Sirius tool: Conversion of clinical study data into OMOP model and implementation of data quality monitoring of wearable sensor data

Vojtech Huser MD, PhD[1], Esteve Verdura MS[2], Michael N. Lubke[1], MS, Bhavna Adhin, MS[1]
[1] Pfizer, Inc, [2] Clarivate

Background

Optimal data representation of human clinical study data is an ongoing medical informatics challenge. The Observational Medical Outcomes Partnership (OMOP) common data model (CDM) has been used to aggregate data across multiple studies to facilitate analysis that is portable across various datasets.[1] We chose to pursue data quality assessment (DQA) against OMOP transformed data and not the original source data.

Our project focuses on digital health studies that utilize wearable sensors. The significance of digital health technologies has been growing recently.[2] Activity monitoring is probably the most advanced digital health domain.[3]

Data for wearable sensors is often received and organized into files per subject per study event. The goal of data quality assessment is to look at data file presence (all files present for all study events for all data types for all study participants) and at data file content (files adhere to rules that investigate data format, data density, value plausibility or consistency across data types).

Methods

We created a data quality assessment framework written in Python and we named it Sirius. The rationale for developing a data quality assessment platform for clinical study data was to generalize and move away from an approach where each study has a custom script. Sirius uses a modular function approach and its library of functions is extensible to cover different wearable sensor devices, different data file formats or different data quality analytical tasks. Sirius data quality rules are defined on individual study level using Yet Another Markup Language (YAML) syntax. Sirius targets a low-code approach to DQA rule authoring. Sirius execution can be automated for different intervals (e.g., daily or monthly execution) and results can be aggregated into a dashboard. The OMOP model is used to standardize disparate studies into standard event structure against which rules can be later written.

Results

Phase 1 of Sirius development took 10 months using a set of six studies with wearable sensor data. In phase 2, the library of functions was expanded and it was applied in 16 studies. Sirius uses three layers of config files. (1) *Study configuration* defines study-level metadata. For example, number of study subjects, storage locations to be monitored, or list of expected data sources. (2) *Preprocessing actions configuration* defines what data transformation should be applied to individual data sources. All actions have an input device data file (or set of files) and generate an output file (typically orders of magnitude smaller in size). Finally, (3) *Rule configuration* defines individual rules that evaluate to true (compliant) or false (data error or warning or notification). Actions and rules rely of an extensible set of modular functions. Multiple actions can be chained together to achieve in steps the necessary data transformation (output of one action becomes input for subsequent action; final action provides input for a data quality rule or for a human review).

We comment below on selected Sirius rule or action functions:
File Name Parsing: Sirius creates observation events based on parsing the file names that contain the sensor data. This function converts unstructured set of files into database of events assigned to participant and linked to timestamps (OMOP observation table events). For studies where consecutive numbering of visits is used (e.g., visit1 instead of absolute date), it assigns symbolic dates to each visit such that it can be represented in the OMOP model. For large sensor data with high frequency of data (more than one data event per minute or hour), the individual rows within sensor file are not converted into OMOP. Subsequent data quality rules then use this OMOP event data to evaluate presence of data per study protocol. An example of a rule is: five cough recording files are present per each visit per each subject.
Temporal Data Compliance: Sirius can analyze temporal patterns in device data files to detect periods of time when expected sensor data are missing (e.g., participant did not wear the sensor or battery exhausted) or have outlier values.
Device-specific custom format transformation: used for sensors using proprietary format (e.g., .bin for ActiGraph watch).

After a set of data actions and rules are finalized, Sirius use includes review of study report (Boolean rules, fully-computerized) and human-assisted review of aggregated data in files outputted by Sirius actions.

Discussion

Size, scope and data review considerations: Wearable sensors data can be extremely large due to high collection frequency. Additionally, raw sensors data may need to be processed into endpoint data (e.g., calculating sleep onset latency from actigraphy data). Case Report Form (CRF) study data or claims data are usually smaller. Established mechanisms exist for review of CRF study data. Because wearable sensor data from a single patient event can be larger than all CRF study data, different approaches to data monitoring may be employed. The file nature of device data (and not a database) represents the main difference for data review in contrast to CRF or EHR data. Similarly to raw imaging and raw genomic data, some level of data granularity stays outside the OMOP model. Researchers must set the level of granularity that is brought into the model. For example, only endpoint data could be imported or the mechanism of data pointer observation events may be used (implemented by Sirius). For example, imaging procedure billing events can be viewed as data pointer event to data in PACS. OMOP model may possibly be extended to better facilitate linkage between large source data (kept outside CDM) and CDM-captured data.
Future work: We expect more evolution of the Sirius tool (support for novel wearable sensors, more sophisticated functions for file content assessment and rule authoring improvements).

Conclusion

We developed a data quality framework for wearable sensor data that automates and improves data monitoring tasks. We also demonstrate that event-based OMOP common data model can facilitate data quality rule authoring for clinical study data. We discuss scope boundary considerations for raw and derived data for large biomedical data.

References

1. Roeder C, Sadowski K, Solovyev P, Araujo S. Clinical Trial Data Conventions for the OMOP CDM. In: OHDSI Symposium. ; 2020. https://www.ohdsi.org/2020-global-symposium-showcase-5/

2. FDA. Framework for the Use of Digital Health Technologies in Drug and Biological Product Development. Accessed May 16, 2023. https://fda.gov/digitalhealth

3. Rist C, Karlsson N, Necander S, Da Silva CA. Physical activity end-points in trials of chronic respiratory diseases: summary of evidence. ERJ Open Res. 2022;8(1):541-2021. doi:10.1183/23120541.00541-2021