ASSESSING NLP ACCURACY: FOCUS ON ANATOMIC PATHOLOGY



Laura A. Evans¹, Jack W. London, PhD², Matvey B. Palchuk, MD, MS, FAMIA^{1,3} ¹TriNetX, Cambridge, MA; ²Sidney Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, PA; ³Harvard Medical School, Boston, MA



INTRODUCTION

- Manual annotation of documents is the gold standard for NLP quality assessment^{1,2}
- We introduce a quality assurance method targeted to anatomic pathology reports that involves comparing extracted data to Tumor Registry records
- We propose to use a confusion matrix to interpret the results of this comparison

NLP

Row Labels	C00	C01 C02 C04 C05 C06 C0	07 C08 C09 C10 C11 C12 C13	.3 C14 C15 C16 C17 C18 C19 C20 C21 C22 C23 C24 C25 C26 C30 C31 C32	C33 C34 C37 C38 C40 C41 C42	C44 C47 C48 C49 C50	C51 C52 C53 C54 C55 C56 C57 C60 C61 C62 C63 C64 C65	C66 C67 C68 C69 C70 C71 C72 C73 C74 C76 C77 Grand Total
C00		9			1	1		1
C01		174 39 1 1	5 1	4	1	2		5
C02		8 134 1 1 1	4	1 1	1	1		1 1 3
C03		1 1 2			19	1		

C03						24
C04			2 5 1 1		1	59
C05	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$					48
C06			5 19 1 2		3 2	63
C07			3 2 2 1		1 5	160
C08					4	27
C09	3 1 2 234 1		1 1 3 1		11	267
C10		1				33
C11		1			2 1 4	62
C12		5			Ζ	19
C13		3	1 1			19
C14	3	6		1	1 2	3
C15		16 3				470
C17		15 2		1 1		158
C18		2 8		2 2 1 7 5 1 1	2 4 14	1436
C19	2 71 59 26 6		4 1		3	172
C20	5 153 32 337 12 7	1	3 1 2 2 1 11	1 1 3	4 1 1 1 7	587
C21	1 1 10 28 2		1 2 1	1	2 2	51
C22	1 569 3 17	1 1	3 1 5 7 5 1 1 7	1 1 1	1 2 1 10	639
C23	1 1 15 40 8		1 1 1	1	2 1	72
C24	34 17 <u>36</u> 4 75	60 1	1 1 2		2	233
C25	1 73 23 1 <u>166</u> 4 38	835	12 1 10 2 14 11	1 3	1 1 3 21	1221
C26	4 2 24 2	1	4 1 2	1	1 1 1	44
C30	1	18 3	1 6 2 1		1 6 2	41
C31	3 1 1	3 13	1 5 10 2 1		3	44
C32	2 1 1 1 4 1 3 5 2 3 1	300 1	1 3 4 1 6		4 1 10	355
C33	1 2 1 2 5 1 1 50 1	3				4
C34	1 2 1 2 5 1 1 50 1	1 1 5 1	<u>1825</u> <u>52</u> 4 <u>36</u> 12 2 6	5 1 16 1	1 1 18 2 9 21 186	2269
C37					1	9
C38	1					58
C40	1 1		2 50 18 1 1		1 6	84
(42			4 5 24 1524 5	5 1 2 1	$1 \qquad 1 \qquad 1 \qquad 4 \qquad 18$	1606
C44	9 16 2 1 1 1 6 1 2 15	1 1	20 4 12 21 501 1 1 17	6 1 1 4	1 7 14 3 1 267 247	1185
C47		1	3	1	5 1	11
C48	2 4		1 6 1 13 1	1 1 2 6 1	1 2	42
C49	1 1 1 1 3	1 1 1	10 1 7 7 5 20 1 1 2	1 1 1 1	2 1 139 4	216
C50	1 2 1 7	1 1 1	21 3 5 13 10 41 1 4 49	999 1 8 1 1 1 1	3 1 3 55 76	5262
C51	1 1		2 1 4 2	67 3 4 1	1 1 4 3	96
C52			2	1 17 1	2 1 1	25
C53	2		2 1 1 2	2 1 10 134 4 68 2	3 3 4	241
C54	2 1 1	1 1	3 1 4	3 6 19 99 468 3 1 1	1 2	617
C55			4 1 1		1	23
C56	3 5 1 5		1 1 17 6	4 2 1 8 139 14	1 7 4	219
C57	1		2		1 1	16
C58			1	2		3
C60	1 3 1 1 2 4			15 4218 2 2	18 5 2 1 1 1 13	15
C61 C62		1	4 <u>3 9 15</u> / 1			4315
C63			5 I I			1
000				1		1



Figure 1. Confusion Matrix. Count of patients with ICD-O site codes from Tumor Registry on the vertical axis and from NLP on the horizontal. Darker color indicates larger number of patients.

METHODS

- Anatomic sites (coded to ICD-O topography) were extracted from pathology reports using Information Discovery text analytics platform (Averbis GmbH, Germany)
- We focused on a cohort of patients who have site data in results of the NLP pipeline as well as in the Tumor Registry
- We included patients who have one tumor in the Tumor Registry (to reduce the number of false positives) and excluded patients who

RESULTS

The major diagonal axis indicates patients whose site data is in agreement between Tumor Registry and NLP, and we can presume that NLP has behaved correctly. In this analysis, 80.8% of patients in this dataset are found along the diagonal.

We noted prominent vertical lines at sites that are common locations for solid tumor metastasis. A vertical line means that the NLP is identifying multiple sites in disagreement with the Tumor Registry record. Manual review of sample documents confirmed that in majority of these cases pathology reports were describing a metastatic site instead of the primary tumor location. Having discovered this error, we can introduce the necessary corrections into the NLP pipeline.

have non-malignant tumors in NLP data

• We created a table with Tumor Registry site codes on the vertical axis and NLP on the horizontal, with the cells shaded according to the number of patients falling into the cell (see Figure 1)

CONCLUSION

We propose to utilize the confusion matrix to review the correctness of NLP on a much larger corpus than would be feasible if we relied on traditional manual annotation techniques. This approach is limited to cases where an independent source of similar data is available. Additionally, this visualization makes it easier to identify trends or areas of concern that can subsequently be examined in greater detail.

REFERENCES

¹ Savova GK, Tseytlin E, Finan S, et al. DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. Cancer Res. 2017 Nov 1;77(21):e115–8. ² Jones B, South B, Shao Y, et al. Development and Validation of a Natural Language Processing Tool to Identify Patients Treated for Pneumonia across VA Emergency Departments. Appl Clin Inform. 2018 Jan 3;09(01):122-8.

AMIA 2018 ANNUAL SYMPOSIUM | SAN FRANCISCO, CA