Leveraging Observational Medical Outcomes Partnership (OMOP) Data to Populate Disease Registries

James P. McGlothlin
RSM US LLP
Dallas, TX USA
jamie.mcglothlin@rsmus.com

Timothy Martens
The Heart Center
Cohen Children's Medical Center
Queens, NY
tmartens1@northwell.edu

Abstract— **Health care disease registries and procedural registries serve a vital purpose in support of research and patient quality. However, it requires a significant level of clinician effort to collect and submit the data required by each registry, and there are over 1000 common patient registries. In previous research, we have evaluated using supervised learning in conjunction with generative artificial intelligence to generate accurate content for disease registries. However, one of the largest challenges was to extract complete and meaningful data from the electronic medical record in a format that enabled the generative Artificial Intelligence (AI) tools. Standards like HL7 and Fast Healthcare Interoperability Resources (FHIR) were insufficient and burdensome. In this project, we propose using the new Observational Medical Outcomes Partnership (OMOP) data standard to acquire this data. Our Electronic Medical Record (EMR) software provides access to this data in the cloud without requiring extraction and transformation. The goal of this project is to utilize this data and technology to improve the population of disease registry records.**

Keywords- population health; OMOP; congenital heart disease; thoracic surgery; artificial intelligence.

I. INTRODUCTION

Patient registries are essential tools for improving healthcare quality, supporting clinical research, and ensuring patient safety. In the U.S. alone, there are over 1,000 active registries tracking patient outcomes and provider performance [1]. A mid-sized pediatric hospital in the U.S. identified 29 registries in which it actively participates, requiring more than 45,000 staff hours annually—including over 3,000 hours of physician time—for data abstraction. This reflects a significant investment of skilled clinical labor.

Despite the high resource demand, opting out of registry participation is not feasible. Registries are not only crucial for advancing research and public health, but they also influence financial incentives. Many registries contribute to provider and hospital performance ratings that impact reimbursement models, such as the Merit-Based Incentive Payment System (MIPS) from Centers for Medicare & Medicaid Services (CMS) [2].

Recent advances in generative AI and Large Language Models (LLMs) offer promising opportunities to reduce this burden [3]. Research has demonstrated that this technology can generate accurate structured information from unstructured clinical text without extensive retraining [4].

The greatest challenge has been to extract the complete patient record into clinical text useful for generative AI. Furthermore, the registry fields have to be explained in detail. In our previous research [3], we leveraged the FHIR interface to extract this data. However, we found that this required significant configuration and custom development, and did not provide the complete data view we needed to fully automate populating the disease registries.

In this proposal, we address this challenge by leveraging the OMOP. We access this data through a cloud interface with no extraction. We assert that this solution will improve the disease registry data and reduce the manual effort.

The rest of this paper is organized as follows. Section II describes the registry and problem domain. Section III describes the OMOP standard and how we apply this protocol. In Section IV, we present conclusions.

II. REGISTRY

For this research, we have decided to limit ourselves to one specific registry. The Society of Thoracic Surgeons (STS) Congenital Heart Surgery Database (CHSD) is a comprehensive, multicenter clinical registry developed to collect and analyze data on congenital heart operations in the United States and participating international institutions [11]. Established in 1994, the STS CHSD aims to improve the quality of care and outcomes for pediatric and adult patients with congenital heart disease by facilitating evidence-based practice through benchmarking, quality improvement, and clinical research. The database captures detailed procedural, demographic, and outcomes data on nearly all types of congenital heart surgeries, including perioperative morbidity, mortality, and resource utilization. The inclusion of both preoperative risk factors and postoperative outcomes supports accurate risk stratification and performance evaluation across participating centers.

The STS CHSD plays a pivotal role in advancing pediatric cardiac surgery by enabling collaborative research, supporting public reporting, and guiding institutional quality improvement efforts. It is one of the largest and most robust congenital heart surgery registries globally, with contributions from over 120 centers performing thousands of procedures annually. Risk-adjusted outcomes reporting is facilitated through the use of standardized nomenclature and

analytic models, such as the STAT (Society of Thoracic Surgeons-European Association for Cardio-Thoracic Surgery) Mortality Categories. The registry has informed numerous peer-reviewed publications, contributing to the development of clinical guidelines and performance standards. As a cornerstone of outcomes research in congenital cardiac surgery, the STS CHSD continues to evolve with data validation enhancements, increased interoperability with electronic medical records, and integration with other congenital heart disease registries to support lifelong patient care.

The STS CHSD registry is very complex. There are more than 1000 required fields. Recent audit data (from the 2022 update) indicates that across just 11 audited centers, approximately 9,128 individual data field entries were assessed for completeness and accuracy during a single harvest period [5].

III. OMOP

The OMOP Common Data Model (CDM) is a standardized data framework developed by the Observational Health Data Sciences and Informatics (OHDSI) initiative to facilitate the systematic analysis of disparate observational health data sources. OMOP enables the transformation of heterogeneous clinical data—such as EMRs, claims data, and disease registries—into a unified format with standardized terminologies and data structures. By harmonizing disparate datasets into the OMOP CDM, researchers and institutions can conduct large-scale, reproducible analyses across diverse populations and settings [6]. The model supports a wide range of research, from drug safety surveillance to comparative effectiveness studies, while leveraging standardized vocabularies like SNOMED CT [12], RxNorm [13], and Logical Observation Identifiers Names and Codes (LOINC) [14] to ensure semantic interoperability.

Utilizing OMOP for disease registries offers significant advantages in terms of scalability, interoperability, and analytical rigor. Disease-specific registries, such as those for cardiovascular disease, diabetes, or rare conditions, can be mapped to the OMOP CDM to facilitate multi-institutional research, enable longitudinal patient tracking, and integrate with broader health data networks. This transformation allows for the use of standardized analytic tools developed by the OHDSI community, including cohort definitions, prediction models, and outcome analysis frameworks, thereby accelerating hypothesis testing and real-world evidence generation. Aligning disease registries with OMOP facilitates regulatory-grade data analysis and supports initiatives such as learning health systems and precision medicine by creating interoperable data ecosystems capable of continuous quality improvement and discovery [7].

ATLAS is an open-source tool that facilitates the design and execution of analyses on standardized, patient-level, observational data in OMOP. “ATLAS enables users to define cohorts using concepts derived from standard vocabularies such as SNOMED, ICD, and LOINC, ensuring that all participants share a common understanding of clinical events and data formats.” [8].

Epic is the predominant EHR vendor in USA [15]. In the new version of Epic, patient records are materialized in OMOP format using the Microsoft Fabric cloud environment. Leveraging this solution allows us to access the data through shortcuts without creating additional API calls, extractions or transformations [9][10]. This enables us to apply generative AI tools directly to the OMOP data in order to generate the information to populate the registry.

IV. CONCLUSIONS

Historically, disease registries have been populated through manual chart abstraction. Recent advances in interoperability and generative AI have narrowed the gap towards automated record keeping but have still fallen short. In this project, we propose utilizing OMOP to complete the data flow and enable full automation.

REFERENCES

[1] R. Gliklich, N. Dreyer, and M. Leavy, eds. "Registries for evaluating patient outcomes: a user’s guide.", 2014.

[2] S. Blumenthal, "The Use of Clinical Registries in the United States: A Landscape Survey,” EGEMS, 2017, p. 26.

[3] J. McGlothlin and T. Martens, “Using Artificial Intelligence and Large Language Models to Reduce the Burden of Registry Participation,” in HealthINF, 2025.

[4] A. Thirunavukarasu et al., "Large language models in medicine." Nature medicine 29.8, 2023, pp. 1930-1940.

[5] J. Jacobs et al., "Introduction to the STS National Database Series: outcomes analysis, quality improvement, and patient safety," The Annals of thoracic surgery 100.6, 2015, pp. 1992-2000.

[6] I. Reinecke et al., "The usage of OHDSI OMOP–a scoping review," German Medical Data Sciences 2021: Digital Medicine: Recognize– Understand–Heal, 2021, pp. 95-103.

[7] P. Biedermann, "Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases." BMC medical research methodolog, 2021, p. 238.

[8] “Introduction to Atlas”, Albert Einstein College of Medicine, https://einsteinmed.edu/uploadedFiles/centers/ICTR/new/introduction-to-atlas-manual.pdf , 2021.

[9] D. Ghosh, Mastering Microsoft Fabric: SAASification of Analytics. Springer Nature, 2024.

[10] B. Mohapatra et al., "Data Integration, Data Export, and Analytics in Dataverse," Deep Dive into the Power Platform in the Age of Generative AI: Architectural Insights and Best Practices for Intelligent Business Solutions. Berkeley, CA: Apress, 2024, pp. 159-224.

[11] J. Jacobs, "The society of thoracic surgeons congenital heart surgery database: 2019 update on outcomes and quality." The Annals of thoracic surgery 107.3, 2019, pp. 691-704.

[12] E. Chang and J. Mostafa, "The use of SNOMED CT, 2013-2020: a literature review." Journal of the American Medical Informatics Association,2021, pp. 2017-2026.

[13] S. Liu et al., "RxNorm: prescription for electronic drug information exchange.", IT professional, 2005, pp. 17-23.

[14] C. McDonald, "LOINC, a universal standard for identifying laboratory observations: a 5-year update." Clinical chemistry, 2003, pp. 624-633.

[15] R. Johnson, "A comprehensive review of an electronic health record system soon to assume market ascendancy: EPIC." J Healthc Commun 1.4, 2016, p. 36.