Knowledge Graph of Breast Cancer Prevention and Treatment: based on Literature Data

Shuyan Jin, Haobin Liang, Wenxia Zhang, Huan Li

Submitted to: JMIR Medical Informatics
on: August 26, 2023

Table of Contents

Knowledge Graph of Breast Cancer Prevention and Treatment: based on Literature Data

Shuyan Jin[1] PhD; Haobin Liang[2]; Wenxia Zhang[1]; Huan Li[1]

[1]Department of Prevention Care Shenzhen Maternity and Child Healthcare Hospital Shenzhen CN
[2]School of Economics and Statistics Guangzhou University Guangzhou CN

Corresponding Author:
Shuyan Jin PhD
Department of Prevention Care
Shenzhen Maternity and Child Healthcare Hospital
No. 2004, Hongli Road, Futian District, Shenzhen
Shenzhen
CN

Abstract

Background: The incidence of breast cancer has remained high and continues to rise since the 21st century. Consequently, there has been a significant increase in research efforts focused on breast cancer prevention and treatment. Despite the extensive body of literature available on this subject, systematic integration is lacking. To address this issue, knowledge graphs have emerged as a valuable tool. By harnessing their powerful knowledge integration capabilities, knowledge graphs offer a comprehensive and structured approach to understanding breast cancer prevention and treatment.

Objective: We aimed to build a knowledge graph to integrate information related to breast prevention and treatment.

Methods: We used MESH terms to search for clinical trial literature on breast cancer prevention and treatment published on PubMed between 2018 and 2021. We downloaded triplet data from SemmedDB and matched them with the retrieved literature to obtain triplet data for the target articles. We visualized the triplet information using NetworkX for knowledge discovery.

Results: Within the scope of literature research in the past five years, Malignant neoplasm appeared most frequently (42.32%). Pharmacotherapy (19.22%) was the primary treatment method, with Paclitaxel (9.83%) being the most commonly used therapeutic drug. Through the analysis of the knowledge graph, we have discovered that there exists a complex network relationship between the treatment methods, therapeutic drugs, and preventive measures for different types of breast cancer, rather than a simple linear correlation.

Conclusions: This study constructed a knowledge graph for breast cancer prevention and treatment, which enabled the integration and knowledge discovery of relevant literature in the past five years. Researchers can gain insights into treatment methods, drugs, preventive knowledge regarding adverse reactions to treatment, and the associations between different knowledge domains from the graph.

(JMIR Preprints 26/08/2023:52210)
DOI: https://doi.org/10.2196/preprints.52210

Preprint Settings

Original Manuscript

Knowledge Graph of Breast Cancer Prevention and Treatment: based on Literature Data

Shuyan Jin[1*], Haobin Liang[2], Wenxia Zhang[1], Huan Li[1]

Affiliations:
[1] Department of Prevention Care, Shenzhen Maternity and Child Healthcare Hospital, Guangzhou, China
[2] School of Economics and Statistics, Guangzhou University, Guangzhou, China

*** Correspondence:**
Shuyan Jin, PhD
Shenzhen Maternity □ Child Healthcare Hospital
No. 2004, Hongli Road, Futian District, Shenzhen
China
Phone: (86)13823373712
Email: 13823373712@163.com

Abstract
Background: The incidence of breast cancer has remained high and continues to rise since the 21st century. Consequently, there has been a significant increase in research efforts focused on breast cancer prevention and treatment. Despite the extensive body of literature available on this subject, systematic integration is lacking. To address this issue, knowledge graphs have emerged as a valuable tool. By harnessing their powerful knowledge integration capabilities, knowledge graphs offer a comprehensive and structured approach to understanding breast cancer prevention and treatment.
Objective: We aim to integrate literature data on breast cancer treatment and prevention, build a knowledge graph, and provide support for clinical decision-making.
Methods: We used MESH terms to search for clinical trial literature on breast cancer prevention and treatment published on PubMed between 2018 and 2022. We downloaded triplet data from SemmedDB and matched them with the retrieved literature to obtain triplet data for the target articles. We visualized the triplet information using NetworkX for knowledge discovery.
Results: Within the scope of literature research in the past five years, Malignant neoplasm appeared most frequently (42.3%, 587/1,387). Pharmacotherapy (19.2%, 267/1,387) was the primary treatment method, with trastuzumab (12.4%, 113/908), being the most commonly used therapeutic drug. Through the analysis of the knowledge graph, we have discovered that there exists a complex network relationship between the treatment methods, therapeutic drugs, and preventive measures for different types of breast cancer.
Conclusions: This study constructed a knowledge graph for breast cancer prevention and treatment, which enabled the integration and knowledge discovery of relevant literature in the past five years. Researchers can gain insights into treatment methods, drugs, preventive knowledge regarding adverse reactions to treatment, and the associations between different knowledge domains from the graph.

Keywords: knowledge graph; breast cancer; treatment; prevention

Introduction

Breast cancer is the most common malignant tumor in women worldwide, with a reported death toll exceeding 600,000 in 2018 alone [1]. Breast cancer has emerged as the most prevalent cancer and a primary cause of mortality among women. The global incidence of new cases of female breast cancer witnessed a sharp increase from 1.05 million in 2000 to 2.09 million in 2018 [2]. In 2020, global cancer burden data revealed that new breast cancer cases reached 2.26 million, constituting 11.7% of all newly diagnosed cancer cases worldwide. The newly reported mortality cases numbered 0.68 million, representing 6.9% of global newly reported death cases [3]. Factors such as old age, young age at menarche, family history of breast cancer, smoking, and drinking alcohol increase the risk of breast cancer [4-6]. On the contrary, regular physical exercise, breastfeeding, regular work and rest, and intake of fruits, vegetables, whole grains and dietary fiber can appropriately reduce the risk of breast cancer [7]. Various treatment methods are used for breast cancer patients, including surgery, radiation therapy, endocrine therapy, chemotherapy, and targeted therapy. So far, most countries have primarily focused on population education for breast cancer prevention, including encouraging increased physical activity, controlling body mass index, and limiting alcohol intake [8]. Despite the increasing number of research literature, a large amount of literature on breast cancer prevention and treatment has not been systematically integrated. Knowledge graph technology allows for the independent connection and integration of disparate literature, resulting in a more comprehensive and cohesive knowledge framework.

Knowledge Graph is a knowledge repository proposed by Google in 2012 to enhance the functionality of search engines. It describes concepts and their relationships in the real world using triplets in the form of entity-relation-entity [9]. Knowledge graphs can integrate information from diverse sources and domains, including text, databases, and web pages, and intricately interlink them. These integrations serve to mitigate information silos, fostering the establishment of a more comprehensive knowledge framework. Knowledge graphs have been widely used in various fields, such as medicine, network security, journalism, finance, and education [10]. Knowledge graphs in the biomedical domain have applications in studies related to disease associations [11], genomics [12], drug interactions [13], and offering support for physicians in formulating individualized treatment regimens [14]. Presently, there exist well-established knowledge graphs, including DisGeNET [15], which integrates information on the associations between genes and diseases; DrugBank [16], a comprehensive bioinformatics and cheminformatics knowledge base; and ClinVar [17], a compilation of genetic variation information from diverse laboratories worldwide. One study extracted breast cancer-related features from Chinese breast cancer mammography reports and built a knowledge graph for diagnosing breast cancer by combining diagnosis and treatment guidelines and insights from clinical experts [18]. Another study integrated triples from clinical guidelines, medical encyclopedias, and electronic medical records to build a breast cancer knowledge graph [19]. Despite a small number of scholars having constructed knowledge graphs for breast cancer, the varied emphases and diverse data sources employed render their applicability limited. A knowledge graph specifically focused on the prevention and treatment of breast cancer has not been constructed at present. Therefore, this study primarily collects information related to the prevention and treatment of breast cancer to construct a knowledge graph.

In the biomedical field, there are already mature tools (e.g., SemRep) for extracting knowledge from medical texts. SemRep is a natural language processing program based on UMLS, which performs operations such as text tokenization, syntactic analysis, part-of-speech disambiguation, phrase mapping, semantic predicate normalization, and syntactic constraints [20]. It extracts entities and relationships from biomedical texts and outputs triplets stored in SemmedDB [21]. SemMedDB currently encompasses details on approximately 96.3 million predications derived from all PubMed citations (around 29.1 million citations) and serves as the foundation for the Semantic MEDLINE application [22]. We downloaded the entity and relationship data provided by the SemMedDB. NetworkX is an open-source library for Python, primarily designed for creating, analyzing, and visualizing complex network structures. NetworkX plays a significant role in knowledge visualization, facilitating users in intuitively presenting and comprehending intricate knowledge graphs or network data.

Methods

Data Source

We conducted a search on PubMed using MESH terms: “breast cancer”, “prevention”, and “treatment”, covering the period from January 1, 2018, to December 31, 2022, and the study type was clinical trials. A total of 3,589 articles were retrieved. We obtained the entity and relationship data from SemMedDB.

Data Processing and Construction of Knowledge Graph

We matched the PMIDs of the retrieved articles with the database and extracted the corresponding triplet information. We initially obtained 33,060 SPO triplets of data.
Next, we made improvements according to the SPO cleaning principles proposed by Fiszman (relevance, connectivity, novelty, and significance) [23]. We combined them with expert manual screening to ensure that the selected SPO triplets have a higher relevance. In the improved process, we did not predefine semantic patterns. Instead, we used a series of cleaning operations to select core SPO triplets and connected SPO triplets, eliminating SPO triplets lacking specific information and those that appeared only once in the frequency. The specific process is as follows:
1. In the same article, there may be repeated occurrences of identical SPO triplets. To maintain equal contribution from each article, we counted the repeated SPO triplets once within the same article.
2. To ensure statistical reliability, we calculated the occurrence frequency of each SPO triplet across different articles. SPO triplets with low occurrence frequencies may lack statistical significance. Therefore, we filtered SPO triplets with frequencies greater than or equal to two.
3. Based on expert domain knowledge, we manually screened the selected SPO triplets with frequencies greater than or equal to two to identify those of research value.
Finally, we obtained 25,449 SPO triplets data. We imported the filtered SPO triplets information into the NetworkX for visual analysis to explore knowledge and information related to breast cancer prevention and treatment.

All analyses were conducted in Python program (version 3.11.3), primarily utilizing Pandas, Matplotlib, WordCloud, and NetworkX packages [24-27].

Results

Summary of Included Literatures

A total of 3,589 articles were published in 618 different journals. Among them, 191 papers were published in the same journal, while 293 journals had only one article published. The journals were ranked based on the number of publications, and the top 100 journals accounted for 2,631 articles, which is 73.30% of the total.

Semantic Relationships and Semantic Patterns

We mainly summarize semantic associations into three types: treatment and prevention, influencing/associated factors, and related diseases (Supplementary Table 1). Regarding treatment and prevention, the relationships include TREATS, ADMINISTERED_TO, USES, and PREVENTS, representing treatment drugs, surgeries, and preventive measures for breast cancer. Regarding influencing/associated factors, the relationships include ASSOCIATED_WITH, AFFECTS, and CAUSES, which represent diseases' impact and etiological factors. Regarding related diseases, the relationship COEXISTS_WITH represents the coexistence between different diseases. In the semantic patterns involving treatment (TREATS), the *topp-TREATS-neop* and *topp-TREATS-podg* have appeared over 1000 times.

Summary of SPO Triples

In terms of breast tumors, malignant neoplasms had the highest frequency, accounting for 42.3% (587/1,387) of the total, followed by triple negative breast neoplasms (4%, 56/1,387) and Her2-positive carcinoma of the breast (4%, 54/1.387) (Figure 1 and Supplementary Figure 1).

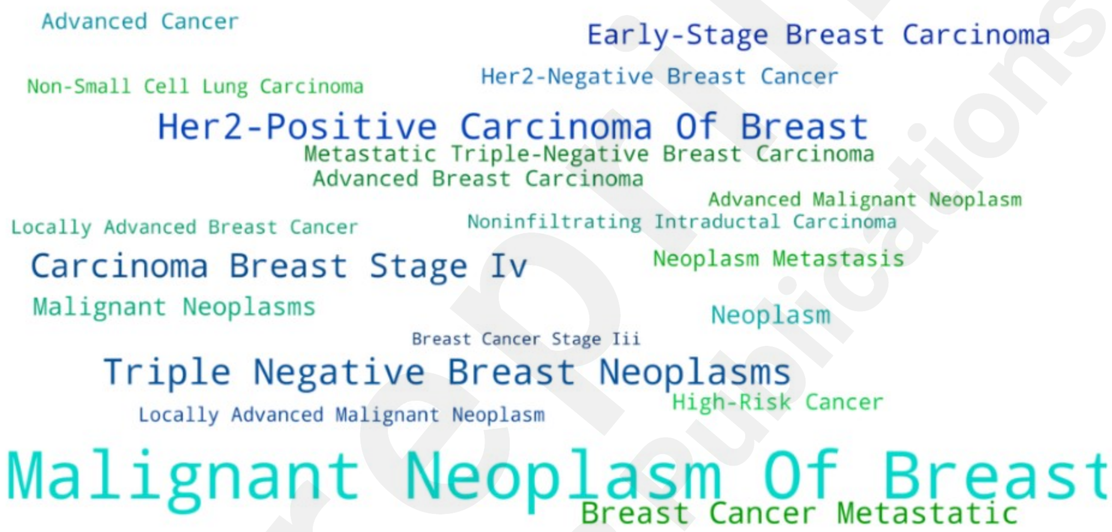


Figure 1 Breast cancer subtypes and stages.

Pharmacotherapy is the most common treatment method, accounting for 19.2% (267/1,387) of the overall frequency. Additionally, other high-frequency treatment modalities include adjuvant chemotherapy (4%, 54/1,387), radiation therapy (3%, 48/1,387), neoadjuvant therapy (6%, 88/1,387), and hormone therapy (3%, 68/1,387) (Figure 2 and Supplementary Figure 2). In breast cancer treatment drugs, trastuzumab (12.4%, 113/908), capecitabine (6%, 56/908), aromatase inhibitors (5%, 41/908), immunologic adjuvants (5%, 41/908) and paclitaxel (4%, 36/1,387) have a relatively high frequency of occurrence (Figure 3 and Supplementary Figure 3).

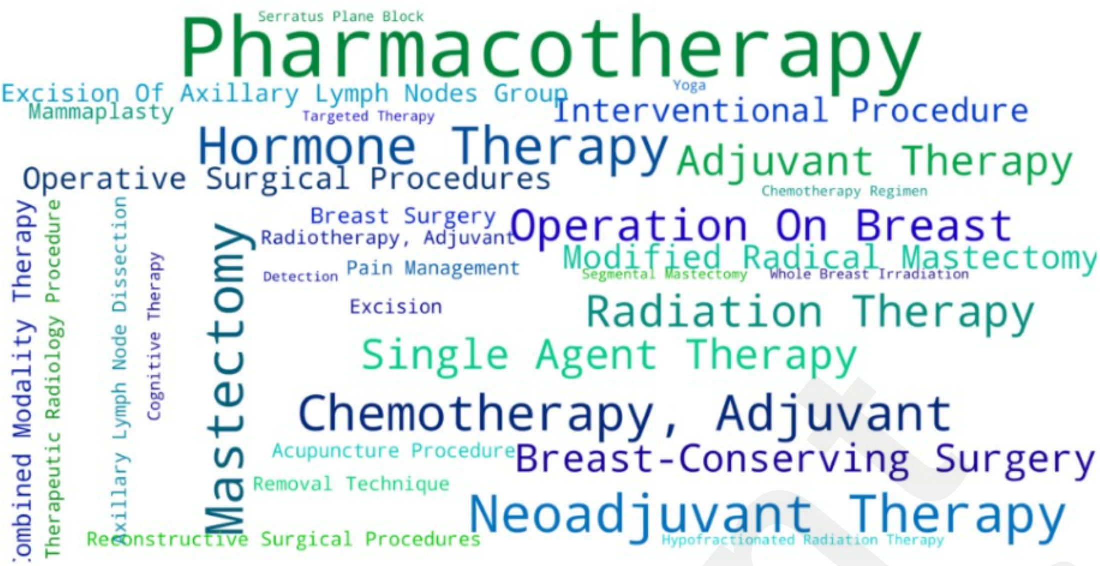


Figure 2 Treatment of breast cancer.

Figure 3 Drugs for breast cancer.

Breast Cancer Knowledge Graph

We visualized the SPO triples and displayed three subgroups: breast cancer treatment methods, therapeutic drugs, and relevant preventive measures. Figure 4 shows the relationship between different subtypes and stages of breast cancer and treatment methods. In different subtypes of breast cancer, the highest frequency is observed in malignant neoplasm of the breast, with pharmacotherapy having the highest frequency among various treatment modalities. Different subtypes simultaneously correspond to multiple treatment modalities; likewise, a single treatment modality corresponds to multiple breast cancer subtypes.

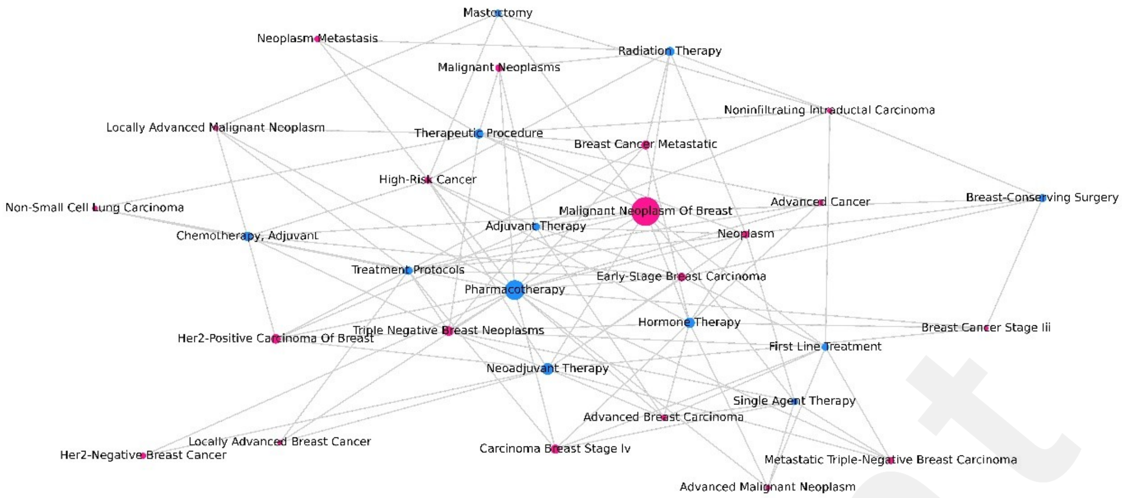


Figure 4 Relationship between different subtypes and stages of breast cancer and treatment methods.

Figure 5 shows the relationship between different subtypes and stages of breast cancer and drugs. Among the therapeutic drugs for breast cancer, trastuzumab has the highest frequency and corresponds to the most types of breast cancer. Capecitabine, paclitaxel□aromatase inhibitors and immunologic adjuvants also have relatively high frequencies. In comparison, immunologic adjuvants have the fewest connections with different types of breast cancer.

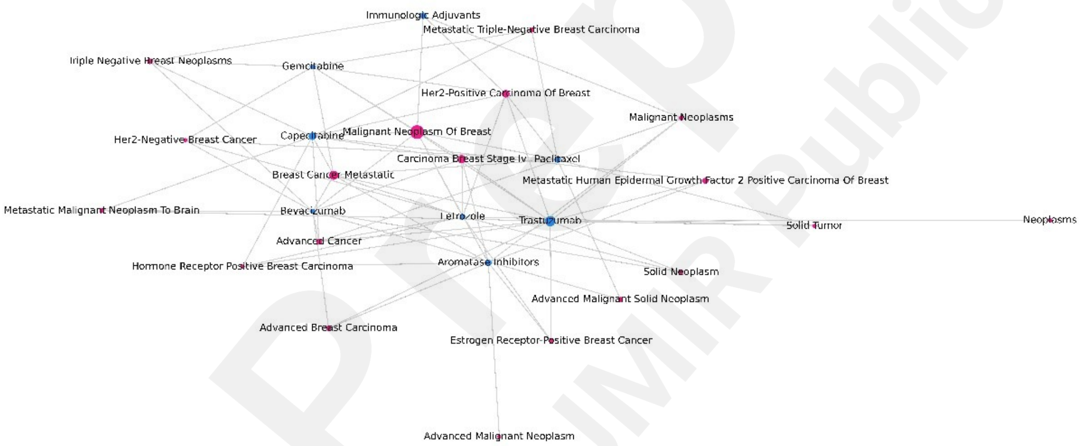


Figure 5 Relationship between different subtypes and stages of breast cancer and drugs.

Figure 6 shows the relationship between breast cancer treatment and adverse reactions. The pharmacotherapy is associated with neuropathy, onycholysis, heart neutropenia failure, alopecia, febrile neutropenia, anemia, stomatitis, leukopenia, thrombocytopenia, premature menopause and gastrointestinal dysfunction. Additionally, multiple nodes are connected, forming multiple pathways, such as pharmacotherapy-febrile neutropenia-adjuvant chemotherapy; pharmacotherapy-leukopenia-breast cancer therapeutic procedure-osteoporosis.

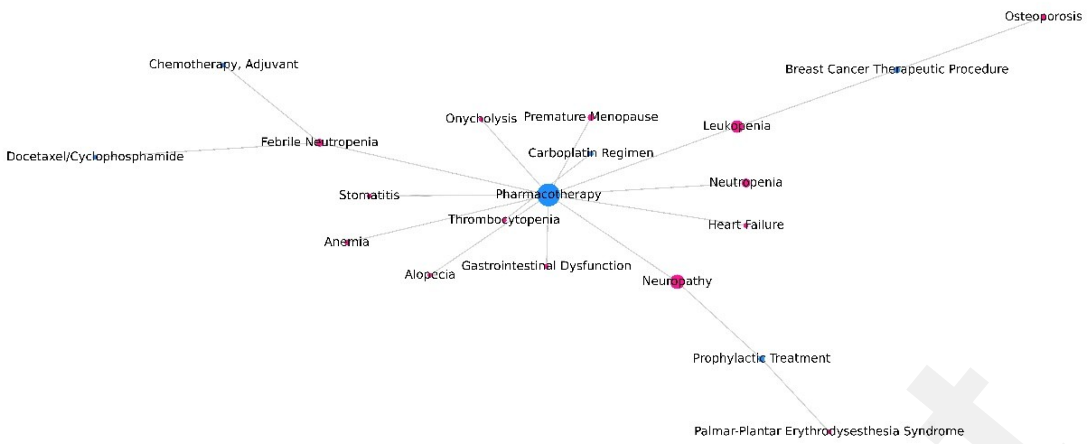


Figure 6 Relationship between breast cancer treatment and adverse reactions.

Figure 7 shows the relationship between adverse events after between breast cancer treatment and preventive measures. Peripheral neuropathy is associated with cryotherapy, low-level laser therapy, compression procedure, acupuncture procedure, pharmacotherapy and massage. Lymphedema is associated with resistance education, axillary lymph node dissection, physical therapy, excision of axillary lymph nodes group, and drainage of lymphatics. Early radiation dermatitis is associated with topical administration and bleomycin/cisplatin/methotrexate protocol. In addition, there are some adverse reactions with relatively few treatment measures, such as stomatitis-diet, alopecia-scalp cooling.

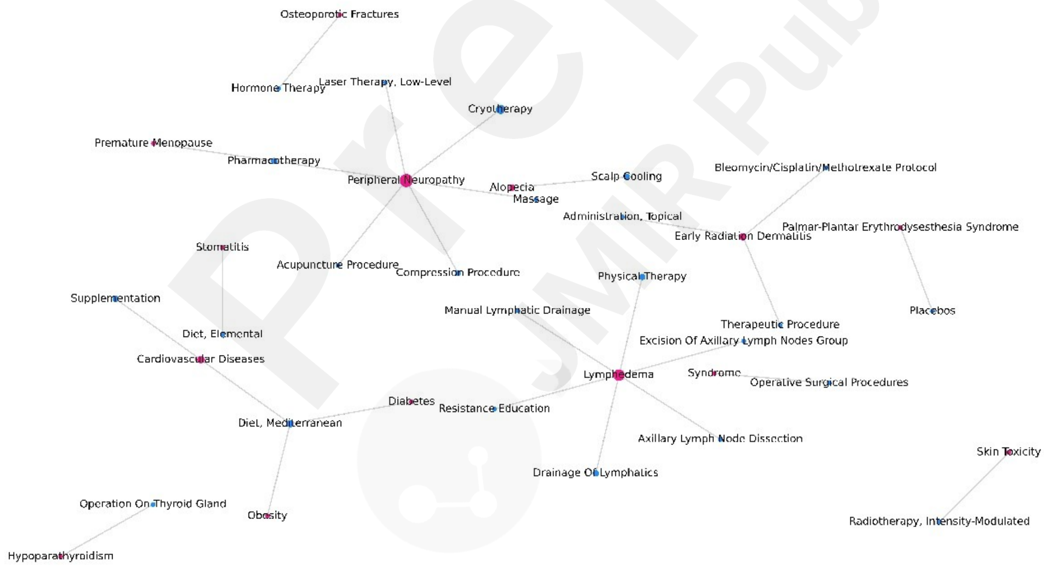


Figure 7 Relationship between adverse reactions after breast cancer treatment and preventive measures.

We performed a relationship visualization to understand better the association between types of breast cancer, treatments, drugs, and genes. Figure 8 intuitively reflects the high frequency of malignant neoplasm of the breast, pharmacotherapy, and trastuzumab. In addition, breast malignant tumors are associated with multiple genes, such as the PIK3CA gene, PDGFRB gene, PTEN gene, and ERBB2 gene.

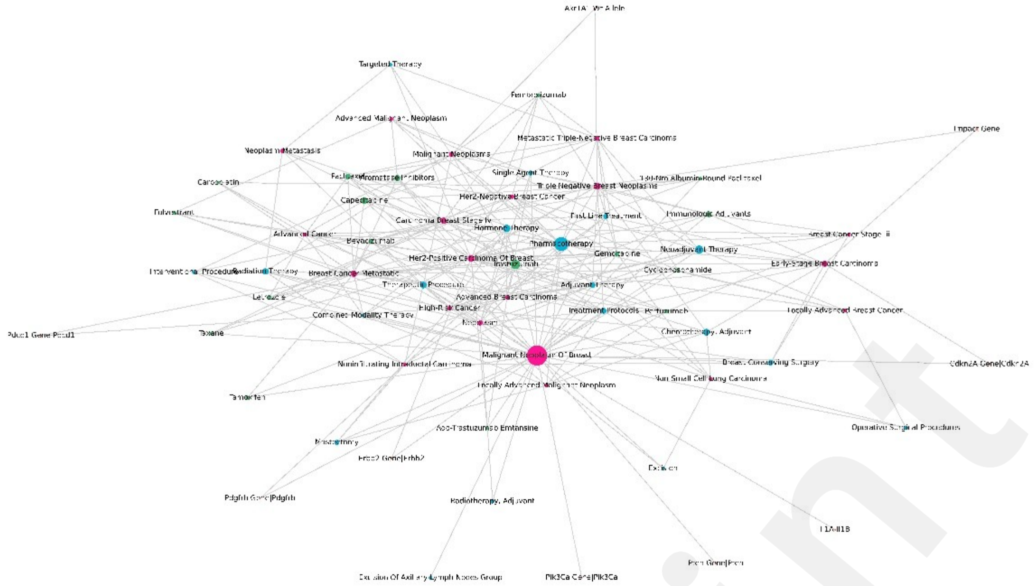


Figure 8 Relationship between types of breast cancer, treatments, drugs, and genes.

Discussion

Knowledge graphs constructed in this study helps researchers understand the research hotspots in breast cancer over the past five years. The complex network involving treatment methods, drugs, adverse reactions, preventive measures, and genes in breast cancer can assist clinicians in making decisions that comprehensively consider multiple aspects, which aids in making decisions that are most beneficial to patients. Additionally, the knowledge graph allows for personalized considerations based on specific genes for individualized patients.

This study found that from 2018 to 2022, breast malignancies appeared most frequently in the literature and were the primary concern for researchers. Research interest in TNBC is higher than other subtypes. This phenomenon may be due to the higher risk of recurrence and poor prognosis of early-stage TNBC patients [10], making it a subject of greater concern to clinicians and researchers. Among treatment modalities, pharmacotherapy receives the highest attention. Pharmacotherapy for breast cancer primarily involves chemotherapy, endocrine therapy, targeted therapy [28]. Compared to traditional surgery and radiotherapy, pharmacotherapy can more precisely intervene in the growth and division of cancer cells by targeting specific molecules or cellular structures, which reduces damage to normal cells and allows for the formulation of personalized treatment plans based on the patient's genotype and molecular characteristics [29]. Medications circulating through the bloodstream can also act on cancer cells throughout the body, preventing cancer cell metastasis. These advantages of pharmacotherapy may be related to the heightened emphasis on pharmacotherapy over the past five years. Trastuzumab receives the highest attention in breast cancer pharmacotherapy, which is a specific cancer-targeting medication used in the treatment of cancers characterized by elevated levels of human epidermal growth factor receptor 2 (HER2) protein [30].

Pharmacotherapy is associated with various adverse reactions, including neutropenia, neuropathy, onycholysis, heart failure, alopecia, and febrile neutropenia. Among these adverse reactions, peripheral neuropathy and lymphedema have the most corresponding preventive and treatment measures, with lymphedema being a common complication after surgery [31]. However, there is limited research on how to prevent and treat the potential adverse reactions of pharmacotherapy, and further studies are needed. Various adverse effects of breast cancer treatment may reduce patients' adherence to treatment. Therefore, when clinicians choose different treatments and drugs, they should pay close attention to their potential adverse reactions and how to prevent or mitigate them.

In existing knowledge graphs related to breast cancer, one study from China constructed a knowledge graph using electronic medical records, clinical guidelines, and expert opinions, primarily focusing on breast cancer diagnosis [18]. Another study by Chinese scholars also employed data from various sources, including clinical guidelines, medical encyclopedias, and electronic medical records, to construct a knowledge graph primarily applied to medical knowledge question-answering and medical record retrieval [19]. These studies utilized data from multiple sources, including structured, unstructured, and semi-structured data. Data extraction and accuracy face challenges. Therefore, they employed neural network models for training and calculated a series of metrics to ensure data accuracy. For instance, they utilized BERT + Bi-LSTM + CRF for textual data to achieve named entity recognition. In this study, SemMedDB was used as the data source, and the database was constructed by extracting semantic information from PubMed using SemRep, which demonstrated good performance in biomedical text [32].

In summary, the knowledge graph constructed in this study for breast cancer treatment and prevention encompasses information on different stages, subtypes of breast cancer, treatment modalities, medications, adverse reactions, and preventive measures. This knowledge forms a complex network, providing clinical practitioners with a comprehensive and referenced knowledge base. We recommend that clinical practitioners apply our research findings in several aspects. Firstly, clinicians can gain insights into the current state of breast cancer treatment and prevention research through our study. Additionally, there is a relative lack of preventive measures and strategies for mitigating postoperative and post-medication adverse reactions compared to breast cancer treatment, and more efforts are needed in these areas. Furthermore, our research can assist clinicians in making comprehensive decisions. For instance, when selecting a treatment approach for patients, the knowledge graph facilitates linking to available medications, associated adverse reactions, and measures to mitigate or prevent adverse effects.

Our research still has several limitations. Firstly, SemRep, as a natural language processing program based on UMLS, still exhibits shortcomings. Despite the extensive coverage and scale of the UMLS Metathesaurus, it has a relatively limited ability to recognize entities. There are still areas for improvement in processing natural language texts [20]. Secondly, clinical researchers often prefer causal relationships rather than pure correlations, however, our study can only reveal the connections between pieces of information and cannot determine the magnitude and direction of their effects. Thirdly, with the release of new literature, the knowledge graph also needs to be updated promptly, increasing the burden on researchers. Future improvements should focus on automating the mining of literature data to ensure timely updates to the knowledge graph for breast cancer prevention and treatment, thereby alleviating the burden on researchers.

Conclusion

This study successfully constructed a knowledge graph for breast cancer prevention and treatment by integrating relevant literature from the past five years and conducting knowledge discovery. Through this knowledge graph, researchers can learn about breast cancer treatment methods, medications, and adverse reactions to preventive treatments and gain insights into the relationships between different pieces of knowledge.

Acknowledgements

We thank Feng Xixi for her suggestions at the initial stage of the study.

Conflicts of Interest

None declared.

Abbreviations

SPO: Subject-Predicate-Object
TNBC: Triple Negative Breast Neoplasms
UMLS: Unified Medical Language System
SemmedDB: Semantic MEDLINE Database

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018 Nov;68(6):394-424.
2. Xiao Y, Xia J, Li L, Ke Y, Cheng J, Xie Y, et al. Associations between dietary patterns and the risk of breast cancer: a systematic review and meta-analysis of observational studies. Breast Cancer Res. 2019 Jan 29;21(1):16.
3. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May;71(3):209-249.
4. Thakur P, Seam RK, Gupta MK, Gupta M, Sharma M, Fotedar V. Breast cancer risk factor evaluation in a Western Himalayan state: A case-control study and comparison with the Western World. South Asian J Cancer. 2017 Jul-Sep;6(3):106-109.
5. Badr LK, Bourdeanu L, Alatrash M, Bekarian G. Breast Cancer Risk Factors: a Cross- Cultural Comparison between the West and the East. Asian Pac J Cancer Prev. 2018 Aug 24;19(8):2109-2116.
6. Zhang X, Dong XP, Guan YZ, Ren Me, Guo DL, He YT, et al. Research Progress on Epidemiological Trend and Risk Factors of Female Breast Cancer[J]. Cancer Research on Prevention and Treatment, 2021, 48(01): 87-92.
7. Tan MM, Ho WK, Yoon SY, Mariapun S, Hasan SN, Lee DS, et al. A case-control study of breast cancer risk factors in 7,663 women in Malaysia. PLoS One. 2018 Sep 14;13(9):e0203469.
8. Britt KL, Cuzick J, Phillips KA. Key steps for effective breast cancer prevention. Nat Rev Cancer. 2020 Aug;20(8):417-436.
9. Fiszman M, Rindflesch T C, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. Proceedings of the computational lexical semantics workshop at HLT-NAACL 2004. 2004: 76-83.
10. SOUHU. Advances in adjuvant therapy for triple-negative breast cancer, just look at these 8 key clinical studies. URL: https://www.sohu.com/a/508222106_121118854[accessed 2023-06-25].
11. Feng B, Gao J. AnthraxKP: a knowledge graph-based, Anthrax Knowledge Portal mined from biomedical literature. Database (Oxford). 2022 Jun 2;2022:baac037.
12. Feng F, Tang F, Gao Y, Zhu D, Li T, Yang S, et al. GenomicKB: a knowledge graph for the human genome. Nucleic Acids Res. 2023 Jan 6;51(D1):D950-D956.
13. James T, Hennig H. Knowledge Graphs and Their Applications in Drug Discovery. Methods Mol Biol. 2024;2716:203-221.
14. Lyu K, Tian Y, Shang Y, Zhou T, Yang Z, Liu Q, et al. Causal knowledge graph construction and evaluation for clinical decision support of diabetic nephropathy. J Biomed Inform. 2023 Mar;139:104298.
15. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017 Jan 4;45(D1):D833-D839.

16. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018 Jan 4;46(D1):D1074-D1082.
17. National Library of Medicine. Clinvar. URL: https://www.ncbi.nlm.nih.gov/clinvar/ [accessed 2023-11-18].
18. Li X, Sun S, Tang T, Lu J, Zhang L, Yin J, et al. Construction of a knowledge graph for breast cancer diagnosis based on Chinese electronic medical records: development and usability study. BMC Med Inform Decis Mak. 2023 Oct 10;23(1):210.
19. An B. Construction and application of Chinese breast cancer knowledge graph based on multi-source heterogeneous data. Math Biosci Eng. 2023 Feb 6;20(4):6776-6799.
20. Li XY, Li JL, Li ZY. Integrated medical language system and its application in knowledge discovery. Digital Library Forum, 2019 (9): 24-29.
21. Kilicoglu H, Rosemblat G, Fiszman M, Shin D. Broad-coverage biomedical relation extraction with SemRep. BMC Bioinformatics. 2020 May 14;21(1):188.
22. National Library of Medicine. Access to SemRep/SemMedDB/SKR Resources. URL: https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR.html [accessed 2023-11-18].
23. Fiszman M, Rindflesch T C, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. Proceedings of the computational lexical semantics workshop at HLT-NAACL 2004. 2004: 76-83.
24. Wes McKinney. (2010). pandas: A Foundational Python Library for Data Analysis and Statistics. Python for High Performance and Scientific Computing, 293-296.
25. John D. Hunter. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3):90-95, 2007.
26. Wordcloud. WordCloud for Python documentation. URL: https://amueller.github.io/word_cloud/ [accessed 2023-12-25].
27. Hagberg A, Swart P, S Chult D. Exploring network structure, dynamics, and function using NetworkX[R]. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
28. SOUHU. The difference between breast cancer radiotherapy, targeted therapy and chemotherapy!. URL: https://www.sohu.com/a/280208482_790163 [accessed 2023-11-18].
29. Nagini S. Breast Cancer: Current Molecular Therapeutic Targets and New Players. Anticancer Agents Med Chem. 2017;17(2):152-163.
30. Cancer Research UK. Trastuzumab. URL: https://www.cancerresearchuk.org/about-cancer/treatment/drugs/trastuzumab [accessed 2023-11-18].
31. Bernas M, Thiadens SRJ, Smoot B, Armer JM, Stewart P, Granzow J. Lymphedema following cancer therapy: overview and options. Clin Exp Metastasis. 2018 Aug;35(5-6):547-551.
32. Kilicoglu H, Rosemblat G, Fiszman M, Shin D. Broad-coverage biomedical relation extraction with SemRep. BMC Bioinformatics. 2020 May 14;21(1):188.

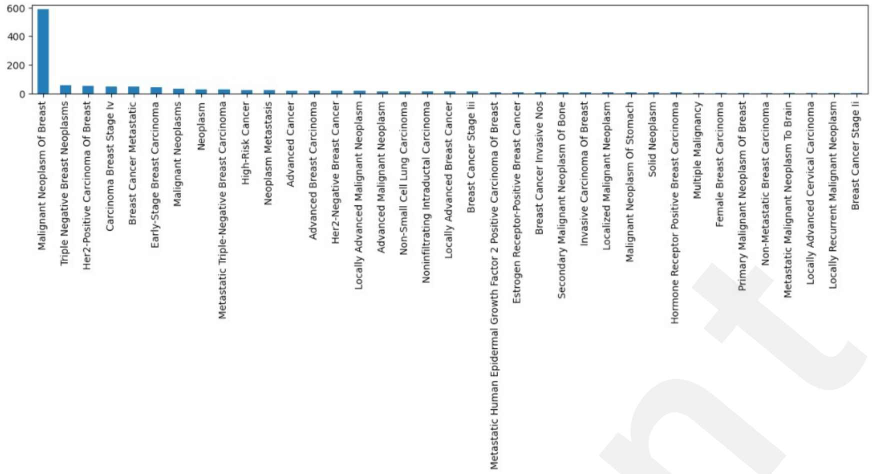
Supplementary Files

Supplementary Table 1. Semantic Relationship and Semantic Schema of Breast Cancer.

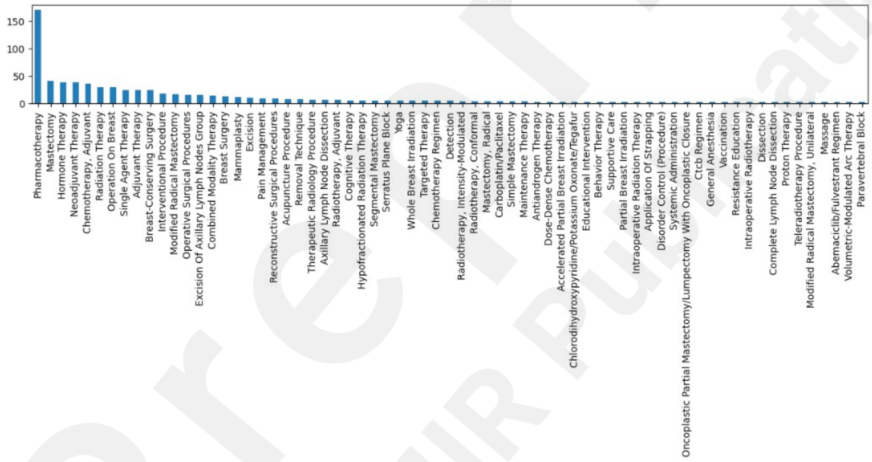
Type	Predicate	Semantic Schema Example	Frequency	SPO Example
Treatment				
	TREATS	topp - TREATS - neop	1389	Radiation therapy TREATS Malignant neoplasm of breast
		topp - TREATS - podg	1128	Radiation therapy TREATS Patients
		phsu - TREATS - neop	908	Pegfilgrastim TREATS Malignant neoplasm of breast
		phsu - TREATS - podg	896	Acetaminophen TREATS Patients
		orch - TREATS - neop	345	Tamoxifen TREATS Malignant neoplasm of breast
		topp - TREATS - popg	298	Chemotherapy, Adjuvant TREATS Woman
		phsu - TREATS - popg	209	Paclitaxel TREATS Population Group
	ADMINISTERED_TO	topp - ADMINISTERED_TO - humn	614	Pharmacotherapy ADMINISTERED_TO Patients
		orch - ADMINISTERED_TO - humn	323	Epirubicin ADMINISTERED_TO Patients
	USES	topp - USES - phsu	479	Adjuvant therapy USES Aromatase Inhibitors
		topp - USES - aapp	155	Therapeutic procedure USES Bevacizumab
Prevention	PREVENTS	phsu-PREVENTS-sosy 41	41	Vitamin B 12 PREVENTS Pain
		topp-PREVENTS-dsyn 36	36	Administration, Topical PREVENTS Early radiation dermatitis
		topp-PREVENTS-sosy 44	44	Music Therapy PREVENTS Fatigue
Influencing/Related factors				
	ASSOCIATED_WITH	aapp - ASSOCIATED_WITH - neop	237	Trastuzumab ASSOCIATED_WITH Malignant neoplasm of breast
		bact - ASSOCIATED_WITH - dsyn	1	Akkermansia muciniphila ASSOCIATED_WITH Obesity
	AFFECTS	topp - AFFECTS - ortf	9	Exercise, Aerobic AFFECTS Cardiac function

		comd - AFFECTS - neop	6	Somatic mutation AFFECTS Malignant neoplasm of urinary bladder
	CAUSES	neop - CAUSES - neop	3	Melanoma CAUSES Malignant neoplasm of lung
		topp - CAUSES - dsyn	39	Pharmacotherapy CAUSES Onycholysis
Related diseases				
	COEXISTS_WITH	fndg-COEXISTS_WITH-dsyn	23	Sedentary lifestyle COEXISTS_WITH Disease
	PRECEDES	dsyn - PRECEDES - patf	1	Skin toxicity PRECEDES Mucositis
		dsyn - PRECEDES - dsyn	1	Hypertensive disease PRECEDES Neutropenia

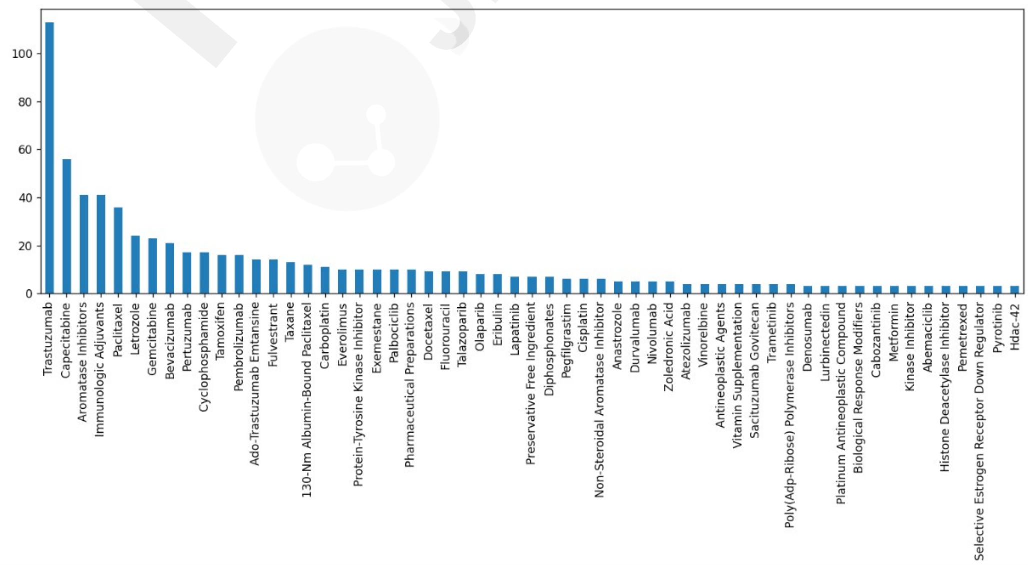
Abbreviations: aapp-Amino Acid, Peptide, or Protein; bact-Bacterium; bpoc-Body Part, Organ, or Organ Component; comd-Cell or Molecular Dysfunction; dsyn- Disease or Syndrome; humn-Human; neop-Neoplastic Process, orch-Organic Chemical; patf-Pathologic Function; phsu-Pharmacologic Substance; podg- Patient or Disabled Group; popg- Population Group; topp-Therapeutic or Preventive Procedur.



Supplementary Figure 1. Different subtypes and stages of breast cancer.



Supplementary Figure 2. Treatments of breast cancer.



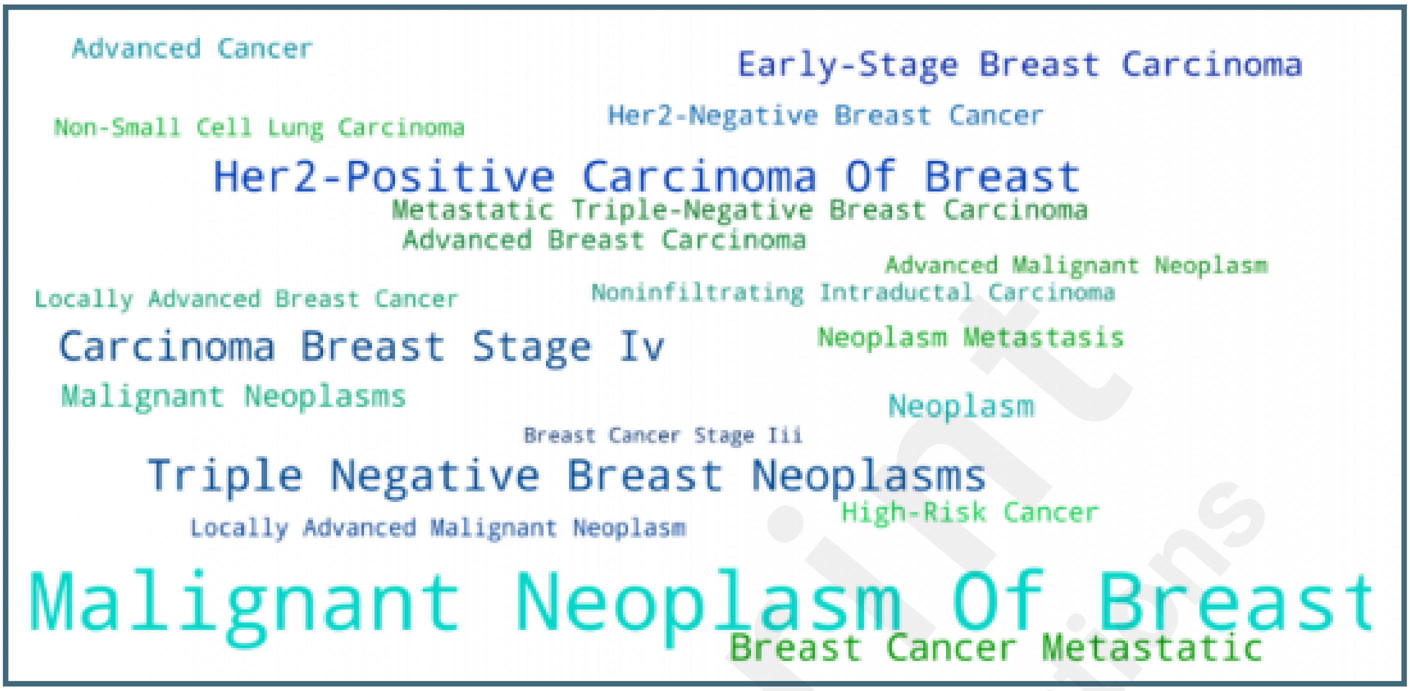
Supplementary Figure 3. Drugs of breast cancer.

Supplementary Files

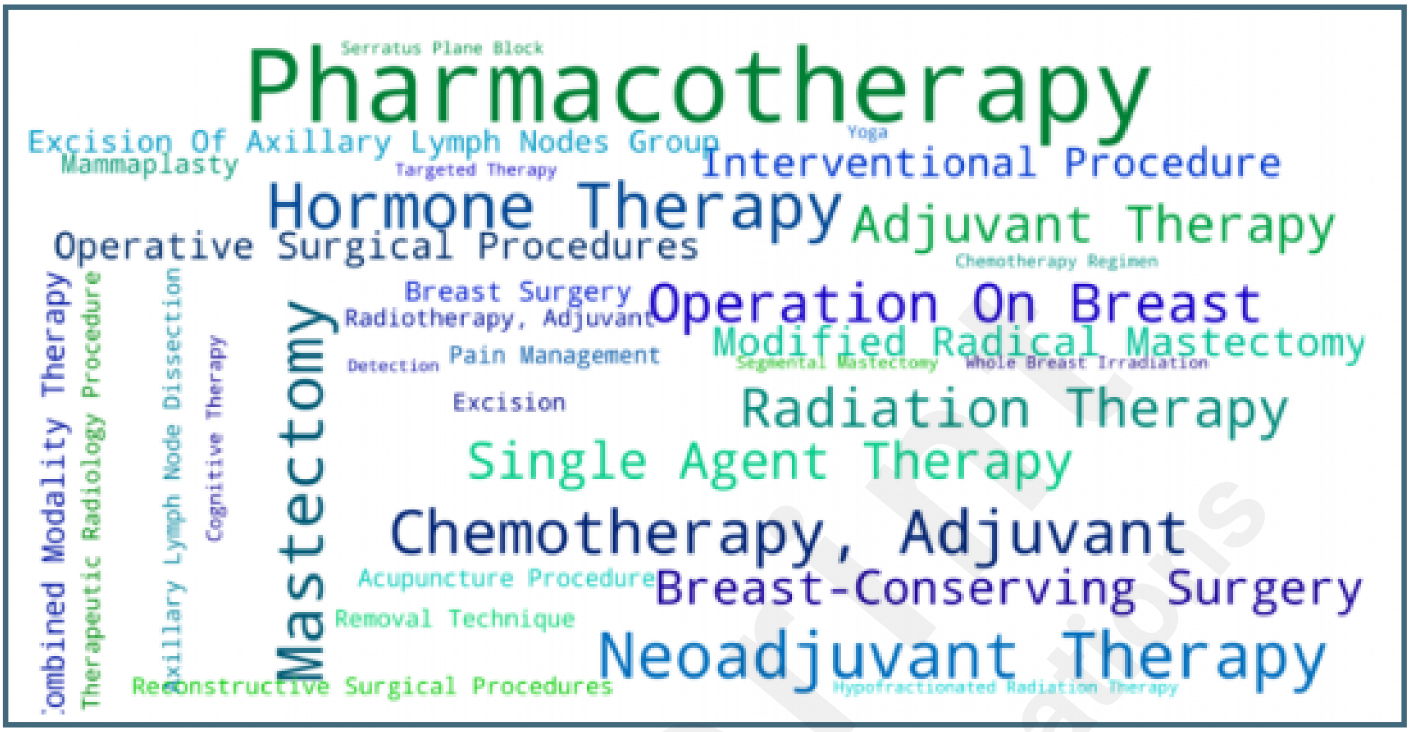
Tracked manuscript.
URL: http://asset.jmir.pub/assets/b81743817d7945c98e33d39197478878.docx

Figures

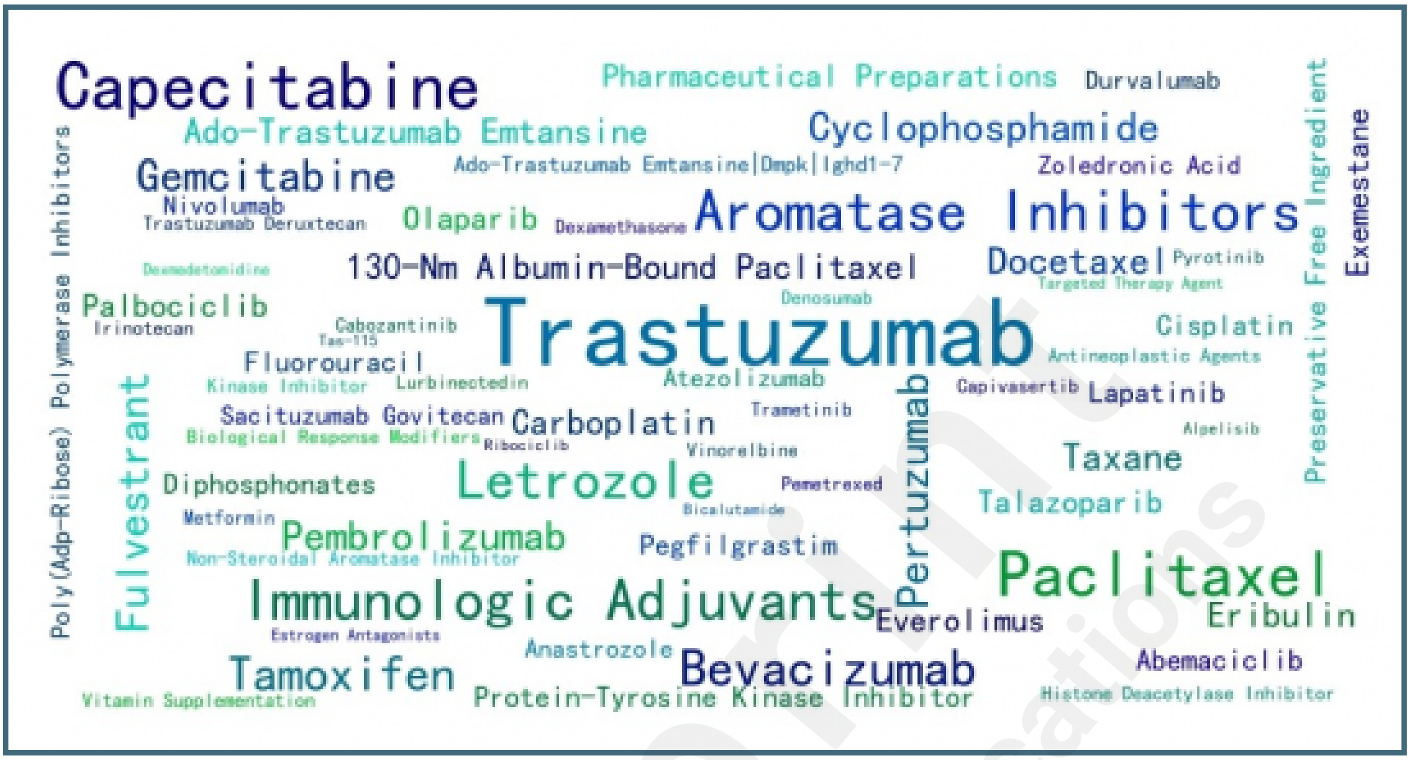
Breast cancer subtypes and stages.



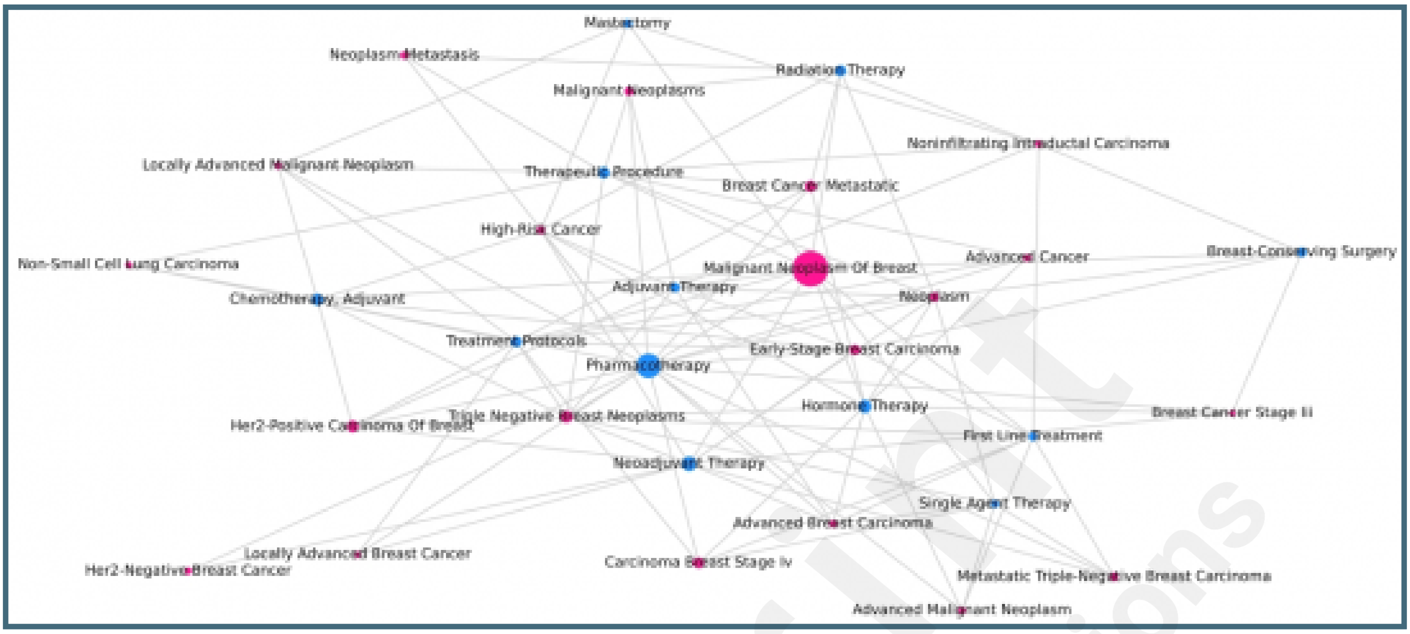
Treatment of breast cancer.



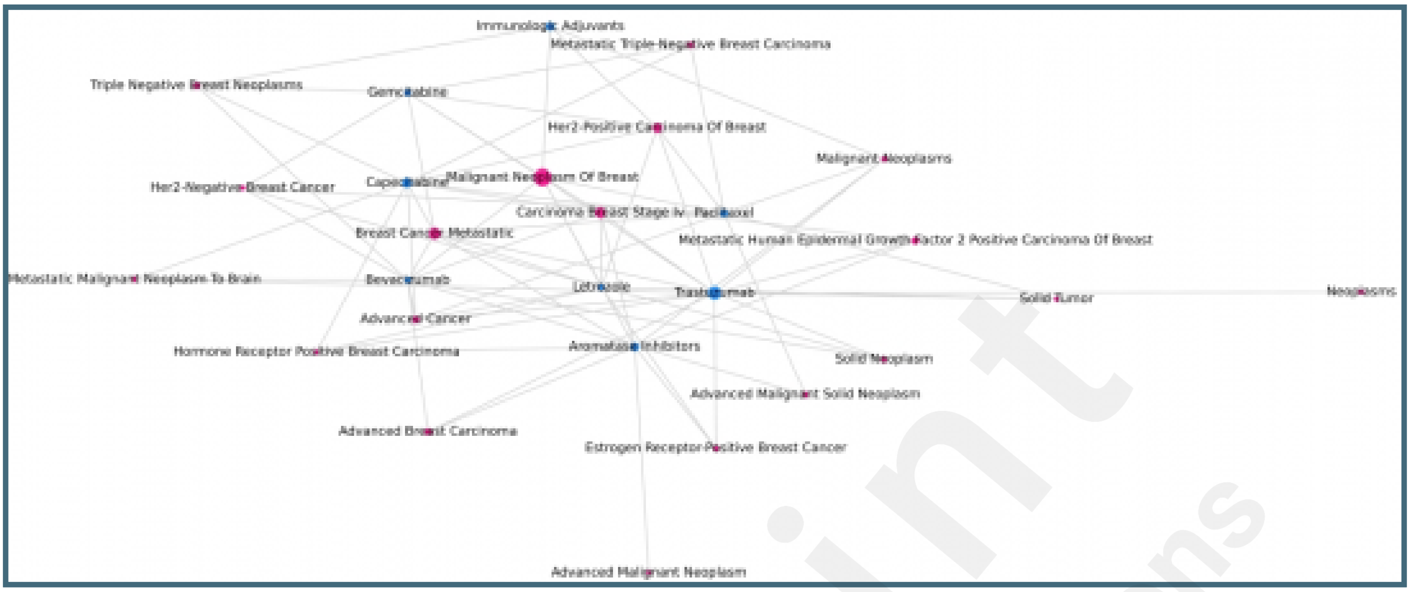
Drugs for breast cancer.



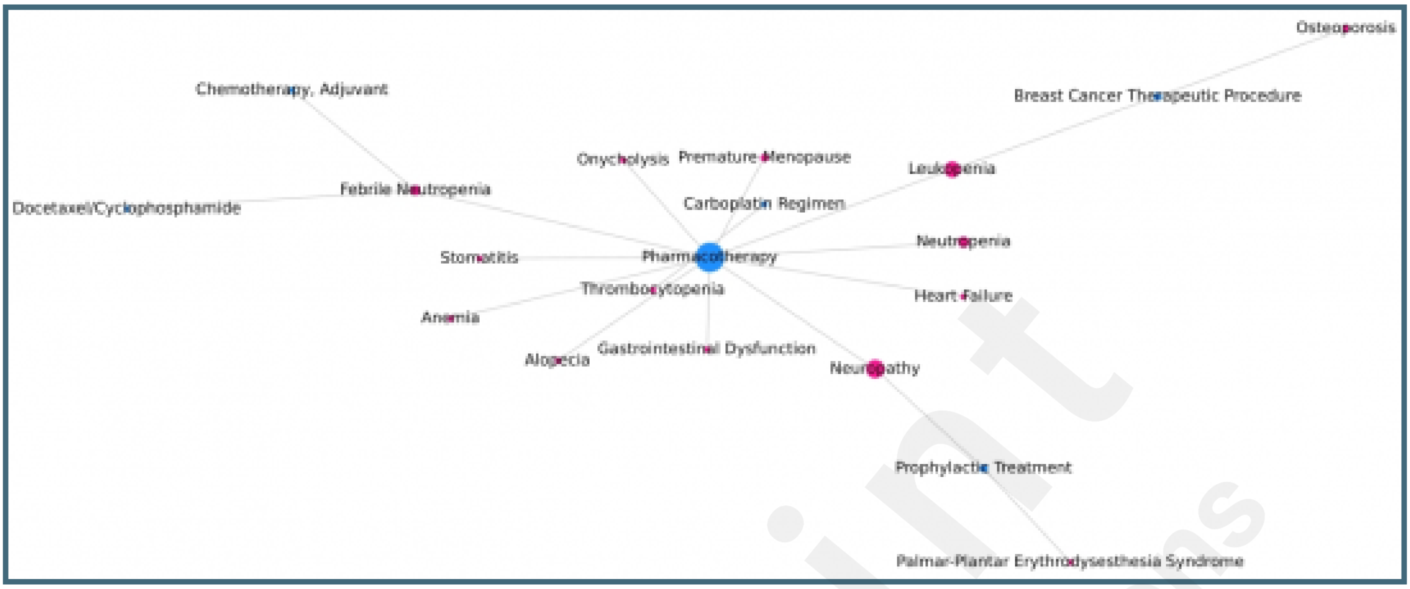
Relationship between different subtypes and stages of breast cancer and treatment methods.



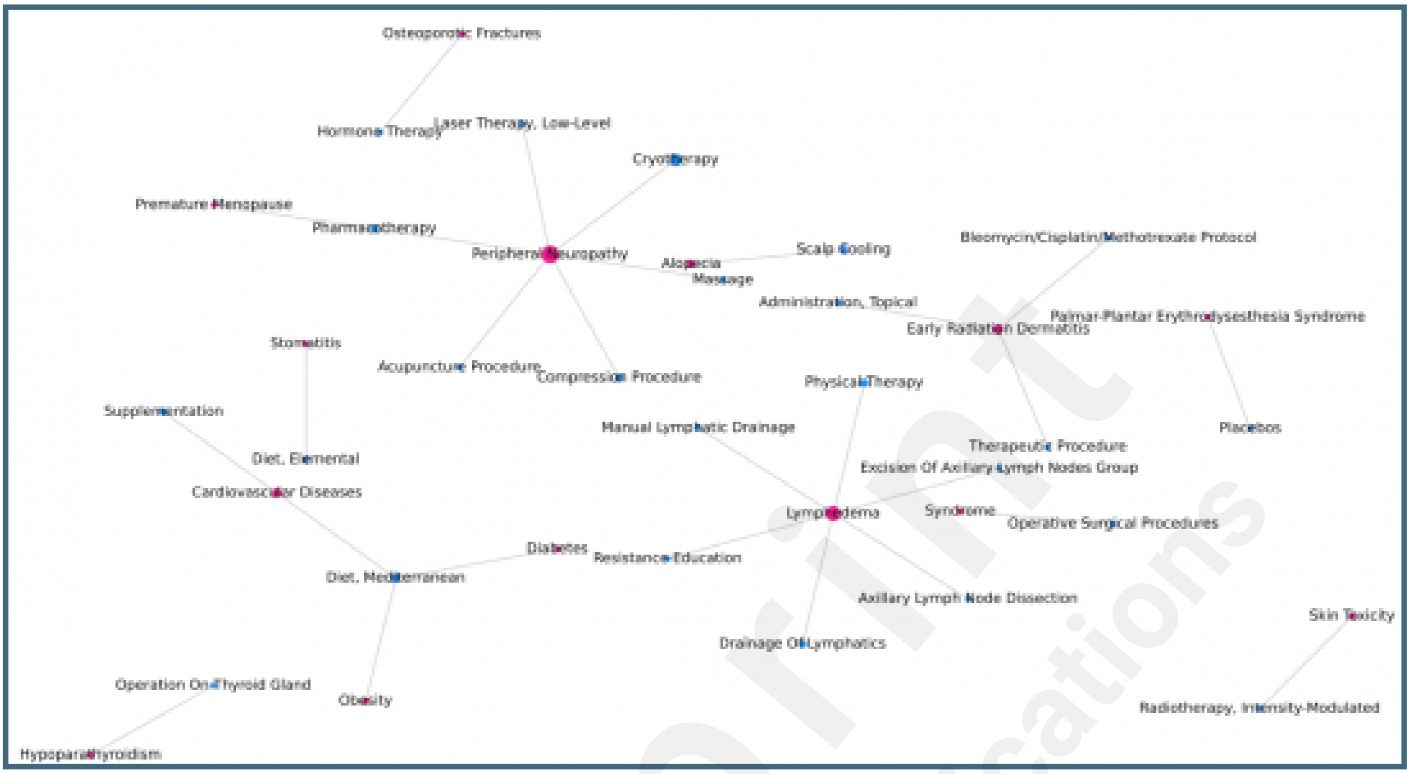
Relationship between different subtypes and stages of breast cancer and drugs.



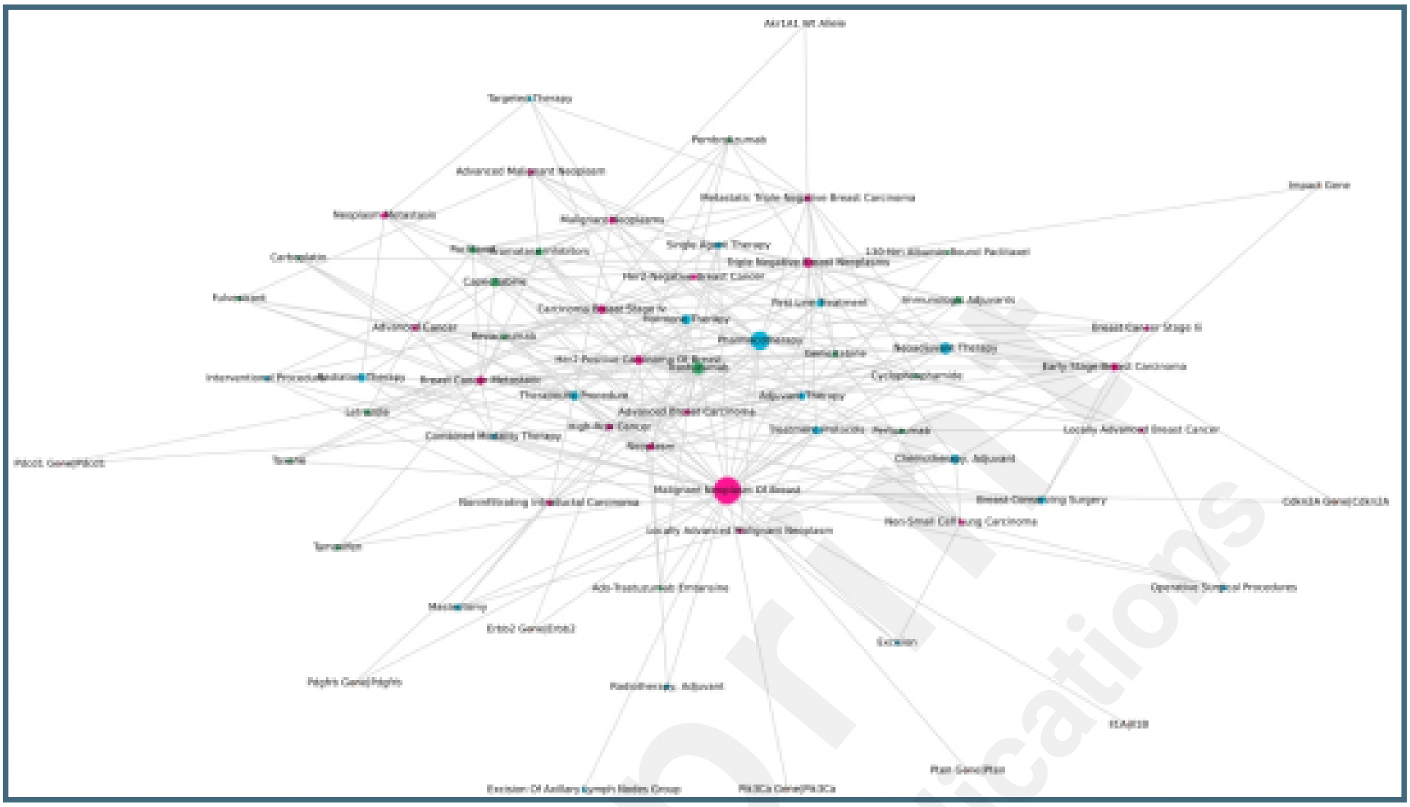
Relationship between breast cancer treatment and adverse reactions.



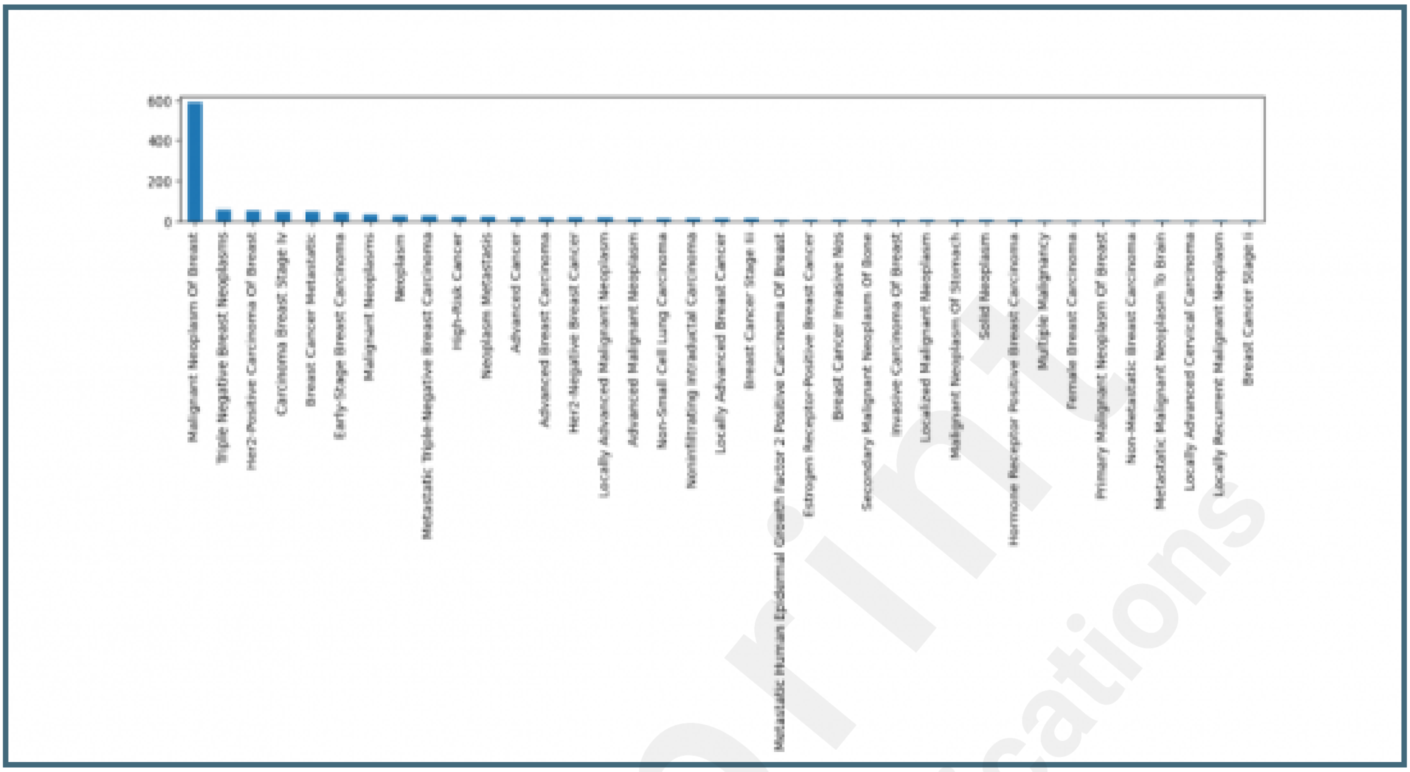
Relationship between adverse reactions after breast cancer treatment and preventive measures.



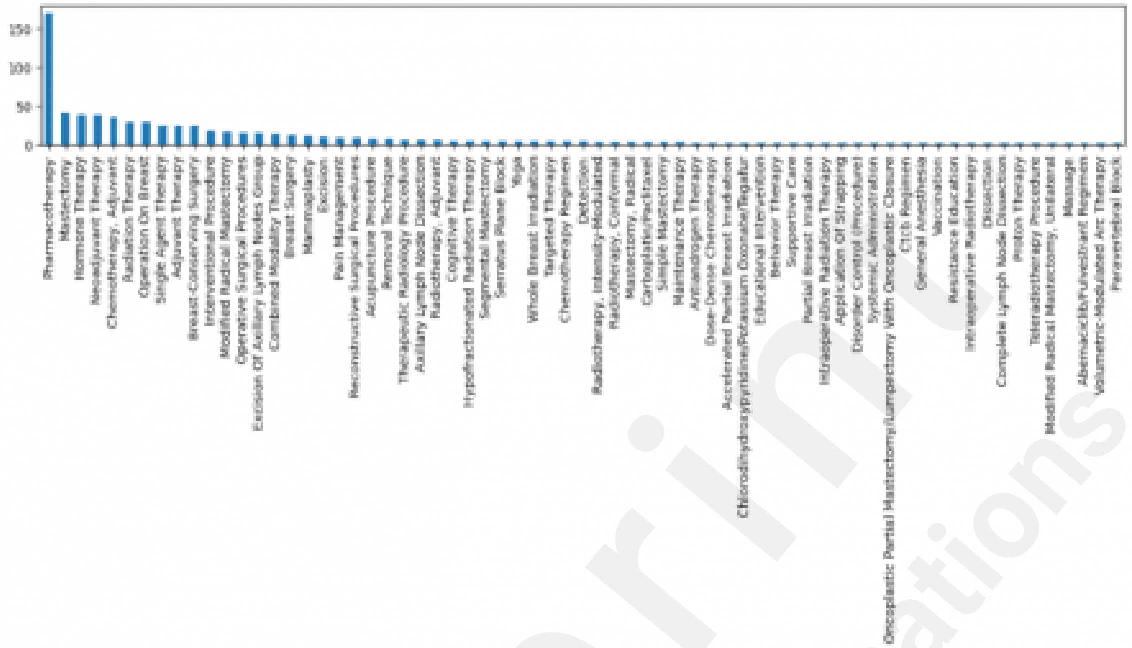
Relationship between types of breast cancer, treatments, drugs, and genes.



Different subtypes and stages of breast cancer.



Treatments of breast cancer.



Drugs of breast cancer.