Automated Production of Research Data Marts from a Canonical Fast Healthcare Interoperability Resource (FHIR) Data Repository: Applications to COVID-19 Research

Leslie A Lenert [1,2], Andrey V. Ilatovskiy[1,2], James Agnew[3], Patricia Rudsill[1,2], Jeff Jacobs[2], Duncan Weatherston[3] and Kenneth Deans[2]

[1]Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC, USA
[2]Health Sciences South Carolina, Columbia, SC, USA
[3]Smile CDR, Toronto, CN

For inquiries contact:

Leslie A. Lenert MD, MS, FACP, FACMI
Assistant Provost for Data Science and Informatics
Medical University of South Carolina
22 West Edge Suite 13
Charleston, SC 29425
Email: Leslie.Lenert@gmail.com
Telephone: 843-792-4268

Abstract

Objective: The rapidly evolving COVID-19 pandemic has created a need for timely data from the healthcare systems for research. To meet this need, several large new data consortia have developed that require frequent updating and sharing of electronic health record (EHR) data in different common data models (CDMs) to create multi-institutional databases for research. Traditionally, each CDM has had a custom pipeline for extract, transform and load operations for production and incremental updates of data feeds to the networks from raw EHR data. However, the demands of COVID-19 research for timely data are far higher, and the requirements for updating faster than previous collaborative research using national data networks. New approaches need to be developed to address these demands.

Methods: In this paper, we describe the use of the Fast Healthcare Interoperability Resource (FHIR) data model as a canonical data model and the automated transformation of clinical data to the Patient-Centered Outcomes Research Network (PCORnet) and Observational Medical Outcomes Partnership (OMOP) CDMs for data sharing and research collaboration on COVID-19.

Results: FHIR data resources could be transformed to operational PCORnet and OMOP CDMs with minimal production delays through a combination of real-time and post-processing steps, leveraging the FHIR data subscription feature.

Conclusions:The approach leverages evolving standards for the availability of EHR data developed to facilitate data exchange under the 21st Century Cures Act and could greatly enhance the availability of standardized datasets for research.

Introduction

The COVID-19 pandemic has illustrated the need for reliable rapidly accessible data from electronic health record (EHR) systems for research on risk factors, predictive models, and evaluation of emerging diseases. Moreover, the lack of reliable large data sets has led to spurious research findings early in the COVID-19 [1]. Two of the largest consortia leverage existing infrastructure for shared data collaboration. The National COVID Collaborative Cohort (N3C) [2] is an alliance among Clinical and Translational Research Grant Awardees sponsored by the National Center to Advancing Translational Science (NCATS). This network leverages past experiences and infrastructure from the Accrual to Clinical Trials Network [3]. N3C's preferred data model for accepting results is the Observational Medical Outcomes Partnership (OMOP) model maintained by the Observational Health Data Sciences and Informatics (OHDSI) collaborative [4]. However, N3C accepts data in a variety of formats. A second consortium is based around the Patient-Centered Outcomes Research Network (PCORnet) [5] and leverages prior investments on comparative effectiveness research across this large research network [6] to create its database. There also are private large research networks, for example, TriNetX [7], that maintain a large data network of patients with COVID-19 for research from its clinical trial eligibility network [8]. The FDA maintains several large networks for safety evaluation of drugs and devices that are also being applied to problems identified during the pandemic [9]. In addition, some of the same partners in N3C are also using the integrating informatics for integrating biology and bedside (i2b2) platform [10] to study COVID-19. Many networks have overlapping membership and, as a result, members have to maintain duplicative data production processes.

As the pandemic has evolved rapidly, so have the requirements for rapid data updating in these large networks. Minimizing the lag between production of the data through the care of patients using an EHR system and its availability for research increases the relevance of the network to the evolving set of problems seen with COVID-19. In the prior modes of operation, MUSC's ACT and PCORnet data networks could be used for data operations with lags of three or more months for production, and cycles for new releases of data sets every quarter. In the COVID era, the specifications for the N3C network call for 2-week production cycles for data releases and one-month lags between the closure of a record and its availability within the network. More current data might be even more valuable as new variants of the virus and new therapies emerge. This is a challenging task that requires the automation of processes for analytic database production.

Production of data for each network is, in itself, a multistep pipeline process that involves mapping and transformation of data to the preferred data model of a research network. Work for data production for different networks is often done in parallel, which is logistically challenging and consumes limited resources. Sometimes work is done in series, mapping from source data to one data model, and then another, which could potentially result in a loss of data through compression or inaccuracies in mapping. In this paper, we describe the use of the Fast Healthcare Interoperability Resources (FHIR) standard data model as a *canonical model* for initial storage of the data for subsequent transformation to more analytics-oriented models (OMOP and PCORnet) as well as an architecture for multiple simultaneous largely automated translations from FHIR to these two CDMs. This is a particularly important task as the 21st Century Cures Act [11] will require availability from EHRs in FHIR standards for the United States Core Data for Interoperability standards [12], which, while evolving, already has many of the required elements for research CDMs.

Methods

The approach taken to standardize a data production pipeline for multiple analytic CDMs from FHIR builds on one of the central tools for the FHIR paradigm: a clinical data repository (CDR) designed to store, persist, retrieve, and deliver FHIR resources. A widely used implementation for this operation is the open-source HAPI FHIR engine [13]. We build our system based on the Smile CDR platform [14] that is powered by the HAPI FHIR engine. This platform can accept data in a variety of formats: JSON or XML encoded FHIR objects, HL7 v2.x messages, flat comma-delimited files, and transform these data elements to FHIR resources that are stored in a proprietary relational format, for efficient search and retrieval. Alternatively, some vendors persist FHIR resources using a “data lake” approach, with extensive indexing but a minimal transformation of the data [15].

A standards-specified feature of FHIR CDRs is automated tooling to allow *subscriptions* to specific FHIR resources [14,16]. Subscriptions in the FHIR standard are triggers attached to FHIR data resources. Creating or updating a resource triggers a function that allows copying and transmission of the resource data object to another system. A common use for subscriptions in the FHIR standard is for notification of events. For example, if a patient is registered in an emergency department, a new FHIR resource with the registration information is created, and this then triggers sending a copy of the FHIR resource to another system via FHIR API with JSON payload or other interoperability protocol. This results in the second computer system becoming “aware” of the notification of registration.

We adapted this feature for use in data transformation in our "Federated on FHIR (F-on-F)" architecture [17]. F-on-F is an architecture that replaces HSSC's legacy cross-institutional integrated data repository with a series of linked FHIR data repositories with a single centralized master person index maintained by subscriptions. F-on-F also uses subscriptions to admission discharge and transfer data for ADT notification and for automated updating of local repositories with centralized data on mortality, geocodes and social determinants of health. A full description of F-on-F is beyond the scope of this paper.

At the individual site level, whenever a new FHIR data resource is created in the clinical FHIR repository, in our system, a copy of the resource is created in a second linked FHIR repository using the subscription mechanism. However, rather than persisting the object in the proprietary database format of the vendor, we co-develop with Smile CDR rule-based transformations implemented in Java to persist data in two different targeted analytical models, OMOP and PCORnet. This results in a system of linked CDMs updated within milliseconds. The approach is illustrated in Figure 1. Both OMOP and PCORnet models are extended to deal with identified data and data elements not covered by the CDM specifications. These are kept in separate tables to preserve the functioning of CDM quality inspection and analytics software; the approach also preserves subject anonymity. The specific examples of FHIR Patient resource mapping to OMOP CDM tables and extensions are illustrated in Supplemental Table S1.

The real-time transformation to analytic CDMs poses additional obstacles due to the transactional nature of the EHR data that evolves and expands for days or even weeks after a given patient encounter, while analytic data models assume a static self-consistent set of data. For example, the OMOP model assumes that each patient has a date of birth (known at least with a year precision); in the EHR data the demographics details might be missing and such patients should be removed from the OMOP instance. Another example, in the OMOP model both visits and associated clinical facts have patient IDs, allowing transitivity violations that occur in the EHR data due to patient merges. If data are incomplete or inconsistent at some point in time, and evolve to completeness and correctness as time passes, the data in the live CDM instances are kept in sync with the source and all updates will be propagated via pipeline. The issue is minimal for study feasibility queries, however, for longitudinal data analysis these inconsistencies need to be resolved. The solution we adopted was a production pipeline with separate static OMOP and PCORnet instances for post-processing to reconcile the data.

Another issue in design was maintaining the robustness of the process to evolutionary changes in the CDMs. Ongoing changes in CDMs are the rule rather than the

exception. The OHDSI group releases a new set of OMOP vocabularies weekly, with changes ranging from adding a few new concepts to complete redesign of the domain organization. Many other networks require frequent vocabulary updates (e.g. once a month for N3C). In case of a major change, the post-processing using an Extract Transform and Load (ETL) approach implemented via SQL was flexible enough to accommodate rapid changes in vocabularies, in contrast to the Java code transformations used in subscription-based mappings. Essentially, the post-processing allows a quasi-incremental approach to the vocabulary updates: the clinical tables built with the old vocabularies (within the pipeline) are combined with the fresh set of vocabularies and all deviations from the new vocabularies are corrected, with the majority of the data being untouched. This flexibility requires preserving enough "rawness" of the clinical data so updates to an analytic CDM do not require a complete rebuilding of the database. The drawback is that "live" versions of the database produced through subscriptions cannot be used for complex analyses without post-processing (although, again, quick study "feasibility" queries, such as counts of patients with specific e-phenotypes, are possible).

Data security is maintained using a variety of approaches. SmileCDR supports direct queries of the FHIR database using Smart-on-FHIR authentication [18]. OMOP and PCORnet databases have been extended with patient data elements in separate data tables. Databases retain their original clinical temporal labels and as such are not truly de-identified data sets. Access to data is controlled by governance, including investigator data use agreements and by honest brokers who produce data sets based on institutional review board approved protocols. Additionally, PCORnet and OMOP access tools (SAS and Atlas) are restricted to the standard CDM tables with limited identifying information.

To test the feasibility of this architecture, we converted the Medical University of South Carolina research data warehouse (RDW) and operational production of PCORnet and OMOP CDMs for MUSC to use a novel process based on this concept model. Specific versions details of this implementation: FHIR version is 3.0.2; OMOP CDM version is currently 5.3.1; and the PCORnet CDM is version 6.0. The database for data management operations and for PCORnet queries is Oracle 19.6.0. Oracle tools are used for maintenance and data manipulation. The Smile CDR also uses Oracle to persist the data but it is not limited to this platform. At MUSC analytic operations for OMOP transform and persist reference data sets using SQL Server 2019. The FHIR CDR is assessed remotely via virtual private network linkages to the HSSC's data center at Clemson University. SAS, used to run the PopMedNet queries, is version 9.4.

The transformation process and maintenance pipeline was based upon the export of a series of flat pipe-delimited files that were loaded into the FHIR CDR but supports many options for importing EHR data, including HL7 FHIR transactions and HL7 v2.x transactions. In this instance, large export files were then processed to instantiate the repository with pre-existing data. An incremental updating approach based on extracting new data into a flat-file from the RDW was also developed. This incremental update is extracted daily, and loaded into the FHIR repository. The FHIR repository and “live” CDMs are updated incrementally. In FHIR, any changes from the previously persisted data are recorded as new versions of the resources. OMOP and PCORnet "live" versions of the databases only store the most recent value. Restoration of data elements in the FHIR database automatically results in updating of the other two CDMs via the subscription mechanism. The staging and production OMOP and PCORnet instances for release are rebuilt de novo for quarterly releases. The environments for FHIR, OMOP, and PCORnet require about 2, 0.5 and 0.4 terabytes respectively.

Composite time for database production including the application of post-processing steps in the pipeline was observed. For the OMOP instance, data quality measures were computed using both the in-house reports and OHDSI Achilles v1.6.7 and DQD v1.0(develop) tools. For the PCORnet instance, standardized database quality assessment routines were computed and applied. Results of prior assessments of data quality for PCORnet certification were compared to this new approach for the generation of the database.

The iterative data quality assessment cycles resulted in numerous improvements, that were mostly implemented as post-processing steps with the expectation that they will be pushed upstream into the main pipeline if they are not constrained by the transactional nature of the data. For instance, the mappings to the expected terminologies were gradually improved in terms of completeness and correctness and then could be applied at any stage. In contrast, the data clean up steps that remove data of insufficient quality were limited to the analytic production instances only since future data updates to the "live" instances might improve the data quality and thus save these bits of information.

Work on and participation in the HSSC data warehouse program is conducted under IRB PRO0009273.

Results

Figure 2 shows the time delays with different approaches to capture of "raw" EHR data. Flat file exports result in delays of days while data accumulate for export at each stage. Export from our Epic EHR to Epic's Clarity database occurs nightly. Data are extracted from this database daily and stored in a linked clinical model based on EHR data model with minimal transformation, and then exported as flat files for conversion to FHIR. There are other available approaches such as HL7 v2.x or FHIR data streams for other contributing data partners. Early data products support (blue shading in the figure) trial feasibility studies (counts); final products meet network quality requirements and support longitudinal analyses. Processing for conversion to FHIR, PCORnet and OMOP occurs at HSSC's Clemson database facility designed to support multiple institutions, each with their own segmented, but linked, FHIR infrastructure.

Table 1 shows the main administrative and clinical data domains implemented in the F-on-F architecture, applied to the MUSC data for the period from 2014-07-01 to 2020-12-31. The results reveal two primary findings. First, the data entry location indicates how the data are represented in each layer. In most cases, there is a 1:1 correspondence between the CDMs, but certain domains are not straightforward. For instance, the vital signs and the lab results are separate domains but exist as common Observation resources in FHIR, stored in both the Measurement and Observation tables in OMOP, and in the individual tables only in PCORnet.

Table 1. Comparison of Source RDW, FHIR, OMOP and PCORnet CDMs for MUSC

	Source	FHIR		OMOP		PCORnet	
Domain	**Count**	**Resource**	**Count**	**Table**	**Count**	**Table**	**Count**
Patient	1,078,964	Patient	1,063,886	Person	1,059,009	Demo-graphic	1,063,891
Visit	10,746,491	Encounter	10,636,834	Visit Occurrence	10,628,243	Encounter	10,636,928
Diagnosis	18,402,862	Condition	17,593,910	Condition Occurrence	14,254,546	Diagnosis	17,594,043
Procedure	23,029,835	Procedure	22,246,999	Procedure Occurrence	16,838,402	Procedures	22,247,280
MedOrder	37,948,002	MedicationRequest	32,693,681	Drug Exposure	31,829,609	Prescribing	32,703,095
MedAdmin	84,450,696	Medication Administra-tion	43,577,863	Drug Exposure	39,562,252	Med-Admin	39,574,769
Vital	16,917,012	Observation	16,127,660	Measurement + Observation	16,128,048	Vital	16,127,920

Lab	137,984,163	Observation	124,870,259	Measurement + Observation	154,831,038	Lab_result_CM	119,992,129

Second, the majority of the metrics are highly consistent between all stages of the pipeline. There are about 11M visits for slightly over 1M patients, associated with 18M diagnoses and 22M procedures. The differences between the CDMs are due to several factors. Most of them are common to all domains: there are certain source data entries that are not processed into the pipeline; on the other hand, the OMOP and PCORnet CDMs require certain data entries to be removed. The most dramatic differences are OMOP-specific due to the domain assignments (e.g. a diagnosis code might be classified as an observation concept) and one-to-many mappings.

In addition to the main resources/tables shown on Figure 1, the F-on-F architecture implementation involved other resources/tables, including linked resources based on the relational model (e.g. FHIR DiagnosticReports), extensions to capture the data elements not covered by the standard specifications, and mapping-related supporting elements (e.g. FHIR ConceptMaps). The specific examples of FHIR to OMOP mapping are illustrated in Supplemental Table S1.

With regard to specifics of data quality, the OHDSI DQD had 3312 data quality checks, including conformance and plausibility tests, and the OMOP instance passed 3092 (93%). Some of the failed checks were due to DQD technical errors (submitted on github). Some data issues cannot be resolved without significant effort for a limited impact (e.g. medication mapping is still somewhat incomplete, with an additional 20K+ codes needed for comprehensive coverage for the last five percent by volume of medication prescriptions and administrations). The PCORnet data quality was verified after each quarterly refresh by executing the Empirical Data Check (EDC) SAS package was provided by the PCORnet Distributed Research Network Operations Center. There were 1450 data checks validating that all data elements were populated as expected, column lengths were correct, mappings conformed to the specification, relationships were logical, referential integrity was respected and more. Our PCORnet instance passed 1424 (98%) of those checks. An extended metrics report (generated by the custom SQL) is provided in the Supplementary Table S2.

Initial loading of the backlog of 5-years of EHR data into the FHIR instance required several weeks. However, once loaded, the approach resulted in significant improvements in the time required to produce quarterly updates of the PCORnet instance and improved the concurrency of data in refreshes. As shown in Figure 3, preparation time was far less. Significant work was still required in post-processing but now could be focused on data quality.

Discussion

The sustainability of research networks for COVID-19 and emerging disorders is an important issue. Costs arise in part due to custom data modeling, ETL tasks, and repetitive data integration tasks required for operations. Further, research infrastructure has to be replicated at each site, requiring significant additional investments. A unique feature of the design is that rather than rely upon the FHIR representation as to the means for query and retrieval of data, the architecture uses FHIR subscriptions to trigger continuous transformations to other common data models that are maintained in synchrony. This approach also allows the selection of data for specific subsets of patients. Linked databases are available for a query with minimal time lag behind the data source for queries for counts and other simple operations. When an analytical-grade quality of the instance is required, the post-processing can be applied on-demand to produce analytical data sets.

The primary innovations in this work are the use of FHIR as an initial canonical data model and FHIR subscription protocols for the transformation and synchronization of multiple data models. In future work, we will explore the use of subscription models to distribute data across networks and to maintain shared data elements, such as mortality status and social determinants of health data. We will also explore the use of this approach to federate clinical data across sites by maintaining a single master patient identifier and consistent supporting demographic information.

Prior approaches to the problem of maintenance of multiple linked data models in a data repository have focused on other canonical models and *automation* of data both transformation for queries and production of data sets. For example, work from the i2b2 group at Harvard has used the i2b2 data format as the canonical representation to support dynamic ETL from that format to the OMOP and to FHIR [19]. Ong and co-authors use the OMOP model as their canonical representation of data and support dynamic queries mapped from the PCORnet CDM [20]. Choi and colleagues [21] have developed automated mapping functions from OMOP to FHIR for computation.

There is also prior work with the use of FHIR as a meta map for ETL operations between clinical data models. Pfaff and coauthors [22] describe the use of CampFHIR, a tool for guiding ETL for conversions of different models. FHIR concept representation aids and speeds a largely manual ETL process. This approach guides current N3C efforts [2]. The F-on-F approach described herein was developed contemporaneously with CampFHIR [17] with both efforts benefiting from collaborations. F-on-F differs in that it is focused on ongoing data production through automated maintenance of parallel CDM instances. Data transformations are implemented in Java code and executed in

near real time. The FHIR CDR servers also provide reference storage and master data management of EHR data and a live canonical representation of the data. The approach may allow us to better understand what is lost in translation.

F-on-F combines near real time data transformations with post-processing for production release databases. We envision the scope of real time processing as being pragmatic, scaled to use cases intended. For example, if a site wanted to run OMOP models for prediction of sepsis from clinical data, then real time processing might need to be expanded to meet those needs.

The use of FHIR as a data source from EHRs is important from a policy perspective as Information Blocking Statutes stemming from the 21st Century Cures Act specify a range of clinical data to be available from EHRs for downloading and data exchange (the USCDI) [11]. They further require that EHRs respond to queries for USCDI elements in the FHIR standard both at an individual patient query level and population level starting in December of 2022. Providers who cannot offer data access in this format may be subject to fines for “information blocking”. EHR vendors must offer these capabilities to maintain certification of their systems as compliant with Meaningful Use regulations. As a result, many barriers for data production for research and safety will be overcome if the starting point for data transformation operations for computational models is the FHIR standard.

The broad future availability of data in the FHIR standard raises the question of whether other analytically oriented models are necessary. OMOP and PCORnet are highly evolved models refined for their purpose [23]. As analytical models, they are optimized for efficient and unbiased analysis of large volumes of longitudinal normalized data. The FHIR model is an object-oriented data model, focused on the accurate expression of clinical events, not computation. F-on-F envisions a best of both worlds approach, with flexible representation and optimized computation.

Limitations

The above approach is standards-based but leverages proprietary extensions of the FHIR subscription specification. As discussed above, there are inherent limitations in the speed with which clinical data can be integrated into any analytic model such as OMOP or PCORnet. For example, orders or laboratory data cannot be linked to an encounter that does not (yet) exist. The use of the persistence module for the transformation of data is novel and computationally efficient but implements rules in compiled Java code, where changes may be more difficult. Ultimately, some manual

ETL was still deemed optimal in the production pipeline; however, future work may reduce these requirements.

Summary

The use of FHIR standard as a canonical representation of clinical data with the subsequent dynamic transformation to other research CDMs for analytics is a practical approach to accelerate the availability of data for research and may be particularly useful for evolving diseases such as COVID-19. While it is theoretically possible to fully automate transformation to near real-time versions of OMOP or PCORnet databases, it is more practical given the evolving nature of data to take a staged approach for models for longitudinal data analysis applications.

Acknowledgments

Funding: This project was supported by grants to Health Sciences South Carolina from the Duke Endowment and by the National Center for Advancing Translational Sciences of the National Institutes of Health under Grant Number UL1 TR001450.

Competing Interests: Dr. Lenert, Dr. Ilatovskiy, Ms. Rudsill, Mr. Jacob, and Mr. Deans have no competing interests Mr. Agnew and Mr. Weatherston have financial interests in Smile CDR and are employed by the company

Contributorship: Dr. Lenert designed the study, helped obtain funding, and wrote major parts of the manuscript. Dr. Illatovsky, Ms. Rudsill, and Mr. Jacobs contributed to the conduct of the study and the writing of the manuscript. Mr. Deans helped obtain funding for the study and contributed to the writing of the manuscript. Mr. Agnew and Mr. Weatherston contributed to the design, the conduct of the study, and to the writing of the manuscript.

Data availability: The data underlying this article will be shared on reasonable request to the corresponding author.

References

1 Piller C. Disgraced COVID-19 studies are still routinely cited. *Science* 2021;**371**:331–2.

2 Haendel MA, Chute CG, Gersing K. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. Journal of the American Medical Informatics Association. 2020. doi:10.1093/jamia/ocaa196

3 Visweswaran S, Becich MJ, D'Itri VS, *et al.* Accrual to Clinical Trials (ACT): A Clinical and Translational Science Award Consortium Network. *JAMIA Open* 2018;**1**:147–52.

4 Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform* 2015;**216**:574–8.

5 Block JP, Marsolo KA, Nagavedu K, *et al.* Characteristics of 24,516 patients diagnosed with COVID-19 illness in a national clinical research network: Results from PCORnet®. bioRxiv. 2020. doi:10.1101/2020.08.01.20163733

6 PCORnet: Progress, Challenges, and Opportunities Ahead. 2015.https://www.pcori.org/blog/pcornet-progress-challenges-and-opportunities-ahead (accessed 24 Nov 2019).

7 TriNetX Readies its Real-World Data Platform and Global Network of Healthcare Organizations to Support COVID-19 Clinical Research. 2020.https://trinetx.com/covid-19/ (accessed 31 Jan 2021).

8 Taquet M, Luciano S, Geddes JR, *et al.* Bidirectional associations between COVID-19 and psychiatric disorder: retrospective cohort studies of 62 354 COVID-19 cases in the USA. *Lancet Psychiatry* 2021;**8**:130–40.

9 FDA Sentinel System's Coronavirus (COVID-19) Activities. 2020.https://www.sentinelinitiative.org/assessments/coronavirus-covid-19 (accessed 31 Jan 2021).

10 Brat GA, Weber GM, Gehlenborg N, *et al.* International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *npj Digital Medicine* 2020;**3**:109.

11 Mandl KD, Gottlieb D, Mandel JC, *et al.* Push Button Population Health: The SMART/HL7 FHIR Bulk Data Access Application Programming Interface. *NPJ Digit Med* 2020;**3**:151.

12 U.S. Core Data for Interoperability (USCDI) | Interoperability Standards Advisory (ISA). https://www.healthit.gov/isa/us-core-data-interoperability-uscdi (accessed 24 Nov 2019).

13 Braunstein ML. SMART on FHIR. Health Informatics on FHIR: How HL7's New API is Transforming Healthcare. 2018;:205–25. doi:10.1007/978-3-319-93414-3_10

14 Braunstein ML. FHIR. In: Braunstein ML, ed. *Health Informatics on FHIR: How HL7's New API is Transforming Healthcare*. Cham: : Springer International Publishing 2018. 179–203.

15 Azure API for FHIR. https://azure.microsoft.com/en-us/services/azure-api-for-fhir/ (accessed 14 Feb 2021).

16 Subscription - FHIR v4.0.1. https://www.hl7.org/fhir/subscription.html (accessed 24 Nov 2019).

17 Health Sciences South Carolina lights up a FHIR-based clinical data repository. 2018.https://www.healthcareitnews.com/news/health-sciences-south-carolina-lights-fhir-based-clinical-data-repository (accessed 30 Jan 2021).

18 Mandel JC, Kreda DA, Mandl KD, *et al.* SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *J Am Med Inform Assoc* 2016;**23**:899–908.

19 Klann JG, Phillips LC, Herrick C, *et al.* Web services for data warehouses: OMOP and PCORnet on i2b2. *J Am Med Inform Assoc* 2018;**25**:1331–8.

20 Ong TC, Kahn MG, Kwan BM, *et al.* Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading. *BMC Med Inform Decis Mak* 2017;**17**:134.

21 Choi M, Starr R, Duke J. OMOP on FHIR as an Enabler for Analytics-As-A-Service. In: *AMIA*. 2017.

22 Pfaff ER, Champion J, Bradford RL, *et al.* Fast Healthcare Interoperability Resources (FHIR) as a Meta Model to Integrate Common Data Models: Development of a Tool and Quantitative Validation Study. *JMIR Med Inform* 2019;**7**:e15199.

23 Weeks J, Pardee R. Learning to Share Health Care Data: A Brief Timeline of Influential Common Data Models and Distributed Health Data Networks in U.S. Health Care Research. *EGEMS (Wash DC)* 2019;**7**:4.

Figure Legends

Figure 1. Adapting a FHIR CDR for real-time ETL to OMOP and PCORnet CDMs.

Figure 2. Computational pipeline for simultaneous multi-CDM production. Shaded boxes show technically available computational products. Approximate delays (relative to the previous step) are shown in the bottom.

Figure 3. Production timelines for PCORnet database release

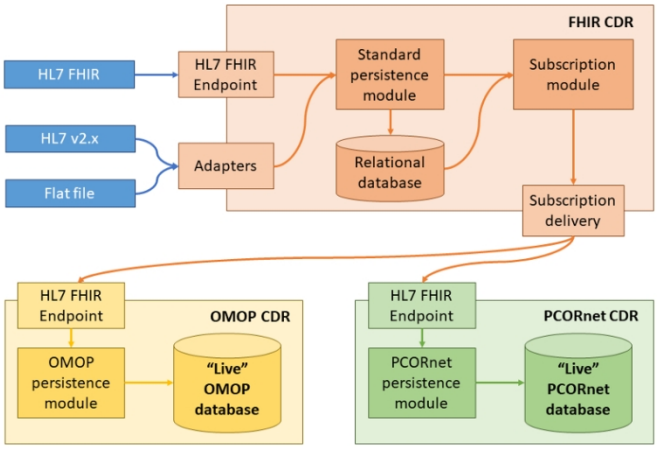

Figure 1. Adapting a FHIR CDR for real-time ETL to OMOP and PCORnet CDMs.

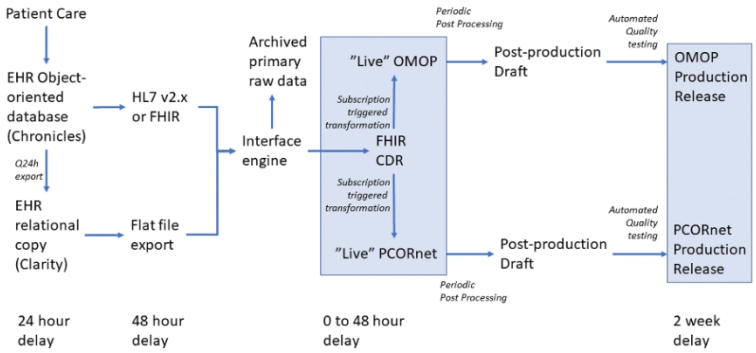

Figure 2. Computational pipeline for simultaneous multi-CDM production. Shaded boxes show technically available computational products. Approximate delays (relative to the previous step) are shown in the bottom.

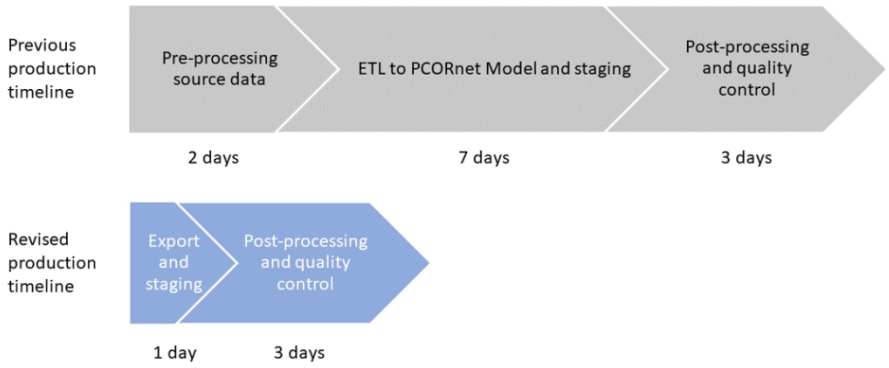

Figure 3. Production timelines for PCORnet database release

FHIR STU3		OMOP CDM v5.3.1			OMOP HSSC extensions		
Patient (standard)	Patient (extensions)	Person	Death	Location	HSSC_Person	HSSC_Death	HSSC_Location
id		person_id	person_id		person_id	person_id	
birthDate		year_of_birth					
		month_of_birth					
		day_of_birth					
		birth_datetime					
gender		gender_source_value			gender_source_system		
		gender_source_concept_id					
		gender_concept_id					
	race	race_source_value			race_source_system		
		race_source_concept_id					
		race_concept_id					
	ethnicity	ethnicity_source_value			ethnicity_source_system		
		ethnicity_source_concept_id					
		ethnicity_concept_id					
identifier.system[identifier.type.coding.code == "MR"]					med_rec_num_system		
identifier.value[identifier.type.coding.code == "MR"]					med_rec_num		
identifier.value[identifier.system == "http://hl7.org/fhir/sid/us-ssn"]					social_security_number		
identifier.value[identifier.system == "EMPI"]					empi_id		
active					status_source_value		
name.family					last_name		
name.given[0]					first_name		
name.given[1]					middle_name		
name.prefix					prefix_name		
name.suffix					suffix_name		
telecom.value[telecom.system == "phone" && telecom.use == "home"]					home_phone		
telecom.value[telecom.system == "phone" && telecom.use == "mobile"]					mobile_phone		
telecom.value[telecom.system == "phone" && telecom.use == "work"]					work_phone		
telecom.value[telecom.system == "email"]					email		
maritalStatus					marital_status_source_value		
					marital_status_source_system		
generalPractitioner					pcp_provider_id		
communication.language					language_source_value		
					language_source_system		
	militaryStatus				military_status_source_value		
					military_status_source_system		
	religion				religion_source_value		
					religion_source_system		
managingOrganization					source_care_site_id		
deceasedBoolean			death_date				
			death_datetime				
						deceased_flag	
deceasedDateTime			death_date				
			death_datetime				
	causeOfDeath		cause_source_value			cause_source_system	
						cause_source_display	
	deathDatePrecision					death_date_precision	
deceasedSource						source_of_value	
address		location_id		location_id			location_id
address.line[1]				address_1			
address.line[2]				address_2			
address.city				city			
address.district				county			
address.state				state			
address.postalCode				zip			
address.country							country_source_value
address.use							location_use
address.type							location_type

7/1/2014- 12/31/2020		Source		FHIR		PCORnet		OMOP		Comments
Domain	**Measurement**	**Count**	**%**	**Count**	**%**	**Count**	**%**	**Count**	**%**	
Demographic	Total	1,078,964		1,063,886		1,063,891		1,059,009		Source: calculated based on the encounter admit date. FHIR: calculated based on the encounter admit date; includes stubs. OMOP: Patients are removed when they have no encounters in the valid date range or have no DOB (includes stubs). PCORnet: Patients are removed when they have no encounters in the
	Age group: 0-4 years	59,686	5.5%	58,631	5.5%	58,368	5.5%	58,599	5.5%	
	Age group: 5-14 years	130,508	12.1%	129,159	12.1%	129,267	12.2%	129,192	12.2%	the index date: death
	Age group: 15-21 years	91,657	8.5%	89,765	8.4%	89,638	8.4%	89,829	8.5%	date if any or 1/1/2021,
	Age group: 22-64 years	542,996	50.3%	529,395	49.8%	529,968	49.8%	530,472	50.1%	whichever is first
	Age group: 65+ years	254,068	23.5%	249,614	23.5%	248,100	23.3%	250,917	23.7%	
	Age group: missing	46	0.0%	7,321	0.7%	8,550	0.8%	0	0.0%	OMOP: patients without DOB are deleted
	Ethnicity: Hispanic	38,055	3.5%	37,620	3.5%	37,529	3.5%	37,632	3.6%	
	Ethnicity: Not Hispanic	938,418	87.0%	925,297	87.0%	924,083	86.9%	927,188	87.6%	
	Ethnicity: Missing	102,491	9.5%	100,969	9.5%	102,279	9.6%	94,189	8.9%	
	Sex: Female	592,313	54.9%	580,722	54.6%	580,083	54.5%	581,814	54.9%	
	Sex: Male	484,836	44.9%	474,139	44.6%	473,529	44.5%	475,431	44.9%	
	Sex: Missing OR Other	1,815	0.2%	1,714	0.2%	10,279	1.0%	1,764	0.2%	

	Race: American Indian or Alaska Native	1,459	0.1%	1,437	0.1%	1,416	0.1%	1,418	0.1%
	Race: Asian	9,276	0.9%	9,118	0.9%	9,089	0.8%	9,106	0.9%
	Race: Black or African American	305,540	28.3%	300,660	28.3%	300,297	27.8%	301,644	28.5%
	Race: Native Hawaiian or Other Pacific Islander	890	0.1%	875	0.1%	873	0.1%	874	0.1%
	Race: White	641,421	59.4%	631,810	59.4%	630,996	58.5%	633,007	59.8%
	Race: Other	66,275	6.1%	64,125	6.0%	64,004	5.9%		in standard vocabularies
	Race: Unknown	38,853	3.6%	36,380	3.4%	54,759	5.1%		in standard vocabularies
	Race: Patient Refused	2,550	0.2%	2,440	0.2%	2,457	0.2%		in standard vocabularies
Death	Total	27,521	2.6%	27,441	2.6%	57,139	5.3%	28,318	2.7% cleaned up. PCORnet:
	death date	25,639	2.4%	25,533	2.4%	25,041	2.3%	25,543	2.4% OMOP: HSSC extension
	death flag	1,882	0.2%	1,908	0.2%	1,878	0.2%	2,775	0.3% OMOP: HSSC extension
	Death data from state	0	0.0%	0	0.0%	50,589	4.7%	0	0.0% only loaded in PCORnet
Address History	address data)	1,073,928	99.5%	1,056,691	99.3%	1,055,001	99.2%	1,055,556	99.7% v5
	History	5,036	0.5%	7,195	0.7%	8,890	0.8%	3,453	0.3% removed since cannot be
Vital	Weight and Smoking status	16,917,012		16,127,660		16,127,920		16,128,048	
	Height: total records	3,693,984		3,659,567		3,659,656		3,659,527	
	Height: mean	62		62		62		62	in
	Height: median	65		65		65		65	in
	Patients with Height Vital	707,019	65.5%	689,090	64.8%	689,094	64.8%	689,001	65.1%
	Weight: Total Records	5,824,820		5,859,912		5,860,024		5,860,476	
	Weight: Mean	157		157		157		157	lb
	Weight: Median	164		164		164		164	lb
	Patients with Weight Vital	773,257	71.7%	753,754	70.8%	753,755	70.8%	753,650	71.2%
	Smoking status: total records	7,398,208		6,608,181		6,608,240		6,608,045	

	Everyday	127,510	11.8%	92,279	8.7%	92,280	8.7%	92,265	8.7%	
	Smoking status: Former	161,100	14.9%	153,670	14.4%	153,670	14.4%	153,652	14.5%	Patients can contribute
	Smoking status: Never	571,414	53.0%	549,732	51.7%	549,735	51.7%	549,690	51.9%	to more than one
	smoking statuses	148,144	13.7%	141,381	13.3%	36,729	3.5%	141,333	13.3%	category.
	(including status NI)	885,857	82.1%	835,815	78.6%	862,750	81.1%	835,705	78.9%	
Encounter	Total	10,746,491		10,636,834		10,636,928		10,628,243		FHIR: includes stubs.
	Patients with Encounters	1,079,094	100.0%	1,063,886	100.0%	1,063,891	100.0%	1,055,314	99.7%	
	Encounters per Patient	9.96		10.00		9.92		10.04		
	Encounter Type: Ambulatory	9,899,185	92.1%	9,789,238	92.0%	9,789,329	92.0%	9,781,035	92.0%	
	Encounter Type: Emergency	543,478	5.1%	535,972	5.0%	535,976	5.0%	535,842	5.0%	
	Encounter Type: Inpatient	303,828	2.8%	301,706	2.8%	301,705	2.8%	301,504	2.8%	
	Encounter Type: Missing	0	0.0%	9,918	0.1%	9,918	0.1%	9,862	0.1%	
	Distinct Providers	35,916		20,120		12,623		12,604		
	Provider	17,554	0.2%	70,583	0.7%	42,331	0.4%	70,166	0.7%	
	Encounters with Clinical data	10,340,028	96.2%	10,164,847	95.6%	10,167,576	95.6%	10,143,916	95.4%	
Diagnosis	Total	18,402,862		17,593,910		17,594,043		14,254,546		condition_occurrence
	Patients with Diagnoses	822,026	76.2%	782,614	73.6%	782,622	73.6%	727,589	68.7%	condition_occurrence
	Patients without Diagnoses	229,570	21.3%	281,272	26.4%	281,269	26.4%	331,434	31.3%	condition_occurrence
	Diagnoses per Patient	17.06		16.5		16.5		13.5		condition_occurrence
	Code: ICD-10	15,770,936	85.7%	14,987,411	85.2%	14,987,544	85.2%	12,179,407	85.4%	condition_source_conce
	Code: ICD-09	2,631,926	14.3%	2,606,499	14.8%	2,606,499	14.8%	2,067,484	14.5%	condition_source_conce
	Principal DX: yes	5,743,251	31.2%	5,395,330	30.7%	940,657	5.3%	155,275	1.1%	OMOP: HSSC extension
	Principal DX: no	12,671,248	68.9%	12,198,580	69.3%	5,251,737	29.8%	295,316	2.1%	OMOP: HSSC extension
	Principal DX: missing	0	0.0%	0	0.0%	11,401,649	64.8%	13,803,955	96.8%	OMOP: HSSC extension
	DX present of arrival: yes	2,673,399	14.5%	2,604,915	14.8%	2,604,940	14.8%	2,648,793	18.6%	OMOP: HSSC extension

	DX present of arrival: no	364,145	2.0%	1,201,539	6.8%	1,201,561	6.8%	537,275	3.8%	OMOP: HSSC extension
E in RDW	DX present of arrival: Other	868,129	4.7%	0	0.0%	0	0.0%	0	0.0%	FHIR: treated as 'N'.
	(No information or Unknown)	14,497,189	78.8%	13,787,456	78.4%	13,787,542	78.4%	11,068,478	77.6%	OMOP: HSSC extension
	Distinct Providers	0		0		8,710		0		associated with
	Missing Provider	18,402,862	100.0%	17,593,910	100.0%	15,223	0.1%	14,254,546	100.0%	associated with
Procedure	Total	23,029,835		22,246,999		22,247,280		16,838,402		procedure_occurrence
	Patients with Procedures	906,448	84.0%	896,231	84.2%	896,234	84.2%	894,212	84.4%	procedure_occurrence
	Patients without Procedures	174,134	16.1%	167,655	15.8%	167,657	15.8%	164,797	15.6%	procedure_occurrence
	Procedures per Patient	21.3		20.9		21		15.9		procedure_occurrence
	Encounters with Procedures	8,666,204	80.6%	8,602,766	80.9%	8,602,836		8,205,810	77.2%	procedure_occurrence
	Code: ICD-09	177,766	0.8%	154,236	0.7%	154,236	0.7%	72,573	0.4%	standard ICD9Proc
	Code: ICD-10	448,457	1.9%	432,168	1.9%	432,176	1.9%	431,371	2.6%	standard ICD10PCS
	Code: CH	22,421,619	97.4%	21,047,534	94.6%	21,105,533	94.9%	15,426,992	91.6%	standard HCPCS and CPT4
	Code: Other	0	0.0%	613,061	2.8%	555,335	2.5%	907,466	5.4%	including ones from the
	Distinct Providers	7,352		7,349		7,349		7,118		procedure_occurrence
	Missing Provider	79	0.0%	16	0.0%	15	0.0%	187,864	1.1%	procedure_occurrence
Lab	Total	137,984,163		124,870,259		119,992,129		154,831,038		and lab results are filtered
	Patients with labs	692,459	64.2%	627,743	59.0%	627,327	59.0%	756,467	71.4%	results from
	Distinct LOINC codes	3,365		0		2,477		2,491		LOINC. OMOP: lab orders
	Mapped to LOINC code	93,552,391	67.8%	0	0.0%	115,116,446	95.9%	113,805,533	73.5%	LOINC. OMOP: lab orders
Med Orders	Total	37,948,002		32,693,681		32,703,095		31,829,609		orders as provided in the
	Patients with MedOrders	823,961	76.4%	778,504	73.2%	778,250	73.2%	777,920	73.5%	orders as provided in the
	Mapped to RxNorm	33,578,829	88.5%	0	0.0%	31,101,640	95.1%	26,256,778	82.5%	RxNorm. OMOP:
	Distinct RxNorm codes	2,240		0		6,197		4,783		RxNorm. OMOP:

Med Admin	Total	84,450,696		43,577,863		39,574,769		39,562,252		statuses are filtered out.
	Patients with MedAdmins	490,185	45.4%	459,434	43.2%	458,375	43.1%	458,268	43.3%	admins as provided in the
	Mapped to RxNorm	67,412,510	79.8%	0	0.0%	39,574,769	100.0%	34,138,226	86.3%	RxNorm. OMOP:
	Distinct RxNorm codes	1,576		0		7,866		2,334		RxNorm. OMOP: