**CHoRUS**

**Broad Standard Operating Procedures for Data Contributing Sites**

**Purpose**

This document is for data contributing sites to better understand what is expected to fulfill contractual obligations and what assistance you can expect from each module. This document describes the types of activities leading to the production of harmonized structured and unstructured EHR data, Images, and Monitor data.

Structured EHR data:

1. Identify the source of the data

Some sites have access to and existing repository, while other sites need to formulate a data extraction request to authorized entities (such as some data group within their CTSI).

1. Formulate the specs of the data extraction

The specification includes (i) a description of the population of patients , (ii) the scope of the extraction, which specifies what data is needed over what period of time. In the pilot phase (April-June 2023), the request is for data from 25 patients from hospital admission to hospital discharge. The time when these patients were admitted does not matter, not does their admitting diagnoses. But we do request that their length of stay exceeds 24 hours, irrespective of outcome. The specs also describe which data elements are needed. Table 1 provide this list. It is understood that not all data are available for the duration of the hospital admission. Please document version number of these specs , date of the request and format you receive the data in. This could be as elementary as a series of comma delimited files or excel spreadsheets, or as sophisticated as fully mapped OMOP tables, depending on the expertise of your data team(s). While CHoRUS will eventually require you to produce a fully “mapped” or harmonized dataset, one of the purposes of this initial request is precisely to evaluate gaps and readiness level of data generating sites.

1. Identify relevant metadata

Please include here data that can fundamentally change the interpretation of the data above. For example, the make/model of bedside monitor/ventilator, infusion pump system used, how blood pressure and temperature are typically measured (instrument and anatomical location), POC glucose device most commonly used. These metadata may be difficult to collect and there is (not yet) a codified system for it.

1. Process the request

Your data team will execute the data extraction, or these specs will be forwarded to the designated group within your institution. Please document when data are received.

1. Process the data - deidentification

If this has not been taken care of in the process of extraction, please develop a process for the de-identification of all personality identifiable information. You IRB might have allowed preserving elements of dates. This step will require your data team to have honest broking privileges and to create various “cross-walk” or mapping files such that the data can be ultimately remapped for future purposes. The research team does not have access to those maps, but they may be used, for example, to link to long-term outcomes (SSDMF) or to tools extracting geographic determinants of health using full addresses or zip codes.

1. Process the data – Quality

Please develop a process to verify that each piece of data requested is included in the data you receive, as well as the degree of missingness of variables requested. CHoRUS will recommend that you develop a process, following the recommendations of the Standards Module expert, to map your data extraction to standard OMOP tables using tools of your choice. The PERSEUS OHDSI toolkit is recommended, as it includes a fully fledged Data Quality Dashboard, in addition to assist in the customization of the Extraction-Transformation-Loading (ETL) process to your own extract.

1. Upload the extract to the CHoRUS cloud

Please upload all data tables, and metadata files to your site-specific cloud enclave within the CHoRUS cloud environment. Fiull specifications of this process is forthcoming.

These steps are for guidance only. Yet they provide a map of the typical steps necessary to produce a quality data extraction. Your site process will have variations on those steps. A description of those variations should also accompany your extract, so CHoRUS fully understands the process at each site and is in the best possible position to help local teams.

|  |  |  |
| --- | --- | --- |
| EHR domain | Specific variables | OMOP clinical tables (CDM 5.4) |
| Person | Age, Race, ethnicity, death (Y/N), death date, discharge destination, address | PERSON\_ID, DEATH |
| Visit & location | Dates of all location changes, type of location | OBSERVATION\_PERIOD, VISIT\_OCCURENCE and VISIT\_DETAIL |
| Measurements and assessments | Labs, vital signs, input/outputs, transfusions, other Tier 1 flow sheet elements | MEASUREMENT, OBSERVATIONS |
| Diagnoses and problems | Admission diagnosis, Discharges diagnoses, | CONDITION\_OCCURENCE |
| Procedures | X-rays, Dialysis, CRRT, EGD, EEG Bronchoscopies, Specific surgeries. Most of these will have a corresponding note/report) | PROCEDURE\_OCCURENCE |
| Devices | Endotracheal tubes, central lines, pacemakers, EVD, etc. | DEVICE\_EXPOSURE |
| Drugs | Instances of drug administration (as opposed to drug orders) | DRUG\_EXPOSURE |
| Specimens | Cultures, pathology, cytology, etc | SPECIMEN |

Unstructured EHR data

As specified in the exploratory calls with sites, CHoRUS is requesting a tokenized version of the notes, not the notes or reports themselves. Here “tokens” refer to clinical entities (e.g., Diagnoses, Procedures, Medications, etc.) mentioned in text along with associated metadata (e.g., the Document ID, OMOP Athena Concept Code associated with the clinical entity, negation, certainty)

*Final deliverable*

The expected final product is the table note-nlp of the OMOP5.4 Common Data Model. The specific output should follow the following standard <http://ohdsi.github.io/CommonDataModel/cdm54.html#NOTE_NLP>.

*Initial extraction*

The initial extraction will be site dependent but will follow the general process. A list of patients will be generated from a structured EHR data. To each admission corresponds admission and discharge dates. The initial step is to request/perform an extraction of all reports and notes generated during the visit corresponding to these dates. The type of notes/reports to generate pertain

|  |  |  |
| --- | --- | --- |
| Tier 1 | Tier 2 | Tier 3 |
| Notes:   * History and Physical/admission * Progress notes * Medical Consults * OR reports * Discharge summaries   Reports:   * Radiology * Cardiac echo * EEG * Surgical pathology/Cytology | Notes:   * Nursing assessments * Social work * Behavioral health * PT * OT * Nutrition * Wound care   Reports:   * PFTs * EMGs * Other |  |

*De-identification (Optional)*

The second step in the process is to produce a de-identified version of the UEHR data. The process of de-identification may be transparent to investigators, as institutional data scientist may already provide a de-identified version of all UEHR. More likely though, it will be the investigators task to produce this de-identified version. De-identification is the process of producing a version of the note devoid of PHI. If elements of dates are preserved, the result is a “limited” dataset. If elements of dates are not preserved, but dates are time shifted, the resulting dataset is “safe harbor”. This is a general definition of those terms and relevant to all data, not only UEHR.

Investigators are encouraged to find out whether there is an institution-favored method of de-identifying text, or an institutional policy pertaining to verify the result of text de-identification. We note that CHoRUS is not requesting de-identified text. Free software to de-identify text is available. Examples include NLM Scrubber (https://lhncbc.nlm.nih.gov/scrubber/) and UCSF’s Philter (<https://github.com/BCHSI/philter-ucsf>). As these approaches are computational/algorithmic, additional evaluation is required. Such an evaluation would need gold standard annotation of a subset of notes (the i2b2 de-identification task would be a good reference for such an annotation task) and comparison of results. In the absence of a specific institutional policy, it is generally recommended that at least 25 notes of each type be manually reviewed to verify the results of the de-identification method.

At this stage, it is important to become familiar with the OMOP NOTE and NOTE\_NLP table structures. Every note de-identified will require a unique “note\_id” , to be linked with a date, person and an occurrence, as well as a note type and possibly subtypes, at the very least. Thus, in the process of de-identification, enough information must be preserved to link each note with the other data domains.

Importantly, the basic contract data generation sites executed with CHoRUS specified that a limited dataset, with elements of dates preserved, should be delivered to CHoRUS. However, many institutions will not allow this to happen. Please, communicate with CHoRUS whether your contract specifies, and you will be delivering a limited or safe harbor dataset. See additional information on this topic under the structured EHR rubric.

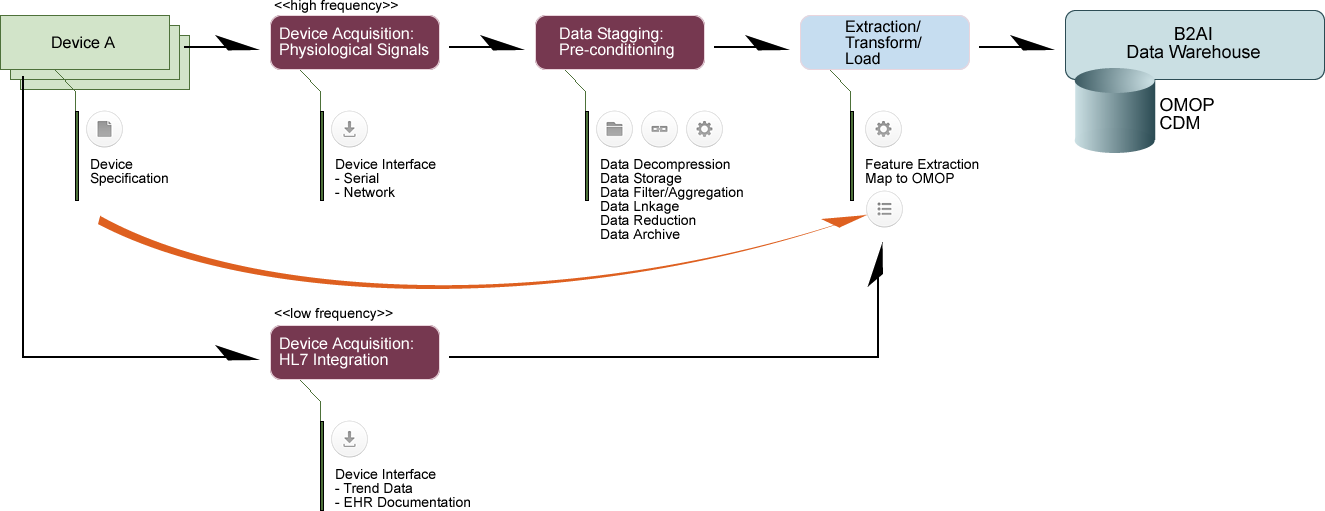
*Producing standard concepts (tokens) from notes – the OHNLP pipeline*

The ultimate UEHR data deliverable to CHoRUS is an OMOP NOTE\_NLP table, which, for each note\_id, will produce possible a large number of concepts. There is an excellent series of office hours put forth by Andrew Wen on the installation or the OHNLP software (<http://ohnlp.org/>; <https://github.com/chorus-ai/OHNLP4CHoRUS>; ), and verification of its performance. It is strongly encouraged that you carefully review these lectures, as well as provided examples.

Waveform data

.

**Overview**



***Figure 1. Generalized Device Data Acquisition Process***

Device data acquisition, standardization, and ingestion into a Bridge2AI OMOP Data Warehouse can consist of two types of data: a) High frequency signals like multi-lead electrocardiogram, EEG, arterial blood pressure, respiratory, and other vitals commonly produced by patient monitors, ventilators, and other physiological signal sensing devices at the patient bedside (reported at greater than 1 sample per second); and b) Low frequency data include trending physiological measurements, EHR captured/documented measurements (hourly blood pressure readings, alarms, alarm settings, etc). High frequencies are often captured through dedicated device gateway integration solutions through direct serial or networked data capture. Low frequency device data are more often captured through HL7 transactions and manual EHR documentation by clinicians and nursing staff.

**Device Meta Data**

Two sets of device meta data will be needed from participating B2AI data acquisition sites representing high frequency data and low frequency data. For high frequency data, meta data may include the following:

**Table 1. High Frequency Device Meta Data (per signal to be contributed)**

|  |  |  |  |
| --- | --- | --- | --- |
| Meta Data | Description | Examples | Notes |
| Label | Physiological signal label available for capture through serial or network connections. | “II ECG”, “respiratory rate”, “heart rate”, etc | Labels are not assumed to comply with any particular standard |
| Source | Device identifier – manufacturer, model, department, floor, bed. UDI if available | Phillips, Drager MediBus, AviaA, Respironics Esprit V60 | Source data can facilitate linking of device to EHR records |
| Sampling Rate | Sampling frequency of the physiological signal | 62 sps, 125 sps, 250 sps, 500 sps, 4000 sps | Respiratory, blood pressure, and other low-frequency signals are typically sampled under 250 sps. EEG, ECG, ABP, and other signals in which morphology analysis is common are sampled at 125 sps or greater. |
| Amplitude Resolution | Amplitude resolution | In terms of analog to digital converter (A/D) units, uV, mV, mmHg, beats per minute, breaths per minute | If A/D units, additional A/D meta data will be required to convert A/D units to standard clinical units |
| Baseline | Signal baseline value (virtual zero) | In terms of analog to digital converter (A/D) units | Normalized signals are adjusted to absolute zero in standard units |
| Amplitude Offset | Signal offset value | In terms of analog to digital converter (A/D) units | Normalized signals are adjusted to absolute zero in standard units |
| Gain | A/D Scale factor | Integer or double value | Additional A/D gain applied to the raw signal – used to convert A/D to standard units of measure |
| Length | Monitoring length | Ms, seconds, minutes, hours | Total monitoring length. Data may include periods of leads off, artifacts, etc. |
| Quality | Overall signal quality | Low, Moderate, High | Site algorithm or methods used to assess and label signal quality. Quality may be associated with time periods. |
| Annotations | Indicate whether signal annotations are provided | Beat location, artifact location, quality annotations | May be generated at the data staging (Fig 1) |

For low frequency data, meta data may be derived from the manufacturer documentation. Manufacturers of physiological medical devices provide a data dictionary or specification document listing the signals and features reported by the device transmitted within HL7 transactions. Data dictionary labels are often used to label OBX segments within a HL7 transaction. Data capture and mapping to OMOP can use these device data dictionaries to form the Feature to OMOP ETL mapping document (see example MIMIC-IV – Appendix A; Phillips iX – Appendix B).

**Table 2. Low Frequency Device Meta Data (per signal to be contributed)**

|  |  |  |  |
| --- | --- | --- | --- |
| Meta Data | Description | Examples | Notes |
| Label | Physiological signal label available for capture through serial or network connections. | “II ECG”, “respiratory rate”, “heart rate”, “Alarm x” | Labels are not assumed to comply with any particular standard |
| Source | Device identifier – manufacturer, model, department, floor, bed. UDI if available | Phillips, Drager MediBus, AviaA, Respironics Esprit V60 | Source data can facilitate linking of device to EHR records |
| Sampling Rate | Sampling frequency of the feature | Per min, per hour, per day, date/time | If not continuous, sampling rate will not apply and measurements/observations will use the HL7 date/time reported in the OBX segment. |
| Amplitude Resolution | Amplitude resolution | uV, mV, mmHg, beats per minute, breaths per minute | Assumes standard clinical units |

Imaging data

Data Processing Steps

1. Identify data extraction partners and methods (e.g., CTSI)
2. Identify the sources of the data (e.g., PACS, site research repository)
3. Determine which image modalities would need to be extracted, date range, and size
4. Confirm the type of data format you will receive (e.g., DICOM)
5. Confirm the type of IDs used to link this data to the patient (e.g., name) or other hospital ID (e.g., MRN)
6. Map local relevant metadata with CHORUS metadata using OMOP format
7. Identify metadata not covered by CHORUS that would be relevant
8. Data deidentification and cross-walk to PHI
9. Data quality review (e.g., data missingness)
10. Upload imaging data to the CHoRUS cloud
11. Upload all data tables and metadata files to your site-specific cloud enclave within the CHoRUS cloud environment.

Imaging Metadata OMOP CDM

Files will be organized hierarchically by patient, modality, study\_UID, and, series\_UID.

**Data Summary by Modality**

We are asking for a summary of radiology occurrence count and radiology image count by imaging modality (e.g., CT, MR). Within those groups, we would like to have a breakdown of manufacturers and body part imaged (e.g., brain, chest).

**Table X. Radiology Occurrence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field | Required | Data type | | Description |
| image\_occurrence\_id (PK) | Yes | integer | The unique key is given to an imaging study record (often referred to as the accession number or imaging order number) | |
| person\_id (FK) | Yes | integer | The person\_id of the Person for whom the procedure is recorded. This may be a system-generated code. | |
| procedure\_occurrence\_id (FK) | Yes | integer | The unique key is given to a procedure record for a person. Link to the Procedure\_occurrence table. | |
| visit\_occurrence\_id (FK) | No | integer | The unique key is given to the visit record for a person. Link to the Visit\_occurrence table. | |
| anatomic\_site\_concept\_id (FK) | No | integer | Anatomical location of the imaging procedure by the medical acquisition device (gross anatomy). It maps the ANATOMIC\_SITE\_SOURCE\_VALUE to a Standard Concept in the Spec Anatomic Site domain. This should be coded at the lowest level of granularity. | |
| wadors\_uri | No | varchar (max) | A Web Access to DICOM Objects via Restful Web Services Uniform Resource Identifier on study level. | |
| local\_path | Yes | varchar (max) | Universal Naming Convention (UNC) path to the folder containing the image object file access via a storage block access protocol. (e.g., \\Server\Directory) | |
| image\_occurrence\_date | Yes | date | The date the imaging procedure occurred. | |
| image\_study\_UID | Yes | varchar (250) | DICOM Study UID | |
| image\_series\_UID | Yes | varchar (250) | DICOM Series UID | |
| modality | Yes | varchar (250) | DICOM-defined value (e.g., US, CT, MR, PT, DR, CR, NM) | |

**Table X. Radiology Image**

|  |  |  |  |
| --- | --- | --- | --- |
| Field | Required | Data type | Description |
| image\_feature\_id (PK) | Yes | integer | The unique key is given to an imaging feature. |
| person\_id (FK) | Yes | integer | The person\_id of the Person table for whom the  the procedure is recorded. This may be a system-generated code. |
| image\_occurrence\_id (FK) | Yes | integer | The unique key of the Image\_occurrence table. |
| table\_concept\_id | Yes | integer | The concept\_id of the domain table that feature is stored in Measurement, Observation, etc. This concept should be used with the table\_row\_id. |
| table\_row\_id | Yes | integer | The row\_id of the domain table that feature is stored. |
| image\_feature\_concept\_id | Yes | integer | Concept\_id of standard vocabulary—often a LOINC or RadLex of image features |
| image\_feature\_type\_concept\_id | Yes | integer | This field can be used to determine the provenance of the imaging features (e.g., DICOM SR, algorithms used on  images) |
| image\_finding\_concept\_id | No | integer | RadLex or other terms of the groupings of image  feature (e.g., nodule) |
| image\_finding\_num | No | integer | Integer for linking related image features. It should not be interpreted as an order of clinical relevance. |
| anatomic\_site\_concept\_id | No | integer | This is the site on the body where the feature was found. It maps the ANATOMIC\_SITE\_SOURCE\_VALUE to a Standard Concept in the Spec Anatomic Site domain. |
| alg\_system | No | varchar  (max) | URI of the algorithm that extracted features,  including version information |
| alg\_datetime | No | datetime | The date and time of the algorithm processing. |