NLP Workshop
JUNE 15, 2017

Lessons learned from NLP implementations at FDA

Mitra Rocca

FDA, CDER, OTS

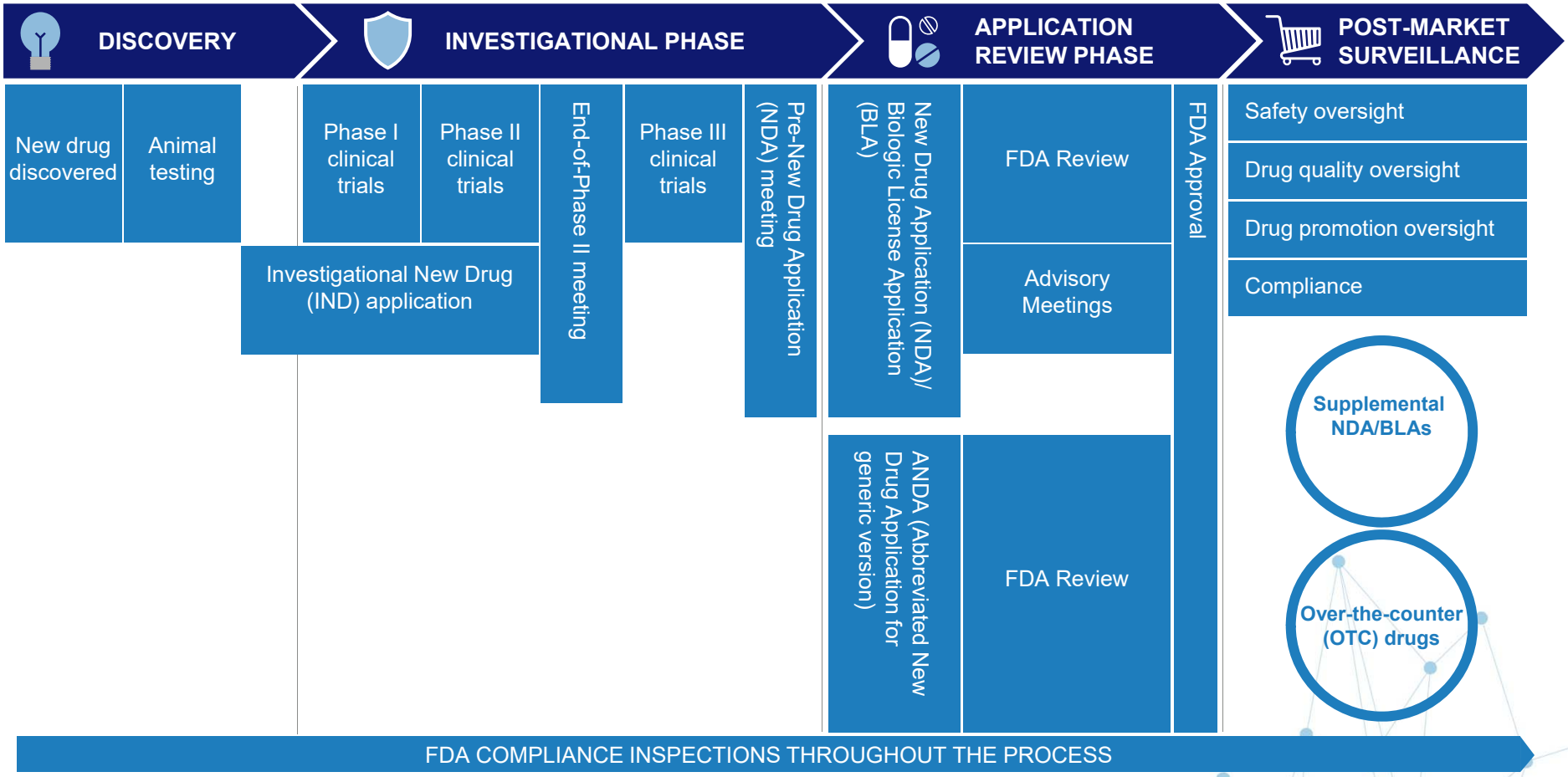
Topics

- Types of Data at FDA
- Examples of NLP Applications and Lessons Learned

Types of Data at FDA

- Investigational New Drug (IND) Application Information
- New Drug Application (NDA) / Biologics License Application (BLA) and 510K Information
- Post-Marketing Information
- FDA Regulated Medical Products Properties
- Sponsor Information
- Clinical Information
- Regulations and Guidances
- Other Miscellaneous Data

Data Submitted and Reviewed in the Drug Lifecycle



Center for Drug Evaluation and Research (CDER)

NLP Application and the DASH Database

Overview

DASH is a regulatory science database that contains marketing application information derived from CDER review and regulatory documentation.

Approach

- Apply NLP tools to extract information from the files
- Map terms to controlled terminologies (e.g. RxNorm, SNOMED CT and LOINC)
- Populate the DASH database proactively

Challenges /Lessons Learned

- Data needs to be put into a format that is analyzable
- High level of accuracy is required

Key Players

CDER OTS Knowledge Management Team

Exploring Adverse Event Term Normalization to MedDRA for Safety Data Reported to ClinicalTrials.gov to Support Pharmacovigilance

Overview

Explore automated methods for mapping Adverse Drug Events (ADE) safety information reported to ClinicalTrials.gov to the Medical Dictionary for Regulatory Activities (MedDRA) by analyzing data extracted from the Aggregate Analysis of ClincalTrials.gov database

Approach

USAGI, a JAVA-based, open source, natural language processing tool developed by the OHDSI community was applied to a subset of clinical trials extracted from the CT.gov results database to map adverse event (AE) terms to MedDRA terms.

Key Players

CDER OTS Data Mining Team

Challenges /Lessons Learned

Limited AE term resolution: The team manually mapped a subset of AE terms to MedDRA that could not be resolved using the USAGI NLP tool

Representing Regulations and Guidance in RDF(Reg2RDF)

Overview

Build a searchable resource of Title 21 is the portion of the Code of Federal Regulations (21 CFR) that links to other regulations, guidances and regulatory processes.

Approach

Leverage Unstructured information management architecture (UMIA) Apache Solr NLP algorithm for

- Processing documents at indexing time
- Processing queries at searching time

Key Players

CDER OTS, OND, OSI

Challenges /Lessons Learned

A tremendous amount of information and knowledge is dormant within unstructured data.
Challenges:
Ability to extract knowledge from the massive heterogeneous data sets and providing 'actionable intelligence'

R-Based, Web-Based Tools to Map Biomedical Text to SNOMED CT

Overview

Identify submissions and applications related to a particular indication in preparation for Advisory committee, Congressional inquiries, workshops, Office level requests. Current search tools require extensive manual effort to obtain this information.

Approach

Develop a Web-based tool that maps indication verbatim text to SNOMED CT:

- Maps strings to the hierarchical (parent and child) relationships and synonyms established in the **SNOMED CT**
- Real-time natural language processing (**NLP**) and a similarity measure (SIM)
- Interactive R-based and Shiny-based applications, which are approved for use at FDA

Key Players

CDER OND

Challenges /Lessons Learned

- NLP does not overcome many errors in text (missing data, spelling errors, punctuations,…)
- Significant iterative manual curation of errors is required upfront
- Many key medical concepts in the indication do not map to a term in the SNOMED CT (e.g., imaging terminology)
- SNOMED CT has many duplicated and redundancy terms

NLP Application to the FDA Labels

Overview

Focus on extraction of Adverse Drug Reactions (ADRs) from the prescription drug labels in Health Level Seven (HL7) Structured Product Labels (SPL).

Approach

Apply NLP tools to automatic extraction of ADRs from drug labels.
Two possible applications will be enabled:

1. Compare the ADRs present in labels from different manufacturers for the same drug, and
2. Perform post-marketing safety analysis (pharmacovigilance) by identifying new ADRs not currently present in the labels.

Challenges /Lessons Learned

Ongoing project:

https://bionlp.nlm.nih.gov/tac2017adversereactions/

Key Players

CDER OND and OSE

Overview: CDER Patient-Centered Outcomes Research Trust Fund (PCORTF) Project: Source Data Capture from EHRs: Using Standardized Clinical Research Data

Overview

- Demonstrate the EHR-to-EDC single-point data capture approach in a ISPY-2/3 oncology trials, using standards
- Leverage mHealth technologies for capturing electronic patient reported outcomes (ePRO)

Approach

Application of NLP to unstructured narrative text in pathology reports and other reports in EHRs such as: discharge summaries, operative notes and procedure Summaries, radiology reports, history and physicals and progress notes.

Challenges /Lessons Learned

High level of Accuracy is required for extraction of EHR data into eCRF

Key Players

FDA CDER OTS and OSP, USCF

OneSource

"enter the right clinical data once, use many times"

OneSource Core & Apps

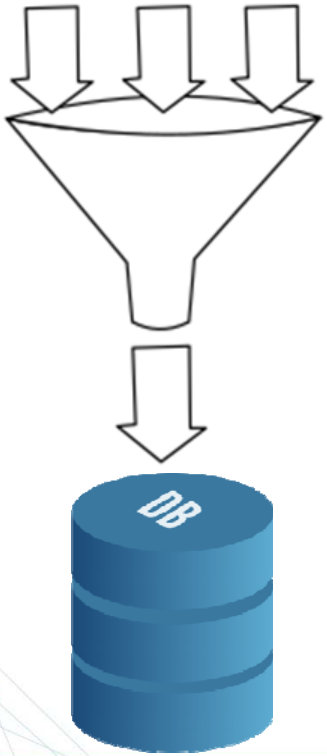
Enter the 'right' data once
Using dynamic checklists for data capture and rendering

Good quality clinical care, clinical trials, registries, quality improvement, researchers, scientists, payors, regulators and others all require the same data elements...

Registries
Clinical Trials
Trial Matching
Regulators
Decision Support
Biomedical Research
Payors
Ancillary Patient Care Systems
Biopharma & Device Orgs
Home Care
Athena
Patient Engagement
Personalized Care
WISDOM
Clinical Efficiency

Vision: Enter Once; Use Many

data from clinical care



Uses

Coordination of Patient Care

Checklists for consistency, correctness, quality, and safety

Facilitate compilation of medical records

Review of patient history with only relevant data

Adapted from UCSF eSource Checklist Overview for OS-ASPE

Center for Biologics Evaluation and Research (CBER)

Processing of Safety Reports
Event-based Text-miner of Health Electronic Records (ETHER)

Overview

- Deconstruction of Adverse Event descriptions:
  - ✓ Extracts **clinical features** (e.g. diagnoses) from safety reports.
  - ✓ Extracts **time information** and **associates** it with the **clinical features**.
- Summarization of Adverse Event information:
  - ✓ Creates textual and tabular summaries.
  - ✓ Visualizes the temporal associations of the extracted features.

Key Players

CBER

Approach

ETHER uses natural language processing to summarize pertinent information in an AE report

Challenges /Lessons Learned

The challenges in implementing automated duplicate classification, which only reinforces the intention to deploy the initial version of the algorithm in a semi-automated fashion, where medical reviewers at the FDA will be provided with lists of cases worth considering for potential linkage in advance of aggregate analyses.

Patient-Centered Outcomes Research Trust Fund (PCORTF) Collaborative Project – CDC and FDA

Overview

- To develop a Natural Language Processing (NLP) Workbench on a shared web service platform.
- Provide access to open source machine learning tools needed to develop and share language models and other algorithms that map unstructured clinical text to standardized coded data.

Approach

- Develop a Natural Language Processing (NLP) Workbench that utilizes Web Services for analyzing unstructured clinical information.
- Cancer pathology data and surveillance data for blood products and vaccines will be used as pilot domains to demonstrate the functionality of the NLP Workbench Web Services

Challenges /Lessons Learned

Development of an NLP Web Service freely available to state public health agencies and researchers which will result in improving the quality of data available to them.

Key Players

CBER

PANACEA – Support Advanced Analysis Build and Analyze Report Networks

Overview

Pattern-based and Advanced Network Analyzer for Clinical Evaluation and Assessment (PANACEA) is used to create report networks. PANACEA uses network analysis to construct networks made of "element nodes" such as exposures (drugs and vaccines) and outcomes (adverse events), or networks made of nodes representing cases (AE reports)

Approach

PANACEA uses NLP and network analysis to facilitate the analysis of clinical data by deconstructing text into components such as products, diagnoses, symptoms, and time stamps.

PANACEA automates the translation of terms into standard medical dictionaries such as MedDRA and ICD; and by visualizing data and aiding in pattern recognition

Key Players

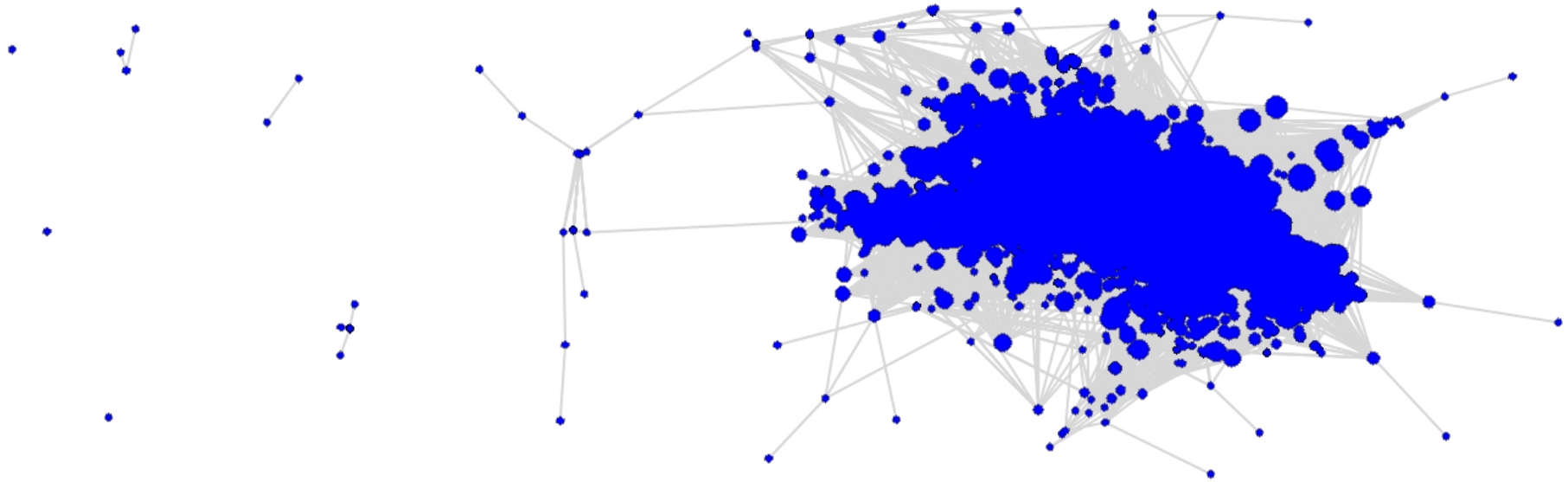
CBER

Challenges /Lessons Learned

- When applied to AE report data, the tools allow reviewers to rapidly evaluate potential associations between products and AEs.
- CBER continues to use, evaluate, and iteratively improve tool functions and algorithms.

PANACEA – Support Advanced Analysis

Build and Analyze Report Networks

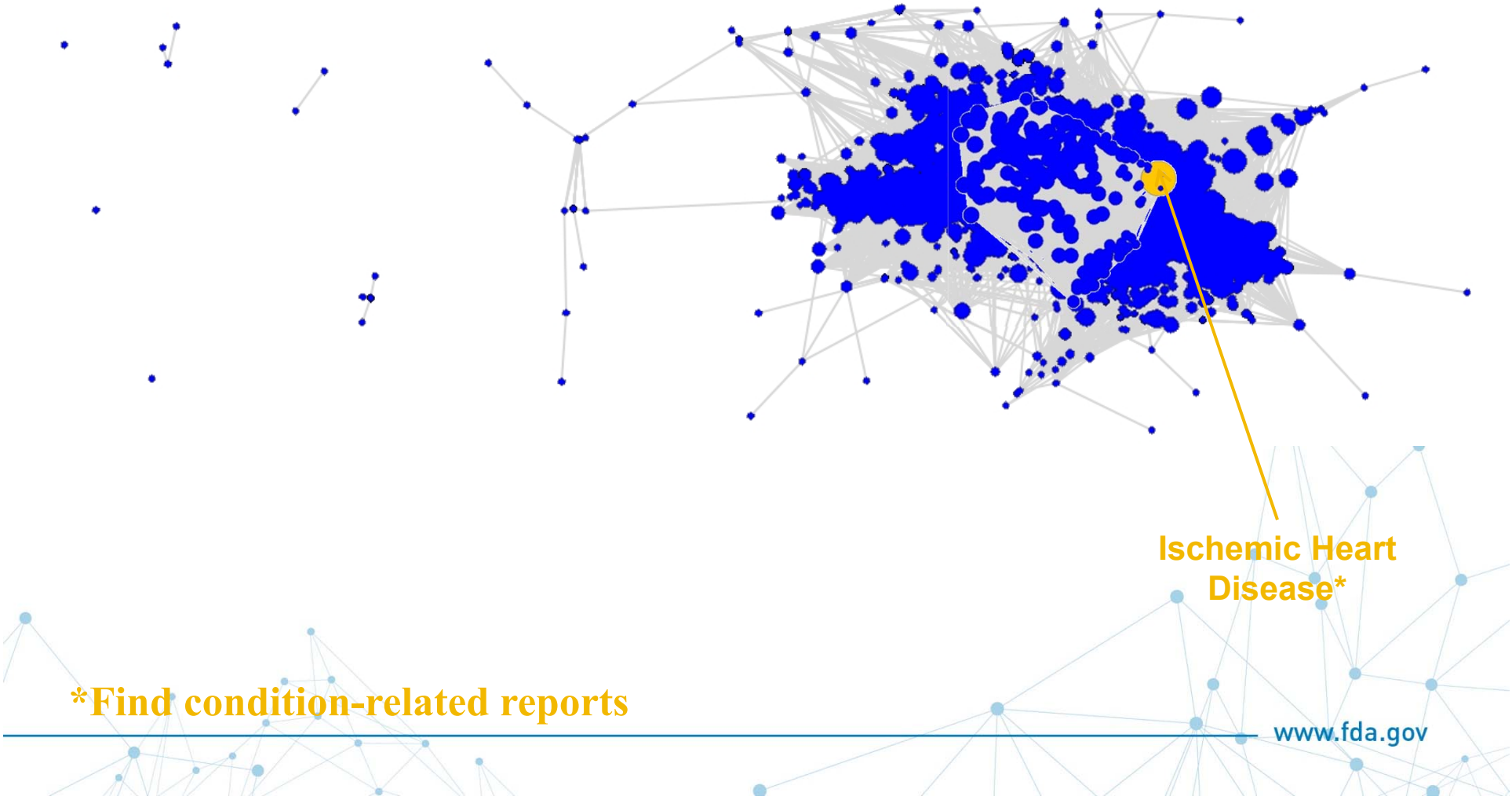


Pattern-based and Advanced Network Analyzer for Clinical Evaluation and Assessment (PANACEA) is used to create **report networks** :

- **Dots** and **lines** represent the **reports** and their **connections**, respectively.
- The **connections** are created when two reports contain the same information.

PANACEA – Support Advanced Analysis

Build and Analyze Report Networks



*Find condition-related reports

Center for Devices and Radiological Health (CDRH)

NLP Applications at CDRH

Overview

CDRH applies various NLP tools to classify unstructured text.

Approach

The CDRH team applies NLP for solving various classification problems in order to identify the causality in a chain of events.
Example: Reuse of endoscopic devices in patients
Apply NLP tools to study causality in scenarios, where the device might be contaminated and infection is transmitted from one patient to another patient. Identify the causality: ineffective sterilization (human factor), manufacturer design issues, etc.

Challenges /Lessons Learned

NLP helps with identifying cases and with the classification problems where chain of events and sequence matters.

Key Players

CDRH

National Center for Toxicological Research (NCTR)

Application of NLP in next generation sequencing data analysis

Overview

Next-generation sequencing (NGS) technologies have provided researchers with vast possibilities in various biological and biomedical research areas. Major gaps currently exist in NGS data analysis and data interpretation.

Approach

Built a framework to pursue data mining and analyze the genetic and biomarker identification on NGS datasets by topic modeling.

Key Players

NCTR

Challenges /Lessons Learned

- The large amounts of data produced by NGS technologies present a significant challenge for data analysis and interpretation.
- Efficient data mining strategies can be applied and are in high demand for large scale comparative and evolutional studies on the NGS datasets.

Data mining for safety surveillance of the FDA adverse event reporting systems

Overview

Focus on the data mining and safety signal detection of FDA drug adverse events.

New safety signals were identified when comparing with the currently available information in various sources.

Drug groups derived from the drug associations were generated which can be used to predict potential adverse event.

Approach

The NCTR team has collected about 10 years' reports from FAERS database, and a total of 63082 drug adverse event pairs were identified as the significant association between 936 drugs and 10316 adverse events. A random network algorithm was applied and 14 drug groups were obtained.

Key Players

NCTR

Challenges /Lessons Learned

- The application of NLP in data mining enables deep analysis and better understanding of FAERS database.
- Topic modeling could be advantageously applied to the large datasets of biological or medical research.

Big data analysis on drug-induced cardiovascular disorders and assessment of the racial/ethnic disparities and gender differences

Overview

Assess the potential differences of diverse minority populations and sex differences that have the drug-induced Cardiovascular disease (CVD) by big data analysis using multiple-layered medical information from diverse resources.

Data on the association between drugs and gender/race/ethnic-differentiated CVD risk factors is sparse and will be identified in the proposed study.

Approach

Novel statistics model development to modify disproportionality analysis and Empirical Bayes Geometric Mean (EBGM) analysis used for Gender/racial/ethnic-differentiated adverse drug events in adverse event sources of data (EHRs, FAERS, PharmaPendium, …) .

Challenges /Lessons Learned

The analysis of large amounts of multivariate data to discover the hidden patterns and the relationships between patterns presents big challenges in both analysis methodology and data interpretation.

Key Players

NCTR
CDER
CVM
VA/UAMS

Center for Tobacco Products (CTP)

NLP Applications at CTP

Overview

CTP provides market authorization for the tobacco products and applies NLP for:

- Clustering the ingredient data from regulated industry.
- Analyze documents (internal marketing, excel spreadsheet, published article, emails, …)

Key Players

CTP

Approach

Application of topic modeling and K-clustering as tools for data mining and analysis.

Challenges /Lessons Learned

Apply NLP to simplify the review process

Summary

- Natural language processing can enhance regulatory science.
- High level of Accuracy is required when applying NLP tools to clinical and non-clinical unstructured data for regulatory decision making.
- Ability to operationalize the NLP tools at FDA and implementing appropriate platforms.

Acknowledgements

- *Robert Ball*
- *ShaAvhrée Buckman-Garner*
- *Fred Sorbello*
- *Rashedul Hasan*
- *Joe Tonning*
- *Christopher Leptak*
- Leposava Antonovic
- Jonathan Stallings
- *Mark Walderhaug*
- Kory Kreimeyer
- *Taxiarchis Botsis*
- *Isaac Chang*
- *Wen Zou*
- *Weizhong Zhao*
- *Deborah* Sholtes
- Thomas Heiman

Thank You

Backup Slides

R-Based, Web-Based Tools to Map Biomedical Text to SNOMED CT Terminology

Jonathan D. Stallings, Military Fellow, OND, IO
Leposava Antonovic, Science Policy Analyst
Chris Leptak, Associate Director Regulatory Science Program

Problem: Divisions are routinely asked to identify submissions and applications related to a particular indication in preparation for Advisory committee, Workshops, Office level requests, Congressional inquiries. Current search tools are antiquated and require extensive manual effort.

Solution: Develop a Web-based tool that maps indication verbatim text to a standardized medical dictionary:

- Maps strings to the hierarchical (parent and child) relationships and synonyms established in the **SNOMED CT**
- Real-time natural language processing (**NLP**) and a similarity measure (SIM)
- Interactive R-based and Shiny-based applications, which are approved for use at FDA

Approach:

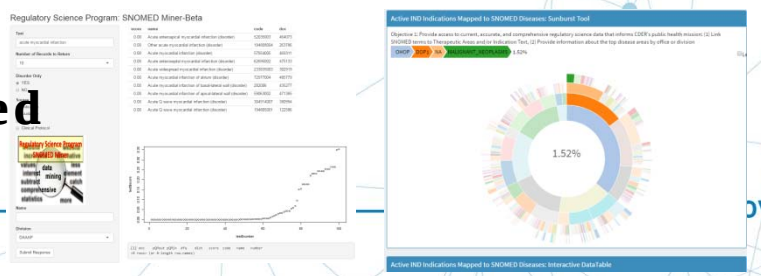
Prior Work Statistics
- SNOMED CT description/relationship files
- *tm* Package: snomed_corpus
- *ngram* package: N-gram library (n = 1-4)
- Corpus Statistics: Importance for each SNOMED term (*It*) and N-gram (*Iw*)

Real-Time Processing
- Natural Language Processing : lower, punctuation, stopwords, whitespace, stemmed, iterative process to correct errors in indications
- *data.table* package: Merge description and ngram tables
- *stringdist* package: Similarity Scores based on Optimal String Alignment

Post-Filters
- DT[sim > 0.01] # perfect match of ngram
- DT[min(CONCEPTID), by = ngram] # lower CONCEPTID
- DT[min(SCORE), by = ngram] # lowest score bias longer

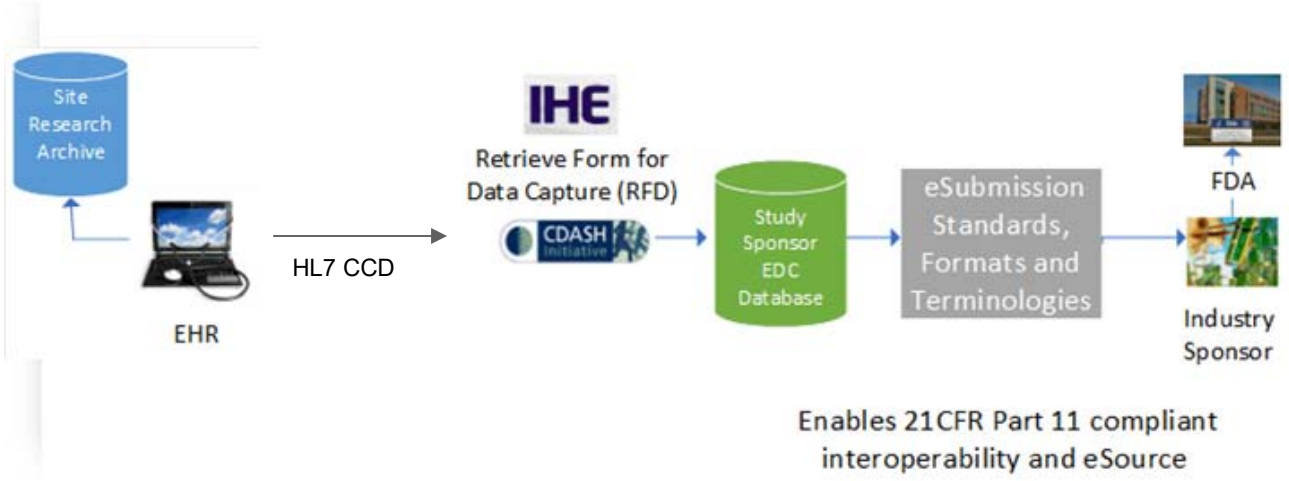
Optional-Filters
- Relationships via SNOMED Hierarchy files: preferred, synonyms, parents, children
- *stringr* package: DT[str_detect(Disorder, findings, situations, organisms, morphologic abnormality)]
- Reduce Intersect of Terms, upper triangle

Web-Based Shiny Tools



Source Data Capture from EHRs Project Goals:

- Demonstrate the EHR-to-EDC single-point data capture approach, using the RFD solution in a FDA-regulated clinical research environment.
- Provide the PCOR researchers with a cloud-based, HIPAA and 21CFRPart 11-compliant tool to seamlessly integrate EHR and EDC systems.



Advanced Methodologies – Why?

Spontaneous Reporting Systems – Processes

Reporter

1. State	2. County where administered	3. Date of birth	4. Patient age	5. Sex
MD	USA	mm dd yy		■ M

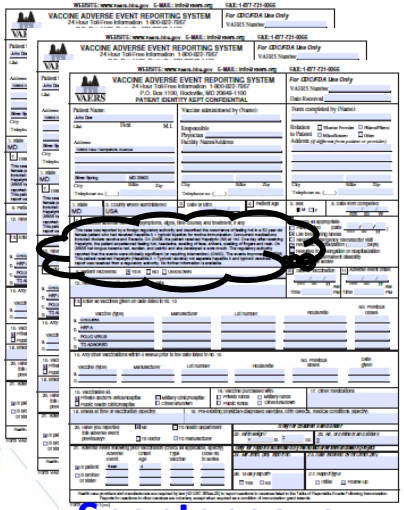
7. Describe adverse events(s) (symptoms, signs, time course) and treatment, if any

This case was reported by a foreign regulatory authority and described the occurrence of feeling hot in a 52 year old female patient who had received hepatitis A + typhoid injection for routine immunization. Concurrent medications included cholera vaccine and Revaxis. On 2/4/05, the patient received Hepatyrix (IM) at 1ml. One day after receiving Hepatyrix, the patient experienced feeling hot, headache, swelling of face, shivers, swelling of fingers and rash. On 2/6/05 her tongue became red, swollen, and painful and she developed a sore mouth. The regulatory authority reported that the events were clinically significant (or requiring intervention) (OMIC). The events improved on 2/7/05. This patient received Hepatyrix (Hepatitis A + Typhoid vaccine) not seperate hepatitis A and typhoid vaccines. This report was received from a regulatory authority. No further information is available.

8. Chec
Patie
Life th
Requ
Requ
Resul
Resul
None

9. Patient recovered ■ YES NO UNKNOWN

10. Date



Spontaneous Reports

Code Assignment

Spontaneous Reporting System

Expert Review

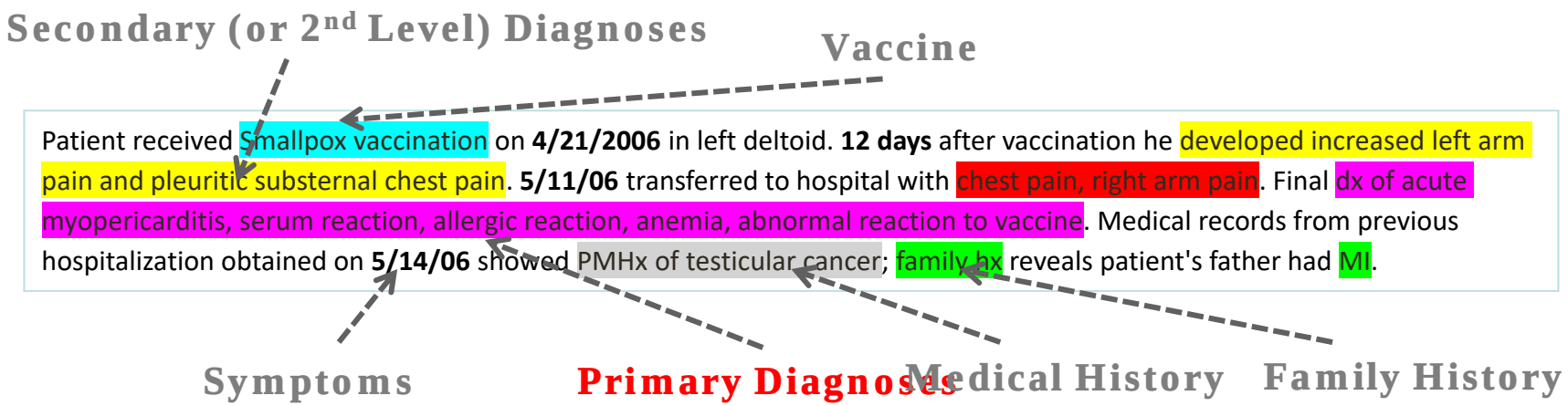
www.fda.gov

Processing of Safety Reports

Event-based Text-miner of Health Electronic Records (ETHER)

- Deconstruction of Adverse Event descriptions:
  - Extracts **clinical features** (e.g. diagnoses) from safety reports.
  - Extracts **time information** and **associates** it with the **clinical features.**
- Summarization of Adverse Event information:
  - Creates textual and tabular summaries.
  - Visualizes the temporal associations of the extracted features.
- Deconstruction supports other tasks:
  - Query-based selection of reports.
  - Conversion of free text to medical codes.
  - Other advanced analysis, such as Network Analysis.

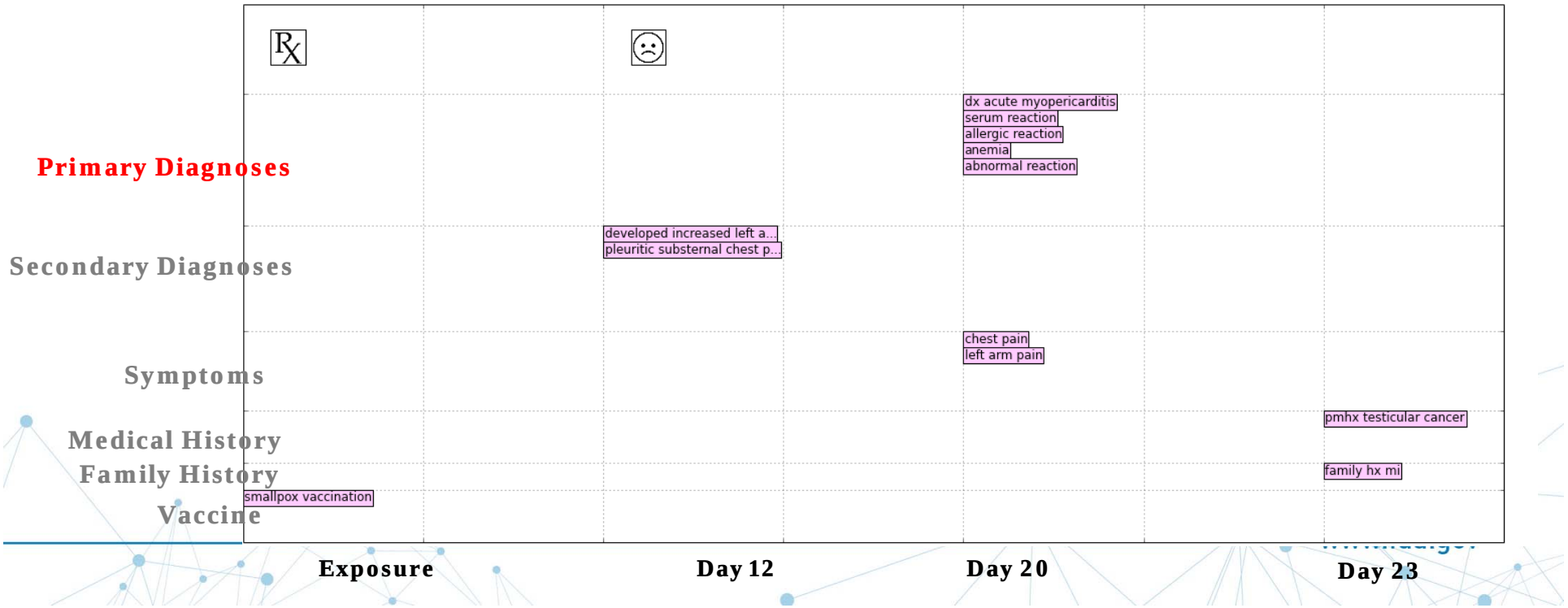
Feature Extraction



Feature Type	Feature Text
Vaccine	smallpox vaccination
Second Level Diagnosis	developed increased left arm pain and pleuritic substernal chest pain
Symptom	chest pain, left arm pain
Primary Diagnosis	dx acute myopericarditis, serum reaction, allergic reaction, anemia, abnormal reaction
Medical History	pmhx testicular cancer
Family History	family hx mi

Visualize Extracted Information with Time Plots

Patient received Smallpox vaccination on **4/21/2006** in left deltoid. **12 days** after vaccination he developed increased left arm pain and pleuritic substernal chest pain. **5/11/06** transferred to hospital with chest pain, right arm pain. Final dx of acute myopericarditis, serum reaction, allergic reaction, anemia, abnormal reaction to vaccine. Medical records from previous hospitalization obtained on **5/14/06** showed PMHx of testicular cancer; family hx reveals patient's father had MI.



ETHER – Information Summarization

Build Structured Summaries for a Group of Cases

Structured Tabular Summaries

Case ID	Age	Sex	Products	Calculated Onset	Diagnosis	Secondary Diagnosis	Medical History	Concomitant Medications
1234	75	M	THYMOGLOBULIN	3 days	died	pyrexia and…	skin cancer, cerebral…	ceftazidime…
2345	45	M	THYMOGLOBULIN	the same day	severe aplastic anemia	pneumonitis	history prostatic hypertrophy	filgrastim
3456	5	M	THYMOGLOBULIN	5 days	pyelonephritis	severe anemia anaemia	history included aortic stenosis	ampicillin
4567	50	M	THYMOGLOBULIN	1 month	pneumonia	cytomegalovirus viraemia		Vancomycin
5678	44	F	THYMOGLOBULIN	1 month and 7 days	acute myeloid leukaemia	unspecified allergic reaction	history right kidney cyst	Busulfan
6789	5	M	THYMOGLOBULIN	8 days		haemophagocytic syndrome		famotidine
7890	71	F	THYMOGLOBULIN	9 days	severe aplastic anemia		history aplastic anemia	methylprednisolone
2341		M	THYMOGLOBULIN	the same day				phenylephrine