



OS13

Leveraging Open-Source Python Tools to Improve Efficiency and Transparency in Clinical Trials

Sohan Lal

Nov2025

www.geninvo.com

North America | UK | Europe | India



Agenda

- Why Efficiency & Transparency Matter
- Why Python?
- Python's Popular Libraries For Pharma
- NLP & Machine Translation Using Python
- Clinical Documents QC/Redaction using AI/ML
- Leveraging Python Libraries To Handle PDF/Docx
- Practical Examples of Multilingual Redaction
- Q&A



Why Efficiency & Transparency Matter

Empowering Patients Through Trust and Safety

- **Informed decisions** can be made about treatments, especially when patients and doctors have access to comprehensive trial results.
- Trust in the healthcare system is strengthened when patients know that data—positive or negative—is not being hidden.

Accelerating Scientific Discovery for Researchers

- English document use as a reference to redact sensitive content in other languages.
- Translating the redacted English content into languages such as German, French, Spanish.

Driving Efficiency and Credibility in the Industry

- **Regulatory alignment:** Agencies like the FDA and EMA increasingly require data sharing, making transparency a compliance necessity.
- **Improved credibility:** Transparency combats the perception of selective reporting or data manipulation.



www.geninvo.com
North America | UK | Europe | India

GENINVO™

Why Python?



- **Open-source and Extensible**
- **Reduces Manual Effort and Automation**
- **Supports Global Collaboration**
- **Accelerates Workflows and Saves time**
- **Machine Learning & Predictive Modelling**
- **Visualization and Reporting**

Python's Popular Libraries For Pharma

Pandas

Numpy

SASPy

PymuPDF2 / Python-
Docx

PyDICOM

NLTK

faker

Spacy

Bert/Mbert

NLP & Machine Translation Using Python

Named Entity Recognition (NER)

- Identifies and redacts sensitive entities (e.g., names, medical conditions, addresses) in documents.
- Ensures compliance with privacy regulations.
- Example: Redacting "John Smith" and "heart disease" in clinical documents.

Semantic Analysis

- Overcomes limitations of traditional keyword matching techniques.
- Avoids over-redaction and under-redaction.
- Understands the underlying meaning of text, ensuring redactions are contextually appropriate.
- Example: In clinical trial documents, terms like "heart disease" or "sensitive medical history" are flagged for redaction, ensuring sensitive data is protected.

Multilingual NLP Models

- Models like BERT process multiple languages, handling grammatical and syntactical differences.
- Essential for ensuring accurate redaction of sensitive data across both original and translated documents.
- Preserves redactions in multilingual environments, ensuring data protection across languages.



Automating Redaction Process Using AI/ML

Redacting Using English as a Reference Point

- English redactions may not translate directly into other languages, requiring systems to identify equivalent sensitive information in translated documents.
- Ensuring that redaction practices in the English version align with equivalent meanings in non-English versions is complex.

Example of Redacted Translation

- **Original English Text:**

"John Smith, a 45-year-old male with a history of heart disease, was admitted."

- **After Redaction in English:**

"[REDACTED], a 45-year-old male with a history of [REDACTED], was admitted."

- **After Machine Translation to German:**

"[REDACTED], ein 45-jähriger Mann mit einer Vorgeschichte von [REDACTED], wurde aufgenommen."

Machine Translation for Multilingual Redaction

- Preserves redactions across languages, ensuring the integrity of sensitive information maintained in translated documents.
- Machine translation tools like Google Translate, DeepL, and domain-specific systems help translate redacted documents into languages with varying sentence structures (e.g., German).
- Ensure redactions remain intact, and context is preserved, despite the differences in syntax.

Technological Solutions: NLP & Machine Translation

Natural Language Processing (NLP)

- NLP enables machines to understand, interpret, and generate human language, including the identification of sensitive information.
- Automates the identification and redaction of sensitive information in documents.
- Processes context, syntax, and semantics to ensure correct redaction across languages.
- Key NLP Components:
 - Named Entity Recognition (NER)
 - Semantic Analysis
 - Multilingual NLP Models



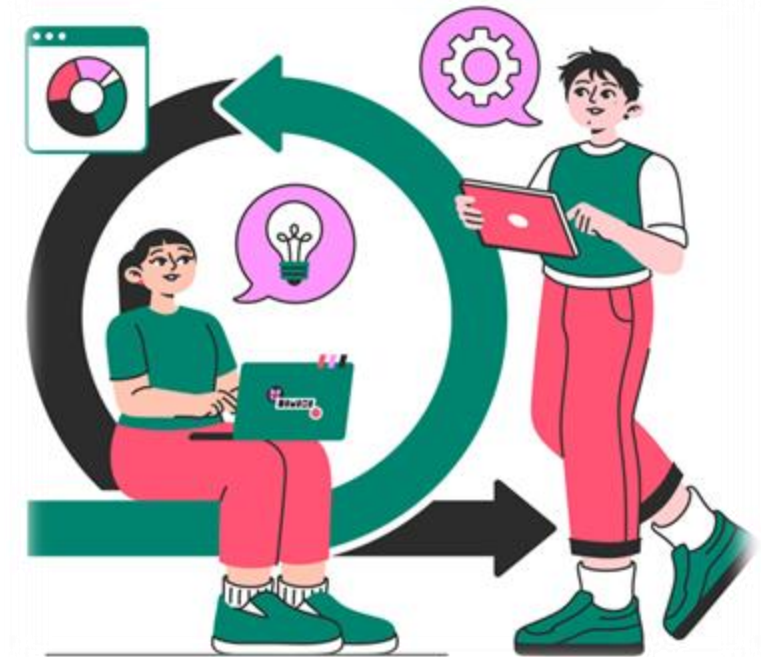
www.geninvo.com
North America | UK | Europe | India

GENINVO™

Technological Solutions: NLP & Machine Translation

Machine Learning NLP Python Libraries

- **python-docx** enables automated creation and modification of Word documents, allowing for the generation of clinical study reports (CSRs), safety summaries, and patient reports. It supports flexible text formatting, tables, image embedding, and metadata extraction for regulatory compliance and version tracking. It can also integrate with NLP tools for extracting key insights from clinical trial documents.
- **PyMuPDF(fitz)** specializes in extracting and transforming PDF content—text, images, tables, and metadata (e.g., title, author, creation date)—into formats like JSON, CSV, HTML, and XML. It supports advanced search and pattern recognition, aiding in precise data extraction and document analysis.



Examples of PDF output

```
import fitz # PyMuPDF

# Open the PDF file
doc = fitz.open("clinical_trial_report.pdf")

# Extract text from the first page
page = doc.load_page(0) # 0-based index for the first page
text = page.get_text("text")

print(text)
```

```
# Open the PDF file
doc = fitz.open("clinical_trial_report.pdf")

# Extract text from the first page
page = doc.load_page(0) # 0-based index for the first page
text_blocks = page.get_text("blocks") # Extract text in blocks
```

```
[In [366]: text
Out[366]: 'ALEPRO trial \nENGOT-0V70/BGOG-0V70 \n
\nVersion 1.0 (8 Feb 2023) \n \n \nPROTOCOL SYNOPSIS \n
\nTitle of clinical Trial («Trial») \nPhase II, open
label, multicenter study of abemaciclib and letrozole in
\nEstrogen receptor positive rare ovarian cancer
\nProtocol \nShort \nTitle \nAcronym \nALEPRO \nTrial
Phase (I, II, III, IV) \n2 \nSponsor name \nUniversity
Hospitals Leuven (UZ Leuven) \nCoordinating Investigator
\nEls Van Nieuwenhuysen \n Contact Address CI
\nHerestraat 49, 3000 Leuven \n Contact Email CI
\nEls.vannieuwenhuysen@uzleuven.be \n Contact
Phone CI \n0032/16342531 \nEudraCT number
```

```
~',
(76.58399963378906,
142.09996032714844,
518.1669311523438,
165.49998474121094,
'Title of clinical Trial («Trial») \nPhase II, open
abel, multicenter study of abemaciclib and letrozole in
\nEstrogen receptor positive rare ovarian cancer \n',
3,
0),
(76.58399963378906,
173.0599822998047,
201.63734436035156,
183.61997985839844,
'Protocol \nShort \nTitle \n',
4,
0),
```

Examples of Python-Docx output

```
from docx import Document

# Function to read and extract text from a .docx file
def read_docx(file_path): # Load the .docx file
    doc = Document(file_path)
    text = ""
    for para in doc.paragraphs:
        text += para.text + "\n" # Add a newline after each paragraph

    return text

file_path = 'SAP.docx' # Replace with your file path
doc_text = read_docx(file_path)

print(doc_text)
```

```
from docx import Document

# Function to read tables from a .docx file
def read_docx_tables(file_path):
    doc = Document(file_path)
    tables_data = []
    for table in doc.tables:
        table_data = [] # Loop through each row in the table
        for row in table.rows:
            row_data = [] # Loop through each cell in the row
            for cell in row.cells:
                row_data.append(cell.text.strip())
            table_data.append(row_data)
        tables_data.append(table_data)
    return tables_data

tables = read_docx_tables(file_path)
print(tables)
```

```
In [12]: print(doc_text)
title page      1
TABLE OF CONTENTS      2
List Of Abbreviations      4
1.      INTRODUCTION      5
2.      OBJECTIVES      5
3.      STUDY OVERVIEW      5
3.1     Study Design      5
3.2     Sample Size      7
3.3     Randomization and Unblinding Procedures      8
4.      STUDY ENDPOINTS/Outcomes      8
5.      HYPOTHESES TESTING      9
6.      ANALYSIS SUBSETS      10
6.1     Safety Population      10
6.2     Modified Intent to Treat Population (mITT Population)      10
```

```
[['Abbreviation or special term', 'Explanation'],
 ['AE', 'Adverse Event'],
 ['ANOVA', 'Analysis of Variance'],
 ['BMI', 'Body Mass Index'],
 ['CFR', 'Code of Federal Regulations'],
 ['eCRF', 'Electronic Case Report Form'],
 ['CRO', 'Contract Research Organization'],
 ['DCF', 'Data Clarification Forms'],
 ['EC', 'Ethics Committee'],
 ['ED', 'Early Discontinuation'],
 ['EDC', 'Electronic Data Capture'],
 ['EOT', 'End of Treatment'],
 ['FDA', 'Food and Drug Administration'],
 ['GCP', 'Good Clinical Practice'],
 ['hCG', 'Human Chorionic Gonadotropin'],
 ['HEENT', 'Head, Eyes, Ears, Nose and Throat'],
```

Examples of Python-Docx output1

CSR File : CSR Final_1.0_13oct2021(headings)_updated_table.docx

(67/4,115 [1.6%] subjects), and "time window deviations" (23/4,115 [0.6%] subjects).

Protocol deviations and COVID-19 protocol deviations are summarized in Table 11 and Table 14.1.1.5, respectively.

Table 11 Protocol Deviations (Safety Population)

Type	XYZ1234	Placebo	Total
Category [n (%)]	(N=3,082)	(N=1,033)	(N=4,115)
Any Important Protocol Deviation	1,837 (59.6)	591 (57.2)	2,428 (59.0)
Any Major Protocol Deviation	157 (5.1)	48 (4.6)	205 (5.0)
03 - Investigational Product	63 (2.0)	22 (2.1)	85 (2.1)
06-Visit Window	53 (1.7)	16 (1.5)	69 (1.7)
05-Endpoint Data	28 (0.9)	11 (1.1)	39 (0.9)
09-Prohibited Co-Medication	11 (0.4)	3 (0.3)	14 (0.3)
02-Exclusion Criteria	2 (0.1)	0	2 (0.0)
04-Assessment Safety	2 (0.1)	0	2 (0.0)

QC Rule

Source to CSR Source Table CSR to Source

Select Table *

Table 14.1.1.4 Protocol Deviations (Safety Population)

Table 14.1.1.4 Protocol Deviations (Safety Population)

Type Category [n (%)]	XYZ1234	Placebo	Total
(No Match)	(N=3082)	(N=1033)	(N=4115)
Any Important PD (Match)	1837 (59.6) (Match)	591 (57.2) (Match)	2428 (59.0) (Match)
Any Major PD (Match)	157 (5.1) (Match)	48 (4.6) (Match)	205 (5.0) (Match)
Any Major PD 03-Investigational Product (Match)	63 (2.0) (Match)	22 (2.1) (Match)	85 (2.1) (Match)
Any Major PD 06-Visit Window (Match)	53 (1.7) (Match)	16 (1.5) (Match)	69 (1.7) (Match)

Examples of Multi-lingual Redaction

1. SYNOPSIS	
Name of Study Drug	PPD
RVT-3101	
Name of Active Ingredient	
Recombinant human monoclonal antibody (immunoglobulin gamma-1 with kappa light chains, IgG1 kappa) directed against Tumor necrosis factor-like Ligand 1A (TL1A)	
Title of Study	PPD
A Phase 2, Multicenter, Double-blind, Two-arm Study of Subcutaneous RVT-3101 for the Treatment of Subjects with Moderate to Severe Active Crohn's Disease	
Study Center(s)	PPD
This study will be conducted in up to approximately 150 sites worldwide. Worldwide Clinical Trials, a contract research organization, will oversee operational aspects of this study on behalf of Telavant, Inc., the Sponsor of the study .	
Phase of Development	

1. SINOPSIS	
Nombre del fármaco del estudio	PPD
RVT-3101	
Nombre del principio activo	
Anticuerpo monoclonal humano recombinante (inmunoglobulina gamma-1 con cadenas ligeras kappa, IgG1 kappa) dirigido contra el ligando 1A (TL1A) similar al factor de necrosis tumoral	
Título del estudio	PPD
Estudio de fase II, multicéntrico, doble ciego y de dos grupos sobre RVT-3101 subcutáneo para el tratamiento de sujetos con enfermedad de Crohn activa moderada a grave	
Centro(s) del estudio	PPD
Este estudio se llevará a cabo en aproximadamente un máximo de 150 centros de todo el mundo. Worldwide Clinical Trials, una organización de investigación por contrato, supervisará los aspectos operativos de este estudio en nombre de Telavant, Inc., el promotor del estudio .	

Figure 1 illustrates an example of redaction in an English document, where the words "Study" and "Drug" are marked for redaction. In the preview copy, these keywords are highlighted to indicate their selection. In the redacted copy, the words "Study" and "Drug" are removed, and their absence is clearly highlighting redacted areas, ensuring transparency and consistency in the redaction process.

Figure 2 illustrates an example of redaction in a Spanish document, where the document undergoes redaction by importing redaction strategies from the corresponding English version of the document. This ensures that the same redaction protocols applied to the English document are consistently applied to the Spanish version, maintaining uniformity in the redaction process across different languages.

Questions & Answers

[Transforming clinical trials with open-source tech |
pharmaphorum](#)

[Unlocking Insights in Clinical Trials with Python](#)





Ideas
Delivered
Intuitively

Thank You!

+1-833-GEN-INVO (436-4686) | info@geninvo.com | www.geninvo.com

