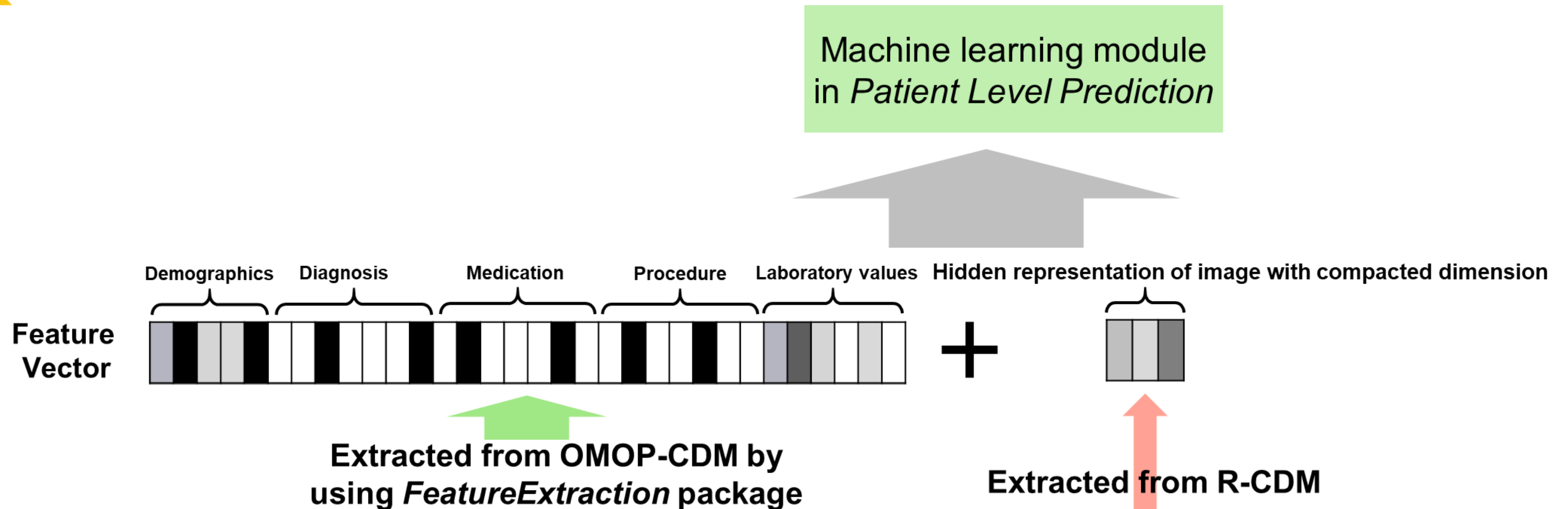
PK	Radiology_occurrence_ID	VARCHAR(255)
	Radiology_occurrence_date	DATE
N	Radiology_occurrence_datetime	DATETIME
	Person_ID	VARCHAR(64)
FK,N	Condition_occurrence_ID	INT
FK	Device_concept_id	VARCHAR(25)
	Radiology_modality_concept_id	VARCHAR(5)
N	Person_orientation_concept_id	VARCHAR(10)
	Radiology_protocol_concept_id	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

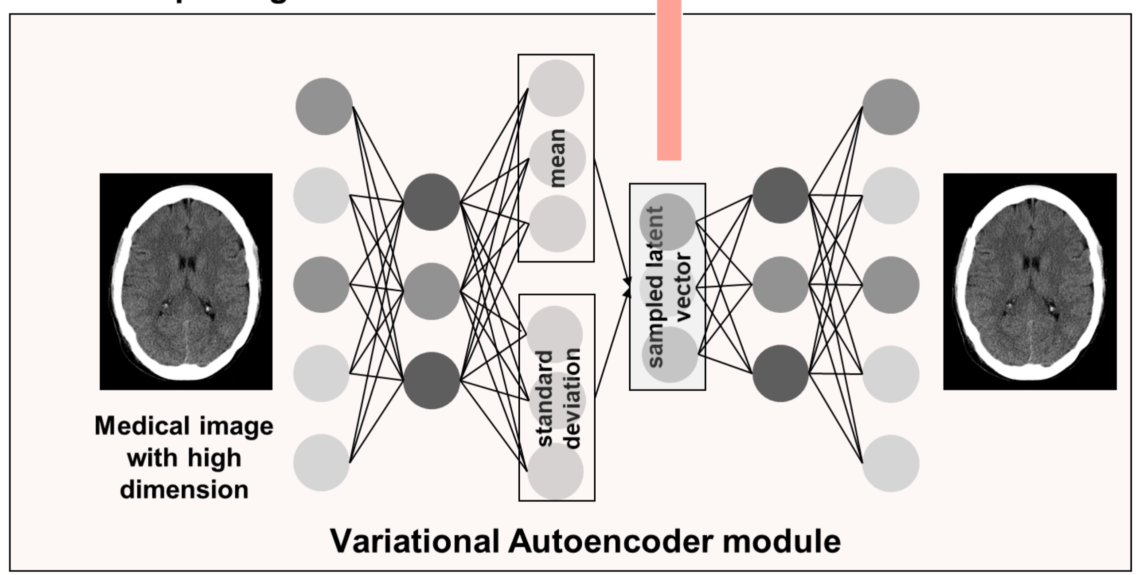
Radiology_Image

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

Combining structured medical information with hidden features from



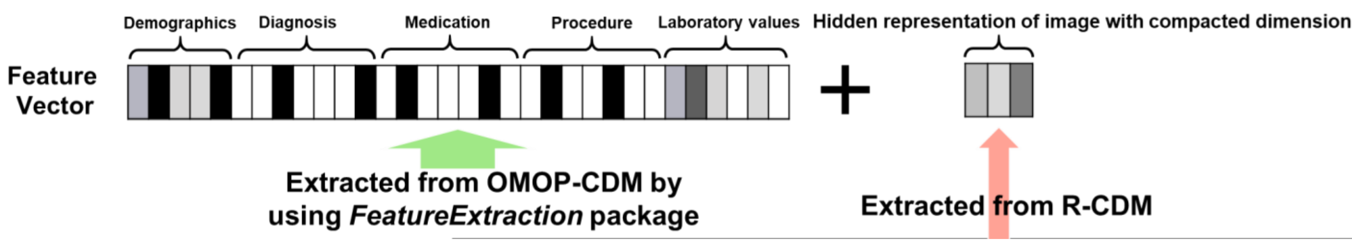


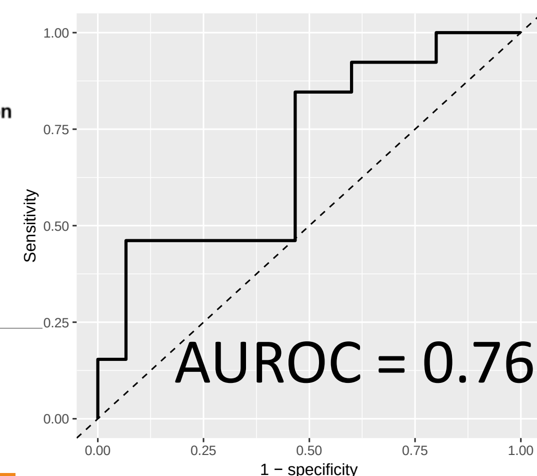
Since **usual radiology images have very high dimension**, we propose to **use hidden representation of images with compacted dimension** by using deep learning auto-encoder

Pilot Study: Prediction of poor functional outcome in ischemic stroke

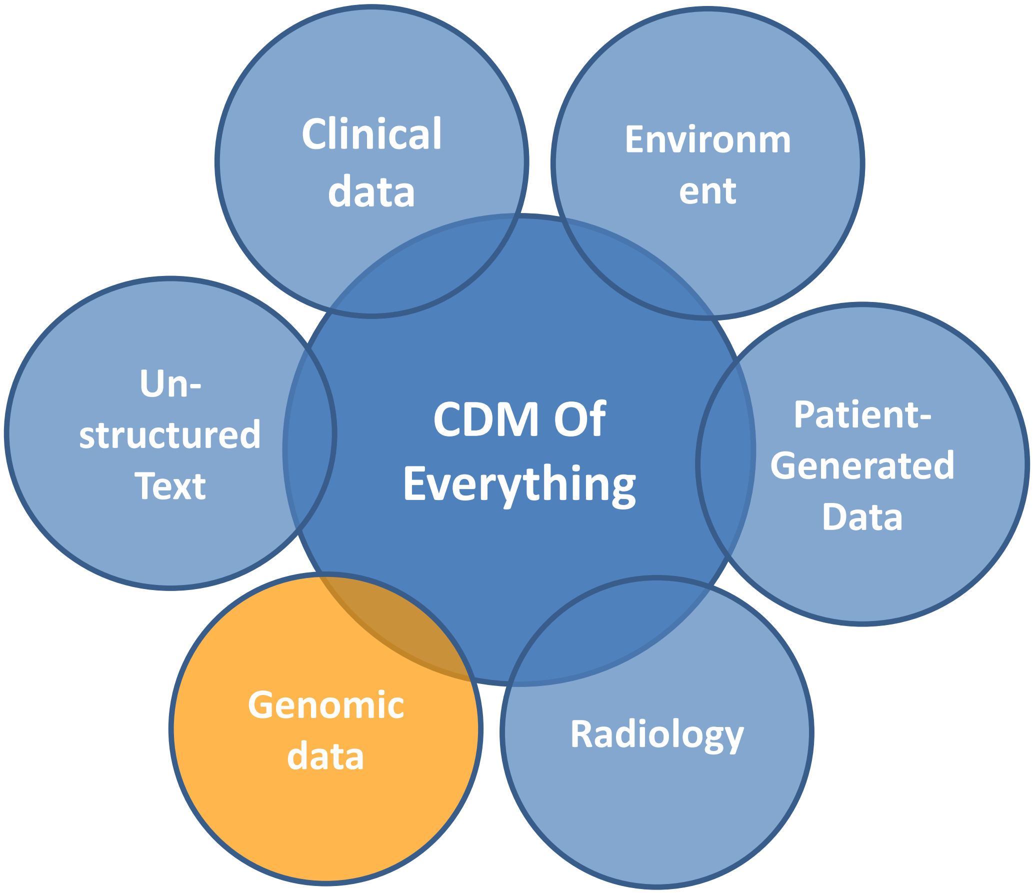
Study Design

- Target cohort: The patients **with ischemic stroke** (n= 141)
- Outcome: **Poor functional outcome** 3 months after stroke, which defined by modified Rankin Scale more than 3 (n= 64)
- Machine learning algorithm: Lasso logistic regression
- Covariates: age group, gender, index year, and procedures combined with **latent feature vector extracted from non-contrast phase of brain CT**



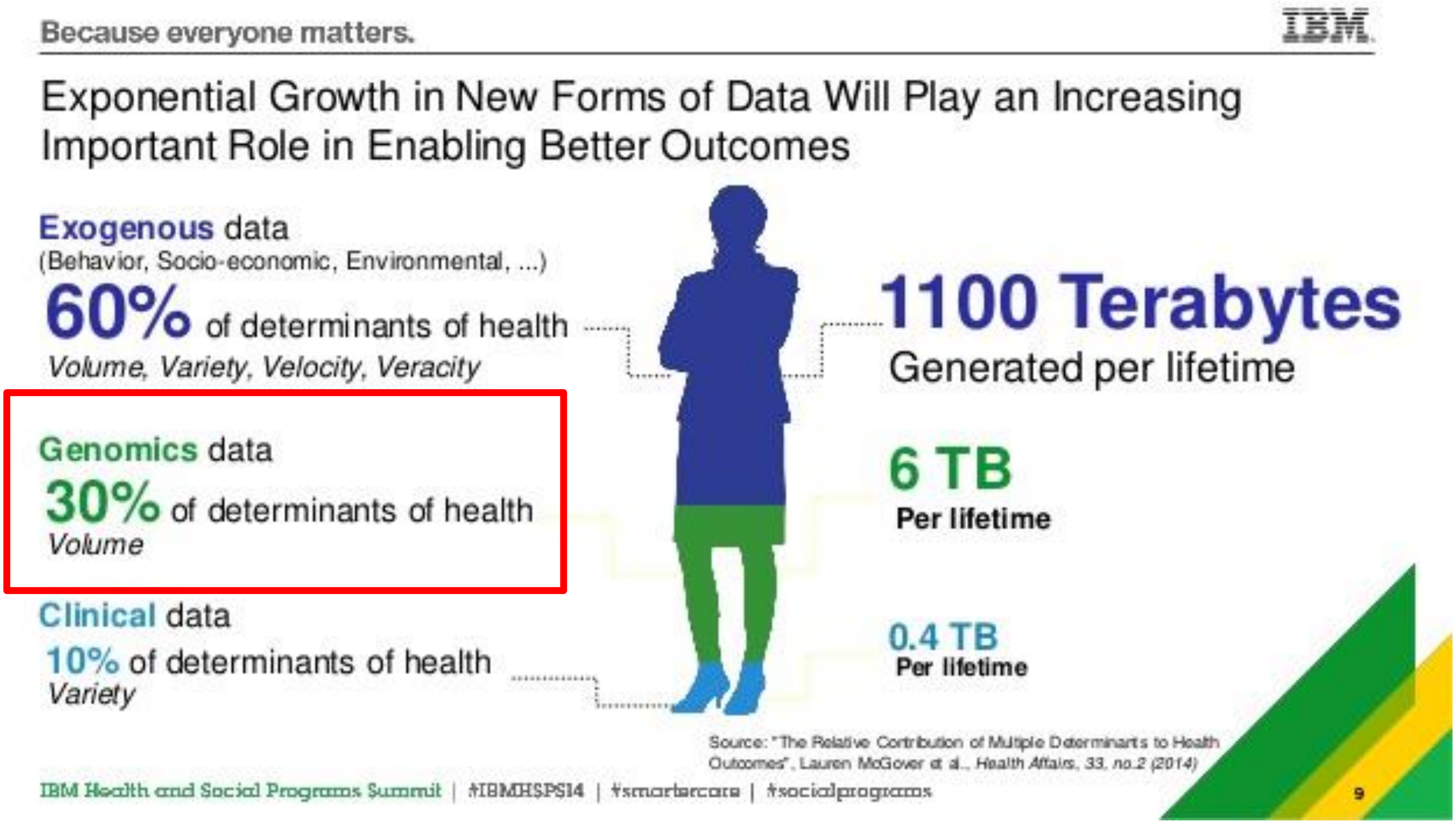


Common Data Model of Everything in Medicine



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
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)
IBM Health and Social Programs Summit | #IBMHSPS14 | #smartercare | #socialprograms
9

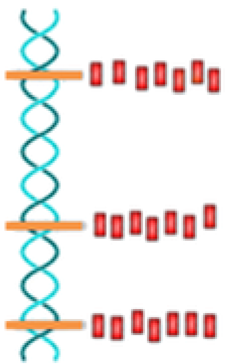
Background: Surge of genomic data

- 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

Genomic Test Process

GENOMIC TEST → TARGET GENE → VARIANT OCCURRENCE → VARIANT ANNOTATION

Next Generation Sequencing (NGS)



EGFR

KRAS

BRAF

p.Leu858Arg

Exon19 Deletion

p.Thr790Met

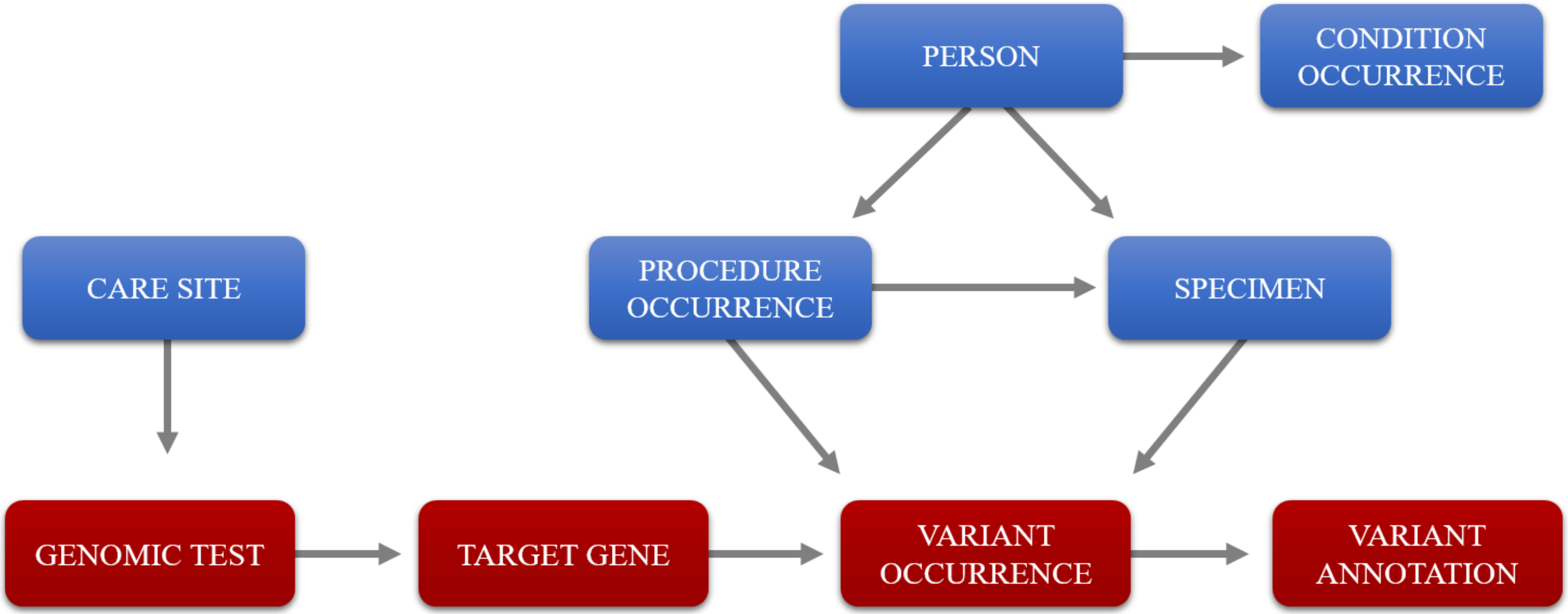
[Clinvar] Pathogenic

[PolyPhen] Damaging

[SIFT] Deleterious

[GNOMAD] 0.686792

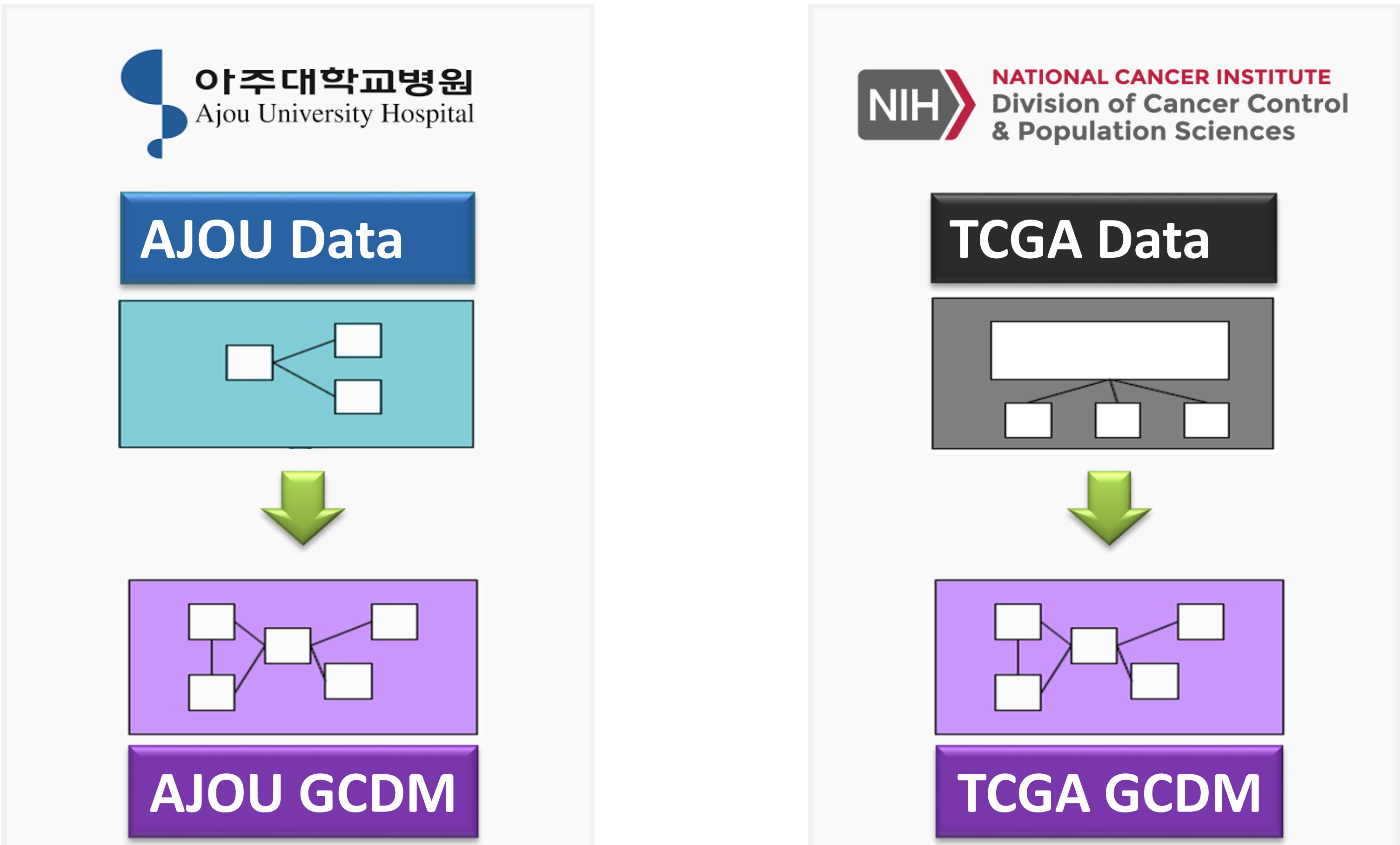
Genomic CDM (G-CDM) Structure



Schematic diagram of the relationship between the tables that make up the GCDM.

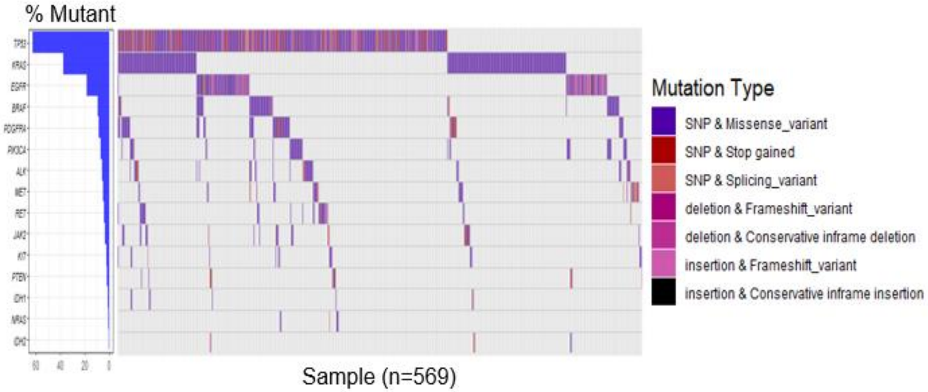
Conversion of G-CDM

- The data structures of the two institutes were unified.

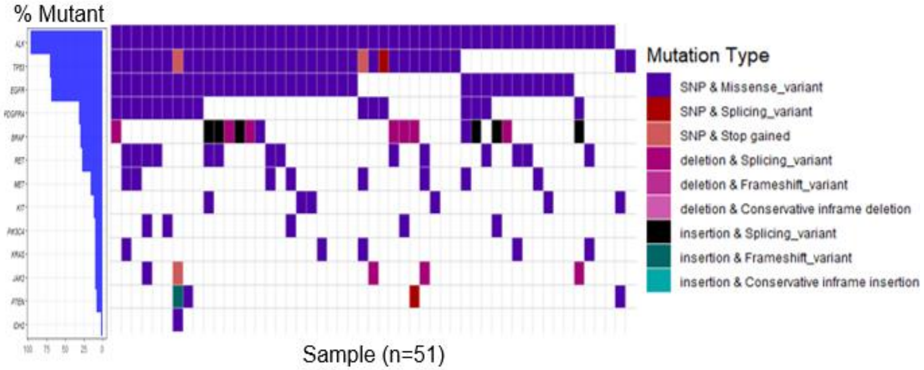


Study Results:
Waterfall plot of adenocarcinoma and squamous cell carcinoma of lung

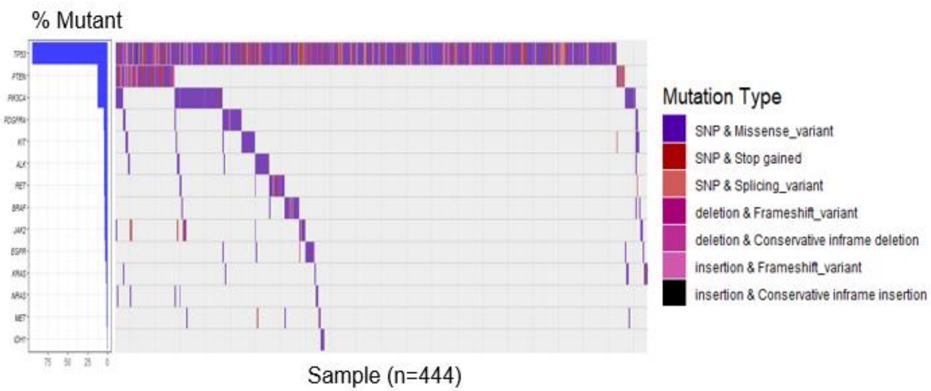
TCGA **LUAD**



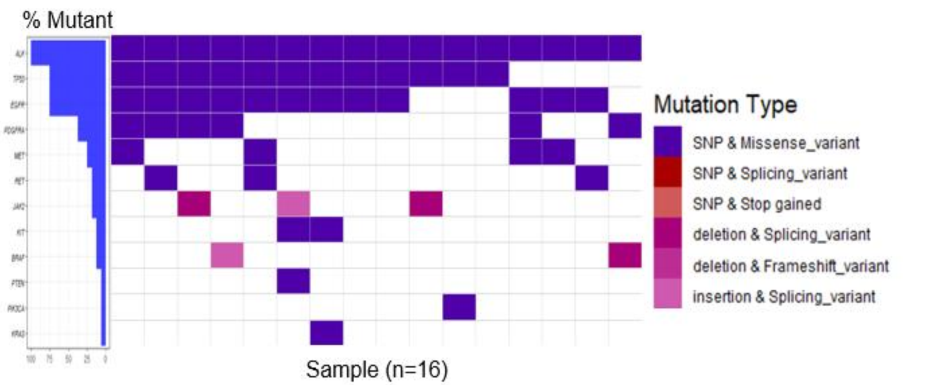
AJOU **LUAD**



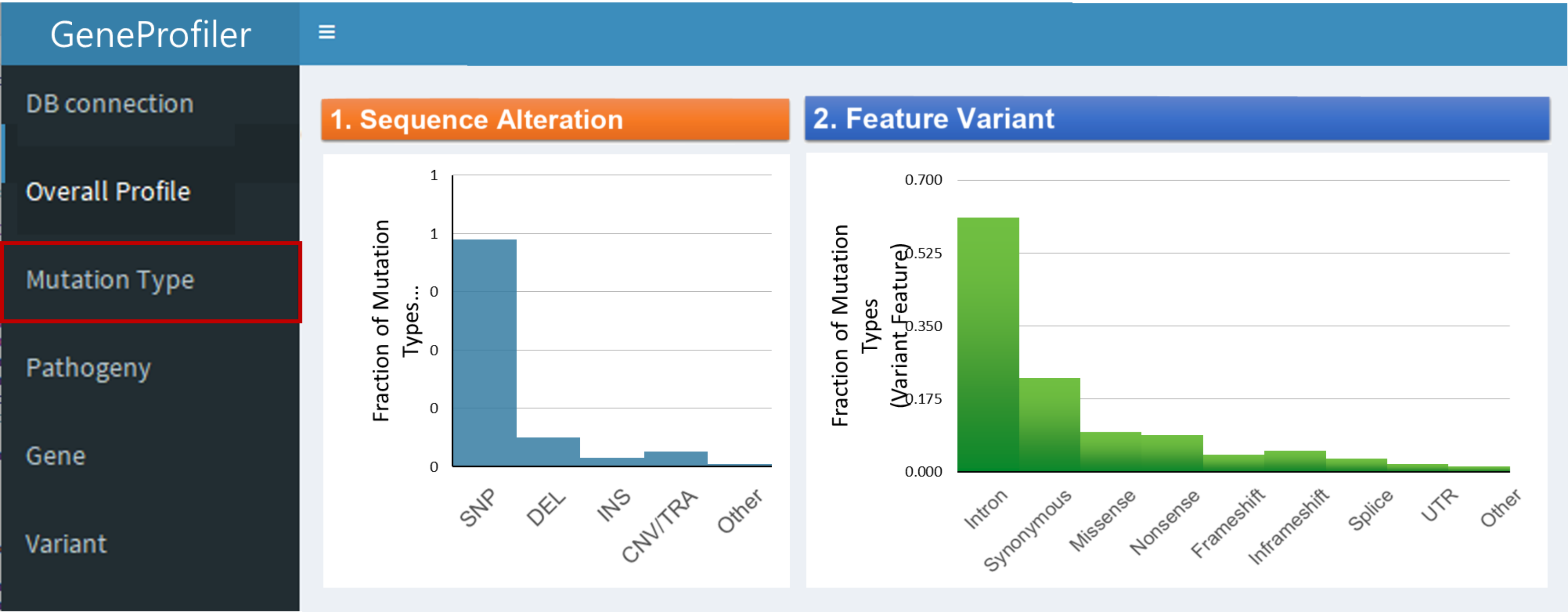
TCGA **LUSC**



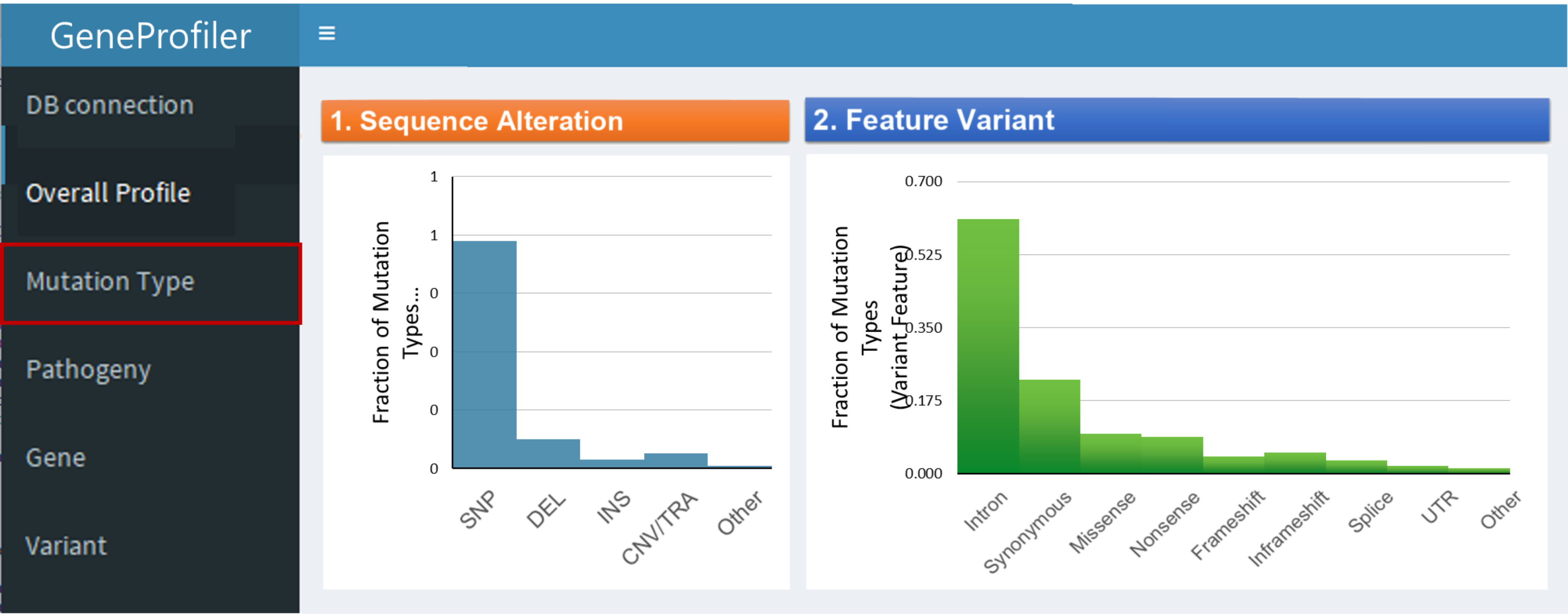
AJOU **LUSC**



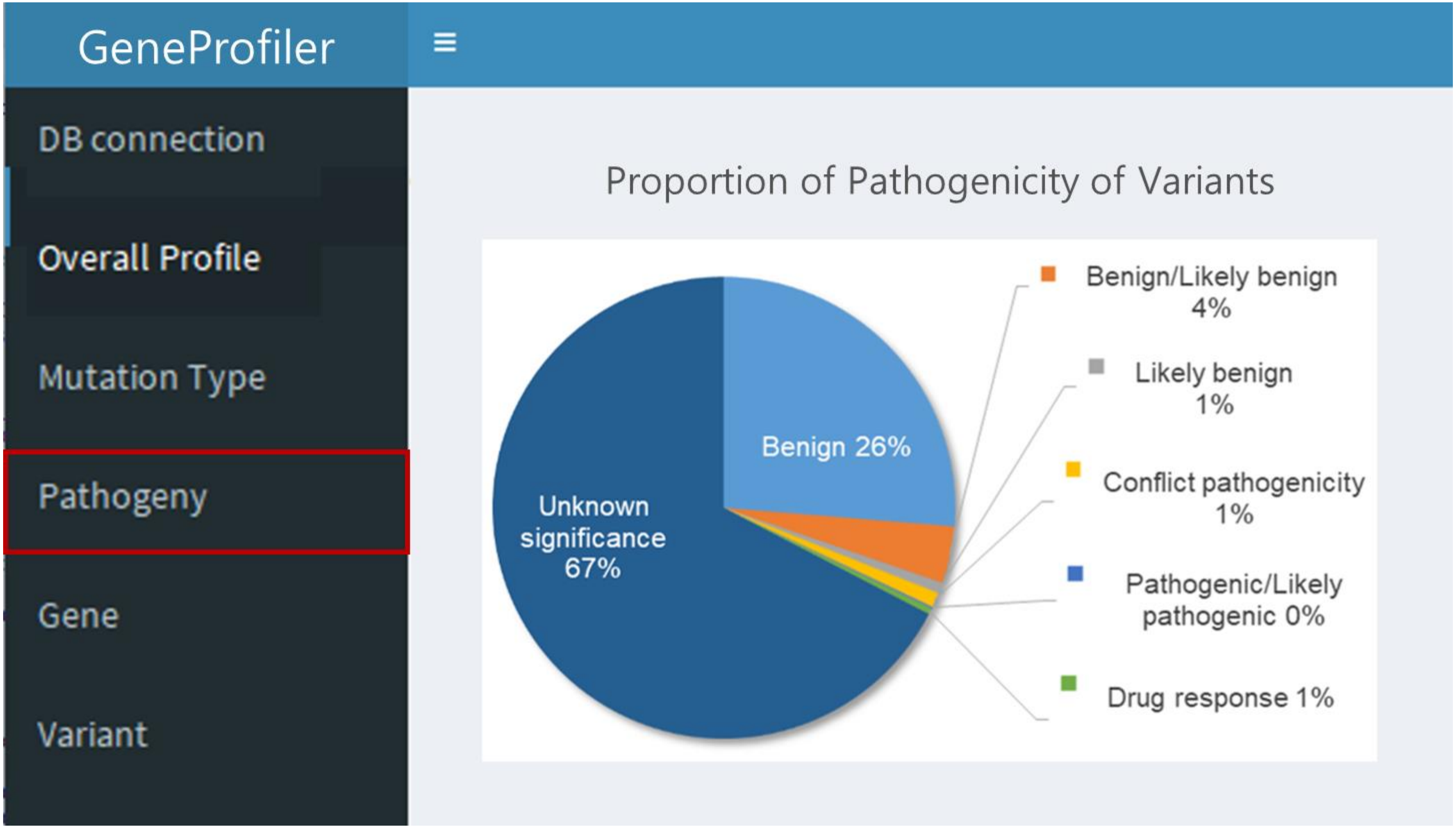
Gene profiler:
Visualization of structural and functional variant



Gene profiler:
Visualization of structural and functional variant

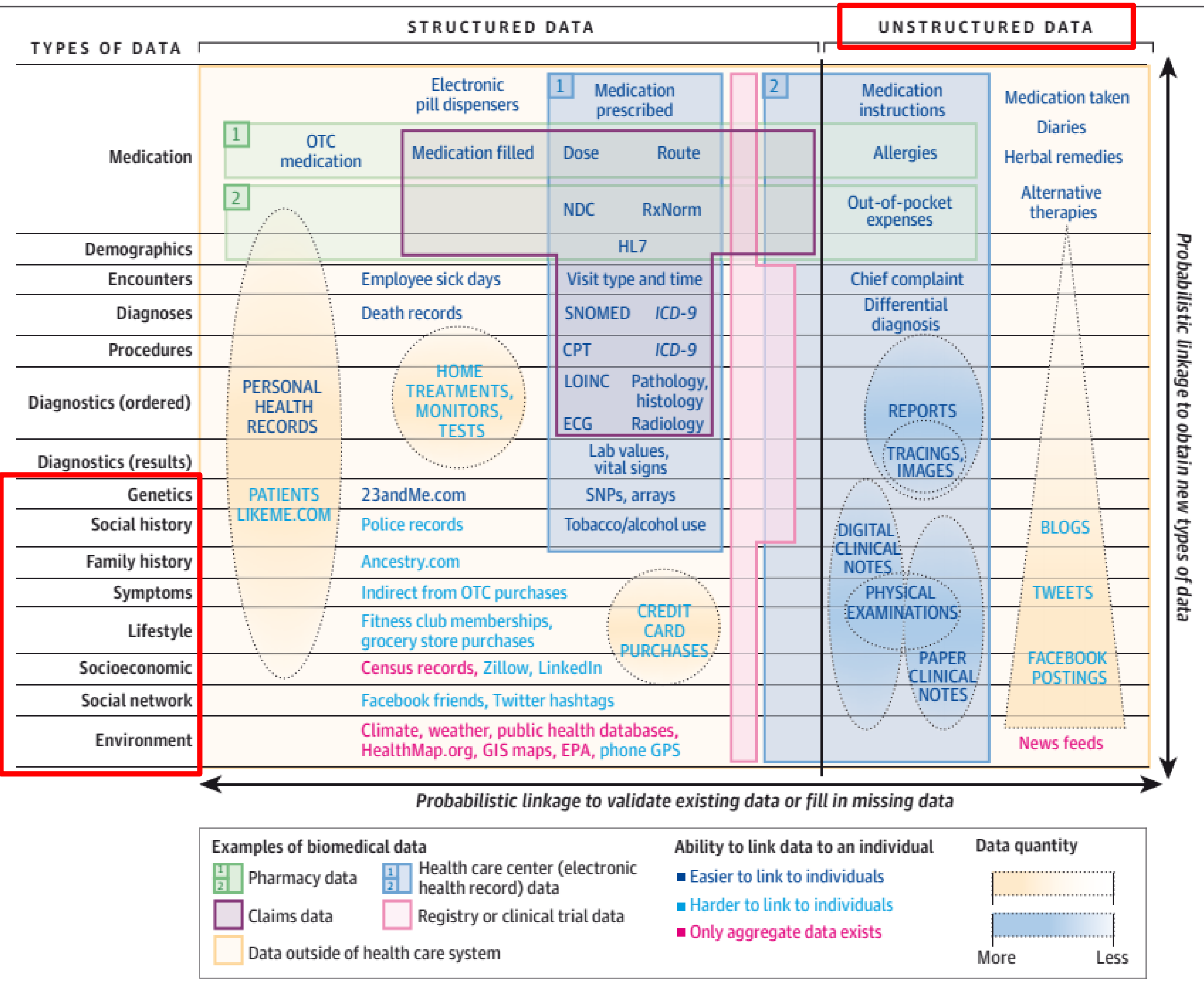


Gene profiler: Visualization of proportion of pathogenicity of variants



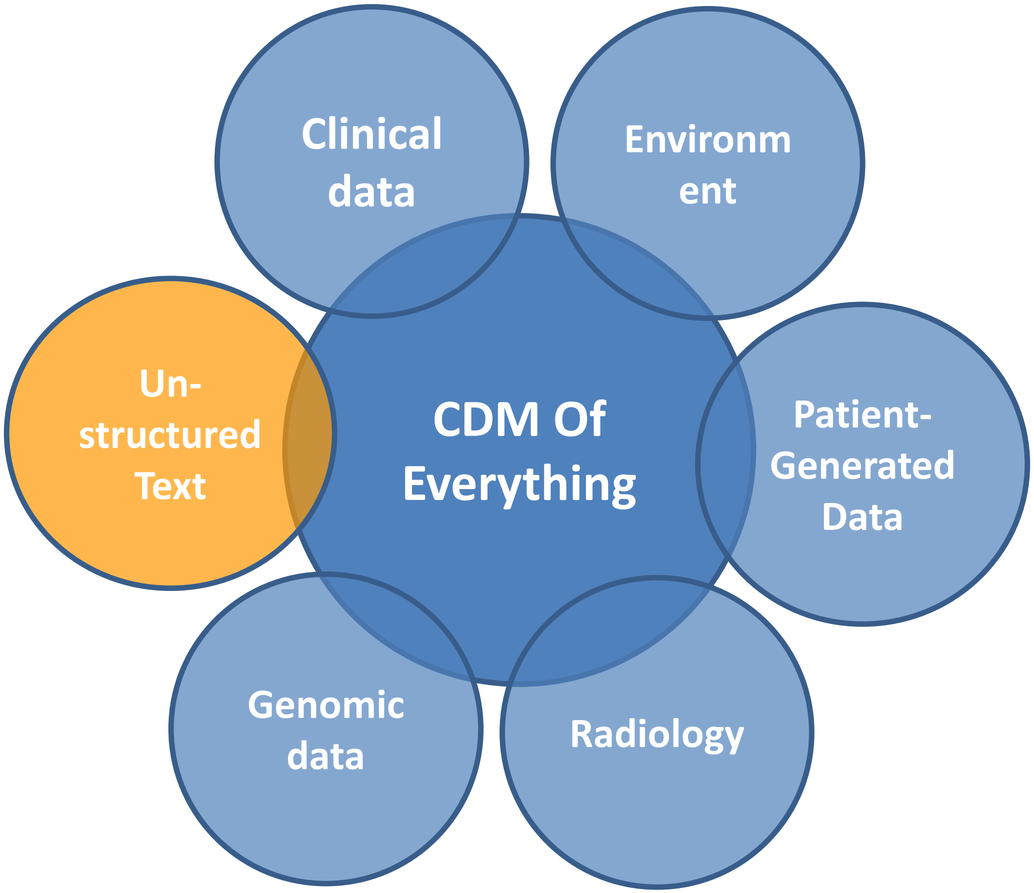
Connecting the missing link for big biomedical data

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



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

Common Data Model of Everything in Medicine



Giup Jang[1], Seng Chan You, MD[2], Dongsu Park[2], Youngmi Yoon, PhD[3], Rae Woong Park, MD, PhD[2,4]

[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

Cross-Language Natural Language Processing based on OMOP-CDM

- We aim to develop the **cross-language natural language processing (NLP) module** for medical free-text in OMOP-CDM by using topic modeling.
- To demonstrate the feasibility, we build prediction model for 30-day readmission through emergency room by **combining features from structured clinical data and unstructured free-text in discharge note** in OMOP-CDM

Overall Process

Tokenization of medical free-text based on medical dictionary

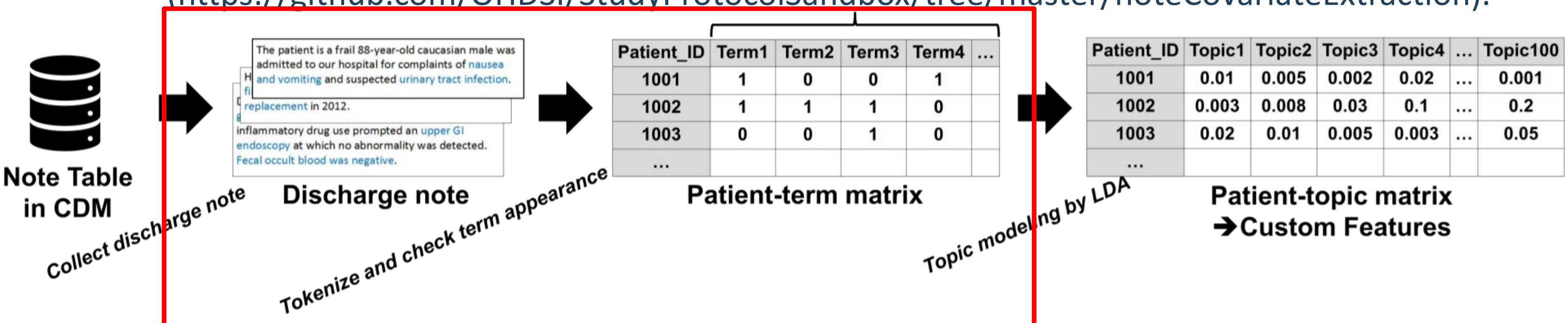
- Tokenization is to divide a sentence into a minimum number of meaningful units.

Topic modeling by Latent Dirichlet Allocation (LDA)

- Topic modeling allows a document to have multiple topics and to analyze the characteristics of the document in more detail than common cluster method.
- LDA is one of the topic modeling algorithm, and it is highly modular and can be easily extended.

Extracting features from note

- Values for each topic estimated from the note by topic modeling were allocated into individual covariates. We developed *noteCoavariateExtraction* function, which is compatible with OHDSI tool ecosystem (https://github.com/OHDSI/StudyProtocolSandbox/tree/master/noteCovariateExtraction).



Patient_ID	Term1	Term2	Term3	Term4	…
1001	1	0	0	1	
1002	1	1	1	0	
1003	0	0	1	0	
…					

Patient_ID	Topic1	Topic2	Topic3	Topic4	…	Topic100
1001	0.01	0.005	0.002	0.02	…	0.001
1002	0.003	0.008	0.03	0.1	…	0.2
1003	0.02	0.01	0.005	0.003	…	0.05
…						

Overall Process

Tokenization of medical free-text based on medical dictionary

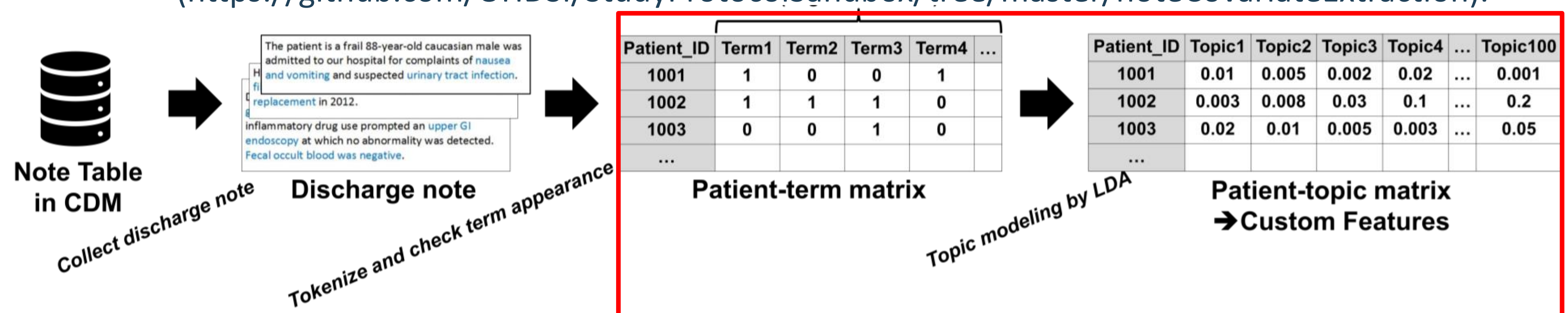
- Tokenization is to divide a sentence into a minimum number of meaningful units.

Topic modeling by Latent Dirichlet Allocation (LDA)

- Topic modeling allows a document to have multiple topics and to analyze the characteristics of the document in more detail than common cluster method.
- LDA is one of the topic modeling algorithm, and it is highly modular and can be easily extended.

Extracting features from note

- Values for each topic estimated from the note by topic modeling were allocated into individual covariates. We developed *noteCoavariateExtraction* function, which is compatible with OHDSI tool ecosystem (https://github.com/OHDSI/StudyProtocolSandbox/tree/master/noteCovariateExtraction).



Patient_ID	Term1	Term2	Term3	Term4	...
1001	1	0	0	1	
1002	1	1	1	0	
1003	0	0	1	0	
...					

Patient_ID	Topic1	Topic2	Topic3	Topic4	...	Topic100
1001	0.01	0.005	0.002	0.02	...	0.001
1002	0.003	0.008	0.03	0.1	...	0.2
1003	0.02	0.01	0.005	0.003	...	0.05
...						

Overall Process

Tokenization of medical free-text based on medical dictionary

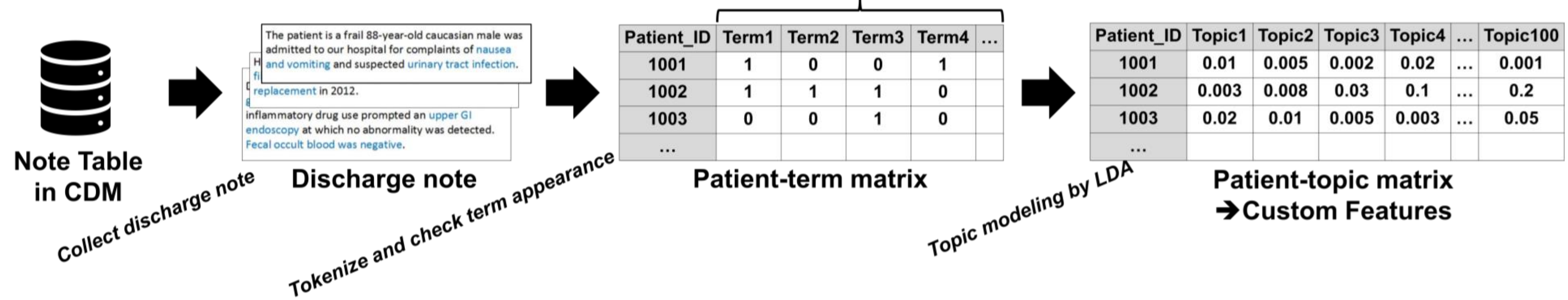
- Tokenization is to divide a sentence into a minimum number of meaningful units.

Topic modeling by Latent Dirichlet Allocation (LDA)

- Topic modeling allows a document to have multiple topics and to analyze the characteristics of the document in more detail than common cluster method.
- LDA is one of the topic modeling algorithm, and it is highly modular and can be easily extended.

Extracting features from note

- Values for each topic estimated from the note by topic modeling were allocated into individual covariates. We developed *noteCoavariateExtraction* function, which is compatible with OHDSI tool ecosystem (https://github.com/OHDSI/StudyProtocolSandbox/tree/master/noteCovariateExtraction).



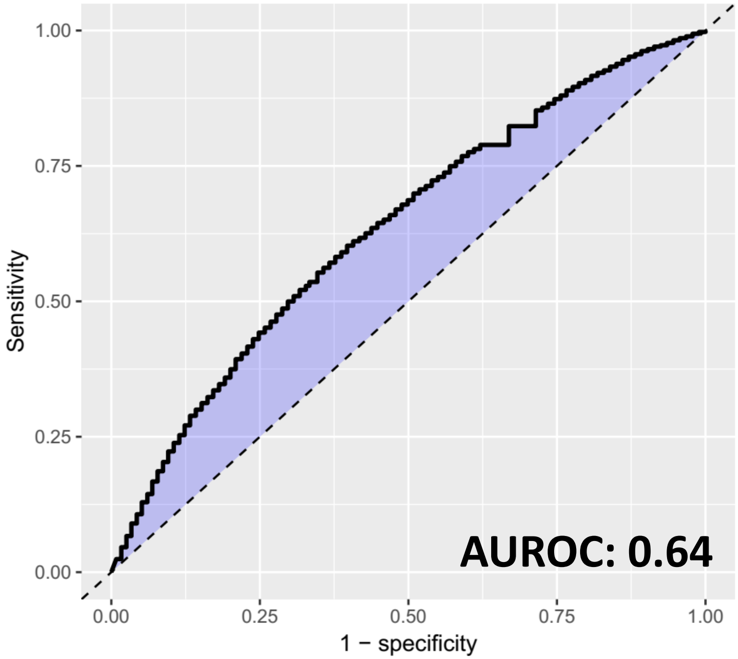
Patient_ID	Term1	Term2	Term3	Term4	…
1001	1	0	0	1	
1002	1	1	1	0	
1003	0	0	1	0	
…					

Patient_ID	Topic1	Topic2	Topic3	Topic4	…	Topic100
1001	0.01	0.005	0.002	0.02	…	0.001
1002	0.003	0.008	0.03	0.1	…	0.2
1003	0.02	0.01	0.005	0.003	…	0.05
…						

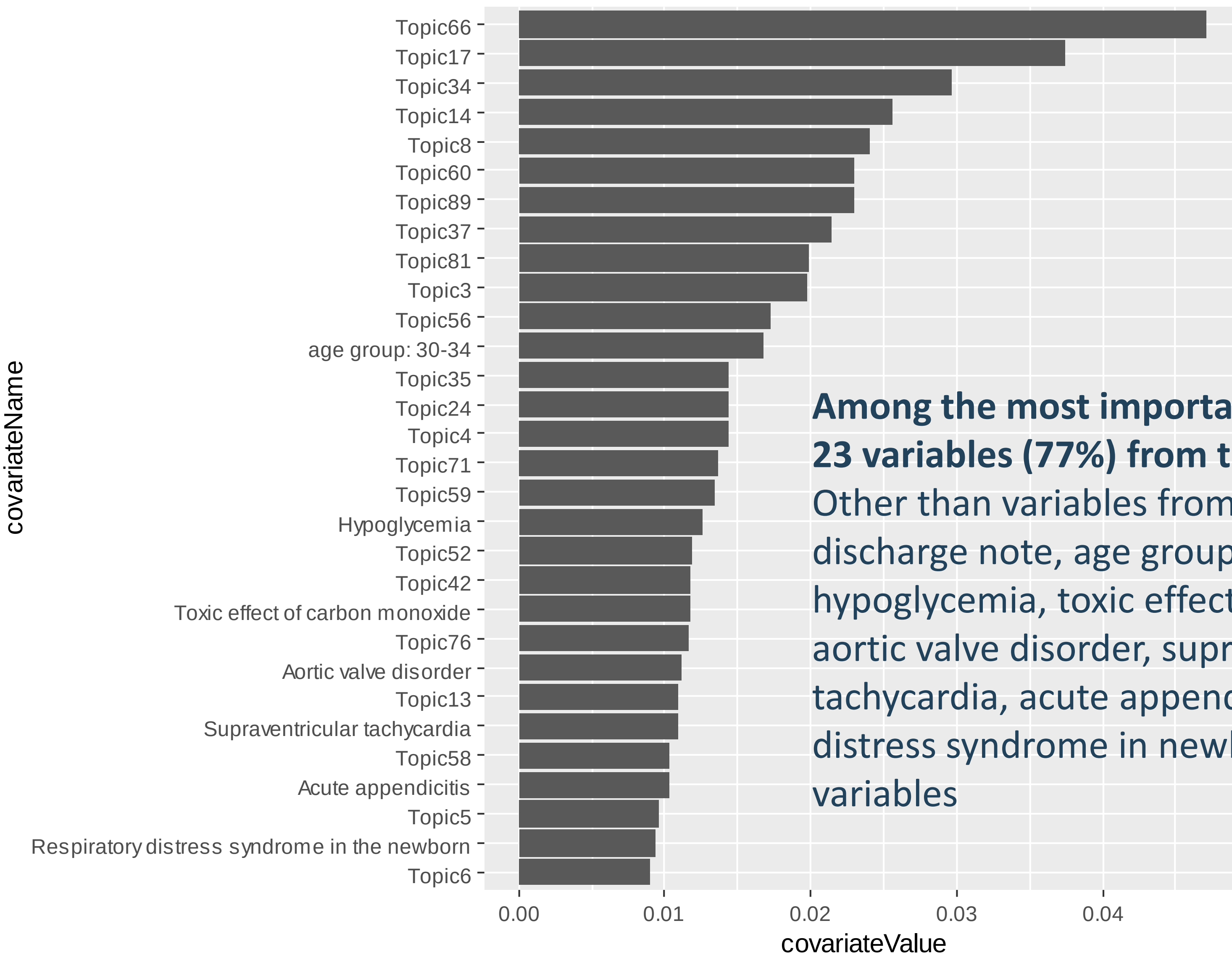
Experiments

- Building prediction model for readmission through emergency room to validate feasibility and usefulness of proposed NLP process
- Target cohort at risk: Subjects who admitted to the hospital and stayed 7 days or more from 1st January 2005 to 1st December 2017.
- Outcome cohort: Subjects who readmitted through emergency room within 30 days after discharge
- Covariates from conventional CDM: Gender, age group, race, ethnicity, index year, index month and condition within 30 days

Experiments-result

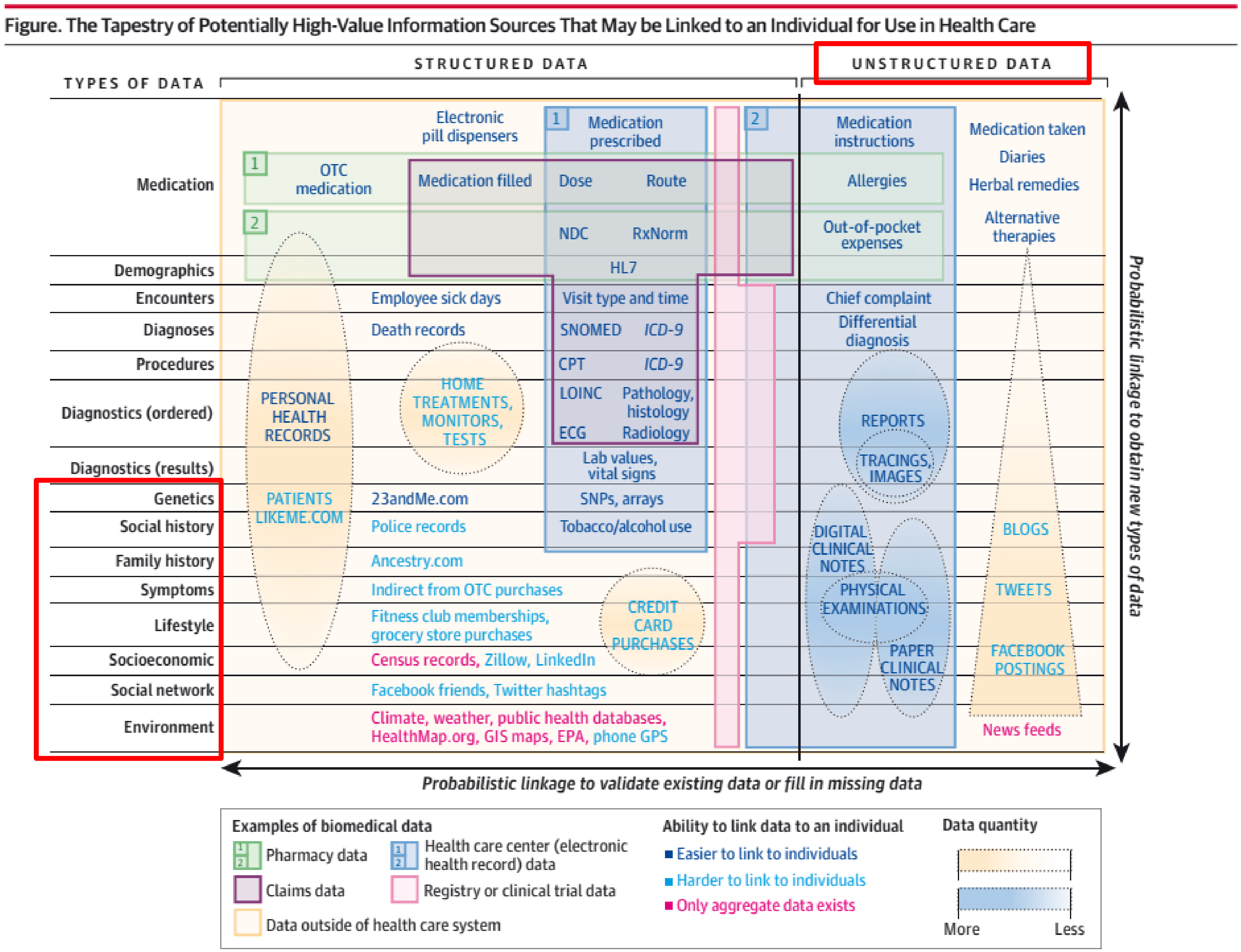


Top 30 most important values



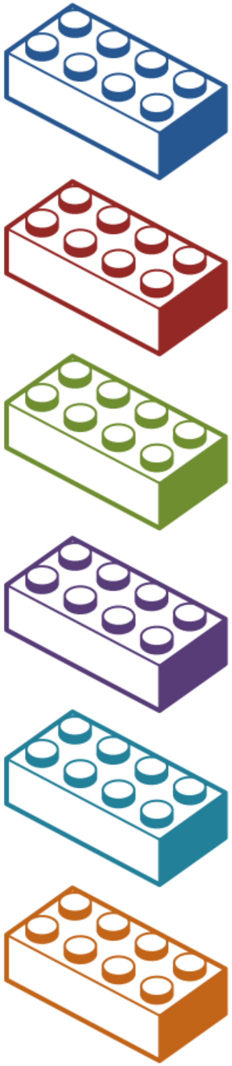
Among the most important top 30 variables, 23 variables (77%) from topics were included. Other than variables from the free-text of discharge note, age group 30 to 34, hypoglycemia, toxic effect of carbon monoxide, aortic valve disorder, supraventricular tachycardia, acute appendicitis and respiratory distress syndrome in newborn were important variables

Connecting 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 in CDM



Conditions

Drugs

Procedures

Measurements

Observations

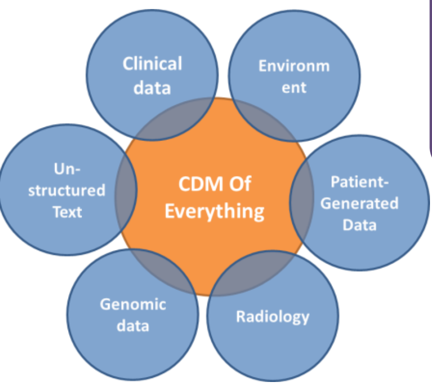
Visits

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 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

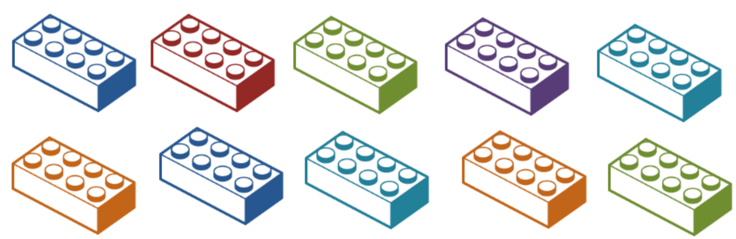
Automatically extract large sets of features for user-specified cohorts using data in the CDM.

Patient Level Prediction

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

Cohort Method

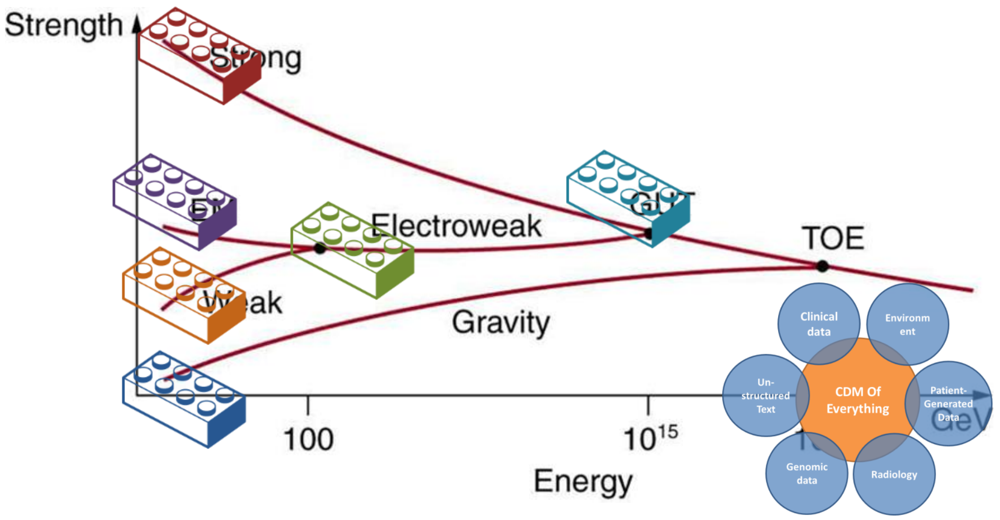
New-user cohort studies using large-scale regression for propensity and outcome models



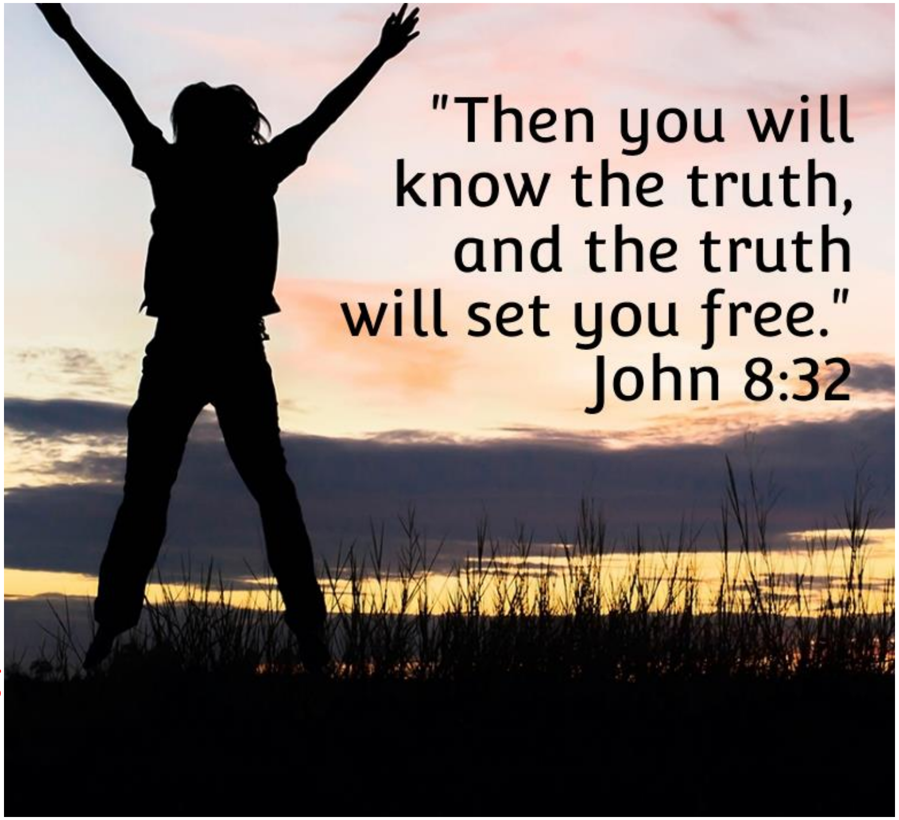
Symmetry in medical data

- 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




Thank
You
for your time