ASPE
ASSISTANT SECRETARY FOR PLANNING AND EVALUATION
OFFICE OF HEALTH POLICY

OS PCOR TF
OFFICE OF THE SECRETARY
PATIENT-CENTERED OUTCOMES RESEARCH TRUST FUND

Standardizing Narrative Text for Public Health Research

Development of a Natural Language Processing (NLP) Web Service for Public Health

Narrative Clinical Data is Valuable, but Difficult to Analyze

Narrative text data, such as a doctor's description of a medication plan, is commonly found in clinical reports, pathology reports, and electronic health records (EHRs), and serves as a rich source of information. By definition, unstructured text is not standardized and is therefore difficult for researchers to extract and analyze. NLP offers one solution by training computers and other technologies to understand human language and convert it into standard, coded data. The combination of NLP tools and computer processing means that vast quantities of data can be analyzed faster and more efficiently.

Key Achievements: Creating an Open Source Tool for Narrative Analysis

This project developed an NLP Web Service or workbench, called the Clinical Language Engineering Workbench (CLEW). CLEW is a cloud-based web service that hosts NLP and machine learning tools that researchers can use to convert unstructured clinical data into standardized coded data. Once converted, data can be analyzed for public health and clinical research and surveillance.

The project teams piloted CLEW on two types of clinical information—cancer data from cancer pathology reports and safety surveillance data extracted from medical safety reports—and have made their technical documentation publicly available via **GitHub**. Insights are also shared via publications on **NLP** in general and in **adverse events reporting**, specifically.

Anticipated Impact

Standardized data are foundational to effective and efficient data exchange, linking, analysis, and aggregation of clinical and other data for surveillance, research, and clinical decision-making. The ability to convert information from narrative text to standard, coded data can optimize data quality by improving the accuracy, timeliness, and completeness of the data, thus making the data analysis-ready. CLEW increases the accessibility of NLP technology by providing simple, user-friendly tools for those with varying levels of NLP expertise. CLEW applications can be expanded to other clinical areas outside of oncology and medical products.

QUESTIONS THIS PROJECT ANSWERS:

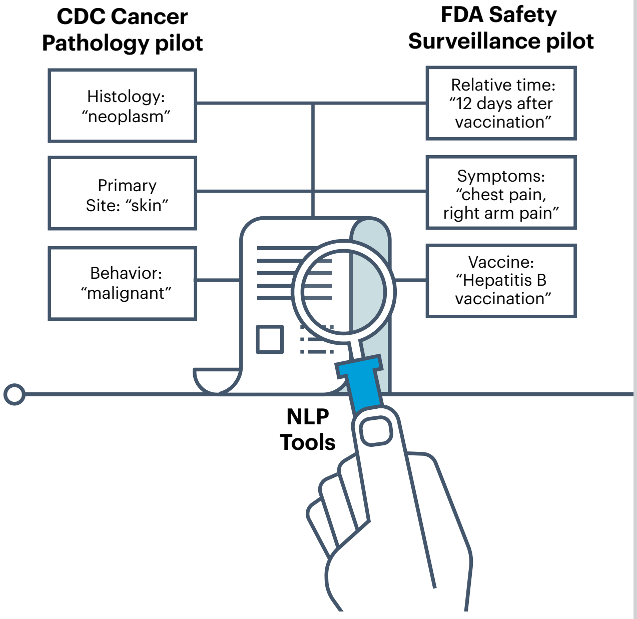
How does NLP technology map cancer pathology information to the International Classification of Diseases (ICD) oncology standard coding system?

How can an NLP system translate unstructured, free-text data submitted to existing agency surveillance systems?

How can NLP help researchers gather temporal and clinical information from unstructured data?

How can this NLP workbench leverage publicly funded data systems for research?

NLP Tools Extract Valuable, Specific Information from Unstructured Text


CDC Cancer Pathology pilot
FDA Safety Surveillance pilot
Histology: "neoplasm"
Primary Site: "skin"
Behavior: "malignant"
Relative time: "12 days after vaccination"
Symptoms: "chest pain, right arm pain"
Vaccine: "Hepatitis B vaccination"
NLP Tools

In the pilots, CLEW extracted information, such as histology, primary site, and behavior, from cancer pathology reports and information related to safety surveillance from medical safety reports.

PROJECT ACHIEVEMENTS:

Publicly Available Products	Brief Description
Clinical Language Engineering Workbench (CLEW)	A cloud-based, open source, NLP Workbench Web Service that hosts NLP and machine learning tools, clinical NLP services, and the opportunity for tool development currently available to researchers on the CDC website.
CLEW Prototype Source Code and Documentation	This product provides the software code and installation instructions for CLEW available on the **CDC** and **FDA's** respective GitHub accounts.
Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of Biomedical Informatics 2017.*	A systematic review of existing clinical NLP systems that generate structured information from unstructured free text, all identified open-source NLP and machine learning tools, frameworks, and systems.
Generation of an annotated reference standard for vaccine adverse event reports. *Vaccine 2018.*	A publication on how to create an annotated dataset for training NLP models.

LEAD AGENCIES:

This project is a collaboration between the U.S. Centers for Disease Control and Prevention (CDC) and the Food and Drug Administration (FDA).

MEET THE PROJECT TEAM:

Sandy Jones
Public Health Advisor
Cancer Surveillance Branch
Division of Cancer Prevention and Control
National Center for Chronic Disease
Prevention and Health Promotion
Centers for Disease Control and Prevention

Mark Walderhaug, Ph.D.
Associate Office Director
Office of Biostatistics & Epidemiology
Center for Biologics Evaluation & Research
U.S. Food and Drug Administration

For more information, visit:
aspe.hhs.gov/development-natural-language-processing-nlp-web-service-public-health-use

ABOUT OS PCORTF

The Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS PCORTF) is a portfolio of approximately 50 projects, including the projects spotlighted in this document, to build data capacity for PCOR on topics such as opioids, value-based care, mortality data, real world evidence, and the interoperability of electronic health records. Managed by the Office of the Assistant Secretary for Planning and Evaluation, the principal advisor to the Secretary of the U.S. Department of Health and Human Services on policy development, OS PCORTF provides for the coordination of relevant federal health programs to build data capacity for comparative clinical effectiveness research, including the development and use of clinical registries and health outcomes research networks, in order to develop and maintain a comprehensive, interoperable data network to collect, link, and analyze data on outcomes and effectiveness from multiple sources including electronic health records. To learn more, visit https://aspe.hhs.gov/patient-centered-outcomes-research-trust-fund. For questions, please email OSPCORTF@hhs.gov.