N3C/MIMIC-IV

MONITOR WAVEFORM FEATURE EXTRACTION

From: Manlik Kwong Date: 2/18/2021

To: Team

MIMIC-IV Waveform Source Data

The MIMIC-IV waveform database is organized in a folder structure where individual patients are first grouped into a root folder (Figure 1) based on the first 2 digits of its numeric identifier (ex p10, p11, p12, etc). Within each top level group folder contains subfolders representing a patient case (ex /p10/p10002348/) group. Within the patient case group folder are one or more dataset subfolders (ex /p10/ p10002348/ p10002348-21240306-0140/). Within the dataset folder there are three classes of data files. The first is the "master" header file having the same same dataset folder name and ".hea" extension (ex /p10/ p10002348/ p10002348-21240306-0140/ p10002348-21240306-0140.hea). This file is the manafest of individual data segment files contained in this folder. The second class is a CSV format that represents reported time-series vitals data (ex /p10/ p10002348/ p10002348-21240306-0140/ p10002348-21240306-0140n.csv). The third class of files represent continuous waveform data divided into one or more segments as indicated by a numeric count at the end of the filename. Each continuous waveform data segment is represented as a paired header and data files (*.hea and *.dat respectively) representing a block of data. The data segment filename consists of the root identifier followed by a underscore and zero padded 4 digit counter (ex /p10/ p10002348/ p10002348-21240306-0140/ p10002348-21240306-0140_**0001**.hea and /p10/ p10002348/ p10002348-21240306-0140/ p10002348-21240306-0140_**0001**.dat). The continuous waveform data will vary in length and content as well as its sampling rate and analog to digital conversion (A/D) settings. The header file is an ASCII text file providing meta information about the data segment – what signals are available (ECG, respiratory, counts, etc), reference start date/time, sample rate, A/D parameters, and compression method (10 available compression algorithms). As illustrated in Figure 1, case p10002348-21240306-0140 contains 13 data segments at different points in time starting in 2/14/2021 at 16:5:57 at the start of the first data segment and ending at the 13th data segment that starts at 2/14/2021 at 17:39:48. Each case contained in the MIMIC-IV WFDB dataset will vary in its start date/time and number of individual data segments and their respective duration per data segment.

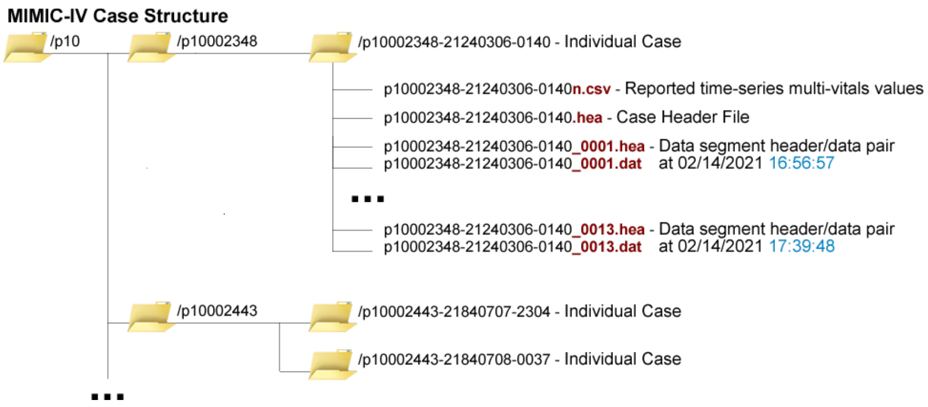

MIMIC-IV Case Structure
/p10
/p10002348
/p10002348-21240306-0140 - Individual Case
p10002348-21240306-0140n.csv - Reported time-series multi-vitals values
p10002348-21240306-0140.hea - Case Header File
p10002348-21240306-0140_0001.hea - Data segment header/data pair
p10002348-21240306-0140_0001.dat at 02/14/2021 16:56:57
...
p10002348-21240306-0140_0013.hea - Data segment header/data pair
p10002348-21240306-0140_0013.dat at 02/14/2021 17:39:48
/p10002443
/p10002443-21840707-2304 - Individual Case
/p10002443-21840708-0037 - Individual Case
...

Figure 1 MIMIC-IV WFDB Folder Structure

In addition to raw signal waveform data, some data segment files provide counts and trending measurements produced by the source patient monitor device such as Premature Ventricular Contraction (PVC) counts, alarms status, and rhythm status annotations.

The above MIMIC-IV waveform organization and encoding approach is referred to as the MIMIC Waveform Database (WFDB). While some data decoding and visualization tools are available on PhysioNet/MIMIC websites to work with the WFDB data, the tools are largely limited to the Linux environment and designed for direct human user interaction. Programmatic tools (ex Java-based code) that is more appropriate for automated Extration/Transform/Load (ETL) needs have not been updated since 2014.

The aims of this project is to integrate both clinical and waveform MIMIC-IV data into an OMOP CDM require the setup of an automated ETL process for initial conversion of MIMIC-IV data to OMOP as well as support future contributions. This suggests the need for building a data process pipeline that can accomadate both current WFDB format data as well as future source data.

Method

There are two methods employed in this project, one to handle the time-series vitals data contained in the CSV file. The second method deals with the continuous waveform data.

Continuous CSV vitals data – there are 21 possible vitals supported in the MIMIC-IV CSV vitals file (see Table 1). Not all the vitals however contain reported values or are mapped to an OMOP LOINC concept at this time.

Column Name	Mapped to OMOP Table	Units	Notes
time	Yes – measurement date/time	milliseconds	From case reference date/time
ABPd	Yes – Invasive diastolic arterial pressure	mmHg	**SNOMED.4354253**
ABPm	Yes - Invasive mean arterial pressure	mmHg	**SNOMED. 4108290**
ABPs	Yes – Invasive systolic arterial pressure	mmHg	**SNOMED. 4353843**

Delta QTc	No	msec	
HR	Yes - Heart rate	bpm	LOINC.3027018
NBPd	Yes – Non-invasive diastolic arterial pressure	mmHg	**SNOMED.4068414**
NBPm	Yes - Non-invasive mean arterial pressure	mmHg	**SNOMED.4108289**
NBPs	Yes – Non-invasive systolic arterial pressure	mmHg	**SNOMED.4354252**
PPV	No - Pulse Pressure Variation	%	
PVC	Yes - Premature ventricular contractions [#]	/min	LOINC.21490839
Pulse (ABP)	Yes - Ambulatory blood pressure monitor study report	bpm	LOINC.37021110
Pulse (NBP)	No – Pulse rate	bpm	**LOINC.4301868**
Pulse (SpO2)	Yes - Pulse oximetry	bpm	SNOMED.4098046
QT	Yes - Q-T interval	msec	LOINC.3025809
QT-HR	No - Averaged HR used for QTc	bpm	
QTc	Yes - Q-T interval corrected	msec	LOINC.3026258
RR	Yes – RR (Respiratory Rate)	rpm	LOINC.36310358
ST-III	Yes - ST amplitude.J point+60 ms Lead III	mm	LOINC.3035359
ST-V	No – non-specific chest lead	mm	
SpO2	Yes - Oxygen saturation in Blood	%	LOINC.3013502
btbHR	Yes - Heart rate | Heart | Specific heart rate (instantaneous heart rate)	bpm	**LOINC.1003329**

Table 1 MIMIC-IV CSV Reported Vitals

For those vitals indicated in Table 1 above as being mapped, a Java-based translation program (WFDBCSV2OMOP.java – Figure 2) was created to convert those vitals mapped to OMOP in Table 1 above to a formatted OMOP ETL CSV load file (Appendix A) consisting of rows representing an OMOP measurement entry. The OMOP ETL CSV is then read and each measurement is loaded into the common data model. In long data segments that are over 1 hour in length, the number of vitals generated can be over 1 million measurements. Depending on the particular use case, the volume of data at its native resolution may not be necessary. The WFDBCSV2OMOP program can be set to derive a subset of measurements based on the needs of the particular implementer using four execution parameters:

WFDBCSV2OMOP Parameter	Description
Start offset (*o*)	How many minutes of data to skip over – default is 0 mins (start from the beginning)
Report duration (*r*)	How many minutes to report measurements – default is to report all measurements

Summary period (*s*)	The number of minutes to summarize the measurement. Measurements are therefore reported out every ***s*** minutes instead of the native reporting resolution – default is to report at the native resolution.
Summary method (*mean, median, first, last*)	If measurements are summarize, indicates the method used – default is ***median***

Table 2 WFDBCSV2OMOP execution parameters

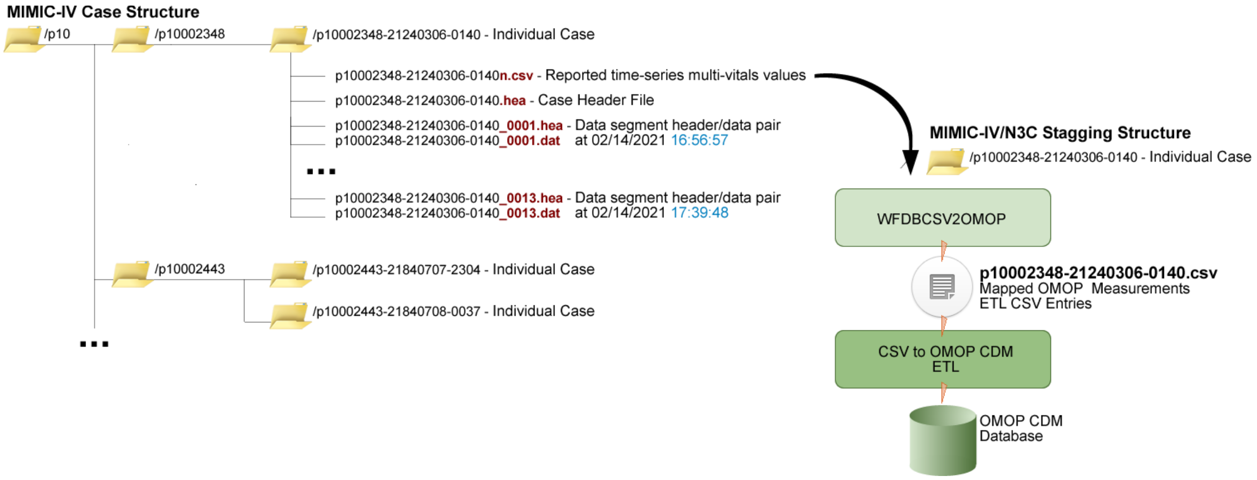

MIMIC-IV Case Structure
/p10
/p10002348
/p10002348-21240306-0140 - Individual Case
p10002348-21240306-0140n.csv - Reported time-series multi-vitals values
p10002348-21240306-0140.hea - Case Header File
p10002348-21240306-0140_0001.hea - Data segment header/data pair
p10002348-21240306-0140_0001.dat at 02/14/2021 16:56:57
...
p10002348-21240306-0140_0013.hea - Data segment header/data pair
p10002348-21240306-0140_0013.dat at 02/14/2021 17:39:48
/p10002443
/p10002443-21840707-2304 - Individual Case
/p10002443-21840708-0037 - Individual Case
...
MIMIC-IV/N3C Stagging Structure
/p10002348-21240306-0140 - Individual Case
WFDBCSV2OMOP
p10002348-21240306-0140.csv
Mapped OMOP Measurements
ETL CSV Entries
CSV to OMOP CDM
ETL
OMOP CDM
Database

Figure 2 WFDBCSV2OMOP.java Example

Continuous Waveform Data (*.hea and *.dat) - In previous work involving multiple clinical sites, both hospital and pre-hospital (ambulance and fire department) settings where a diverse mix of patient care records (PCRs), electronic health records (EHRs), and continuous medical devices (defibrillators, electrocardiographs, patient monitors, etc) are used, an object oriented adaptive approach was developed to deal with the diversity of data sources – electronic Patient Care Record (ePCR) systems and medical devices. The approach consists of three stages: 1) Device adapters; 2) Data Staging; and 3) ETL to a target common data model (CDM) storage. The three components are further executed and manged by a common pipeline process that operate on application programming interfaces (APIs) following accepted software design and implementation practices.

1. ***Device adapters*** – these are object oriented designed/programmed software components for converting specific source data into a common staging format. These components further implement a common application programming interface (API) providing an abstraction layer to separate the specific characteristics of a specific source device from a set of functions (software behaviors) that is common across all participating devices. For example signals are captured and stored using a specific sampling rate (sps – samples per second). The Device Adapter software is required to implement an API function to provide the sampling rate (ex ***`public int getSampleRate()`***). Other API functions include getting the signal values as an array of integer values (***`int[] getSignal()`***), amplitude resolution (***`getResolution()`***), etc.

2. ***Data Staging*** – the pipeline is designed to be document-centric or tranactional. That is each step consumes a transaction (representing a patient case or data segment within a case), performs a relatively narrow/well-defined action of work, and produces a document or transaction to be consumed by the next task. This design approach allows unit and blackbox testing and validation of various components within the pipeline. The proprietary or binary (WFDB) data is first transformed and staged as an XML transaction file using the ccsiecg.dtd (http://tipi-is.com/dtd/ccsiecg.dtd) XML schema previously developed and used for electrocardiographic and multi-parameter data. This is an open format that contains meta data (same data found in the *.hea files), decoded and normalized waveform (converted to standard units – uV/LSB, mmHg, respiratory beats per minute, etc), and derived measurements (not originally available in the WFDB).
3. ***Extraction/Transform/Load (ETL)*** – consists of two ETL programs. The first use existing electrocardiograph measurements and computerized interpretation statements to OMOP CDM maps to convert the staged case XML transaction file to a CSV formatted file containing row level measurements and observation entries – by specifying the source, source value, and standard OMOP CDM concept to be mapped to. The second ETL consumes the CSV file (Appendix B) and makes entries into the MIMIC-IV OMOP CDM tables.

Throughout the pipeline process, the staging (XML) and ETL (CSV) files retain the same filename of the source WFDB header/binary data files (Figure 3). The CSV therefore contains source values that can be used (ex visit_details) to reference back to the XML or WFDB source data depending on the end-user needs or preference in working with decoded XML data or the binary WFDB signal data.

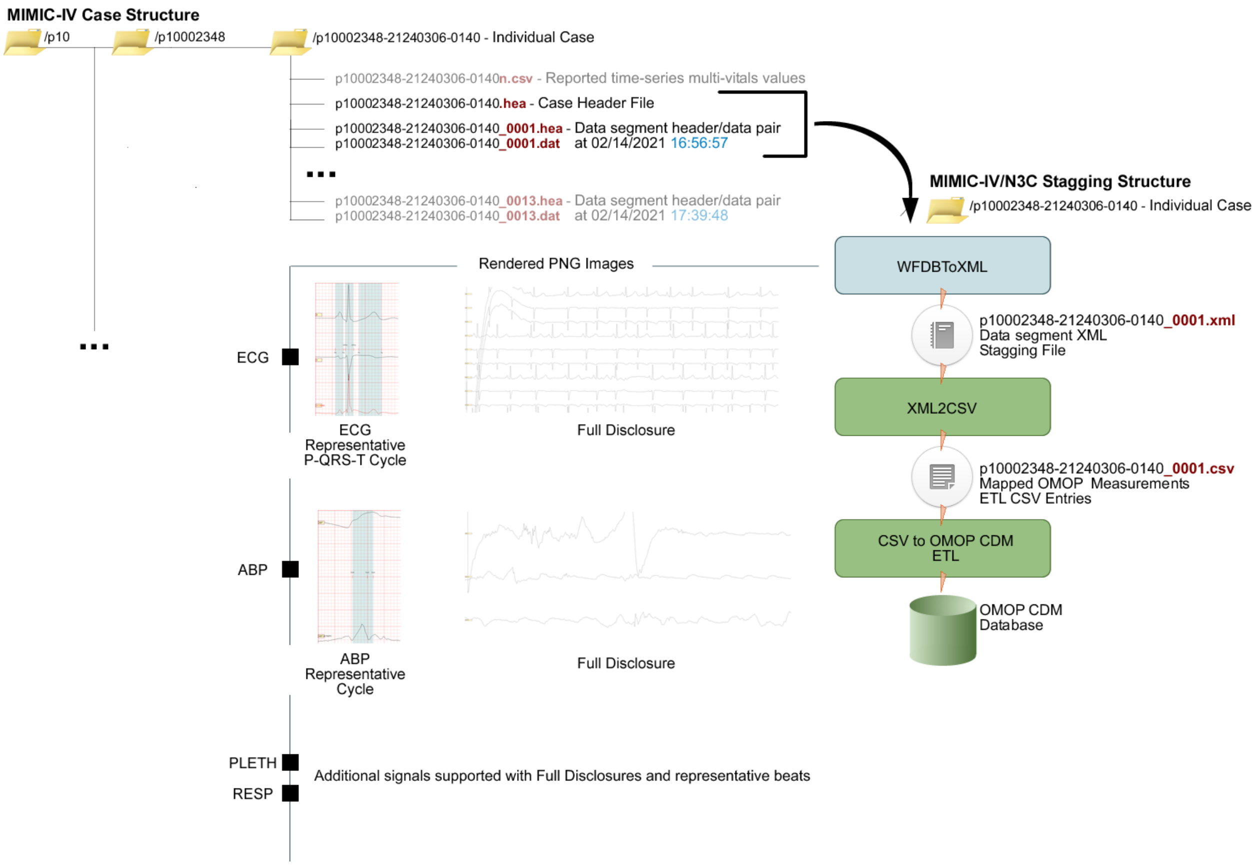

MIMIC-IV Case Structure
/p10
/p10002348
/p10002348-21240306-0140 - Individual Case
p10002348-21240306-0140n.csv - Reported time-series multi-vitals values
p10002348-21240306-0140.hea - Case Header File
p10002348-21240306-0140_0001.hea - Data segment header/data pair
p10002348-21240306-0140_0001.dat at 02/14/2021 16:56:57
...
p10002348-21240306-0140_0013.hea - Data segment header/data pair
p10002348-21240306-0140_0013.dat at 02/14/2021 17:39:48
...
MIMIC-IV/N3C Stagging Structure
/p10002348-21240306-0140 - Individual Case
Rendered PNG Images
WFDBToXML
p10002348-21240306-0140_0001.xml
Data segment XML
Stagging File
XML2CSV
p10002348-21240306-0140_0001.csv
Mapped OMOP Measurements
ETL CSV Entries
CSV to OMOP CDM
ETL
OMOP CDM
Database
ECG
ECG
Representative
P-QRS-T Cycle
Full Disclosure
ABP
ABP
Representative
Cycle
Full Disclosure
PLETH
RESP
Additional signals supported with Full Disclosures and representative beats

Figure 3 WFDBToXML.java Staging and Transform to OMOP

Staging Data

The Staging Data is an XML format transaction file representing each case WFDB data segment. For example for case p10/ p10002348-21240306-0140, the segment header p10002348-21240306-0140_0001.hea and binary data p10002348-21240306-0140_0001.dat are decoded and merged into a single p10002348-21240306-0140_0001.xml file using the schema defined at http://tipi-is.com/dtd/ccsiecg.dtd. The XML schema structure consists of key blocks:

1. Meta – contains the p10002348-21240306-0140_0001.hea data values as well as adapter (#1) normalization information.
2. Signal – for each waveform signal, a separate block of decoded comma delimited signal values are provided along with its sampling rate and normalized amplitude units. With the signal data stored as a simple comma delimited list of amplitude values, it can be easily copied and pasted into analysis programs like R, MatLab, or EXCEL for visualization or signal processing/exploration. For signals like electrocardiographic and arterial blood pressure

(episodic events), the Signal section also includes a signal averaged representative single cycle waveform based on identitified events (ex QRS detector). This is a derived waveform that is made available to represent the data segment but not presently used directly in the MIMIC OMOP CDM. However, derived measurements using the representative waveform such as QRS width, R-wave amplitude, ST at J-point, etc are saved to the CSV staging file for ETL import into the MIMIC OMOP CDM. In the case of electrocardiographic signals, the Signal XML block will also contain detected beat locations/annotations used to calculate the mean and median heart rates, dominant beat classification, and multiple family (Normal vs Premature Ventricular Contractions) formation.

3. Derived Signal – in situations where multiple electrocardiographic leads are available for a MIMIC data segment (ex MCL1, II, V), additional derived signals may be available in the XML staging file such as a derived vector magnitude (VCGMag) signal that combines all available electrocardiographic leads. The VCGMag is primarily used for QRS detection where available. For single lead electrocardiograph signals, an derived activity function is used.

Signal Normalization

All the signal processing algorithms used in this project were repurposed software originally developed to handle 500 sps 12-lead ECG records captured in pre-hospital and emergency departments. ECG signals and other continuous waveform signals (ABP, PLETH, Respiration) can vary in their sampling rate from 62.5 to 250 samples per second. MIMIC-IV signal can also have different analog to digital (A/D) gain and baselines values. First the baseline offset is removed from each signal if present. Second, the A/D gain is used to scale the signal to its indicated units. For ECG signals which are in mv units, the signal is further scaled to uV/LSB resolution. To reduce the effort by end-users of the signals contained in the staged XML version of the MIMIC-IV WFDB data, signals are up-sampled using linear interpolation to bring signal sampling rates up to a minimum of 125 samples per second to eliminate the use of floating point math on time and amplitude calculations involved in analysyses and graphic presentations. ECG signals in data segments less than 1 hour in duration are further up-sampled to 500 sps. ECG signals that are more than 1 hour in length will be up-sampled to 250 sps due to memory constraints to support long data arrays. The original signals are retained within the staged XML file and labeled with a ".ORIGINAL" (ex ID="II.ORIGINAL") and the normalized signal is labled as indicated (ex ID="II"). Therefore the user can work with either signals – original or normalized. All derived features however are based on the normalized signal as the units of measure are in milliseconds and microvolts typical of how data is presented to a clinician.

Extracted Features

Two types of extracted features were produced in this project. The first pertains to electrocardiographic signals. The second derived features are derived respiratory rate. For MIMIC-IV cases with one or more electrocardiographic signal (ex MCL1, II, III, V, etc) a QRS detector was executed on the entire data segment signal using an activity function. For single lead ECG records, an activity function used a forward/backward bandpass FIR filter on the sum of weighted first and second difference values. For multi-lead ECG records, a merged activity function similar to deriving a vector managnitude and forward/backward bandpass FIR filter was used to detect QRS beat locations. In both single and multi-lead conditions, the bandpass frequencies are 0.5 and 100.0 Hz.

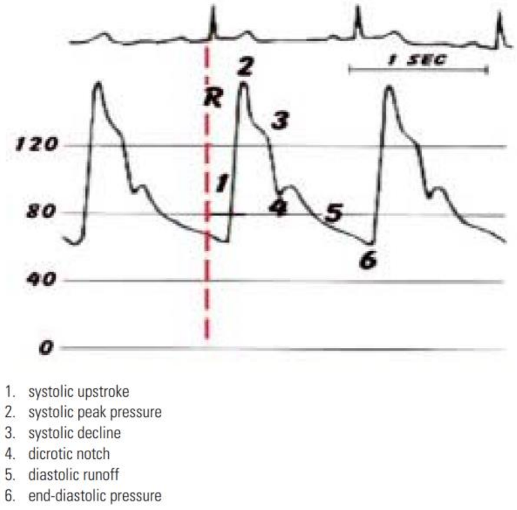
The QRS detector is tuned to be highly sensitive (error on the side of false positives) and rely on a subsequent QRS morphology classifier algorithm ("family formation") to identify the dominant waveform (QRS complex) from which a signal averaged representative beat is formed. A signal averaged representative beat is formed by overlaying all members of a "family" and taking the average amplitude values over a one second window with the fiducial point is centered at 200ms. In general, the fiducial point is typically represented by the dominant R or S wave of the QRS cycle. However in this design we account for errors due to sampling rates (temporal) and noise (vertical) by first aligning members of a family using the QRS morphology instead of using a single point like the peak of the R or S wave. Once the signal averaged representative beat is formed, the P-wave, QRS, ST, and T-wave features are derived and stored in the XML schema file representing the WFDB data segment.

The second set of derived features is the repiratory rate using a forward/backward bandpass FIR filtered line crossing algorithm to calculate the respiratory rate at 1 minute intervals using a moving 30 second window. The derived respirtatory rate is reported out as a simple average and Alpha-trimmed Mean (depending on the number of line crossings is a 10% upper and lower trimmed mean or median).

If arterial blood pressure (ABP) or plethysmographic (PLETH) signals are available, a single cycle representative beat is formed using its timing relationship with the QRS complex Figure 1. While we do not at present have standard LOINC concepts for ABP or PLETH derived features, we do provide single cycle onset/offset locations to facilitate user algorithms to calculate features such as systolic peak pressure, systolic decline, dicotic notch, etc. The calculated ABP and PLETH onset and offset location is provided in the signal's first family portrait measurements block within the XML file:

- GLOBAL SYSONSET (Sysb) in samples
- GLOBAL SYSOFFSET (Syse) in samples

In future work we will attempt to provide various morphology features illustrated in Figure 4.



(From Mark JB: Atlas of Cardiovascular Monitoring. New York, Churchill Livingstone, 1998: figure 8-1)

Figure 4 Moxham et al - Understanding Arterial Pressure Waveforms

In addition to the basic ABP and PLETH landmark measurements, a representative waveform image will also be rendered as part of the staged derived output from normalization and staging for import into the OMOP common data model (Figure 5).

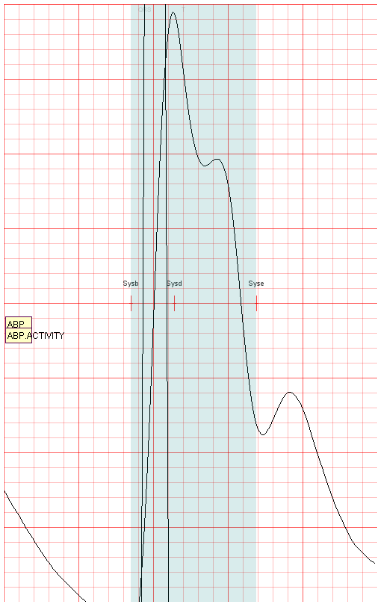


Figure 5 ABP Representative PNG image

Similarly, each data segment that contains valid data will include rendered PNG formatted image of the representative beat. At the bottom of these images is the rendering of the accompanying activity function used to identify key landmarks of the representative beat and shaded accordingly (see Appendix D for an example image set). The ECG, respiratory, and blood pressure continuous waveforms are also rendered as a full-disclosure PNG image file to provide users a visual resource to evaluate and assess the quality of the WFDB segment and therefore the appropriateness of including the derived features for analysis. Some WFDB segments may have excessive artifacts (positional changes, leads off episodes, etc) that should be excluded for analysis. The XML schema also provide signal quality metrics in the signal data block (<CCSI:signalEnergy>, <CCSI:highFreqQA>, <CCSI:blwQA> - baseline wander, and <CCSI:snrQA> - signal to noise ratio) to further provide indicators of signal quality for each signal. These data quality metrics are not presently imported into the OMOP CDM but available for analytics and algorithm develoment.

Derived ECG Measurements

Due to incomplete coverage of the LOINC coding for ECG features a subset of measurements derived and contained in the staged XML data segment file will be included in the OMOP common data model. Table 3 below lists example derived measurements for lead II and its OMOP concept map. All standard leads (I-III, aVR, aVL, aVF, V1-V6) except "V" and "MCL" are represented similarly by LOINC concepts. The OMOP concepts selected were based on expert review and current mapping of Tufts Medical Center's 12-lead ECG measurements produced by Philips TC70 class electrocardiographs. The mapping has been extended to cover other manufacturers like GE 12SL and GLASGOW measurements and computerized interpretation statements.

Derived Feature Name	Mapped to OMOP Measurements Table As	Units	OMOP Concept ID
GLOBAL HEART RATE	Heart rate	beats/min	LOINC.3027018

GLOBAL QT	Q-T interval	milliseconds	LOINC.3025809
GLOBAL QTc	Q-T interval corrected	milliseconds	LOINC.3026258
GLOBAL QRS Dur	QRS duration	milliseconds	LOINC.3022022
II QRS DUR	QRS duration in lead II	milliseconds	LOINC.3017568
GLOBAL PR INT	P-R Interval	milliseconds	LOINC.3007794
II I AMP	QRS initial amplitude in lead II	uV	LOINC.3025713
II F AMP	QRS terminal amplitude in lead II	uV	LOINC.3015593
II Q AMP	Q wave depth in lead II	uV	LOINC.3009332
II Q DUR	Q wave duration in lead II	milliseconds	LOINC.3020600
II R AMP	R wave amplitude in lead II	uV	LOINC.3002640
II R DUR	R wave duration in lead II	milliseconds	LOINC.3010930
II S AMP	S wave amplitude in lead II	uV	LOINC.3015703
II S DUR	S wave duration in lead II	milliseconds	LOINC.3012847
II STJ60	ST amplitude.J point+60 ms Lead II	uV	LOINC.3035117
II STJ80	ST amplitude.J point+80 ms Lead II	uV	LOINC.3008999
II TPOS AMP	T wave amplitude in lead II	uV	LOINC.3025057

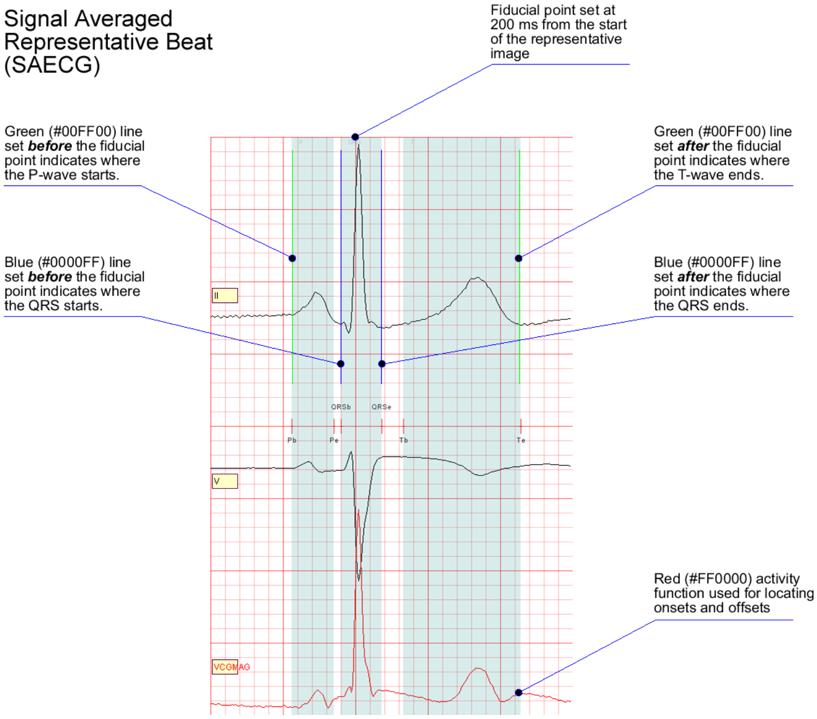
Table 3 ECG Derived Features Mapping to OMOP Concepts – Lead II

PhysioNet WFDB Backwards Linkage

Both the time series measurement and derived measurement records inserted in the OMOP database will use the OMOP visit_details table to centrally manage the linkage of measurements, observations, and conditions to its PhysioNet WFDB source. In the above summary, the staging XML and rendered PNG image files are contained in a folder of the same name as the WFDB source. Further the filenames of the XML and PNG files retain the same WFDB data segment file name with extensions to help identify the image as representing the ABP, PLETH, or RESP (respiratory) signals. No extension refers to the ECG signals. Similarly, the OMOP measurement records has a foreign key to the visit_occurrence record which in turn is linked to a visit_detail containing the same WFDB source reference from which the measurement was derived. While the use of the OMOP visit_detail database table to provide linkage to its source is somewhat outside the typical definition of the visit_detail, it was implemented in this way as a relatively clean and low-impact approach to referencing external objects like the WFDB binary source (similar to referencing an external image resource).

Data Quality

In this pilot study and proof of concept, we adapted existing 12-lead ECG algorithms to normalize and extract features on various continuous signals from the MIMIC-IV WFDB data set. Numerous adaptations in the original filters, activity functions, and measurements have produced acceptable results within this context and aims. In order improve the quality of derived features a mechanism was implemented to allow reviewers to graphically annotate the ECG signal averaged waveform image to indicate by expert inspection where four key landmarks should be placed (Figure 6). Macromedia's Fx image editor was used to annotate the representative ECG PNG file. Any PNG file editor can be used as the annotation is a straight line drawn in the upper ¼ of the PNG file using two colors – Green (#00FF00) or Blue (#0000FF). The annotated SAECG image can then be read by the program called **OnsetOffsetAdjImage**.java (Figure 7) to update the P-wave onset, QRS onset/offset, and T-wave offset locations and re-measure the P-wave, QRS, ST, and T-wave amplitudes and duration features and update the staging XML document accordingly. These annotated onset/offset corrections are then propagated throughout subsequent data segment automatically if the ECG waveform did not significantly change in morphology.



Signal Averaged Representative Beat (SAECG)

Fiducial point set at 200 ms from the start of the representative image

Green (#00FF00) line set ***before*** the fiducial point indicates where the P-wave starts.

Green (#00FF00) line set ***after*** the fiducial point indicates where the T-wave ends.

Blue (#0000FF) line set ***before*** the fiducial point indicates where the QRS starts.

Blue (#0000FF) line set ***after*** the fiducial point indicates where the QRS ends.

Red (#FF0000) activity function used for locating onsets and offsets

Figure 6 Image-based Review/Editing and Data Quality Control

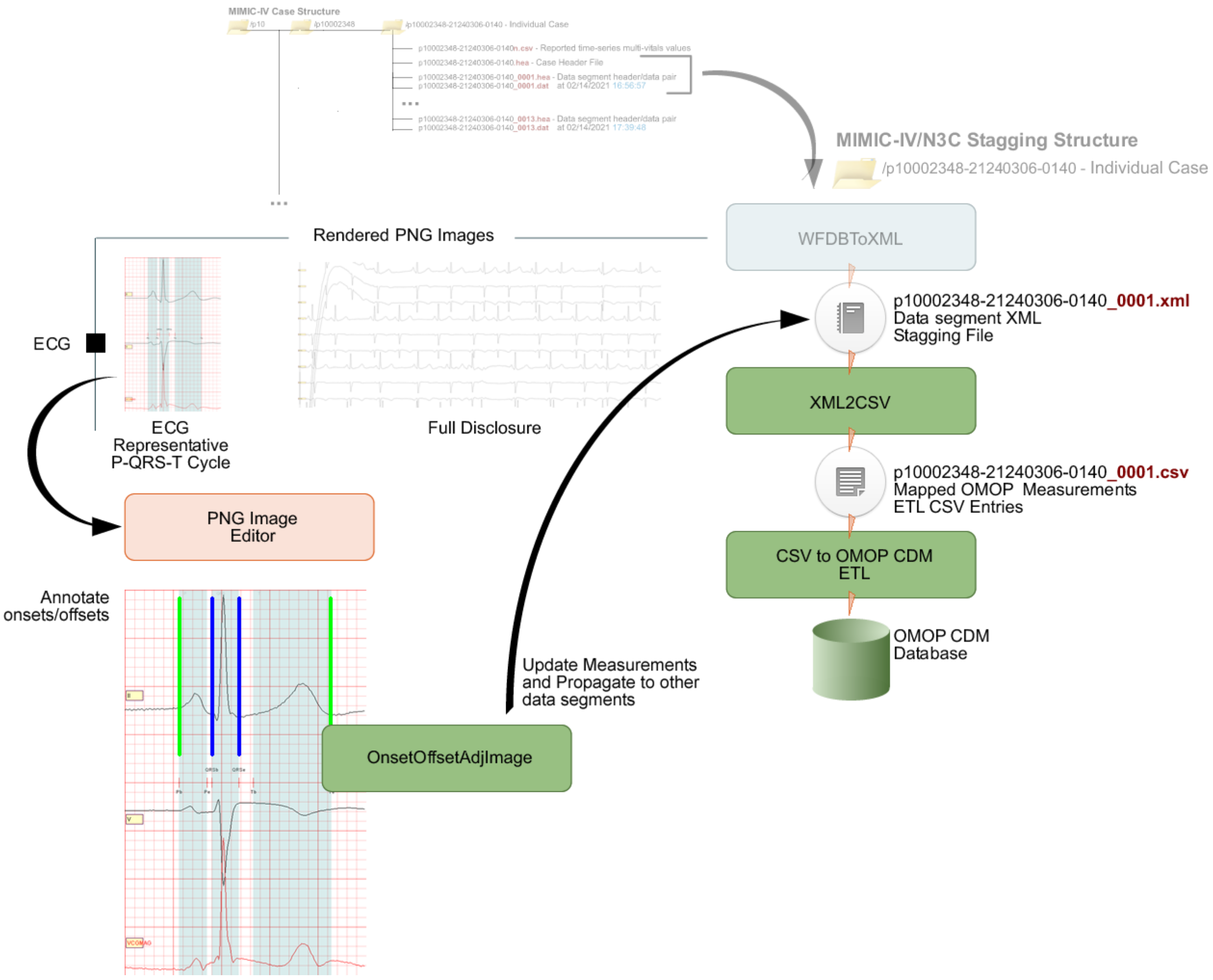

MIMIC-IV Case Structure
/p10
/p10002348
/p10002348-21240306-0140 - Individual Case
p10002348-21240306-0140n.csv - Reported time-series multi-vitals values
p10002348-21240306-0140.hea - Case Header File
p10002348-21240306-0140_0001.hea - Data segment header/data pair
p10002348-21240306-0140_0001.dat at 02/14/2021 16:56:57
...
p10002348-21240306-0140_0013.hea - Data segment header/data pair
p10002348-21240306-0140_0013.dat at 02/14/2021 17:39:48
...
MIMIC-IV/N3C Stagging Structure
/p10002348-21240306-0140 - Individual Case
WFDBToXML
p10002348-21240306-0140_0001.xml
Data segment XML
Stagging File
XML2CSV
p10002348-21240306-0140_0001.csv
Mapped OMOP Measurements
ETL CSV Entries
CSV to OMOP CDM
ETL
OMOP CDM
Database
Rendered PNG Images
ECG
ECG
Representative
P-QRS-T Cycle
Full Disclosure
PNG Image
Editor
Annotate
onsets/offsets
Update Measurements
and Propagate to other
data segments
OnsetOffsetAdjImage

Figure 7 Review and onset/offset Annotations

Discussion and Future Work/Direction

[Talk about other signals not addressed in this demonstration work. Include discussion regarding extracted feature resolutions – every 1 min or every 5 mins etc – what are the use cases to help determine these feature issues. Long term algorithm related tracking and quality control issues – XML schema does provide fields to identify the algorithm/version. This can then be linked to GitHub source packages. Talk about support for comparitive research on algorithms and human reviewers.]

Dissemination Artifacts

The MIMIC-IV/N3C project GitHub will include the following contributions and software support for working with the waveform data:

1. iMedical Solutions LLC core, XML, and ECG signal processing libraries in the form of Java jar-files. These libraries for the underlying signal handline and feature derivation tasks.

2. Tufts Medical Center – Clinical Translational Science Institute (Tufts CTSI) Java OMOP libraries in the form of a binary jar-file.
3. Other open source libraries such as Apache's POI, XML, and image rendering. These are specific versions used in this project. Newer open source versions have not been qualified or verified.
4. N3C MIMIC-IV ETL source and jar-file containing WFDB to XML ETL components and pipeline code. The software includes data segment decoding, signal normalization, and top-level beat detection and feature derivation. The software also include XML data segment and representative beat image review and editing – re-derive ECG features from manual onset/offset review.
5. XML to CSV conversion using a OMOP CDM concept map definitions - EXCEL file.
6. Example XML signal extraction program
7. Example XML measurements extraction program

In addition to the above software artifacts, documentation and example development setup user guide will be provided on the GitHub.

Limitations

The translation of wdb/WDB MIMIC-IV does not include counts data like PVC (premature ventricular complex) that is contained in the WFDB data segments. Some counts are captured through the MIMIC-IV CSV time series reported data instead. As listed in Table 1, not all MIMIC-IV time series reported data is presently mapped to an OMOP standard concepts. In future iteration, a custom vocabulary may need to be constructed to support all available data.

Similarly, derived P-QRS-T features for leads labeled "V" or "MCL" are not imported into the OMOP CDM as there are no standard LOINC concept to map features for these signals. A new custom vocabulary may be required to fully support these non-specific chest leads.

The derived ECG features are based on algorithmic and one member of the project team's subjective determination of the P, QRS, and T wave onsets and offsets. A panel of cardiology experts approach is required in order to establish a "gold" standard basis for the derived ECG measurements. This is a significant undertaking and outside the present scope of the MIMIC-IV to OMOP demonstration project.

Appendix A OMOP Vitals CSV ETL Input Format

COLUMN NAME	REQUIRED	DESCRIPTION
CASE ID	Yes	Root case ID – ex p10002348
SEGMENT NAME	Yes	Data segmentment name – ex p10002348-21240306-0140
DATE-TIME	Yes	Date/time of the measurement - Format yyyy-MM-dd HH:mm:ss
SRC NAME	Yes	Source name
CONCEPT ID	Yes	Measurement concept ID
CONCEPT NAME	Yes	Measurement concept name
VALUE	Yes	Value of the measurement
UNIT CONCEPT ID	Yes	Units concept ID
UNIT CONCEPT NAME	Yes	Units concept name
VISIT DETAIL - SOURCE	No	Default – "csv"
VISIT DETAIL - START FROM MINUTES	No	Start offset in minutes
VISIT DETAIL - REPORT MINUTES	No	Report length in minutes
VISIT DETAIL - SUMMARIZE MINUTES	No	Summary length in minutes
VISIT DETAIL - METHOD	No	Summary method

Table A.1 OMOP Vitals ETL CSV Format

Example: Native Reporting Frequency

```
Case ID,Segment Name,Date-time,Src Name,Concept ID,Concept Name,Value,Unit Concept ID,Unit Concept Name,Visit Detail - Source,Visit Detail - Start from
minutes,Visit Detail - Report minutes,Visit Detail - Sumarize minutes,Visit Detail - Method
p10002348,p10002348-21240306-0140,2124-03-06 01:40:00,PVC [/min],21490839,Premature ventricular contractions [#],0.0,8483,counts per minute,csv,0,-
1,0,NONE
p10002348,p10002348-21240306-0140,2124-03-06 01:40:00,Pulse (SpO2) [bpm],4098046,Pulse oximetry,68.0,4118124,beats/min,csv,0,-1,0,NONE
p10002348,p10002348-21240306-0140,2124-03-06 01:40:00,RR [rpm],36310358,RR,18.0,4117833,breaths/min,csv,0,-1,0,NONE
p10002348,p10002348-21240306-0140,2124-03-06 01:40:00,SpO2 [%],3013502,Oxygen saturation in Blood,94.4,8554,percent,csv,0,-1,0,NONE
p10002348,p10002348-21240306-0140,2124-03-06 01:40:01,btbHR [bpm],3027018,Heart rate,73.39,4118124,beats/min,csv,0,-1,0,NONE
p10002348,p10002348-21240306-0140,2124-03-06 01:40:03,PVC [/min],21490839,Premature ventricular contractions [#],0.0,8483,counts per minute,csv,0,-
1,0,NONE
p10002348,p10002348-21240306-0140,2124-03-06 01:40:03,Pulse (SpO2) [bpm],4098046,Pulse oximetry,68.0,4118124,beats/min,csv,0,-1,0,NONE
p10002348,p10002348-21240306-0140,2124-03-06 01:40:03,RR [rpm],36310358,RR,18.0,4117833,breaths/min,csv,0,-1,0,NONE
p10002348,p10002348-21240306-0140,2124-03-06 01:40:03,SpO2 [%],3013502,Oxygen saturation in Blood,94.4,8554,percent,csv,0,-1,0,NONE
p10002348,p10002348-21240306-0140,2124-03-06 01:40:06,PVC [/min],21490839,Premature ventricular contractions [#],0.0,8483,counts per minute,csv,0,-
1,0,NONE
p10002348,p10002348-21240306-0140,2124-03-06 01:40:06,Pulse (SpO2) [bpm],4098046,Pulse oximetry,68.0,4118124,beats/min,csv,0,-1,0,NONE
p10002348,p10002348-21240306-0140,2124-03-06 01:40:06,RR [rpm],36310358,RR,18.0,4117833,breaths/min,csv,0,-1,0,NONE
p10002348,p10002348-21240306-0140,2124-03-06 01:40:06,SpO2 [%],3013502,Oxygen saturation in Blood,94.3,8554,percent,csv,0,-1,0,NONE
p10002348,p10002348-21240306-0140,2124-03-06 01:40:09,PVC [/min],21490839,Premature ventricular contractions [#],0.0,8483,counts per minute,csv,0,-
1,0,NONE
p10002348,p10002348-21240306-0140,2124-03-06 01:40:09,Pulse (SpO2) [bpm],4098046,Pulse oximetry,68.0,4118124,beats/min,csv,0,-1,0,NONE
```

Example: Summarized Reporting Frequency – start from beginning on the first 120 minutes at 5 minute summary intervals using median

```
Case ID,Segment Name,Date-time,Src Name,Concept ID,Concept Name,Value,Unit Concept ID,Unit Concept Name,Visit Detail - Source,Visit Detail - Start from
minutes,Visit Detail - Report minutes,Visit Detail - Sumarize minutes,Visit Detail - Method
```

```
p10002348,p10002348-21240306-0140,2124-03-06 01:45:00,HR [bpm],3027018,Heart rate,81.0,4118124,beats/min,csv,0,120,5,MEDIAN
p10002348,p10002348-21240306-0140,2124-03-06 01:45:00,PVC [/min],21490839,Premature ventricular contractions [#],0.0,8483,counts per minute,csv,0,120,5,MEDIAN
p10002348,p10002348-21240306-0140,2124-03-06 01:45:00,Pulse (SpO2) [bpm],4098046,Pulse oximetry,71.0,4118124,beats/min,csv,0,120,5,MEDIAN
p10002348,p10002348-21240306-0140,2124-03-06 01:45:00,RR [rpm],36310358,RR,19.0,4117833,breaths/min,csv,0,120,5,MEDIAN
p10002348,p10002348-21240306-0140,2124-03-06 01:45:00,SpO2 [%],3013502,Oxygen saturation in Blood,94.4,8554,percent,csv,0,120,5,MEDIAN
p10002348,p10002348-21240306-0140,2124-03-06 01:45:00,btbHR [bpm],3027018,Heart rate,73.39,4118124,beats/min,csv,0,120,5,MEDIAN
p10002348,p10002348-21240306-0140,2124-03-06 01:50:00,HR [bpm],3027018,Heart rate,74.0,4118124,beats/min,csv,0,120,5,MEDIAN
p10002348,p10002348-21240306-0140,2124-03-06 01:50:00,PVC [/min],21490839,Premature ventricular contractions [#],0.0,8483,counts per minute,csv,0,120,5,MEDIAN
p10002348,p10002348-21240306-0140,2124-03-06 01:50:00,Pulse (SpO2) [bpm],4098046,Pulse oximetry,74.0,4118124,beats/min,csv,0,120,5,MEDIAN
p10002348,p10002348-21240306-0140,2124-03-06 01:50:00,RR [rpm],36310358,RR,17.0,4117833,breaths/min,csv,0,120,5,MEDIAN
p10002348,p10002348-21240306-0140,2124-03-06 01:50:00,SpO2 [%],3013502,Oxygen saturation in Blood,94.6,8554,percent,csv,0,120,5,MEDIAN
p10002348,p10002348-21240306-0140,2124-03-06 01:50:00,btbHR [bpm],3027018,Heart rate,58.79,4118124,beats/min,csv,0,120,5,MEDIAN
p10002348,p10002348-21240306-0140,2124-03-06 01:55:00,HR [bpm],3027018,Heart rate,70.0,4118124,beats/min,csv,0,120,5,MEDIAN
p10002348,p10002348-21240306-0140,2124-03-06 01:55:00,PVC [/min],21490839,Premature ventricular contractions [#],0.0,8483,counts per minute,csv,0,120,5,MEDIAN
p10002348,p10002348-21240306-0140,2124-03-06 01:55:00,Pulse (SpO2) [bpm],4098046,Pulse oximetry,70.0,4118124,beats/min,csv,0,120,5,MEDIAN
```

Appendix B OMOP Continuous Waveform ETL CSV Input Format

COLUMN NAME	REQUIRED	DESCRIPTION
CASE ID	Yes	Root case ID – ex p10002348
SEGMENT NAME	Yes	Data segmentment name – ex p10002348-21240306-0140
DATE-TIME	Yes	Date/time of the measurement - Format yyyy-MM-dd HH:mm:ss
SRC NAME	Yes	Source name
CONCEPT ID	Yes	Measurement concept ID
CONCEPT NAME	Yes	Measurement concept name
VALUE	Yes	Value of the measurement
UNIT CONCEPT ID	Yes	Units concept ID
UNIT CONCEPT NAME	Yes	Units concept name
VISIT DETAIL - SOURCE	No	Default – "csv"
VISIT DETAIL - START FROM MINUTES	No	Start offset in minutes
VISIT DETAIL - REPORT MINUTES	No	Report length in minutes
VISIT DETAIL - SUMMARIZE MINUTES	No	Summary length in minutes
VISIT DETAIL - METHOD	No	Summary method

Table B.1 OMOP Continuous Waveform (WFDB) ETL CSV Format

Example:

```
Case ID,Segment Name,Date-time,Src Name,Concept ID,Concept Name,Value,Unit Concept ID,Unit Concept Name,Visit Detail - Source,Visit Detail - Start from
minutes,Visit Detail - Report minutes,Visit Detail - Sumarize minutes,Visit Detail - Method
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,II S DUR,3012847,S wave duration in lead
II,22,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,GLOBAL QTc,3026258,Q-T interval
corrected,438,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,II QRS DUR,3017568,QRS duration in lead
II,92,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,II R DUR,3010930,R wave duration in lead
II,58,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,GLOBAL HEART RATE,3027018,Heart
rate,72,4118124,beats/min,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,II TPOS AMP,3025057,T wave amplitude in lead II,288,9461,bel
microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,GLOBAL QRS Dur,3022022,QRS
duration,112,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,II Q AMP,3009332,Q wave depth in lead II,-67,9461,bel
microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,II F AMP,3015593,QRS terminal amplitude in lead II,0,9461,bel
microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,II STJ60,3035117,ST amplitude.J point+60 ms Lead II,-16,9461,bel
microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,II I AMP,3025713,QRS initial amplitude in lead II,-26,9461,bel
microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,II S AMP,3015703,S wave amplitude in lead II,-45,9461,bel
microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,II R AMP,3002640,R wave amplitude in lead II,1193,9461,bel
microvolt,xml,0,0,0,NONE
```

```
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,GLOBAL PR INT,3007794,P-R Interval,200,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,II Q DUR,3020600,Q wave duration in lead II,24,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,II STJ80,3008999,ST amplitude.J point+80 ms Lead II,5,9461,bel microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0001,1969-12-31 16:00:00,GLOBAL QT,3025809,Q-T interval,412,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,II S DUR,3012847,S wave duration in lead II,22,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,GLOBAL QTc,3026258,Q-T interval corrected,436,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,II QRS DUR,3017568,QRS duration in lead II,104,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,II R DUR,3010930,R wave duration in lead II,42,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,GLOBAL HEART RATE,3027018,Heart rate,74,4118124,beats/min,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,II TPOS AMP,3025057,T wave amplitude in lead II,232,9461,bel microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,GLOBAL QRS Dur,3022022,QRS duration,104,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,II Q AMP,3009332,Q wave depth in lead II,-94,9461,bel microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,II F AMP,3015593,QRS terminal amplitude in lead II,0,9461,bel microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,II STJ60,3035117,ST amplitude.J point+60 ms Lead II,-67,9461,bel microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,II I AMP,3025713,QRS initial amplitude in lead II,0,9461,bel microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,II S AMP,3015703,S wave amplitude in lead II,-49,9461,bel microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,II R AMP,3002640,R wave amplitude in lead II,1107,9461,bel microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,GLOBAL PR INT,3007794,P-R Interval,14,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,II Q DUR,3020600,Q wave duration in lead II,40,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,II STJ80,3008999,ST amplitude.J point+80 ms Lead II,-48,9461,bel microvolt,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,GLOBAL QT,3025809,Q-T interval,406,9593,millisecond,xml,0,0,0,NONE
p10002348-21240306-0140,p10002348-21240306-0140_0002,1969-12-31 16:01:44,II T AMP,3025057,T wave amplitude in lead II,232,9461,bel microvolt,xml,0,0,0,NONE
```

Appendix C WFDB to XML Software Dependencies

The MIMIC-IV to OMOP Software release version 0.1 (GitHub ....) depends on the following technologies and versions.

PACKAGE NAME	VERSION	NOTES
IMEDICAL SOLUTIONS LLC – IMSLIB.JAR	1.8	Core Java components including file handling, signal processing, and math functions
IMEDICAL SOLUTIONS LLC – CCSICORE.JAR	1.8	Core ECG, signal processing, and analytics core library
IMEDICAL SOLUTIONS LLC – CCSIXML.JAR	1.8	ECG, signals, and annotation XML data model
IMEDICAL SOLUTIONS LLC – CCSIDB.JAR	1.8	Core database access components
IMEDICAL SOLUTIONS LLC – CCSICALIPER.JAR	1.8	ECG and signal analytics such as QRS detectors, morphology analytics, and measurements – Calipers
IMEDICAL SOLUTIONS LLC – CCSIGRH.JAR	1.8	Graphics and reporting components
IMEDICAL SOLUTIONS LLC – CCSICHE.JAR	1.8	High level image renderers to produce PNG image of signals
TUFTS CTSI – CTSIWORKBENCH.JAR	1.8	High level CSV and EXCEL file reader and writers
TUFTS CTSI – COHORT.JAR	1.8	OMOP 5.3 and 6.x ETL components
TUFTS CTSI – MIMICWF.JAR AND JAVA SOURCE FILES (GITHUB ....)	1.8	MIMIC-III and MIMIC-IV WFDB decoders and ETL components
GSON IS AN OPEN-SOURCE JAVA LIBRARY TO SERIALIZE AND DESERIALIZE JAVA OBJECTS TO JSON - GSON-2.6.2.JAR	2.6.2	[May not be needed/used]
ORACLE MYSQL JDBC DRIVER - MYSQL-CONNECTOR-JAVA-8.0.18.JAR	8.0.18	JDBC driver – used connect to a OMOP database to lookup concept names (optional)
APACHE POI JAVA LIBRARIES FOR READING AND WRITING FILES IN MICROSOFT OFFICE FORMATS - POI-3.12-BETA1-20150228.JAR, POI-OOXML-3.12-BETA1-20150228.JAR, POI-OOXML-SCHEMAS-3.12-BETA1-20150228.JAR	3.12	Read and write EXCEL files (optional)
APACHE PDFBOX USED TO CREATE, RENDER, PRINT, SPLIT, MERGE, ALTER, VERIFY AND EXTRACT TEXT AND META-DATA OF PDF FILES - PDFBOX-APP-2.0.3.JAR	2.03	Optional
ORACLE J2SE 1.8 DEVELOPERS KIT - JDK-8U261-WINDOWS-X64	1.8.261	Latest build version

Table C.1 MIMIC-III/IV WFDB ETL Software Dependencies

Due to the historic nature of many of the iMedical Solutions LLC and Tufts CTSI components that were originally developed for previous projects and purposes, some of the optional components are required for compilation even if these components are not actually used in the MIMIC to OMOP ETL tools.

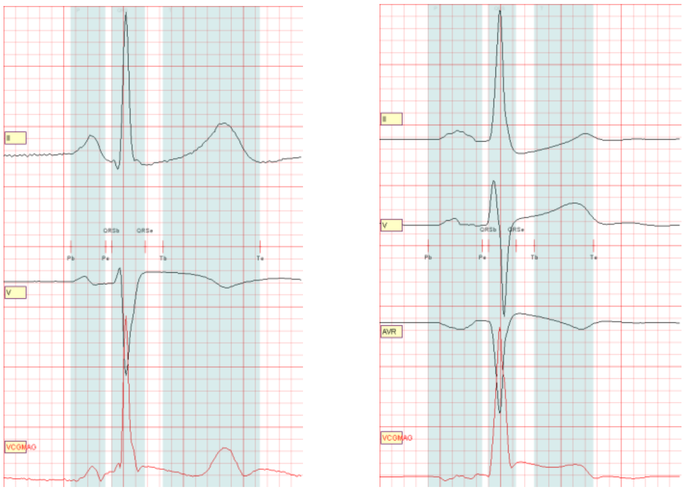
The Integrated Devlelpment Environment (IDE) used to build these ETL programs used Eclipse Version: Photon Release (4.8.0) Build id: 20180619-1200. Execution used Oracle's J2SE 1.8.261 or later run-time environment on a Windows 10 computer.

Software source control used a private repository on bitbucket.com and Atlassian's Sourcetree version 2.6.10.0 on a Windows 10 computer.

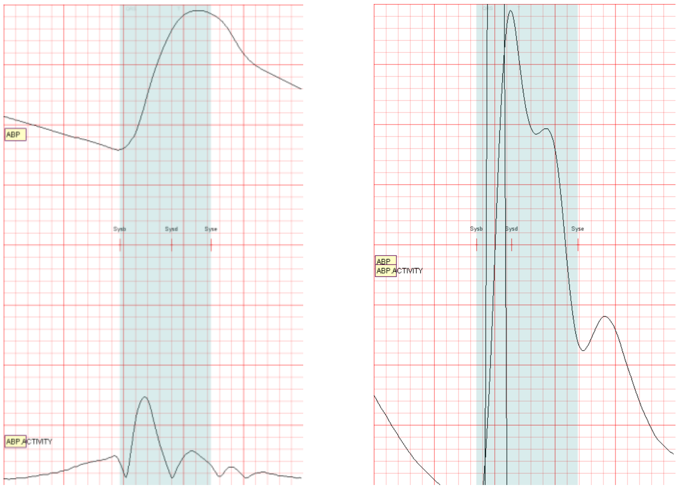
Appendix D WFDB data segment PNG image example

ECG, ABP, and PLETH signals will render a representative signal averaged single cycle waveform and its activity function. The image width is 500 pixels wide, matching the upper sampling rate of 500 samples per second. The main X-axis is divided in 100 ms intervalsl and small boxes in 20 ms per typical ECG report formats. The Y-axis is in mV divisions – or 1 small box is 1mm – typical of standard ECG presentations.

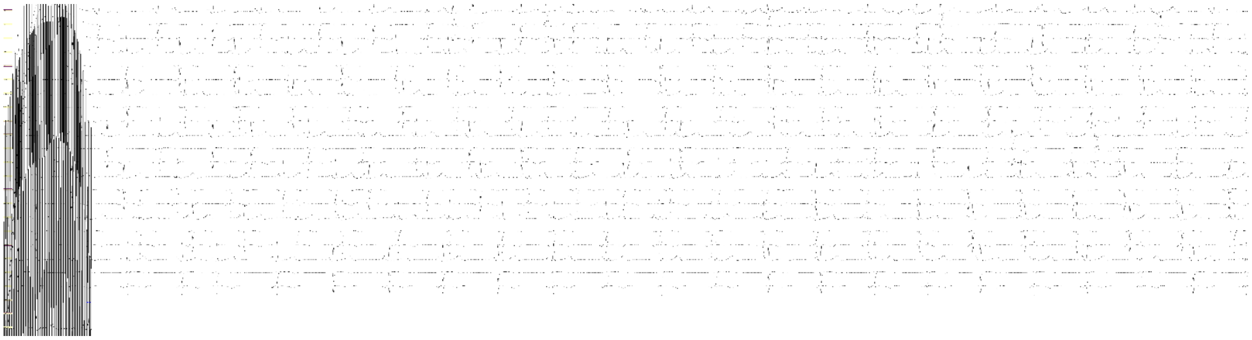
ECG Signal Averaged Representative Waveform (beat):



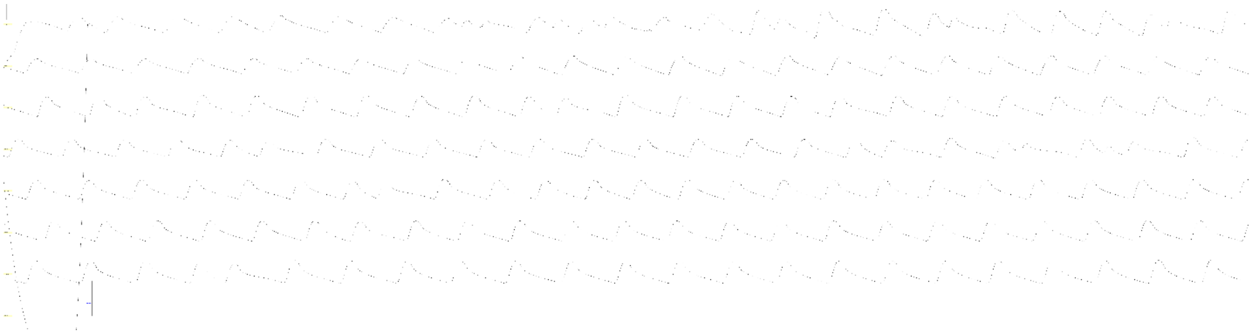
ABP Signal Averaged Representative Waveform (beat):



ECG Full Disclosue:



ABP Full Disclosure:



Appendix E Example Code For Working with the MIMIC-IV XML Staged File

The following is a code fragment for extracting lead III signal and writing it out to standard out.

```
com.ccsi.xml.io.XMLLoader xmlloader = new com.ccsi.xml.io.XMLLoader();
ECG ecg = new ECG(xmlloader.load(“\\p10\\p10002348\\p10002348-21240306-0140\\p10002348-21240306-
0140_0001.xml",false));
for (String lid : ecg.getLeadIDs()) {
   if (lid.equals(“III”)) {
      for (int d : ecg.getLead(lid).getSignalData()) {
         System.out.println(d);
      }; // end for
   };    // end if
};       // end if
```

The following is a code fragment for extracting lead III dominant representative beat signal and writing it out to standard out. Each signal within an ECG record contains an “Album” of one or more P-QRS-T morphology “Portrait”s. A portrait is a 1 second wide representative beat formed from signal averaging its members aligned horizontally (time alignment) and vertically (amplitude offsets). The portrait that contains the most number of members is designated as the “First Portrait” and in most cases the dominant or “normal” morphology for the patient. Exceptions is a bigeminy rhythm where the ectopic beats outnumber the normal beats. Additional second stage classification analysis is not included at this time for this project.

```
com.ccsi.xml.io.XMLLoader xmlloader = new com.ccsi.xml.io.XMLLoader();
ECG ecg = new ECG(xmlloader.load(“\\p10\\p10002348\\p10002348-21240306-0140\\p10002348-21240306-
0140_0001.xml",false));
for (String lid : ecg.getLeadIDs()) {
   if (lid.equals(“III”)) {
      for (int d : ecg.getLead(lid).getAlbum().firstPortrait().getSignalData()) {
         System.out.println(d);
      }; // end for
   };    // end if
};       // end if
```