OMOP Common Data Model Extension for Genomic Data

Seo Jeong Shin, MS[1,2], Seng Chan You, MD, MS[1], Jin Roh, MD[3], JungHyun Byun, MS[1], Seungbin Oh[1], Clair Blacketer, MPH[4], Rae Woong Park, MD, PhD[1, 5]

[1]Dept. of Biomedical Sciences, Graduate School of Ajou University, Suwon, Korea; [2]BK21 Plus Program, Department of Biomedical Sciences, Graduate School of Ajou University; [3]Dept. of Pathology, Ajou University Hospital, Suwon, South Korea; [4]Janssen Research & Development, LLC, Raritan, New Jersey; [5]Dept. of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea

Background

- Since March 2017 in Korea, the cancer panel test for patient with cancer or rare genetic disease started to be applied to medical insurance.
- Until recent practices in medical research field, clinical and genomic data have been harvested and stored by individual disease or institution. In a Distributed Research Network (DRN), each data holder who uses the OMOP Common Data Model can conduct a statistically powerful study without violation of private information by using the same code and integrating results of the analysis.
- We designed and present a highly scalable standard for clinical - genome convergence database.

Methods

- The International Organization for Standardization (ISO) document ISO/TS20428 defines the required/optional data fields for structured reporting the next-generation sequencing (NGS) result to clinicians or patients. We compared ISO/TS20428 document with the OMOP-CDM v5.0 to identify the relationship between their fields, and design a new database schema.
- New tables for genomic data were created in order to minimize redundancy and include the maximum information required for analysis.
- After converting both clinical and genomic data into OMOP-CDM extension model, descriptive statistics was conducted to compare variants between target and comparator cohorts. As a pilot study, we set advanced stage of lung cancer (stage IIA or more) as a target cohort and non-advanced stage of lung cancer as a comparator cohort.
- The PatientLevelPrediction package was used to predict the stage of lung cancer by adding genomic features to the clinical features through FeatureExtraction package.

Results

- In the existing OMOP-CDM v5.0, 7 clinical data, 3 health system data, and 3 derived elements tables were designated as clinical information tables to be linked to the genome information table. The table for the genomic data was expanded by creating three tables; Sequencing tables, Variant_occurrence table and Variant_annotation table (Figure 1).
- Specification and sample data of each table is uploaded at http://bit.ly/2hA1ioi.

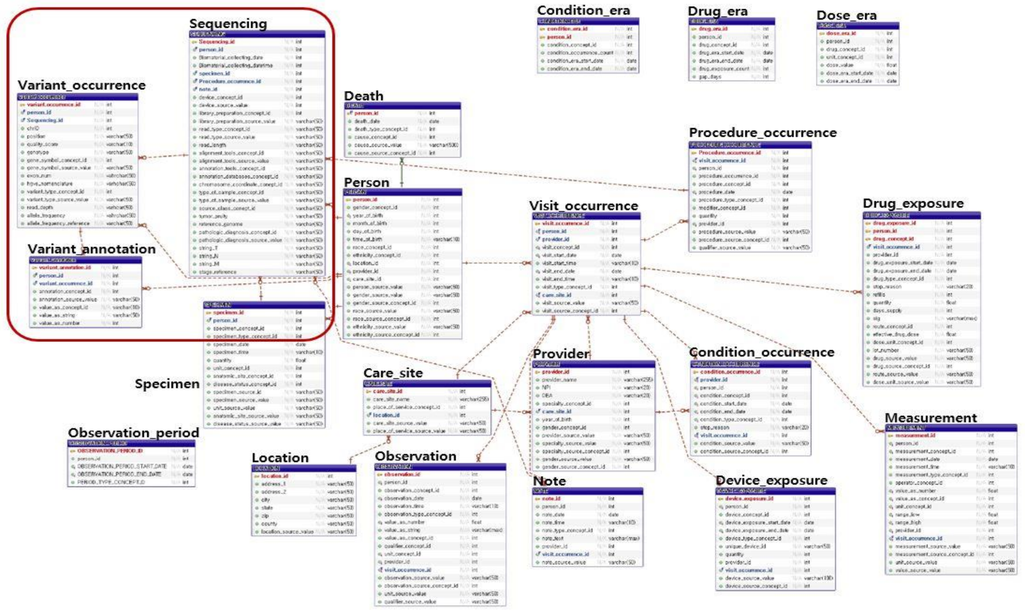


Figure 1. OMOP-CDM extension model for genomic data

- Several results of descriptive statistics of three tables in the extended genomic model are shown in Figure 2~3.

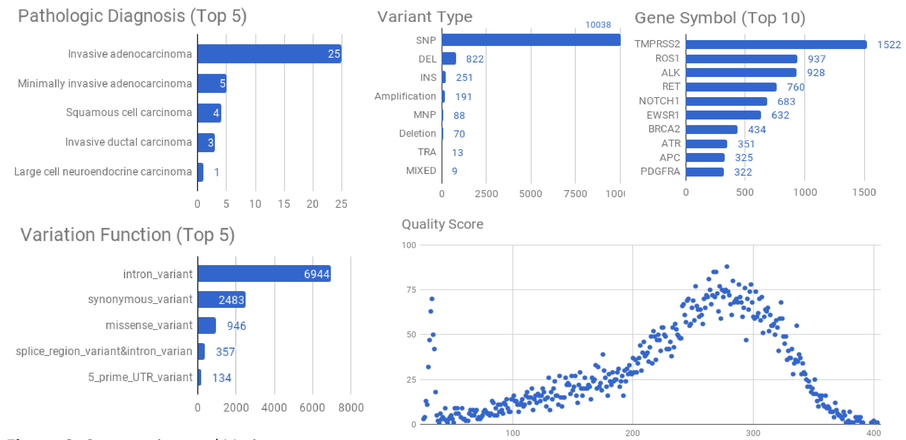


Figure 2. Sequencing and Variant_annotation Table Descriptive Statistics; Pathologic diagno-sis and Variant Function

Figure 3. Variant_occurrence Table Descriptive Statistics; Variant Type, Gene Symbol (Top 10), and Quality Score

- The most different mutation occurrence rate (p=0.002) between lung cancer stage was PDGFRB, which is already well-known as related to growth of cancer cell (Fig. 4).

Patient Group by Cancer Stage

Group	Stage
Comparator Cohort	Stage IA, Stage IB
Target Cohort	Stage IIA, Stage IIB, Stage IIIA

Annotation Information of Top 1 Mutation

VAR_ID	VAR_TYPE	CHRID_1	START_POSITION	GENE1_SYMBOL	HGVS_NOMENCLATURE
1016	SNP	chr5	149515397	PDGFRB	c.85A>T

Variant ID Mapping Table (part)

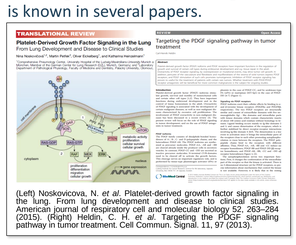
VAR_ID	VARIANT_TYPE	CHRID_1	CHRID_2	START_POSITION	END_POSITION	GENE1_SYMBOL	GENE2_SYMBOL	HGVS_NOMENCLATURE
1	Translocation	chr1	chr20	171308889	39750709	FMO4	TOP1	#N/A
2	Insertion	chr6	#N/A	117647581	0	ROS1	#N/A	c.5367-5dupT
3	SNP	chr10	#N/A	43572753	0	RET	#N/A	c.47T>C
4	Amplification	chr1	#N/A	115251095	115258845	NRAS	#N/A	#N/A
5	Deletion	chr9	#N/A	5073681	0	JAK2	#N/A	c.1777-7delT
6	Translocation	chr1	chr20	171308889	39750709	TOP1	FMO4	#N/A
7	Insertion	chr17	#N/A	7578517	0	TP53	#N/A	c.412dupG
8	MNP	chr1	#N/A	156834550	0	NTRK1	#N/A	c.318_319delGGinsTT
9	SNP	chr1	#N/A	156849989	0	NTRK1	#N/A	c.2205+40G>A
10	SNP	chr12	#N/A	58142253	58145511	CDK4	#N/A	#N/A

Target cohort has significantly more Variant ID 1016 mutation occurrence rate than Comparator.

Top 10 Mutations

#	Variant ID	p-value
1	ID1016	0.002
2	ID264	0.006
3	ID1394	0.007
4	ID965	0.009
5	ID545	0.023
6	ID489	0.028
7	ID1012	0.029
8	ID109	0.030
9	ID769	0.030
10	ID596	0.043

PDGF Signaling Pathway in Lung Cancer is known in several papers.



(Left) Noskovicova, N. *et al.* Platelet-derived growth factor signaling in the lung. From lung development and disease to clinical studies. American journal of respiratory cell and molecular biology 52, 263–284 (2015). (Right) Heldin, C. H. *et al.* Targeting the PDGF signaling pathway in tumor treatment. Cell Commun. Signal. 11, 97 (2013).

Figure 4. Calculating difference of specific mutant occurrence rate between two cohorts of lung cancer patients.

- With application of PatientLevelPrediction package, predictive model was constructed corresponding to existing OHDSI research pipeline.
- For the pilot study, data of lung cancer patients who have had NGS test was used to predict whether their cancer stage is advanced or not.
- When we use both clinical and genomic data, the prediction accuracy was higher (AUROC 0.87) compared to when we use only clinical data (AUROC 0.82) (Fig. 5).

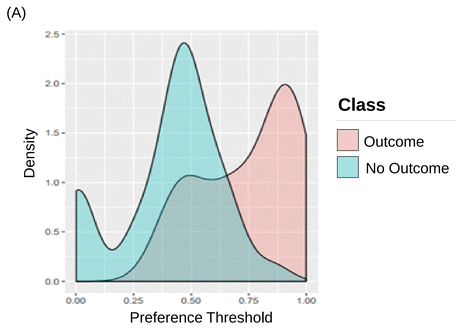


Figure 5. Predicting advanced stage lung cancer patient using both clinical and genomic data. (A) Density plot over preference threshold. (B) Box plot shows that the two class can be easily distinguished among the lung cancer patients.

Conclusions

- We have constructed a database that can systematically manage the genetic variation data of patients who have been examined by NGS. It consists of three tables newly designed to contain genomic information.
- Through this study, we demonstrated that this database can be used not only to manage data quality but also to analyze genomic data integrated with clinical data.
- It is expected that this genomic extension model can be suggested as a prototype database standard which can be easily extended to domestic and overseas medical institutions.

- Acknowledgement : This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HI16C0992). Also, this research was supported by a grant of the Korea company SKT.
- Conflict of interest : None

Contact: lucid900921@naver.com