Research Article

Kai-Po Chang, Yen-Wei Chu*, John Wang*

Analysis of hormone receptor status in primary and recurrent breast cancer via data mining pathology reports

https://doi.org/10.1515/med-2019-0013
received August 16, 2018; accepted December 5, 2018

Abstract: Background: Hormone receptors of breast cancer, such as estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (Her-2), are important prognostic factors for breast cancer. Objective: The current study aimed to develop a method to retrieve the statistics of hormone receptor expression status, documented in pathology reports, given their importance in research for primary and recurrent breast cancer, and quality management of pathology laboratories. Method: A two-stage text mining approach via regular expression-based word/phrase matching, was developed to retrieve the data. Results: The method achieved a sensitivity of 98.8%, 98.7% and 98.4% for extraction of ER, PR, and Her-2 results. The hormone expression status from 3679 primary and 44 recurrent breast cancer cases was successfully retrieved with the method. Statistical analysis of these data showed that the recurrent disease had a significantly lower positivity rate for ER (54.5% vs 76.5%, p=0.001278) than primary breast cancer and a higher positivity rate for Her-2 (48.8% vs 16.2%, p=9.79e-8). These results corroborated the previous literature. Conclusion: Text mining on pathology reports using the developed method may benefit research of primary and recurrent breast cancer.

Keywords: Breast cancer; Hormone receptor; Primary cancer; Recurrent cancer; Text mining.

1 Introduction

Electronic pathology reports, an important component of electronic health records [1], often document valuable data for research and quality control [2]. For breast cancer, the data documented in pathology reports is especially important since the expression statuses of hormone receptors, such as the estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (ErbB2 or Her-2), are immunohistochemically examined [3–5] and documented in pathology reports. Expression of these markers not only affects prognosis [6, 7] but also has implications on the choice of hormone therapy and chemotherapy [8, 9]. The importance of these markers in the treatment of breast cancer has been widely recognized and incorporated in the most recent (eighth) version of the American Joint Committee on Cancer staging system [10].

The statistics on ER, PR, and Her-2 immunostaining results are important quality indicators for pathology laboratories. Comparison of these statistics with the literature, may highlight possible problems in the quality of immunohistochemistry. For this reason, the accreditation process of the College of American Pathologists requires yearly ER, PR, and Her-2 immunostaining results [11]. Moreover, due to the possible value of the hormone receptor expression status on the prediction of local recurrence [12], statistics of hormone receptor expression status is also valuable for research on recurrent breast cancer.

To obtain hormone receptor expression statistics, some hospitals utilize a synaptic pathology report system [13–15] in which pathologists enter structured data. Since

***Corresponding author: Yen-Wei Chu,** Biotechnology Center, Agricultural Biotechnology Center, Institute of Molecular Biology, National Chung Hsing University, Taichung 402, Taiwan
Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung 402, Taiwan, Tel.: +886-4-22840338 #7041, Email: ywchu@nchu.edu.tw
John Wang, Department of Pathology, China Medical University Hospital, Taichung 404, Taiwan, Tel.: +886-4-22052121 #2598, E-mail: d96085@mail.cmuh.org.tw
Kai-Po Chang, Ph.D. Program in Medical Biotechnology, National Chung Hsing University, Taichung 402, Taiwan
Kai-Po Chang, Department of Pathology, China Medical University Hospital, Taichung 404, Taiwan

this approach requires a behavioral change by pathologists, it is not widely applied, and pathology reports remain stored in pure text form. To retrieve these free-text data, a text mining approach must be used to avoid manual work. However, most general medical text mining utilities do not specifically process tokens about immunohistochemical (IHC) findings [16. 17], and the few tools that handle IHC data use sophisticated natural language processing (NLP) methods, such as subgraph mining and factorization [18, 19], which require computing powers that are not affordable in the general hospital information system. Moreover, even within these sophisticated tools, no function can yet discriminate between primary and recurrent disease, which makes a comparison of hormone receptor expression between primary and recurrent disease impossible with these tools.

Since focused mining of IHC study data is of small scope, simple methods, such as word/phrase matching, concept-match scrubbing [20], and semantic, grammar-based, concept finding [21], if combined with clinical knowledge, may still achieve good results at a small scale. The current study describes a method to mine IHC data from pathology reports documenting either primary or recurrent breast cancer, using regular expression-based word/phrase matching.

2 Materials and methods

2.1 Data retrieval and pre-processing

All pathology reports issued by the China Medical University Hospital (CMUH) from 2013 to 2017 were exported in pure text form, by a client mentored by Mr. Chi-Sung Wei from the Department of Information. The patient data within the text file was then automatically de-identified by the method described by Neamatullah *et al.* [22] to eliminate violation of privacy and ethical concern. To accelerate the information retrieval from reports, the text files were then archived into a client-side SQLite3 database. Figure 1 illustrates the data retrieval and pre-processing steps. Due to the nature of the study, no ethical approval was required.

2.2 Searching for primary and local recurrent breast cancer cases

Most pathology reports in our institute, as in many other hospitals, are written according to the format suggested

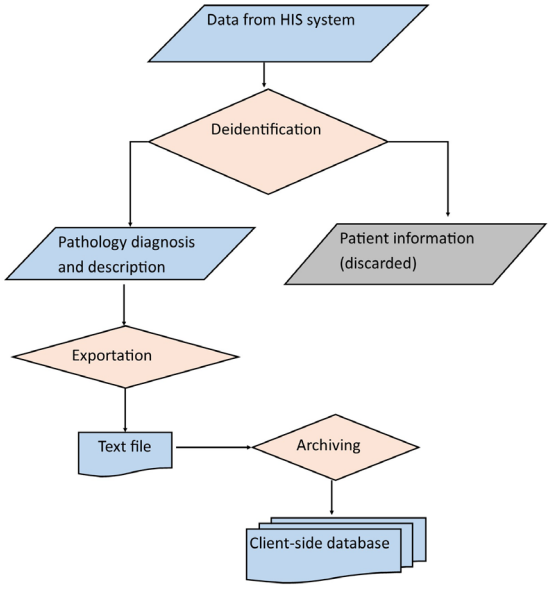

Figure 1: Data retrieval and pre-processing steps.

by "Rosai and Ackerman's Surgical Pathology" [23], and the diagnoses are written in the following format: "Organ name, side/position, surgical procedure, diagnosis". Tokens in the diagnosis section of the pathology report can then be searched according to this format. In the method designed in the current study, the program will first search for organ names "breast" or "chest wall", the sites in which primary or recurrent breast cancer occur. ". If the organ name matches the previous criteria, the program will then search for tokens related to invasive carcinoma.

For pathology reports starting with the organ name "breast", the program will first match for the keyword "carcinoma". If the keyword "carcinoma" is present in a diagnosis, the program will then check if a diagnosis contains keywords that represent carcinoma *in situ* in the World Health Organization definition [24]. All cases that contain the token "carcinoma" and do not contain keywords that represent carcinoma *in situ* will then be included in the breast cancer case list. Cases that represent local, ipsilateral recurrent carcinoma are detected by presence of token "recurrent" or "recurrence". All cases represent new biopsy at the site of recurrent that confirmed the diagnosis once the recurrence occurred.

For pathology reports starting with the organ name "chest wall", all cases with the token "carcinoma" and "breast origin" will be included for analysis and marked as recurrent breast cancer, since all these cases represent local, ipsilateral recurrent carcinoma. Figure 2 shows the search protocol.

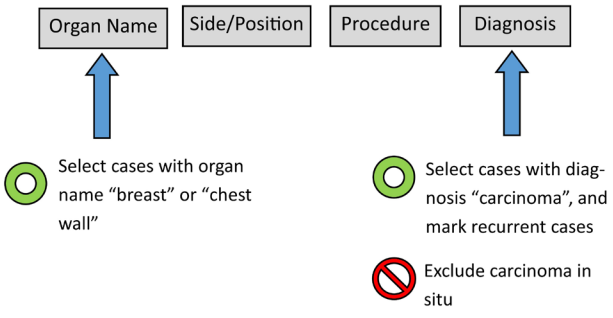

Figure 2: Protocol to search for primary and recurrent breast cancer cases.

2.3 Two-stage data mining approaches for hormone receptor data

For mining of hormone receptor status examined by immunohistochemistry, a two-stage mining approach was designed in the current study, by first extracting the paragraph that may contain IHC study data and then attempting to retrieve ER, PR, and Her-2 results from the mined paragraph. This approach, which is depicted in Figure 3, enhances the execution speed and minimizes the extraction error by matching only a small target, rather than the whole report for IHC study data.

2.3.1 Identification of paragraphs containing IHC study results

To optimize executing speed, a two-step regular expression matching engine for IHC study extraction was designed. In the first step, the program will attempt to match three common forms of IHC study result expression. The first form of reporting IHC study results consists of a separate paragraph in the pathology report, written in multiple rows separated by a line break (Figure 4). In this procedure, every different marker is placed on a new row. The second form comprises a separate paragraph in the pathology report written without a line break. In this approach, the different markers are separated by commas (Figure 5). The third protocol consists of a sentence in the microscopic description, as shown in Figure 6. The identification of paragraphs therefore involves matching the text with one of following regular expression patterns: ["[Ii]mmunohisto.*\)", "[Aa]ncillary.*\)", and "[Ii]mmunostain.\)".

Paragraphs extracted from this step will then undergo extraction of the IHC study result (section 2.3.2).

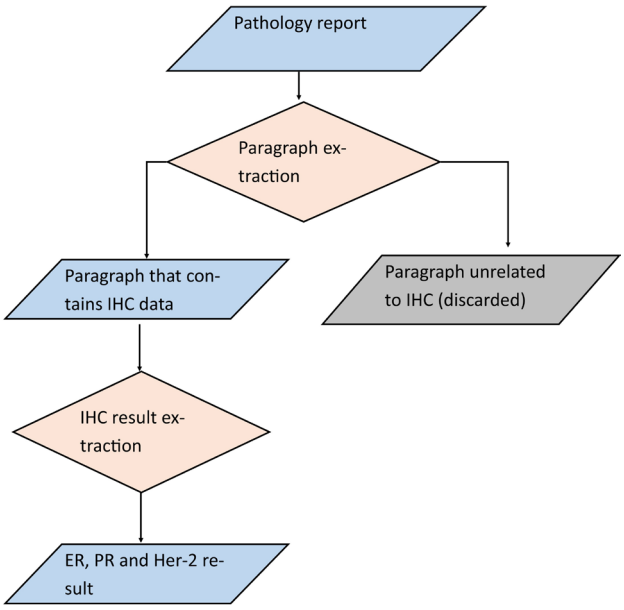

Figure 3: Protocol for mining of hormone receptor data.

2.3.2 Extraction of IHC study results

To extract the results of each separate marker can be a difficult task since there can be unlimited ways to write these results. For institutes that are routinely accredited by the College of American Pathologists (CAP), such as our institute, the format of reporting ER, PR, and Her-2 results is regulated by guidelines [25, 26]. Therefore, in the method described herein, the ER, PR, and Her-2 results are matched and extracted according to the guidelines.

For ER and PR, it is required that pathologists first report the positivity findings. If the result is positive, the expression percentage should be documented. For pathologists who comply to the guideline, it would result in three patterns: "ER/PR (positive, __%)", "ER/PR: positive, __%", and "ER (positive)". The paragraphs containing ER/PR results are parsed by matching the following regular expression: "er\ *[\:\(]" and "pr\ *[\:\(]", while the percentages are by matching the following regular expression: "[0-9]+\%"

For Her-2 results, pathologists must report both the positivity (positive, equivocal, negative) and score (0, 1+, 2+, and 3+). Compliance with this guideline, would result in two patterns: "Her-2/Her2/HER2/HER-2 (positive/equivocal/negative, 0/1+/2+/3+ or score 0/1/2/3)" and "Her-2/Her2/HER2/HER-2 (positive/equivocal/negative, 0/1+/2+/3+ or score 0/1/2/3, weak/moderate/strong staining in __%)". The paragraphs containing Her-2 result are parsed by the matching the following regular expression

```
1.  Diagnosis 1
2.  Diagnosis 2
3.  Diagnosis 3
…… (Other diagnoses)
n. Immunohistochemical study/Ancillary study for invasive tumor cells:
   ER: positive/negative, __%
   PR: positive/negative, __%
   Her-2: positive/equivocal/negative, score 0/1+/2+/3+
…… (Other immunohistochemical study)
n+1. Pathologic staging:
```

Figure 4: Reporting immunohistochemical study result as a solitary paragraph with multiple rows.

```
1.  Diagnosis 1
2.  Diagnosis 2
3.  Diagnosis 3
…… (Other diagnoses)
n. Immunohistochemical study/Ancillary study for invasive tumor cells shows E
(positive/negative, __%), PR(positive/negative, __%), Her-2(positive/equivoca
score 0/1+/2+/3+), and …… (Other immunohistochemical study).
n+1. Pathologic staging:
```

Figure 5: Reporting immunohistochemical study result as a solitary paragraph, with different studies separated by commas.

```
Microscopically, the breast shows invasive…… some description ….. Immunohistc
study/Ancillary study for invasive tumor cells shows ER(positive/negative, __
PR(positive/negative, __%), Her-2(positive/equivocal/negative, score 0/1+/2+/
(Other immunohistochemical study). Breast elsewhere shows …….
```

Figure 6: Reporting immunohistochemical study result as a sentence in the microscopic description.

pattern: “her-*2\ *[\:\(]”, while the Her-2 scores are parsed by matching the following regular expression pattern: “score\ [0-9]+”, “[0-9]\+”.

2.4 Recording of results

The results are exported into a comma-separated text (CSV) file by the program, recording each case in the form: “case ID, cell type, ER result, PR result, Her-2 result, Recurrent or Primary disease”. If there is a failed extraction, the result is recorded as “None”.

2.5 Validation of results

All cases and IHC study results were reviewed by two board-certificated pathologists (Kai-Po Chang and John Wang), for validation.

2.6 Statistical analysis

For comparison of the hormone receptor results between primary and recurrent breast cancer, Pearson’s chi-squared test with Yates’ continuity correction was done using the MASS package of R 3.5.1 for Windows 10

3 Results

3.1 Total case number

Our program identified a total of 3806 invasive breast carcinoma cases, of which 3762 were primary breast cancer, and 44 were local recurrent breast cancer. The cases were all correctly identified and verified.

3.2 Effectiveness of IHC study result detection and extraction

The ER IHC study was done in 3768 of 3806 breast cancer cases. The expression status result was correctly extracted in 3723 cases, yielding a sensitivity of 98.8%. The PR IHC study was done in 3733 of 3806 breast cancer cases. The expression status result was correctly extracted in 3685 cases (98.7% sensitivity). The Her-2 IHC study was done in 3738 of 3806 breast cancer cases. The expression status result was correctly extracted in 3679 cases (98.4% sensitivity). The result is shown in Table 1.

3.3 Comparison of hormone receptor expression between primary and recurrent breast cancer

Of 3723 cases with ER IHC study results, 3679 had primary disease, and 44 had local recurrent disease. Among 3679 primary disease cases, 2814 were ER-positive (76.5%). Among 44 recurrent disease cases, 24 were ER-positive (54.5%). It indicates that primary disease is is more prone to be ER-positive than recurrent disease (χ^2=10.374, df=1, p=0.001278). The result is shown in Table 2.

Of 3685 cases with machine-identified PR IHC study results, 3642 were primary disease, and 43 were local recurrent disease. Among 3642 primary disease cases, 2299 were PR-positive (63.1%). Among 43 recurrent disease cases, 18 were PR-positive (41.9%). It indicates that primary disease is more prone to be PR-positive than recurrent disease (χ^2=7.3467, df=1, p=0.006719). The result is shown in Table 3.

Of 3679 cases with machine-identified Her-2 IHC study results, 3636 had primary disease, and 43 had recurrent disease. Among 3642 primary disease cases, 589 were Her-2-positive (defined as score 3+) by IHC criteria, 1024 were Her-2-equivocal (defined as score 2+), and 2023 were Her-2-negative (defined as score 0 or 1+; 1074 with score 0 and 949 with score 1+). The overall Her-2 positivity is therefore 16.2% in primary breast cancer. Among 43 recurrent disease cases, 21 were Her-2-positive (48.8%), 12 were Her-2-equivocal (27.9%), and 10 were Her-2-negative (23.2%; 6 with score 0 and 4 with score 1+). It indicates the recurrent disease is more prone to be Her-2 positive than recurrent disease (χ^2=35.449, df=3, p=9.79e-8). The result is shown in Table 4.

This observation that the recurrent disease is more prone to be ER-negative, PR-negative, and Her-2-positive is consistent with the previous literature [12].

4 Discussion

4.1 Sensitivity issue of regular expression-based word/phrase matching

The proposed program failed to detect and extract hormone receptor data in some cases (1.2% for ER, 1.3% for PR, and 1.6% for Her-2). Most of these were caused by failure to detect phrases containing the IHC study result. During the manual examination of the failed cases, at least three alternative ways other than our target are found. These alternative patterns included "The immunohistochemical study.....", "Ancillary study for tumor cells.....", "Tumor immunoprofile:", and *vice versa*. There are simply too many ways to express the IHC study result, so the simple regular expression-based word/phrase matching strategy

Table 1: Summary of results of the extraction of immunohistochemical study result data.

Marker	Total sample number	Number of results correctly detected	Sensitivity
ER	3768	3723	98.8%
PR	3733	3685	98.7%
Her-2	3738	3679	98.4%

Table 2: Difference in ER expression between primary and recurrent breast cancer.

ER result	Primary breast cancer	Recurrent breast cancer
Positive	2814	24
Negative	865	20
Positive rate	76.5%	54.5%
χ^2=10.374, df=1, p=0.001278		

Table 3: Difference in PR expression between primary and recurrent breast cancer.

PR result	Primary breast cancer	Recurrent breast cancer
Positive	2699	18
Negative	1343	25
Positive rate	63.1%	41.9%
χ^2=7.3467, df=1, p=0.006719		

Table 4: Difference in Her-2 expression between primary and recurrent breast cancer.

Her-2 result	Primary breast cancer	Recurrent breast cancer
Score 0	1074	6
Score 1	949	4
Score 2	1024	12
Score 3	589	21
Positive rate (positive defined as score 3+)	16.2%	48.8%
χ^2=35.449, df=3, p=9.79e-8		

cannot be expected to match all pathologic reports that contain these data.

In some cases, the IHC study paragraph was identified, but the program still failed to extract one or more of the hormone receptor data. In this situation, the issue was in the expression pattern of the hormone receptors. For example, the program aimed at the phrase "ER (positive, 90%)" or "ER: positive, strong expression in 90%", for extraction of ER positivity and percentage, but this is not the only pattern that can be used to write the ER expression. For example, some pathologists prefer a narrative form, and report the ER expression status as "The tumor is positive for ER". For positivity, some pathologists prefer the shorter phrase "ER+" rather than "ER positive". These variant writing habits of pathologists accounted for some cases in which the program failed to detect ER or PR data.

Writing habit is an even more serious issue in Her-2 data extraction because this tumor marker can be expressed in numerous ways, including "Her-2/*neu*", "HER2", or "Her-2". Moreover, not all pathologists report the Her-2 result according to the CAP recommendation. For instance, if a Her-2 IHC slide is positive according to the regulation, it should be reported as "Her-2: positive, score 3+" or "Her-2: positive, score 3/3". However, some pathologists merely report it as positive, without mentioning the score, or write the score as "+++" rather than "3+". For human interpreters, these expressions are readable without any difficulty but, for the machine, it represents a problem.

In general, though the proposed program achieved acceptable sensitivity, regular expression-based text mining does have its limitations.

4.2 Possible direction for further development of NLP engine for hormone receptor status mining

This study demonstrated the limitations of the conventional expression-based approach. To solve the issue of variable syntax used in pathology reports, distributional semantic modeling [27] may be a solution. For distributional semantic modeling, a large enough text corpus is first built from pathology reports. The semantic modeling engine then analyzes relationships between words in the corpus. In this way, similar phrases that are close to the mining target can be learned and included in the mining algorithm. Distributional semantics have been used successfully in searching the medical literature when an appropriate corpus is given [28], and it would have a role in the mining of hormone receptor data as well.

Another promising method for hormone receptor status mining is neurolinguistics. Since the syntax variations in electronic health records are due to differences in expression of natural language, some researchers have achieved a good result in data mining of medical texts, by applying models that imitate the human brain [29, 30]. Given the difficulty in mining hormone receptor status also comes from syntax variability, neurolinguistics modeling may further increase the sensitivity of data detection described in the current study.

In conclusion, the proposed method achieved a good result in mining the hormone receptor status from breast cancer cases, but it can still be improved. With the progression of NLP technology and computing power, the obstacle encountered in this study may eventually be resolved.

4.3 Limitation of hormone receptor status comparison between primary and recurrent breast cancer via data mining

The present article demonstrated a significant difference in the hormone receptor expression between primary and recurrent breast cancer, which is consistent with previous literature [12]. However, the statistical power of this result

is limited, due to the relatively small sample number of recurrent breast cancer cases. Two inherent limitations of acquiring cases of recurrent disease via data mining of pathology reports exist. The first one is the scarcity of recurrent breast cancer cases in pathology archives because some of the recurrent cancer cases are diagnosed by imaging, and no tissue is sent to the pathology department for validation. The other limitation is the under-reporting of recurrent diseases. Even if a case of recurrent breast cancer is sent to the pathology department for tissue confirmation, it may not be correctly reported as recurrent disease.

To solve these issues, the case number can be expanded by combining data from multiple institutes, to increase the total number of recurrent diseases or deeper mining of the health record database can be performed, to directly detect recurrent disease from the patient history, thereby avoiding the problem of under-reporting. With future expansion of the database and improved data mining algorithms, data mining can have a great impact on recurrent breast cancer research.

Conflict of interest: The authors declare no conflict of interest.

Acknowledgments: This research was supported by the Ministry of Science and Technology, Taiwan, R.O.C. [grant numbers 106-2221-E-005-077-MY2, 107-2634-F-005-002, and 107-2321-B-005 -013], and the National Chung Hsing University and Chung-Shan Medical University [grant number NCHU-CSMU-10705].

K.P.C. and Y.W.C conceived the program. K.P.C, Y.W.C, and J.W. performed the tests and validated the results. All authors analyzed the data, and contributed to writing and editing the manuscript. All authors approved the final document.

References

[1] Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010; 363(6): 501-504. [https://doi.org/10.1056/NEJMp1006114]

[2] Nagtegaal ID, van Krieken JHJM. The role of pathologists in the quality control of diagnosis and treatment of rectal cancer—an overview. Eur J Cancer 2002; 38(7): 964-972. [https://doi.org/10.1016/S0959-8049(02)00056-4]

[3] Carlson RW, Moench SJ, Hammond ME, et al. HER2 testing in breast cancer: NCCN Task Force report and recommendations. J Natl Compr Cancer Network 2006; 4(Suppl 3): S1-22.

[4] Harvey JM, Clark GM, Osborne CK, Allred DC. Estrogen receptor status by immunohistochemistry is superior to ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. J Clin Oncol 1999; 17(5): 1474-1481. [https://doi.org/10.1200/jco.1999.17.5.1474]

[5] Nadji M, Gomez-Fernandez C, Ganjei-Azar P, Morales AR. Immunohistochemistry of estrogen and progesterone receptors reconsidered. Am J Clin Pathol 2005; 123(1), 21-27. [https://doi.org/10.1309/4WV79N2GHJ3X1841]

[6] Paik S, Hazan R, Fisher ER, et al. Pathologic findings from the National Surgical Adjuvant Breast and Bowel Project: prognostic significance of erbB-2 protein overexpression in primary breast cancer. J Clin Oncol 1990; 8(1): 103-12. [https://doi.org/10.1200/JCO.1990.8.1.103]

[7] Sotiriou C, Neo S, McShane LM, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. PNAS 2003; 100(18): 10393-10398. [https://doi.org/10.1073/pnas.1732912100]

[8] Baselga J, Norton L, Albanell J, Kim Y, Mendelsohn J. Recombinant humanized anti-HER2 antibody (HerceptinTM) enhances the antitumor activity of paclitaxel and doxorubicin against HER2/neu overexpressing human breast cancer xenografts. Cancer Res 1998; 58(13): 2825-2831. Erratum in: Cancer Res 1999; 59(8): 2020.

[9] Ellis MJ, Coop A, Singh B, et al. Letrozole is more effective neoadjuvant endocrine therapy than tamoxifen for ErbB-1- and/or ErbB-2-positive, estrogen receptor-positive primary breast cancer: evidence from a phase III randomized trial. J Clin Oncol 2001; 19(18): 3808-3816. [https://doi.org/10.1200/JCO.2001.19.18.3808]

[10] Giuliano AE, Connolly JL, Edge SB, et al. Breast cancer—Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. CA Cancer J Clin 2017; 67: 290-303. [https://doi.org/10.3322/caac.21393]

[11] Hammond MEH, Hayes DF, Wolff AC, Mangu, PB, Temin S. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. J Oncol Pract 2010; 6(4): 195-197. [https://doi.org/10.1200/JOP.777003]

[12] Nguyen PL, Taghian AG, Katz, MS, et al. Breast cancer subtype approximated by estrogen receptor, progesterone receptor, and HER-2 is associated with local and distant recurrence after breast-conserving therapy. J Clin Oncol 2008; 26(14): 2373-2378. [https://doi.org/10.1200/JCO.2007.14.4287]

[13] Casati B, Haugland HK, Barstad GMJ, Bjugn, R. Implementation and use of electronic synoptic cancer reporting: an explorative case study of six Norwegian pathology laboratories. Implementation Sci. 2014; 9: 111. [https://doi.org/10.1186/s13012-014-0111-2]

[14] Leong AS. Synoptic/checklist reporting of breast biopsies: has the time come? Breast J 2001; 7(4), 271-274. [https://doi.org/10.1046/j.1524-4741.2001.21001.x]

[15] Srigley JR, McGowan T, MacLean A, et al. Standardized synoptic cancer pathology reporting: a population-based approach. J Surg Oncol 2009; 99(8): 517-24. [https://doi.org/10.1002/jso.21282]

[16] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001; 17-21. [PMCID: PMC2243666] [PMID: 11825149]

[17] Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. JAMIA 2010; 17(5): 507-513. [https://doi.org/10.1136/jamia.2009.001560]

[18] Luo Y. Towards unified biomedical modeling with subgraph mining and factorization algorithms [dissertation]. Cambridge (MA): Massachusetts Institute of Technology; 2015.

[19] Luo Y, Sohani AR, Hochberg EP, Szolovits P. Automatic lymphoma classification with sentence subgraph mining from pathology reports. JAMIA 2014; 21(5): 824-832. [https://doi.org/10.1136/amiajnl-2013-002443]

[20] Berman JJ. Concept-match medical data scrubbing. How pathology text can be used in research. Arch Pathol Lab Med 2003; 127(6): 680-686. [https://doi.org/10.1043/1543-2165(2003)127<680:CMDS>2.0.CO;2]

[21] Nassif H, Woods R, Burnside E, Ayvaci M, Shavlik J, Page D. Information extraction for clinical data mining: a mammography case study. Proceedings of the 2009 IEEE International Conference on Data Mining Workshops (ICDMW 2009) (pp. 37-42); 2009 Dec 6; Miami, Florida. Available from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3676897/ [https://doi.org/10.1109/ICDMW.2009.63]

[22] Neamatullah I, Douglass MM, Lehman L, et al. Automated de-identification of free-text medical records. BMC Med Inf Decis Making 2008; 8(1): 32. [https://doi.org/10.1186/1472-6947-8-32]

[23] Rosai J. Rosai and Ackerman's surgical pathology. 10th Ed. China: Elsevier Inc. 2011. Available from http://books.google.com/books?id=1CKX7aGBbUsC&pgis=1

[24] Lakhani SR, Ellis IO, Schnitt SJ, Tan PH van de Vijver MJ, Eds. World Health Organization classification of tumours. WHO classification of tumours of the breast. 4th Ed. Lyon, France: International Agency for Research on Cancer (IARC) 2014.

[25] Fitzgibbons P, Murphy DA, Hammond MEH, Allred DC, Valenstein PN. Recommendations for validating estrogen and progesterone receptor immunohistochemistry assays. Arch Pathol Lab Med 2010; 134(6): 930-935. [https://doi.org/10.1043/1543-2165-134.6.930]

[26] Wolff AC, Hammond MEH, Hicks DG, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer. American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. J Clin Oncol 2013; 31(31): 3997-4013. [https://doi.org/10.1200/JCO.2013.50.9984]

[27] Marelli M, Bentivogli L, Baroni M, Bernardi R, Menini S, Zamparelli R. SemEval-2014 task 1: evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 1–8); 2014 Aug 23-24; Dublin, Ireland. Available from http://alt.qcri.org/semeval2014/cdrom/pdf/SemEval001.pdf

[28] Pakhomov SVS, Finley G, McEwan R, Wang Y, Melton GB. Corpus domain effects on distributional semantic modeling of medical terms. Bioinformatics 2016; 32(23): 3635-44. [https://doi.org/10.1093/bioinformatics/btw529]

[29] Duch W, Matykiewicz P, Pestian J. Neurolinguistic approach to natural language processing with applications to medical text analysis. Neural Networks 2008; 21(10): 1500-1510. [https://doi.org/10.1016/j.neunet.2008.05.008]

[30] Matykiewicz P, Duch W, Zender PM. Crutcher KA, Pestian JP. Neurocognitive approach to clustering of PubMed query results. In: Köppen M, Kasabov N, Coghill G, Eds. Advances in Neuro-Information Processing. ICONIP 2008. Lecture notes in computer science. Vol. 5507. Berlin/Heidelberg, Germany: Springer 2009; pp. 70-79. [https://doi.org/10.1007/978-3-642-03040-6_9]