SMEs optimization with high precision data ingestion of CAPriCORN CDM onto OMOP at AllianceChicago

Andrew Hamilton, RN, BSN, MS[1], Amro Hassan, MSA, MSE[1], Davera Gabriel, RN FHL7 FAMIA[2], Guy Tsafnat PhD FAIDH[2]
AllianceChicago[1], Evidentli[2]

Background

AllianceChicago is a not-for-profit organization that amalgamates data from 81 community health centers in the Chicago metropolitan area and beyond[1]. AllianceChicago participates in several research and public health networks including as a member of the CAPriCORN group[2], a Chicago-based, regional patient-centered research network, and the All of Us[3] network. AllianceChicago experiences significant challenges to effective data utilization due to the diversity of member organizations' implemented data warehouses based on divergent medical record systems. In order to participate in network analysis, data sources are harmonized to a CAPriCORN data model that is derived from the PCOR Common Data Model (PCOR)[4]and further translated into other common data models, primarily the OMOP Common Data Model (OMOP)[5].

Transformation of the data from the source systems to the CAPrICORN CDM and translation from that into OMOP, are time consuming processes that require scarce and specialized skills including familiarity with both PCOR and OMOP and a strong command of SQL and other data manipulation languages. Alliance Chicago evaluated Evidentli's Piano platform and its unique AI-supported ETL development environment that is claimed to provide significant acceleration and strong translation fidelity [6]. To this end, Alliance Chicago embarked on a study to transform a temporal subset of data in the CAPriCORN model into OMOP as required by the All of Us network.

Methods

The study source data came from the Athena Practice EMR System, which employs GE Centricity coded observations (HDID)[7]. A cohort of 1000 patients were identified in the Capricorn Data Mart, and 100,000 records representing all available attributes for this cohort were extracted. The project team augmented these extracted records with Observation terms and Procedure code names and their definitions which provided a more detailed and context-rich understanding of the patient data within the CAPriCORN Data Mart. The project team developed a CDM structure mapping by selecting corresponding tables/fields from the Capricorn CDM that correspond to existing tables/fields in the OMOP CDM (v5.4) which provided the foundation for the ETL comparison (Figure 1). Twelve Capricorn tables were mapped to 11 OMOP CDM tables and for each table, individual fields (columns) were also mapped.

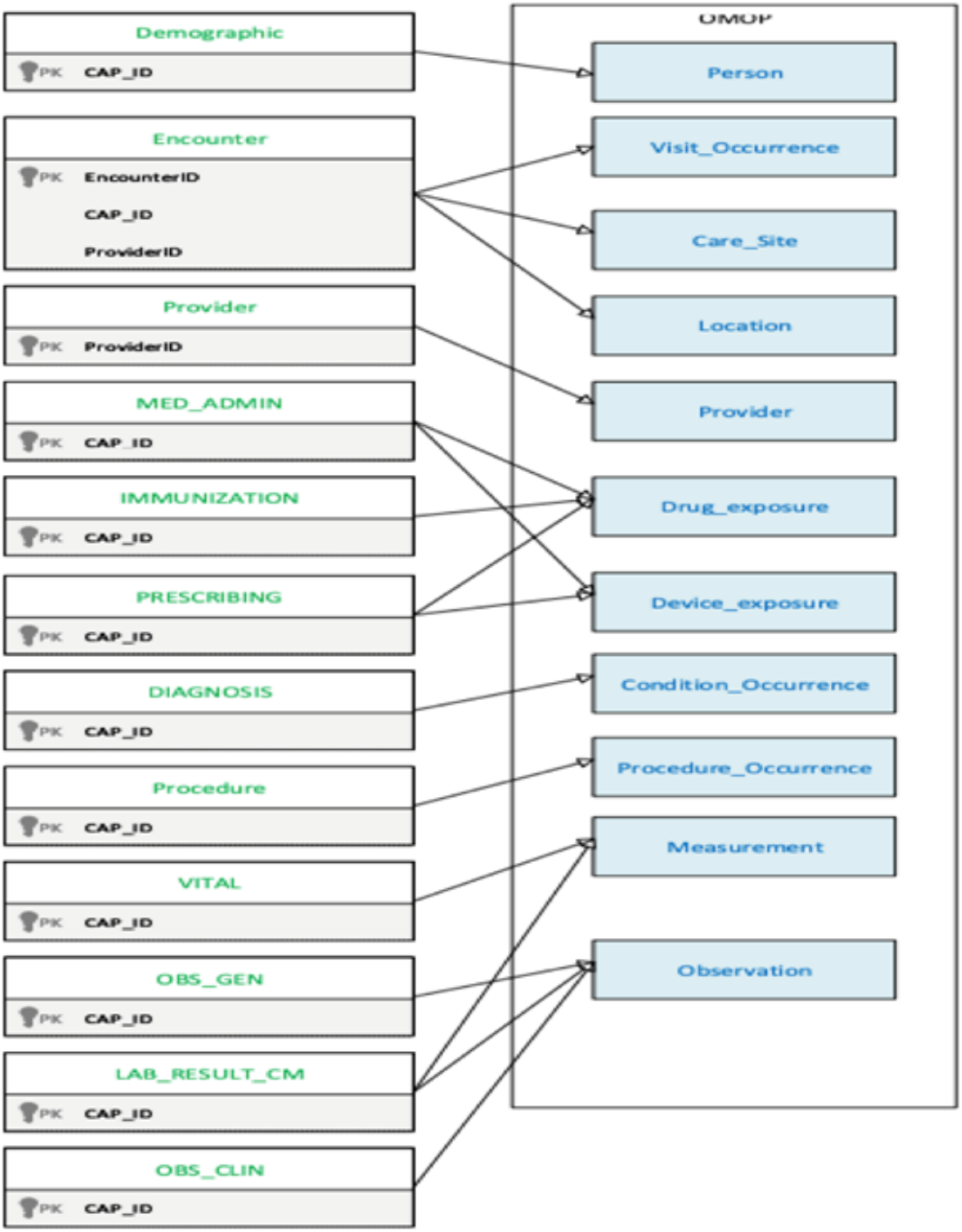

Figure 1. Mappings at a tabular level between the Capricorn and OMOP definitions

The Piano data ingress workflow utilized a series of automated operations starting from data acquisition, advancing through transformation and quality control, and culminating in the storage of processed data. Extracted patient records were staged on the PIANO back-end database on Postgres where two "scheduler" nodes optimize the start of the data processing sequence. Data is then introduced via a source node, marking the ingress of data into the workflow. This step is followed by transformation and quality control measures. The workflow sequence underlines a structured approach to refining the data,

ensuring it meets predefined quality and format standards before it is funneled to the final storage phase, embodied by the "store" node.

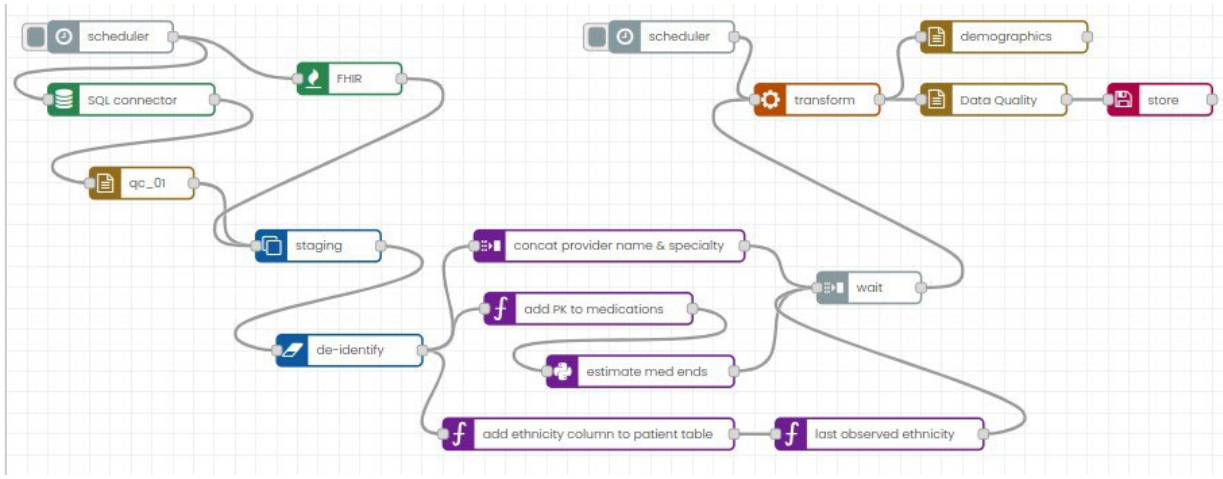

Figure 2. a screen capture of the Piano data transformation workflow.

The Piano data ingestion optimization process included automatic disaggregation of location information from source encounter tables and automating additional context added to Procedures, Observation & Measurement domains. The "care_site" node executed SQL operations that initially check for the existence of a "care_site" table within the targeted database, proceeding to recreate the table afresh populated with unique "FACILITYID" records from source Encounter tables. This ensures the singular representation of facility identifiers, streamlining data related to care sites. Furthermore, a diagnosis preprocessing node features a SQL script designed to handle the manipulation of a concatenation column within the staged DIAGNOSIS table. Here, the operation involves the addition of a new column if it's found that an existing one either doesn't exist or isn't system-generated, with this new column deriving its content through the aggregation of various data elements, signifying advanced data manipulation to aid in analytical processes.

Another preprocessing operation addresses the cleaning of data across multiple tables within a staged database. This involves the standardization of null representations to a consistent SQL NULL across various columns in all tables enhancing data quality and consistency crucial for reliable analysis. Lastly, the "changing datatypes" node describes SQL code that fundamentally alters the data types of columns across several tables. This transformation includes updating numerical data to double precision and date columns to the date data type, alongside normalizing specific columns to ensure uniformity.

As a component of the transformation process, Piano automatically added temporal dimension tables for Observations and Conditions and eliminated duplicate data resulting from qualitative and quantitative recording of the same observation. Piano aggregated data from multiple source tables into one OMOP table, and disaggregated one source table into multiple target OMOP tables, automatically depending on the main concept in each row. During the transformation the project team observed the tool execution time at 59.06 seconds consistently

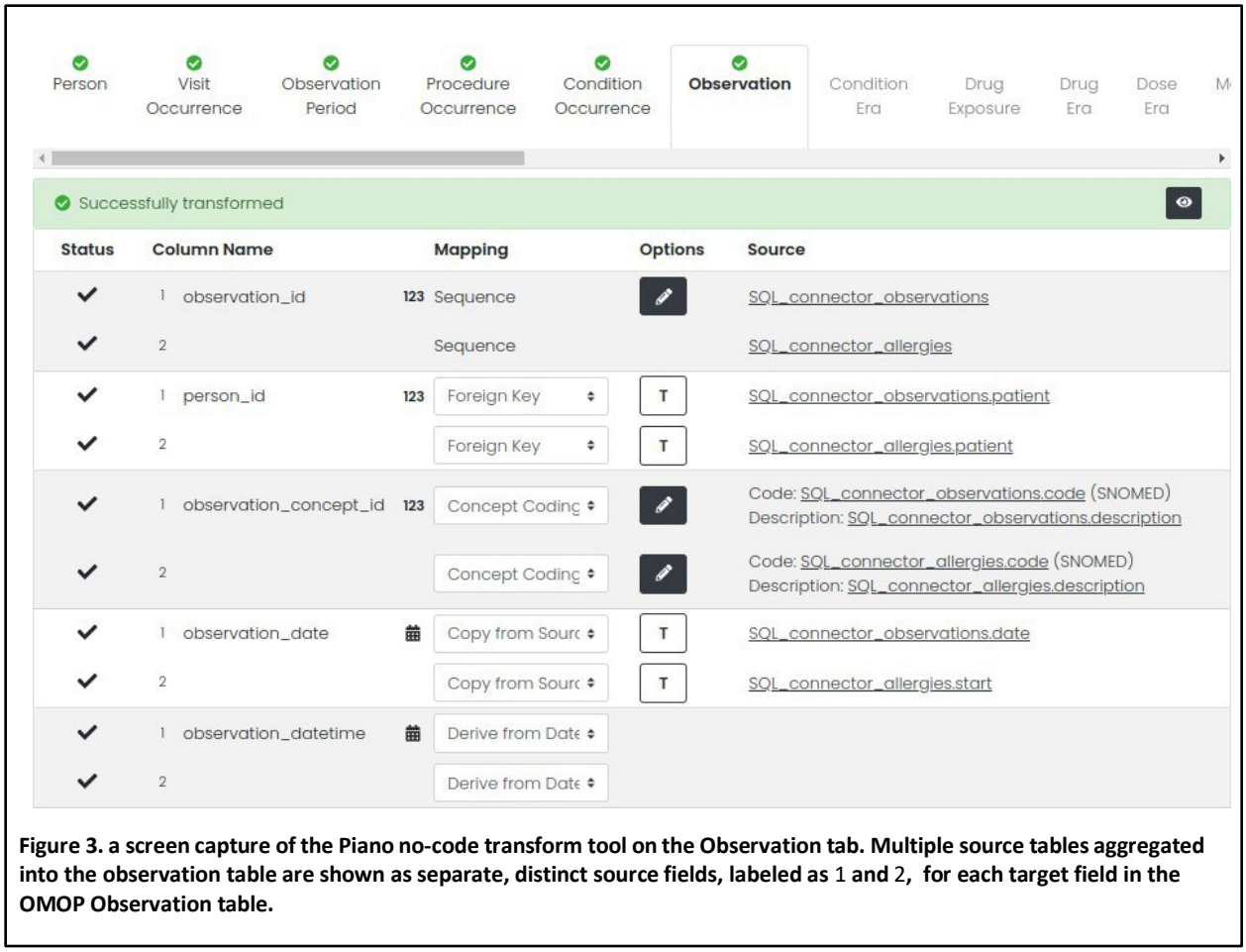

Figure 3. a screen capture of the Piano no-code transform tool on the Observation tab. Multiple source tables aggregated into the observation table are shown as separate, distinct source fields, labeled as 1 and 2, for each target field in the OMOP Observation table.

Results

The absolute resource reduction required to develop a data transformation workflow using Piano was less than 1.8% of the resources previously used for the same task without Piano (4.5 person days, compared to 252). This difference is not adjusted for a less experienced team, new to OMOP and PCOR, that used Piano compared to the baseline team. Manual verification of the resulting OMOP dataset, shows that the team provided the mappings for just 0.4% of all mappings, with the remaining 99.6% provided by the AI.

Subjective feedback from the team indicated that they felt that Piano saved a lot of time by simplifying aggregation, disaggregation and lookups that normally require significant SQL coding. They felt the simplified interfaces and automated SQL statement generation increased focus on the mapping task, and reduced hard-to-debug errors that are typical of ETL development using a language such as SQL. Additionally, the team found using the transformation tool allowed them to become proficient with the OMOP CDM while producing the mappings.

Conclusion

In light of these results, the data transformation team strongly recommended Alliance Chicago adopt the Piano platform to transform CAPriCORN data onto OMOP. Further, they recommended Alliance Chicago investigate transformation of the source data directly into OMOP, which holds the potential to greatly reduce data transformation resources by using the OMOP data to create the CAPriCORN data sets.

References

1. AllianceChicago. AllianceChicago; 14 Sep 2023 [cited 10 Jun 2024]. Available: https://alliancechicago.org/

2. CAPriCORN - research network - the chicagoland area. In: CAPriCORN [Internet]. 21 Aug 2019 [cited 24 May 2024]. Available: https://www.capricorncdrn.org/about/

3. All of Us Research Program. In: All of Us Home [Internet]. [cited 10 Jun 2024]. Available: https://allofus.nnlm.gov/

4. The National Patient-Centered Clinical Research Network. In: The National Patient-Centered Clinical Research Network [Internet]. 16 Oct 2020 [cited 10 Jun 2024]. Available: https://pcornet.org/

5. Observational Health Data Sciences, Informatics. The book of OHDSI. 11 Jan 2021 [cited 10 Jun 2024]. Available: https://ohdsi.github.io/TheBookOfOhdsi/

6. Tsafnat G. Transforming Clinical Data .More Accurately than Humans - Piano's Automapper. 2019.

7. Using Centricity Electronic Medical Record Meaningful Use Reports. GE Healthcare; 2013. Available: https://centricityusers.com/wp-content/forum-file-uploads/davidshower/Using-Centricity-Electronic-Medical-Record-Meaningful-Use-and-Quality-Reports.pdf