August 2019

Clinical Language Engineering Workbench (CLEW) Final Report

Contents

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the author's agencies (CDC, FDA). The authors have no conflicts of interest related to this work to disclose.

CLEW Team Members

Team Member Organizations	Team Members
Center for Biologics Evaluation and Researcher Food and Drug Administration	Taxiarchis Botsis – Project Co-Lead Mark Walderhaug – Project Co-Lead Kory Kreimeyer Abhishek Pandey Matthew Foster Richard A. Forshee
Cancer Surveillance Branch Division of Cancer Prevention and Control Centers for Disease Control and Prevention	Sandy Jones – Project Co-Lead Joe Rogers Wendy Blumenthal Temitope Alimi
Northrop Grumman	Steve Campbell – Project Manager Fred Sieling – Project Manager Marcelo Caldas Sanjeev Baral
Health Language Analytics Global (sub-contract with Northrop Grumman)	Jon Patrick
Engility Corporation	Wei Chen Guangfan Zhang Wei Wang
Vassar College (sub-contract with Northrop Grumman)	Keith Suderman

1 Project Overview, Goals, and Objectives

The 2009 American Recovery and Reinvestment Act included multiple measures to modernize our nation's infrastructure, one of which is the Health Information Technology for Economic and Clinical Health (HITECH) Act. The HITECH Act supports the concept of Electronic Health Record (EHR) Meaningful Use, an effort led by the Centers for Medicare & Medicaid Services and the Office of the National Coordinator for Health IT. HITECH proposed the Meaningful Use of interoperable EHRs throughout the United States' health care delivery system as a critical national goal. Because of these initiatives, the clinical care community has moved to EHR systems that utilize standardized common data elements (CDEs). Despite this move, a significant amount of clinical information in the EHR continues to be unstructured textual data, which limits both the primary and secondary usage of the contents of these reports.

The overall goals of this initiative were to:

- Develop a generalized natural language processing (NLP) web service to convert unstructured clinical information to structured and standardized coded data.
- Pilot the NLP workbench web services using cancer data and blood products and vaccine surveillance data.
- Update the NLP workbench web services based on pilot results and provide technical documentation that will describe the requirements for expansion of the NLP workbench web services to meet additional domain needs.

The primary project objective was to provide a mechanism to "translate" free-text data into a structured form that enables the harnessing of clinical prose for surveillance and research purposes. Patient-Centered Outcomes Research (PCOR) researchers, federal agencies, and public health agencies will have access to NLP tools and shared services to translate clinical textual information into standardized formats to enable the use of classical data analytics methods.

2 Background

In the United States, central cancer registries collect, manage, and analyze longitudinal data about cancer cases and deaths. These data are collected from multiple sources such as hospitals, laboratories, physician offices, and independent diagnostic and treatment centers. While there have been strides through Meaningful Use and other activities to implement standardized EHR systems, parts of the medical record, laboratory reports, and other clinical reports still use free-form narratives. Information in this format cannot be used in any form of statistical analysis. Similarly, a considerable amount of clinical information submitted to FDA's spontaneous reporting systems is either not coded at all or not linked to the appropriate Medical Dictionary for Regulatory Activities (MedDRA) terms.

Hospital reporting of cancer cases has been highly standardized over the past decade but as patient care has expanded beyond the hospital setting, central cancer registries have had to

incorporate data from non-standard systems containing large amounts of unstructured data. Over the past decade, CDC has worked closely with the College of American Pathologists to develop synoptic reports for standardized reporting of pathology and cancer biomarker data. The use of these templates would standardize the majority of cancer data that pathology and genetic laboratories report to central cancer registries. However, an incentive does not currently exist for pathology laboratories to change their current laboratory information systems to use these templates; thus, the anatomic pathology and cancer biomarker data reported to central cancer registries continue to be free-form narrative reports.

The data contained in these unstructured pathology reports, biomarker reports, and physician EHR reports are critical for maintaining a complete cancer surveillance system for analysis of population health and adequate patient care. However, the process of abstracting cancer data is labor intensive and expensive, requiring human intervention to abstract critical pieces of information such as histology, primary site, behavior, grade, laterality, and more. The use of NLP tools can reduce the resources needed for analyzing unstructured data and increase the completeness, timeliness, and accuracy of cancer surveillance data, as well as capturing a broader variety of data from these reports.

FDA's Adverse Event Reporting System (FAERS) and Vaccine Adverse Event Reporting System (VAERS) are spontaneous reporting systems in which pharmaceutical manufacturers, medical practitioners, patients, and their representatives submit data regarding the safety of drugs, vaccines, and biologics. The reported information supports important surveillance tasks such as the examination of safety concerns related to marketed products, the evaluation of manufacturer compliance with reporting regulations, and multiple research activities both within and outside FDA. The adverse event (AE) description in VAERS and FAERS data is coded as MedDRA codes. In both systems, a considerable amount of clinical information in the AE narrative is either not coded at all, such as the medical and family history, or not linked to the MedDRA codes, such as the date of onset. In particular, the exact time information for each event (and the code(s) that represent it) is not fully captured from the AE narrative and stored in a structured VAERS or FAERS data field. Both systems also often have multiple reports for the same event, which can affect surveillance. Manual review of each report is often the only way to trace these duplicates. However, after unstructured fields have been transformed into structured data, detection and accounting for duplicate data samples becomes trivial. It is therefore important to develop services for getting the most from the VAERS and FAERS narratives by structuring the unstructured data, improving their quality, and better supporting data requests from the research community.

Objective	Deliverables
Identify existing NLP architectural methods and tools	Conduct an environmental scan and literature review of existing tools, methods, and architectures for consideration.
Use CDC's Innovation Research & Development (R&D) Lab to host NLP Workbench web services	Review architecture of CDC's R&D Lab and work with CDC's Informatics Innovation Unit to develop a plan for NLP Workbench web services architectural design.
Design NLP web service	• Facilitate a requirements-gathering process to direct the architectural design and functionality of the generalized NLP web service ad NLP components. • Write a technical report that includes the architectural design, functional requirements, business rules, NLP algorithms, and any other systems information required.
Develop a pilot version of the NLP web service	The NLP web service will be developed, hosted, and maintained in CDC's Public Health Community Platform (PHCP).
Build structured datasets for piloting the web service	• Gather functional requirements for eMaRC Plus to interface with the NLP web service. • Document the enhancements needed. • Develop a pilot version of eMaRC Plus that will interface with the pilot version of the NLP web service.
Develop a core NLP rule-based approach	Develop source code and a corresponding report, which will be part of the CLEW Technical Report.
Evaluate performance of the pilot version of the NLP web service	Document lessons learned.
Train medical experts	Outline training for the medical experts to use the NLP web service.
Provide guidance to expand the NLP web service	Write a guidance document that explains how to expand the NLP web service to other domains.
Create a protocol to generate testing and validation datasets	Document generation of the testing and validation datasets used to evaluate, update, and make final adjustments to the NLP web service.
Publish the final NLP Workbench web service	Publish the production source code as well as user and technical documentation on GitHub.
Write the final ASPE report	

3 Major Accomplishments

3.1 Environmental Scan (ES)

The NLP ES includes combined results from a systematic literature review with a comprehensive multi-channel review covering researchers and institutions, NLP challenges, and government activities. The ES identified 54 existing open-source tools. We selected the eligible tools on the basis of availability (a tool is open-source, downloadable, and source code exists) and relevance

(a tool supports the processing of clinical texts, such as the ones included in adverse event or pathology reports; it generates standardized and/or coded output; or has advanced capabilities). These tools were subsequently categorized into complete systems, applications, and NLP components and further evaluated on three aspects (development activity, popularity, and framework used) representing the importance and applicability of the tools.

The development of the Clinical Language Engineering Workbench (CLEW) platform was guided by the ES findings and the pilot use cases, with initial focus on selected tasks around the processing of safety surveillance and cancer data. A catalogue of tools that were identified in the ES, identified by stakeholders, or revealed during our ongoing monitoring of the NLP community is included in the CLEW.

The final ES results were published in the *Journal of Biomedical Informatics* (see https://doi.org/10.1016/j.jbi.2017.07.012).

3.2 Design of CLEW Platform

The CLEW is an open platform environment where different techniques can be harnessed for resolving specific use cases. This platform provides clinical NLP services as well as open-source NLP and machine learning tools to develop, experiment with, and refine clinical NLP models. The infrastructure is created for sharing new tools with the wider clinical NLP community, assembling NLP tools into a processing workflow, and generating training files for feeding machine learning algorithms to develop language models. The CLEW was initially developed and hosted on CDC's Innovation Research and Development (R&D) lab for pilot purposes. More information about CDC's R&D lab can be found at www.philab.cdc.gov/index.php/services/cloud/.

See Figure 1 below for a high-level conceptual architectural design of the CLEW. Components of the Language Application (LAPPS) Grid project, funded by the U.S. National Science Foundation, have been leveraged in development of the CLEW architecture. The LAPPS Grid project has made progress toward interoperability of NLP tools and data, as well as creating a comprehensive network of web services and resources within the general NLP community. The CLEW is a service-oriented architecture and supports Simple Object Access Protocol (SOAP) and REpresentational State Transfer (RESTful) services to utilize various NLP tools and modules to engineer clinical NLP pipelines.

Below are functional requirements gathered during a Patient-Centered Outcomes Research Trust Fund (PCORTF) NLP Workbench Stakeholder Meeting webinar held on April 26, 2017 and a face-to-face NLP and Machine Learning Workshop held on December 18–19, 2017, on CDC's campus in Atlanta, Georgia.

- Build on existing efforts.
- Be modular, freely available, and open-source.
- Satisfy multiple use cases and applications for various clinical subdomains.

- Encode clinical data to multiple terminologies.
- Provide interoperability between tools.
- Use a service-oriented architecture using RESTful.
- Identify candidate service consumers and providers.
- Provide an environment that enables the development of an NLP pipeline by combining existing services and tools.

The CLEW provides a wide range of functionality and flexibility depending on the end user's capabilities and expertise with using NLP and machine learning techniques. The CLEW supports three types of end users:

- Novice users: These users have a high-level conceptual understanding of NLP. They can access educational materials on NLP and machine learning techniques and interact with existing services offered on the CLEW.
- Intermediate users: These users have detailed technical knowledge to interface programmatically with pre-defined shared services on the CLEW, submit data to a service, and process expected outputs back to the local application on behalf of its users.
- NLP expert users: These experts can use the LAPPS Grid platform instance in the CLEW to develop and share NLP solutions, pipelines, and tools that address specific issues.

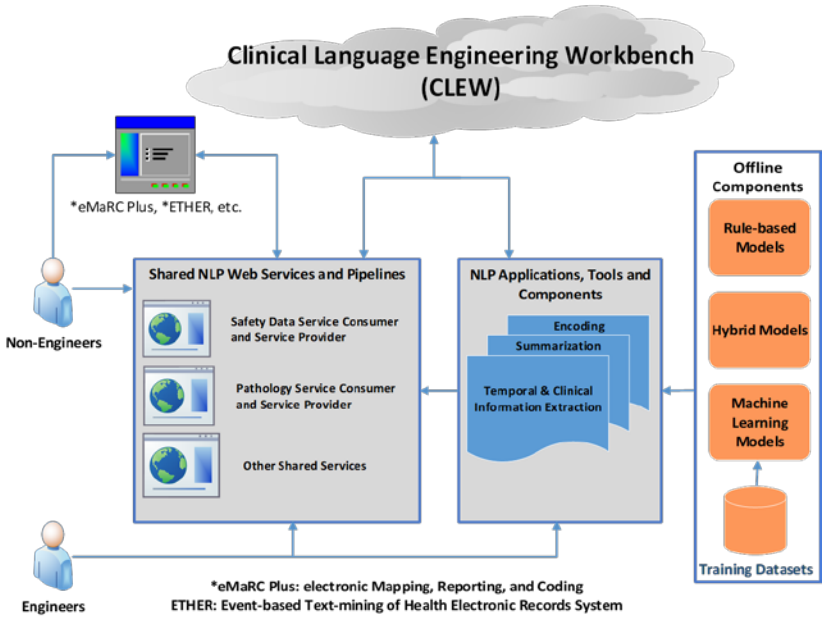

Figure 1. High-level architecture design of the CLEW.

Clinical NLP pipelines and associated services can be catalogued on the Workbench for use as a public service, or, if desired, downloaded and implemented within the user's local environment. The development of the CLEW using the LAPPS Grid is the first attempt to create a collaborative environment for educating, sharing, developing, testing, and implementing NLP and machine learning methods and solutions in the clinical NLP community.

3.3 Pilot Projects on the aCLEW

As part of this project, CDC and FDA completed separate pilot projects to test the use of the CLEW focused on cancer pathology and safety surveillance data, respectively. The CLEW Technical Report provides detailed results from the two pilot projects. Below are some highlights from the pilot use cases.

3.3.1 Cancer Pathology Pilot

The cancer pathology pilot focused on processing unstructured cancer pathology reports for four specific cancer sites: prostate, (female) breast, colorectal, and lung cancers. The goal of the pilot was to use the CLEW LAPPS Grid instance to develop NLP machine learning models to identify five key data elements: body site of the cancer, histology, grade of disease, behavior of the cancer, and laterality (side of the body where the cancer is located). Four national laboratories each provided 125 de-identified pathology reports for each of the four cancer sites, for a total of 500 cancer pathology reports provided per laboratory. Due to a formatting issue with one dataset, the Cancer Care Ontario cancer registry provided a set of de-identified cancer pathology reports to bring the total cancer pathology training corpus dataset to 2,000 reports.

Existing open-source annotation tools were reviewed for possible use, but the Health Language Analytics Global (HLA-G) in-house Visual Annotator tools were considered the best solution for the project. As a result, the HLA-G made a version of the Visual Annotator tool available for use on the CLEW. In order to develop a Clinical Entity Recognition (CER) gold standard, two subject matter experts from the Cancer Care Ontario cancer registry created an annotated training set for use in developing and testing the CER services. An iterative process of annotation and modeling was performed until an acceptable rate of false positives and false negatives was reached.

The CLEW LAPPS Grid was used to develop pipelines for each of the four tools used: Stanford, OpenNLP, GATE, and cTAKES. Each pipeline showed systematic weaknesses, which were corrected by developing pre- and post-processing modules. These modules were incorporated into the CLEW LAPPS Grid for use by others. After the pipelines were finalized to their maximum efficiency, a pathology service was developed for each pipeline. The services address the needs of cancer registries and other clinical researchers to convert text into categorical data or codes. The CER service identifies the same information that may be written differently in pathology reports and outputs standardized terms. The Pathology Coding Service converts the standardized text output from the CER service to standardized national codes defined in the International Classification of Diseases for Oncology 3rd Edition (ICD-O-3). The four CER services can be used to demonstrate differences between different pathology datasets using different pipelines. See Table 1 below for results from the four CER services.

Table 1. Overall precision, recall, and F-score results from four CER services across training batches.

Train Model	Stanford Full Batch Model				OpenNLP Full Batch Model				GATE Full Batch Model				cTAKES Full Batch Model			
Test Batch	P	R	F-Score	N	P	R	F-Score	N	P	R	F-Score	N	P	R	F-Score	N
1	99.44	98.71	99.07	19833	99.46	98.68	99.07	19831	99.42	98.76	99.09	19834	99.57	99.09	99.33	19834
2	95.08	92.54	93.80	17570	96.22	93.14	94.65	16525	93.51	91.01	92.24	17530	92.06	90.40	91.23	17529
4	91.33	90.26	90.79	12699	90.74	88.81	89.76	11099	87.82	86.59	87.20	12716	82.85	83.49	83.17	12709
5	90.99	92.27	91.63	11545	90.93	91.75	91.34	11244	85.95	86.73	86.34	11626	85.08	86.83	85.95	11619
Full	94.93	94.00	94.46	61647	95.26	93.92	94.59	58699	92.8	91.79	92.29	61706	91.21	91.10	91.15	61691

Central cancer registries use an application developed by CDC called electronic Mapping, Reporting and Coding (eMaRC) Plus to receive and process cancer pathology and biomarker data as unstructured narrative data in HL7 version 2 message format. Since 2006, eMaRC Plus has provided integrated rule-based text mining functionality to translate text to code. With improvements in processing unstructured data in the NLP community, expansion in the current eMaRC Plus functionality would allow for implementation of machine learning methodologies to improve the quality of coding. As a project deliverable, eMaRC Plus was enhanced to interface with the CLEW CER and Pathology Coding services to process unstructured pathology data and return coded data for primary site, histology, behavior, grade, and laterality. eMaRC Plus now allows users to use either the rule-based text mining method, the CLEW web services for CER and pathology coding, or both methods of coding the attributes when importing pathology reports. This option allows current eMaRC Plus users to continue using the application to process pathology reports that contain any type of cancer cases, while the CLEW service developers target and fine-tune the language model to target lung, breast, prostate, and colorectal cancers. A feedback mechanism was included in eMaRC Plus for users to communicate any annotation or coding errors identified, so that the CLEW CER and Pathology Coding services can be improved.

3.3.2 Safety Surveillance Pilot

As part of this project, FDA developed rule-based NLP pipelines using a combination of tools and components including cTAKES, BioPortal, and FDA's Event-based Text-mining of Health Electronic Records (ETHER) system. See Figure 2 below for an overview of the different pipelines developed with different tool combinations. All pipelines and tools have been incorporated into the CLEW prototype for use by others.

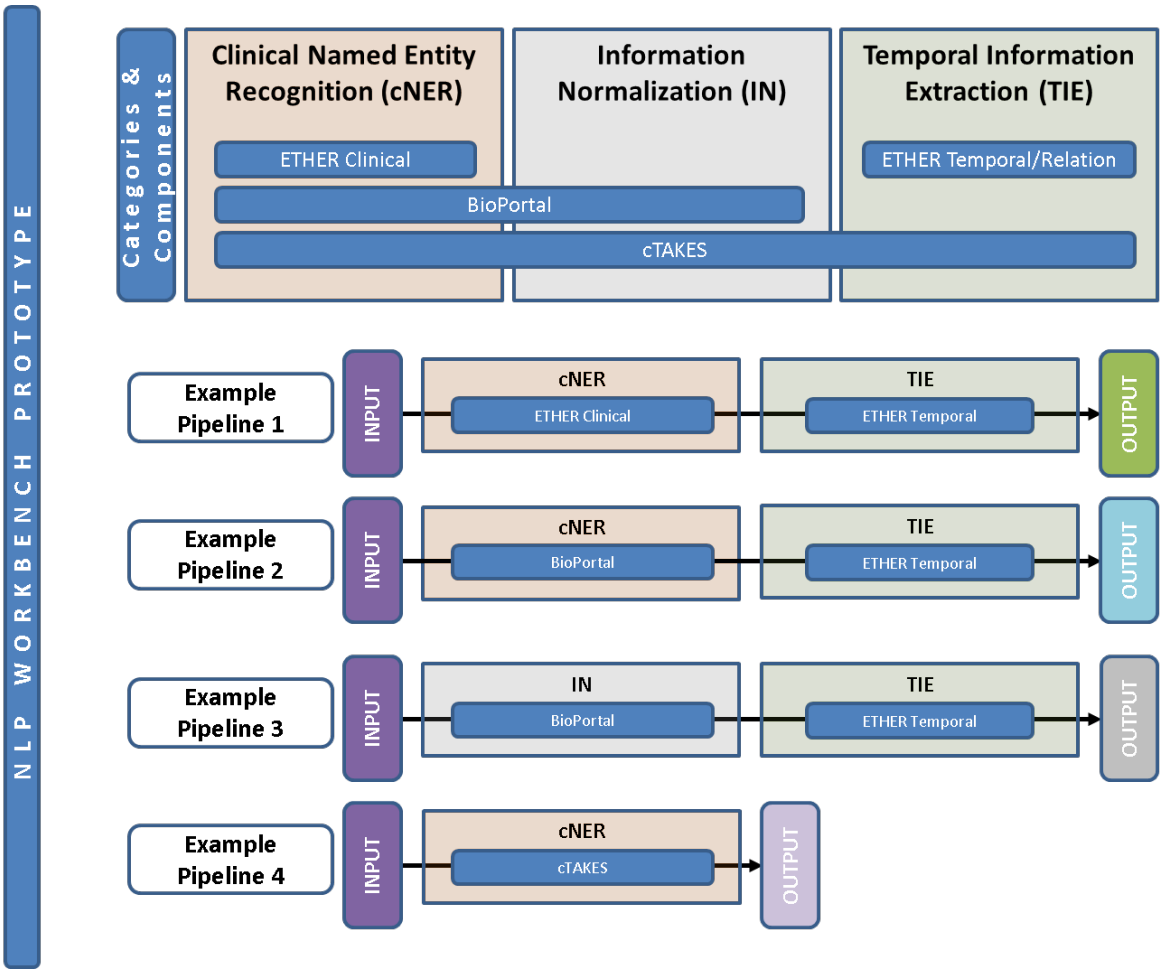

Figure 2. An outline of the CLEW prototype. The integrated tools are shown in the top boxes, where they have been categorized according to their functionality. Four example pipelines are also shown that demonstrate some of the possible combinations of tools that can be made in the CLEW prototype to process input text and produce annotated output.

ETHER was modified to be compatible with the CLEW architecture by separating some of its key components so they could be executed independently. The separate services include the extraction of clinical features, the extraction of temporal expressions, and the creation of temporal associations between features and expressions. Each service requires a plain text input and returns an XML output file listing the identified features, expressions, or associations. See Table 2 below for SER performance evaluation results of pipelines developed using ETHER's rule-based method, Conditional Random Field (CRF) machine learning techniques, and a hybrid approach using both rule-based and machine learning methods for SER.

Table 2. SER performance evaluation results: CRF (FID = 8) model trained on 206 reports and tested with 500 reports; tested with 294 reports otherwise.

	Gold Size	System Size	Recall	Precision	F-Score	Type-Accuracy
ETHER	11216	6900	0.849 9523/11216	0.791 5455/6900	0.819	0.644 6135/9523
CRF	11216	8709	0.769 8626/11216	0.708 6163/8709	0.737	0.664 5731/8626
Hybrid	11216	8726	0.929 10424/11216	0.748 6523/8726	0.829	0.702 7315/10424

4 Lessons Learned

In this section, we will review several critical lessons that were learned from this project. A complete list of lessons is detailed in the Lessons Learned Document developed as part of the project deliverables.

This was intended to be a collaborative development project that would allow staff from FDA and CDC to have equal access to servers and tools required to complete the tasks. When this project was proposed, CDC and Association of State and Territorial Health Organizations' Public Health Community Platform (PHCP) was identified as an appropriate home to develop and host the products from this project. However, security challenges across federal agencies limited the work on the PHCP. To move the project forward, CDC's Research and Development Laboratory was selected as a temporary place to support development and testing activities. Since Research and Development Lab servers are not open to the public, special access permissions were granted to FDA and contract staff to complete this project. For future projects, a permanent cross-agency host server allow for the CLEW to be fully implemented, reviewed, tested, and deployed across multiple domains.

Software development collaboration across U.S. government agencies and other organizations did not meet expectations of the project team including connectivity, available software development tools, server hosting, scalability, administrative support, and usage policy. CDC's Informatics Innovation Unit (IIU) environment encountered administrative delays, limited server hosting options and no plans for scalability. As a solution, we switched to Amazon Web Services (AWS); however, there continued to be challenges.

Many existing open-source tools, frameworks, libraries, and applications would be beneficial to include in the CLEW for development, testing, and sharing. Integrating all of these tools would take more time and effort than was allowed as part of this project. The LAPPS Interoperability Format (LIF) helps in this area, but it was still challenging to integrate UIMA-based tools such as cTAKES into pipelines. Additional work needs to be completed to address interoperability across tools, so that tools and solutions can be used across frameworks to strengthen NLP and

machine learning models and pipelines. The cancer pathology pilot project demonstrated that it is important to develop different pipelines using different tools to evaluate and identify the best model that performs at a level of sensitivity and specificity to meet the need.

5 Publications and Presentations

The CLEW team presented project information at several venues during and after the two-year project period. FDA has published two journal articles, and CDC has one publication in progress.

5.1 Oral Presentations

- OS ASPE PCORTF Education Webinar, June 24, 2019, Presentation Title: **Development of Natural Language Processing (NLP) Web Services**
- InterAgency NLP and Machine Learning Webinar, November 13, 2018, Presentation Title: **CLEW: A Collaborative NLP Workbench**
- AMIA Annual Conference NLP Special Interest Group, November 8, 2018, Presentation Title: **CLEW: Clinical Language Engineering Workbench**
- NLP and Machine Learning Workshop, December 19, 2017, Presentation Title: **Clinical Language Engineering Workbench (CLEW) Demonstration**
- NAACCR 2017 Annual Conference, June 2017, Presentation Title: D**evelopment of Natural Language Processing (NLP) Workbench Web Services**
- Annual Cancer Research Seminar, June 19, 2017, Presentation Title: **Advancing Innovation and Convergence in Cancer Research: Federal Cancer Moonshot**
- OS PCORTF NLP Workbench Stakeholder Webinar, April 26, 2017, Presentation Title: **Natural Language Processing (NLP) Workbench Web Services Stakeholder Meeting**
- NLP and Machine Learning Workshop, December 9, 2016, Presentation Title: **CDC/FDA Natural Language Processing (NLP) Web Service Project**
- North American Association of Central Cancer Registries (NAACCR) 2016 Annual Conference, June 2016, Presentation Title: **Development of a Natural Language Processing (NLP) Web Service for Structuring and Standardizing Unstructured Clinical Information**

5.2 Journal Publications

- Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, Forshee R, Walderhaug M, Botsis T. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of Biomedical Informatics* 2017;73:14–29. DOI: 10.1016/j.jbi.2017.07.012.

- Foster M, Pandey A, Kreimeyer K, Botsis T. Generation of an annotated reference standard for vaccine adverse event reports. *Vaccine* 2018;36(29):4325–4330. DOI: 10.1016/j.vaccine.2018.05.079.

5.3 GitHub Publication of CLEW Source Code and Documentation

- CDC Public GitHub: https://github.com/CDCgov/NLPWorkbench
- FDA Public GitHub: https://github.com/FDA/

6 Future Considerations

Through the ES, this project found that many tools and solutions have been developed across the clinical community to translate unstructured clinical text to structured coded data. It became clear by the middle of year one that this project would not be able to identify the best NLP and machine learning solutions available, but rather find a way to harness all of the valuable work that has been done by the clinical NLP community across multiple organizations. This project found that there were too many solutions available to be able to test and evaluate every possible combination.

This project has developed a valuable platform that will encourage the clinical NLP community to continue to collaborate more and share their work in a central repository. This could enable development of NLP and machine learning solutions to provide more cost-efficient and quality structured data for use across clinical, academic, government, and public health organizations. The CLEW can minimize duplicate solution development and provide additional clinical domains with tools and services that may not have been available to them previously.

The two years for this project was not a sufficient amount of time to fully implement and test all desired functionality. One of the most critical challenges this project encountered was identifying a permanent home for the CLEW. A location that encourages collaboration and sharing of NLP and machine learning solutions across federal agencies, public health, and clinical researchers would allow for full implementation of the CLEW.

The additional work that remains for the CLEW is to:

- Form a Governance Committee that includes representation from across federal agencies, academia, and public health and clinical researchers to oversee the CLEW's development and use.
- Further integrate the LAPPS Interoperability Format (LIF) for clinical NLP to achieve interoperability across NLP frameworks such as UIMA and GATE.
- Expand the CLEW LAPPS Grid instance to include additional services, clinical NLP, machine learning, and deep learning tools, pipelines, and services for use.
- Implement and test use of CLEW LAPPS Grid to develop solutions for other clinical domains.

7 Summary

This project completed a comprehensive environmental scan to identify all existing open-source clinical NLP tools available across different domains. A subset of these tools, chosen for their applicability to specific use cases for safety surveillance and cancer pathology processing, are available for use on the CLEW through a Tools Catalogue. The CLEW platform leveraged components of the LAPPS Grid to address interoperability issues that prevented NLP pipelines from being developed and tested using components of tools from across different NLP frameworks such as UIMA and GATE. The final safety surveillance and cancer pathology NLP pipelines and web service APIs have been made available on the CLEW for use by other researchers. The cancer pathology and safety surveillance end-user applications have been expanded to use the CLEW APIs. The annotated datasets that were developed as part of the pilot projects have been made available for other NLP training activities. A detailed list of lessons learned was provided in a separate document.

The CLEW, including the LAPPS Grid instance, provides researchers with access to tools and datasets to enhance analytic capabilities that will accelerate clinical innovation. The availability of extensible NLP web services can be useful and valuable across agencies and in public health and research communities. The NLP conventions included in the CLEW environment can be used to advance surveillance and research.

Significant resources are spent on NLP and machine learning activities across local, state, and federal public health communities and in the clinical and academic research communities. This project has developed an initial framework that can serve as a central repository to educate researchers and share clinical NLP and machine learning tools, pipelines, and web services. This project provides the resources to extract structured information accurately from narrative data, so the information can be used to develop better public health models to improve patient care.