OMOP Waveform Extension: A Schema for Integrating Physiological Signals and Derived Features into the OMOP CDM

Jared Houghtaling[1], Polina Talapova[1,4], Brian Gow[2], Manlik Kwong[1], Andrew J King[3], Benjamin Moody[2], Mike Kriley[3], Tom Pollard[2], Andrew E Williams[1]

[1] Tufts University School of Medicine, Boston, MA, USA

[2] Massachusetts Institute of Technology, Cambridge, MA, USA

[3] University of Pittsburgh, Pittsburgh, PA, USA

[4] SciForce, Ukraine

Background

Physiological waveforms signals - such as electrocardiograms (ECG), electroencephalograms (EEG), and arterial pressure or pulse oximetry tracings - contain rich clinical information not captured in conventional structured electronic health records, which typically store only low-resolution or intermittent time-series data. Incorporating details about these continuous signals into a standardized data model can enable more holistic analyses and improve predictive modeling.[1] Studies show that combining waveform data with traditional EHR features yields superior prognostic accuracy over either modality alone.[2,3] Despite their value, OMOP Common Data Model (CDM) version 5.4 lacks the dedicated structure to comprehensively catalog high resolution physiological waveforms, and associated annotations and metadata. Most large observational databases omit or silo waveform files due to this lack of standard integration.[4] This gap limits multi-site research that leverages physiologic signals alongside clinical observations. To address this need, the NIH Bridge2AI CHoRUS consortium developed an OMOP Waveform Extension schema. These extension tables will fully represent waveform acquisition events, associated signal metadata, derived physiological features, and storage references, enabling integration of these data into the OMOP CDM alongside traditional clinical records. These extensions will directly address informatics gaps for high level cohort discovery.

Methods

We designed the OMOP Waveform Extension through an iterative multi-institutional effort, aligning with OMOP CDM conventions for table structure, keys, and standardized vocabularies, and drawing upon prior work in extending the model for imaging data standardization.[5] The new design introduces four extension tables to support waveform integration: waveform_occurrence, waveform_registry, waveform_channel_metadata, and waveform_feature. Each table has a primary key (e.g. waveform_occurrence_id), links to existing person/visit contexts via person_id and timestamps, and uses standard concept identifiers to classify waveform modalities, channels, and feature types (Figure 1).

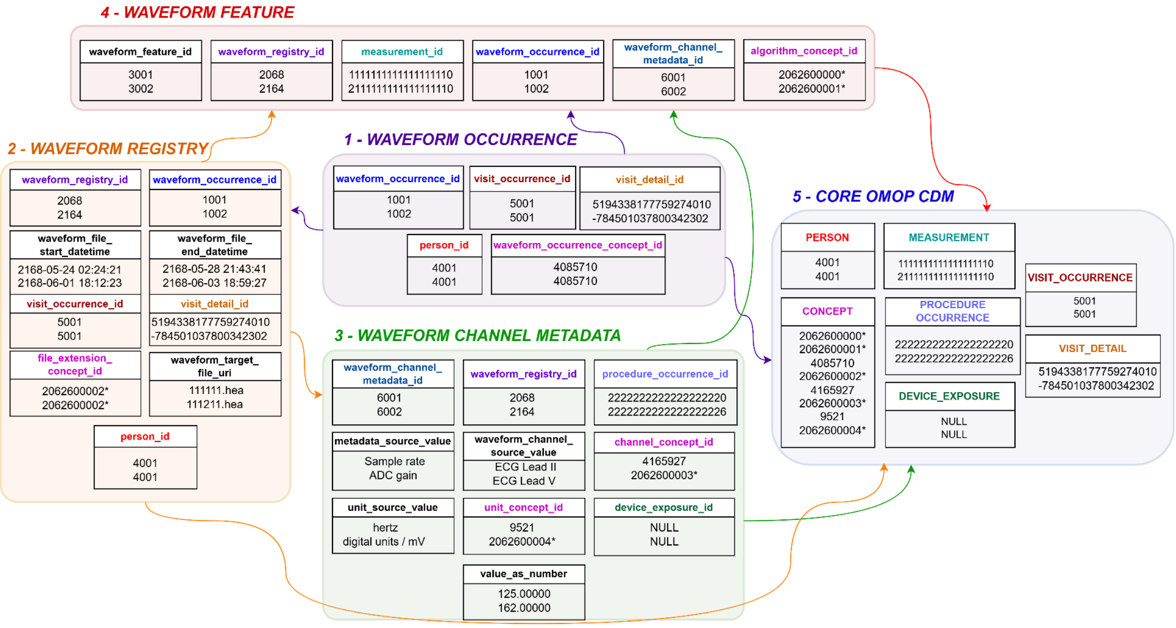

Figure 1. Example of a waveform recording within the relational structure. Arrows indicate key dependencies, while concept IDs marked with an asterisk (*) represent custom 2-billion concepts created for terms not present in the OHDSI Standardized Vocabularies. Arrow placement is schematic and may not reflect exact field-to-field connections.

Modeling decisions were informed by mapping exemplar datasets - including ICU monitor waveforms from MIMIC-IV and in-house telemetry data - into the proposed schema and refining it for completeness and consistency.

Schema Design:

- **1. waveform_occurrence:** Records each waveform recording event for a patient, including timing (start/end) and a waveform type concept (e.g. ECG vs. plethysmography). This table anchors the waveform in the clinical timeline (linking to encounters or procedures when applicable).
- **2. waveform_registry:** Indexes individual waveform files and storage references. Each entry corresponds to a waveform file or data object (with identifier, format, duration), allowing raw signals to reside outside the CDM. waveform_occurrence entries reference the registry to retrieve details about the waveform recording.
- **3. waveform_channel_metadata:** Describes individual channels within a multi-channel recording. Each record specifies a channel's type (mapped to a standard concept, e.g. "Lead II ECG" or "arterial pressure wave"), sampling rate, units, and other parameters, and links to its parent waveform_occurrence.
- **4. waveform_feature:** Contains quantitative features or measurements derived from waveforms (e.g. heart rate or QT interval from an ECG). Each feature is associated with a waveform_occurrence (and optionally a specific channel) and carries a standard concept ID for its meaning. Feature values are stored in this table (optionally linked to existing Observation or Measurement records).

Two additional tables were proposed but excluded from the initial schema. **waveform_annotation** (for time-indexed labels on signals, such as arrhythmia events) was deferred to limit complexity - such annotations can be handled via existing OMOP CDM tables. **waveform_representation** (for storing raw waveform samples in the CDM) was omitted due to volume and performance concerns. Instead, the combination of waveform_registry and waveform_occurrence allows reference to external signal files while capturing essential metadata within OMOP.

Results

We implemented the extension and evaluated it using waveform data from three sites. A preliminary implementation using selected waveform recordings from MIMIC-IV and partner hospitals demonstrated the schema's ability to represent diverse signal types (ECG, arterial pressure, photoplethysmography, respiratory waveforms) with sampling frequencies in the tens to hundreds of Hz. The schema accommodated diverse recording durations and multi-channel configurations (up to ~8 channels per recording) without information loss. Metadata in the waveform_channel_metadata table captures detailed information about the source and nature of these signals such as channel type, sampling rate, gain setting, and recording methods. The waveform_registry table links each waveform_occurrence to an external file or database entry (including large PhysioNet archives), confirming that raw signals can remain outside the CDM while metadata and links reside within it.

We also demonstrated that physiologic features extracted from waveforms can be stored and queried via this extension. For example, heart rate values computed from continuous ECG recordings were stored as waveform_feature entries with standard concept IDs, enabling them to be queried alongside conventional vital signs in the measurement table. Similarly, derived features such as QT interval or heart rate variability can be represented and linked to specific waveform occurrences and channels. This allows waveform-derived metrics to be seamlessly integrated into cohort definitions and analytic workflows alongside other OMOP measurements.

While current OHDSI tools do not yet support the extension tables natively, the schema is designed to align with OMOP conventions and can be integrated into future tool development with minimal disruption, expanding the model's analytical scope.

Conclusions

We developed and evaluated a novel OMOP CDM extension to support physiologic waveform data, filling a critical gap in the standard model. The OMOP Waveform Extension provides a semantically rich yet practical framework to integrate continuous signals (e.g. ECG, EEG, hemodynamic waveforms) with clinical data, enabling both single-site implementation and scalable multi-center research. This work enables multimodal analyses by allowing physiologic time-series features to coexist with patients' clinical profiles in OMOP. Our evaluation with MIMIC-IV and partner hospital data demonstrates the schema's flexibility and value for representing real-world waveform collections. The extension's alignment with OHDSI conventions ensures that waveform data can be leveraged with familiar tools and methods. Going forward, we will refine the model (e.g. adding needed vocabulary for novel features and exploring annotation support) and work with the OHDSI community to encourage adoption. By incorporating waveform data into OMOP, this extension paves the way for advanced multi-modal research and improved patient monitoring and outcomes.

References

1. Monfredi OJ, Moore CC, Sullivan BA, Keim-Malpass J, Fairchild KD, Loftus TJ, Bihorac A, Krahn KN, Dubrawski A, Lake DE, Moorman JR. Continuous ECG monitoring should be the heart of bedside AI-based predictive analytics monitoring for early detection of clinical deterioration. Journal of electrocardiology. 2023 Jan 1;76:35-8.
2. Mathis MR, Engoren M, Williams AM, et al. Prediction of postoperative deterioration in cardiac surgery patients using electronic health record and physiologic waveform data. Anesthesiology. 2022;137(5):621-632. DOI: 10.1097/ALN.0000000000004345.
3. Huang SC, Pareek A, Seyyedi S, et al. Fusion of medical imaging and electronic health records using deep learning: a systematic review. NPJ Digit Med. 2020;3(1):136. DOI: 10.1038/s41746-020-00341-z.
4. Johnson AEW, Bulgarelli L, Pollard TJ, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10(1):1. DOI: 10.1038/s41597-022-01899-x.
5. Park WY, Jeon K, Sippel Schmidt T, et al. Development of Medical Imaging Data Standardization for Imaging-Based Observational Research: OMOP Common Data Model Extension. J Imaging Inform Med. 2024;37(2):899-908. PMID: 38315345.