

An AI Implementation Science Study to Improve Trustworthy Data in a Large Healthcare System

Benoit L. Marteau[†], Andrew Hornback[†], Shaun Q. Tan[†], Christian Lowson[§], Jason Woloff[§], May D. Wang[†]

[†] Georgia Institute of Technology, Atlanta, GA 30332

Email: benoitmarteau@gatech.edu; ahornback6@gatech.edu; yzhong307@gatech.edu; maywang@gatech.edu

[§] Shriners Hospitals for Children, Tampa, FL, USA

Email: christian.lowson@shrinenet.org; jason.woloff@shrinenet.org

Abstract—The rapid growth of Artificial Intelligence (AI) in healthcare has sparked interest in Trustworthy AI and AI Implementation Science, both of which are essential for accelerating clinical adoption. Yet, barriers such as strict regulations, gaps between research and clinical settings, and challenges in evaluating AI systems hinder real-world implementation. This study presents an AI Implementation case study within Shriners Children’s (SC), a large multisite pediatric system, showcasing the modernization of SC’s Research Data Warehouse (RDW) to OMOP CDM v5.4 within a secure Microsoft Fabric environment. We introduce a Python-based data quality assessment tool compatible with SC’s infrastructure, an extension of OHDSI’s R/Java-based Data Quality Dashboard (DQD) that integrates Trustworthy AI principles using the METRIC framework. This extension enhances data quality evaluation by addressing informative missingness, redundancy, timeliness, and distributional consistency. We also compare systematic and case-specific AI implementation strategies for Craniofacial Microsomia (CFM) using the FHIR standard. Our contributions include a real-world evaluation of AI implementations, integration of Trustworthy AI in data quality assessment, and evidence-based insights into hybrid implementation strategies, highlighting the need to blend systematic infrastructure with use-case-driven approaches to advance AI in healthcare.

Index Terms—health informatics, FHIR, OMOP-CDM, data standard, data harmonization, data quality, trustworthy AI, AI implementation science

I. INTRODUCTION

Artificial Intelligence (AI) has made significant progress over the past decade, driven by advancements in technology and adoption. In healthcare, this has led to a growing emphasis on Trustworthy AI (TAI), aimed at mitigating uncertainty and improving data and model transparency, as well as AI Implementation Science, which identifies barriers and practical solutions to AI adoption in clinical settings.

However, implementing AI in healthcare remains challenging due to strict data privacy regulations and the multimodal nature of patient data, including time series, monitoring, imaging, genomics, and structured or unstructured Electronic Health Records (EHRs) [1]–[3]. These challenges are amplified in large multisite healthcare systems. Current research addresses these issues by improving model generalizability through data standardization, federated learning, or foundation models [4]–[8].

Given the choice between working on AI model implementation or data quality assessment and improvement, we prioritized building the foundations for high-quality data, recognizing that model performance depends on the quality of the input data. To support our research, we adopted two standards, the Fast Healthcare Interoperability Resources (FHIR) standard for data harmonization and the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) for data standardization [9]. We identify several critical potential weaknesses in current research that hinder the implementation and widespread adoption of AI in healthcare. There is a gap in AI research conducted in controlled environments with curated datasets, and real-world implementation [10]. Most existing frameworks prioritize model evaluation over implementation or data quality improvement, overlooking the nuanced challenges and opportunities involved in deploying AI in clinical practice [11]. The challenges lie in the multidimensionality of data quality evaluation, encompassing the adherence to technical standards, the fidelity of data in representing real-world phenomena, and the usefulness of data for both AI models and users. While tools like Observational Health Data Sciences and Informatics (OHDSI)’s Data Quality Dashboard (DQD) support structured evaluations within OMOP CDM, broader frameworks such as Measurement Process, Timeliness, Representativeness, Informativeness, Consistency (METRIC) offer more comprehensive but abstract guidance, leaving implementation details to users [12], [13]. Although current AI Implementation Science emphasizes systematic, generalizable frameworks, these often fall short when applied to specific healthcare use cases, which require tailored approaches.

To address these gaps, we collaborated with Shriners Children’s (SC), a large multisite pediatric healthcare system with over 22 hospitals across North America. SC provides an ideal case study due to its diverse, multimodal data and its multiple specialties, including craniofacial disorders, burns, and orthopedics, offering unique case studies. Our key contributions are as follows:

- We provide Real-World Evidence (RWE) of data infrastructure standardization and modernization from an AI Implementation Science study in a real-world, multisite, and multimodal healthcare system.

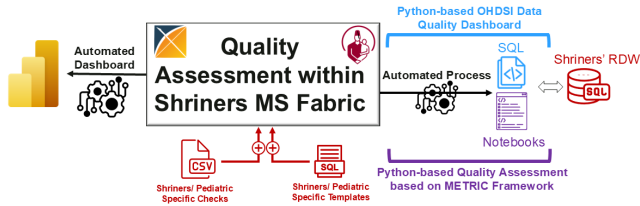


Fig. 1. Overview of our adaptation and implementation of OHDSI Data Quality Dashboard (DQD) within Shriners Children's MS Fabric environment.

- We extend the OHDSI DQD standard data quality evaluation to include TAI approaches and the METRIC framework.
- We provide evidence-based details and insights into the differences and similarities between systematic and case study-specific implementations.

II. BACKGROUND

A. SC Data Infrastructure

SC established the Shriners Health Outcomes Network (SHONet) to build a Research Data Warehouse (RDW) following the OMOP CDM. The SHONet initiative began as a means to leverage SC's data, thereby enhancing SC's clinical efficacy studies and enabling its clinicians to conduct comprehensive patient cohort analyses. SC utilizes standardized Extract-Transfer-Load (ETL) processes to map its data from its Cerner Millennium and newer Epic System EHR systems. SC RDW is currently housed in a secure Microsoft (MS) Azure environment with more than 240 Gigabytes (GB) of data and billions of data points, accessible only via Azure Virtual Machines (VMs) controlled via a Role-Based Access Control (RBAC) mechanism.

Recently, SC adopted MS Fabric, enabling researchers to access real-world data within a secure environment that complies with the Health Insurance Portability and Accountability Act (HIPAA) [14]. MS Fabric integrates various data services, such as Lakehouse data storage that leverages Spark Data Frames for large-scale data processing. The MS Fabric notebook provides an interactive coding environment to enable data engineers and scientists to develop programs and AI models with direct access to the data. For this study, SC copied its RDW into an MS Fabric workspace environment, which we used to perform the various experiments and analyses.

B. OHDSI OMOP CDM

OHDSI developed the OMOP CDM data standard to enable large-scale collaboration between healthcare institutions [15]. This standard revolves around the OMOP CDM Concept Code ID, which represents various concepts encountered in medical practice, including procedures, measurements, drugs, devices, observations, and conditions. To support the OMOP CDM, OHDSI developed a standard vocabulary to harmonize the different classification and code systems that medical professionals use in their practice (e.g., International Classification of Diseases (ICD)-9, ICD-10, Systematized Nomenclature of Medicine - Clinical Terminology (SNOMED-CT), Current Procedural Terminology, 4th Edition (CPT4), etc) [16].

OHDSI also developed a tool suite to facilitate the ETL process between source EHR data and OMOP CDM databases, or to evaluate the quality of an OMOP CDM database with the DQD [12].

C. HL7 FHIR

FHIR is a modern interoperability standard developed by Health Level Seven (HL7) that enables the structured exchange of healthcare data using internet-based technologies. The standard defines a set of modular data components, called Resources, that represent common clinical concepts. These Resources, such as Patient, Observation, Medication, and Condition, can be accessed and manipulated through Representational State Transfer (RESTful) Application Programming Interfaces (API), typically in JavaScript Object Notation (JSON) or eXtensible Markup Language (XML) formats, allowing for flexible and scalable data integration across healthcare systems [9].

D. METRIC Framework

The METRIC framework published by Schwabe et al. in 2024 plays a crucial role in the development of TAI systems by providing a structured framework to assess the quality of training data, with the assumption that high-quality and reliable data is key to robust and ethical AI systems, implying that we can't have TAI without trustworthy data [13]. The authors developed this framework based on a systematic review of scientific studies to increase adoption of AI in healthcare, providing guidelines for AI scientists, engineers, end-users, and regulators. However, this framework does not provide any specific software or tools contrary to OHDSI DQD, and therefore, its implementation relies on the developer. This framework comprises five dimensions: 1) **MEasurement Process** measures the uncertainty related to data acquisition, from sensors to human-induced error and source credibility; 2) **Timeliness** ensures that the data is up to date with the latest knowledge and standard (e.g., ICD9 vs. ICD10); 3) **Representativeness** measures how well the data represents a target population; 4) **Informativeness** evaluate the amount of information represented by the data (e.g., data redundancy, data missingness); and 5) **Consistency** measures dataset consistency regarding standards.

E. Case Study: Craniofacial Microsomia

Craniofacial Microsomia (CFM) is a complex congenital condition characterized by the underdevelopment of the ear, mandible, and associated facial structures. Care and treatment often require long-term, multidisciplinary care across psychosocial, surgical, and other domains. Given the variability in phenotypic presentation and treatment trajectories, managing CFM presents significant challenges for care coordination, data standardization, and clinical decision-making. One crucial task is evaluating the impact of surgeries and CDM on patient mental health. This makes CFM an ideal use case for exploring how AI implementation science, driven by FHIR, can support personalized care planning, automate data integration across

specialties, and potentially enhance longitudinal tracking of outcomes.

III. AI IMPLEMENTATION SCIENCE

AI implementation science is an emerging field that focuses on bridging the gap between the development of AI models and their practical and ethical integration into real-world healthcare settings. Unlike traditional AI research, which often emphasizes model performance in controlled environments, it focuses on how these tools are adopted, utilized, and evaluated in complex clinical workflows. Another specificity of AI implementation science is its heavy reliance on engineering. However, unlike AI engineering, which focuses on building the tools, AI implementation science is the systematic study of how to integrate evidence-based practices/ interventions/ approaches into real-world practice to bridge the gap of the well-documented “know-do-gap” [17].

A. Modernization and Evaluation of SC RDW

SC developed its RDW around 2015, coinciding with the development of the OMOP CDM versions 5.1 and 5.2. Due to the complexity associated with developing new ETL pipelines, SC RDW still adheres to the same version, thereby limiting its potential use for collaborative research or the implementation of peer-reviewed tools [18]–[20]. Moreover, SC personalized some tables to fit their needs, specifically with data related to the ETL process version, care site specialty, a table specific to Patient-Reported Outcome Measurements (PROM) Observations, and pain mitigation medications.

To modernize SC RDW to the latest OMOP CDM version 5.4, we started by mapping SC’s RDW tables and columns to version 5.4. We then identified the relevant functions and logic of the DQD developed by OHDSI, which includes the generation of SQL scripts based on templates. We then selected Python as the programming language as it is natively implemented and supported by multiple environments (such as MS Fabric or Databricks). We then implemented our version of the DQD within MS Fabric, validating the SQL scripts generated by our implementation with the SQL scripts generated by the original DQD by OHDSI. We ultimately evaluated the quality of SC RDW before and after modernization. Essentially, we tried to answer the following Research Questions (RQs):

- **RQ1:** How does modernizing SC OMOP CDM databases influence its data quality assessment using OHDSI DQD?

1) AI Implementation Challenges and Opportunities:

OHDSI developed its tools using the R and Java programming languages, both of which are technically supported by MS Fabric. However, these tools require a specific version of Java that is incompatible with MS Fabric, presenting an implementation barrier. Previous research either had the option to create a whole infrastructure compatible with this version of Java or manually run the tools on the database [20], [21]. These tools usually comprise three parts: a part that uses templates to generate SQL scripts, a part that interacts with the database, and a part that acts as a dashboard with the use of web applications. We propose an alternative solution

to the current OHDSI R-based implementation: converting the first part (SQL generation) to Python to be used as a package, while leaving the second and third parts as use-case dependent implementations. Specifically, we allow the user to change the interaction mechanism and dashboard visualization easily. This choice is based on the dependence between the interaction mechanism and dashboard visualization on the environment. For example, MS Fabric has specific APIs to interact with the databases, and we used Power BI to create our dashboard. We acknowledge that this still leaves some implementation to the user; however, this should be mitigated over time as more researchers implement different interaction mechanisms and dashboard visualizations across different environments and share them. We provide an overview of our implementation and conversion of OHDSI’s DQD in **Fig. 1**. We validated our implementation by comparing the SQL scripts generated and results obtained with the original OHDSI DQD. This iterative process enabled us to identify and fix code and logic mistakes in our implementation. Our current implementation yields the same SQL scripts and results as the R-based DQD. We then integrated our converted code into an automated pipeline within SC MS Fabric and created an interactive dashboard that regularly monitors SC RDW quality over time.

B. TAI-based Implementation Evaluation

We extended OHDSI’s DQD with approaches inspired by TAI research, such as the METRIC framework. We believe that such a framework proposes a natural extension of OHDSI DQD, with additional evaluation dimensions. The primary challenge lies in applying these abstract evaluation concepts to real-world data. This is why in this study, we selected and implemented four quality assessments based on evaluation dimensions from the METRIC framework. Specifically, we selected Informative Missingness, Timeliness, and Distribution Consistency, as we could translate these evaluation concepts into specific research questions:

- **RQ2 (Informative Missingness):** Are missing data presenting a specific pattern based on their type (e.g., procedure, condition) or hospital site?
- **RQ3 (Timeliness):** Are all unique source data mapped to the same concepts (e.g., are mappings of the same concept different based on the version of the code (ICD9 vs. ICD10))?
- **RQ4 (Distribution Consistency):** Is data distribution uniform across the different hospital sites?

We think that not all the assessments proposed by the METRIC framework can be applied to a systematic implementation evaluation. For example, evaluating Informativeness and Representativeness partially requires expertise and is case study dependent (e.g., what is the target population). The measurement process also requires manual expert investigation, notably to understand the potential causality chain that led to a device or human-induced error. That being said, some of these limitations to systematic evaluations can be overcome by using multiple data sources as a reference. For example, the application of Natural Language Processing (NLP) techniques

to clinical notes may help mitigate human-induced errors and enhance credibility by providing multiple sources, as demonstrated by previous studies [22].

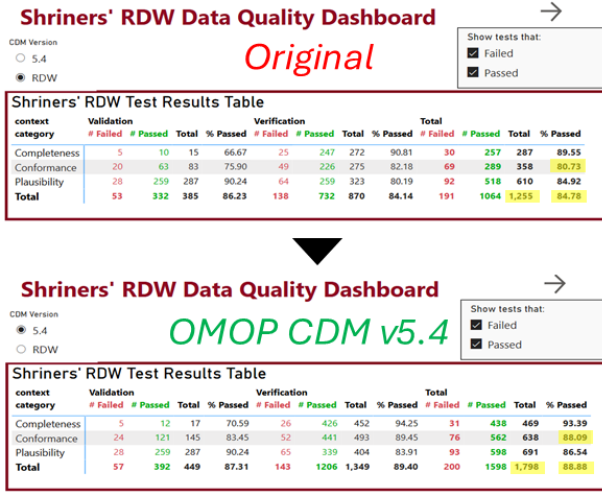


Fig. 2. We modernized Shriners Children’s Research Data Warehouse OMOP CDM database, and observed an improvement of the OHDSI DQD assessment, with more tests performed post-modernization.

C. Systematic vs. Case Study

To evaluate the benefits of our systematic approaches in a specific case study, we chose to work on a case study on CFM. For this case study, we had access to only a subset of SC’s RDW, which limited our ability to perform all the analyses that could be conducted using systematic approaches, notably regarding multisite analysis. However, this use case enabled us to assess the usefulness of the data, using AI models to analyze the impact of the patients’ diagnoses and procedures on their mental health. This means that for our case study, we first had to define a cohort definition to retrieve the a priori relevant data with the help of clinicians. Moreover, clinicians were heavily involved in the data quality assessment and pre-processing, notably in our case study, providing important insights on how codes were generated. Specifically, the fact that multiple procedure codes are generated for a single surgery, the change of code vocabulary between ICD-9 and ICD-10 in 2015, and the evaluation of the relevance of the procedure and psychiatric diagnosis to the case study.

We also leveraged this case-study to evaluate the impact of data harmonization on AI performance. The AI model was trained using patients’ diagnoses to classify whether they had a psychiatric-related diagnosis or not. We collaborated with clinicians to generate the list of relevant psychiatric diagnoses. We then identified the diagnoses and procedure codes of every patient, and created one-hot encoding as our input features. This means that we have as many features as different diagnoses and procedure codes, using “1” representing a code linked with the patient. We repeated the operation using the source codes (e.g., ICD, SNOMED, CPT4, ...) and then using the harmonized OMOP CDM concept codes. We present the results of the AI performance using either source or OMOP CDM concept codes in the next section.

This case-study enabled us to answer the following questions:

- **RQ5:** Does Data Harmonization impact AI model performance? Due to the use of multiple similar vocabulary (e.g., ICD9 and ICD10), using the OMOP CDM concept codes to harmonize the data reduces the effective number of unique source codes required to represent the data (which means that two source codes in different vocabulary represent the same concept, and therefore are mapped to the same OMOP CDM concept code).
- **RQ6:** Does reducing the number of concept code to represent a data improves AI model performance (using OMOP CDM concept relationship by grouping OMOP CDM concept codes together in supersets)?

Ultimately, we focused on FHIR for the case study, as we believe this standard to be more appropriate for the concrete adoption of AI, not only with EHR data, but with multimodal data. Indeed, although the OMOP CDM is more suitable for storing observational EHR data, FHIR enables the creation of interactive and user-friendly web applications that can directly interact with the database, rather than merely visualizing it (e.g., through dashboards). To create the FHIR resources, we determined the exact nature of the data and how it could be converted into FHIR for AI-based integration with FHIR applications. The study dataset integrated patient demographics, clinical conditions, and surgical procedures, all mapped to HL7® FHIR® resources. The Patient resource captured demographics, using standard extensions for race and ethnicity. The Condition resource encoded diagnoses using ICD-10 and OMOP terminologies. The Procedure resource included both standardized OMOP codes for analytics and local source codes as custom URIs to ensure data fidelity. All Condition and Procedure records were linked to the corresponding Patient resource via the subject field.

IV. RESULTS

A. Improving SC Data Infrastructure

Our implementation of the OMOP CDM v5.4 was partially successful, as we were able to map SC’s existing RDW to OMOP CDM v5.4. However, numerous data points are missing, notably because the OMOP CDM v5.4 is more comprehensive than previous versions, resulting in some fields being left blank.

B. Implementation Evaluation Analysis

We observed that the modernization had a positive impact according to the DQD, with a general quality test success rate improvement of 4% (from 84.78% to 88.88%), and 8% conformance improvement (from 80.73% to 88.09%), validating our **RQ1**. However, we were not able to achieve 100% conformance, notably because multiple data points were in the wrong table/ category (e.g., an observation in the procedure table), meaning that further investigation is required. We also developed our dashboard using Power BI, showing pre- and post-modernization DQD results in a user-friendly and accessible interactive application, which we show **Fig. 2**.

We implemented several dimensions of the METRIC framework evaluation. We present the specific approaches and results we obtained for the different RQs:

For **RQ2 (Informative Missingness)**, we assessed data completeness across different hospital sites, calculated for each column as the proportion of non-null, non-zero entries. We then compared these metrics across sites and data types. The divergence in data completeness between sources, illustrated in **Fig. 3**, suggests that the data are not missing completely at random and that missingness is dependent on the data source.

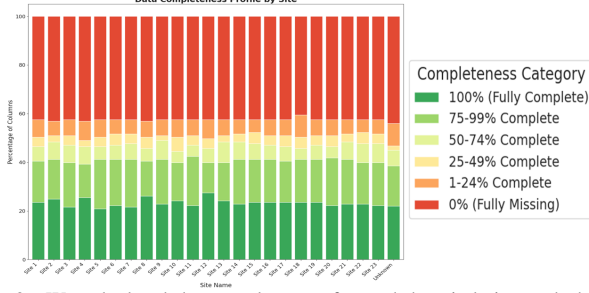


Fig. 3. We calculated the completeness for each hospital site, and observed a slight difference in completeness.

For **RQ3 (Timeliness)**, we compared the overlap between ICD-9 and ICD-10 procedure codes. We found that only half of the ICD-9 codes shared a common mapping with an ICD-10 code (**Table I**). This limited overlap suggests that AI models are at risk of performance degradation when encountering data with different distributions of these coding systems. While ICD-10 codes were more prevalent, likely due to their greater comprehensiveness [23], the poor mapping indicates a broader potential risk that may affect other clinical vocabularies and warrants further investigation.

# unique concept mapped from	Both ICD9 and ICD10	ICD9 Only	ICD10 Only
	7,125	7,754	205,329

TABLE I

OVERLAPPING BETWEEN ICD9-ICD10 AND OMOP CDM CODES

For **RQ4 (Distribution Consistency)**, we analyzed the distribution of different data types (e.g., procedures, conditions) across hospital sites. **Fig. 4** shows that the data distributions varied between sites, likely reflecting different clinical specializations. However, consistent cross-site patterns were observed: observations (e.g., vital signs) were uniformly the most prevalent data type, while clinical notes were the least. We propose several hypotheses to explain this: (1) not all notes are entered into the EHR or may be omitted during the ETL process; (2) clinicians may input multiple billing codes for procedures to ensure accuracy, whereas notes may not be duplicated in the same way; and (3) a single note often summarizes multiple observations, procedures, or conditions.

C. Case Study Specificity

We identified key challenges and implementation barriers inherent to our use case.

1) **Data Retrieval:** We retrieved patient data using CFM-specific ICD-9 and ICD-10 codes. However, we do not expect that a systematic OMOP CDM-based data quality assessment would have benefited data retrieval, as clinicians are familiar

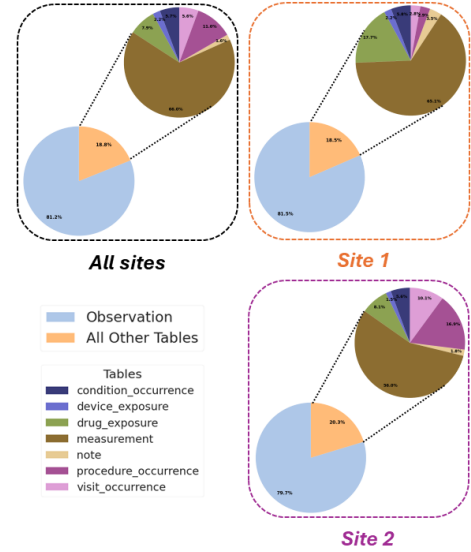


Fig. 4. We represented the distribution of the different data for different data sources (hospital sites). We can observe that the distribution differs from one site to another.

with ICD-9 and ICD-10 codes, but not necessarily with the OMOP CDM concept code. To nuance this, the OMOP CDM provides a centralized repository containing data from all hospital sites, which facilitates the retrieval of all CFM patients across all SC systems. Moreover, the implementation of an automated cohort discovery tool (such as OHDSI Atlas) might greatly benefit data retrieval [24].

2) **Data Fidelity and Usefulness:** We observed that data fidelity and usefulness were more critical than their compliance with the OMOP CDM. For example, we identified limitations for systematic approaches regarding the assessment of data fidelity and usefulness for AI models. We think that a good evaluation of data usefulness would be represented by the upper bound of any AI model performance, meaning that it is case-study dependent.

3) **CFM Case-Study and FHIR Implementation:** The CFM study enabled us to assess the impact of data harmonization on AI model performance. We used the Area Under the Receiver Operating Characteristic (AUROC) curve to evaluate our model performance, with the label being the patient's presence or absence of psychiatric diagnosis. We represent the results for 3 AI models: RandomForest (RF), eXtreme Gradient Boosting (XGBoost), and Adaptive Boosting (AdaBoost) [25]–[27], using 5-fold Cross-Validation (CV). We observed two phenomena based on **Fig. 5**: (1) data harmonization do not significantly impact model performance with a mean AUROC of 71.3% using source medical codes, and 70.0% using OMOP CDM codes, validating **RQ5**, and (2) reducing the number of OMOP CDM concept code features decreased our AI model performance, suggesting that although making the task "easier" (by reducing the number of features), we also lose granularity in the data. Further investigation is required as we did not validate **RQ6**, notably since we applied a brute-force approach in our AI model implementation. These results, however, are encouraging, as they provide additional evidence that data harmonization won't negatively impact AI model

performance, while increasing interoperability and facilitating collaboration. We still believe that better approaches could leverage the OMOP CDM relationships to increase AI performance.

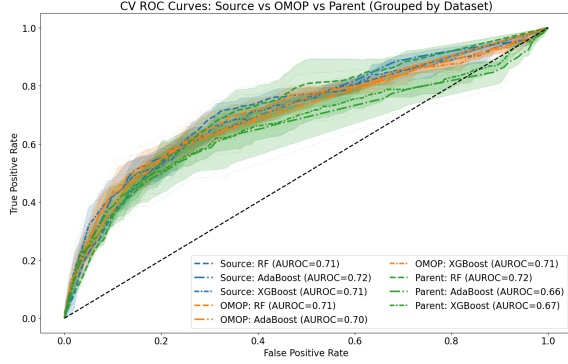


Fig. 5. We show the AUROC for different models when using source codes (blue) vs. using harmonized OMOP CDM concept code (orange) vs. using supersets of OMOP CDM concept codes (green).

The implementation of FHIR within the SC MS Fabric environment presents significant challenges. SC MS Fabric is closed to external API connections, mitigating the implementation of FHIR servers capable of supporting FHIR Resource exchange. Moreover, FHIR does not rely on OMOP CDM concept code ID, but rather on the source codes (e.g., ICD-9, ICD-10), meaning that mapping from the OMOP CDM and FHIR is not straightforward, and might depend on the data context (e.g., specific ICD-9 codes map to different OMOP CDM concept code IDs, or vice-versa). This complexity limits the ability to implement FHIR systematically. However, we believe that FHIR remains the best option for implementing AI in case studies. Despite its limited scope, this work serves as a foundation by creating specific FHIR resources that support future research and development of FHIR infrastructure compatible with MS Fabric within SC’s secure environment.

V. DISCUSSIONS

The focus on AI implementation science in healthcare has increased in the last few years, with a positive impact on AI adoption by healthcare stakeholders (e.g., patients, providers, insurers) [28]. Most importantly, this study demonstrates that implementing AI in a multisite healthcare system is not a trivial task and that systematic approaches are insufficient for adoption of AI in healthcare. We believe that data distribution drift is one of the most critical risks that can compromise AI implementation, as it can significantly reduce AI performance (e.g., sudden population migration to one site) [29]. Therefore, it is crucial to have an effective monitoring mechanism. AI implementation is a complex task, often involving all stakeholders in a highly iterative development and implementation process. Based on our effort, we identified some critical variables that impact AI implementation.

1) **Environmental:** The implementation depends on the resources available (e.g., cloud servers, computation capabilities), the existing infrastructure and workflows, which might impact usability of specific tools/ code (e.g., OHDSI DQD

and Java within MS Fabric), and the type of interaction the end-user (e.g., clinicians) will have with the system.

2) **Data Access:** Data Access plays a significant role in AI Implementation. It depends on the local regulations in place, the existing data infrastructure, as well as standardization and harmonization. For example, access to raw data or access to only a subset will drive the scope and adoption of AI (e.g., restricted access to data will limit AI implementation).

3) **Expertise Access:** As mentioned above, AI implementation is a collaborative enterprise involving many stakeholders with different expertise. However, the need from the primary end-users, such as patients and medical providers, will drive the need for implementation, while engineers and data scientists lead the technical implementation.

To contextualize our findings within medical data quality research, we benchmarked our experience against the large-scale European Health Data and Evidence Network (EHDEN) consortium [30]. This comparison yields two insights for our study. First, our experience mirrors theirs in identifying data model conformance with the highest number of data quality issues, confirming that this is a common challenge in OMOP CDM implementation. Second, we observed a data quality gap, with the overall data quality scores improvement across the mature EHDEN network being higher than in our system implementation. We argue this gap is not a limitation but rather empirical evidence for the key implementation variables discussed previously, namely: **Environmental**, **Data Access**, and **Expertise**. Several key factors can explain the discrepancy. While there are apparent differences in the maturity of the data infrastructure, inherent variations in patient cohorts, and methodological differences can explain some of the differences, as the EHDEN analysis utilized a different and smaller subset of DQD tests (with an average of 881 tests per organization in the EHDEN study vs. 1798 tests performed in ours). This highlights that while general patterns are shared, a direct comparison of quality scores across institutions can be misleading without accounting for local context.

Ultimately, an ideal AI Implementation Science framework should encompass both systematic and case study needs, with a hybrid approach to AI implementation. Moreover, as demonstrated by the METRIC framework and TAI, AI Implementation Science researchers should draw inspiration from other fields with similar problems (e.g., finance for anomaly and data drift detection, or Extended Reality (XR) implementation frameworks in healthcare). In addition to TAI, AI Implementation Science should incorporate approaches and techniques from Safe AI, Actionable AI, or Responsible AI (STAR-AI).

Future Work: There are several limitations to this study, notably the fact that we did not have access to the raw EHR data (pre-ETL process). This limits our ability to evaluate data fidelity, although we aim to apply AI-based techniques for improved data anomaly detection. We also plan to utilize clinical notes, combined with NLP techniques and AI models such as Large Language Models (LLMs), to obtain a secondary data source, thereby increasing confidence and trust in the

data by reducing its uncertainty. As a next step, we will conduct a more formal comparison with external healthcare system, and perform a dedicated usability study of our tools, using the System Usability Scale (SUS) and open-ended surveys to gather feedback from end-users, namely SC's data engineers and physicians [31]. Lastly, we plan to explore AI Implementation Science from the perspective of AI models.

VI. CONCLUSION

This study highlights the methods and importance of integrating AI implementation science principles into the deployment of TAI within SC, a large, multi-site healthcare system. As AI technologies continue to advance, their capabilities will not be judged solely on accuracy, but also on their real-world impact. The integration of AI systems into clinical workflows, their acceptance by end-users, and their governance in accordance with ethical and regulatory standards will be key factors in determining the real-world clinical impact that AI implementation science addresses. Accordingly, our findings suggest that technical performance alone is not sufficient to ensure clinical utility; instead, attention must also be paid to factors such as data interoperability, clinician engagement, workflow alignment, and transparency of model outputs.

INTERNAL REVIEW BOARD NOTE

For the systematic approach, the work was undertaken as a Quality Improvement Initiative at Shriners Hospitals for Children and, as such, was not formally supervised by an Institutional Review Board (IRB).

The CFM case study involving human subjects was conducted in accordance with the ethical standards outlined in the Belmont Report and received approval from the Georgia Institute of Technology, IRB approval number H21297.

REFERENCES

- [1] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical ai," *Nature medicine*, vol. 28, no. 9, pp. 1773–1784, 2022.
- [2] A. Bohr and K. Memarzadeh, "Current healthcare, big data, and machine learning," in *Artificial intelligence in healthcare*. Elsevier, 2020, pp. 1–24.
- [3] S. Gerke, T. Minssen, and G. Cohen, "Ethical and legal challenges of artificial intelligence-driven healthcare," in *Artificial intelligence in healthcare*. Elsevier, 2020, pp. 295–336.
- [4] R. AlSaad, A. Abd-Alrazaq, S. Boughorbel, A. Ahmed, M.-A. Renault, R. Damseh, and J. Sheikh, "Multimodal large language models in health care: applications, challenges, and future outlook," *Journal of medical Internet research*, vol. 26, p. e59505, 2024.
- [5] G. Kumar, S. Basri, A. A. Imam, S. A. Khawaja, L. F. Capretz, and A. O. Balogun, "Data harmonization for heterogeneous datasets: a systematic literature review," *Applied Sciences*, vol. 11, no. 17, p. 8275, 2021.
- [6] O. Y. Chén and B. Roberts, "Personalized health care and public health in the digital age," *Frontiers in digital health*, vol. 3, p. 595704, 2021.
- [7] R. S. Antunes, C. André da Costa, A. Küderle, I. A. Yari, and B. Eskofier, "Federated learning for healthcare: Systematic review and architecture proposal," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 4, pp. 1–23, 2022.
- [8] Y. He, F. Huang, X. Jiang, Y. Nie, M. Wang, J. Wang, and H. Chen, "Foundation model for advancing healthcare: challenges, opportunities and future directions," *IEEE Reviews in Biomedical Engineering*, 2024.
- [9] FHIR. [Online]. Available: <https://fhir.org/>
- [10] R. A. El Arab, M. S. Abu-Mahfouz, F. H. Abuadas, H. Alzghoul, M. Almari, A. Ghannam, and M. M. Seweid, "Bridging the gap: From ai success in clinical trials to real-world healthcare implementation—a narrative review," in *Healthcare*, vol. 13, no. 7. MDPI, 2025, p. 701.
- [11] J. G. You, T. Hernandez-Boussard, M. A. Pfeffer, A. Landman, and R. G. Mishuris, "Clinical trials informed framework for real world clinical implementation and deployment of artificial intelligence applications," *NPJ Digital Medicine*, vol. 8, no. 1, p. 107, 2025.
- [12] M. G. Kahn, T. J. Callahan, J. Barnard, A. E. Bauck, J. Brown, B. N. Davidson, H. Estiri, C. Goerg, E. Holve, S. G. Johnson *et al.*, "A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data," *Egems*, vol. 4, no. 1, p. 1244, 2016.
- [13] D. Schwabe, K. Becker, M. Seyferth, A. Klaß, and T. Schaeffter, "The metric-framework for assessing data quality for trustworthy ai in medicine: a systematic review," *NPJ Digital Medicine*, vol. 7, no. 1, p. 203, 2024.
- [14] Microsoft. [Online]. Available: <https://www.microsoft.com/en-us/microsoft-fabric>
- [15] G. Hripesak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek *et al.*, "Observational health data sciences and informatics (ohdsi): opportunities for observational researchers," *Studies in health technology and informatics*, vol. 216, p. 574, 2015.
- [16] OHDSI. [Online]. Available: <https://athena.ohdsi.org/>
- [17] C. M. Wray and C. D. Jones, "Bridging the know-do gap in hospital care transitions," *JAMA Internal Medicine*, vol. 183, no. 5, pp. 424–425, 2023.
- [18] N. Koscielniak, D. Jenkins, S. Hassani, C. Buckon, J. S. Tucker, S. Sienko, and C. A. Tucker, "The shonet learning health system: infrastructure for continuous learning in pediatric rehabilitation," *Learning Health Systems*, vol. 6, no. 3, p. e10305, 2022.
- [19] I. Reinecke, M. Zoch, C. Reich, M. Sedlmayr, and F. Bathelt, "The usage of ohdsi omop—a scoping review," *German Medical Data Sciences 2021: Digital Medicine: Recognize–Understand–Heal*, pp. 95–103, 2021.
- [20] B. L. Marteau, A. Hornback, Y. Zhong, C. Lowson, J. Woloff, B. M. Smith, C. Hilton, and M. D. Wang, "Improving a large healthcare system research data warehouse using ohdsi's data quality dashboard," in *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2024, pp. 1–8.
- [21] D. M. Lima, J. F. Rodrigues-Jr, A. J. Traina, F. A. Pires, and M. A. Gutierrez, "Transforming two decades of epr data to omop cdm for clinical research," in *MEDINFO 2019: Health and Wellbeing e-Networks for All*. IOS Press, 2019, pp. 233–237.
- [22] J. Wang, N. Abu-el Rub, J. Gray, H. A. Pham, Y. Zhou, F. J. Manion, M. Liu, X. Song, H. Xu, M. Rouhizadeh *et al.*, "Covid-19 signsym: a fast adaptation of a general clinical nlp tool to identify and normalize covid-19 signs and symptoms to omop common data model," *Journal of the American Medical Informatics Association*, vol. 28, no. 6, pp. 1275–1283, 2021.
- [23] M. Topaz, L. Shafran-Topaz, and K. H. Bowles, "Icd-9 to icd-10: evolution, revolution, and current debates in the united states," *Perspectives in Health Information Management/AHIMA, American Health Information Management Association*, vol. 10, no. Spring, p. 1d, 2013.
- [24] OHDSI. [Online]. Available: <https://atlas-demo.ohdsi.org/>
- [25] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [26] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [27] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771–780, p. 1612, 1999.
- [28] C. A. Longhurst, K. Singh, A. Chopra, A. Atreja, and J. S. Brownstein, "A call for artificial intelligence implementation science centers to evaluate clinical effectiveness," p. A1p2400223, 2024.
- [29] B. Sahiner, W. Chen, R. K. Samala, and N. Petrick, "Data drift in medical machine learning: implications and potential remedies," *The British Journal of Radiology*, vol. 96, no. 1150, p. 20220878, 2023.
- [30] C. Blacketer, E. A. Voss, F. DeFalco, N. Hughes, M. J. Schuemie, M. Moinat, and P. R. Rijnbeek, "Using the data quality dashboard to improve the ehden network," *Applied Sciences*, vol. 11, no. 24, p. 11920, 2021.
- [31] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.