❗

# Considerations for Using Data Elements from the EHR

This document presents the collective experience of UF bioinformaticians and research analysts with using data from the UF Health Electronic Health Record (EHR).

As collaborations between the Integrated Data Repository (IDR) and the Center for Data Solutions (CDS) grew, it became a point of frequent discussion how certain data elements were “repeat offenders” in terms of certain pre- and post-processing issues and/or considerations. Therefore, volunteers from the IDR and CDS have broadly categorized those issues, listed data elements of interest, and provided explanations of the root causes wherever possible.

## IDR Data Sources

- UF Health EHR (aka Epic)
- Social Security Death Index
- Tumor Registry
    - Gainesville data refreshed monthly, Jacksonville data refreshed annually
- Other Linkages
    - Trauma Registry
    - Consent2Share Registry
    - CTSI Biorepository
    - Health Street
    - Axium

# Categories

💡 Data Issues

## 1) Sparsely-populated

- Data is collected in the UF Health EHR (EPIC), however not many entries are recorded per patient
- Examples:
    - **Patient-reported data**
    - Newly-introduced data elements
    - Data element is not recorded at every visit (or depends on the nature of the visit)

---

## 2) Incomplete

- Data exists (or may exist) outside of the UF Health EHR
    - We only have data that is recorded at UF Health (other diagnoses, tests or medications etc. from other institutions are not included in our system. Thus may not reflect the full medical history of patients.)
- Examples:
    - **Patient-reported data**
    - Death data
    - Out of UF Health Network elements
        - Healthcare encounters
        - Lab results
        - Imaging results
    - Pre-Epic Data
        - GNV Data started June 2011

        - JAX Data started March 2013

## 3) Inconsistent

- Lack of standardized data formatting in UF Health EHR
- Examples:
    - **Patient-reported data**
        - e.g. non-smoker in a previous encounter, ex-smoker in a later encounter (lack of standardized definitions of non-smoker/ex-smoker)
    - **Categorical Data**
        - Categories may be subject to revision over time (e.g. ethnicity, race)
    - Measurement units (e.g. height, weight, temperature, etc.)
    - Lab values (qualitative vs. quantitative values, numerous units of measure)

## 4) Human Input Error

- Erroneous values due to human and/or input issues present in the UF Health EHR
- Examples:
    - BMI of 100 or 0

> 💡 Communication Issues

## 5) Data Misinterpretation or Discrepancy

- Issues that arise from not being familiar with or knowing the data sources
- Examples:
    - Not all providers (or study staff) have access to the same version of Epic
    - Epic frontend is **not** the Epic backend (IDR analysts pull data from the backend database tables)

    - *Scheduled* diagnosis tests may **not** indicate that the patient has a *confirmed* diagnosis
    - Timeframe of data elements (full medical history vs encounter-based data)

# Data Elements with Special Considerations

- **Social History** (Sparsely populated) (Inconsistent)
    - Alcohol
    - Tobacco
    - Years of Education
    - Drug use
    - Employment History
    - Issues:
        - **Patient-reported data**
        - Unreliable input via Epic's front-end
        - Present only for patients with certain types of encounters where social history would be logged
- **Vitals** (Human Input Error)
    - Issues:
        - Human input error (zero-values or below zero values), missing data due to lack of entry
- **Diagnosis** (Data Misinterpretation or Discrepancy)
    - Issues:
        - *Scheduled* diagnosis tests may **not** indicate that the patient has a *confirmed* diagnosis
            - **Sources**: Problem List, Encounter, Hospital Billing, Professional Billing.

            - “Encounter” source type can include scheduled diagnosis tests

- **Sexual Orientation and Gender Identity** - SOGI (Sparsely populated)
    - Issues:
        - SOGI was recently added to Epic around Mar-Apr 2019, and may not be present in earlier data. Also, SOGI is unreliably logged in EPIC’s front-end.
- **BMI** (Sparsely populated)
    - Issues:
        - Height is often missing from EPIC.
        - Incompatible units may be used (cm, kg, lbs, etc.)
        - Different sources for calculating BMI (weight, height, etc.)
        - Cross tabulation
- **Ethnicity** (Sparsely populated) (Inconsistent)
    - Issues:
        - **Categorical Data**
        - Relatively new, race/ethnicity were previously combined
            - Consequently, Hispanic used to be a "Race" in the past; it is now an "Ethnicity"
        - Ethnicity has a large degree of missing data, and whether ethnicity is recorded depends on the nature of the encounter.
- **Death dates** (Incomplete)
    - Issues:
        - Incomplete if the death date is not recorded in Epic or SSDI
- **Data from other institutions** (Incomplete)
    - Outside lab tests
    - Prescriptions
    - Other diagnoses or encounters from non-UF facilities

    - Outside imaging
- **Respiratory / Airway Data** (Human Input Error)
    - Issues:
        - The data depends on input from nurses or RTs into the EPIC flowsheet data
            - e.g. Intubation stop and start dates may be unreliable, reintubation may not be documented.
- **Prior to Admission (PTA) med list** (Inconsistent)
    - Issues:
        - PTA Medication Lists are currently difficult to capture with respect to specific encounters. Epic stores the current PTA Medication List.
- **Clinical Notes** (Data Misinterpretation or Discrepancy)
    - Smart text
    - Questionnaire (MCQ, structured data)
    - Issues:
        - There are multiple forms of 'notes' in EPIC (physician and other care provider notes, smart text, questionnaires) that are a mix of structured and unstructured data. Asking for 'the notes' requires more specificity to accomplish project goals.
- **Lack of updated death dates for patients** (Incomplete)
    - SSDI death date
    - Epic death date
- **Race** (Inconsistent)
    - Issues:
        - **Categorical Data**
        - Race is not coded per the latest NIH guidelines (American Indian/Alaska Native, Asian, Native Hawaiian or Other Pacific Islander,

Black or African American, White, More than One Race, Unknown or Not Reported)

    - Hispanic used to be a "Race" in the past; it is now an "Ethnicity"
        - Categorization seems to have changed when Epic started

- **Home or Patient Address** (Incomplete) (Human Input Error)
    - Issues:
        - May not be currently updated in the UF Health system due to the last point of contact not in recent years, etc.
        - Patients move and do not update their address. Address is only updated at times of certain medical encounters
- **Location** (Data Misinterpretation or Discrepancy)
    - Hospital location, patient address, etc.
- **Lab Values** (Inconsistent) (Human Input Error) (Data Misinterpretation or Discrepancy)
    - Urine Drug Screen
    - Glucose Lab Tests
        - Some glucose labs are tolerance tests, which are not good measures of glucose levels
    - Fasting Glucose
        - Some fasting glucose labs are simply labeled as "glucose"
    - Issues:
        - Labs like the urine drug screen may have a mix of results. Other labs may be the same lab test but reported in different units.
- **Insurance or Payer Type** (Inconsistent)
    - Issues:
        - Payer type has different hierarchy levels and potential subcategories, which results in inconsistency.
            - Sources:

        - Clinical Encounters
        - Hospital
- **Timeframe of Data Elements** (Data Misinterpretation or Discrepancy)
    - Issues:
        - As opposed to real-life thinking of a patient’s longitudinal history, EPIC contains timestamps of various data elements in different places (full medical history vs encounter-based data).
        - Hospital diagnoses will be coded towards the end of the encounter. Please be wary of this when apply timeframes to diagnoses.
- **Date/Time stamps, phone numbers, hospital rooms vs beds** (Inconsistent)
    - Issues:
        - Formatting is not always consistent, leading to issues with post-processing in programs like SAS and Excel

# Not Available Yet

- Family History
    - May be available via free text or via surrogate ICD-9-CM or ICD-10-CM codes, however both sources are unreliable
    - Questionnaire that patients are asked to fill in clinics
    - Only asked at specific appointments and clinics
- ICD Family History Codes
    - Not regularly used

## Change Log

### 2021-03-04

- Added **[IDR Data Sources]** section
- Separation of **[Data Issues]** and **[Communication Issues]** sub-sections
- **[Problematic]** category to **[Inconsistent]** and **[Human Input Error]** categories
- **[Researcher Misunderstanding]** category to **[Data Misinterpretation or Discrepancy]** categories
- **[Categorial Data]** and **[Patient-Reported Data]**
- Aggregate list of all data elements instead of by category
- Added information  on Glucose Labs in "Lab Values" data element
- Added **[Not available Yet]** section