Name:	Clair Blacketer
Affiliation:	Janssen Research & Development
Email:	mblacke@its.jnj.com
Presentation type (select one):	Poster

Converting SEER-Medicare to the OMOP Common Data Model

Margaret S. Blacketer, MPH[1,2], Jennifer L. Duryea, MPH[3], Amy Matcho[1,2], Jenna Reps[1,2]

[1]Janssen Research and Development, Raritan, NJ, [2]Observational Health Data Sciences and Informatics (OHDSI), New York, NY, [3]Outcomes Insights Inc., Westlake Village, CA

Abstract

SEER-Medicare data is a critical data resource for studying cancer incidence and survival in the US. The combination of registry and claims information make it a challenge to ETL properly because of the more complex internal logic of the database. We outline the technical difficulties encountered and proposed solutions for converting this database to the OMOP CDM version 5.

Introduction

The Surveillance, Epidemiology, and End Results (SEER) is a program overseen by the National Cancer Institute (NCI) and is the leading source of cancer incidence and survival rates in the US[1]. It is currently the only cancer registry that collects information on the stage of cancer at the time of diagnosis and patient survival[1]. This repository links SEER registry data to Medicare claims for individuals providing longitudinal healthcare data for Medicare-eligible enrollees and creating one of the most comprehensive data sources for studying healthcare utilization and outcomes in a US cancer population. This makes it an ideal candidate for conversion to the OMOP CDM version 5[2].

Methods

SEER Medicare data was analyzed for B-cell cancers from 1991 to 2011. We began the ETL conversion process to the OMOP CDM version 5[2] by referencing the guidelines for the Medicare synthetic public use data (SynPUF) created by Danese et al.[3]. The SynPUF dataset is missing many of the variables present in SEER-Medicare, but we were provided a good start for the basic logic and enabled us to build upon it. The following sections describe ETL problems unique to the SEER Medicare.

Understanding Enrollment

SEER Medicare variables are stored in the Patient Entitlement and Diagnosis Summary File (PEDSF). Monthly indicators of enrollment for Part A, Part B, HMO (Part C), and Part D Medicare plans were available. Payer plan records for each type of Medicare plan were created for each patient, where a positive indicator creates a date range spanning the month. Observation periods are created by calculating the months where the patient had both Part A and Part B Medicare coverage but did not have HMO coverage since any claims paid by the third-party HMO vendor are not available in the Medicare files. This is unique to other past ETLs where the observation period is calculated as the earliest date and latest date in the payer plan period record[4].

Visit Logic

There are four tables that contribute to visit creation in SEER-Medicare: MEDPAR, OUTSAF, NCH and DME. MEDPAR contains all Part A institutional inpatient claims, including skilled nursing facilities, OUTSAF contains all Part B institutional outpatient claims, NCH contains all Part B physician/supplier reimbursement claims and DME contains all Part B claims for medical equipment. The NCH and DME files also contain denied claims so those must be removed before assigning visits.

Claim dates, place of service values, and provider values were used to combine multiple claims together in one visit record. A hierarchy was used when creating the visit logic where claim dates from MEDPAR > OUTSAF > NCH or DME. This means if an OUTSAF claim's dates were contained entirely inside a MEDPAR claim's dates, then the OUTSAF claim would be combined with the MEDPAR claim and the combination would become an inpatient visit. We used the same date logic and added a place of service requirement when combining NCH and DME claims

with OUTSAF claims. Visits were then created for the remaining NCH and DME claims using place of service values to determine whether they should be categorized as emergency or outpatient.

Recording Costs

A significant departure from past ETLs was how SEER Medicare cost variables were converted into the CDM. SEER Medicare contains payment information for billed HCPCS, DRG, and NDC codes and the CDM has associated procedure cost, drug cost, and visit cost tables. However, CDM vocabularies require source codes be mapped to different analytic tables based on the definition of the source code, which could decouple the cost record with the procedure, drug, or visit record. For example, a HCPCS code could be mapped to the Observation table, resulting in decoupling the procedure cost record from the original HCPCS code. To address this issue, all HCPCS codes were stored in the Procedure_Occurrence table where associated procedure cost records were referenced. If the vocabularies suggested a source code should live in a table other than the Procedure_Occurrence table, the source code was copied to the assigned table with the correct concept_id. A copy of the source code stayed in the Procedure_Occurrence table with a procedure_concept_id = 0. To prevent double-entry of cost values, cost records with Revenue Code 0001 were removed from the database, since this code is an administrative code used on claims to indicate a "total cost" figure.

Cancer Information in SEER Variables

One of the main advantages of SEER Medicare data is the extensive information on a patient's cancer diagnosis. However, these SEER variables do not have a current mapping in the OHDSI Vocabularies. In this version of the ETL, the SEER variable name, along with the cancer file, and cell value was added to the Observation table. No mapping of CONCEPT_IDs or relevant table mappings were done. All PEDSF variables were copied.

Results

We compared the demographic breakdown of the patients before and after conversion in order to estimate the effectiveness of the ETL and there is less than a 0.01% difference in gender, race and ethnicity between the raw and CDM versions of SEER-Medicare.

Limitations

There are over 2,000 variables in the PEDSF table that we did not map to a standard vocabulary and instead chose to put in the OBSERVATION table. The relevant SEER variables in the PEDSF that are specific to clinical information on the cancer diagnosis need to be identified and mapped to relevant CDM tables. During the course of creating this ETL there was also a change in OMOP CDM v5 from using four separate cost tables to one comprehensive cost table. In the next version of the ETL we will address both of these issues as it will make the SEER-Medicare data even more useful to analysts and compliant with the new version 5 of the CDM.

Conclusion

We were able to successfully map the SEER-Medicare data to the OMOP CDM version 5^2 with <0.01% loss of person data though most of the logic was focused on the conversion of the Medicare claims. Additional work will need to be done to properly ETL SEER variables into the CDM.

References

1. Overview of the SEER Program. 2016; http://seer.cancer.gov/about/overview.html Accessed 13Jun2016.
2. OMOP Common Data Model. [Webpage]. 2015; http://www.ohdsi.org/data-standardization/the-common-data-model/, 20 Jul 2015.
3. Danese M, Voss E, Duryea J, et al. Feasibility of Converting the Medicare Synthetic Public Use Data Into a Standardized Data Model for Clinical Research Informatics. AMIA 2015.
4. Observational Health Sciences and Informatics Example ETLs. [Webpage]. 2015; http://www.ohdsi.org/web/wiki/doku.php?id=documentation:example_etls, 25 Jun 2015.