# Report on Research Paper Retrieval from PubMed

#### 1. Introduction

This report outlines the approach, methodology, and results of the Python program developed to fetch research papers from PubMed based on user-specified queries. The program aims to identify papers with authors affiliated with pharmaceutical or biotech companies and output the results in a structured CSV format.

## 2. Approach

The program automates the retrieval of research papers from PubMed using the Entrez API. The approach consists of the following key components:

- 1. **User Query Input:** The user provides a query term related to a research topic (e.g., "cancer research").
- 2. **PubMed API Search:** The program sends the query to the PubMed database using the NCBI Entrez API.
- 3. **Data Filtering:** The program extracts metadata such as paper ID, title, year of publication, authors, and affiliations.
- 4. **Affiliation Analysis:** The extracted affiliations are checked for references to pharmaceutical or biotech companies.
- 5. **Output Generation:** The final processed data is saved as a CSV file or displayed in debug mode.

# 3. Methodology

The methodology involves the following structured steps:

#### 3.1 Data Retrieval

- The Entrez API is used to fetch research paper metadata from PubMed.
- The search query is processed and formatted to optimize retrieval.

#### 3.2 Data Extraction

- The program extracts metadata including:
  - Paper ID (PMID)
  - o Title
  - Year of publication
  - Author names
  - Author affiliations

### 3.3 Affiliation Analysis

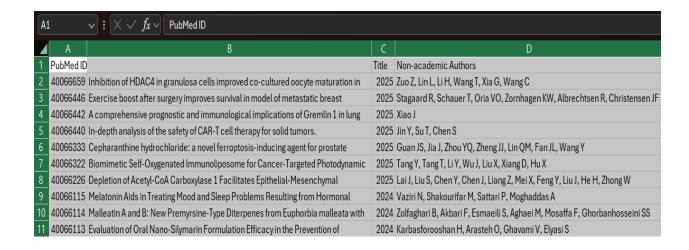
- The affiliations of authors are analysed using keyword matching techniques.
- Keywords such as "pharmaceutical", "biotech", "biopharma", and company names are used for identification.

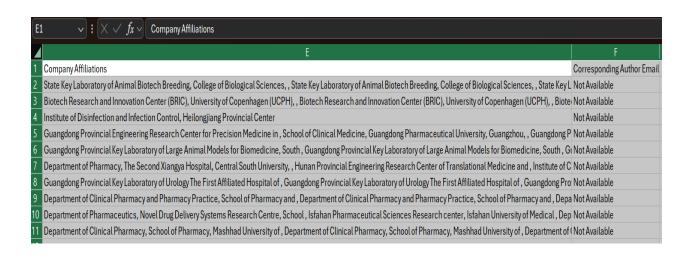
### 3.4 Output Generation

- The extracted and filtered data is saved in CSV format.
- Debug mode allows real-time viewing of processed results in the console.

#### 4. Results

The program successfully retrieved research papers relevant to the specified query and identified papers associated with pharmaceutical or biotech companies. Sample results include:





#### 5. Conclusion

The implemented Python program effectively automates research paper retrieval and affiliation analysis for pharmaceutical and biotech-related studies. It provides a structured and efficient approach to identifying relevant scientific publications, aiding researchers in literature reviews and industry-related research monitoring. Future enhancements could include natural language processing (NLP) for improved affiliation analysis and automated company name recognition.

**Keywords:** PubMed API, Research Paper Retrieval, Affiliation Analysis, Pharmaceutical Industry, Biotech Companies, Python Automation