Data Processing in BioSense v2.0 – Welcome

Questions may be submitted via the chat feature. Please try to hold questions until the Q&A portion of the Webinar. Any questions we cannot answer in today's Webinar will be answered and posted to the Collaboration Web Site. Technical questions related to Webinar difficulties may be submitted via the chat feature at any time.

BioSense 2.0
Public Health Surveillance Through Collaboration
https://biosen.se

Data Processing in BioSense v2.0

Mike Alletto

BioSense Redesign Team

July 22, 2014

Meeting Topics

- **Raw data processing**
  - Errors that can prevent files/records from being processed
  - Extraction rules
- **Binning process rules**
  - 100-mile rule
  - 24-hour sliding window

File Names and Protections

- **File names must comply with naming standards**
  - Any change in file naming convention will prevent a file from being processed
  - Historical files must also comply with conventions
  - {2-Letter State}_{Provider Acronym}_{8-Digit Date}_{2-Digit Hour}_{File Number}.{File-Type Suffix}
  - CA_DH_20131028_21_001.hl7
- **Protections must be set to "664"**
  - rw- rw- r--

Facility IDs

- Facility IDs are provided to BioSense in the facility spreadsheet along with other data relating to the facility.
- Facility spreadsheet format should not be modified.
- Facility IDs in MSH4.2 (or 4.1) must match IDs in spreadsheet exactly.
- If a facility ID in HL7 records is not recognized, the record will be placed in the exceptions table.
- Note: Data from the facility spreadsheet are placed in the record within the locker instead of data from the HL7 message.

Patient ID

- Patient ID (PID) is extracted from the first field in the following list:
  1) PID.2.1,
  2) PID.3.1,
  3) PID.4.1,
  4) PID.18.1,
  5) PID.19.1

- If PID is not present, the record will be moved to the exceptions table.

Extraction Rules – Date/Time

- **Earliest date/time is the earliest of the following:**
  - OBX.14.1—Observation
  - PV1.45.1—Discharge
  - PV1.44.1—Admit
  - PR1.5.1—Procedure
  - PID.29.1—Death
  - EVN.2.1—Recorded Event
  - DG1.5.1—Diagnosis
  - MSH.7.1—Message

Extraction Rules – Chief Complaint

- PV2.3.1—Admit Reason
- PV2.3.2—Admit Reason Text
- PV2.3.5—Admit Reason Alt Text
- OBX.5.1,OBX.5.8, OBX.5.9—Chief Complaint
  - LOINC 11292-0 or 8661-1
- OBX.5.1—Triage Notes
  - LOINC 54094-8
- OBX.5.1,OBX.5.2—Diagnosis Impression
  - LOINC 44833-2 or 11300-1

Binning

- **The process of binning**
  - Identify and remove duplicate records
  - Consolidate records into visits
  - Identify syndromes and subsyndromes

- **Two paths for binning are employed**
  - IDC 9/10
  - Free text

Chief Complaint and Diagnosis

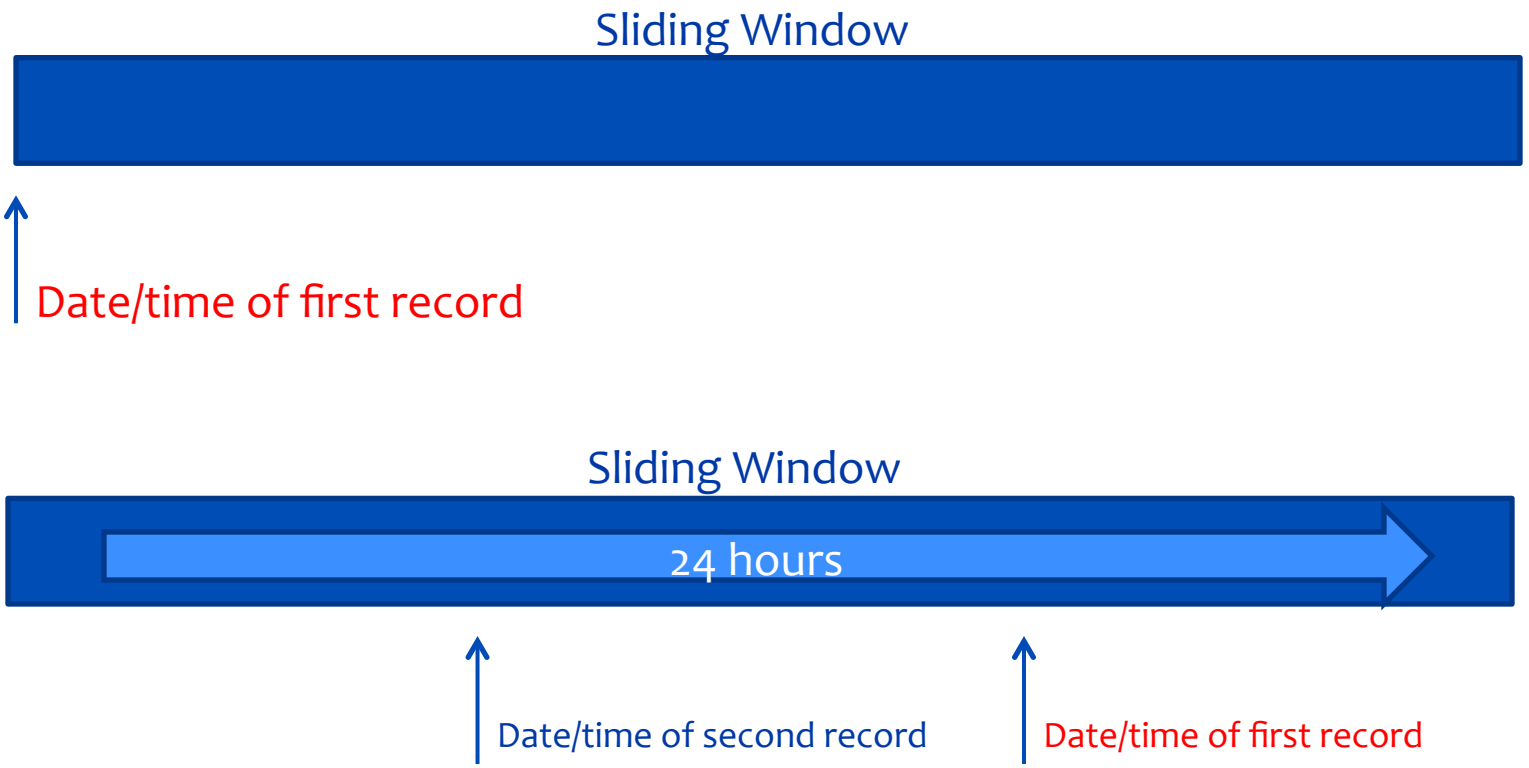
- If DG1 segment is present, then ICD code is used to identify the syndrome.
- If no ICD code is found, then chief complaint (free text) is parsed to identify the syndrome.
- If no diagnosis is present and no chief complaint is present, then the visit is counted but no syndrome is listed.

Special Rules Processing

- **100-mile rule**
  - If a patient's zip code is more than 100 miles away from the facility zip code, then the facility zip code is used to locate the syndromes.

- **24-sliding window rule**
  - Records are grouped into visits by facility, patient ID, and earliest date/time.
  - Date/time is sliding 24-hour window.

24-Hour Sliding Window



Frontend ETL Process

Query ADM

Process Visit Records

Load Into MongoDB

Query ADM

- **Set of four SQL queries against the Analysis Data Mart database:**
  - VA, DOD
  - State/local (RT): CC (chief complaint), DX (diagnosis codes)
- **Result is CSV of visits**

Process Visit Records

- Aggregate data: compute visit totals by syndrome, for all combinations of place, age group, and gender. Via Map/Reduce algorithm.
- Line-level data: captures original ADM visit, but coded by age group, location, jurisdiction, for searchability. Includes age value, CC/DX text, facility info.
- Both cases: assign data owner jurisdiction based on facility location.

Load Into MongoDB

- Delete existing records for the date/source.
- Load in new records.
- Aggregate visit counts drive the visualization (time series & map).
- Line level visits available for CSV export where authorized.

Timing

- Process executes daily at 10:00 AM and 6:00 PM Eastern time, for all sources, for visits today and yesterday.
- Backfill operation loads previous 7 days, starting at 10:00 PM Eastern, nightly.
- Custom date ranges can be reloaded for specific jurisdictions to comply with ad hoc requests.

BioSense 2.0
Public Health Surveillance Through Collaboration
https://biosen.se

Upcoming BioSense v2.0 Webinars

- The BioSense Redesign Team is currently working to plan a Webinar for August. The topic is yet to be determined.

- For more information, please visit our Collaboration Web Site www.biosenseredesign.org

- If you have any suggestions for future Webinars, please contact us at BioSenseProgram@cdc.gov