Unveiling Breast Cancer Causes Through Knowledge Graph Analysis and BioBERT-Based Factuality Prediction

Hasna El Haji[1][a], Nada Sbihi[1], Kaoutar El Handri[2,3][b], Adil Bahaj[1], Mohammed Elkasri[1], Amine Souadka[4][c] and Mounir Ghogho[1][d]

[1] *TiCLab, International University of Rabat, Rabat, Morrocco*

[2] *MedBiotech Laboratory, Faculty of Medicine and Pharmacy (FMPR), University Mohammed V, Rabat, Morocco*

[3] *Aivancity School of AI & Data for Business & Society, Cachan, France*

[4] *Surgical Oncology Department, National Institute of Oncology, University Mohammed V, Rabat, Morocco*

Keywords: Knowledge Graph, Breast Cancer, Bioinformatics, BioBERT, Factuality Prediction.

Abstract: Worldwide, millions of women are affected by breast cancer, with the impact significantly worsened in underserved regions. The profound effect of breast cancer on women's health has driven research into its causes, with the aim of developing methods for the prevention, diagnosis, and treatment of the disease. The significant influx of research on this subject is overwhelming and makes manual exploration arduous, which motivates automated knowledge exploration approaches. Knowledge Graphs (KGs) are one of these approaches that attracted significant attention in the last few years for their ability to structure and present knowledge, making it easier to explore and analyze. Current KGs that include causes of breast cancer are deficient in contextual information, highlighting the uncertainty of these causes (facts). In this work, we present a method for extracting a sub-graph of breast cancer causes and fine-tuning BioBERT to evaluate the uncertainty of these causes. Our automated approach, which simulates human annotation, computes uncertainty scores based on textual factuality and assesses cause reliability using a Closeness Score. We also create a web-based application for easy exploration[a].

[a] https://bckg.datanets.org/

1 INTRODUCTION

Breast cancer is a complex and multifaceted disease, affecting approximately 2.3 million women worldwide each year, according to the "WHO" (World Health Organization, 2024). With about one in eight women diagnosed during their lifetime (American Cancer Society, 2024), the disease significantly impacts quality of life, particularly in disadvantaged countries. This has led to extensive research in various disciplines to explore prevention (El Haji et al., 2024), diagnosis, and prognosis (El Haji et al., 2023). One crucial research focus is identifying the leading causes of breast cancer, as addressing the root cause can increase the success of treatments. However, abundant research papers on this topic can make manually exploring findings challenging and require multiple specialists' involvement. Automated knowledge extraction methods have been studied to answer this need.

There has been significant progress in automated knowledge extraction from scientific literature, particularly in the biomedical field (Hogan et al., 2021), (Zheng et al., 2021), (Wang et al., 2023). The aim is to organize and structure knowledge to facilitate exploration and discovery (Bahaj et al., 2022). In KGs, facts are organized into triplets, where each triplet comprises a head entity and a tail entity linked by a semantic relation. Mathematically, a KG $\mathcal{G}$ can be presented as a triplet $\mathcal{G} = (\mathcal{V}, \mathcal{R}, \mathcal{E})$, where $\mathcal{V}$ is the set of entities, $\mathcal{R}$ is the set of relations and $\mathcal{E}$ is the set of entities such that $\mathcal{E} = \{(h,r,t)|h,t \in \mathcal{V}, r \in \mathcal{R}\}$.

KGs can be utilized for knowledge exploration using visualization tools, improving search engine results, question-answering systems, recommendation

[a] https://orcid.org/0009-0000-0437-2731
[b] https://orcid.org/0000-0001-6732-2627
[c] https://orcid.org/0000-0002-1091-9446
[d] https://orcid.org/0000-0002-0055-7867

El Haji, H., Sbihi, N., El Handri, K., Bahaj, A., Elkasri, M., Souadka, A. and Ghogho, M.
Unveiling Breast Cancer Causes Through Knowledge Graph Analysis and BioBERT-Based Factuality Prediction.
DOI: 10.5220/0013179700003911
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2025) - Volume 2: HEALTHINF*, pages 141-148
ISBN: 978-989-758-731-3; ISSN: 2184-4305
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

systems (El Handri and Idrissi, 2020), and tasks such as node classification and link prediction (Huang et al., 2019), (Guo et al., 2020), (Rossi et al., 2021). In biomedicine, research in biomedical KGs has seen a surge in recent years. This is due to the influx of new biomedical publications in multiple disciplines on various medical concepts. Existing biomedical KGs in literature are either domain-specific or general. COVID19KG (Kejriwal, 2020) is a domain-specific KG with knowledge about COVID19. SemMedDB (Kilicoglu et al., 2012) is a general-purpose biomedical KG that contains a wide range of biomedical concepts. However, KGs assume the accuracy and reliability of their facts. Nonetheless, uncertainty can impact them (Sosa and Altman, 2022). This uncertainty arises from various sources of noise in the source data or in the extraction process. Uncertain knowledge graphs (UKGs) have been developed, whereby a confidence score is assigned to each fact, quantifying its validity (Chen et al., 2019). Formally, these UKGs assign a confidence score s_l to each triple $l = (h, r, t)$. This confidence score reflects the level of veracity of that fact in the KG.

In this direction, language modeling has gained considerable attention recently. The creation of the transformer-based architecture and self-supervised pretraining made considerable strides to have more reliable models. BERT (Devlin et al., 2018) is one such model. BERT uses a transformer-based architecture relying on a masked language modeling training process. BERT initiated a research direction, which extended from general language modeling to more domain-specific language models in law (Sun, 2023), biomedicine (Chen et al., 2023), marketing (Li et al., 2022) and other domains. These variants are generally fine-tuned versions of BERT, where a domain-specific corpus is curated and labeled, then used to retrain BERT. Language model fine-tuning is essential in modeling small datasets (Zhang et al., 2020) and low-resource languages (Hangya et al., 2022). One specific application of these language models is factuality prediction, which predicts the veracity of an event expressed in a sentence (Veyseh et al., 2019). Factuality prediction is generally formulated as a classification problem, where the model is tasked to predict the factuality class of a sentence (Kilicoglu et al., 2017). Other approaches, such as formulating the problem as a regression problem, where a factuality value that ranges between -3 and 3 is provided for each sentence. Negative and positive values signify negative and positive polarity, respectively (Stanovsky et al., 2017). The main difficulty in factuality prediction stems from the nature of factuality itself. Practically, factuality is expressed using certain cue words, and corresponding to their dependence on different sentence elements, the sentence factuality gets affected (Jiang and de Marneffe, 2021). This syntactic dependency referred to as text paradigmatics is generally challenging to model since most models focus on modeling syntax with no regard to long-range dependencies.

To address this gap, advancements in Natural Language Processing (NLP) have been significantly driven by novel pre-training techniques, such as those presented in XLNet, ERNIE, and ELECTRA. XLNet (Yang et al., 2019), with its generalized autoregressive pretraining, enhances language understanding by capturing bidirectional context through permutation-based training, which surpasses traditional models in handling complex language tasks. ERNIE (Zhang et al., 2019) further builds on this foundation by incorporating informative entities, offering a more context-aware representation that is particularly effective for tasks involving entity-level understanding. Meanwhile, ELECTRA (Clark et al., 2020) introduces a discriminator-based pre-training approach, improving the efficiency and effectiveness of language models by focusing on distinguishing real from generated tokens. In this evolving landscape, our study contributes by fine-tuning BioBERT (Lee et al., 2020) specifically for factuality prediction of triplets, leveraging its domain-specific pre-training for biomedical text.

Our work builds on these foundational approaches by applying BioBERT for factuality prediction. Aditionally, our work proposes a metric to evaluate the reliability of triplets within our knowledge sub-graph. This enables visibility and clarity for physicians and scientists working on breast cancer. Furthermore, we design a web application that serves as an interactive exploration platform displaying the entire sub-graph.

Our approach uniquely advances existing methodologies by addressing key limitations in handling uncertainty and factuality prediction within biomedical knowledge graphs. By integrating BioBERT—a domain-specific language model—with the construction of a breast cancer causes knowledge sub-graph, we introduce a novel framework that not only predicts factuality but also quantifies the reliability of triplets through a Closeness Score. This dual-layered approach significantly enhances the interpretability and trustworthiness of the extracted causal relationships. Unlike existing models that assume the veracity of their facts, our methodology explicitly evaluates and categorizes factuality, making it particularly impactful for applications in biomedical research where uncertainty can hinder effective decision-making.

The paper is structured as follows: section 2 details our methodology for constructing the knowledge sub-graph and evaluating the factuality of triplets. Section 3 presents our experiments and results, while sections 4 and 5 provide discussion and concluding remarks on the added value of our approach.

2 METHODOLOGY

The framework of our method consists of subgraph construction (extraction of triplets), fine-tuning BioBERT (Lee et al., 2020) for factuality prediction and triplet reliability evaluation based on Closeness Score. We detail each part in the subsections 2.1, 2.2 and 2.3, respectively. The overall workflow is shown in Figure 1.

2.1 Knowledge Sub-Graph Construction

For constructing our knowledge sub-graph of breast cancer causes, we utilize a generic biomedical literature database known as SemMedDB (Kilicoglu et al., 2012). SemMedDB is a valuable resource for integrating and analyzing information from biomedical literature. It captures relationships between concepts such as genes, diseases, drugs, and other biomedical concepts using multiple semantic relations that express general association, cause, influence and other relations. SemMedDB provides a rich network of interconnected biomedical knowledge and it includes contextual details like negation and modality and allows tracing back to PubMed articles via unique identifiers, enabling researchers to explore relationships and generate hypotheses grounded in biomedical literature.

Even though an API is available, we download all the resources to enable offline work and ensure efficient local processing. We extract triplets related to breast cancer causes for sub-graph construction and source sentences for factuality prediction. Each triplet represents a relationship between a cause and breast cancer. We also retrieve the paper's PubMed Identifier (PMID) and its citation count to further evaluate the relevance of the extracted triplets.

We are interested in the causes of breast cancer, thus we filter the extracted triplets related to breast cancer by retaining only those with the relationship pattern:
"Subject" $<$ *causes* $>$ "Breast cancer".
Each "Subject" is identified by a Concept Unique Identifier (CUI), a unique code used to retrieve information about entities across various medical databases. In our context, the object "breast cancer" refers specifically to female breast cancer, which can appear under different synonyms in the biomedical literature. We utilize UMLS to extract these synonyms (Rossi et al., 2021).

2.2 Fine-Tuning BioBERT for Factuality Prediction

Although the automatic construction of KGs offers a more practical alternative to manual annotation, it suffers multiple pitfalls. For example, according to SemMedDB, secondary Malignant neoplasm of the skin causes breast cancer. But, upon closer examination of the sentence: " We present the rare clinical entity of a breast cancer which was first diagnosed due to the skin metastasis away from the breast tumor". It is clear that the sentence does not allow us to confirm this relationship with certainty. Therefore, retrieving the sentence corresponding to each triplet is essential as it enriches our knowledge sub-graph.

Given the large number of triplets, it is difficult to manually assign a factuality score to each one. As a result, we use a language model to predict the factuality score of a triplet using the source sentence as input. To construct the factuality model, we use a factuality dataset (Kilicoglu et al., 2017). This dataset is built using 500 PubMed abstracts. Sentences in these abstracts were annotated to extract seven factuality values: fact, probable, possible, doubtful, counteract, uncommitted, conditional. Although the dataset contains 500 abstracts, the number of triples (subject, relation, object) extracted from them amounts to 3,149 (Kilicoglu et al., 2017). The dataset is unbalanced, with 87% of the sentences classified as facts, 4.5% probable, 3.8% uncommitted, 2% possible, 1.8% counterfact, 0.25% Doubtful, and just 1 case conditional. We removed the last two classes due to their minimal representation.

Sentences can act as explanatory variables in a predictive model, with factuality as the target variable. In this context, the objective is to leverage the information within the sentences to predict the factuality or outcome of interest. We use BioBERT (Lee et al., 2020) as the basis for constructing our predictive model since it is a common and effective approach, especially for tasks involving natural language processing and understanding. We fine-tune the pre-trained BioBERT model to adapt it to our specific NLP task: factuality classification. We use the "WeightedRandomSampler" sampling strategy, assigning higher probabilities to samples from underrepresented classes. This approach helps prevent the model from being biased towards the major-

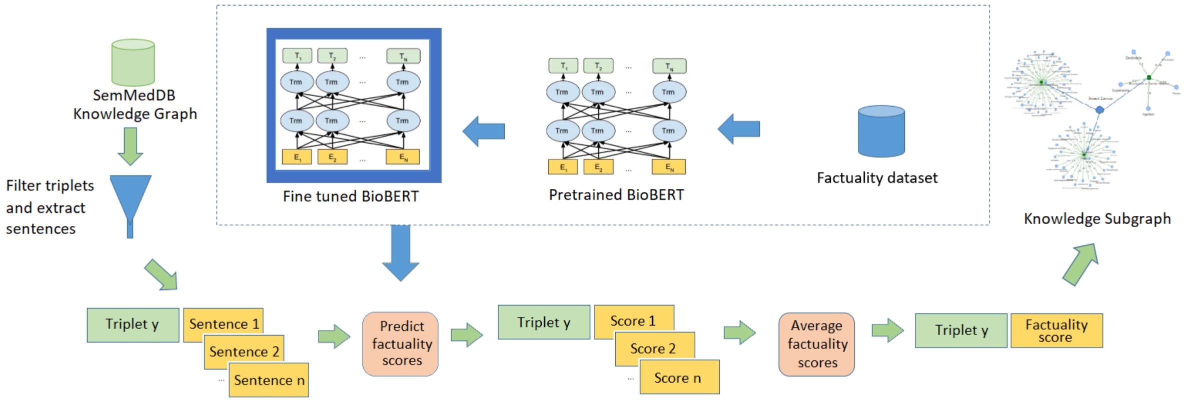


Figure 1: Sub-graph construction and factuality prediction: Triplets are extracted from the SemMedDB knowledge graph, filtered to include only cases where the "Subject" causes "breast cancer," followed by the extraction of sentences related to each triplet. For each triplet *y*, fine-tuned BioBERT is used to predict the factuality score for each sentence from which the triplet was derived. The factuality scores are averaged across sentences corresponding to the same triplet, resulting in a unique score assigned to each triplet (indicating the degree of factuality of a subject that causes breast cancer).

ity class and improve its ability to learn from the minority class samples. After testing different parameter values, the following combination gives us the best validation accuracy: a learning rate of 1e-5, bach size of 8, and number of epochs of 10.

As proposed by (Kilicoglu et al., 2017), we convert factuality classes to numeric scores as illustrated in Table 1. Finally, to calculate the factuality of a breast cancer cause, we compute the average of all factuality scores predicted from sentences involving that cause. This approach may be biased, as some breast cancer causes are cited numerous times while others are cited only once. Consequently, we propose further evaluation in the next section to assess cause reliability.

Table 1: Conversion of factuality classes to numeric scores (Kilicoglu et al., 2017).

Class	Factuality score
Fact	1
Probable	0.75
Possible	0.5
Uncommitted	0.25
Counterfact	0

2.3 Triplet Reliability Evaluation Based on Closeness Score

The assessment of triplet reliability in knowledge graphs is a multifaceted challenge that requires evaluating both the impact and contextual relevance of extracted relations. Previous studies have highlighted the role of citation counts as a proxy for significance in research evaluation, where more frequently cited works are often considered more impactful (Waltman, 2016). On the other hand, contextualized language models, such as BERT and its biomedical variants like BioBERT, have been increasingly used to assess the relevance of extracted triplets by capturing semantic similarity within specific contexts (Soares et al., 2019). Despite these advances, the integration of citation-based metrics with contextual relevance scores for triplet reliability assessment has not been fully explored.

We evaluate triplet reliability based on Closeness Score, which quantifies the degree of proximity of the triplet to the core objective of the publication. By assessing the closeness between the triplet's mention and the publication title, we can determine how relevant the triplet is to the study's main focus. This ensures that the triplet is not merely referenced as the result of other research but is central to the study itself.

For each triplet T_k, we identify the papers i that have cited it and retrieve the titles $Title_i$ of these papers using the PubMed API. Then, using BioBERT embeddings, we calculate the similarity score between each title and the sentence $S(T_k)$ from which the triplet was extracted in that paper. The overall triplet closeness score is computed as the mean of the similarity scores across all citing papers, where N is the number of citing papers (Equation 1).

$$Closeness(T_k) = \frac{1}{N}\sum_{i=1}^{N} Similarity(Title_i, S(T_k)) \quad (1)$$

For an effective visual representation of our findings, we design a web application using the Python framework Dash.

3 EXPERIMENTS AND RESULTS

As depicted in the framework in Figure 1, we represent all the potential causes of breast cancer as a knowledge sub-graph and we assign a degree of factuality to each cause linked to breast cancer. After filtering the triplets, there are 1,039 subjects identified as causing breast cancer. In terms of their types, the most frequently detected are hazardous or poisonous substances, genes or genomes, amino acids, peptides, and proteins. Figure 2 illustrates the most commonly identified types of causes in the literature; however, this does not necessarily imply that these causes are the most influential. Conversely, the figure highlights also types that are less studied and need further research and exploration by the scientific community.

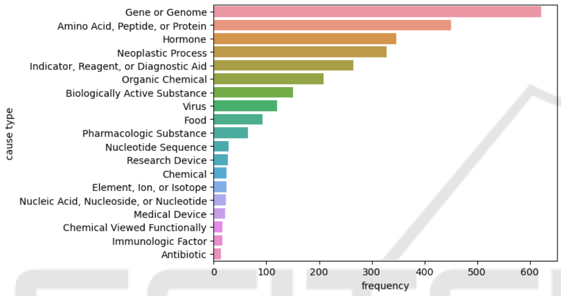


Figure 2: Top 20 breast cancer cause types.

As for factuality prediction, Table 2 summarizes the performance metrics of the fine-tuned BioBERT model for factuality prediction across various classes. The model demonstrates strong overall accuracy (0.927). It also showed very good performance in the "Fact" class, with a precision of 0.970, recall of 0.918, and an F1-score of 0.944, reflecting its high accuracy in identifying factual statements.

For the "Probable" class, the model shows a precision of 0.526 and a recall of 0.833, resulting in an F1-score of 0.645. This indicates that while the model is fairly effective at recognizing probable statements, there is room for improvement in its precision.

The "Possible" class has a precision of 0.600 and recall of 0.500, leading to an F1-score of 0.545. Despite achieving an overall accuracy of 0.927, the model's performance in predicting possible statements is less robust compared to the "Fact" class, suggesting that further refinement is needed.

The "Uncommitted" class presents a precision of 0.412 and recall of 0.778, with an F1-score of 0.538. This reflects moderate performance, indicating that while the model can identify uncommitted statements to some extent, its precision remains relatively low.

Finally, for the "Counterfact" class, the model exhibits a precision of 0.286 and recall of 0.667, resulting in an F1-score of 0.400. This performance suggests that the model has challenges in accurately identifying counterfactual statements, highlighting an area for potential improvement.

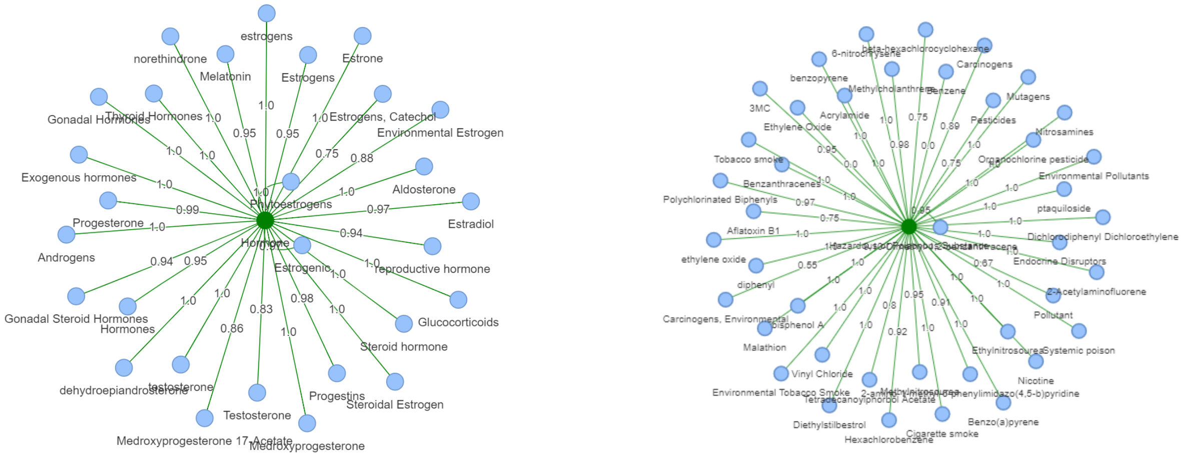
Table 2: Performance metrics of the fine-tuned BioBERT for factuality prediction.

Class	Precision	Recall	F1-Score	Accuracy
Fact	0.970	0.918	0.944	0.927
Probable	0.526	0.833	0.645	
Possible	0.600	0.500	0.545	
Uncommitted	0.412	0.778	0.538	
Counterfact	0.286	0.667	0.400	

Our web interface serves as an interactive exploration platform and consists of a dropdown list where the user can choose "All" to display the entire knowledge sub-graph or select a specific segment of the sub-graph (Figure 3). If the user is interested in viewing only the causes related to hormones (Figure 3a), they should select the "Hormones" option. This will display the causes categorized as "Hormones" each with the predicted factuality score. Similarly, Figure 3b highlights factors related to hazardous or poisonous substances that contribute to breast cancer. These visualizations incorporate factuality scores, offering valuable insights into the confidence levels associated with various causal relationships. These are just two sub-graphs of breast cancer causes, with the remaining causes accessible through our web application.

Finally, we notice that some breast cancer causes are cited numerous times while others are cited only once. Figure 4 depicts the most frequently cited causes. As shown, some causes have been cited over 200 times. However, among the 1,039 causes we extracted, over 90% are cited fewer than five times.

Figure 5 presents the calculated Closeness Scores for the extracted triplets. The x-axis shows the index of each triplet after sorting to provide a clear visual representation of how the scores change across the most to least relevant or popular triplets. Figure 5 shows that the closeness score starts high, around 0.90, and gradually decreases as the triplet index increases. This indicates that the first few triplets are more closely aligned with the main objective of the publication. As the index increases, the triplets become less relevant. Plotting all 1039 triplets could lead to a dense or unreadable chart. Limiting the view to the top 80 keeps the visualization clean and interpretable.



(a) Hormones

(b) Hazardous or poisonous substances

Figure 3: Segments of the knowledge sub-graph of breast cancer causes with corresponding factuality scores, displayed as screenshots from our web application.

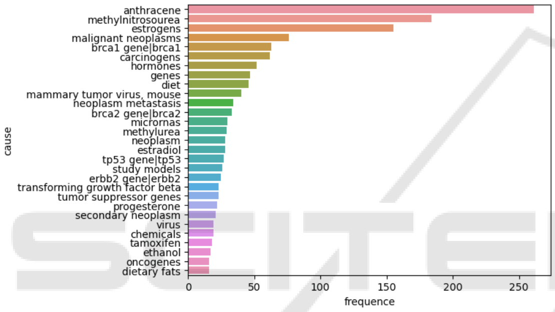


Figure 4: Causes of breast cancer with the highest citation frequency (frequency corresponds to the number of research articles from which the cause and its associated sentence were extracted).

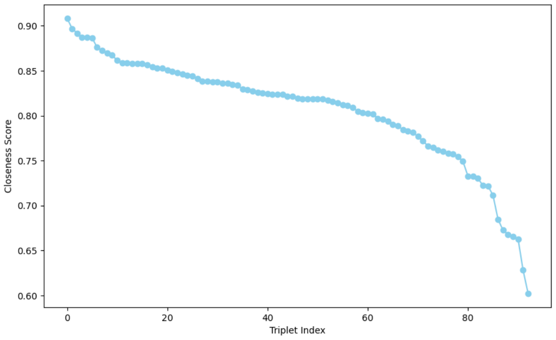


Figure 5: Closeness Scores: Measurement of the alignment between triplets and the primary focus of the scientific article in which they are mentioned.

4 DISCUSSION

Our framework (Figure 1) aids in summarizing the literature on the causes of breast cancer by providing a factuality score automatically computed using BioBERT. This factuality is subsequently categorized as reliable or unreliable based on the number of citations of the relationship and the alignment of sentences revealing the triplets with those papers. Here, we discuss the significance of our findings and their implications.

Figure 2 highlights the most frequently identified types of breast cancer causes in the biomedical literature, such as hazardous substances and genetic factors. This distribution not only reflects existing research priorities but also uncovers underexplored areas, such as less-studied molecular factors. These findings emphasize the need for targeted investigations in less-documented categories, guiding future research toward filling critical gaps in the literature.

The visual segments in Figure 3, focusing on hormones and hazardous substances, demonstrate the breadth of our knowledge sub-graph and its capacity to categorize causal relationships with associated factuality scores. This categorization enables researchers to focus on specific domains while maintaining a comprehensive understanding of the underlying relationships. The integration of factuality scores into these segments provides clarity on the reliability of each causal link, enhancing trust in automated knowledge extraction systems.

Figure 4 underscores the uneven citation distribution among the identified breast cancer causes. While some causes are supported by extensive literature, others are based on minimal citations, indicating variability in research focus. This variability emphasizes the importance of combining citation counts with contextual evaluation, such as the Closeness Score, to ensure a balanced representation of causes in the knowledge graph.

As depicted in Figure 5, the Closeness Score provides a quantitative measure of triplet relevance to the core objective of each publication. Higher scores correspond to triplets directly aligned with the study's primary focus, while lower scores represent peripheral relationships. This metric demonstrates the effectiveness of integrating contextual relevance with citation-based significance, addressing a key limitation in existing knowledge graph methodologies.

As for BioBERT, while it performs well in identifying factual statements, its performance varies across different classes, particularly concerning precision and F1-score. The results highlight areas where the model's predictive capabilities could be improved, especially for less frequently occurring classes.

Several aspects of our work are still under development. For instance, the model struggles with detecting negation. For example, in the sentence, "These findings do not support a role for HAAs from meat or NAT2 in the etiology of breast cancer," the algorithm incorrectly classifies the factuality of meat causing breast cancer as probable, disregarding the negation.

Another challenge is illustrated in the relationship between tea and breast cancer. In the sentence, "The role of tea in the aetiology of breast cancer is controversial," the algorithm fails to interpret the term "controversial" correctly. Although the relationship is a fact, the algorithm misclassifies it as probable due to its misunderstanding of the term.

Incorporating crowd-sourcing could improve the reliability of this knowledge sub-graph. Volunteers could assist in classifying relationships and validating the classes detected by the model, with these corrections serving as training examples for further refinement.

5 CONCLUSION

Ultimately, this study showcases the potential of artificial intelligence to revolutionize our understanding of breast cancer. We introduce a knowledge sub-graph to illustrate the causes contributing to breast cancer. By mapping potential causal relationships and assigning factuality scores to these links, the sub-graph provides a valuable resource for researchers and clinicians. The interactive web application enhances usability, allowing for customized data exploration. While offering a comprehensive overview, the sub-graph also highlights areas for further investigation, including expanding the representation of viral factors and elucidating the underlying biological mechanisms. Furthermore, we plan to include gene interactions and causal antecedents in our knowledge sub-graph to better understand the causal pathways leading to breast cancer.

The results underscore the importance of collaborative efforts across disciplines to address the challenges posed by breast cancer. By combining this knowledge sub-graph with clinical data, researchers can develop more accurate risk prediction models and improve patient care. Future research could expand the sub-graph's scope, validate causal relationships through experimental studies, and explore integrating additional data from domain-specific sources.

Limitations and Future Directions

As a limitation, this study lacks a comprehensive comparison with state-of-the-art models. To address this, we are adapting ERNIE and XLNet to our factuality dataset for prediction, and we will compare their performance with BioBERT to assess potential improvements in factuality classification.

Another limitation of the current study is the reliance on a single biomedical knowledge graph (SemMedDB) for triplet extraction. Future work will involve integrating additional datasets from domain-specific sources, such as clinical trial repositories, genetic databases, and epidemiological studies. These diverse datasets will enrich the knowledge sub-graph, capturing a wider range of causal relationships and improving the comprehensiveness of our approach.

Finally, the inclusion of real-world clinical data, such as electronic health records (EHRs) and patient registries, will provide context-specific insights into breast cancer causes and their interactions. This integration will allow for personalized causal inference, aligning our findings more closely with clinical practice.

REFERENCES

American Cancer Society (2024). American cancer society. https://www.cancer.org/. Accessed: 2024-06-30.

Bahaj, A., Lhazmir, S., Ghogho, M., and Benbrahim, H. (2022). Covid-19-related scientific literature exploration: Short survey and comparative study. *Biology*, 11(8):1221.

Chen, Q., Du, J., Hu, Y., Keloth, V. K., Peng, X., Raja, K., Zhang, R., Lu, Z., and Xu, H. (2023). Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv preprint arXiv:2305.16326*.

Chen, X., Chen, M., Shi, W., Sun, Y., and Zaniolo, C. (2019). Embedding uncertain knowledge graphs. In

Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 3363–3370.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

El Haji, H., Sbihi, N., Guermah, B., Souadka, A., and Ghogho, M. (2024). Epidemiological breast cancer prediction by country: A novel machine learning approach. *PLOS ONE*, 19(8):e0308905.

El Haji, H., Souadka, A., Patel, B. N., Sbihi, N., Ramasamy, G., Patel, B. K., Ghogho, M., and Banerjee, I. (2023). Evolution of breast cancer recurrence risk prediction: a systematic review of statistical and machine learning–based models. *JCO Clinical Cancer Informatics*, 7:e2300049.

El Handri, K. and Idrissi, A. (2020). Parallelization of topk algorithm through a new hybrid recommendation system for big data in spark cloud computing framework. *IEEE Systems Journal*, 15(4):4876–4886.

Guo, Q., Zhuang, F., Qin, C., Zhu, H., Xie, X., Xiong, H., and He, Q. (2020). A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3549–3568.

Hangya, V., Saadi, H. S., and Fraser, A. (2022). Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11993–12006.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., et al. (2021). Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37.

Huang, X., Zhang, J., Li, D., and Li, P. (2019). Knowledge graph embedding based question answering. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 105–113.

Jiang, N. and de Marneffe, M.-C. (2021). He thinks he knows better than the doctors: Bert for event factuality fails on pragmatics. *Transactions of the Association for Computational Linguistics*, 9:1081–1097.

Kejriwal, M. (2020). Knowledge graphs and covid-19: opportunities, challenges, and implementation. *Harv. Data Sci. Rev*, 11:300.

Kilicoglu, H., Rosemblat, G., and Rindflesch, T. C. (2017). Assigning factuality values to semantic relations extracted from biomedical research literature. *PloS one*, 12(7):e0179926.

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., and Rindflesch, T. C. (2012). Semmeddb: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Li, P., Castelo, N., Katona, Z., and Sarvary, M. (2022). Language models for automated market research: A new way to generate perceptual maps. *Available at SSRN 4241291*.

Rossi, A., Barbosa, D., Firmani, D., Matinata, A., and Merialdo, P. (2021). Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(2):1–49.

Soares, L. B., FitzGerald, N., Ling, J., and Kwiatkowski, T. (2019). Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Sosa, D. N. and Altman, R. B. (2022). Contexts and contradictions: a roadmap for computational drug repurposing with knowledge inference. *Briefings in Bioinformatics*, 23(4):bbac268.

Stanovsky, G., Eckle-Kohler, J., Puzikov, Y., Dagan, I., and Gurevych, I. (2017). Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 352–357.

Sun, Z. (2023). A short survey of viewing large language models in legal aspect. *arXiv preprint arXiv:2303.09136*.

Veyseh, A. P. B., Nguyen, T. H., and Dou, D. (2019). Graph based neural networks for event factuality prediction using syntactic and semantic structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4393–4399.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of informetrics*, 10(2):365–391.

Wang, Z., Nie, H., Zheng, W., Wang, Y., and Li, X. (2023). A novel tensor learning model for joint relational triplet extraction. *IEEE Transactions on Cybernetics*, 54(4):2483–2494.

World Health Organization (2024). World health organization. Accessed: 01:08:2024, url = https://www.who.int/en,.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., and Artzi, Y. (2020). Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. (2019). Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

Zheng, W., Wang, Z., Yao, Q., and Li, X. (2021). Wrtre: Weighted relative position transformer for joint entity and relation extraction. *Neurocomputing*, 459:315–326.