Extraction of breast cancer biomarker status using natural language processing

Paul Dexter*

Indiana University School of Medicine,
Indianapolis IN, USA
and
Regenstrief Institute,
Indianapolis IN, USA
and
Eskenazi Health,
Indianapolis IN, USA
Email: prdexter@regenstrief.org
*Corresponding author

Jinghua He

Merck & Co., Inc.,
Kenilworth, NJ, USA
Email: Jinghua_he@merck.com

Jarod Baker

Regenstrief Institute,
Indianapolis IN, USA
Email: bakerjar@regenstrief.org

George Eckert

Indiana University School of Medicine,
Indianapolis IN, USA
Email: geckert@iu.edu

Abby Church

Indiana University Health,
Indianapolis IN, USA
Email: akchurch@iu.edu

Ning Jackie Zhang

Seton Hal University,
South Orange, NJ, USA
Email: ning.zhang@shu.edu

Copyright © 2019 Inderscience Enterprises Ltd.

Abstract: We employed natural language processing (NLP) algorithms to extract estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor 2 (HER2) receptor status for females with breast cancer using unstructured (free text) EMR data, and to determine the prevalence of triple negative breast cancer in the Indiana network for patient care (INPC) population. We identified female patients in INPC with a history of breast cancer over a ten year period who had at least five oncology notes or one related pathology document. Based on manual chart review, our NLP algorithms for extracting ER, PR, and HER2 receptor status performed well with sensitivity 87.5% to 92.6%, specificity 88.6% to 95.8%, positive predictive values (PPV) 82.4% to 99.0%, and negative predictive values (NPV) 85.2% to 97.7%. This study confirmed our primary hypothesis that NLP algorithms are effective in identifying important breast cancer biomarkers in patients with breast cancer using unstructured data.

Keywords: NLP algorithms; effective; breast cancer biomarkers; breast cancer.

Reference to this paper should be made as follows: Dexter, P., He, J., Baker, J., Eckert, G., Church, A. and Zhang, N.J. (2019) 'Extraction of breast cancer biomarker status using natural language processing', *Int. J. Computational Medicine and Healthcare*, Vol. 1, No. 1, pp.112–120.

Biographical notes: Paul Dexter has been a Research Scientist at Regenstrief Institute for more than 20 years with a focus on adapting Regenstrief's information systems for both clinical and research purposes and has conducted multiple trials related to computerise clinical reminder systems. He served as the Chief Medical Information Officer at Wishard/Eskenazi hospital for 12 years. He helped implement a robust research IT infrastructure that includes tools related to decision support, natural language processing, and the efficient performance of identified and de-identified data queries. He currently serves as the Co-Chair of NHGRI's Implementing Genomics in Practice (IGNITE) consortia.

Jinghua He obtained his PhD in Pharmaco Epidemiology from Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida. He also received his Master of Public Health degree from College of Public Health and Health Profession, University of Florida. He is currently a Principal Scientist at Center for Observation and Real-World Evidence, Merck Inc. & Co, Kenilworth, NJ, USA.

Jarod Baker obtained his MBA in Healthcare Management from Indiana Wesleyan University. He currently serves as the Program Manager of the Merck-Regenstrief Institute collaborative partnership which is currently in its eighth year. The partnership seeks to improve the health of patients through data analytics, health care innovation, education and research.

George Ecker is the Biostatistician Supervisor in the Department of Biostatistics at the Indiana University School of Medicine. He has more than 25 years of experience after graduating from The Ohio State University with a Master of Applied Statistics degree. He has extensive collaborative work with researchers in dentistry, hypertension, and health services research

Abby Church is the Project Manager in the Department of Biomedical Informatics at Regenstrief Institute during the study. She received her MPH from the IU School of Medicine in 2012.

Ning Jackie Zhang started his faculty career in the Doctoral Program in Public Affairs in the University of Central Florida, Orlando, FL after graduating from the Department of Health Administration with a PhD in 2003. He received tenured in 2009 and continued to teach as an Associate Professor till 2014 when he moved to Seton Hall University, South Orange, NJ. At SHU, he was promoted to a Full Professor in 2014 and the Associated Dean for Academic Affairs in 2018. He became the Editor-in-Chief of *the International Journal of Computational Medicine and Healthcare* in 2018.

1 Introduction

The National Cancer Institute (2018) defines a biomarker as "a biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease." Various biomarkers are important in diagnosis, prognosis, and predicting the response to certain therapies. They are particularly important to drug discovery, providing information on drug efficacy and safety (Fan et al., 2006; Voduc et al., 2010). Because biomarkers can predict drug efficacy more quickly than conventional clinical endpoints, they can accelerate product development (Foukakis and Bergh, 2016).

Biomarkers such as estrogen receptors (ER), progesterone receptors (PRs), and human epidermal growth factor 2 (HER2) receptors have been especially important in the case of breast cancer (Carey et al., 2014). Assessment of ER, PR, and HER2 receptor status is a routine component in the workup of patients with breast cancer, and essential to determining the need and type of adjuvant therapy (Howlader et al., 2014). Given their importance to prognosis, biomarkers are also often used in clinical trials to stratify patients into randomised groups when testing novel treatments (Koboldt et al., 2012). Insights regarding these biomarkers have led to new therapies for breast cancer, such as the selective ER modulators (SERMs) and selective ER downregulators (SERDs) (Maximov et al., 2013).

Breast cancer is the leading cause of cancer death in women worldwide. In the USA, breast cancer is the most common female cancer and the second most common cause of cancer death in women (American Cancer Society, 2018). A pharmaceutical company in USA has conducted a number of trials searching for more effective breast cancer treatments. An ongoing clinical research study is evaluating the investigational study drug, pembrolizumab (MK-3475) for treating metastatic 'triple-negative' breast cancer – or breast cancer that lacks all three biomarkers, the ER, PR, and HER2 receptor biomarkers.

The widespread availability of large electronic medical records (EMRs) provides new possibilities for efficiently identifying patient with particular characteristics, such as women with a history of 'triple negative' breast cancer. However, others have noted that such cancer characteristics are poorly captured in structured clinical data, and instead commonly found only in free text clinical documents such as pathology reports (Liao et al., 2015). Culling the desired data from these free text sources can be challenging and requires natural language processing (NLP).

To such ends, the Regenstrief Institute created an advanced text-mining and NLP platform for the development and validation of methods for extracting critical

information from free-text documents. In this protocol, we describe a study leveraging this platform and utilising unstructured free text data to create an NLP phenotype for breast cancer biomarkers.

The primary objectives of the study are as follows:

1 to determine the performance of NLP algorithms for extracting ER, PR, and HER2 receptor status for patients with breast cancer using unstructured (free text) EMR data

2 to determine the prevalence of triple negative breast cancer in the INPC population

The primary hypothesis is that NLP algorithms data will effectively identify the ER, PR, and HER2 receptor status for patients with breast cancer using unstructured EMR data.

2 Study design

Per the protocol, the Indiana network for patient care (INPC) was queried to identify a cohort of breast cancer patients. Breast cancer was identified via ICD9 and ICD10 code between 1 January 2006 and 31 December 2015. Furthermore, patients were required to meet a secondary criteria of having pathology or oncology text documents available to be included in the cohort. This requirement was at least one pathology document or at least five oncology documents in the period of 90 days before or up to one year following the first breast cancer diagnosis. Finally, only females over eighteen years of age at the time of the first breast cancer diagnosis were included in the cohort.

These criteria generated a cohort of 13,310 patients, whose text reports (all available text reports, not just pathology/oncology) were then imported to nDepth, the Regenstrief NLP platform, for review. nDepth was then used to develop an algorithm to identify the biomarker status of the patient's breast cancer. Using regular expressions and the nDepth platform, a set of heuristics were developed to identify and extract specific sentences from the text reports containing relevant biomarker information. These relevant sentences were then classified according to the order of certain keywords contained therein. The most common sentence types were then given a score of –1, 0, or 1 for each of the three biomarkers. For example, a sentence like 'the patient has ER/PR positive breast cancer' would be classified as ER-PR-POS, and then assigned a score of ER = 1, PR = 1, HER2 = 0. Since a single patient often has dozens of relevant documents, each potentially containing several relevant sentences, a single patient could have hundreds of relevant sentences, each with their own score. Each patient was given an overall biomarker score determined by taking the summation of scores over all associated sentences. If the score was positive, then the patient is interpreted as having a positive biomarker, if negative it is interpreted as a negative biomarker, and if zero, it is interpreted as being unavailable in the data or otherwise unknown.

To validate the performance of the algorithm, 200 patients were each manually reviewed by two chart reviewers. In the case of chart reviewer disagreement, the patient was manually reviewed by the principal investigator and those results used in the analysis. Due to technical issues with the nDepth platform, only 194 patient reviews were usable. One other patient was excluded due to multiple primary tumours with differing biomarker profiles (something for which the algorithm cannot compensate).

3 Results

Performance of the NLP algorithms for extracting ER, PR, and HER2 were evaluated via chart review. Two chart reviewers cross-validated the results. Table 1 shows the cross validation agreement for negative, positive and indeterminate results. Agreement between chart reviewers appears to be high: kappa scores are 0.90 for ER, 0.90 for PR, and 0.93 for HER2, respectively.

Table 1 Cross-validation of NLP algorithms

		ER		
		Reviewer 2		
		Negative	*Indeterminate*	*Positive*
Reviewer 1	Negative	20	0	2
	Indeterminate	2	58	2
	Positive	2	3	105
		PR		
		Reviewer 2		
		Negative	*Indeterminate*	*Positive*
Reviewer 1	Negative	29	1	0
	Indeterminate	3	62	2
	Positive	3	3	91
		HER2		
		Reviewer 2		
		Negative	*Indeterminate*	*Positive*
Reviewer 1	Negative	87	2	0
	Indeterminate	4	82	0
	Positive	2	0	17

Performance statistics for the NLP algorithms demonstrates good performance of the NLP algorithm for all evaluation parameters. Among them, the sensitivity ranges from 87.5% to 92.6%; the specificity ranges from 88.6% to 95.8%; the positive predictive values (PPV) range from 82.4% to 99.0% and the negative predictive values (NPV) range from 85.2% to 97.7%.

Table 2 Performance of NLP algorithms

		Reviewer			
		Sensitivity	*Specificity*	*PPV*	*NPV*
NLP	ER	89.91	95.83	98.99	85.19
	PR	92.55	88.57	97.75	93.94
	HER2	87.50	91.40	82.35	97.70

Details of the agreement between the NLP algorithms and chart reviews are shown in Table 3. The agreements between the NLP and chart reviewers appear to be high. The

kappa agreement values for ER, PR and HER2 are 86.7%, 88.3% and 89.2%, respectively.

Table 3 Cross-validation of NLP algorithms with all data

		ER		
		Reviewer		
		Negative	*Indeterminate*	*Positive*
NLP	Negative	23	2	2
	Indeterminate	1	57	9
	Positive	0	1	98
		PR		
		Reviewer		
		Negative	*Indeterminate*	*Positive*
NLP	Negative	31	1	1
	Indeterminate	4	61	6
	Positive	0	2	87
		HER2		
		Reviewer		
		Negative	*Indeterminate*	*Positive*
NLP	Negative	85	0	2
	Indeterminate	7	82	0
	Positive	1	2	14

After validation of the NLP algorithms, the algorithm was applied to all available data. Prevalence of each combination of the three biomarkers was estimated. All patients were female. The minimum age was 18 and the maximum was greater than 90, with mean (SD) = 60 (14).

Biomarker combination frequencies, excluding patients who had at least one indeterminate biomarker, are shown in Table 4.

Table 4 Numbers of biomarker combinations

ER	*PR*	*HER2*	*N*	*%*
–	–	–	1,190	17.20
–	–	+	457	6.60
–	+	–	48	0.69
–	+	+	14	0.20
+	–	–	559	8.08
+	–	+	185	2.67
+	+	–	3,915	56.58
+	+	+	552	7.98

Table 5 presents frequencies of all biomarker combinations, including all patients (includes indeterminate biomarkers, ‘i’).

Table 5 Number of biomarker combinations for all patients

	ER	*PR*	*HER2*	*N*	*%*
-1-1-1	–	–	–	1,190	8.94
-1-10	–	–	i	276	2.07
-1-11	–	–	+	457	3.43
-10-1	–	i	–	13	0.10
-100	–	i	i	59	0.44
-101	–	i	+	12	0.09
-11-1	–	+	–	48	0.36
-110	–	+	i	5	0.04
-111	–	+	+	14	0.11
0-1-1	i	–	–	24	0.18
0-10	i	–	i	11	0.08
0-11	i	–	+	9	0.07
00-1	i	i	–	159	1.19
000	i	i	i	4,243	31.88
001	i	i	+	75	0.56
01-1	i	+	–	17	0.13
010	i	+	i	16	0.12
011	i	+	+	3	0.02
1-1-1	+	–	–	559	4.20
1-10	+	–	i	139	1.04
1-11	+	–	+	185	1.39
10-1	+	i	–	118	0.89
100	+	i	i	180	1.35
101	+	i	+	41	0.31
11-1	+	+	–	3,915	29.41
110	+	+	i	990	7.44
111	+	+	+	552	4.15

Table 6 shows frequencies of each biomarker, excluding patients who were indeterminate for the specific biomarker.

Table 6 Frequency distribution of biomarkers

		N	*%*
ER	–	1,709	24.70
	+	5,211	75.30
PR	–	2,391	34.55
	+	4,529	65.45
HER2	–	5,712	82.54
	+	1,208	17.46

Table 7 presents frequencies of each biomarker, including all patients (includes indeterminate biomarkers, 'i').

Table 7 Frequency distribution of biomarkers including all patients

		N	*%*
ER	–	2,074	15.58
	i	4,557	34.24
	+	6,679	50.18
PR	–	2,850	21.41
	i	4,900	36.81
	+	5,560	41.77
HER2	–	6,043	45.40
	i	5,919	44.47
	+	1,348	10.13

4 Discussions

Employing an INPC cohort of adult female patients with a history of breast cancer, we confirmed that the performance of our NLP algorithms for extracting ER, PR, and HER2 receptor status was good with sensitivity 87.5% to 92.6%, specificity 88.6% to 95.8%, PPV 82.4% to 99.0%, NPV 85.2% to 97.7%.This study further supports our primary hypothesis that NLP algorithms could be effective in identifying important breast cancer biomarkers in patients with breast cancer using unstructured data.

We also determined biomarker combination frequencies for the entire adult female breast cancer cohort. An important subset of breast cancer patients who are eligible for the investigational study drug pembrolizumab (MK-3475) are those with 'triple-negative' metastatic disease. For patients with evidence of biomarker information for all three receptors in unstructured data, we found an incidence of 17.2%. It is notable that this compares very favourably to previously described rates for triple negative breast cancer of 15–20% (Kumar and Aggarwal, 2016).

It is concluded that the NLP algorithms using regular expressions on unstructured EMR data would be an effective method of identifying adult female breast cancer patients for chemotherapy trials that rely on ER, PR, and HER2 receptor status criteria. Future studies should further examine the performance and feasibility to use NLP algorithms in identifying patients for clinical trials based on unstructured EMR data across various cancers and broader populations.

Acknowledgements

The authors wish to thank Joel Martin, Jon Duke, Fangqian Ouyang and Kristina Knapp for their invaluable contributions to the current article.

References

American Cancer Society (2018) *Cancer Facts & Figures 2018*, American Cancer Society, Atlanta.

Carey, L.A., Cheang, M.C.U. and Perou, C.M. (2014) 'Chapter 29: genomics, prognosis, and therapeutic interventions', in Harris, J.R., Lippman, M.E., Morrow, M. and Osborne, CK. (Eds.): *Diseases of the Breast*, 5th ed., Wolters Kluwer Health Adis., Lippincott Williams & Wilkins.

Fan, C., Oh, D.S., Wessels, L., Weigelt, B., Nuyten, D.S., Nobel, A.B. and Perou, C.M. (2006) 'Concordance among gene-expression-based predictors for breast cancer', *New England Journal of Medicine*, Vol. 355, No. 6, pp.560–569.

Foukakis, T. and Bergh, J. (2016) in Dizon (Ed.): *Prognostic and Predictive Factors in Early, Non-Metastatic Breast Cancer*.

Howlader, N., Altekruse, S.F., Li, C.I., Chen, V.W., Clarke, C.A., Ries, L.A. and Cronin, K.A. (2014) 'US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status', *JNCI: Journal of the National Cancer Institute*, Vol. 106, No. 5.

Koboldt, D.C., Fulton, R.S., McLellan, M.D., Schmidt, H., Kalicki-Veizer, J., McMichael, J.F., Fulton, L.L., Dooling, D.J., Ding, L., Mardis, E.R. and Wilson, R.K. (2012) 'Comprehensive molecular portraits of human breast tumours', *Nature*, Vol. 490, No. 7418, pp.61–70.

Kumar, P. and Aggarwal, R. (2016) 'An overview of triple-negative breast cancer', *Arch Gynecol Obstet*., Vol. 293, pp.247–69, DOI: 10.1007/s00404-015-3859-y.

Liao, K.P., Cai, T., Savova, G.K., Murphy, S.N., Karlson, E.W., Ananthakrishnan, A.N., Gainer, V.S., Shaw, S.Y., Xia, Z., Szolovits, P., Churchill, S. and Kohane, I. (2015) 'Development of phenotype algorithms using electronic medical records and incorporating natural language processing', *BMJ*, Vol. 350, PMCID: PMC4707569, DOI: 10.1136/bmj.h1885.

Maximov, P.Y., Lee, T.M. and Jordan, V.C. (2013) 'The discovery and development of selective estrogen receptor modulators (SERMs) for clinical practice', *Curr. Clin. Pharmacol.*, Vol. 8, No. 2, pp.135–155, PMCID: PMC3624793, DOI: 10.2174/1574884711308020006.

National Cancer Institute (2018) *NCI Dictionary of Cancer Terms* [online] https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker.

Voduc, K.D., Cheang, M.C., Tyldesley, S., Gelmon, K., Nielsen, T.O. and Kennecke, H. (2010) 'Breast cancer subtypes and the risk of local and regional relapse', *Journal of Clinical Oncology*, Vol. 28, No. 10, pp.1684–1691.