Development of a Novel Tool for the Retrieval and Analysis of Hormone Receptor Expression Characteristics in Metastatic Breast Cancer via Data Mining on Pathology Reports

Kai-Po Chang
China Medical University Hospital
John Wang
China Medical University Hospital
Cheng-Hsi Liao
Taichung Armed Forces General Hospital
Yen-Wei Chu (ywchu@nchu.edu.tw)
National Chung Hsing University College of Engineering https://orcid.org/0000-0002-5525-4011

Research article

Keywords: breast cancer; hormone receptor; text mining; natural language processing

Posted Date: August 4th, 2019

DOI: https://doi.org/10.21203/rs.2.11642/v1

Version of Record: A version of this preprint was published at BioMed Research International on May 27th, 2020. See the published version at https://doi.org/10.1155/2020/2654815.

Abstract

Background Information about the expression status of hormone receptors such as estrogen receptor (ER), progesterone receptor (PR) and Her-2 is crucial in the management and prognosis of breast cancer. Therefore, the retrieval and analysis of hormone receptor expression characteristics in metastatic breast cancer may be valuable in breast cancer study. Methods Herein, we report a text mining tool based on word/phrase matching that retrieves hormone receptor expression data of regional or distant metastatic breast cancer from pathology reports. It was tested on pathology reports at the China Medical University Hospital from 2013 to 2018. Results The tool showed specificities of 91.6% and 63.3% for the detection of regional lymph node metastasis and distant metastasis, respectively. Sensitivity in immunohistochemical study result extraction in these cases was 98.6% for distant metastasis and 78.3% for regional lymph node metastasis. Statistical analysis on these retrieved data showed significant differences in PR and Her-2 expression between regional and metastatic breast cancer, which is compatible with previous studies. Conclusion In conclusion, our study shows that metastatic breast cancer hormone receptor expression characteristics can be retrieved by text mining. Algorithm designed in this study may be useful in future studies about text mining in pathology reports.

Background

Breast cancer is the second most lethal cancer worldwide, accounting for 626,679 deaths in 2018 [1]. These fatalities are primarily due to its potential to metastasize, with 28.8% of patients experiencing axillary lymph node metastases [2] and 20-30% of patients experiencing subsequent distant metastasis even if the cancer is found in an early stage [3]. Therefore, study on the behavior of metastatic breast cancer is of particular importance in breast cancer treatment and public health. During the previous two decades of medical advancement, numerous novel molecular targets, such as LIFR [4], PI3K [5] and aldehyde dehydrogenase-1 [6], have been studied for prognosis prediction and target therapy for metastatic breast cancer, but none of them have proven to be more valuable than the long-standing markers estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (erbb-2 or Her-2). According to recent studies, molecular subtypes luminal A, luminal B, Her-2 and triple-negative, which are determined by these markers, are still relevant to the treatment and prognosis of metastatic breast cancer [7–10].

As important markers of special value, ER, PR and Her-2 expression are routinely examined by immunohistochemical study [11–13] on all invasive breast cancer slides and are documented in pathology reports. Combined with the fact that occurrences of lymph node or distant metastatic breast cancer are frequently sampled for pathologic examination [14], a pathology report database may be an important resource for the hormone receptor expression status of metastatic breast cancer. However, extraction of these data can be a tedious task. Unlike surgical pathology reports for primary breast cancer, in which pathologists are required to report in certain forms [15] or a synaptic report system [16–18], there are no required forms for reporting metastatic carcinoma in most institutions, and most of these reports stay in free text form. Retrieving these data requires text mining approaches to avoid tedious manual work. As we have discussed in a previous article [19], most general medical text mining utilities do not process immunohistochemical study results [20, 21], while those that do process immunohistochemical data use advanced natural language processing (NLP) methods [22, 23] and therefore will not be available in general hospital information system (HIS).

This difficulty can be solved by using simpler methods such as word/phrase matching, concept-match scrubbing [24] and semantic grammar-based concept finding [25] with clinical knowledge. We have shown in a previous publication [19] that regular expression-based word/phrase matching can be used to mine hormone receptor data for primary

and recurrent breast cancer. In this article, we show that the text mining algorithm described in the previous publication can also be applied to metastatic breast cancer.

Methods

Data Retrieval and Preprocessing

All pathology reports issued at the China Medical University Hospital (CMUH) from the years 2013 to 2018, estimated 200,000 reports were first exported into pure text form. The patient data within the text file was then automatically deidentified using the method described by Neamatullah et al.[26] to eliminate violation of privacy and ethical concerns. A Python script [27] was designed to extract the pathology diagnosis and description columns from the text files and build a client-side database using SQLite3 [28]. The data retrieval and preprocessing steps are shown in figure 1.

Figure 1. Data retrieval and preprocessing steps.

Retrieval of metastatic breast cancer cases

The authors first manually reviewed 50 pathology reports documenting regional lymph node metastatic breast cancer and 50 pathology reports documenting distant metastatic breast cancer. From these reports, it was seen that most pathology reports documenting a metastatic carcinoma had either “carcinoma, metastatic” or “carcinoma, involved” in the diagnosis. Those of breast origin were described as “breast origin” or “breast primary”. Regional lymph node metastatic tumors were described as “soft tissue, axillary” or “lymph node, axillary”, while distant metastatic tumor were described in the pattern “<any organ name other than axillary tissue>, <procedure>, carcinoma, metastatic/involved, breast origin”.

Based on these results, we designed our metastatic breast cancer finding algorithm according to the following strategy:

Each line from the diagnostic column is matched with the phrase “carcinoma, metastatic”, “carcinoma, involved” or any phrase indicating metastatic carcinoma by a regular expression engine. If any of the lines matched one of the patterns, the report is passed to the next step for further processing.

When one of the lines in the diagnosis indicates metastatic carcinoma, that line is checked for the presence of phrases that indicate breast origin, such as “breast primary” or “breast origin”. Any report that shows a match in these phrases is passed into the next step for examination.

For reports that show evidence of metastatic carcinoma of breast origin, the whole diagnostic column is checked for the presence of signs of primary breast cancer. If any of the lines from the diagnostic column shows any phrase that represents primary breast cancer, the report is excluded from further analysis.

Metastatic sites are parsed and recorded by another regular expression engine.

490 reports documenting metastatic disease (359 regional metastases, 131 distant metastases) are retrieved in this step. The search protocol is shown in figure 2.

Figure 2. Protocol for searching metastatic breast cancer cases.

Identification of paragraphs containing immunohistochemical study results

A two-step regular expression matching engine for immunohistochemical study extraction, as described in our previous study on extracting immunohistochemical result of primary and recurrent breast cancer [19], was utilized. In the first step, the program attempted to match common forms in which pathologists express immunohistochemical study results.

There is, however, a significant difference between identification of immunohistochemical study in primary/recurrent breast cancer and metastatic breast cancer. When reporting metastatic carcinoma, pathologists in our institution usually document immunohistochemical study results in the description rather than diagnostic column; therefore, searching immunohistochemistry-containing paragraphs in the current study only involved parsing the description column (figure 5) but not the diagnosis column (figure 3-4). This approach can optimize the searching process without sacrificing sensitivity.

Figure 3. Reporting immunohistochemical study results as a solitary paragraph with multiple rows.

Figure 4. Reporting immunohistochemical study results as a solitary paragraph, with different studies separated by commas.

Figure 5. Reporting immunohistochemical study results as a sentence in the microscopic description.

Paragraphs extracted from this step will then undergo the following steps for immunohistochemical study result extraction.

Extraction of immunohistochemical study results

In institutes that are routinely accredited by the College of American Pathologists (CAP), such as our institute, the reporting format of ER, PR and Her-2 result is regulated by guidelines [29, 30]. Therefore, in our method, the results of ER, PR and Her-2 result are matched and extracted according to those guidelines.

For ER and PR, positivity is required. If the result is positive, the expression percentage should be reported. Therefore, there would be three patterns: “ER/PR (positive, __%)”, “ER/PR: positive, __%”, and “ER (positive)”.

For Her-2 results, both positivity (positive, equivocal, negative) and score (0, 1+, 2+ and 3+) are required. Therefore, there would be 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 __%”.

Recording of results

The results are exported into a csv file by the program, recording each case in the form: “case ID, metastatic site, ER result, PR result, Her-2 result”. If there is a failed extraction, the result is recorded as “None”.

Validation of results

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

Statistical analysis

For comparison of hormone receptor results between different metastatic sites, Pearson's Chi-squared test with Yates' continuity correction was done with the MASS package of R version 3.5.1 under Windows 10.

Results

Detection of metastatic breast cancer cases

Our program labeled 131 pathology reports as describing distant metastatic breast cancer, of which 83 were correctly labeled, resulting in a specificity of 63.3%. There were 359 pathology reports labeled as describing regional lymph node metastatic breast cancer, of which 329 were correctly labeled, resulting in a specificity of 91.6%. Sensitivity could not be determined, since there is no cancer registry data for metastatic carcinoma. The results are summarized in table 1.

Among the 83 cases of distant metastatic cancer, the metastatic sites include nonregional lymph node (22 cases), bone (20 cases), brain (12 cases), liver (8 cases), gastrointestinal tract (8 cases), lung (7 cases), uterus (1 case), pleura (1 case), pelvic cavity (1 case), ovary (1 case), mediastinum (1 case) and urinary bladder (1 case). The results are summarized in table 2.

Table 1 (see Supplementary Files). Summary of the results of metastatic breast cancer detection.

Table 2 (see Supplementary Files). Summary of metastatic sites.

Immunohistochemical study result detection and extraction

In the 83 cases documenting distant metastatic disease, the program detected immunohistochemical study results in 65 cases, with an error in documentation of the immunohistochemical study result in 1 case, resulting in a sensitivity of 78.3% and a specificity of 98.4%. In 329 cases documenting regional lymph node metastatic diseases, the program correctly detected immunohistochemical study results in 316 cases, resulting in a sensitivity of 98.1% and a specificity of 100%. The results are documented in table 3.

Among the 64 cases of distant metastatic cases with correctly detected immunohistochemical study results, all were tested for ER, 52 were tested for PR, and 58 were tested for Her-2. Of the cases tested for ER, 36 (62.0%) were positive, and 28 (38.0%) were negative. Of the cases tested for PR, 12 (23.0%) were positive, and 40 (67.0%) were negative. Of the cases tested for Her-2, 23 (39.6%) were positive (score 3+), 11 (19.0%) were equivocal (score 2+), and 24 (41.4%) were negative (score 1+ or 0). The results are shown in table 4.

Among the 322 cases of regional lymph node metastatic cases with correctly detected immunohistochemical study results, 308 were tested for ER, 91 were tested for PR, and 303 were tested for Her-2. Of the cases tested for ER, 198 were positive, and 110 were negative. Of the cases tested for PR, 52 were positive, and 29 were negative. Of the cases tested for Her-2, 103 were positive (score 3+), 95 were equivocal (score 2+), and 112 were negative (score 1+ or 0). The results are shown in table 5.

Table 3 (see Supplementary Files). Summary of results of the extraction of immunohistochemical study result data.

Table 4 (see Supplementary Files). Summary of immunohistochemical study results of distant metastatic tumors.

Table 5 (see Supplementary Files). Summary of immunohistochemical study results of regional metastatic tumors.

Comparison of hormone receptor expression between lymph node metastatic breast cancers

After applying chi-square tests to the above results, it was concluded that distant metastatic tumors had a significantly higher probability to be Her-2 positive and PR negative than did regional metastatic tumors, while there was no significant difference between ER expression in regional and distant metastatic disease. For details, please see tables 6-8.

Table 6 (see Supplementary Files). Difference of ER expression between distant and regionally metastatic breast cancers.

Table 7 (see Supplementary Files). Difference of PR expression between distant and regionally metastatic breast cancers.

Table 8 (see Supplementary Files). Difference of Her-2 expression between expression between distant and regionally metastatic breast cancers.

Our observation that distant metastatic tumors are more prone to be Her-2 positive and PR-negative may be consistent with previous studies that Her-2 positive and PR-negative tumor have higher incidence of distant metastasis [31, 32].

Comparison of hormone receptor expression between major metastatic sites

According to our data, compared with bone and brain metastatic disease, lung metastatic disease has a tendency to be more ER-positive and Her-2 positive, which is consistent with previous studies [31]. However, there is no statistically significant difference in the chi-square analysis, which is probably due to low a sample number. Details are shown in tables 9-11.

Table 9 (see Supplementary Files). ER expression status of major metastatic sites.

Table 10 (see Supplementary Files). PR expression status of major metastatic sites.

Table 11 (see Supplementary Files). Her-2 expression status of major metastatic sites.

Discussion

Specificity issue of distant metastatic cases detection

The most significant flaw in our approach on metastatic breast cancer mining is its low specificity in distant metastatic cases. Of the 47 cases in which the program marked the report as a metastatic carcinoma but it actually was not, most (35) of them were documenting soft tissue or skin of the chest wall involved in recurrent breast cancer, in which the case should have been labeled as recurrent disease, not metastatic disease. Of the remaining wrongly marked cases, 11 of the 12 were due to a particular special habit of some pathologists when reporting negative sentinel lymph nodes, in which a phrase “s/p breast cancer” is inserted to the diagnosis to specify that the patient

has undergone previous surgery for breast cancer. The last case is an endometrial curettage report, in which the pathologist noted in the diagnosis that the patient was under Tamoxifen treatment for breast cancer.

Chest wall recurrent cases misinterpreted as metastatic carcinoma occurred most often, but they may be the most easily handled. In our previous publication [19], we developed an algorithm that detects recurrent carcinoma at either the breast or chest wall. If combined with that algorithm, chest wall recurrent cases can be easily filtered out. The cases in which the pathologist mentioned breast cancer in otherwise nonmalignant reports is a more difficult issue, since interpretation of that phrase will require semantic understanding of the pathology report.

To solve this problem, rule-based approaches, such as one described by Hur et al. [33] for mining biomedical literature, and another described by Yang et al. [34] for mining hospital records, may be developed. However, since the pathology reports are written quite liberally, it is questionable whether specific rules can be built to fit theoretically infinite numbers of possible writing combinations on a pathology report. A more recent text-mining method is distributional semantic modeling [35]. In this method, corpora of text are first given, and the relationships between all words, including similarity and relatedness, is measured by vector-assisted analysis of coexistence in the corpus. This approach maybe more feasible, since this method would recognize the semantics of pathology reports. Subgraph mining [22, 23, 36] that deconstructs the whole pathology report into higher order elements (subgraphs) maybe helpful as well. With recent advancements in text mining technology, new methods will emerge, and the problem encountered in our study may be overcome.

Further Research Directions

This study confirmed the concern in our previous publication that nonstandardized pathology report may pose a difficulty in text mining, but we have discussed in the previous paragraph that it can be solved. By altering regular expression patterns, multiple forms of pathology report writing can be parsed and mined. Another issue mentioned in our previous publication, variation in reporting immunochemical study result, is nevertheless still not solved. Since we only have reports from one institution, it is unknown if our program works in pathology reports elsewhere. Therefore, for researchers in text mining, exploring the various forms in which hormone receptors such as ER, PR and Her-2 are expressed may be an interesting and realistic research target. As we have stated above, the detection of metastatic disease, because of its difficulty, is also a potential research project.

Conclusions

In conclusion, our program showed that in metastatic breast cancer, the ER, PR and Her-2 immunohistochemical study data can be mined using simple word/phrase matching assisted by regular expression. Algorithm designed in this study may be useful in future studies about text mining in pathology reports.

Declarations

Availability of data and materials

The data generated by the script is provided in additional file 1.

Sample code was deposited in GitHub https://github.com/medchem/breastmeta/

Authors' contributions

K.P.C. and Y.W.C conceived and designed the program; K.P.C, Y.W.C and C.H.L. performed the tests; K.P.C and J.W. validated the results; all authors analyzed the data; all authors wrote, edited and approved the manuscript.

Ethics approval and consent to participate

Since these data are from a deidentified pathology report database, ethics approval may not be necessary.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by :

Ministry of Science and Technology, Taiwan, ROC under grant number 106-2221-E-005-077-MY2,107-2634-F-005-002 and 107-2321-B-005 -013. This funding is supported for the salaries on research assistants and graduate students;

National Chung Hsing University and Chung-Shan Medical University under grant number NCHU-CSMU-10705. This funding is applied to research equipment, stationery, and printing fee

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. BrayF, FerlayJ, SoerjomataramI, SiegelRL, TorreLA, JemalA. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68:394–424. doi:10.3322/caac.21492.

2. LeeJH, KimSH, SuhYJ, ShimBY, KimHK. Predictors of Axillary Lymph Node Metastases (ALNM) in a Korean Population with T1-2 Breast Carcinoma: Triple Negative Breast Cancer has a High Incidence of ALNM Irrespective of the Tumor Size. Cancer Res Treat. 2010;42:30. doi:10.4143/crt.2010.42.1.30.

3. AbeO, AbeR, EnomotoK, KikuchiK, KoyamaH, MasudaH, et al. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: An overview of the randomised trials. Lancet. 2005;365:1687–717. doi:10.1016/S0140-6736(05)66544-0.

4. HungM-C, GuptaS, RezaeianAH, SunY, WeiY, MaL, et al. LIFR is a breast cancer metastasis suppressor upstream of the Hippo-YAP pathway and a prognostic marker. Nat Med. 2012;18:1511–7. doi:10.1038/nm.2940.

5. Gonzalez-AnguloAM, Ferrer-LozanoJ, Stemke-HaleK, SahinA, LiuS, BarreraJA, et al. PI3K Pathway Mutations and PTEN Levels in Primary and Metastatic Breast Cancer. Mol Cancer Ther. 2011;10:1093–101. doi:10.1158/1535-

7163.mct-10-1089.

6. CharafeJauffretE, GinestierC, IovinoF, TarpinC, DiebelM, EsterniB, et al. Aldehyde dehydrogenase 1-positive cancer stem cells mediate metastasis and poor clinical outcome in inflammatory breast cancer. Clin Cancer Res. 2010;16:45–55. doi:10.1158/1078-0432.CCR-09-1630.

7. TelliML. Triple-negative breast cancer. Mol Pathol Breast Cancer. 2016;363:71–80. doi:10.1007/978-3-319-41761-5_6.

8. AkramiM, TahmasebiS, ZangouriV, HosseiniS, TaleiA. Metastatic Behavior of Breast Cancer Subtypes. Multidiscip Cancer Investig. 2017;1 Supplementary 1:0–0. doi:10.21859/mci-supp-102.

9. TranB, BedardPL. Luminal-B breast cancer and novel therapeutic targets. Breast Cancer Research. 2011;13. doi:10.1186/bcr2904.

10. MalikF, IthimakinS, DayK, IgnatoskiK, ZenQ, ThomasD, et al. Abstract 3470: HER2 drives luminal breast cancer stem cells in the absence of HER2 amplification: Implications for efficacy of adjuvant trastuzumab. Cancer Res. 2012;72 8 Supplement:3470–3470. doi:10.1158/1538-7445.am2012-3470.

11. NadjiM, Gomez-FernandezC, Ganjei-AzarP, MoralesAR. Immunohistochemistry of estrogen and progesterone receptors reconsidered: Experience with 5,993 breast cancers. Am J Clin Pathol. 2005;123:21–7. doi:10.1309/4WV79N2GHJ3X1841.

12. Harvey JM, Clark GM, Osborne CKAD. Estrogen receptor status by immunohistochemistry is superior to ligand binding assay for predicting response to adjuvant therapy in breast cancer. J Clin Oncol. 1999;17:1474–81.

13. CarlsonRW, MoenchSJ, HammondME, PerezEA, BursteinHJ, AllredDC, et al. HER2 testing in breast cancer: NCCN Task Force report and recommendations. J NatlComprCancNetw. 2006;4 Suppl 3:S1-22.

14. TsaoMN, RadesD, WirthA, LoSS, DanielsonBL, GasparLE, et al. Radiotherapeutic and surgical management for newly diagnosed brain metastasis(es): An American Society for Radiation Oncology evidence-based guideline. Pract Radiat Oncol. 2012;2:210–25. doi:10.1016/j.prro.2011.12.004.

15. TobiasJ, ChilukuriR, KomatsoulisGA, MohantyS, SioutosN, WarzelDB, et al. The CAP cancer protocols – a case study of caCORE based data standards implementation to integrate with the Cancer Biomedical Informatics Grid. BMC Med Inform Decis Mak. 2006;6:25. doi:10.1186/1472-6947-6-25.

16. CasatiB, HauglandHK, BarstadGMJ, BjugnR. Implementation and use of electronic synoptic cancer reporting: An explorative case study of six Norwegian pathology laboratories. Implement Sci. 2014;9:111. doi:10.1186/s13012-014-0111-2.

17. LeongASY. Synoptic/checklist reporting of breast biopsies: Has the time come? Breast J. 2001;7:271–4.

18. SrigleyJR, McgowanT, MacleanA, RabyM, RossJ, KramerS, et al. Standardized synoptic cancer pathology reporting: A population-based approach. Journal of Surgical Oncology. 2009;99:517–24. doi:10.1002/jso.21282.

19. ChangKP, ChuYW, WangJ. Analysis of hormone receptor status in primary and recurrent breast cancer via data mining pathology reports. Open Med. 2019;14:91–8. doi:10.1515/med-2019-0013.

20. SavovaGKG, MasanzJJ, OgrenPVP, ZhengJ, SohnS, Kipper-SchulerKC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Informatics Assoc. 2010;17:507–13. doi:10.1136/jamia.2009.001560.

21. AronsonAR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings. AMIA Symposium. 2001. p. 17–21. doi:D010001275 [pii].

22. LuoY. Towards Unified Biomedical Modeling with Subgraph Mining and Factorization Algorithms. Massachusetts Institute of Technology; 2014.

23. LuoY, SohaniAR, HochbergEP, SzolovitsP. Automatic lymphoma classification with sentence subgraph mining from pathology reports. J Am Med Inform Assoc. 2014;21:1–9. doi:10.1136/amiajnl-2013-002443.

24. BermanJJ. Concept-match medical data scrubbing: How pathology text can be used in research. Arch Pathol Lab Med. 2003;127:680–6. doi:10.1043/1543-2165(2003)127<680:CMDS>2.0.CO;2.

25. NassifH, WoodsR, BurnsideE, AyvaciM, ShavlikJ, PageD. Information Extraction for Clinical Data Mining: A Mammography Case Study. In: 2009 IEEE International Conference on Data Mining Workshops. 2009. p. 37–42. doi:10.1109/ICDMW.2009.63.

26. NeamatullahI, DouglassMM, LehmanLH, ReisnerA, VillarroelM, LongWJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak. 2008;8:32. doi:10.1186/1472-6947-8-32.

27. RossumGV. The Python Language Reference Manual, Release 3.5.1. Network Theory; 2003.

28. OwensM. The definitive guide to SQLite. 2006. https://www.google.com/books?hl=zh-TW&lr=&id=VsZ5bUh0XAkC&oi=fnd&pg=PR17&dq=sqlite&ots=u77Ngm44A6&sig=w46spF6bGTnbgPwuEbfrlvxUBhY. Accessed 19 May 2019.

29. FitzgibbonsPL, MurphyDA, HammondMEH, AllredDC, ValensteinPN. Recommendations for validating estrogen and progesterone receptor immunohistochemistry assays. Arch Pathol Lab Med. 2010;134:930–5. doi:10.1043/1543-2165-134.6.930.

30. WolffAC, HammondMEH, HicksDG, DowsettM, McShaneLM, AllisonKH, 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;31:3997–4013. doi:10.1200/JCO.2013.50.9984.

31. NguyenPL, TaghianAG, KatzMS, NiemierkoA, Abi RaadRF, BoonWL, 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. doi:10.1200/JCO.2007.14.4287.

32. Savci-HeijinkCD, HalfwerkH, HooijerGKJ, HorlingsHM, WesselingJ, van deVijverMJ. Retrospective analysis of metastatic behaviour of breast cancer subtypes. Breast Cancer Res Treat. 2015;150:547–57. doi:10.1007/s10549-015-3352-0.

33. HurJ, SchuylerAD, StatesDJ, FeldmanEL. SciMiner: Web-based literature mining tool for target identification and functional enrichment analysis. Bioinformatics. 2009;25:838–40.

34. YangH, SpasicI, KeaneJA, NenadicG. A Text Mining Approach to the Prediction of Disease Status from Clinical Discharge Summaries. J Am Med Informatics Assoc. 2009;16:596–600.

35. MarelliM, BentivogliL, BaroniM, BernardiR, MeniniS, ZamparelliR. SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment. In: Proceedings of the 8th International Workshop on Semantic Evaluation, {SemEval@COLING} 2014. 2014. p. 1–8. http://clic.cimec.unitn.it/marco/publications/marelli-etal-semeval14-task1.pdf%0Ahttp://aclweb.org/anthology/S/S14/S14-2001.pdf.

36. LuoY, XinY, HochbergE, JoshiR, UzunerO, SzolovitsP. Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. J Am Med Informatics Assoc. 2015;22:1009–19. doi:10.1093/jamia/ocv016.

Tables

Due to technical limitations, tables 1-11 are only available as a download in the supplemental files section.

Figures

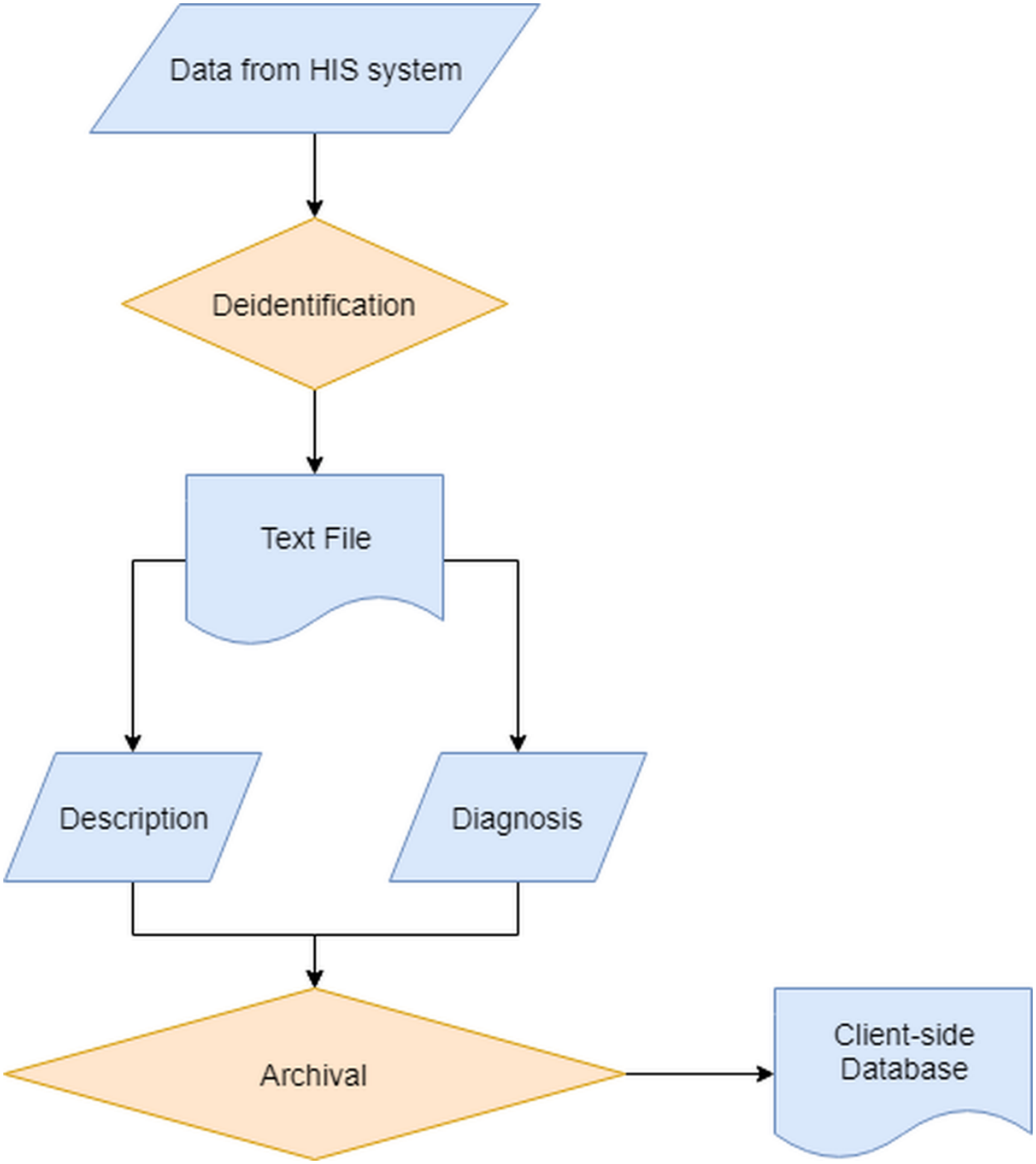

Figure 1

Data retrieval and preprocessing steps.

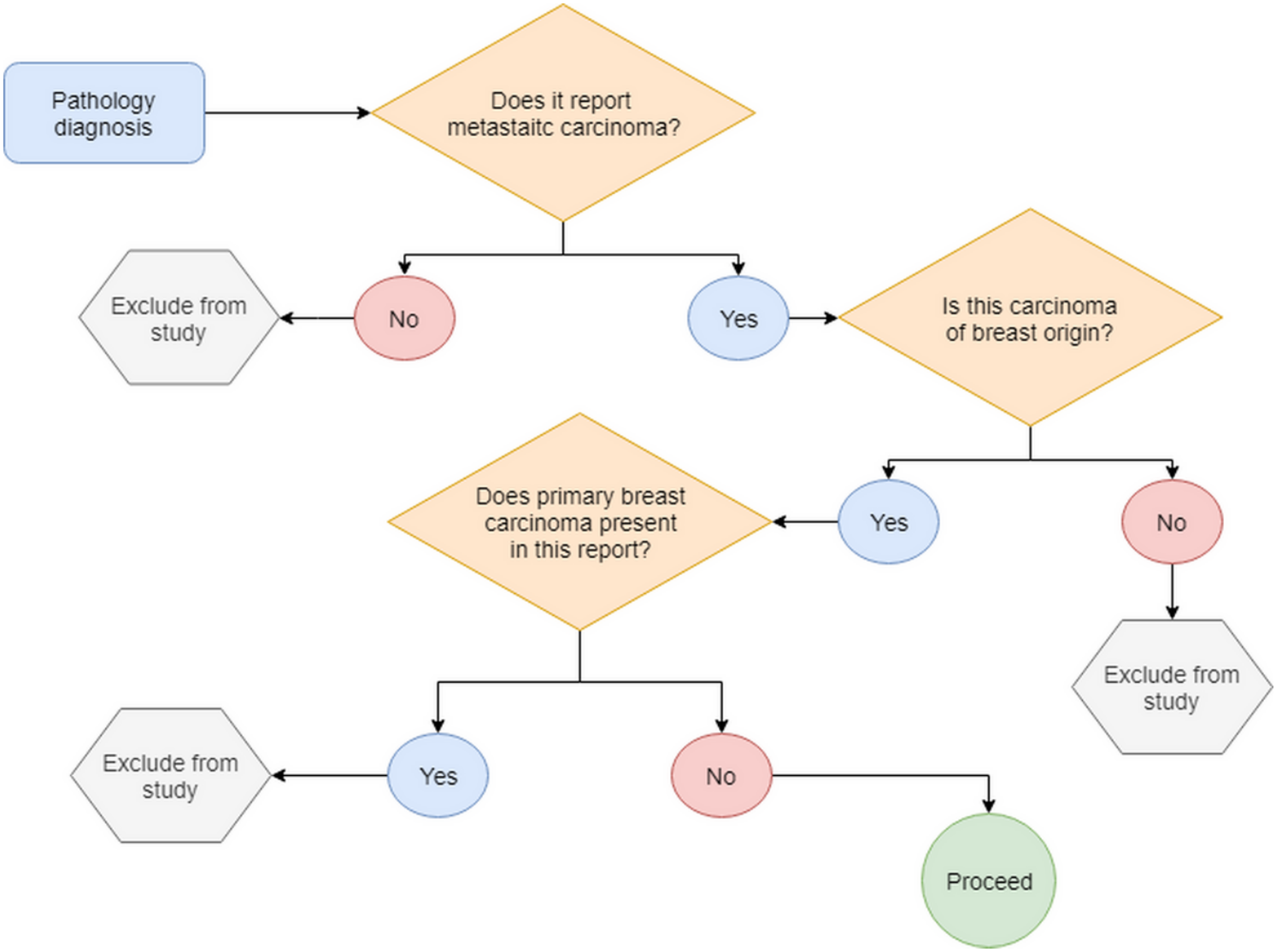

Figure 2

Protocol for searching metastatic breast cancer cases.

```
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 3

Reporting immunohistochemical study results 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 ER
(positive/negative, __%), PR(positive/negative, __%), Her-2(positive/equivocal/negative,
score 0/1+/2+/3+), and …… (Other immunohistochemical study).
n+1. Pathologic staging:
```

Figure 4

Reporting immunohistochemical study results as a solitary paragraph, with different studies separated by commas.

```
Microscopically, the breast shows invasive…… some description ….. Immunohistochemical
study/Ancillary study for invasive tumor cells shows ER(positive/negative, __%),
PR(positive/negative, __%), Her-2(positive/equivocal/negative, score 0/1+/2+/3+), and ……
(Other immunohistochemical study). Breast elsewhere shows …….
```

Figure 5

Reporting immunohistochemical study results as a sentence in the microscopic description.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table3.pdf
- OriginalResearchData.xlsx
- Table11.pdf
- Table2.pdf
- Table1.pdf
- Table8.pdf
- Table10.pdf
- Table5.pdf
- Table9.pdf
- Table4.pdf
- Table6.pdf
- Table7.pdf