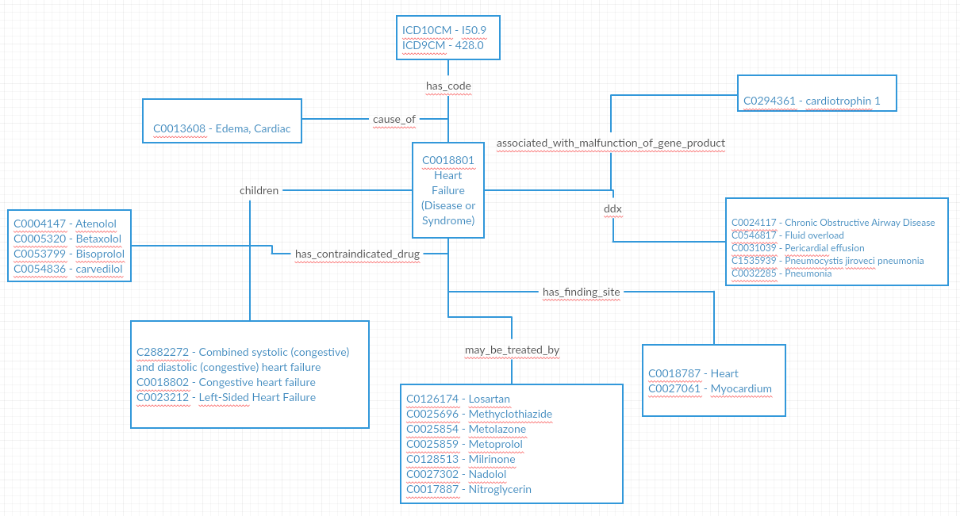
1. Mappings from Clarity + ORIS + Billing + ARKS + other sources to the UMLS (unified medical language system - <http://www.ncbi.nlm.nih.gov/books/NBK9684/#ch02.I21_Overview> )
   1. Mappings -
      1. Dx to ICD9/ICD10/SNOMED
      2. Proc/Surg to CPT/HCPCS
      3. Labs to LOINC
      4. Flowsheets to LOINC
      5. Meds to NDC/NDF-RT/RXNORM
      6. >900,000 terms mapped to custom concepts
         1. CCF providers (Clarity+ORIS+Billing+ARKS) to custom UMLS concepts (I combine all of the prov\_id’s for providers that are the same person)
         2. CCF rooms/beds/locations
   2. Atoms
      1. Contains all synonyms & codes for each ConceptID
      2. Concepts contains the preferred term (typically the most descriptive term)
   3. Relationships/Hierarchies (See appendix 1)
      1. Parent/child relationships
      2. MEDICATION “may\_treat” DIAGNOSIS
      3. MEDICATION “has\_active\_ingredient” PHARMACOLOGIC SUBSTANCE
      4. LAB COMPONENT “analyzes” CELL
   4. Attributes
      1. NDCs
      2. Discrete doses
   5. Custom concepts
      1. Atoms created for both raw value & cleaned term
         1. E.g.
            1. Providers – “has\_specialty”, “part\_of”
            2. Locations – “part\_of”, “has\_location”, “has\_specialty”
      2. >500,000 custom concepts created
2. Clean data
   1. Only 9% of columns are free text; every column possible is mapped to UMLS concepts
   2. Approximately 180 tables from Clarity (Epic) & other sources are used to generate clean database
   3. Hospital visits (Admissions)
      1. Providers/department/service area/room/bed mapped to custom UMLS concepts
      2. Free-text visit reason mapped to UMLS terms
      3. Includes admit/discharge/ED arrival/arrival times
   4. Allergies
      1. Mapped to UMLS terms
   5. Appointments
   6. CurrentMedications
      1. Medications mapped to UMLS terms
      2. Denotes whether or not medications were active at a given encounter
   7. Diagnoses
      1. Combine all sources of diagnosis information into 1 diagnoses table
         1. 28 different sources
         2. Includes free-text diagnoses as well
         3. Primary dx + entry order + present on admission indicators
         4. Lists of ICD9s/ICD10s are split into each individual diagnosis
      2. Also includes a Charlson Comorbidity index calculator
   8. Encounters
      1. Providers/department mapped to custom UMLS concepts
      2. Free-text visit reason mapped to UMLS terms
      3. Encounter type/appointment status/cancellation reason/etc mapped to UMLS concepts
   9. Flowsheets
      1. 4.6 billion rows
      2. All mapped to UMLS concepts
   10. Identities (All patients’ various MRNs)
       1. Identity type mapped to custom UMLS concept for the hospital the MRN came from
   11. Medications
       1. Medications mapped to UMLS terms
       2. Generics mapped to UMLS terms
       3. Order/start/end/discontinue times included
       4. Dose/unit/frequency/route/etc mapped to UMLS terms
       5. Parsed SIGs are available
   12. Notes (500 million rows)
       1. Contains all information on every note in Epic
       2. Can be easily linked to the full note text
       3. >150 million notes
       4. ~20 million procedure results parsed into key-value pairs
   13. Patients
       1. Both standard and difficult to identify test patients are removed via various methods
       2. Death information from SSDI and Ohio Death Index added
       3. Addresses geocoded
   14. Procedures
       1. All mapped to REAL CPT codes
   15. Results
       1. 1.6 billion rows
       2. All lab components mapped to UMLS/LOINC terms which contain –
          1. Component – what is being measured
          2. Property – the characteristic or attribute of the component (e.g. mass concentration, ratio, etc.)
          3. Time – interval of time over which the observation was made
          4. Specimen – the specimen upon which the observation was made
          5. Scale – how to the observation was quantified or expressed (e.g. quantitative, qualitative, ordinal, nominal)
          6. Method – measurement method (e.g. enzyme-linked immunoassay, high power microscope, test strip)
   16. SocialHistory
   17. Surgeries
       1. Both from ORIS (back to 1991) & OpTime
       2. Full ORIS data is stored on 1 set of backup tapes in a format CCF does not support anymore
   18. SurgicalHistory
       1. Free text dates & surgeries are parsed
   19. UniversalChargeLine
3. Other data sources
   1. ARKS/PHDS (Intraop anesthesia data)
   2. DDWMart + Archive, FlashMart (aka billing data)
   3. TSI\_Archive (billing back to 1989)
   4. KP – PROs & QOL measures
   5. HealthQuest (additional PROs)
   6. Foundation Medicine (cancer sequencing)
   7. Critical care transport log
   8. Polysmith – Sleep studies
   9. Provation – GI scope studies
   10. Cancer staging – TNM clinical & pathological + AJCC stage (version 7)
   11. Social security death index
   12. Ohio Death Index (contains causes of death and many other useful variables)
   13. Compiled list of >100,000 abbreviations & acronyms
   14. Census & Wikipedia lists of common first and last names for test patient identification
4. Parsing/Natural language processing (NLP)
   1. Notes parsed into sections & individual measures –
      1. A generic note parser was written to handle almost every type of procedure note
         1. Autopsies, AMET/CMET notes, blood pressures, cancer staging - pathology, cardiac caths, cardiac MRIs, colonoscopies, culture labs, cystoscopies, diagnositic radiology reports, echos, EKGs, endoscopies, GI imaging, heart rates, LVEFs, mammographies, Mohs surgery reports, PAP test results, PFTs, PVRs, renal scans, sleep studies, surgical pathology results, & urodynamics
      2. Can write regular expressions right in SQL server via a CLR we wrote
   2. NLP
      1. We can take any note and extract parts of speech, negation, phrases, word dependencies, identified terms (diagnoses, procedures, medications etc.) which output as UMLS ConceptIDs, conditional statements
      2. Temporal phrase identification with reference date
         1. Reference date = 6/1/2017
         2. Sentence = “On 5/23 the patient had a heart attack.”
         3. Temporal output = DATE – “05/23/2017”
      3. Sentiment analysis
      4. Utilizes LRI’s high power compute cluster to run NLP on ~3 million notes/day
      5. Words/phrases by sentiment/polarity/conditionality/certainty
      6. Temporal phrases
      7. Structured abstracts labels aka section headers
      8. Can run NLP on ~2 million notes per day
   3. Parsed free-text SIGs
      1. >300,000 distinct SIGs parsed
5. DataCoordinator
   1. .NET website & toolset for managing data import/export, mapping terms to the UMLS, validation of mappings, data exploration, etc.
   2. 6.8 million terms mapped to UMLS ConceptIDs
   3. Counts for the number of times identifiers have been used in various data source available for each individual term
   4. >35 billion individual data points which have been translated into UMLS ConceptIDs
   5. SSIS packages for staging, cleaning & mapping data are automatically generated
      1. 600 tables are staged to generate 20 data mart tables
   6. >200 data sources mapped to UMLS ConceptIDs
6. SQL db of various quality measures & their corresponding ICD9/ICD10/SNOMED/LOINC/CPT/etc codes
   1. AHRQ CCS diagnosis & procedure categories (both single & multilevel)
   2. APR to MSDRG crosswalk
   3. CMS ICD9 to ICD10 (and 10 to 9) GEMs
   4. CMS ICD10 to DRG
   5. GPRO ACO codes
   6. HEIDS NDC codes
   7. CMS Hierarchical Condition Categories
   8. NLM value sets
   9. NQF measures
   10. Ohio APR DRG weights
   11. USHIK value sets
7. Geospatial tools (sql-based & web-based for geocoding & connecting to census/American Community Survey data as well as the ability to identify patients living in SNFs/retirement homes)
   1. Including cleaned & validated addresses for all patients’ address changes + CCF facilities
   2. Most health care facility addresses for identifying patients living in a health care facility
   3. Distance calculations
   4. American Community Survey data by census tract/block group
      1. Median income
      2. Education
      3. Employment status
      4. Race/Ethnicity
      5. Citizenship
      6. Housing costs
      7. Poverty status
      8. Language
      9. Family structure
      10. Health insurance coverage
      11. Household vehicles/travel methods
      12. Vacancy status
   5. Weather data (temperatures, precipitation, etc)
8. Other tools utilized
   1. Team foundation server (TFS) for project management and source control
   2. H2O.ai for prediction/machine learning modeling on the fly
   3. Data packager for automation of table generation and pushing of data to external databases
   4. R & Python for various data-related tasks
      1. Can be used right inside SQL server
   5. Tableau for data visualization
      1. Created an oData service to consume REDCap data directly with Tableau
   6. Programmatic import/export of REDCap data
   7. SSIS for ETL & data mapping + query plan analysis & optimization
9. What can be done with the optimized data warehouse?
   1. Live real-time population exploration
   2. Subject matter expert-free protocol development
   3. Data flowing from new sources in minutes
   4. Reduced time from protocol to paper
   5. Easy identification of unutilized data sources
   6. Remove reliance on chart reviews
   7. Less wheel reinvention with predefined quality measures and UMLS relationships/hierarchies
   8. Proactively fix data quality issues
   9. Retroactively fix source data issues
10. Future plans
    1. QHS Outcomes data warehouse to Clarity query translation
    2. OMOP translation
    3. I2B2 integration

Appendix 1 – Sample UMLS Relationships (not a full representation of all available relationships)



Appendix 2 – Data Mappings

