GIRA Data Pull Implementation Guide

Spring 2025 Data Refresh

Table of Contents

I. Purpose & Scope

The purpose of this document is to guide eMERGE sites in pulling the outcomes from EHR (GIRA) data elements from their electronic medical records for the eMERGE IV conditions and participants. Sites should use this guide to follow the data dictionary and perform quality control using this checklist to ensure the cleanest data is delivered to the CC (the CC will not be cleaning data). When outcome analysis is done later, it can be decided if certain values should be removed. Data will be collected using OMOP standardization as well as using custom variables.

Between 10 (minimum) and 20 patients (depending on how many patients are in the datasets) should be randomly checked by each site so the medical record can be reviewed to make sure meds, labs, codes, and visits are looking accurate before submitting to PheKB.

Please also keep in mind that this data will be uploaded to eMERGE's record counter application, Tanagra. In order to upload data files to Tanagra, they must not contain any errors. Please follow the notes for each data type in this document and in the QC checklist to ensure the CC received the cleanest data from each site so that it is able to be uploaded to Tanagra.

All outcomes from EHR data files will be uploaded to this data pull's specific PheKB page (phenotype #1720). It is very important that all data files from all sites are titled using format: **Site_DataType_Date (example: VUMC_Person_20230324)** and as a .csv file. The date will allow for proper version control as edited files are added to the PheKB page. Unless a new "implementation" is added to the PheKB page from each site for each new upload, it will not reflect the date the file is uploaded making it imperative that the data pull (or upload) date be reflected on each data file's name.

PheKB will do an initial quality control check against the data dictionary for each data file (demographics, labs, meds, etc). Please edit all data files to reflect the data dictionary exactly so that data files have zero errors when uploaded to PheKB. A programmer at the CC will review all data files once they are uploaded to check for additional errors. All checks that will be done by the programmer are described below in bulleted form for each data file. Data files will be returned to sites with data quality documents for corrections to be made as needed.

II. PERSON

Notes & Quality Control

- Participant demographics will be collected including EMERGE_ID, WITHDRAWAL_STATUS, YEAR_OF_BIRTH, GENDER_CONCEPT_ID (collected for quality control purposes), RACE_CONCEPT_ID, and ETHNICITY_CONCEPT_ID.
- EMERGE_ID is the unique de-identified ID formatted as SITE_ID + 5 digit randomized and will be downloaded directly from R4 by the CC and provided to all sites.
    - All eMERGE IDs included in the PERSON file must exactly match the R4 ID list provided to all sites with the initial data pull request (no IDs should be duplicated).
    - The R4 ID list will also include withdrawal status.
    - Internal mapping between eMERGE ID and MRN may need to occur at sites but only the eMERGE ID will be used when returning the data.
- Gender, race, and ethnicity OMOP concept IDs will be used for this data collection. If needed, please translate any gender, race, and/or ethnicity data values to those stated in the values columns of the PERSON data dictionary.
- Please see the PERSON checklist and confirm all boxes are checked before uploading the data file to PheKB. This includes confirming no ages are outside of the 1900-2024 range and only encoded values for gender, race, and ethnicity are used.

Example Data

EMERGE_ID	WITHDRAWAL_STATUS	YEAR_OF_BIRTH	GENDER_CONCEPT_ID	RACE_CONCEPT_ID	ETHNICITY_CONCEPT_ID
2711111	1	1990	8507	8516	38003564
2722222	1	1956	8532	8515	38003563

III. MEASUREMENT

Notes & Quality Control

- Participant lab data will be collected including EMERGE_ID, AGE_AT_EVENT, MEASUREMENT_CONCEPT_ID, VALUE_AS_NUMBER, VALUE_AS_TEXT, RANGE_LOW, RANGE_HIGH, RANGE_FLAG, UNIT_CONCEPT_ID, UNIT_CONCEPT_AS_TEXT, and ROW_ID.
- Only eMERGE IDs from the R4 ID list (and PERSON file) should be included in the MEASUREMENT file. Please make sure no withdrawn subject data is included in the MEASUREMENT file.
- Only the labs located in file GIRA_Refresh_Labs_20250108 will be collected for participants (all in record during date range 1/1/2017-present).
- AGE_AT_EVENT should be imputed at each site to 3 decimal points for granularity.
    - Age should not be outside of the 0-120 range. Please check for outliers before submitting data.
- Every LOINC name must map to one single LOINC code.
- If a lab has a numeric value, please make sure the text column is left null (Blank=Missing as stated in the DD) and if a lab has a text value, please make sure the numeric column is left null (Blank=Missing as stated in the DD). No single row of data can have both numeric and text values.
- Please make sure that there are no rows in the MEASUREMENT file that do not have either a numeric or text value. If there are no individuals with certain LOINC codes, please leave those out of the data file.
- For the MEASUREMENT_CONCEPT_ID variable, keywords may be used to match LOINC codes to lab names in databases where LOINC does not match lab.
- All lab units should be in lowercase lettering and be free of leading or trailing spaces.
- There should be no duplicated rows in the MEASUREMENT file.
    - Please use ROW_ID to differentiate between identical rows
- Please see the MEASUREMENT checklist and confirm all boxes are checked before uploading the data file to PheKB.

Example Data

EMERGE_ID	AGE_AT_EVENT	MEASUREMENT_CONCEPT_ID	VALUE_AS_NUMBER	VALUE_AS_TEXT	RANGE_LOW	RANGE_HIGH	RANGE_FLAG	UNIT_CONCEPT_ID	UNIT_CONCEPT_AS_TEXT	ROW_ID
2711111	35.785	3007070	95.0		70.0	100.0	2	8840	mg/dl	
2711122	78.934	3016723		7.2%	4.0%	6.4%	3	8753	%	

IV. DRUG EXPOSURE

Notes & Quality Control

- Participant medication data will be collected including EMERGE_ID, AGE_AT_EVENT, DRUG_CONCEPT_ID, DRUG_CONCEPT_NAME, DRUG_RXCUI, RXCLASS_ID, OUTCOMES_CONCEPT, and ROW_ID.
- Only eMERGE IDs from the R4 ID list (and PERSON file) should be included in the drug exposure file. Please make sure no withdrawn subject data is included in the DRUG EXPOSURE file.
- Medication data will be pulled using RxClass codes which have been compiled for each condition in file RxNorm_Class_List_20250423. Please pull all medications that fall under each class listed in the sheet for participants (all in record during date range 1/1/2017-present).
- AGE_AT_EVENT should be imputed at each site to 3 decimal points for granularity.
    - Age should not be outside of the 0-120 range. Please check for outliers before submitting data.
- All medication units should only have lowercase lettering and be free of leading or trailing spaces.
- There should be no duplicated rows in the DRUG EXPOSURE file.
    - Please use ROW_ID to differentiate between identical rows
- Please see the DRUG EXPOSURE checklist and confirm all boxes are checked before uploading the data file to PheKB.

Example Data

EMERGE_ID	AGE_AT_EVENT	DRUG_CONCEPT_ID	DRUG_CONCEPT_NAME	DRUG_RXCUI	RX_CLASS_ID	OUTCOMES_CONCEPT	ROW_ID
2711111	35.785	19019073	Atorvastatin 20 MG oral tablet	617310	C10AA05	Statin	
2711122	78.934	19059793	Metformin 500 MG oral tablet	861007	A10BA02	Drug used in Diabetes	

V. BMI

Notes & Quality Control

- Participant BMI measurement data will be collected including EMERGE_ID, AGE_AT_EVENT, MEAUREMENT_CONCEPT_ID, MEASUREMENT_CONCEPT_NAME, VALUE_AS_NUMBER, UNIT_CONCEPT_ID, UNIT_CONCEPT_NAME, BMI_Z_SCORE, and ROW_ID.
- Only eMERGE IDs from the R4 ID list (and PERSON file) should be included in the drug exposure file.
- Only MEASUREMENT_CONCEPT_IDs provided infile GIRA_Data_Refresh_BMI_Codes_20250108 are included in the BMI data file.
- AGE_AT_EVENT should be imputed at each site to 3 decimal points for granularity.
    - Age should not be outside of the 0-120 range. Please check for outliers before submitting data.
- There should be no duplicated rows in the DRUG EXPOSURE file.
    - Please use ROW_ID to differentiate between identical rows
- Please see the BMI checklist and confirm all boxes are checked before uploading the data file to PheKB.

Example Data

EMERGE_ID	AGE_AT_EVENT	MEASUREMENT_CONCEPT_ID	MEASUREMENT_CONCEPT_NAME	VALUE_AS_NUMBER	UNIT_CONCEPT_ID	UNIT_CONCEPT_NAME	BMI_Z_SCORE	ROW_ID
2711111	35.785	3025315	Hemoglobin A1c	5.6	8753	%	0.45	
2711122	78.934	3004249	Glucose	105	8840	mg/dl	1.12	

VI. ICD

Notes & Quality Control

- Participant ICD data will be collected including EMERGE_ID, AGE_AT_EVENT, ICD_CODE, ICD_FLAG, and ROW_ID.
- Only eMERGE IDs from the master ID list (and PERSON file) should be included in the ICD file.
- All ICD codes in record will be pulled for every participant with the exception of those on the redacted code list included with the initial data pull request by the CC (titled FINAL ICD & CPT Codes for Redaction 20230821).
    - Please see additional notes for many of the codes in the redaction list that impact how the codes are to be excluded from the data files.
- ICD_AGE_EVENT should be imputed at each site to 3 decimal points. All three decimals are needed to keep granularity.
    - Age should not be outside of the 0-110 range.
- There are two numerical choices (9, 10) for the ICD_CODE_FLAG variable. Sites should make sure no other numbers or letters are included in this column.
- There should be no duplicated rows in the ICD file.
    - Please use ROW_ID to differentiate between identical rows
- Please see the ICD checklist and confirm all boxes are checked before uploading the data file to PheKB.

ICD Example Data

EMERGE_ID	AGE_AT_EVENT	ICD_CODE	ICD_FLAG	ROW_ID
2711111	35.785	403.1	9	
2722222	78.934	N18.2	10	

VII. CPT

Notes & Quality Control

- Participant CPT data will be collected including EMERGE_ID, AGE_AT_EVENT, CPT_CODE, and ROW_ID.
- Only eMERGE IDs from the master ID list (and Person file) should be included in the CPT file.
- All CPT codes in record will be pulled for every participant with the exception of those on the redacted code list included with the initial data pull request by the CC (titled FINAL ICD & CPT Codes for Redaction 20230821).
    - Please see additional notes for many of the codes in the redaction list that impact how the codes are to be excluded from the data files.
- CPT_AGE_EVENT should be imputed at each site to 3 decimal points. All three decimals are needed to keep granularity.
    - Age should not be outside of the 0-110 range.
- There should be no duplicated rows in the CPT file.
    - Please use ROW_ID to differentiate between identical rows
- Please see the CPT checklist and confirm all boxes are checked before uploading the data file to PheKB.

CPT Example Data

EMERGE_ID	AGE_AT_EVENT	CPT_CODE	ROW_ID
2711111	11042	43.133	
2722222	11043	76.653	

VIII. VISIT

Notes & Quality Control

- Participant visit data will be collected including EMERGE_ID, AGE_AT_EVENT, VISIT_CONCEPT_ID, and ROW_ID.
- Only eMERGE IDs from the master ID list (and Person file) should be included in the Visit file.
- VISIT_AGE_EVENT should be imputed at each site to 3 decimal points. All three decimals are needed to keep granularity.
    - Age should not be outside of the 0-110 range.
    - Age at event will be the age at the start of the visit.
- Visit data for participants will be collected from 1/1/2017-present using 4 numerical codes for visit type (inpatient, outpatient, emergency, or other). No other values should be in the visit type column.
- Please use the OMOP standard concept IDs in the provider domain which can be found here.
- There should be no duplicated rows in the Visit file.
    - Please use ROW_ID to differentiate between identical rows
- Please see the Visit checklist and confirm all boxes are checked before uploading the data file to PheKB.

Visit Example Data

EMERGE_ID	AGE_AT_EVENT	VISIT_CONCEPT_ID	ROW_ID
2711111	9201	43.133	
2722222	9202	76.653	

IX. ORDERS

Notes & Quality Control

- Participant order data will be collected including EMERGE_ID, AGE_AT_EVENT, ORDER_DESC, ORDER_CONCEPT_NAME, ORDER_TYPE, ORDER_STATUS, ORDER_CLASS, and CANCEL_REASON, and CANCEL_REASON.
- Only eMERGE IDs from the master ID list (and Person file) should be included in the Orders file.
- AGE_AT_EVENT should be imputed at each site to 3 decimal points. All three decimals are needed to keep granularity.
    - Age should not be outside of the 0-120 range.
- An order mapping document titled GIRA_Data_Refresh_Order_Mapping_20250506 will be used. Sites can filter column C by site and find the order description (column B) in the orders from the EHR.

X. REFERRALS

Notes & Quality Control

- Participant referral data will be collected including EMERGE_ID, AGE_AT_EVENT, REFERRAL_DESC, REFERRAL_SPECIALTY, REFERRAL_TYPE, REFERRAL_STATUS, REFERRAL_CLASS, REFERRAL_REASON, and CANCEL_REASON.
- Only eMERGE IDs from the master ID list (and Person file) should be included in the Referrals file.
- AGE_AT_EVENT should be imputed at each site to 3 decimal points. All three decimals are needed to keep granularity.
    - Age should not be outside of the 0-120 range.