EPiC Series in Technology

Volume 3, 2026, Pages 27–35

Proceedings of AI for Accelerated Research Symposium

Identifying High-Risk Cancer Patients on Breast Cancer Pathology Reports with Large Language Models

Trevor Kwan, MSc[1]; Jaimie J. Lee, BSc[2,3]; Raymond T. Ng, PhD[1,4]
[1] Data Science Institute, University of British Columbia (UBC);
[2] Department of Radiation Oncology, BC Cancer; [3] Department of Surgery, UBC;
[4] Department of Compute Science, UBC
trevor.kwan@ubc.ca, jaimie.lee@ubc.ca, rng@cs.ubc.ca

Abstract

Breast cancer subtypes defined by estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status guide treatment decisions, yet manual extraction of these biomarkers from pathology reports is time-consuming and error-prone. We present an end-to-end NLP pipeline that automates high-risk subtype identification (HER2-positive and triple-negative) from digital core-biopsy reports. A corpus of 2,722 reports (2,401 non-synoptic, 321 synoptic) was annotated in Doccano, yielding 16,706 question–answer pairs. Reports were pre-processed and then split using a multi-stratified sampling approach into training (59%), validation (17%), and held-out test (24%) sets. We fine-tuned BioMedBERT on SQuAD 2.0 and then on our domain-specific dataset, employing hyperparameter optimization and prediction post-processing. On the held-out test data, our model achieved 99.79% accuracy on synoptic reports and 98.83% on non-synoptic reports, outperforming human annotators and maintaining robust performance across report formats and biomarker classes. By automatically flagging eligible patients for neoadjuvant chemotherapy triage, this pipeline has the potential to streamline clinical workflows, reduce treatment delays, and improve outcomes for high-risk breast cancer patients.

1 Introduction

Breast cancer is a heterogeneous disease encompassing molecular subtypes that vary in prognosis and treatment response. These subtypes are defined by the expression of biomarkers, including estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). Among them, HER2 positive (HER2+) and triple negative breast cancer (TNBC) are considered high-

risk due to their aggressive growth patterns and increased likelihood of early recurrence and metastasis. HER2+ tumours are driven by overexpression of the HER2 protein, while TNBC lacks expression of ER, PR, and HER2. These high-risk subtypes benefit from expedited triage and initiation of treatment, making their early and accurate identification a critical step in optimizing breast cancer care.

Clinical guidelines recommend that high-risk HER2+ and TNBC patients receive chemotherapy prior to surgery, known as neoadjuvant therapy (NACT). However, approximately one third of eligible stage 2 and 3 high-risk patients receive surgery before chemotherapy, deviating from the ideal treatment sequence. As a result, they miss the chance to benefit from NACT, which can reduce the need for more invasive procedures such as total mastectomy and axillary lymph node dissection. Timely referral to a medical oncologist is therefore essential, as they are responsible for evaluating NACT eligibility and initiating treatment. While some delays are unavoidable such as those resulting from second opinion requests or patient readiness for chemotherapy, many stem from systemic inefficiencies. In particular, suboptimal handoffs between primary care physicians, radiologists, pathologists, oncologists, and surgeons contribute to delays that could be avoided through better coordination.

From the patient perspective, earlier identification of high-risk subtypes supports timely initiation of NACT, increasing the likelihood of breast-conserving surgery and reducing the need for more invasive procedures such as total mastectomy or axillary lymph node dissection. By minimizing delays, the model has the potential to improve outcomes while reducing the physical and emotional burden associated with late or missed intervention.

Natural Language Processing (NLP) has the potential to improve high-risk breast cancer triage by automating the extraction of clinically relevant information from medical text. In particular, transformer-based architectures have been dominating the field, achieving leading performance across numerous NLP benchmarks. Among these, BERT (Bidirectional Encoder Representations from Transformers) has been widely adopted due to its strong performance and well-supported tooling infrastructure. In addition, its relatively compact size, especially when compared to larger autoregressive models like GPT, allows it to be hosted and fine-tuned locally, an important advantage when working with sensitive clinical data under strict privacy constraints.

In current clinical workflows, biopsy reports are the earliest available source of tumour biomarker status. Building on the recent progress in NLP and LLMs, this study aims to develop and evaluate an NLP pipeline for identifying patients with high-risk breast cancer from their breast biopsy reports. By automatically flagging eligible cases for expedited referral to medical oncology, the model has the potential to reduce delays in initiating NACT and optimize clinical outcomes.

2 Methods

2.1 Overview

We developed an NLP pipeline to automate the extraction of biomarker information from core-biopsy cancer reports to enable the triage of high-risk patients. The pipeline comprises three stages: (1) report pre-processing, (2) a question-answering (QA) model for biomarker identification, (3) post-processing to structure and consolidate the outputs. Figure 1 provides a high-level summary of the study design, data sources, and overall NLP workflow.

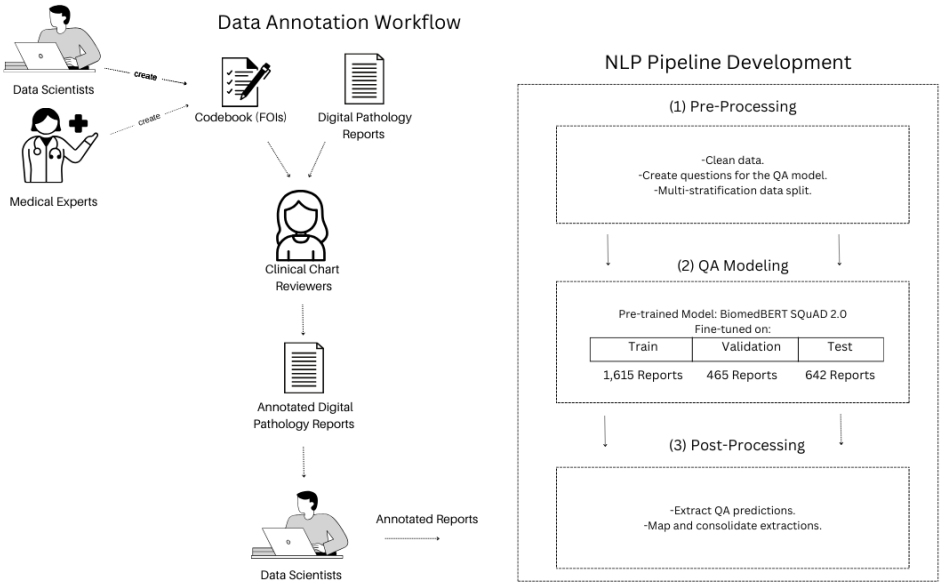

Figure 1: End-to-end workflow for building and evaluating our BERT-based NLP pipeline: (left) expert-driven annotation of biopsy reports; (right) three stages of NLP development: pre-processing, BioMedBERT QA modeling on splits, and post-processing consolidation

2.2 Data Sources

The study population included patients diagnosed with breast cancer between January 1, 2020, and June 1, 2024, in British Columbia (BC), Canada. From these patients, a total of 2,722 breast core needle biopsy reports were sourced from the BC Cancer breast cancer outcomes unit database. Among the reports, 2,401 were non-synoptic and 321 were synoptic. Synoptic reports employ a standardized, highly structured format, whereas non-synoptic reports lack such a structure. Figure 2 provides a visual comparison of these two report types. Institutional ethics approval was acquired prior to study initiation.

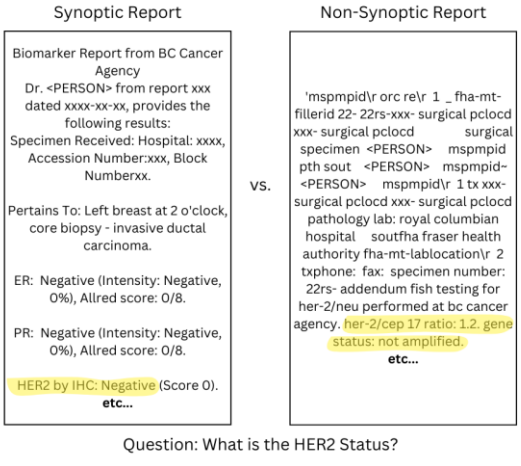

Figure 2: Comparison of synoptic versus non-synoptic reports. The left panel shows a structured, field-based synoptic report with clearly labeled biomarker results (e.g. "HER2 by IHC: Negative"), while the

right panel illustrates an unstructured, non-synoptic report containing the same HER2 information in free-text form. Both panels illustrate the QA prompt: "What is the HER2 Status?"

2.3 Primary Outcome

The primary outcome of this study was the performance of NLP tools in extracting ER, PR, and HER2 status from breast biopsy reports. However, the NLP tools need to deal with the complication that all three biomarkers may not be reported in a single report. Furthermore, in clinical practice, pathology reports are sometimes amended with addenda or supplemented by follow-up testing, which can update biomarker status. Thus, the identification of triple negative breast cancer patients consists of two phases: (i) per-report level extraction of the biomarkers' status; and (ii) aggregation of report-level extraction to give a patient-level status. Because the latter phase is data engineering activities, this study only focuses on per-report level accuracy, if ER, PR and HER2 status are mentioned at all.

2.4 Dataset Curation

The study dataset was created through manual annotation of breast biopsy reports using the Doccano tool on a Docker platform. Biopsy sites included the breast, chest wall, skin, muscle, soft tissue invading the ribs, and skin over the sternum. An annotation scheme was developed in collaboration with breast cancer oncologists, surgeons, and data scientists. Trained annotators labelled spans of text in each report to capture ER, PR, and HER2 status. For ER and PR, annotations included the reported status (positive or negative) as well as scoring metrics when explicitly mentioned, including Allred scores (0-8) and staining intensity scores (0-3). For HER2, annotations captured the reported status (positive, negative, or equivocal) from immunohistochemistry (IHC), with 3+ labelled as positive, 0-1+ as negative, and 2+ as equivocal. In cases where HER2 IHC results were equivocal, follow-up fluorescence in situ hybridization (FISH) results were annotated, with "amplified" labelled as positive and "not amplified" as negative. These annotations served as ground truth for model development and evaluation.

2.5 NLP Pipeline Development

2.5.1. Pre-processing and Data Cleaning

All annotated reports were first imported and cleaned to standardize labeling conventions and correct typo errors. To fine-tune our BERT-based QA model, we created both "mention" and "no mention" examples. Annotators already labeled all spans containing biomarker information, so for any report without a given biomarker label we generated a "no mention" case so the model could learn to recognize absent cases. We formulated seven binary questions, one per biomarker class, and posed each question to every report (e.g. "Is the ER status positive?") during training (see Table 1).

Biomarker	**Question**
ER Status	Is the ER Status positive?
	Is the ER Status negative?
PR Status	Is the PR Status positive?
	Is the PR Status negative?
HER2 Status	Is the HER2 Status positive?

	Is the HER2 Status negative?
	Is the HER2 Status equivocal?

Table 1: Binary question prompts used for each biomarker class during QA model training

2.5.2. Use of Train, Validation, and Test Datasets

The dataset was split into training (59%, n = 1,615), validation (17%, n = 465), and test (24%, n = 642) sets (see Table 2 for the synoptic vs. non-synoptic breakdown). The training and validation splits were used for model fine-tuning. Because the validation set was consulted multiple times during fine-tuning iterations, it no longer served as an unbiased performance estimator. Thus, a held-out test set was reserved for the final, unbiased evaluation.

Split	Non-Synoptic % (# of Reports)	Synoptic % (# of Reports)	Combined % (# of Reports)
Train	53% (1453)	6% (162)	59% (1,615)
Validation	17% (465)	0% (0)	17% (465)
Test	18% (483)	6% (159)	24% (642)
Total	88% (2401)	12% (321)	100% (2,722)

Table 2: Distribution of core-biopsy reports across training, validation, and test splits, with report counts and percentages shown separately for non-synoptic and synoptic formats

2.5.3. Model training and Architecture

We began with BiomedBERT, a BERT version pretrained on PubMed abstracts and full-text articles from PubMedCentral, and first fine-tuned it on the SQuAD 2.0 dataset (~150K question-answer pairs drawn from Wikipedia) to adapt this model for our question-answering task.

To specialize this model for our clinical QA task, we further trained this model on our own dataset of 16,706 question-answer pairs. This domain-specific fine-tuning significantly improves performance in clinical text tasks compared to generic, out-of-the-box language models.

2.5.4. Hyperparameter Optimization

We applied a sliding-window mechanism to accommodate BERT's 512-token input limit, segmenting each report into overlapping chunks. To guarantee that no annotated span straddled two windows, we first analyzed the distribution of answer lengths observing that the vast majority of extracted spans fall below 10 tokens, with a long tail of longer responses. Thus, we adjusted the stride accordingly. Hyperparameters such as learning rate, batch size, number of training epochs, and weight decay were tuned using manual search based on validation performance as well. The final selected values were learning rate = 2e-5, per-device batch size = 16, number of epochs = 5, and weight decay = 0.01.

During inference, we converted the model's raw confidence score into a binary decision: scores above the chosen threshold were classified as "mention", whereas scores at or below the threshold were treated as "no mention". The threshold itself was selected based on maximized validation performance.

2.5.5. Post-processing and Consolidating Predictions

Each report was queried with ten independent, binary questions – one per biomarker category. To handle cases where a single report contained multiple extracted mentions for the same biomarker, we applied a post-processing step that consolidated multiple mentions into a single classification. Figures 3 and 4 illustrate this aggregation logic.

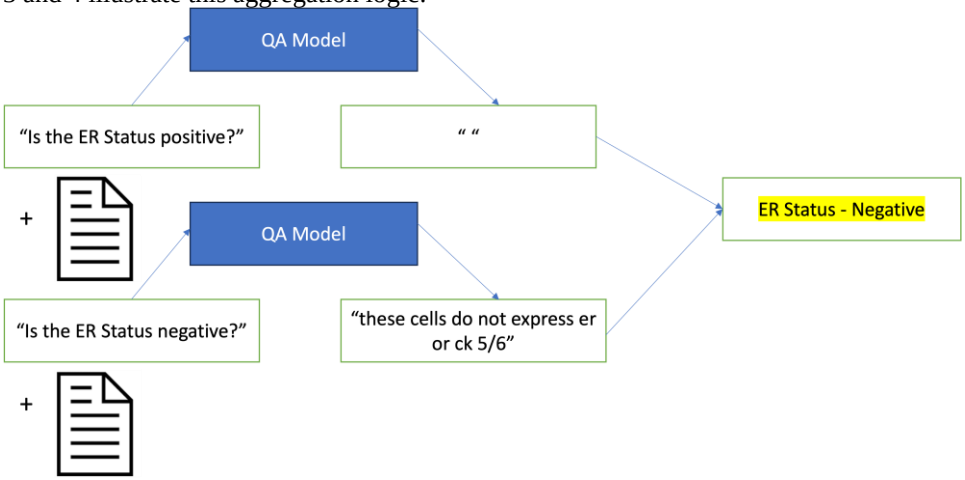

Figure 3: Illustration example of post-processing when only one class is mentioned for the biomarker ER Status. The report is queried with "Is the ER Status positive?" and "Is the ER Status negative?", extracts a single mention span, and results in one consolidated classification label: ER Status – Negative

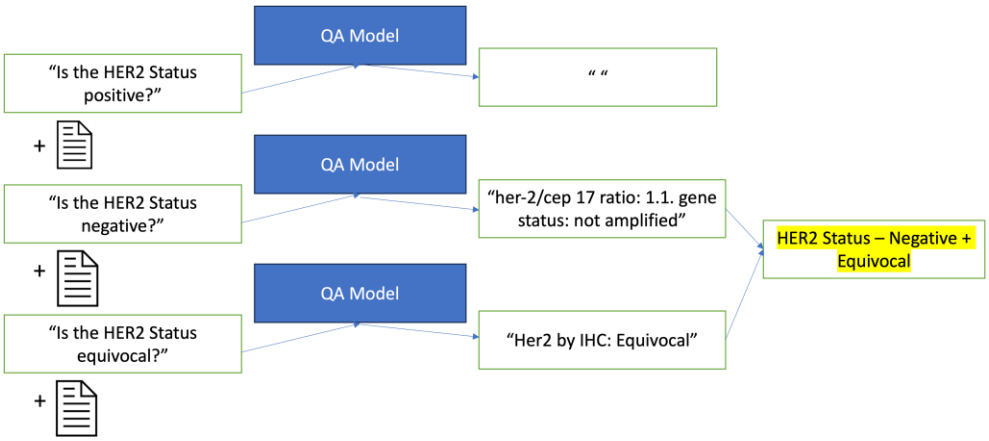

Figure 4: Illustration example of post-processing when multiple classes are mentioned for the biomarker HER2 Status. The report is queried with "Is the HER2 Status positive?", "Is the HER2 Status negative?", and "Is the HER2 Status equivocal?", extracts extracts multiple mention spans that are then merged into a final label: HER2 Status - Negative + Equivocal

3 Results: Model Performance

We evaluated model performance for each biomarker using the following metrics: (i) Yes Mention Accuracy: Percentage of correctly identified "mention" cases among all true mentions.; (ii) No Mention

Accuracy: Percentage of correctly identified "no mention" cases among all true non-mentions; and (iii) Per-Class Accuracy: Percentage of correct classifications within each individual class or combination of classes.

Our QA pipeline surpassed human performance across all biomarker categories, achieving an overall accuracy of 99.79% on the synoptic test set and 98.83% on the non-synoptic test set. As anticipated, the highest accuracy was observed on structured (synoptic) reports, while the model also demonstrated robust performance on unstructured (non-synoptic) reports. Detailed results are presented in Table 3 (non-synoptic) and Table 4 (synoptic).

Biomarker	**Is there a Mention? (Accuracy, Correct/Total)**	**What is the Mention? (Accuracy, Correct/Total)**
ER Status	Yes Mention – 97.90%, 280/286	Positive – 98.64%, 218/221
		Negative – 95.38%, 62/65
	No Mention – 99.49%, 196/197	n/a
PR Status	Yes Mention – 98.64%, 217/220	Positive – 98.04%, 150/153
		Negative – 100.00%, 67/67
	No Mention – 100.00%, 263/263	n/a
HER2 Status	Yes Mention – 97.71%, 214/219	Positive – 100.00%, 18/18
		Negative – 97.87%, 92/94
		Equivocal – 100.00%, 62/62
		Positive + Equivocal – 83.33%, 5/6
		Negative + Equivocal – 94.87%, 37/39
	No Mention – 99.24%, 262/264	n/a

Table 3: QA model performance on the non-synoptic test set (n=483), showing per-biomarker accuracy for both "mention" and "no mention" detection for ER, PR, and HER2 statuses

The performance on the synoptic test set (n = 159), as expected, is even better than that on the non-synoptic reports. For mentions, the overall accuracy is 158/159, and for positive/negative extractions the accuracy is 100% across all three biomarkers. We suppress the details to save space.

4 Discussion and Conclusion

We used NLP empowered by LLMs to extract breast cancer biomarker status and identify high-risk subtypes from breast biopsy reports. Such efforts can expedite the triage of patients who should be prioritized for referral to medical oncology and initiation of NACT.

This study builds on prior work using NLP to extract breast cancer biomarker status from pathology reports by addressing key limitations of earlier methods. Earlier approaches primarily used traditional machine learning methods. While they achieved good performance, they relied heavily on manual feature engineering and often tailored to specific report formats, limiting their generalizability across institutions. In contrast, our BERT-based approach leverages deep contextual representations to more flexibly interpret varying report structures. Additionally, rather than extracting ER, PR, and HER2 in isolation, we integrated all three biomarkers to flag high-risk subtypes most likely to benefit from NACT, making the study more directly useful for guiding treatment decisions. Finally, by training and validating on both synoptic and non-synoptic reports, we demonstrate robust performance across reporting formats, increasing the model's real-world applicability. In ongoing work, a validation involving multiple health authorities is underway.

Automating high-risk breast cancer triage has important clinical implications for improving system-level efficiency and reducing patient burden. Automatically extracting biomarker information from biopsy reports reduces the need for manual chart review, streamlines administrative processes, and may enable real-time flagging of high-risk cases within EHRs. This enhances coordination among primary care providers, oncologists, and surgeons, helping to accelerate referral and treatment planning.

Beyond breast cancer triaging, the methodology presented here can be generalized to other cancer types, particularly when timely initiation of chemotherapy to high-risk patients are essential to their survival. In our ongoing work, we are exploring how to adapt the proposed NLP tools to better management of colon cancer patients.

References

Alafari, F., Driss, M., & Cherif, A. (2025). Advances in natural language processing for healthcare: A comprehensive review of techniques, applications, and future directions. Computer Science Review, 56, 100725. https://doi.org/10.1016/j.cosrev.2025.100725

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), Advances in Neural Information Processing Systems (Vol. 33). Curran Associates, Inc.

Cavalcante, F. P., Zerwes, F. P., Alcantara, R., Millen, E. C., Mattar, A., Antonini, M., Lima, A. D. N., Bines, J., Brenelli, F. P., Novita, G. G., Berretini Junior, A., Szymanski Machado, R. H., DE Souza, A. B. A., Campelo, D. C., da Costa Vieira, R. A., & Frasson, A. L. (2025). Oncological outcomes of breast-conserving surgery versus mastectomy following neoadjuvant chemotherapy in a contemporary multicenter cohort. Scientific reports, 15(1), 9032. https://doi.org/10.1038/s41598-025-93491-7

Chavez Mac Gregor, M., Housten, A., Paredes, E., Malinowski, C., Harris, C., & Giordano, S. H. (2021). A qualitative study informing about barriers and facilitators associated to chemotherapy initiation among breast cancer patients: Next steps for an intervention. Journal of Clinical Oncology, 39(28_suppl), 247. https://doi.org/10.1200/JCO.2020.39.28_suppl.247

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.

Exman, P., & Tolaney, S. M. (2021). HER2-positive metastatic breast cancer: a comprehensive review. Clinical advances in hematology & oncology : H&O, 19(1), 40–50.

Guo, L., Kong, D., Liu, J., Zhan, L., Luo, L., Zheng, W., Zheng, Q., Chen, C., & Sun, S. (2023). Breast cancer heterogeneity and its implication in personalized precision therapy. Experimental hematology & oncology, 12(1), 3. https://doi.org/10.1186/s40164-022-00363-1

Holmes, B., Chitale, D., Loving, J., Tran, M., Subramanian, V., Berry, A., Rioth, M., Warrier, R., & Brown, T. (2021). Customizable Natural Language Processing Biomarker Extraction Tool. JCO clinical cancer informatics, 5, 833–841. https://doi.org/10.1200/CCI.21.00017

Hossain, E., Rana, R., Higgins, N., Soar, J., Barua, P. D., Pisani, A. R., & Turner, K. (2023). Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. Computers in biology and medicine, 155, 106649. https://doi.org/10.1016/j.compbiomed.2023.106649

Hughes, K. S., Zhou, J., Bao, Y., Singh, P., Wang, J., & Yin, K. (2020). Natural language processing to facilitate breast cancer research and management. The breast journal, 26(1), 92–99. https://doi.org/10.1111/tbj.13718

Klug, K., Beckh, K., Antweiler, D., Chakraborty, N., Baldini, G., Laue, K., Hosch, R., Nensa, F., Schuler, M., & Giesselbach, S. (2024). From admission to discharge: a systematic review of clinical natural language processing along the patient journey. BMC medical informatics and decision making, 24(1), 238. https://doi.org/10.1186/s12911-024-02641-w

Korde, L. A., Somerfield, M. R., Carey, L. A., Crews, J. R., Denduluri, N., Hwang, E. S., Khan, S. A., Loibl, S., Morris, E. A., Perez, A., Regan, M. M., Spears, P. A., Sudheendra, P. K., Symmans, W. F., Yung, R. L., Harvey, B. E., & Hershman, D. L. (2021). Neoadjuvant Chemotherapy, Endocrine Therapy, and Targeted Therapy for Breast Cancer: ASCO Guideline. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 39(13), 1485–1505. https://doi.org/10.1200/JCO.20.03399

Lüönd, F., Tiede, S., & Christofori, G. (2021). Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. British journal of cancer, 125(2), 164–175. https://doi.org/10.1038/s41416-021-01328-7

Montagna, G., Mrdutt, M. M., Sun, S. X., Hlavin, C., Diego, E. J., Wong, S. M., Barrio, A. V., van den Bruele, A. B., Cabioglu, N., Sevilimedu, V., Rosenberger, L. H., Hwang, E. S., Ingham, A., Papassotiropoulos, B., Nguyen-Sträuli, B. D., Kurzeder, C., Aybar, D. D., Vorburger, D., Matlac, D. M., Ostapenko, E., … Weber, W. P. (2024). Omission of Axillary Dissection Following Nodal Downstaging With Neoadjuvant Chemotherapy. JAMA oncology, 10(6), 793–798. https://doi.org/10.1001/jamaoncol.2024.0578

Naveed, H., Khan, A., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., & Mian, A.S. (2023). A Comprehensive Overview of Large Language Models. ACM Transactions on Intelligent Systems and Technology, 16, 1 - 72.

Pironet, A., Poirel, H. A., Tambuyzer, T., De Schutter, H., van Walle, L., Mattheijssens, J., Henau, K., Van Eycken, L., & Van Damme, N. (2021). Machine Learning-Based Extraction of Breast Cancer Receptor Status From Bilingual Free-Text Pathology Reports. Frontiers in digital health, 3, 692077. https://doi.org/10.3389/fdgth.2021.692077

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. Neural Information Processing Systems.

Yala, A., Barzilay, R., Salama, L., Griffin, M., Sollender, G., Bardia, A., Lehman, C., Buckley, J. M., Coopey, S. B., Polubriaginof, F., Garber, J. E., Smith, B. L., Gadd, M. A., Specht, M. C., Gudewicz, T. M., Guidi, A. J., Taghian, A., & Hughes, K. S. (2017). Using machine learning to parse breast pathology reports. Breast cancer research and treatment, 161(2), 203–211. https://doi.org/10.1007/s10549-016-4035-1.

Yin, L., Duan, J. J., Bian, X. W., & Yu, S. C. (2020). Triple-negative breast cancer molecular subtyping and treatment progress. Breast cancer research : BCR, 22(1), 61. https://doi.org/10.1186/s13058-020-01296-5.