📄 Project Report: PubMed CLI Tool with TestPyPI Deployment

Author: Vignesh Ravichandran
Project: get-papers-list-Vignesh_R
Duration: 4 days
Status: ✅ Successfully developed, tested, and published to TestPyPI

✅ Objective

To build a Python CLI tool that queries **PubMed**, extracts papers with **at least one non-academic author affiliated with a biotech/pharma company**, and exports results to a structured CSV file. The tool must:

- Support **PubMed's full query syntax**
- Run via the **command line**
- Be packaged and published using **Poetry**
- Be validated via **TestPyPI**

⚙️ Approach & Methodology

1. PubMed API Integration

- Used **NCBI E-Utilities (esearch & efetch)** for data fetching
- Added batching & rate-limiting support
- Supported full query syntax ([dp], AND, OR, MeSH terms)

2. Affiliation Filtering

- Created heuristics using keyword matching:
    - **Academic**: "university", "institute", "hospital", etc.
    - **Biotech/Pharma**: "biotech", "pharma", "inc", known company names
- Parsed XML using xml.etree.ElementTree

3. CSV Export

- Used csv.DictWriter
- Cleaned affiliation info and handled UTF-8 characters
- Included columns: PubMed ID, Title, Date, Company Affiliations, Emails

4. Command-Line Interface

- Added CLI entry point via in pyproject.toml:

```
[tool.poetry.scripts]
get-papers-list = "get_papers.cli:main"
```

- Handled --help, --file, and --debug flags using argparse

5. Error Handling

- Handled HTTP errors, XML parsing errors, I/O errors, malformed data
- Included --debug mode for traceability

6. Testing

- Unit tested:
  - PubMed ID fetcher
  - XML parser
  - Affiliation classifiers
  - CSV exporter
- Used pytest, monkeypatch, and tempfile

7. Packaging & Publishing

- Managed dependencies via **Poetry**
- Published to **TestPyPI**:

  ```
  poetry publish --build -r test-pypi
  ```

- Installed and tested CLI using:

  ```
  pip install --index-url https://test.pypi.org/simple/ --extra-index-url https://pypi.org/simple get-papers-list-Vignesh_R
  ```

📈 Results

- ✅ CLI works as expected with full PubMed queries
- ✅ Exports data for queries like:

bash

CopyEdit

get-papers-list "CRISPR AND 2023[dp]" --file crispr.csv

- ✅ Handles mixed academic & corporate authors
- ✅ Upload verified on TestPyPI: TestPyPI Project Page
- ✅ Full test coverage across major components

🚀 Next Steps

- Automate testing + publishing using **GitHub Actions**
- Publish to **real PyPI** after validation
- Expand affiliation filtering using ORCID/ROR/GRID datasets