Comparison of Actionable Mutations in Cancer Patients using Genomic Common Data Model (G-CDM)

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

Background

- The importance of precision medicine to improve medical care is rapidly increasing, and there are a large number of ongoing large-scale biobanking and genetic testing initiatives.
- Clinical data has been standardized by OMOP-CDM, but it has low coverage of genomic data and cannot reflect the latest trend because OMOP-CDM focuses on Clinical data.
- We had proposed a beginning version of the genomic extension model in the OHDSI Symposium in May, 2018.

Purpose

- This study aims to propose a upgrade version of a genomic common data model (G-CDM) to take full utilize of the existing OMOP-CDM tables and adapt a standard vocabulary system.
- Then we aim to confirm a feasibility of G-CDM by presenting analysis results and tools based on G-CDM.

Methods

Data Model Development

- G-CDM were developed by extending OMOP-CDM for linking clinical data into genomic data. The TCGA and COSMIC databases have been reviewed to define how the variation is described. ISO20428 document, which is a standard format for reporting sequencing results was reviewed to design columns for variant annotation.
- For standard vocabulary, unique symbols and names for human genes approved by HGNC (HUGO Gene Nomenclature Committee) was used as a concept id. HGVS (Human Genome Variation Society) Nomenclature was used to standardize the description of sequence variation occur in each gene.

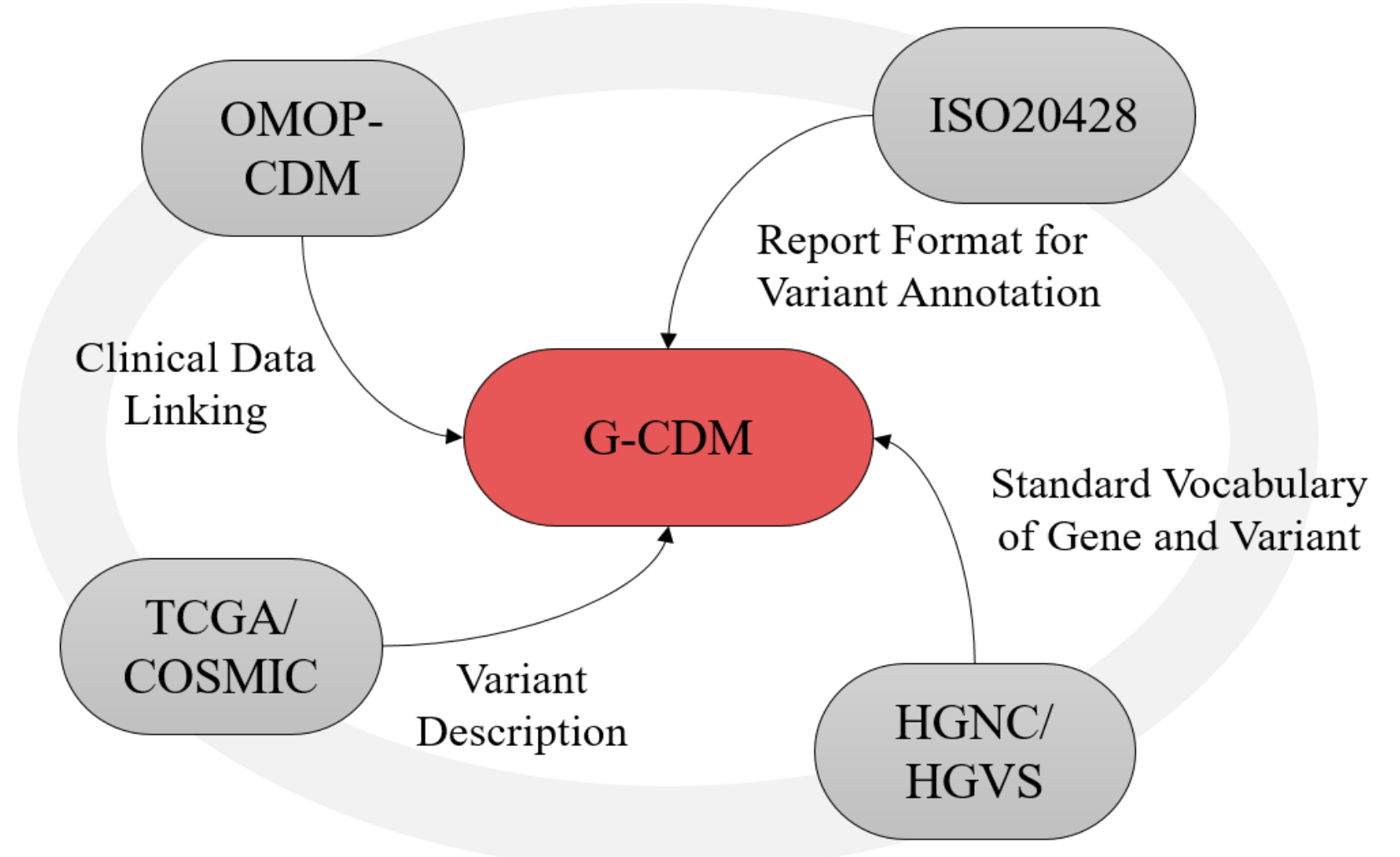


Figure 1. Description about how G-CDM was impacted by other models

Data Source for Model Validation

- In Ajou University Hospital database (AUSOM), 92 lung adenocarcinoma (LUAD) and 22 lung squamous cell carcinoma (LUSC) patients had NGS test from June 2017 to August 2018. TCGA database (TCGA) have 603 LUAD patients and 457 LUSC patients.
- By extracting, transforming, and loading NGS data from both organizations into the G-CDM schema, we were able to reuse the analysis code for both data.

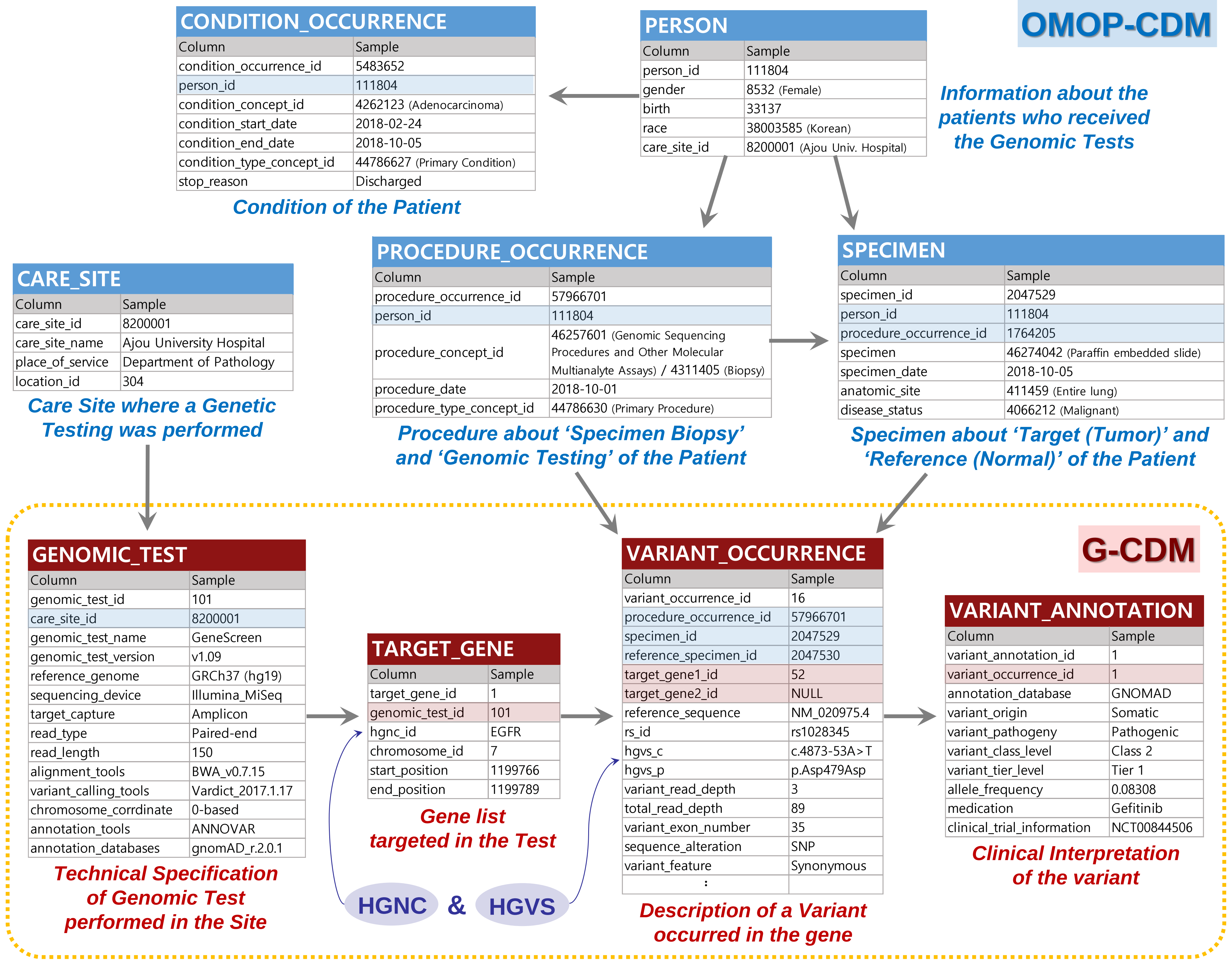
Comparison of Mutations Frequency

- To validate a usability of G-CDM in real medical practice, comparison of mutation occurred in lung cancer patients between different institutions was performed.

Results

Entity-relationship diagram (ERD) of G-CDM as a Genomic Extension of OMOP-CDM

- To store genomic sequencing data and process four tables named as 'Genomic Test', 'Target gene', 'Variant occurrence', and 'Variant annotation' were defined (Figure 2).
- 'Person', 'Condition occurrence', 'Care site', 'Procedure occurrence', and 'Specimen' is already in OMOP-CDM and deal with clinical data that is directly linked to 'Genomic test' and 'Variant occurrence' table.
- HGNC and HGVS vocabulary system were adapted for standardized gene and variant nomenclature in 'Target gene' and 'Variant occurrence' table.



OMOP-CDM

CONDITION_OCCURRENCE

Column	Sample
condition_occurrence_id	5483652
person_id	111804
condition_concept_id	4262123 (Adenocarcinoma)
condition_start_date	2018-02-24
condition_end_date	2018-10-05
condition_type_concept_id	44786627 (Primary Condition)
stop_reason	Discharged

Condition of the Patient

PERSON

Column	Sample
person_id	111804
gender	8532 (Female)
birth	33137
race	38003585 (Korean)
care_site_id	8200001 (Ajou Univ. Hospital)

Information about the patients who received the Genomic Tests

CARE_SITE

Column	Sample
care_site_id	8200001
care_site_name	Ajou University Hospital
place_of_service	Department of Pathology
location_id	304

Care Site where a Genetic Testing was performed

PROCEDURE_OCCURRENCE

Column	Sample
procedure_occurrence_id	57966701
person_id	111804
procedure_concept_id	46257601 (Genomic Sequencing Procedures and Other Molecular Multianalyte Assays) / 4311405 (Biopsy)
procedure_date	2018-10-01
procedure_type_concept_id	44786630 (Primary Procedure)

Procedure about 'Specimen Biopsy' and 'Genomic Testing' of the Patient

SPECIMEN

Column	Sample
specimen_id	2047529
person_id	111804
procedure_occurrence_id	1764205
specimen	46274042 (Paraffin embedded slide)
specimen_date	2018-10-05
anatomic_site	411459 (Entire lung)
disease_status	4066212 (Malignant)

Specimen about 'Target (Tumor)' and 'Reference (Normal)' of the Patient

G-CDM

GENOMIC_TEST

Column	Sample
genomic_test_id	101
care_site_id	8200001
genomic_test_name	GeneScreen
genomic_test_version	v1.09
reference_genome	GRCh37 (hg19)
sequencing_device	Illumina_MiSeq
target_capture	Amplicon
read_type	Paired-end
read_length	150
alignment_tools	BWA_v0.7.15
variant_calling_tools	Vardict_2017.1.17
chromosome_corrdinate	0-based
annotation_tools	ANNOVAR
annotation_databases	gnomAD_r.2.0.1

Technical Specification of Genomic Test performed in the Site

TARGET_GENE

Column	Sample
target_gene_id	1
genomic_test_id	101
hgnc_id	EGFR
chromosome_id	7
start_position	1199766
end_position	1199789

Gene list targeted in the Test

HGNC & HGVS

VARIANT_OCCURRENCE

Column	Sample
variant_occurrence_id	16
procedure_occurrence_id	57966701
specimen_id	2047529
reference_specimen_id	2047530
target_gene1_id	52
target_gene2_id	NULL
reference_sequence	NM_020975.4
rs_id	rs1028345
hgvs_c	c.4873-53A>T
hgvs_p	p.Asp479Asp
variant_read_depth	3
total_read_depth	89
variant_exon_number	35
sequence_alteration	SNP
variant_feature	Synonymous
:	

Description of a Variant occurred in the gene

VARIANT_ANNOTATION

Column	Sample
variant_annotation_id	1
variant_occurrence_id	1
annotation_database	GNOMAD
variant_origin	Somatic
variant_pathogeny	Pathogenic
variant_class_level	Class 2
variant_tier_level	Tier 1
allele_frequency	0.08308
medication	Gefitinib
clinical_trial_information	NCT00844506

Clinical Interpretation of the variant

Figure 2. Schematic diagram of the contents and relationship between the tables that make up the G-CDM. Tables with genomic data (red) and clinical data (blue) is linked with each other by foreign keys (marked in each tables). Not all of columns consisting each table were not shown in order to give a concise description.

Comparison of Mutation using G-CDM

- The different incidence of EGFR mutation by ethnicity (Asian and non-Asian) in patient with lung adenocarcinoma was reported and thus we investigated the mutation frequency between AUSOM and TCGA databases.
- EGFR, BRAF, AKT1, KRAS, PIK3CA and MET genes known for driver mutation in LUAD patients appeared with different frequencies ($p<0.05$) between AUSOM and TCGA databases (Figure 3).
- Activating and resistance mutations of epidermal growth factor receptor (EGFR) in LUAD patient role in clinical response to EGFR tyrosine kinase inhibitor (TKI) such as Gefitinib, Erlotinib, Afatinib, and Osimertinib.

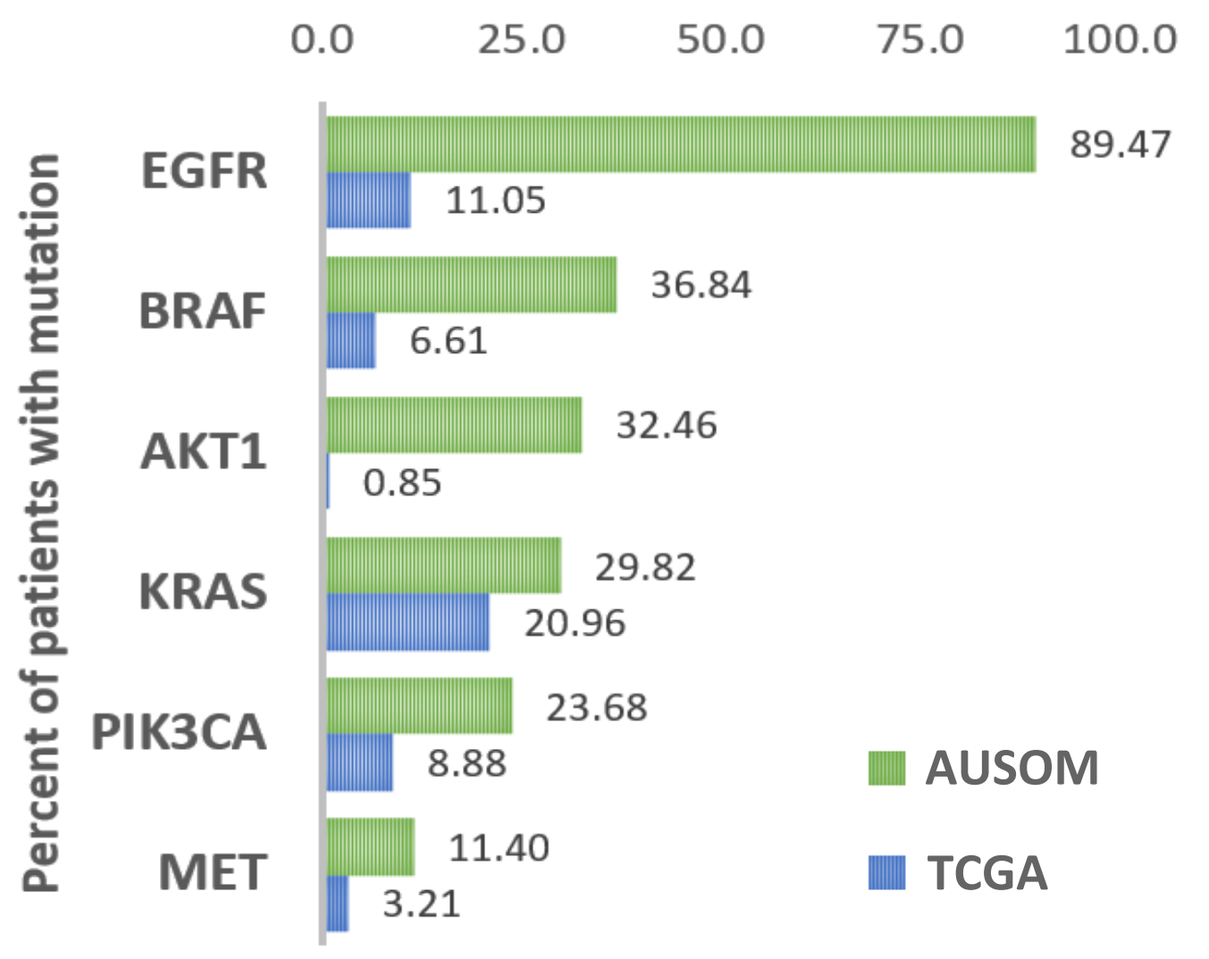


Figure 3. Proportion of patients with mutation in driver genes between AUSOM and TCGA databases.

- Of all EGFR positive LUAD patients, one with mutations that induce drug sensitivity accounted for 88% (AUSOM) and 91% (TCGA). Besides, patients with EGFR mutations that induce drug resistance or acquired drug resistance occupied 12% (AUSOM) and 9% (TCGA) which is similar proportion ($p>0.05$) in both institutes (Figure 4).

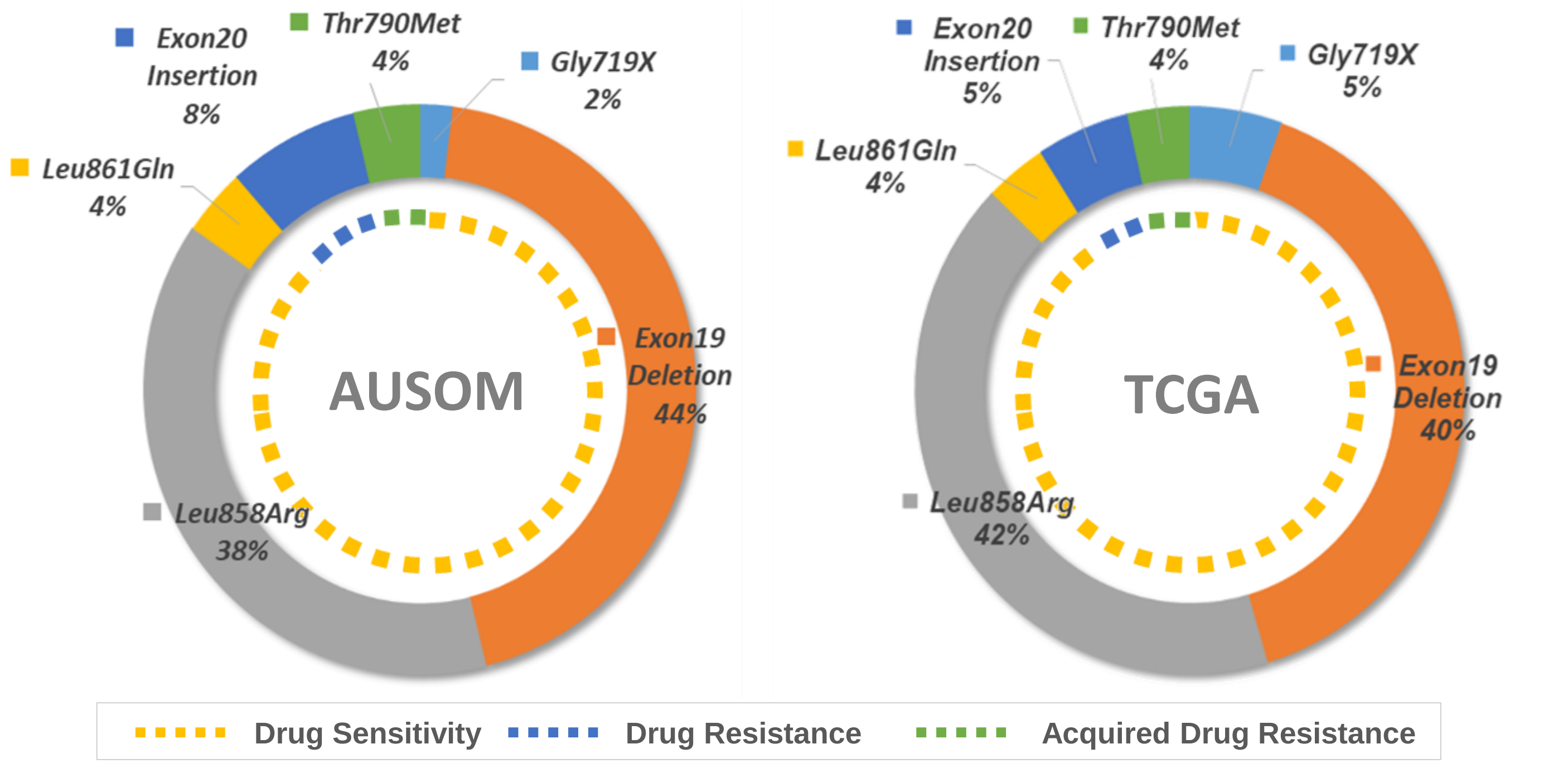


Figure 4. Patient proportion with drug-responsive (actionable) EGFR mutation.

Data Description Tool based on G-CDM

- We made a data description tool 'GeneProfiler' based on genomic data of patients based on G-CDM structure using R shiny package.
- After cohort definition using ATLAS of OHDSI Ecosystem, 'GeneProfiler' provide several plots on diverse point of view.
- Users can get understanding about their genomic data at the angle of 'overall mutation profile', 'proportion of variant types', 'fraction of pathogenic and drug response genes' and 'detailed variant information that have pathogenicity and drug responsiveness' (Figure 5).

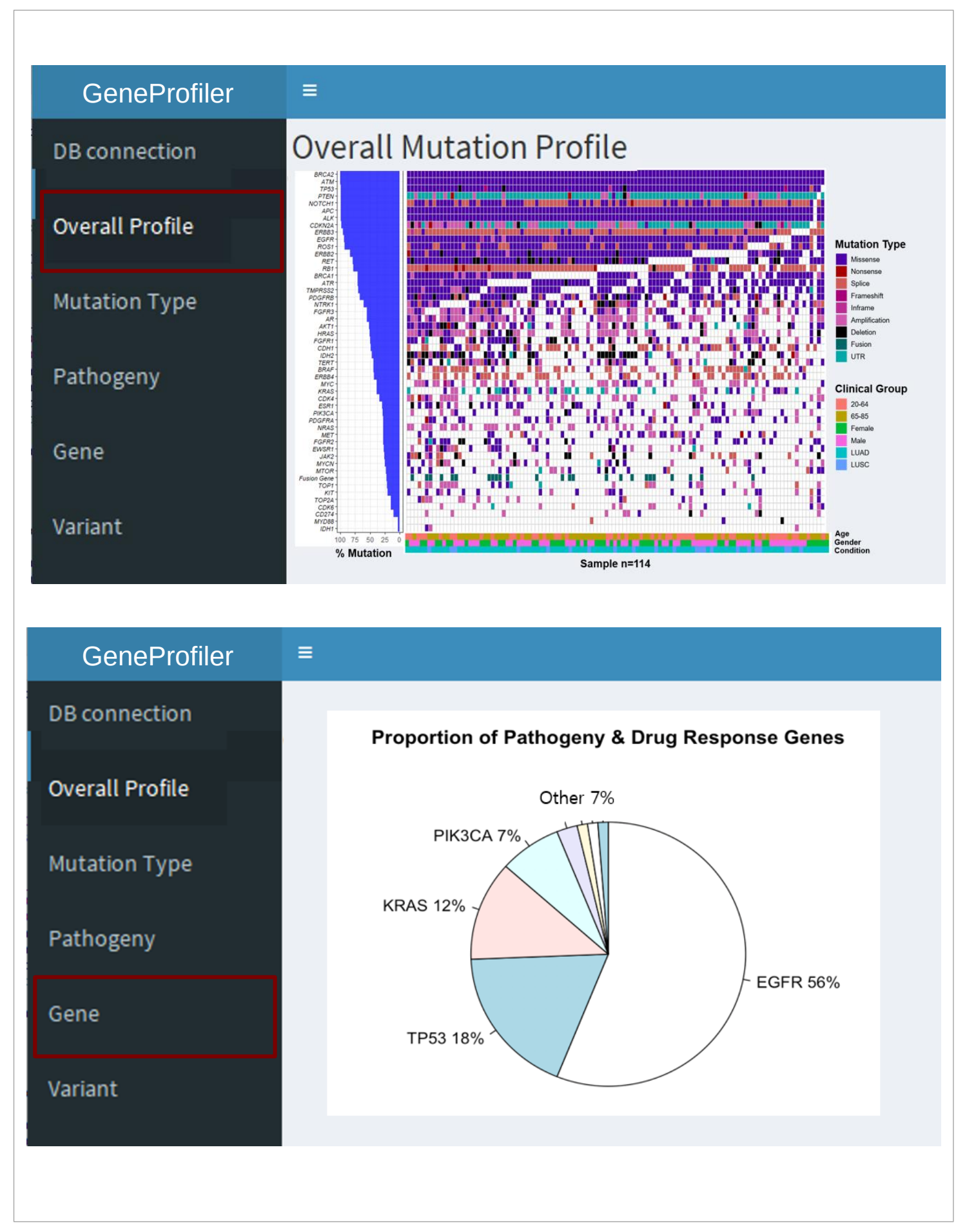


Figure 5. Data description tool 'GeneProfiler' based on G-CDM

Conclusions

- We developed and propose a genomic extension model, G-CDM, to store the data generated from NGS technology and integrate the data into OMOP-CDM.
- G-CDM focused on standardization of recording mutation clearly, not standardization of sequencing pipelines or report formats.
- We compared mutations in different institutes using G-CDM and confirmed that every single institute has to grasp the genomic background of their patients in order to provide a tailored care. Standardized genomic data will enable the sophistication of precision medicine.

Contact: Seo Jeong Shin lucid90sj@gmail.com

This work was supported by the Bio Industrial Strategic Technology Development Program (20001234) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea)