Implementation of Episode-based Oncology OMOP-CDM In Electronic Health Records

July 9th, 2019

Hokyun Jeon

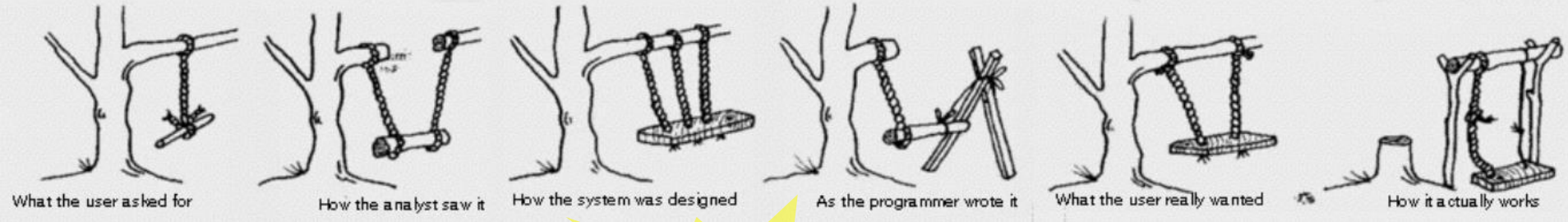
Objectives

- **Prototype of Episode-based Oncology CDM in EHRs**

Objectives

- **Prototype of Episode-based Oncology CDM in EHRs**

-The results do not always match what we planned.



Issues

- **We always got some new issues in prototype**

Objectives

- As a proof-of-concept, we tried to populate EHR-derived oncology data of colon cancer in Episode-based Oncology Extension Model
- We aimed to report the issues regarding to this conversion

Objectives

We tried to develop a toy model in the format of oncology CDM proposal

EPISODE

Field	Content
episode_id	4325345
person_id	John Smith
episode_concept_id	First Occurrence
episode_start_datetime	February 14, 1996
episode_end_datetime	November 18, 1996
episode_object_concept_id	Adenocarcinoma of sigmoid colon
episode_type_concept_id	Algorithm #123

EPISODE_EVENT

Field	Content		
episode_id	4325345	4325345	4325345
condition_occurrence_id	9900145	9900850	
procedure_occurrence_id			456774870
drug_exposure_id			
specimen_id			
note_id			

Cancer diagnosis

CONDITION_OCCURRENCE

Field	Content
condition_occurrence_id	9900145
person_id	John Smith
condition_concept_id	Adenocarcinoma of sigmoid colon
condition_occurrence_start_datetime	February 14, 1996
condition_occurrence_end_datetime	
condition_occurrence_type_concept_id	Cancer Registry

CONDITION_OCCURRENCE

Field	Content
condition_occurrence_id	9900850
person_id	John Smith
condition_concept_id	Adenocarcinoma of sigmoid colon
condition_occurrence_start_datetime	September 15, 1999
condition_occurrence_end_datetime	
condition_occurrence_type_concept_id	EMR

PROCEDURE_OCCURRENCE

Field	Content
procedure_occurrence_id	456774870
person_id	John Smith
procedure_concept_id	Intravenous chemotherapy
procedure_occurrence_start_datetime	November 1, 1996
procedure_occurrence_end_datetime	November 18, 1996
procedure_occurrence_type_concept_id	EMR

Objectives

We tried to develop a toy model in the format of oncology CDM proposal

4. Treatment episode and related event records

EPISODE

Field	Content
episode_id	9900850
person_id	John Smith
episode_concept_id	Treatment Regimen
episode_start_datetime	August 1, 1996
episode_end_datetime	November 18, 1996
episode_parent_id	
episode_number	
episode_object_concept_id	Chemotherapy Treatment
episode_type_concept_id	Cancer Registry
episode_source_value	Chemotherapy
episode_source_concept_id	C25 (NAACCR ID)

EPISODE

Field	Content
episode_id	9900851
person_id	John Smith
episode_concept_id	Treatment Cycle
episode_start_datetime	August 1, 1996
episode_end_datetime	August 28, 1996
episode_parent_id	9900850
episode_number	1
episode_object_concept_id	Chemotherapy Treatment
episode_type_concept_id	EMR
episode_source_value	PACLITAXEL + CARBOPLATIN
episode_source_concept_id	

EPISODE

Field	Content
episode_id	9900852
person_id	John Smith
episode_concept_id	Treatment Cycle
episode_start_datetime	October 15, 1996
episode_end_datetime	November 18, 1996
episode_parent_id	9900850
episode_number	2
episode_object_concept_id	Chemotherapy Treatment
episode_type_concept_id	EMR
episode_source_value	PACLITAXEL + CARBOPLATIN
episode_source_concept_id	

EPISODE_EVENT

Field	Content		
episode_id	9900851	9900851	9900851
condition_occurrence_id			
procedure_occurrence_id			
drug_exposure_id	9900145	9900146	9900147
device_exposure_id			
observation_id			
specimen_id			
note_id			

DRUG_EXPOSURE

Field	Content
drug_exposure_id	9900145
person_id	John Smith
drug_concept_id	Cyclophosphamide
drug_exposure_start_datetime	August 1, 1996
drug_exposure_end_datetime	August 1, 1996
drug_exposure_type_concept_id	EMR
drug_exposure_source_value	Cyclophosphamide 1000 MG Injection

DRUG_EXPOSURE

Field	Content
drug_exposure_id	9900146
person_id	John Smith
drug_concept_id	Doxorubicin hydrochloride
drug_exposure_start_datetime	August 4, 1996
drug_exposure_end_datetime	August 4, 1996
drug_exposure_type_concept_id	EMR
drug_exposure_source_value	Doxorubicin Hydrochloride 50 MG Injection

DRUG_EXPOSURE

Field	Content
drug_exposure_id	9900147
person_id	John Smith
drug_concept_id	Dexamethasone acetate
drug_exposure_start_datetime	August 7, 1996
drug_exposure_end_datetime	August 7, 1996
drug_exposure_type_concept_id	EMR
drug_exposure_source_value	Dexamethasone acetate 8 MG/ML Injectable

Treatment regimen

Contents

We focused on two challenges in implementation of oncology extension model

- **Challenge 1 :**

  Diagnosis code in EHR or claim database does not have detailed information related with cancer diagnosis

- **Challenge 2 :**

  Also, treatment regimen information is not structured in EHR or claim database

Challenge 1

- **Diagnosis code in EHR does not have detailed information related with oncology diagnosis**

  1. Such as topography, histology, and staging

  2. This information should be extracted from medical narrative text in EHR

  3. These text data are usually not machine-readable

Issue :

Histology information should be extracted from the narrative pathology report

Pathology reports

현미부수체 불안정성은 유전자의 기능소실에 의해 발생하는 것으로 알려져 있습니다. 유전성비용종증대장암(HNPCC)의 약 90%, 산발성 대장암의 10-20%에서 MSI가 관찰됩니다. 대장암 이외에도 위암, 난소암, 췌장암, 자궁내막암 등에서도 MSI가 관찰됩니다.

본 검사에서는 5가지 marker (BAT25, BAT26, NR21, NR24, MONO27)에 대한 MSI 분석을 시행하였습니다. 2개 이상의 marker에서 불안정성을 나타낼때 MSI-High로 보고하며, 1개의 marker에서 불안정성을 나타낼때 MSI-Low, 모든 marker에서 불안정성을 나타내지 않을때 MSS(microsatellite stable)로 보고합니다. 단, 3bp 미만의 shift는 결과분석에서 고려하지 않았습니다.

Malignant tumor of ascending colon

(ICD10 : C18.2)

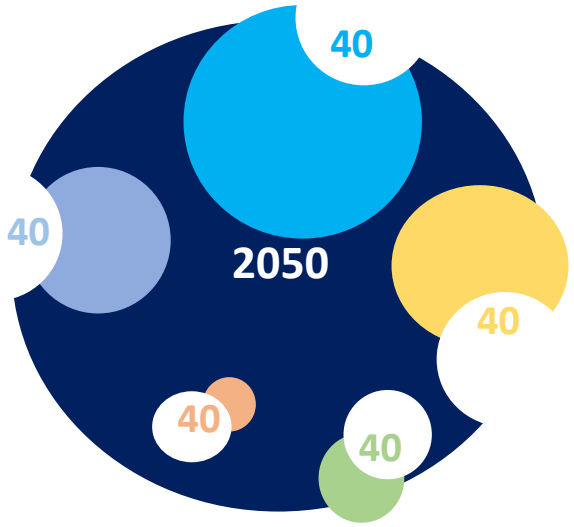
Adenocarcinoma of ascending colon cancer

(ICD-O-3 : 8140/3-C18.2)

- **Histology features of colon cancer should be extracted and curated from narrative text of pathology reports to be machine-readable**

Method :

Subsampling target patients



ICD-10 Code

- C18 (Colon)
- C19 (Rectosigmoid junction)
- C20 (rectum)

Patients who had pathology report in note table (2014~2017)

First, we focused on colorectal cancer patients

My colleague has already curated and reconstructed pathology report from the patients with colon cancer in JSON form (I'll present about it in an hour)

Excluded benign tumor

112

200

Random sampling in 5 topics

JSON form Pathology report review

Method :

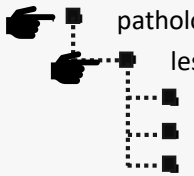
Pathology report of target patients had been structured in JSON form

pathology report

Note id : 1008128
Result: 1. Colon, 30cm from anal verge, (B), biopsy: Adenocarcinoma, moderately differentiated

JSON Note

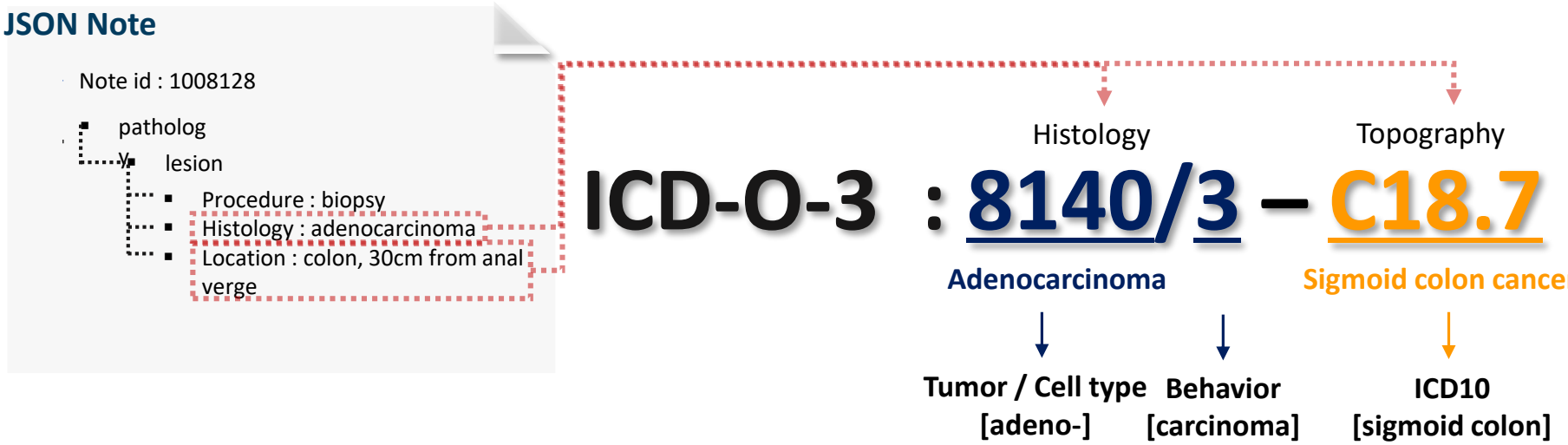
Note id : 1008128



- **The structured JSON form pathology report allows us to get the desired information in a nutshell**

Method :

Histology information were extracted from structured pathology report for target patients



- Topography and Histology information can be extracted from structured pathology report in addition to the primary condition (colon cancer), which enable us to reconstruct the condition concept Ids from ICD10 to ICD-O-3

Result :

Frequency of ICD-O-3 concept IDs

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

In total 112 colorectal cancer patients,

- **14 distinct ICD-O-3 concept IDs were assigned**
- **Most frequent concept ID was adenocarcinoma of sigmoid colon (ICD-O-3 : 8140/3-C18.7)**

Result :

Generation of Disease Occurrence Episode

Episode table

Field	Content
Episode_id	0012321
Person_id	0001234
Episode_concept_id	32528[Disease First Occurrence]
Episode_source_value	**Adenocarcinoma of sigmoid colon**

Episode event table

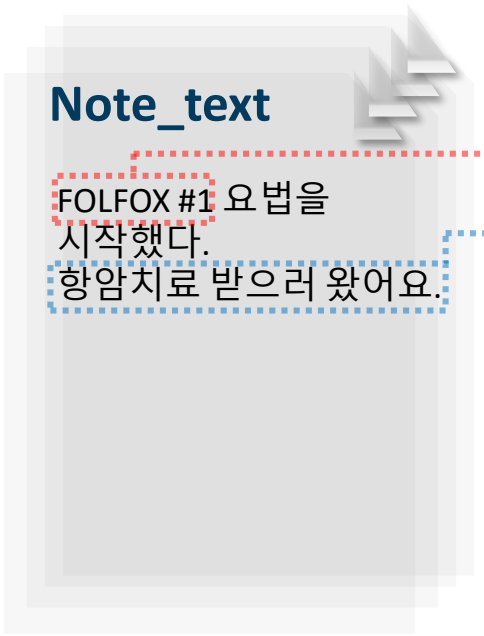
Field	Content
Episode_id	0012321
Condition_occurrence_id	1001234
Procedure_occurrence_id	
Drug_exposure_id	

Condition occurrence table

Field	Content
condition_occurrence_id	1001234
Person_id	0001234
condition_concept_id	4200514 [Adenocarcinoma of sigmoid colon]
condition_occurrence _start_datetime	2014-02-21

Challenge 2

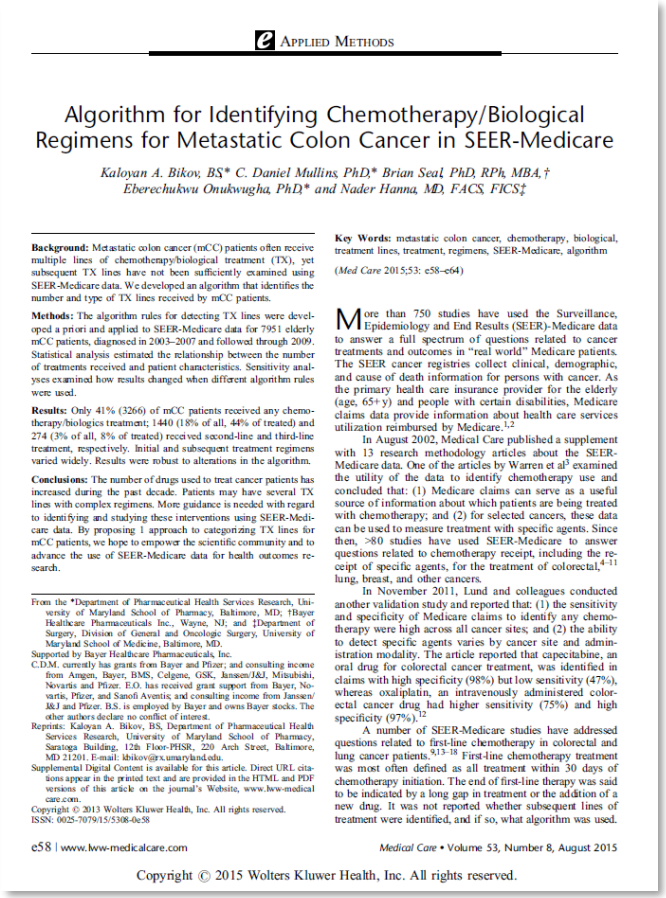
- **Treatment regimen information is not structured in EHR**



1. Treatment regimen is described in note table with other narrative text
2. Some patients did not have even any information about the treatment regimen
3. Which regimen was used or how many cycle did treatment tried were not machine-readable

Method :

Algorithm to extract treatment regimen from drug exposure



APPLIED METHODS

Algorithm for Identifying Chemotherapy/Biological Regimens for Metastatic Colon Cancer in SEER-Medicare

Kaloyan A. Bikov, BS, C. Daniel Mullins, PhD,* Brian Seal, PhD, RPh, MBA,† Eberechukwu Onukwugha, PhD,* and Nader Hanna, MD, FACS, FICS‡*

Background: Metastatic colon cancer (mCC) patients often receive multiple lines of chemotherapy/biological treatment (TX), yet subsequent TX lines have not been sufficiently examined using SEER-Medicare data. We developed an algorithm that identifies the number and type of TX lines received by mCC patients.

Methods: The algorithm rules for detecting TX lines were developed a priori and applied to SEER-Medicare data for 7951 elderly mCC patients, diagnosed in 2003–2007 and followed through 2009. Statistical analysis estimated the relationship between the number of treatments received and patient characteristics. Sensitivity analyses examined how results changed when different algorithm rules were used.

Results: Only 41% (3266) of mCC patients received any chemotherapy/biologics treatment; 1440 (18% of all, 44% of treated) and 274 (3% of all, 8% of treated) received second-line and third-line treatment, respectively. Initial and subsequent treatment regimens varied widely. Results were robust to alterations in the algorithm.

Conclusions: The number of drugs used to treat cancer patients has increased during the past decade. Patients may have several TX lines with complex regimens. More guidance is needed with regard to identifying and studying these interventions using SEER-Medicare data. By proposing 1 approach to categorizing TX lines for mCC patients, we hope to empower the scientific community and to advance the use of SEER-Medicare data for health outcomes research.

From the *Department of Pharmaceutical Health Services Research, University of Maryland School of Pharmacy, Baltimore, MD; †Bayer Healthcare Pharmaceuticals Inc., Wayne, NJ; and ‡Department of Surgery, Division of General and Oncologic Surgery, University of Maryland School of Medicine, Baltimore, MD.

Supported by Bayer Healthcare Pharmaceuticals, Inc.

C.D.M. currently has grants from Bayer and Pfizer; and consulting income from Amgen, Bayer, BMS, Celgene, GSK, Janssen/J&J, Mitsubishi, Novartis and Pfizer. E.O. has received grant support from Bayer, Novartis, Pfizer, and Sanofi Aventis; and consulting income from Janssen/J&J and Pfizer. B.S. is employed by Bayer and owns Bayer stocks. The other authors declare no conflict of interest.

Reprints: Kaloyan A. Bikov, BS, Department of Pharmaceutical Health Services Research, University of Maryland School of Pharmacy, Saratoga Building, 12th Floor-PHSR, 220 Arch Street, Baltimore, MD 21201. E-mail: kbikov@rx.umaryland.edu.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Website, www.lww-medicalcare.com.

ISSN: 0025-7079/15/5308-0e58

Key Words: metastatic colon cancer, chemotherapy, biological, treatment lines, treatment, regimens, SEER-Medicare, algorithm

(*Med Care* 2015;53: e58–e64)

More than 750 studies have used the Surveillance, Epidemiology and End Results (SEER)-Medicare data to answer a full spectrum of questions related to cancer treatments and outcomes in "real world" Medicare patients. The SEER cancer registries collect clinical, demographic, and cause of death information for persons with cancer. As the primary health care insurance provider for the elderly (age, 65+ y) and people with certain disabilities, Medicare claims data provide information about health care services utilization reimbursed by Medicare.[1,2]

In August 2002, Medical Care published a supplement with 13 research methodology articles about the SEER-Medicare data. One of the articles by Warren et al[3] examined the utility of the data to identify chemotherapy use and concluded that: (1) Medicare claims can serve as a useful source of information about which patients are being treated with chemotherapy; and (2) for selected cancers, these data can be used to measure treatment with specific agents. Since then, >80 studies have used SEER-Medicare to answer questions related to chemotherapy receipt, including the receipt of specific agents, for the treatment of colorectal,[4–11] lung, breast, and other cancers.

In November 2011, Lund and colleagues conducted another validation study and reported that: (1) the sensitivity and specificity of Medicare claims to identify any chemotherapy were high across all cancer sites; and (2) the ability to detect specific agents varies by cancer site and administration modality. The article reported that capecitabine, an oral drug for colorectal cancer treatment, was identified in claims with high specificity (98%) but low sensitivity (47%), whereas oxaliplatin, an intravenously administered colorectal cancer drug had higher sensitivity (75%) and high specificity (97%).[12]

A number of SEER-Medicare studies have addressed questions related to first-line chemotherapy in colorectal and lung cancer patients.[8,13–18] First-line chemotherapy treatment was most often defined as all treatment within 30 days of chemotherapy initiation. The end of first-line therapy was said to be indicated by a long gap in treatment or the addition of a new drug. It was not reported whether subsequent lines of treatment were identified, and if so, what algorithm was used.

- Previous study suggested an algorithm identifying therapy regimen applied to SEER-medicare
- Fragmented drug exposure records were leveraged as low level data of treatment

Method :

Algorithm to extract treatment regimen from drug exposure

- Index date was based on the diagnosis of colon cancer in pathology report.

Index date

Drug_exposure_start_date	Drug
2014-08-14	Megestrol Acetate
2014-08-14	calcium polycarbophil
2014-08-14	Lactulose
2014-08-18	Metoclopramide
2014-08-18	Dexamethasone
2014-08-18	Fluorouracil
2014-08-18	Dexamethasone
2014-08-18	calcium polycarbophil
2014-08-18	Irinotecan hydrochloride
2014-08-18	Magnesium Oxide
2014-08-18	Leucovorin
2014-08-20	Atropine Sulfate
2014-08-20	calcium polycarbophil

Method :

Algorithm to extract treatment regimen from drug exposure

- **We screened drugs of interest from drug exposure data**

  fluorouracil, leucovorin, oxaliplatin, capecitabine, irinotecan, cetuximab, and bevacizumab

Drug_exposure_start_date	Drug
2014-08-18	Metoclopramide
2014-08-18	Dexamethasone
2014-08-18	**Fluorouracil**
2014-08-18	Dexamethasone
2014-08-18	calcium polycarbophil
2014-08-18	**Irinotecan hydrochloride**
2014-08-18	Magnesium Oxide
2014-08-18	**Leucovorin**
2014-08-20	Atropine Sulfate
2014-08-20	calcium polycarbophil

Method :

Algorithm to extract treatment regimen from drug exposure

- The drugs of interest on the same drug exposure start date were bundled to determine the regimen.

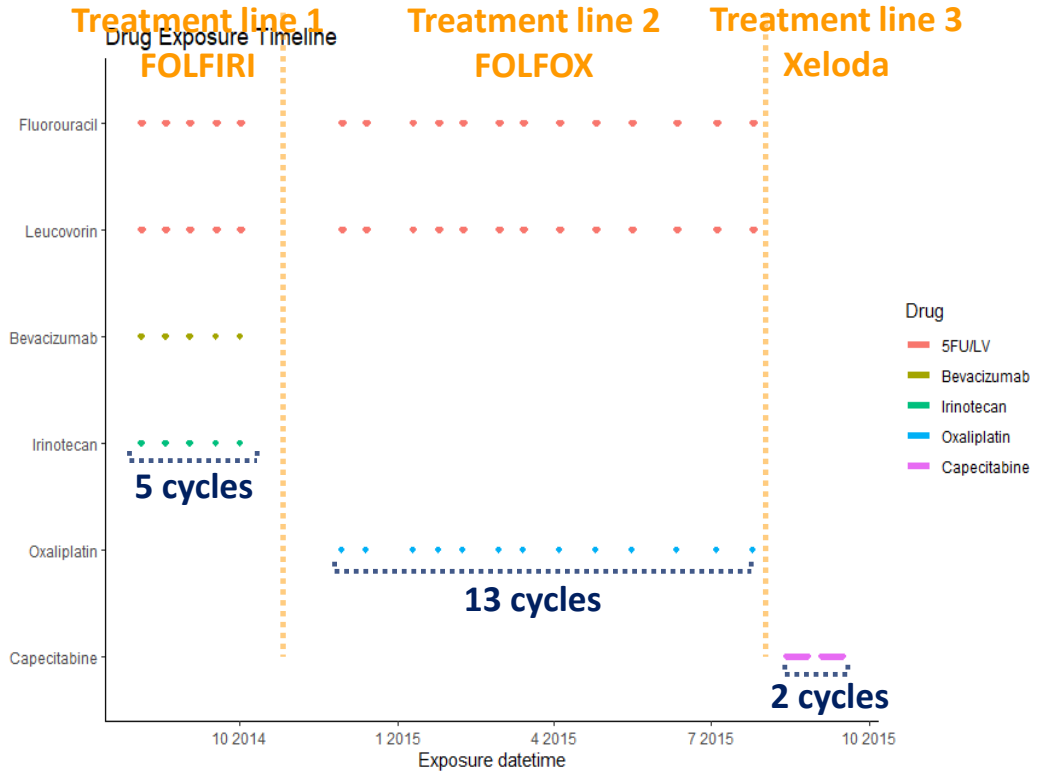
Drug_exposure_start_date	Drug
2014-08-18	Fluorouracil
2014-08-18	Irinotecan hydrochloride
2014-08-18	Leucovorin

FOLFIRI

FOL - folinic acid (leucovorin)
F - fluorouracil (5-FU)
IRI - irinotecan

Method :

Extracted regimen information were used to generate treatment line and total cycle



- Each treatment line lasted until the new drug was used.
- Each bundled drug exposure data were considered as each cycle

Result :

Treatment regimen distribution in target cohort

- Of the 112 colon cancer patients in target cohort, 28 (25%) received FOLFOX as a first-line treatment. Subsequently, 4 (14%) patients received FOLFIRI as a second-line therapy.

- Large portion of patients were not fully followed up in EHRs

Result :

Treatment regimen episode were generated

Episode table

Field	Content
Episode_id	0012321
Person_id	0001234
Episode_concept_id	32531[Treatment Regimen]
Episode_source_value	**FOLFOX**

Episode event table

Field	Content	Content	Content
Episode_id	0012321	0012321	0012321
Condition_occurrence_id			
Procedure_occurrence_id			
Drug_exposure_id	20001234	20001235	20001236

Drug exposure table

Field	Content
drug_exposure_id	20001234
Person_id	0001234
drug_concept_id	1388796 [Leucovorin]

Field	Content
drug_exposure_id	20001235
Person_id	0001234
drug_concept_id	955632 [Fluorouracil]

Field	Content
drug_exposure_id	20001236
Person_id	0001234
drug_concept_id	1318011 [Oxaliplatin]
drug_exposure _start_datetime	2014-02-21

Result :

Treatment regimen episode were generated

Episode table

Field	Content
Episode_id	0012321
Person_id	0001234
Episode_concept_id	32531[Treatment Regimen]
Episode_source_value	**FOLFOX**

Condition occurrence table

Field	Content
condition_occurrence_id	1001234
Person_id	0001234
condition_concept_id	4200514 [Adenocarcinoma of sigmoid colon]
condition_occurrence _start_datetime	2014-02-21

Episode event table

Field	Content	Content	Content	Content
Episode_id	0012321	0012321	0012321	
Condition_occurrence_id				1001234
Procedure_occurrence_id				
Drug_exposure_id	20001234	20001235	20001236	

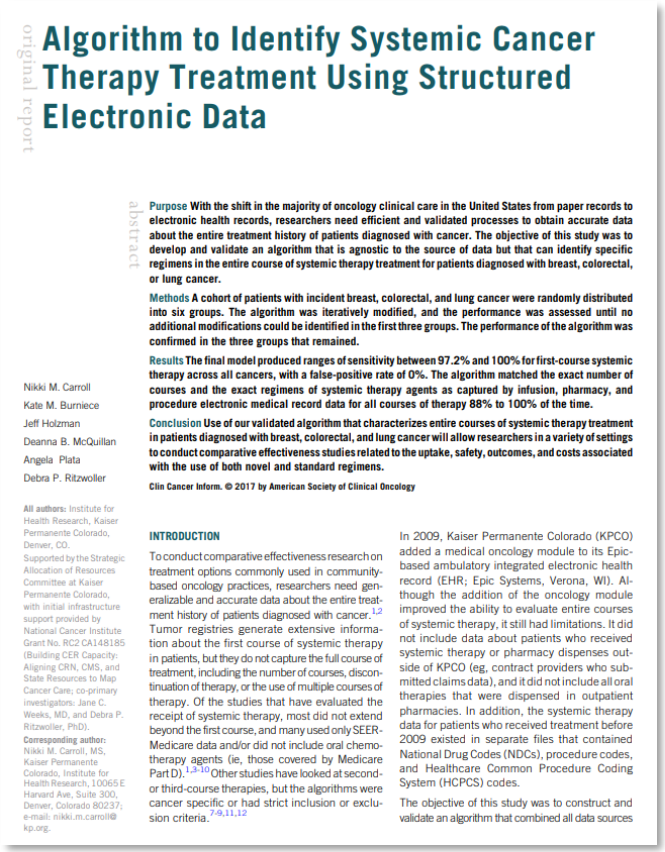
?

We were not sure if the condition occurrence table could be mapped to a treatment regimen episode event table.

Further study

- Recurrence / Progression / Stage / Surgery / Other procedures of treatment would be next step of study
- We aiming to be able to obtain the treatment cycle and regimen automatically.

Further study



original report

Algorithm to Identify Systemic Cancer Therapy Treatment Using Structured Electronic Data

abstract

Purpose With the shift in the majority of oncology clinical care in the United States from paper records to electronic health records, researchers need efficient and validated processes to obtain accurate data about the entire treatment history of patients diagnosed with cancer. The objective of this study was to develop and validate an algorithm that is agnostic to the source of data but that can identify specific regimens in the entire course of systemic therapy treatment for patients diagnosed with breast, colorectal, or lung cancer.

Methods A cohort of patients with incident breast, colorectal, and lung cancer were randomly distributed into six groups. The algorithm was iteratively modified, and the performance was assessed until no additional modifications could be identified in the first three groups. The performance of the algorithm was confirmed in the three groups that remained.

Results The final model produced ranges of sensitivity between 97.2% and 100% for first-course systemic therapy across all cancers, with a false-positive rate of 0%. The algorithm matched the exact number of courses and the exact regimens of systemic therapy agents as captured by infusion, pharmacy, and procedure electronic medical record data for all courses of therapy 88% to 100% of the time.

Conclusion Use of our validated algorithm that characterizes entire courses of systemic therapy treatment in patients diagnosed with breast, colorectal, and lung cancer will allow researchers in a variety of settings to conduct comparative effectiveness studies related to the uptake, safety, outcomes, and costs associated with the use of both novel and standard regimens.

Nikki M. Carroll
Kate M. Burniece
Jeff Holzman
Deanna B. McQuillan
Angela Plata
Debra P. Ritzwoller

All authors: Institute for Health Research, Kaiser Permanente Colorado, Denver, CO.

Supported by the Strategic Allocation of Resources Committee at Kaiser Permanente Colorado, with initial infrastructure support provided by National Cancer Institute Grant No. RC2 CA148185 (Building CER Capacity: Aligning CRN, CMS, and State Resources to Map Cancer Care; co-primary investigators: Jane C. Weeks, MD, and Debra P. Ritzwoller, PhD).

Corresponding author: Nikki M. Carroll, MS, Kaiser Permanente Colorado, Institute for Health Research, 10065 E Harvard Ave, Suite 300, Denver, Colorado 80237; e-mail: nikki.m.carroll@kp.org.

INTRODUCTION

To conduct comparative effectiveness research on treatment options commonly used in community-based oncology practices, researchers need generalizable and accurate data about the entire treatment history of patients diagnosed with cancer.[1,2] Tumor registries generate extensive information about the first course of systemic therapy in patients, but they do not capture the full course of treatment, including the number of courses, discontinuation of therapy, or the use of multiple courses of therapy. Of the studies that have evaluated the receipt of systemic therapy, most did not extend beyond the first course, and many used only SEER-Medicare data and/or did not include oral chemotherapy agents (ie, those covered by Medicare Part D).[1,3-10] Other studies have looked at second- or third-course therapies, but the algorithms were cancer specific or had strict inclusion or exclusion criteria.[7-9,11,12]

In 2009, Kaiser Permanente Colorado (KPCO) added a medical oncology module to its Epic-based ambulatory integrated electronic health record (EHR; Epic Systems, Verona, WI). Although the addition of the oncology module improved the ability to evaluate entire courses of systemic therapy, it still had limitations. It did not include data about patients who received systemic therapy or pharmacy dispenses outside of KPCO (eg, contract providers who submitted claims data), and it did not include all oral therapies that were dispensed in outpatient pharmacies. In addition, the systemic therapy data for patients who received treatment before 2009 existed in separate files that contained National Drug Codes (NDCs), procedure codes, and Healthcare Common Procedure Coding System (HCPCS) codes.

The objective of this study was to construct and validate an algorithm that combined all data sources

- **This paper, which was discussed in the last working group meeting, expected to extends the scope of cancer types**
- **We intend to produce algorithms that are commonly applied to various types of cancer.**

Thank you for listening !