

Report Article

Comparison of the Accuracy of NLP Tools and Manual Investigation for Identification, Extraction, and Classification of Biomarkers in Colorectal Cancer

Mahsa Mirzarazi*

Shahid Beheshti University of Medical Sciences, Tehran, Iran

Received: 27 August, 2025; Accepted: 26 July, 2026

DOI: 10.22037/nbm.v14i4.50096

Abstract

Background and Aim: The rapid expansion of biomedical literature has made the manual identification and extraction of biomarker information increasingly time-consuming. Natural language processing (NLP) tools may facilitate large-scale literature screening and support the generation of research hypotheses. However, their ability to broadly identify diverse biomarker types without predefined target terms remains insufficiently characterized.

Materials and Methods: Fifty original articles on colorectal cancer biomarkers published between 2015 and 2025 were manually reviewed, and reported biomarkers were extracted to establish a reference dataset. The same articles were processed using two biomedical NLP models: d4data/biomedical_ner_all and en_ner_bionlp3cg_md from the SciSpaCy framework. Biomarkers extracted by each model were compared with the manually curated reference set. The proportion of manually identified biomarkers recovered by each model was calculated, and extracted entities not present in the reference set were also assessed. Statistical analyses were performed using an independent-samples t-test.

Results: Both NLP pipelines successfully processed 94% of the 50 articles. The mean matching rates with the manually curated biomarkers were 49.6% for SciSpaCy and 50.7% for BioMedNLP, indicating only a minimal difference between the two models. More than 50% of the manually identified biomarkers were recovered in 46% of the SciSpaCy outputs and 54% of the BioMedNLP outputs. Additional entities not present in the manual reference set accounted for 21.4% and 19.2% of the entities extracted by SciSpaCy and BioMedNLP, respectively. The differences between NLP-derived and manually curated biomarker counts were statistically significant ($P < 0.0001$).

Conclusion: Without predefined biomarker-specific keywords or domain-specific fine-tuning, the evaluated NLP models recovered approximately half of the manually identified biomarkers. Nevertheless, the relatively similar distribution of biomarker categories suggests that these tools may provide useful preliminary overviews of biomarker landscapes. Their most appropriate role may therefore be as screening and ideation-support tools, followed by manual verification for precise biomarker identification and interpretation.

Keywords: Biomedical NLP, Biomarker extraction, Text mining, Artificial Intelligence in biomedical research

*Corresponding Author: Mahsa Mirzarazi, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Email: mirzarazi.m@sbm.ac.ir. ORCID:0000-0001-6058-4368

Extraction, and Classification of Biomarkers in Colorectal Cancer. *Novel Biomed.* 2026;14(4):205-8

Introduction

The rapid growth of biomedical literature has made manual extraction of biomarker data labor-intensive and inefficient. NLP techniques, particularly transformer-based models such as BioBERT (Biomedical Bidirectional Encoder Representations from Transformers), have significantly improved the accuracy of biomedical named entity recognition (NER) tasks compared to general-purpose models¹. Recent advancements also include specialized pipelines capable of extracting biomarker entities — for example, MarkerGenie, which automates the extraction of molecular markers from biomedical texts. Additionally, open-source tools such as biomarker_nlp facilitate the mining of biomarker information for targeted cancer therapies, demonstrating the practical feasibility and relevance of NLP methods in biomarker extraction workflows². Tools like SciSpaCy provide robust, domain-optimized processing pipelines for biomedical text, enabling fast and reliable extraction of named entities³. This study aimed to utilize NLP tools in medical and biomedical research to enhance the speed of study and facilitate ideation for researchers.

Methods

In this study, AI assistance, Python software version 3.10.11, and two models of NLP tool that were used for biomedical named entity recognition: 1) d4data/biomedical_ner_all is a transformer-based biomedical language model (similar to Nature BioBERT), and 2) en_ner_bionlp3_md is provided by the SciSpaCy library, which is built on the Spacy framework and optimized for biomedical text processing.

For statistical analyses, an independent sample t-test was performed using SPSS software version 16.0.

In the first step, 50 original articles published in 2015 - 2025 on colorectal cancer biomarkers were reviewed, and all biomarkers mentioned in each article were manually identified and extracted (“Manual biomarkers”).

In the next step, NLP tools were applied to study, identify, and extract biomarkers from the PDF files of

these 50 articles. The necessary code was implemented and executed in Python.

It is worth noting that to achieve maximum accuracy (approximately 80–85%) in NLP-based extraction, the target terms must be defined as specific keywords. However, in this study, NLP tools were used in a general manner, without predefining biomarkers as targeted keywords⁴. The aim was to evaluate the models’ overall capability to identify different types of biomarkers. If the models demonstrate sufficient capability, they can serve as a valuable complement to researchers, enabling faster literature analysis and facilitating the generation of novel, up-to-date ideas. For example, suppose the models can generally identify and extract all biomarker types and names from the 50 articles. In that case, researchers can determine which biomarkers have been most extensively investigated and which have been least studied.

Finally, the raw data Excel file generated by the NLP tools was organized and classified with AI assistance. AI-assisted spreadsheet organization was used solely to arrange and format extracted entities. Final classification and verification were manually reviewed by the author.

Results

The biomarkers identified by the NLP tools were extracted in two columns labeled SciSpaCy and BioMedNLP. For comparison and analysis, the manually extracted biomarkers were considered the reference standard (ground truth) for performance comparison.

Each model achieved a 94% success rate in processing the 50 articles. Here, success rate refers to the ability of the NLP pipeline to successfully process PDF files and generate extractable outputs, whereas match rate refers to the overlap between NLP-extracted biomarkers and the manually extracted reference biomarkers.

Then, for each row in the two columns, the percentage of biomarkers from the Manual column that were identified was calculated. Among the data from the SciSpaCy and BioMedNLP columns, 46% and 54%, respectively, successfully identified and extracted more than 50% of the biomarkers from the Manual column.

The NLP tools also extracted additional entities that

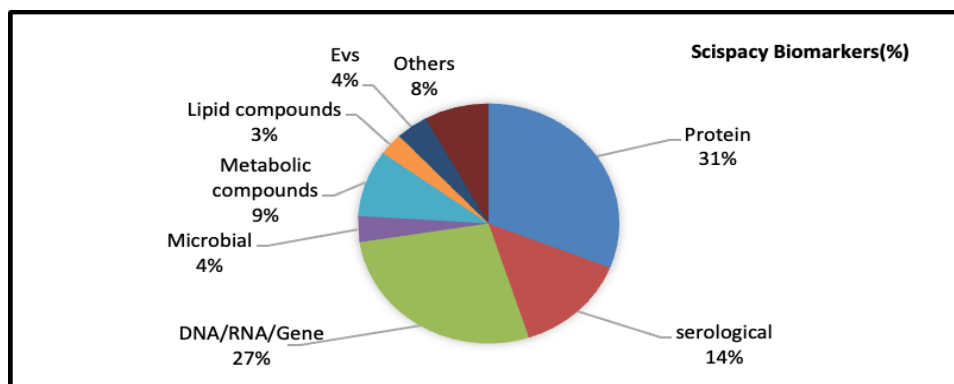


Figure 1. Classification of biomarker types extracted from SciSpaCy column.

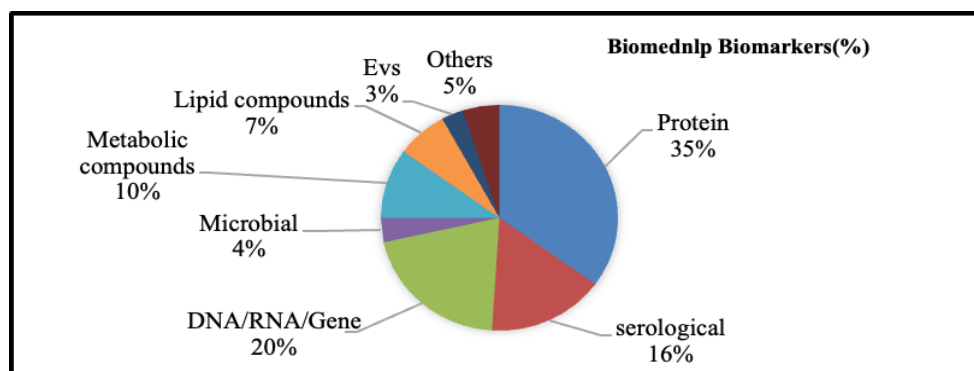


Figure 2. Classification of biomarker types extracted from the BioMedNLP column Isolates Were MBL Positive. A, IMP (10 µg) disc alone; B, disk of IMP (10 µg) plus EDTA.

were not included in the manually extracted reference set. These represented 21.4% and 19.2% of the total extracted entities for the SciSpaCy for the BioMedNLP, respectively.

The overall mean percentage of matches for the SciSpaCy and BioMedNLP columns was 49.6% and 50.7%, respectively, indicating minimal difference between the two NLP tools in their ability to match manual biomarkers.

It should be noted that the differences in the mean match rate between each NLP tool column and the control column were statistically significant ($P < 0.0001$).

Figures 1-3 illustrate the classification of biomarker types extracted from each column, presented graphically.

Discussion

In this study, the NLP tools identified approximal 50% of the biomarkers extracted by the manual investigation. This difference was statistically significant, indicating that the NLP tools had

moderate accuracy in recognizing biomarker names. The results showed that when NLP models such as SciSpaCy³ and BioMedNLP¹ are used to extract any colorectal cancer biomarker—without restricting to specific categories (e.g., proteins, cytokines, metabolites) — and without fine-tuning the models for this particular domain, the outputs achieved an average matching rate of approximately 50% compared to the reference manual data.

Generally, to achieve accuracy above 85%, the scope of the query must be limited and well-defined. Additionally, the model should be fine-tuned, or specialized trained models should be employed. Finally, manual review and refinement of the output are necessary to ensure completeness and accuracy⁴.

The reason for these results is that general or pretrained NLP models typically exhibit limited accuracy when applied to broad or open-ended tasks (such as extracting “all biomarkers”), often producing outputs that are incomplete or contain irrelevant terms. However, in the classification of biomarkers by type (e.g., protein, immunological, metabolic), the NLP tools performed comparably to manual extraction. The

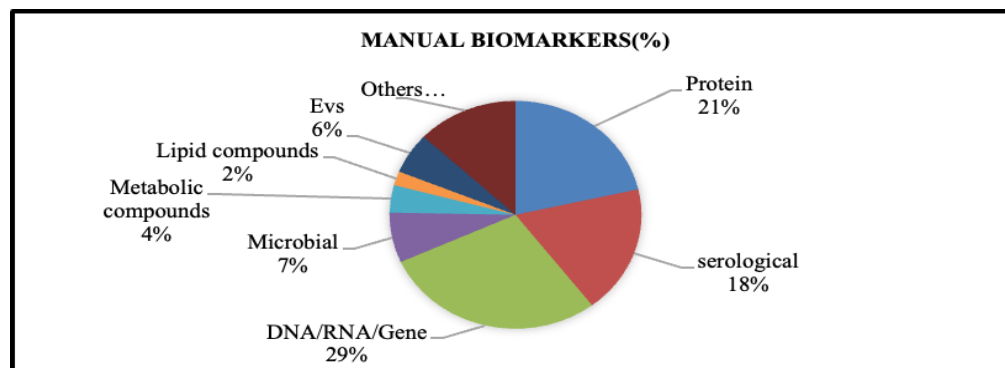


Figure 3. Classification of biomarker types extracted from the Manual column.

percentage distribution of biomarker types was nearly identical between the two methods, as illustrated in Figures 1-3. The results of this study align with prior work showcasing the utility of NLP tools in biomedical contexts. While tools like BioBERT exhibit high accuracy in structured settings (e.g., clinical reports), their performance tends to decline when applied broadly without domain-specific fine-tuning. This mirrors our findings, in which pretrained models, used without predefined keyword constraints, achieved a moderate matching accuracy (~50%) compared to manual extraction. Nevertheless, such tools remain valuable for screening and ideation, helping to streamline manual curation efforts—especially when a full manual review is impractical.

A limitation of the present study is that biomarker-specific performance metrics were not evaluated separately for individual biomarker categories such as proteins, genes, metabolites, or immunological markers. Future studies may investigate category-specific extraction performance using larger datasets and domain-specific model fine-tuning.

Conclusion

These findings suggest that although the NLP tools demonstrated moderate accuracy in extracting specific biomarker names, they provided a relatively reliable overview of the biological patterns present in the articles. Consequently, the NLP tools can be effectively utilized as an initial ideation resource for screening articles or for generating preliminary

research maps. For instance, researchers can first employ NLP to identify common biomarker types and subsequently apply manual methods for precisely extracting specific names and references, thereby reducing analysis time and costs.

Acknowledgment

None.

Conflict of interest

The authors further declare that they have no conflict of interest.

Funding

None.

References

1. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Wren J, editor. *Bioinformatics*. 2020;36(4):1234–40.
2. Lin J, He Y, Madhavan S, Kim C, Boca SM. Biomarker_nlp: A Python Package for Mining Biomarker Information for Targeted Cancer Therapies. *J Open Res Softw*. 2024;12:15.
3. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In: *Proceedings of the 18th BioNLP Workshop and Shared Task* [Internet]. Florence, Italy: Association for Computational Linguistics; 2019 [cited 2025 Aug 11]. p. 319–27. Available from: <https://www.aclweb.org/anthology/W19-5034>
4. He T, Kreimeyer K, Najjar M, Spiker J, Anagnostou V, Botsis T. Artificial Intelligence-assisted Biomedical Literature Knowledge Synthesis to Support Decision-making in Precision Oncology.