[[1]](#footnote-0)

A Comparative Approach for Summarizing PubMed Articles based on PICO Questions for Creating Clinical Case Report

Sabah Mohammed, Hashmath Shaik, and Jinan Fiaidhi

***Abstract*— This electronic document is a “live” template. The various components of your paper [title, text, heads, etc.] are already defined on the style sheet, as illustrated by the portions given in this document.**

***Keyword*— PICO questions, Evidence-Based Medicine, Clinical Cases, Machine Learning, BERT, Bio-BERT.**

# INTRODUCTION

Evidence based medicine (EBM) is centered on identifying the best clinical evidence available both from the clinical practice at the bedside and the scientific evidence from the bench [1]. The process of identifying these evidences involves iterative interrogation to clinical sources and medical literature as well as summarizing the evidences either manually by physician or through assistive summarization methods. The clinical question needs to address the patient case main investigation issues to seek relevant materials and more focused search. However, a well-built clinical question needs to include import keywords and phrases relevant to the clinical case problem or its population effect as well as any information related to possible interventions, comparison of alternative treatments/therapies and the outcome of the used treatments/therapies methods. In this direction, the medical practice often uses the PICO standard to define the key information needed for formulating a well-built clinical question [2]. Actually PICO is the *de facto* standard to improve the specificity and conceptual clarity of clinical questions [3]. Articulating a clinical question based on PICO uses four anatomic parts—**P**roblem/**P**opulation, **I**ntervention, **C**omparison, and **O**utcome (PICO) to facilitate searching for precise answers from open research. Figure 1 illustrates the different types of clinical questions that can be expressed with PICO. Based on identifying these components from any case description, clinician can formulate a PICO question. For example, in a clinical case where *the patient is a female* *diagnosed with SLE. Her rheumatologist suggested Plaquenil medication for her joint pain, but she is interested in alternative therapies because she heard about success a neighbor had with turmeric tea.* The PICO components for this scenario should be as follows:

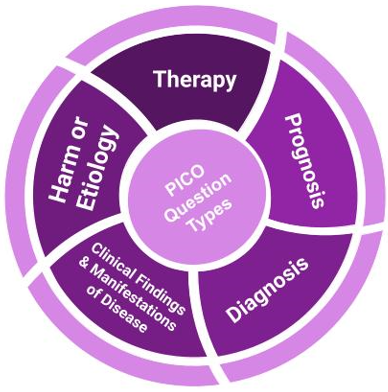
**P-** Patient with SLE

**I-** Turmeric tea

**C-** Plaquenil

**O-** Reduction of joint pain

and the PICO question can be formulated as follows: “*In adult patients with SLE, is consuming turmeric tea more effective than Plaquenil at reducing joint pain?*” The US NIH (National Institute of Health) as well as other medical institutions have developed web applications that uses PICO form to interrogate important clinical literature from many sources including the PubMed covering variety clinical repositories [4] (e.g. Clinical Trials, Randomized Controlled Trails, Systematic Reviews, Practice Guidelines, Meta-Analysis and medical Literature). Only the PubMed repository contains more than 35 million articles and is updated daily [5].



**Fig. 1:** PICO Question Types.

Although there are lots of educational tutorials on formulating PICO questions related to clinical cases, their formulation remains less straightforward [6]. For this reasons, many attempts have been introduced to automatically assist in identifying the PICO components from medical cases descriptions to help in formulating variety of PICO questions. Most important are the machine learning attempts that provided helpful results [7, 8, 9]. However, these attempts does not reveal the most effective machine learning that can identify PICO components without much of the expert intervention. In this paper we are investigating two important issues including the best machine learning technique that can be used for identifying PICO components and the second how to summarize an clinical case so PICO components can be more focused as we as summarizing the articles collected from the PICO search.

# machine Learning Methods for Identifying PICO Components

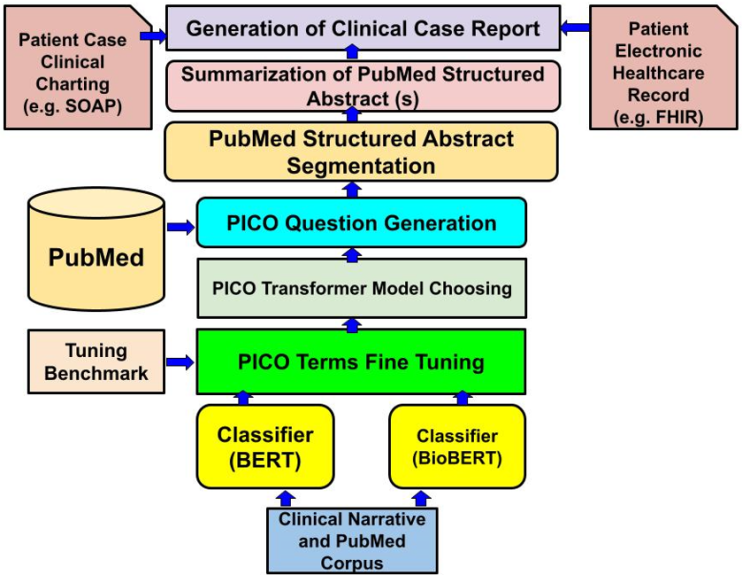
As stated earlier, the fundamental skill required for practicing Evidence-Based Medicine (EBM) is asking well-built PICO clinical questions when presented with any clinical narrative case. However, automatically identifying the PICO components from such clinical narratives requires a special spider-like capability involving what is called as “qualitative evidence syntheses” [10]. This synthesis process necessitates effective machine-learning models that can reliably identify the relevant PICO components.

The use of transformer models [11] (e.g., BERT, BioBERT) provides state-of-the-art results on natural language processing (NLP) tasks such as those required for identifying PICO tags in clinical narratives [12], PubMed structured Abstract segments [13,14,15] and summarization of clinical narratives [16].

However, both BERT and BioBERT are not initially trained to spherically identify PICO terms which makes their performance in this direction lower than the expectation. This necessitates the need for fine-tuning additional training [19].

We fine-tuned these models with a dataset that helps categorize relevant PICO words. After fine-tuning, there were notable improvements in the performance of both models. The accuracy of the BERT model improved from 0.35 to 0.44 after fine-tuning. Similarly, BioBERT's accuracy also improved from 0.40 to 0.52. Moreover, the precision of BioBERT increased from 0.30 to 0.58, and the recall for the BioBERT model improved to 0.52 from 0.09. The F1-Score, which indicates a balance between precision and recall, increased for both models after fine-tuning, marking significant strides in their performance. However, there is still room for further improvement.

Our focus in the next sections builds on this process of PICO term identification to recognize the structured abstract component including aim, methods, results, and conclusions, as well as summarizing individual and multiple structured abstracts. Figure 2 illustrates our qualitative evidence synthesis process for creating clinical case reports.



**Fig. 2:** The Framework for Generating Clinical Case Report based on PICO Questions.

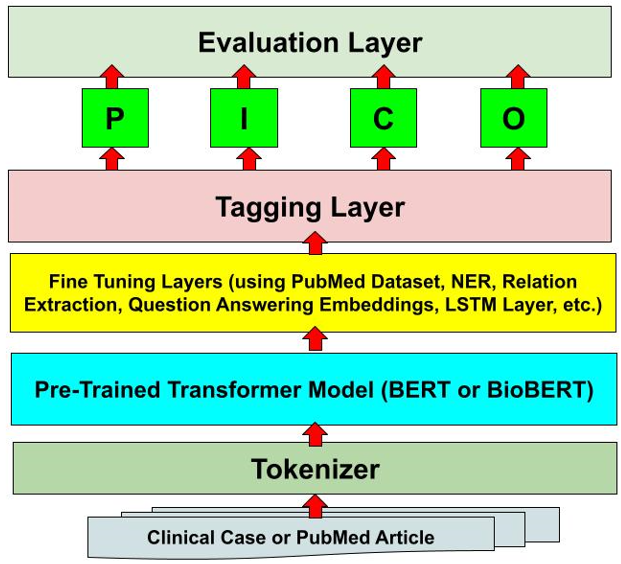
## A. Machine Learning Models Trained using PICO Identified Dataset using BERT and Bio-BERT

In our quest to effectively identify and extract PICO elements from clinical narratives and biomedical literature, we sought to compare the performance of two potent language models: BERT [17] (Bidirectional Encoder Representations from Transformers) and its domain-specific counterpart, BioBERT [18].

BERT, developed by Google, is a transformer-based machine learning technique for natural language processing (NLP) pre-trained using the corpus of 800M words from Toronto Book Corpus and 2500M words from English Wikipedia. The BioBERT model, is designed specifically for biomedical text mining using a dedicated biomedical corpus that includes 4.5B words from PubMed and 13.4B words from PMC. However, BioBERT has gone through a series of improvements. The variant "*biobert-base-cased-v1.2*" was used in this paper for comparison. This variant represents an essential evolution in this sequence, introducing substantial enhancements that increase both the model's performance and its versatility. The "BioBERT-Base v1.2" extends the training corpus by incorporating an additional one million articles from PubMed. This inclusion not only increases the volume of training data but also broadens the spectrum of biomedical sub-domains the model is exposed to, thereby refining its understanding and representation of biomedical language. However, both BERT and BioBERT are not trained to spherically identify PICO terms which makes their performance in this direction lower than the expectation and necessitates the need for a fine-tuning additional training [19]. For this purpose, our fine-tuning started with a dataset that helps categorizing the relevant PICO words such as:

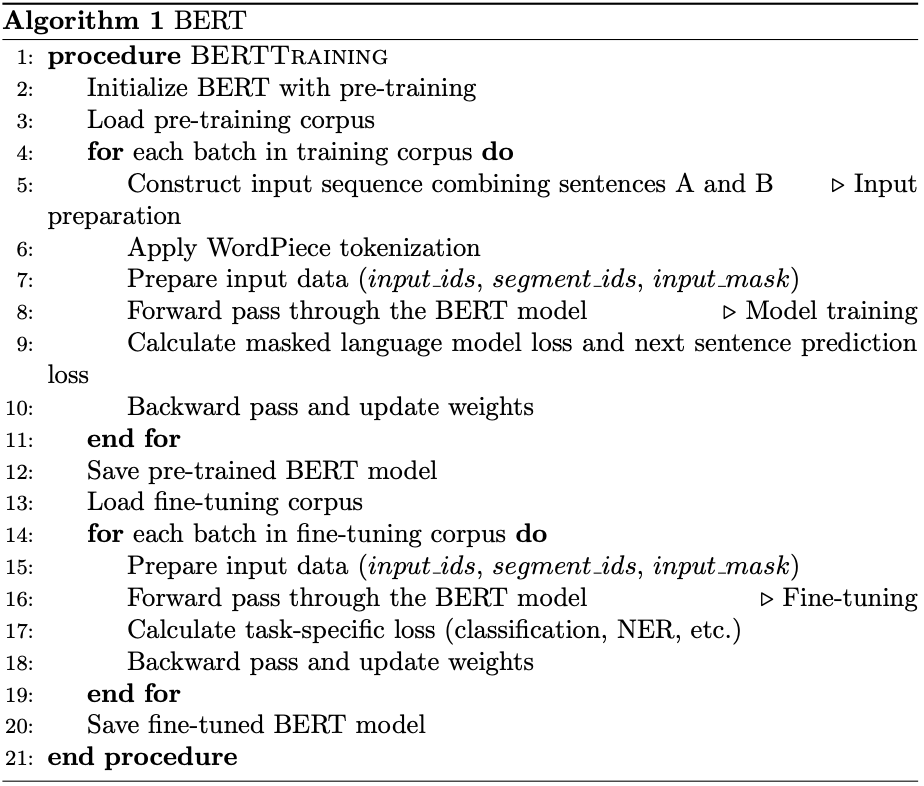
* **Aim (A)** Objective, Background, Purpose, Importance, Introduction, Aim, Rationale, Goal, Context, Hypothesis
* **Participants (P)** Population, Participant, Sample, Subject, Patient
* **Intervention (I)** Intervention, Mediation, Interposition
* **Outcome (O)** Outcome, Measure, Variable, Assessment
* **Method (M)** Method, Setting, Design, Material, Procedure, Process, Methodology
* **Results (R)** Result, Finding
* **Conclusion (C)** Conclusion, Implication, Discussion, Interpretation

This PICO tuning dataset was provided by [20] and for testing we used the tutorial by Juliet Gray of Southern Illinois University [21] that include 12 clinical cases annotated based on PICO relevancy words. Figure 3 illustrates our fine-tuning strategy for enhancing the pre-trained BERT and BioBERT models.

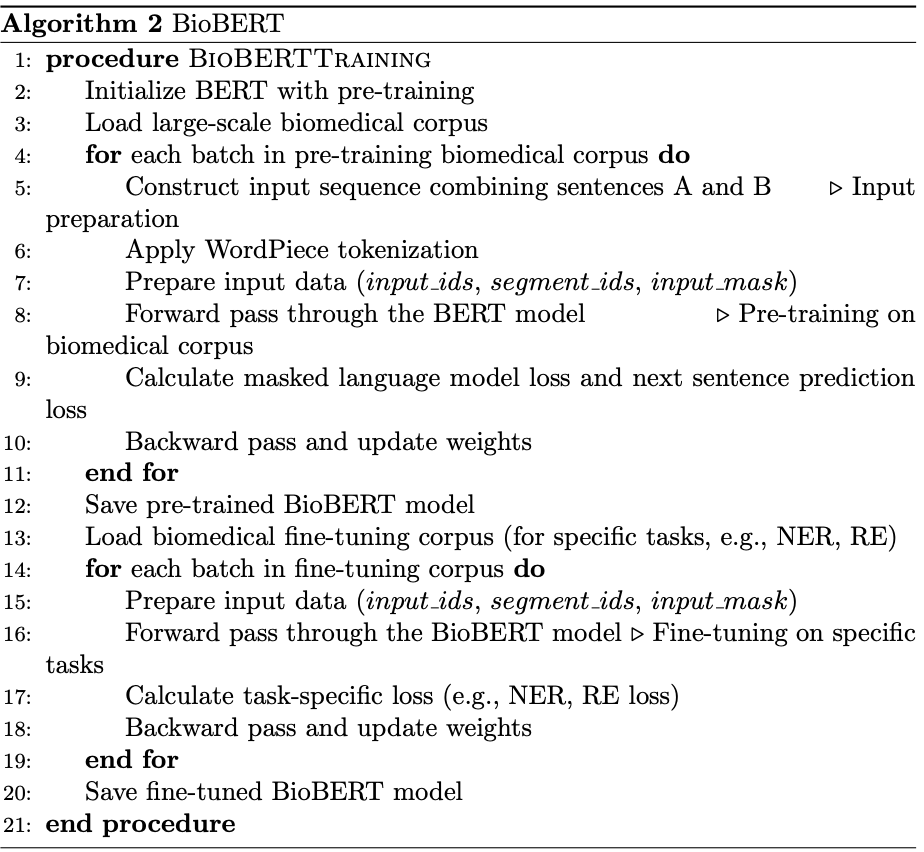


**Fig. 3:** Fine Tuning Strategy for BERT and BioBERT.

A straightforward comparison based on accuracy measures indicates some advantages of BioBERT over the BERT model (see Table 1). However, a closer comparison using different types of correlation between human-identified PICO targets and the predicted terms identified by the fine-tuned models reveals a more decisive superiority of the BioBERT in identifying the PICO terms. Table 2 illustrates the correlation-based comparison between BERT and BioBERT models.



**Fig. 4:** Fine Tuning Strategy for BERT and BioBERT.



**Fig. 5:** Fine Tuning Strategy for BERT and BioBERT.

**Table 1:** Comparing BERT and BioBERT using Accuracy Measures.

| **Model** | **Accuracy** | **precision** | **Recall** | **F1-scores** |
| --- | --- | --- | --- | --- |
| **Bert** | 0.3526315789473684 | 0.14285714285714285 | 0.08333333333333333 | 0.10526315789473685 |
| **Bio-BERT** | 0.401739130434782608 | 0.30434782608695654 | 0.091739130434782608 | 0.20057971014492754 |

**Table 2:** Comparing BERT and BioBERT using Accuracy Measures with fine-tuning technique.

| **Model** | **Accuracy** | **precision** | **Recall** | **F1-scores** |
| --- | --- | --- | --- | --- |
| **Bert** | 0.44473684210526315 | 0.17647058823529413 | 0.125 | 0.14634146341463414 |
| **Bio-BERT** | 0.5208333333333334 | 0.5859375 | 0.5208333333333334 | 0.4934210526315789 |

**Table 3:** Comparing BERT and BioBERT using multiple correlation measures.

| **Model** | **Bert** | **Bio-BERT** |
| --- | --- | --- |
| **MCC** | 0.06712417476562717 | 0.3749306775111925 |
| **Cohen’s Kappa** | 0.038566904711050176 | 0.2977099236641222 |
| **Pearson** | 0.021562203060958043 | 0.4758343846465464 |
| **Spearman’s Rank** | 0.04054318744115211 | 0.5896779102419609 |
| **Kendall Rank** | 0.037206802386340894 | 0.5431688020354455 |
| **Point Biseral** | 0.021562203060958043 | 0.4758343846465464 |

Dataset and Data Preprocessing

The dataset used in this study, introduced by Jin, Di, and Peter Szolovits (2018) [20], is a collection of biomedical literature from PubMed. It is structured to facilitate the detection of PICO elements (Participants/Problem (P), Intervention (I), Comparison (C), and Outcome (O)) — a widely used framework for formulating clinical questions in evidence-based medicine.

This dataset was designed to utilize a Long Short-Term Memory (LSTM) neural network-based model for the automatic identification of PICO elements in a medical text. Each PICO entity in the dataset is tagged, making it an invaluable tool for training machine learning algorithms to detect and classify such elements in biomedical literature.

In terms of structure, the dataset contains original abstracts encapsulated in the 'structured\_abstracts\_PICO' file. Each record begins with a PMID, followed by the original section heading, an assigned gold label for train and test, and the section content. This information is separated by the symbol '|'. The assigned gold label is created based on a keyword check in the section heading.

An alternate version of this dataset has each section content sentence-split using the Stanford CoreNLP toolkit. This arrangement ensures that each line contains only one sentence, and all numeric values are replaced by '@'.

The dataset, upon being loaded into a pandas data frame, is observed to consist of 344,636 records across five columns. The initial distribution of PICO entities is skewed, with 'O' or non-PICO elements accounting for the majority of the dataset. In contrast, the other PICO elements ('I-INTERVENTION', 'I-OUTCOME', 'I-PARTICIPANT', and 'I-POPULATION') are underrepresented.

Addressing Class Imbalance and Model Fine-tuning

Upon an initial examination of the dataset, we identified a significant class imbalance problem, with the 'O' or non-PICO elements considerably outnumbering the other PICO elements ('I-INTERVENTION', 'I-OUTCOME', 'I-PARTICIPANT', and 'I-POPULATION'). This imbalance posed a potential issue for the performance of the machine learning models, as they could easily develop a bias towards the majority class, leading to less than optimal performance on the underrepresented classes.

Addressing this imbalance is crucial, as machine learning models trained on imbalanced datasets could develop a bias towards the majority class, leading to poor generalization performance on underrepresented classes. Particularly in a medical context, such as ours, where every false negative (failure to correctly identify a PICO element) could have serious ramifications, it is imperative to accurately classify all classes.

To achieve a more balanced class distribution, we employed an undersampling technique. This approach reduces the instances of the majority class to match the size of the minority classes. While this method does risk discarding potentially useful information from the majority class, we considered it a necessary trade-off to mitigate bias and improve the model's performance across all PICO entities.

The undersampling resulted in each PICO entity being represented 1,538 times, establishing a more balanced dataset for further analysis and model training. Given the significance of correctly identifying all PICO elements in a medical context, it was critical to address this class imbalance. The solution we employed was an undersampling technique that reduced the instances of the majority class to match the size of the minority classes. Although this method potentially discarded information from the majority class, it was a necessary trade-off to decrease bias and increase the machine learning models' performance across all PICO entities.

Upon implementing the undersampling strategy, we observed a noticeable enhancement in model performance for both BERT and BioBERT. The reduction in bias towards the majority class substantially improved the identification of PICO terms previously underrepresented. Both models' precision, recall, and F1 scores for these entities saw a significant boost, demonstrating the efficacy of this technique (see Table 4).

A graph with blue bars

Description automatically generated

**Fig. 6:** Fine Tuning Strategy for BERT and BioBERT.

However, we did not stop at rebalancing the classes. The next stage involved fine-tuning the pre-existing BERT and BioBERT models on the PICO-identified data. This process was aimed at enabling the models to better recognize and extract the relevant PICO terms. The fine-tuning resulted in improved performance for both models. We observed substantial increases in accuracy, precision, recall, and F1 scores. These results confirmed that both the undersampling technique and the subsequent fine-tuning significantly improved the models' ability to extract PICO components from clinical narratives.

While the performance enhancements were notable, it is important to emphasize that there is still ample room for further advancements (see Table 5). The complex nature of clinical narratives and the inherent challenge of extracting specific PICO components suggest that there will always be a demand for better performance from these models. Future research should focus on refining these techniques and exploring other avenues to enhance the extraction and identification of PICO elements from clinical narratives.

**Table 4:** Comparing BERT and BioBERT using Accuracy Measures with fine-tuning techniques after training on balanced datasets.

| **Model** | **Accuracy** | **precision** | **Recall** | **F1-scores** |
| --- | --- | --- | --- | --- |
| **Bert** | 0.8220930232558139 | 0.5454545454545454 | 0.6 | 0.5714285714285713 |
| **Bio-BERT** | 0.8837547270056568 | 0.8822857330122664 | 0.8837547270056568 | 0.879824017366772 |

**Table 5:** Comparing BERT and BioBERT using multiple correlation measures with fine-tuning techniques after training on balanced datasets.

| **Model** | **Bert** | **Bio-BERT** |
| --- | --- | --- |
| **MCC** | 0.1701777723047615 | 0.6495651646675051 |
| **Cohen’s Kappa** | 0.3749306775111925 | 0.054142204827136275 |
| **Pearson** | -0.05916987686 | 0.2100694374718645 |
| **Spearman’s Rank** | -0.03725695418437184 | 0.15609147789499034 |
| **Kendall Rank** | -0.03181491634683386 | 0.13906290561297782 |
| **Point Biseral** | -0.059169876860193225 | 0.2100694374718645 |

## B. Machine Learning Models that Utilizes other NLP Technologies such as GPT2 and ChatGPT

*Here you will need to identify methods that uses NLP techniques like those components provided by the GP3 or ChatGPT or others to identify PICO components. How do you compare them as well as with the group in A.*

For the classification of PICO elements in this study, we utilized the GPT-2 model, a large transformer-based language model developed by OpenAI. GPT-2 has proven its capabilities in various NLP tasks and is recognized for its exceptional performance in text generation and understanding tasks.

The GPT-2 model was chosen for this task due to its unique architectural characteristics. It is a transformer-based model that uses self-attention mechanisms, which enables it to consider the entire context of a sentence to make predictions, a critical requirement for accurately identifying and classifying PICO elements in the complex medical literature.

Before training, the GPT-2 model was initialized with weights pre-trained on a large corpus of text. The use of these pre-trained weights allows the model to leverage general language understanding learned from a wide variety of texts, serving as a robust starting point for further fine-tuning on our specific task.

The GPT-2 model was then fine-tuned on our balanced dataset, which was prepared through undersampling to mitigate class imbalance. The model was trained to predict the correct PICO label ('O', 'I-INTERVENTION', 'I-OUTCOME', 'I-PARTICIPANT', or 'I-POPULATION') for each token in a sentence. The model learns to understand the context and semantics of the sentence, allowing it to classify each token accurately.

To prevent overfitting, we used a split of the data into training, validation, and testing sets, following an 8:1:1 ratio. The model was trained on the training data, and the performance was periodically evaluated on the validation set. The test set was reserved for the final evaluation of the model.

The use of the GPT-2 model for PICO element classification represents an advanced application of transformer models in the realm of biomedical literature analysis. The results from this study provide valuable insights into the potential of large language models for automated information extraction in medical research.

**Table 6:** Accuracy Measures of GPT-2 with fine-tuning techniques after training on balanced datasets.

| **Model** | **Accuracy** | **precision** | **Recall** | **F1-scores** |
| --- | --- | --- | --- | --- |
| **GPT-2** | 0.904 | 0.9031719308669772 | 0.904 | 0.9033668654969342 |

**Table 7:** Comparing GPT-2 using multiple correlation measures with fine-tuning techniques after training on balanced datasets.

| **Model** | **GPT-2** |  |
| --- | --- | --- |
| **MCC** | 0.6137269544 |  |
| **Cohen’s Kappa** | 0.6740370511218559 |  |
| **Pearson** | 0.7375237269542 |  |
| **Spearman’s Rank** | 0.657678144 |  |
| **Kendall Rank** | 0.6375251137269544 |  |
| **Point Biseral** | 0.6269542 |  |

# Individual Clinical Case Summarization

In this section, you will need to describe different summarization techniques and state which one better. Does the earlier techniques used to identify PICO component helps? Whish techniques is better?

# Summarization of Literature from PICO Search

In this section, you will need to experiment with summarizing the structured abstract of many collected PubMed articles. What are the best suggested techniques and which one is better?

# Conclusion

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression, “One of us (R. B. G.) thanks . . .” Instead, try “R. B. G. thanks”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

1. Burns, Patricia B., and Kevin C. Chung. "Developing good clinical questions and finding the best evidence to answer those questions." Plastic and reconstructive surgery 126, no. 2 (2010): 613.
2. Huang, Xiaoli, Jimmy Lin, and Dina Demner-Fushman. "Evaluation of PICO as a knowledge representation for clinical questions." In AMIA annual symposium proceedings, vol. 2006, p. 359. American Medical Informatics Association, 2006.
3. Cheng, Grace YT. "A study of clinical questions posed by hospital clinicians." Journal of the Medical Library Association 92, no. 4 (2004): 445.
4. Brown, David. "A review of the PubMed PICO tool: using evidence-based practice in health education." Health Promotion Practice 21, no. 4 (2020): 496-498.
5. Ebbert, Jon O., Denise M. Dupras, and Patricia J. Erwin. "Searching the medical literature using PubMed: a tutorial." In Mayo Clinic Proceedings, vol. 78, no. 1, pp. 87-91. Elsevier, 2003.
6. Huang, Ke-Chun, I-Jen Chiang, Furen Xiao, Chun-Chih Liao, Charles Chih-Ho Liu, and Jau-Min Wong. "PICO element detection in medical text without metadata: Are first sentences enough?." Journal of biomedical informatics 46, no. 5 (2013): 940-946.
7. Boudin, Florian, Jian-Yun Nie, Joan C. Bartlett, Roland Grad, Pierre Pluye, and Martin Dawes. "Combining classifiers for robust PICO element detection." BMC medical informatics and decision making 10, no. 1 (2010): 1-6.
8. Chabou, S., and M. Iglewski. "PICO Extraction by combining the robustness of machine-learning methods with the rule-based methods." In 2015 World Congress on Information Technology and Computer Applications (WCITCA), pp. 1-4. Ieee, 2015.
9. Huang, Ke-Chun, Charles Chih-Ho Liu, Shung-Shiang Yang, Furen Xiao, Jau-Min Wong, Chun-Chih Liao, and I-Jen Chiang. "Classification of PICO elements by text features systematically extracted from PubMed abstracts." In 2011 IEEE International Conference on Granular Computing, pp. 279-283. IEEE, 2011.
10. Noyes, Jane. "Never mind the qualitative feel the depth! The evolving role of qualitative research in Cochrane intervention reviews." Journal of Research in Nursing 15, no. 6 (2010): 525-534.
11. Li, Yikuan, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. "A comparative study of pretrained language models for long clinical text." Journal of the American Medical Informatics Association 30, no. 2 (2023): 340-347.
12. Wang, Qianying, Jing Liao, Mirella Lapata, and Malcolm Macleod. "PICO entity extraction for preclinical animal literature." Systematic Reviews 11, no. 1 (2022): 1-12.
13. Li, Jie, Gaihong Yu, and Zhixiong Zhang. "RCMR 280k: Refined Corpus for Move Recognition Based on PubMed." Data Intelligence,MIT Press Direct, April 27 2023.pp 1-23, https://doi.org/10.1162/dint\_a\_00214
14. Hu, Yan, Yong Chen, and Hua Xu. "Towards More Generalizable and Accurate Sentence Classification in Medical Abstracts with Less Data." Springer Journal of Health Informatics Research, (2022).https://doi.org/10.21203/rs.3.rs-2026270/v1
15. Hashimoto, Kai, and Ushio Inoue. "Automatic Generation of Structured Abstracts from Research Papers by using Deep Learning." In 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 424-429. IEEE, 2020.
16. Soleimani, Amir, Vassilina Nikoulina, Benoit Favre, and Salah Ait-Mokhtar. "Zero-Shot Aspect-Based Scientific Document Summarization using Self-Supervised Pre-training." In Proceedings of the 21st Workshop on Biomedical Language Processing, pp. 49-62. 2022.
17. Turchin, Alexander, Stanislav Masharsky, and Marinka Zitnik. "Comparison of BERT implementations for natural language processing of narrative medical documents." Informatics in Medicine Unlocked 36 (2023): 101139.
18. Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36, no. 4 (2020): 1234-1240.
19. Su, Peng, and K. Vijay-Shanker. "Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction." BMC bioinformatics 23, no. 1 (2022): 120.
20. Jin, Di, and Peter Szolovits. "Pico element detection in medical text via long short-term memory neural networks." In Proceedings of the BioNLP 2018 workshop, pp. 67-75. 2018. https://github.com/jind11/PubMed-PICO-Detection
21. Juliet Gray, Picking a PICO, Online Tutorial, Available Online: <https://libguides.siue.edu/c.php?g=333872&p=2244149>

1. \*Research supported by NSERC and MITACS.

   **Sabah Mohammed** is a full Professor of Computer Science with the Lakehead University, Thunder Bay Ontario P7B 5E1, Canada (phone 807-343-8777; fax: 807-7667243; e-mail: mohammed@lakeheadu.ca).

   **Hashmath Shaik** is a Graduate Student with Department of Computer Science, Lakehead University, Thunder Bay, Ontario P7B 5E1, Canada (e-mail: hshaik@lakeheadu.ca).

   **Jinan Fiaidhi** is a full Professor of Computer Science with the Lakehead University, Thunder Bay Ontario P7B 5E1, Canada (e-mail: jfiaidhi@lakeheadu.ca). [↑](#footnote-ref-0)