shaip

CASE STUDY

Oncology Data Precision: Licensing, De-identification, & Annotation for NLP Model Innovation

Revolutionizing Cancer Care with Cutting-Edge NLP Technologies

COMPANY
A cutting-edge data science division of a renowned healthcare leader

INDUSTRY
Healthcare

USE CASE
Improvement of Oncology Research Utilizing NLP and Data De-identification

OUR OFFERING
Data Collection, De-identification, and Complex Annotation Services

KEY STATS

Data Licensing + Data De-id	Non Oncology Domain	Oncology Domain	Oncology Relationships	Negation	NER + Relationship Mapping
10,000 pages	**10,000 pages**	**10,000 pages**	**4500 pages**	**9000 pages**	**1223 pages**

INTRODUCTION

The client, a major player in the healthcare industry, required an advanced NLP solution to process a substantial volume of oncology medical records. As part of a pivotal initiative to refine oncology research, the need to balance detailed data analysis with stringent privacy standards is paramount. This case study outlines our contributions to enhancing the client's research endeavors through high-fidelity data annotation, rigorous de-identification practices, and the application of Natural Language Processing (NLP) techniques, all within the regulatory framework provided by HIPAA.

CHALLENGES

The project required a nuanced understanding of clinical documentation, precise identification of medical entities, and the ability to apply negation labels accurately, all within a secure framework that protects patient privacy according to HIPAA regulations. The endeavor demanded not only technical expertise in handling large volumes of complex data but also a strategic approach to incorporate feedback and maintain quality across all stages of the annotation process.

Detailed Description of Services:

- **Comprehensive Clinical Data Coverage:** Spanning various note types, care settings, and oncological subspecialties, ensuring a robust dataset reflective of diverse clinical scenarios.
- **Rigorous De-identification:** Ensuring all labeled records are de-identified in compliance with HIPAA's Safe Harbor method, assuring client confidence in data privacy and security.
- **Annotation Guidelines:** Creation and implementation of standard data annotation guidelines for preparing Labeled Records in line with HIPAA standards.
- **Advanced Annotation Techniques:** Application of NLP to 10,000 pages of oncology-related records, involving intricate labeling of negation statuses and other relevant details as per prior established guidelines.
- **Rigorous Quality Assurance:** Attain the specified quality standard outlined in the guideline.

SOLUTION

Our approach involved the following key strategies:

Comprehensive Clinical Data Coverage

To tailor the dataset to the client's specific needs, a targeted selection of data was meticulously extracted from Shaip's extensive repository of over 5 million Electronic Health Records. This curated dataset encompassed a variety of note types and care settings, providing a rich and diverse spectrum of clinical scenarios. This ensures a dataset that is not only comprehensive but also highly representative of real-world medical data.

Rigorous De-identification

The process adhered strictly to HIPAA's Safe Harbor method for de-identification, which guarantees the client's confidence in data privacy and security. This involves removing all Protected Health Information (PHI) and replacing it with labeled placeholders, thereby maintaining the utility of the data while protecting patient confidentiality.

De-identification Variables

Category	Subcategory
Name	Patient name, Physician name, Nurse practitioner name, Family member name, Medical center name, Clinic name, Nursing home name, Company name, University name
Age	
Date	Date pattern, Month Year pattern, Day Month pattern, Day Year pattern, Day, Month, Year, Season
Location	Country, State, City, Street, ZIP Code, Room number, Suite number, Floor number
ID	Social security number, Medical record number, Health plan beneficiary number, Account number, Certificate/License number, Biometric id, Record id, Accession number, Vehicle identification number, License plate numberDevice identifiers and serial number
Contact	Telephone number, Fax number, Email address, Web URL, IP address

Example:

On September 25, 2106, at 11:00 am, Mr. Harry Pace, aged 90, was admitted to Forrest General Hospital for a scheduled hip surgery, previously consulted by his primary care physician Dr. Jose Martin, and attended by Kendra Reith, MD. During his stay, he was under the care of Mary Hu, N.P., and Suzan Ray, R.N., with R. Charles Melancon, PA, also being consulted. His operation, conducted on the same day as admission, was successful with no complications reported. Following surgery, Mr. Pace was transferred to Room 202, Floor 2, for recovery. His wife, Emma Pace, was present throughout and was provided with all necessary updates. During his brief stay, his medical records, including MRN MR99062619 and Account KV000014764, were handled according to the standard protocols of Gracewood Nursing Home, his previous residence. He was discharged later the same day to the care of Oakland Outpatient Clinic for further recuperation. Throughout the process, all procedures were documented and secured with adherence to confidentiality standards.

Example: De-identified

On [Date Pattern], at 11:00 am, Mr. [Patient Name], aged [Age], was admitted to [Medical Center Name] for a scheduled hip surgery, previously consulted by his primary care physician Dr. [Physician Name], and attended by [Physician Name] MD. During his stay, he was under the care of [Nurse Practitioner], N.P., and [Nurse Practitioner], R.N., with [Physician Name], P.A., also being consulted. His operation, conducted on the same day as admission, was successful with no complications reported. Following surgery, Mr. [Patient Name] was transferred to Room no. [Room Number], Floor no. [Floor Number], for recovery. His wife, [Family Member Name], was present throughout and was provided with all necessary updates. During his brief stay, his medical records, including MRN [Medical Record Number] and Account [Account Number], were handled according to the standard protocols of [Nursing Home Name], his previous residence. He was discharged later the same day to the care of [Clinic Name] for further recuperation. Throughout the process, all procedures were documented and secured with adherence to confidentiality standards.

Annotation Guidelines & Advanced Annotation Techniques

Shaip was instrumental in establishing and implementation of standard data annotation guidelines ensured that all Labeled Records were prepared consistently and in compliance with HIPAA standards. Moreover 10,000 pages from various medical records were meticulously annotated, with a focus on the detailed labeling of negation statuses and other clinically relevant entities including various oncology subspecialties. The annotation were carried out by a team of expert annotators with specialized knowledge in oncology and data privacy regulations.

Complex Annotation

Category	Subcategory
Date Annotation (Oncology)	Diagnosis Date, Stage Date, Onset, Procedure Date, Med Date Started, Med Date Ended, Radiation Date Started, Radiation Date Ended
Disease (Oncology)	Cancer Problem, Histology, Clinical Status, Body Site, Behaviour, Grade, Cancer Stage, TNM stage, Tumour Marker Test, Dimensions, Code
Treatment (Oncology)	Cancer Medicine, Drug Dosage, Frequency, Cancer Surgery, Surgery Result, Radiation Modality, Radiation Dosage
Genomics	Variation Code, Gene Studied, Method, Specimen
Negation	Negative, Possible Negative, Uncertain, Possible Positive
Clinical NER Relationships	Cancer problem - Body Site, Histology - Body Site, Behaviour - Body Site, Cancer Surgery - Body Site, Radiation Modality - Body Site, Histology - Grade, Cancer Problem - Dimension

Example:

Oncology Clinical Note Statement

"Patient Jane Doe was diagnosed with Stage IIIB non-small cell lung cancer (NSCLC), specifically adenocarcinoma, on 03/05/2023. The cancer is located in the right lower lobe of the lung. It is classified as T3N2M0 according to the TNM staging system, with a tumor size of 5 cm x 3 cm. An EGFR exon 19 deletion was identified through PCR analysis of the tumor biopsy specimen. Chemotherapy with Carboplatin AUC 5 and Pemetrexed 500 mg/m² was initiated on 03/20/2023 and is to be administered every 3 weeks. External beam radiation therapy (EBRT) at a dose of 60 Gy in 30 fractions commenced on 04/01/2023. The patient's treatment is ongoing, and there is no evidence of brain metastases on the recent MRI. The possibility of lymphovascular invasion is yet to be determined, and the patient's tolerance for the full chemotherapy regimen remains uncertain.

Oncology Clinical Note Statement:

Oncology Progress Note

Diagnosis Date: 03/05/2023
Stage Date: 03/12/2023
Onset: 02/01/2023 (date patient first reported symptoms)
Procedure Date: 03/15/2023 (biopsy)
Med Date Started: 03/20/2023 (chemotherapy commencement)
Med Date Ended: Ongoing
Radiation Date Started: 04/01/2023
Radiation Date Ended: Ongoing

Disease (Oncology):

Cancer Problem: Non-small cell lung cancer (NSCLC)
Histology: Adenocarcinoma
Clinical Status: Active treatment
Body Site: Right lower lobe of lung
Behavior: Invasive
Grade: G2 (moderately differentiated)
Cancer Stage: Stage IIIB
TNM Stage: T3N2M0
Tumor Marker Test: Elevated carcinoembryonic antigen (CEA)
Dimensions: 5 cm x 3 cm mass
Code: C34.1 (ICD-10 code for primary malignant neoplasm of right lower lobe of lung)

Treatment (Oncology):

Cancer Medicine: Carboplatin and Pemetrexed
Drug Dosage: Carboplatin AUC 5, Pemetrexed 500 mg/m²
Frequency: Every 3 weeks
Cancer Surgery: Not indicated at current stage
Surgery Result (Cancer surgery only): N/A
Radiation Modality: External beam radiation therapy (EBRT)
Radiation Dosage: 60 Gy in 30 fractions

Genomics:

Variation Code: EGFR exon 19 deletion
Gene Studied: EGFR
Method: Polymerase chain reaction (PCR)
Specimen: Tumor biopsy

Negation:

Negative: No evidence of brain metastases on MRI
Possible Negative: Possible absence of lymph vascular invasion, pending further tests
Uncertain: Uncertain if patient will tolerate full course of chemotherapy
Possible Positive: Possible involvement of the pleura, requires confirmation with imaging

Clinical NER relationships:

Cancer Problem - Body Site: NSCLC located in the right lower lobe of lung
Histology - Body Site: Adenocarcinoma in the right lower lobe of lung
Behavior - Body Site: Invasive tumor in the right lower lobe of lung
Cancer Surgery - Body Site: No surgery indicated for tumor in the right lower lobe
Radiation Modality - Body Site: EBRT targeting the right lower lobe of lung
Histology - Grade: Adenocarcinoma with a Grade of G2
Cancer Problem - Dimension: NSCLC with a mass measuring 5 cm x 3 cm

Rigorous Quality Assurance

Implemented a flexible project management framework that facilitated the effective integration of client feedback while upholding stringent quality standards. A comprehensive quality assurance protocol was enforced, aligning with the guidelines to reach the requisite quality benchmarks. This protocol featured successive rounds of review and verification, securing the precision and dependability of the annotated data. Such meticulous quality oversight is crucial in crafting a dependable NLP solution, vital for informed clinical decision-making and research excellence.

THE OUTCOME

Successfully delivered 10,000 high-quality, De-identified Labeled Records, providing a secure and valuable dataset for the client's NLP model development. The meticulous application of NLP and adherence to HIPAA de-identification standards resulted in a highly refined dataset that will underpin the client's ongoing and future oncology research efforts, ultimately aiming to enhance oncology patient outcomes and care delivery efficiency.

The success of the project illustrates our ability to handle complex medical data with precision, contributing to the client's aim of improving patient care outcomes and accelerating the pace of healthcare innovation.

CUSTOMER TESTIMONIAL

★★★★★

Our partnership with Shaip has been instrumental in advancing our NLP capabilities within the oncology domain. The professional handling of 10,000 medical records, annotated with detailed negation and other clinical entities, demonstrated their commitment to excellence and compliance. Moreover, their commitment to privacy standards like HIPAA have provided us with invaluable resources to drive our AI initiatives of developing a cutting-edge oncological treatments and diagnostics forward.

shaip

Shaip provides high-quality data across multiple data types (text, audio, image & video) to companies looking to build unbiased and high quality AI/ML models. Shaip licenses, collects and annotates data for Healthcare, Conversational AI, Computer Vision and Generative AI/LLM use cases. Going beyond data, Shaip offers a complete Responsible LLM Toolkit to align, evaluate, and enhance large language models using reinforcement learning from human feedback (RLHF). Headquartered in Kentucky with offices in Silicon Valley and India, our global team blends data science expertise with deep industry knowledge. Visit us at www.shaip.com.