Conversion of Diagnosis and Chemotherapy Data in Electronic Health Records to Episode-based Oncology Extension of OMOP-CDM

Hokyun Jeon[1], Seng Chan You, MD, MS[2], Jimyung Park [1], Rae Woong Park, MD, Ph.D[1,2,3]

[1] Department of Biomedical Science, Ajou University Graduate School of Medicine, Suwon, Gyeonggi-do, Republic of Korea;
[2] Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Gyeonggi-do, Republic of Korea;
[3] FEEDER-NET(Federated E-health Big Data for Evidence Renovation Network)

Background

- Recently, oncology working group suggests the reconciliation of cancer data from heterogeneous sources such as cancer registries, Electronic Health Records (EHRs) or clinical trials into newly generated ***'Episode'*** table **(Fig 1)**.
- In **oncology extension Common Data Model (CDM)**, diagnosis includes histology and topography features, and treatment regimen records are curated in *'Episode'* table.
- However, in EHR, it is challenge to generate ICD-O-3 diagnosis and treatment episode compared with other sources.

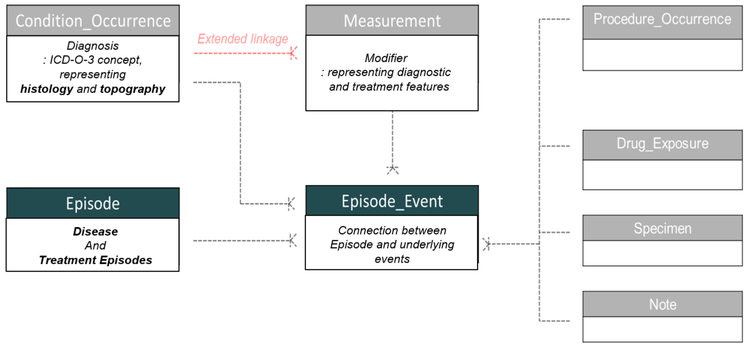


Figure 1. The table relationship schema of oncology CDM from oncology working group proposal[1]. The **modifiers** represent diagnostic and treatment feature in Measurement table. Episode and Episode event table are newly established. **Disease episode** includes diagnostic information, such as first occurrence, Recurrence, aggravation, and alleviation. **Treatment episode** represents the type of procedure / therapy / treatment or the number of repeated cycle.

Purpose

- We purposed to demonstrate conversion of **EHR data** into **oncology extension CDM**. As a pilot example, we targeted **colorectal cancer diagnosis** and their **chemotherapy treatment regimen**.
- We aimed to extract histology and topography features from histology reports and generate **ICD-O-3 code** for diagnosis.
- We tried to generate **a parameter-based algorithm** for extracting each treatment regimen cycle records from drug exposure records. Then, we curated treatment episode into 'Episode' table.

Generating ICD-O-3 diagnosis from JSON form pathology note

- ICD-O-3 codes are consisted of Histology ICD-O code and topography ICD-O code.
- The pathology report of subsampled patients are converted into **Java Script Object Notation(JSON)** form with *'SOCRATex'* package.
- *'SOCRATex'* is NLP package for clinical text.
- The whole codes are available at: https://github.com/ABMI/SOCRATex
- From JSON form notes, histology and topography features are extracted and assigned to each matching **ICD-O- code**. ICD-O-3 codes are generated in combination of respective ICD-O code **(Fig 2)**.

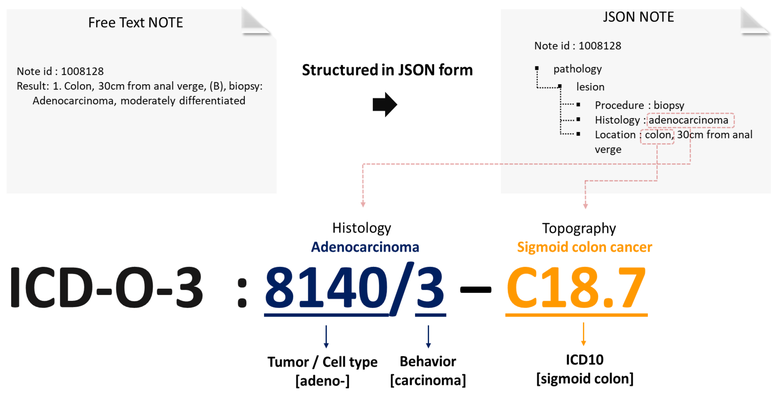


Figure 2. Schematic flow chart for extracting histology and location information of cancer from JSON form notes and generating ICD-O-3 code for diagnosis.

Algorithm for extracting treatment regimen

- In EHRs, treatment regimen is recorded in the **discharge summary** with **narrative** description.
- However, it is challenging to extract treatment regimen from discharge summary, because they are recorded in **a variety terms** and not including **outpatient records**.
- So, we generated algorithm that extract treatment regimen cycle from drug exposure records with **researcher-based parameter settings**.
- The researcher, decided to extract certain treatment regimen from their cohort, can extract each **cycle start date** and **end date** for target regimen and get results about the **distribution of patients** in iterative cycle (**Fig 3,4**).
- In oncology CDM, **HemOnc** is consented as a guideline for treatment regimen, so researchers can refer to HemOnc for setting the parameter. It is available at: https://hemonc.org/wiki/Main_Page
- The treatment cycle extraction algorithm code is available at : https://github.com/ABMI/treatmentCycleExtraction

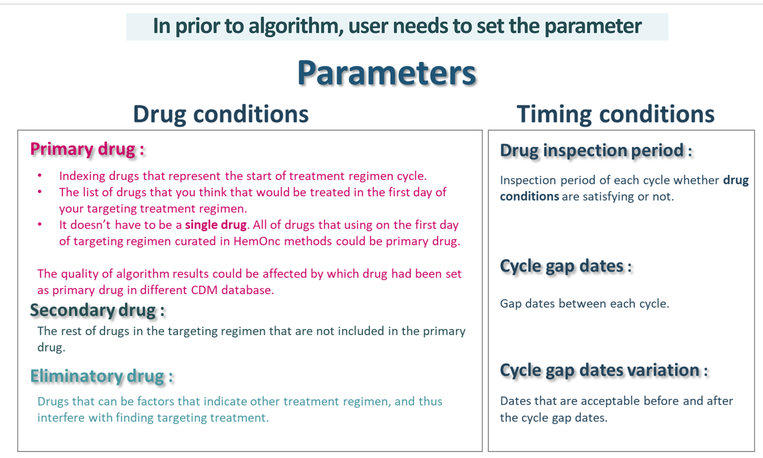


Figure 3. Parameter settings for cycle extraction algorithm. The algorithm extract regimen cycle from drug records, and *'Drug conditions'* are parameters for which drug should be in or not. *'Timing conditions'* define the dates apart between each cycle in same treatment line.

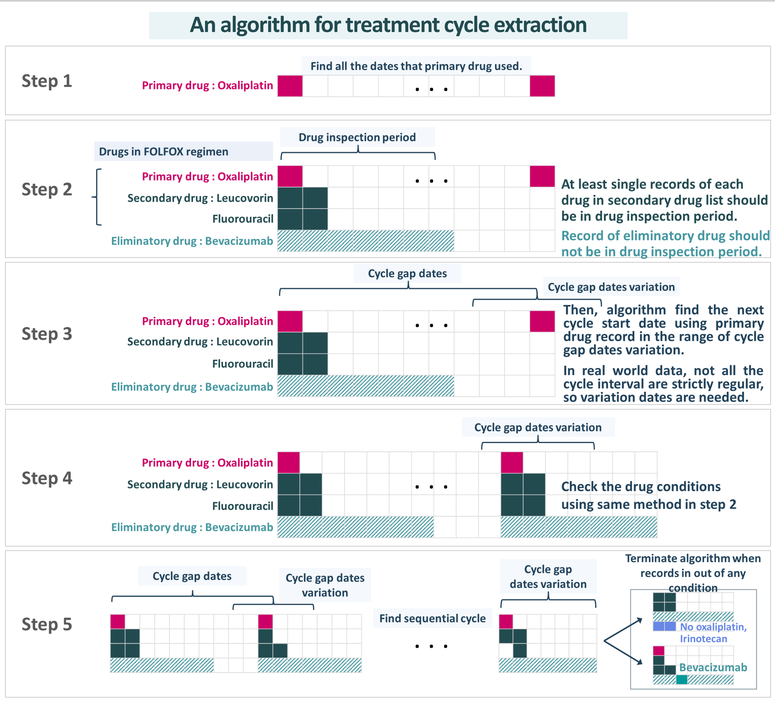


Figure 4. Overall process for extracting treatment cycle from drug records in parameter setting.

Results

- In total 2243 of colorectal cancer patients, we subsampled 200 patients. Among them, **112 patients with carcinoma** were assigned **ICD-O-3 code** according to extracted histology and location features.
- The most prevalent ICD-O-3 code is '***adenocarcinoma of sigmoid colon***' (CONCEPT ID: "44502464"), and it accounts for 33% of ICD-O-3 code diagnoses, and the other diagnosis distribution information are below **(Table 1).**
- ICD-O-3 diagnosis codes are curated in '*condition occurrence*' table and they are linked to '*Episode event*' table which is including subitems of '*Episode*' table **(Fig. 5)**.

ICD-O-3 diagnosis	Concept code	concept ID	N	(%)
Adenocarcinoma of colon	8140/3-C18.9	44502464	12	10.7
Adenocarcinoma of hepatic flexure of colon	8140/3-C18.3	44501932	5	4.5
Adenocarcinoma in tubulovillous adenoma of ascending colon	8263/3-C18.2	44502946	1	0.9
Adenocarcinoma of transverse colon	8140/3-C18.4	44500927	7	6.3
Tubular adenocarcinoma of rectosigmoid junction	8211/3-C19.9	36526362	1	0.9
Adenocarcinoma of ascending colon	8140/3-C18.2	44502439	9	8.0
Adenocarcinoma of cecum	8140/3-C18.0	44504337	2	1.8
Tubular adenocarcinoma of colon	8211/3-C18.9	36530925	1	0.9
Adenocarcinoma of overlapping lesion of colon	8140/3-C18.8	36561605	4	3.6
Adenocarcinoma of rectum	8140/3-C20.9	44500130	16	14.3
Adenocarcinoma of rectosigmoid junction	8140/3-C19.9	44501075	12	10.7
Carcinoma of transverse colon	8010/3-C18.4	44504361	1	0.9
Adenocarcinoma of sigmoid colon	8140/3-C18.7	44504380	37	33.0
Adenocarcinoma of descending colon	8140/3-C18.6	44500497	4	3.6

Table 1. Distribution of generated ICD-O-3 code in 112 sample colorectal patients.

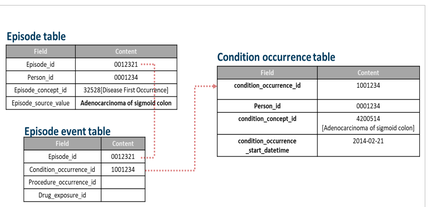
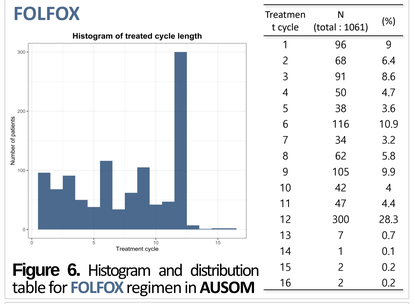


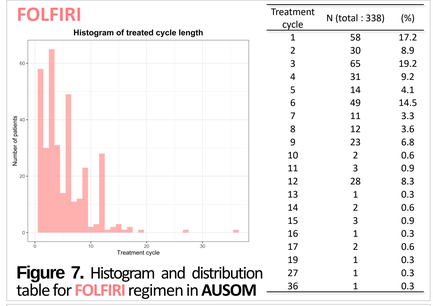
Figure 5. Schematic figure for relationship of disease related table.

- In 6,838 colorectal patients, *FOLFOX* regimen of 1,061 patients and *FOLFIRI* regimen of 338 patients are extracted from **Ajou University School Of Medicine (AUSOM)** database. *FOLFOX* regimen are mostly repeated **12** cycles and *FOLFIRI* regimen are mostly repeated **3** cycles **(Fig 6,7).**
- In 15,958 **National Health Insurance-National Sample Cohorts(NHIS-NSC)** colorectal patients, *FOLFOX* regimen are extracted from 884 patients. It is mostly repeated **12** cycles as well **(Fig 8).**
- Histogram and distribution table are generated as a result of algorithm, and each cycle records are curated in 'Episode' event table with linkage of each drug exposure records **(Fig 9).**



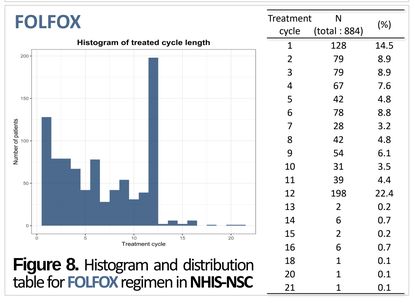
Treatment cycle	N (total : 1061)	(%)
1	96	9
2	68	6.4
3	91	8.6
4	50	4.7
5	38	3.6
6	116	10.9
7	34	3.2
8	62	5.8
9	105	9.9
10	42	4
11	47	4.4
12	300	28.3
13	7	0.7
14	1	0.1
15	2	0.2
16	2	0.2

Figure 6. Histogram and distribution table for FOLFOX regimen in **AUSOM**



Treatment cycle	N (total : 338)	(%)
1	58	17.2
2	30	8.9
3	65	19.2
4	31	9.2
5	14	4.1
6	49	14.5
7	11	3.3
8	12	3.6
9	23	6.8
10	2	0.6
11	3	0.9
12	28	8.3
13	1	0.3
14	2	0.6
15	3	0.9
16	1	0.3
17	2	0.6
19	1	0.3
27	1	0.3
36	1	0.3

Figure 7. Histogram and distribution table for FOLFIRI regimen in **AUSOM**



Treatment cycle	N (total : 884)	(%)
1	128	14.5
2	79	8.9
3	79	8.9
4	67	7.6
5	42	4.8
6	78	8.8
7	28	3.2
8	42	4.8
9	54	6.1
10	31	3.5
11	39	4.4
12	198	22.4
13	2	0.2
14	6	0.7
15	2	0.2
16	6	0.7
18	1	0.1
20	1	0.1
21	1	0.1

Figure 8. Histogram and distribution table for FOLFOX regimen in **NHIS-NSC**

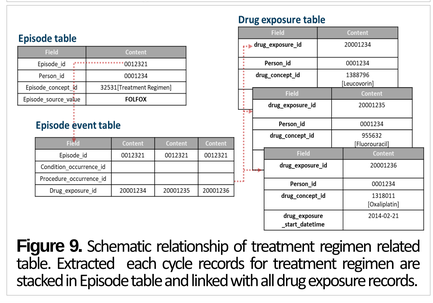


Figure 9. Schematic relationship of treatment regimen related table. Extracted each cycle records for treatment regimen are stacked in Episode table and linked with all drug exposure records.

- The extracted **number of repeated cycle** of *FOLFOX,FOLFIRI* and *XELODA* **in randomly subsampled 100 patients** are compared with **the manual review result of text in the discharge summary**.
- '*Accuracy*' is a proportion of **exactly matching cases** except for '*No information in the discharge note*', and '*RMSE*' is Root Mean Square Error (RMSE) value for cycle number difference between the algorithm result and the summary in the discharge note **(Fig 10).**

Treatment regimen	No information in the discharge note	consistent with discharge note	Accuracy	RMSE
FOLFOX	8	80	87%	0.64
FOLFIRI	21	70	88.60%	1.36
XELODA	65	24	69%	1.5

Figure 10. Algorithm validation with discharge note.

Conclusions

- To our knowledge, this is the first attempt to convert oncology-specific data from EHR to an episode-based oncology extension model of OMOP-CDM in OHDSI network.
- By leveraging information from structured pathology reports and the treatment cycle extraction algorithm, it is possible to populate oncology data in oncology extended CDM.

Acknowledgements : This work 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] and the Bio Industrial Strategic Technology Development Program (20003883) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea). **References :** 1. oncology working group proposal, https://github.com/OHDSI/CommonDataModel/issues/239 / **Conflict of interest :** none / **Contact** : veritas@ajou.ac.kr /