Augmenting the National COVID Cohort Collaborative (N3C) Dataset with Medicare and Medicaid (CMS) Data, Secure and Deidentified Clinical Dataset

Stephanie Hong, FAMIA[1], Thomas Richards, MS[6], Benjamin Amor PhD[2], Tim Schwab[2], Philip Sparks[2], Maya Choudhury[2], Saad Ljazouli, MS[2], Peter Leese, MSPH[4], Amin Manna, MEng[2],Christophe Roeder, MS[3], Tanner Zhang, MD, MS[1], Lisa Eskenazi[1], Bryan Laraway, MS[3], Nirvisha Garara, Rasi Talluri, James Cavallon[1], Eric Kim[1], Shijia Zhang, MS[1],Emir Amaro Syailendra, MD[1], Shawn O'Neil, PhD[3], Davera Gabriel, RN FHL7, FAMIA[1] Sigfried Gold, MS[1], Tricia Francis, MS[1], Andrew Girvin, PhD[2], Emily Pfaff, PhD, MS[4], Anita Walden, MS[3], Harold Lehmann, MD, PhD[1], Melissa Haendel, PhD[3], Ken Gersing MD[5], Christopher G Chute, MD DrPH[1], on behalf of the N3C Consortium

[1] Johns Hopkins University School of Medicine, Baltimore, MD; [2] Palantir Technologies, Denver, CO; [3] University of Colorado Anschutz Medical Campus, Aurora, CO; [4] University of North Carolina, Dept of Medicine, Chapel Hill, NC; [5] National Center for Advancing Translational Sciences, National Institutes of Health; [6] Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD

Background

The National COVID Cohort Collaborative (N3C)[1] is a US-based program that collects electronic medical record (EMR) data of patients with COVID-19 symptoms from over 240 participating care delivery organizations. The N3C Enclave, largest centralized repository of clinical observational data to date, [1] is a continuously refreshed data resource which follows the Observational Medical Outcomes Partnership (OMOP) Common Data Model.[2] However, EMR data may not have information from outpatient settings, such as visit to pharmacies, clinics and non-hospital institutional care facilities as well as telehealth visits, home healthcare, and home equipment and device purchases. To address this, we supplemented the N3C EMR datasets with a comprehensive CMS (Centers for Medicare and Medicaid Services) claims dataset by linking the two datasets using the Privacy Preserving Record Linkage[6] (PPRL) technology. In principle, this method provides a more holistic view of patients' healthcare journey and fills gaps of missing data for patients who received care from multiple providers. In this paper, we describe our work in developing a method to construct the CMS claims data as clinical encounters to complement and augment the OMOP-ified N3C datasets.

Methods

CMS claims record describes the reimbursement for care that the patient received. Medicare claims include inpatient, outpatient, Part D drug prescription, home health, hospice, durable medical equipment, and skilled nursing claims. Medicaid claims include inpatient, long-term care, other service and prescription claims. From the CMS claims data, we reorganized information describing specific claims for care that each patient received with regard to care site, diagnosis and procedure codes, and prescribed medications into a patient-centric data following the OMOP common data model.[2]

The claim files were parsed and reshaped from denormalized tabular format into a long format such that the terminology codes found in the claims were consolidated into a single code column and a single code system column by source code vocabulary type. The code systems in the CMS data include ICD10CM, ICD-10-PCS, HCPCS, CPT4 and NDC codes. The codes in the claim files were spread across multiple columns anywhere from column01 to column45 in multiple rows. Some files were over 4000 columns wide. The claims files were transformed from wide data format into a long format, and then the code map translation crosswalk table was generated using the Code Map Service. Figure 1 below describes the Code Map Service architecture. Then, the generated code map crosswalk mapping table is used to transform the claims data into OMOP CDM.

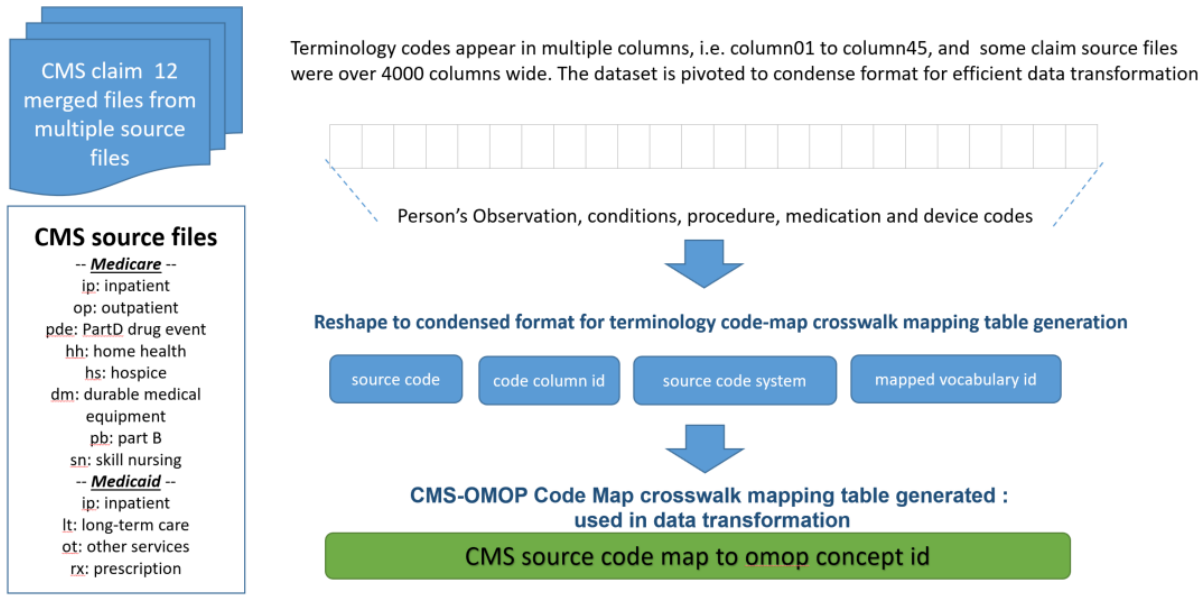

Figure 1 - Code Map service, data transformation into OMOP CDM is performed using generated code map crosswalk mapping table

The visit_occurrence table is constructed using the dates found in the claims files. Visit is based on the admission and discharge dates or service begin and end date reported on the claims data. Contiguous claims filed under weekly or monthly time blocks are merged as one visit. Visit dates across all claim source files are also reviewed and if overlapping dates are found without a gap in days then they are merged into the macro visit dataset with minimum of the earliest date as the macro_visit_start_date and the maximum of the latest date as the macro_visit_end_date. Similar to N3C EMR data, these additional macro visit dates are added to the visit_occurrence table to provide an overlapping timeline of inpatient visit, professional visit, facility, pharmacy, device or durable medical equipment claims data that may exist for the same beneficiary on the claims.

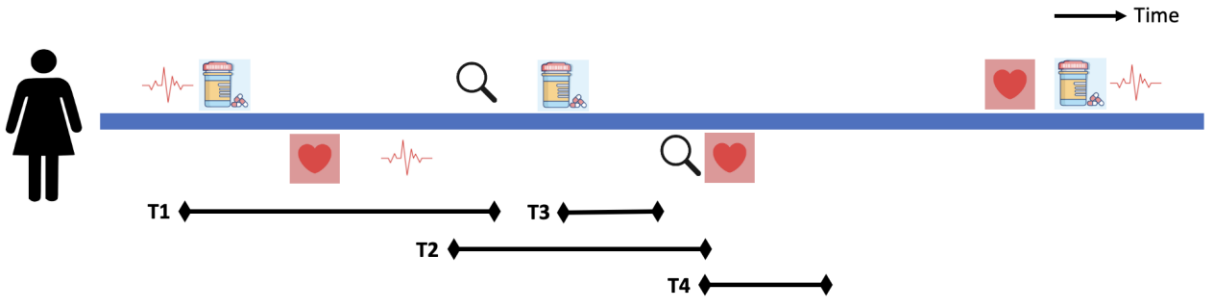

Figure 2 - The figure above shows an example of a merged overlapping or contiguous with zero days gap timeline that is used to build the macro visit timeline, t1_start_date as macro_visit_start_date and t4_end_date as macro_visit_end_date.

The diagnosis, procedures and medication codes present on the claim for a given beneficiary are inserted to the OMOP domains following the OMOP CDM terminology guidelines. The references to the visits are matched based on person, dates and provider. All the terminology codes mentioned in the claims are translated into OMOP standard concepts. Visits are defined using the following key fields from the claims source files.

CMS	Claims Source file	Key fields for visit/encounter
Medicare	Inpatient (IP)	Person, provider, admission date and discharge date. TYPE_ADM and revenue center code determines IP and IPER visits.
Medicare	Outpatient(OPL)	Person, provider, rev_dt
Medicare	Part D (PDE)	Person, provider, rx_dos_dt, rx_end_dt
Medicare	Part B (PB)	Person, provider, expnsdt1, expnsdt2
Medicare	Home Health (HH)	Person, provider, from date, thru date
Medicare	Hospice (HS)	Person, provider, from date, thru date
Medicare	Skilled Nursing (SN)	Person, provider, admission date, max(discharge date)
Medicare	Durable Medical Equipment(DM)	Person, provider, expnsdt1, expnsdt2
Medicaid	Inpatient(IP)	Person, provider, admsn_date, dschrgdt
Medicaid	Long-term care(LT)	Person, provider, srvc_bgn_dt, srvc_end_dt
Medicaid	Other services (OT)	Person, provider, srvc_bgn_dt, srvc_end_dt
Medicaid	Prescriptions(RX)	Mdcd_pd_dt = rx_start_date = rx_end_date

Figure 3 - dates used for the visit construct by claim type

Results

Using the methods we described, we built the Medicare and Medicaid data pipeline to transform the CMS claims datasets to conform with the OMOP common data model (CDM). Converting the clinical concepts in the CMS claims files, originally in a wide format, into a long format made the transformation process efficient and cogent. Claims dates are used to define a patient encounter (i.e. Visit Concept), and each relevant Clinical Concept is then tied to that encounter, establishing a Clinical Event. The Clinical Events are then inserted into OMOP CDM domain tables following the OMOP terminology conventions. As the CMS Data Ingestion and Harmonization pipeline is built, data health checks and OMOP schema validation checks are performed automatically. Next, the Privacy Preserving Record Linkage (PPRL) dataset is then used to "join" each N3C patient with the corresponding CMS patient. Any data from N3C sites that are not participating in the PPRL CMS collection process is filtered out. In cases where N3C person_id is duplicated, where a same patient is found in multiple N3C sites, a Global Person ID (GPI) record identifier is provided for each such N3C patient. Currently, 41 sites are participating in the PPRL linkage, 22 sites have opted into the PPRL CMS linkage, and 26 sites have opted into the PPRL mortality evidence linkage. Among the PPRL-linked N3C patient, on the average from Medicare claims, 60% patients have additional 6.33 procedure concepts, 71% patients have additional 78 condition concepts, 75% patients have additional 21.83 drug concepts, 60% patients have additional 16.48 measurement concepts, 66.9% patients have additional 8.60 observation concepts, and 4.6% patients have additional 6.8 device concepts. Additionally, among the PPRL-linked N3C patient, on the average from Medicaid claims, 20.2% patients have additional 23 procedure concepts, 20.8% patients have additional 33.9 condition concepts, and 21.9% patients have additional 21 drug concepts, 17.8% patients have additional 17.44 measurement concepts,18.3% patients have additional 6.68 observation concepts, and 13.9 patients have additional 6.3 device concepts.

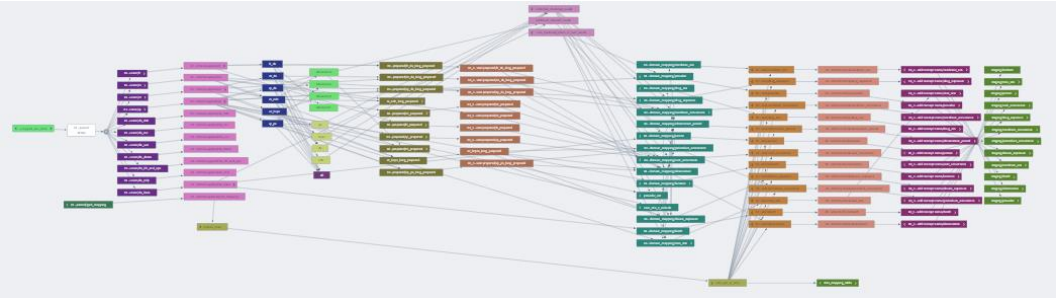

Figure 4 - CMS transformation pipeline. There are two transformation pipelines, one for Medicare and another for Medicaid.

Conclusion

We transformed the CMS claims data into OMOP CDM dataset to supplement the N3C EMR information with a comprehensive CMS claims dataset. Using PPRL linkage N3C EMR data is enriched with visit to the pharmacy, outpatient, Part D drug prescription, home health, hospice, durable medical equipment, and skilled nursing claims data. The PPRL linkage provides additional clinical information for PPRL linked patients to render a more holistic view of their healthcare journey and fills gaps of missing data for those who received care from multiple providers. Longitudinal outcome of patients in the N3C cohort is improved using claims data on therapeutics, comorbid diagnosis, vaccinations, and health care utilization. This data is refreshed monthly for Medicare and annually for Medicaid. It is available for research within the N3C Enclave community.

Acknowledgements

N3C Data Use Request: 9b45ea07-e1e0-497b-bbc1-8520b61e73b8

IRB00327758: National Clinical Cohort Collaborative (N3C): A national resource for shared analytics

The process described in this abstract was conducted with data and tools accessed through the NCATS N3C Data Enclave https://covid.cd2h.org and supported by NCATS U24 TR002306, Axle Informatics Subcontract: NCATS-P00438-B. This work was made possible because of the patients whose data was contributed by partner organizations (covid.cd2h.org/dtas). We gratefully acknowledge the individuals who have contributed to the on-going development of this community resource (covid.cd2h.org/acknowledgements).

References

1. Haendel M, Chute C, Gersing K. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. J Am Med Inform Assoc. 2020 [cited 3 Sep 2020]. doi:10.1093/jamia/ocaa196
2. OHDSI CDM Working Group. OMOP Common Data Model. In: OMOP Common Data Model [Internet]. Dec 2021 [cited 25 Jan 2022]. Available: https://ohdsi.github.io/CommonDataModel/
3. CMS Research Assistance Center Data Variables [Internet]. https://resdac.org/search-data-variables
4. Application of Episode Groupers to Medicare, [Internet] Oct 2013. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/ActiveProjectReports/Active-Projects-Reports-Items/CMS1187287
5. Aronson, J. (n.d.). Landscape Analysis of Privacy Preserving Patient Record Linkage Software (P3RLS). National Cancer Institute Division of Cancer Control & Population Sciences. https://surveillance.cancer.gov/reports/TOP1-PPRLS-Landscape-Analysis.pdf
6. Grannis SJ, Kho A, Phua J, Kasthurirathne SN. Evaluation of token collections and matching models to support privacy-preserving record linkage (PPRL). AMIA 2021 Proc Symp, Fall Symposium (in press).

N3C Publication Intent Form Submitted : MSID:1464.346