Design of Feedback Reports for Evaluating Data Fitness for Use in the Bridge2AI For Clinical Care Research Consortium

Presenter: Jared Houghtaling

Intro:

The Bridge2AI for Clinical Care (B2AI For CC) research consortium consists of fifteen data contributing sites (DCS) led by various researcher teams based on expertise and objectives. Two of these teams, namely Standards and Data Acquisition, are responsible for providing guidance to DCS in their efforts to generate interoperable, multimodal data extracts. One critical tool in this guidance effort has been the creation and dissemination of formal reports that characterize submitted data extracts with respect to: (1) metadata (i.e. number of patients, size of files, etc.) about the delivery , (2) protected health information (PHI) and IRB compliance, (3) extent of cohort capture using validated definitions from the OHDSI PhenotypeLibrary, (4) standard data quality and characterization checks (i.e. DQD[1] & Achilles[2]), and (5) fitness for use in the B2AI For CC consortium and comparisons with other data contributing sites. These reports serve both a prospective role as detailed instructions for sites to iteratively update their data extracts to suit the needs and requirements of the consortium, and a retrospective role to detail a history of the prior extracts those sites have delivered and the associated feedback they have received.

Methods:

a. File Delivery Details

Data contributing sites organized files (e.g. tabular, waveform, and imaging) in structured delivery packets; the AzureStor[4] package in R parsed these packets and retrieved metadata.

b. Checks for PHI

We executed a per-mode PHI scan using tooling developed within the B2AI For CC consortium,[5] and included details from these scans like pass/fail percentages and redacted contextual examples, as their own report section.

c. Capture of Validated Phenotypes

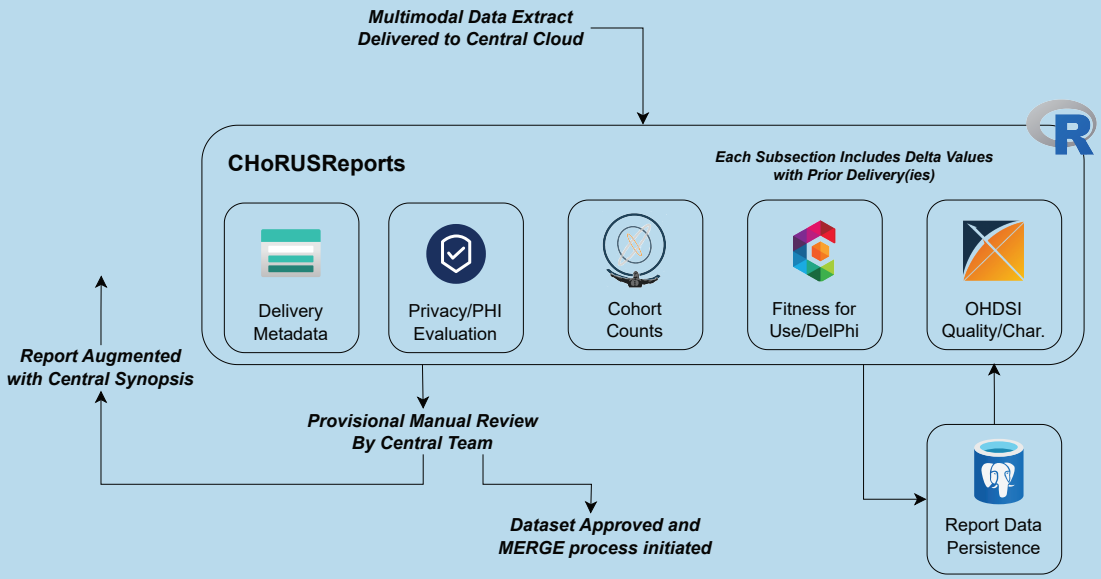
We used WebAPI to programmatically ingest validated cohort definitions from the OHDSI PhenotypeLibrary, and then generated each of those cohorts on the resulting ingested data.

d. Fitness for Use in B2AI For CC

We included data characterization and data quality checks specific to the B2AI For CC consortium; these additional queries incorporate elements from a prioritization process[6] based on critical-care settings.

e. Postgres-Based Report Data Persistence

We stored R objects and report elements as large objects in the central database to calculate and persist changes in data content over time.



Feedback reports have helped to iteratively improve data contributions across this multi-partner consortium; ***emphasis on captured phenotypes as a metric for fitness for use*** has provided valuable context for granular quality and characterization details, and has accelerated targeted updates at sites.

Example Report Table Created From Data Packet Scan

category	file_count	person_count	person_delta	target_percent
ALL Data Modes	12,345	7,864	1,184	78.6
ANY Data Modes	21,546	9,423	2,000	94.3
OMOP Data	13	9,423	2,000	94.3
Imaging Data	17,583	8,654	1,578	86.5
Waveform Data	19,817	9,068	1,651	90.7
Note Data	2	9,253	1,889	92.5
Waveform Data	18,754	7,443	1,517	74.4
Note Data	14,576	6,846	1,254	68.5

Check out the full conference proceeding here:

References:
[1] https://www.github.com/OHDSI/DataQualityDashboard
[2] https://www.github.com/OHDSI/Achilles
[3] https://www.github.com/EHDEN/CDMInspection
[4] https://www.github.com/Azure/AzureStor
[5]Ohno-Machado L et al. Biomed Inform. 2023
[6]Talapova P, et al. 2023 OHDSI Global Symposium. 2023
[7] https://www.github.com/chorus-ai/CHoRUSReports

Results & Discussion:

While the B2AI For CC consortium is still in nascent stages with regard to multimodal data standardization, ingestion, and consolidation, the feedback reports we present here have helped to accelerate iterative data contribution processes and to improve overall data quality. In contrast with the EHDEN consortium that aims to support large-scale federated analyses in which all data remains at each site, B2AI For CC is aggregating and curating a multi-site dataset centrally; the high level of detail, consortium specificity, and cross-site comparisons in reports for B2AI For CC relative to EHDEN[3] reflect this inherent advantage of broad data accessibility in a central location. Moreover, the reports we generate leverage and augment powerful OHDSI resources like the PhenotypeLibrary in order to produce a phenotype-oriented overview of data fitness for use. These reports do not stand on their own; rather, they serve as a tool for the Standards and Data Acquisition teams to use during meetings with individual sites to review those sites' data packets and help to interpret and prioritize updates needed for subsequent deliveries.

Conclusions

- The work described here represents a first step toward a robust tool for enhancing consortium-wide data quality within B2AI For CC; we expect that the software we have developed[7] in support of this reporting functionality holds utility in the OHDSI community and other research consortia with similar objectives regarding data quality and interoperability.

- Much of the work presented here builds on the effort and dedication of so many others in the OHDSI community; we will continue to contribute to - and advocate for - open-source development of these powerful tools, and we plan to continue to share our efforts and experiences along the way.

Authors: *Jared Houghtaling[a], Gilles Clermont[b], Andrew E. Williams[a]*

[a]Tufts Medicine - Institute for Clinical Research and Health Policy Studies (ICRHPS)

[b]University of Pittsburgh - School of Medicine

TuftsMedicine
University of Pittsburgh School of Medicine
OHDSI