Integrating ATLAS Cohorts with DICOM Images and ECG Waveforms to Enrich Real-World Evidence Research

Boudewijn Aasman[1], Selvin Soby[1], Adil Ahmed[2], Shweta Garg[2], Silvie Colman[1], Chandra has Nelapatla[1], Manuel Wahle[1], Parsa Mirhaji[2]

[1] Montefiore Medicine, [2] Albert Einstein College of Medicine at Montefiore

Background

OHDSI's Atlas is a web-based open-source tool which facilitates the design and execution of analysis on standardized, patient-level, observational data. Users can build complex cohorts using structured derived from data either originating from the EHR or claims data. In a different workflow we demonstrate how ATLAS cohorts and cohort-based analytics can be integrated with a scalable NLP infrastructure to enable cohort analytics that combines structured data with unstructured clinical notes. This workflow is meant to complement the ATLAS cohort functionality by empowering researchers to use Atlas to conduct real-world evidence research, cohort-based analysis, and data driven research using a feature-rich dataset combining discrete, non-discrete, and ancillary data points such as DICOM images and ECG waveform data. This functionality natively integrates with ATLAS-NLP workflow and is extensible to accommodate different and new modalities of data in the future.

DICOM images and ECG waveform data are important datasets for real-world evidence and clinical research, but they are typically stored as derived interpretations in separate and siloed databases. It is not trivial to access, retrieve and integrate this information in optimized data models that provide a research-ready access to this information. In the OMOP common data model, various methods are used to categorize mappings and load derived findings from the EHR into specific domains. However, larger datasets like MIMIC-IV contain approximately 800,000 full ECG waveforms for nearly 160,000 unique patients. Additionally, the MICIM Chest X-ray database has approximately 380,000 images corresponding to approximately 225,000 studies that won't naturally fit the common data model and cannot be easily accessed for cohort generation via ATLAS. To enable comprehensive comparative analysis and insights, our project extends Atlas to allow researchers to extract, link with structured clinical and other phenotypic data, and upload raw data files, including all DICOM images and 12-lead ECG waveform data for the specific patient cohorts into a secured high-performance computing enclave for analysis. All data access and retrieval pipelines are linked to a near real-time IRB audit and control system that ensures regulatory compliance and protection of confidential information when records are linked for research purposes.

Methods

Outline for DICOM images

1. Identify the source of DICOM image data in PACS system metadata database, establish understanding of the workflow for image data and related DICOM metadata
2. Developed logic to match image location from metadata database to DICOM file store
3. Einstein Atlas Data Extraction workflow was extended to allow the download of DICOM images and linkage to the patients and their records within OMOP-CDM and ATLAS cohort.
4. Automation and query orchestration pipeline for extraction of DICOM metadata, raw images, extraction of clinical data baskets and uploading the linked data to secured HPC enclave are implemented using Dagster automation framework.
5. The IRB profile that approves linkage and use of DICOM images are deposited to secure HPC enclave to enable users to access the enclave and analyze the data within the scope of an approved IRB, thereby maintaining a secure pipeline end-to-end

Outline for ECG waveforms

6. Identify the source of ECG waveform data in EHR, establish understanding of the workflow for live waveform data to become finalized in patient's chart

7. Configure an interface to receive near real-time ECG waveform XML files prospectively and deposit into a specialized database
8. Set up a one-time process to download all retrospective ECG waveform XML files into specialized database, completing a longitudinal archive of ECG waveforms that prospectively updates in real-time.
9. Atlas Data Extraction workflow was extended to allows access and the download of ECG waveforms for patients in the cohort if it exists.

Results

The DICOM pipeline has successfully compiled a subset of images for 5 separate project cohorts.

The pipeline was also able to provide access to our institution's complete ECG data history for a predictive analytics project and an ongoing clinical trial by cardiology team studying impact and outcomes of using AI and predictive analytics for early recognition of cardiac arrhythmia. This implementation enables researchers to independently extract full waveforms for their patient cohorts using specialized software and visualizers that are capable of viewing and processing waveforms. It empowers researchers to utilize ECG waveforms for predictive models, aiding in near real-time diagnosis augmentation for patient conditions.

Conclusion

This project demonstrates extension of ATLAS cohort functionality to empower and enable researchers to conduct real-world evidence research, cohort-based analysis, and data driven research using rich datasets combining discrete, non-discrete, and ancillary data points such as DICOM images and ECG waveform data. This functionality natively integrates with ATLAS workflows and is extensible to accommodate different and new modalities of data in the future.

References:

1. Choi, S., Joo, H. J., Kim, Y., Kim, J.-H., & Seok, J. (2022). Conversion of Automated 12-Lead Electrocardiogram Interpretations to OMOP CDM Vocabulary. In Applied Clinical Informatics (Vol. 13, Issue 04, pp. 880–890). Georg Thieme Verlag KG. https://doi.org/10.1055/s-0042-1756427
2. Yoo, H., Yum, Y., Park, S. W., Lee, J. M., Jang, M., Kim, Y., Kim, J.-H., Park, H.-J., Han, K. S., Park, J. H., & Joo, H. J. (2023). Standardized Database of 12-Lead Electrocardiograms with a Common Standard for the Promotion of Cardiovascular Research: KURIAS-ECG. In Healthcare Informatics Research (Vol. 29, Issue 2, pp. 132–144). The Korean Society of Medical Informatics. https://doi.org/10.4258/hir.2023.29.2.132
3. Gow, B., Pollard, T., Nathanson, L. A., Johnson, A., Moody, B., Fernandes, C., Greenbaum, N., Berkowitz, S., Moukheiber, D., Eslami, P., Herbst, E., Mark, R., & Horng, S. (2022). MIMIC-IV-ECG - Diagnostic Electrocardiogram Matched Subset (version 0.1). PhysioNet.
4. Johnson, A., Pollard, T., Mark, R., Berkowitz, S., & Horng, S. (2019). MIMIC-CXR Database (version 2.0.0). *PhysioNet*. https://doi.org/10.13026/C2JT1Q.
5. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 6, 317 (2019). https://doi.org/10.1038/s41597-019-0322-0
6. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.
7. Mirhaji, P., Soby, S., Henninger, E., Nelapatla, C., Wahle, M., Aasman, B., Belin, E., Leveraging OHDSI/ATLAS and Open-Source Development to Support Translational Research, Data Science, and Regulatory Compliance. Presented at 2022 OHDSI Symposium in Baltimore, MD.