**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Electronic Medical Record (EMR) Data | |
| **Vertical (area)** | | Healthcare | |
| **Author/Company/Email** | | Shaun Grannis/Indiana University/sgrannis@regenstrief.org | |
| **Actors/Stakeholders and their roles and responsibilities** | | Biomedical informatics research scientists (implement and evaluate enhanced methods for seamlessly integrating, standardizing, analyzing, and operationalizing highly heterogeneous, high-volume clinical data streams); Health services researchers (leverage integrated and standardized EMR data to derive knowledge that supports implementation and evaluation of translational, comparative effectiveness, patient-centered outcomes research); Healthcare providers – physicians, nurses, public health officials (leverage information and knowledge derived from integrated and standardized EMR data to support direct patient care and population health) | |
| **Goals** | | Use advanced methods for normalizing patient, provider, facility and clinical concept identification within and among separate health care organizations to enhance models for defining and extracting clinical phenotypes from non-standard discrete and free-text clinical data using feature selection, information retrieval and machine learning decision-models. Leverage clinical phenotype data to support cohort selection, clinical outcomes research, and clinical decision support. | |
| **Use Case Description** | | As health care systems increasingly gather and consume electronic medical record data, large national initiatives aiming to leverage such data are emerging, and include developing a digital learning health care system to support increasingly evidence-based clinical decisions with timely accurate and up-to-date patient-centered clinical information; using electronic observational clinical data to efficiently and rapidly translate scientific discoveries into effective clinical treatments; and electronically sharing integrated health data to improve healthcare process efficiency and outcomes. These key initiatives all rely on high-quality, large-scale, standardized and aggregate health data. Despite the promise that increasingly prevalent and ubiquitous electronic medical record data hold, enhanced methods for integrating and rationalizing these data are needed for a variety of reasons. Data from clinical systems evolve over time. This is because the concept space in healthcare is constantly evolving: new scientific discoveries lead to new disease entities, new diagnostic modalities, and new disease management approaches. These in turn lead to new clinical concepts, which drives the evolution of health concept ontologies. Using heterogeneous data from the Indiana Network for Patient Care (INPC), the nation's largest and longest-running health information exchange, which includes more than 4 billion discrete coded clinical observations from more than 100 hospitals for more than 12 million patients, we will use information retrieval techniques to identify highly relevant clinical features from electronic observational data. We will deploy information retrieval and natural language processing techniques to extract clinical features. Validated features will be used to parameterize clinical phenotype decision models based on maximum likelihood estimators and Bayesian networks. Using these decision models we will identify a variety of clinical phenotypes such as diabetes, congestive heart failure, and pancreatic cancer. | |
| **Current**  **Solutions** | **Compute(System)** | | Big Red II, a new Cray supercomputer at I.U. |
| **Storage** | | Teradata, PostgreSQL, MongoDB |
| **Networking** | | Various. Significant I/O intensive processing needed. |
| **Software** | | Hadoop, Hive, R. Unix-based. |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Clinical data from more than 1,100 discrete logical, operational healthcare sources in the Indiana Network for Patient Care (INPC) the nation's largest and longest-running health information exchange. |
| **Volume (size)** | | More than 12 million patients, more than 4 billion discrete clinical observations. > 20 TB raw data. |
| **Velocity**  **(e.g. real time)** | | Between 500,000 and 1.5 million new real-time clinical transactions added per day. |
| **Variety**  **(multiple datasets, mashup)** | | We integrate a broad variety of clinical datasets from multiple sources: free text provider notes; inpatient, outpatient, laboratory, and emergency department encounters; chromosome and molecular pathology; chemistry studies; cardiology studies; hematology studies; microbiology studies; neurology studies; provider notes; referral labs; serology studies; surgical pathology and cytology, blood bank, and toxicology studies. |
| **Variability (rate of change)** | | Data from clinical systems evolve over time because the clinical and biological concept space is constantly evolving: new scientific discoveries lead to new disease entities, new diagnostic modalities, and new disease management approaches. These in turn lead to new clinical concepts, which drive the evolution of health concept ontologies, encoded in highly variable fashion. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | | Data from each clinical source are commonly gathered using different methods and representations, yielding substantial heterogeneity. This leads to systematic errors and bias requiring robust methods for creating semantic interoperability. |
| **Visualization** | | Inbound data volume, accuracy, and completeness must be monitored on a routine basis using focus visualization methods. Intrinsic informational characteristics of data sources must be visualized to identify unexpected trends. |
| **Data Quality (syntax)** | | A central barrier to leveraging electronic medical record data is the highly variable and unique local names and codes for the same clinical test or measurement performed at different institutions. When integrating many data sources, mapping local terms to a common standardized concept using a combination of probabilistic and heuristic classification methods is necessary. |
| **Data Types** | | Wide variety of clinical data types including numeric, structured numeric, free-text, structured text, discrete nominal, discrete ordinal, discrete structured, binary large blobs (images and video). |
| **Data Analytics** | | Information retrieval methods to identify relevant clinical features (tf-idf, latent semantic analysis, mutual information). Natural Language Processing techniques to extract relevant clinical features. Validated features will be used to parameterize clinical phenotype decision models based on maximum likelihood estimators and Bayesian networks. Decision models will be used to identify a variety of clinical phenotypes such as diabetes, congestive heart failure, and pancreatic cancer. |
| **Big Data Specific Challenges (Gaps)** | | Overcoming the systematic errors and bias in large-scale, heterogeneous clinical data to support decision-making in research, patient care, and administrative use-cases requires complex multistage processing and analytics that demands substantial computing power. Further, the optimal techniques for accurately and effectively deriving knowledge from observational clinical data are nascent. | |
| **Big Data Specific Challenges in Mobility** | | Biological and clinical data are needed in a variety of contexts throughout the healthcare ecosystem. Effectively delivering clinical data and knowledge across the healthcare ecosystem will be facilitated by mobile platform such as mHealth. | |
| **Security & Privacy**  **Requirements** | | Privacy and confidentiality of individuals must be preserved in compliance with federal and state requirements including HIPAA. Developing analytic models using comprehensive, integrated clinical data requires aggregation and subsequent de-identification prior to applying complex analytics. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Patients increasingly receive health care in a variety of clinical settings. The subsequent EMR data is fragmented and heterogeneous. In order to realize the promise of a Learning Health Care system as advocated by the National Academy of Science and the Institute of Medicine, EMR data must be rationalized and integrated. The methods we propose in this use-case support integrating and rationalizing clinical data to support decision-making at multiple levels. | |
| **More Information (URLs)** | | Regenstrief Institute (http://[www.regenstrief.org](http://www.regenstrief.org)); Logical observation identifiers names and codes (http://[www.loinc.org](http://www.loinc.org)); Indiana Health Information Exchange (http://[www.ihie.org](http://www.ihie.org)); Institute of Medicine Learning Healthcare System (http://www.iom.edu/Activities/Quality/LearningHealthcare.aspx) | |
| **Note:** <additional comments> | | | |

**Note: No proprietary or confidential information should be included**