ASSESSING NLP ACCURACY: FOCUS ON ANATOMIC PATHOLOGY

Laura A. Evans[1], Jack W. London, PhD[2], Matvey B. Palchuk, MD, MS, FAMIA[1,3]
[1]TriNetX, Cambridge, MA; [2]Sidney Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, PA; [3]Harvard Medical School, Boston, MA

TriNetX

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

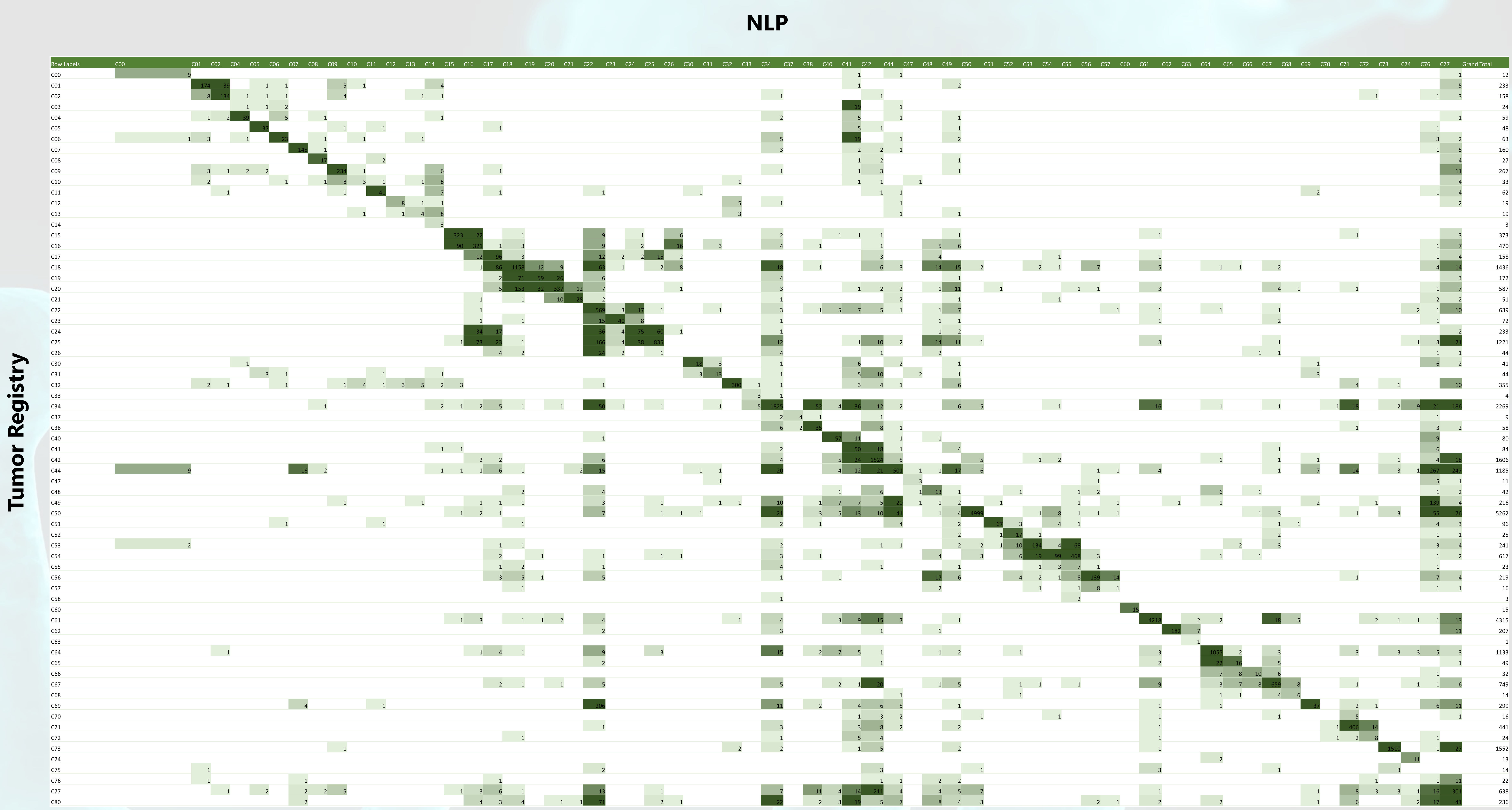

Figure 1. Confusion Matrix. Count of patients with ICD-O site codes from Tumor Registry on the vertical axis and from NLP on the horizontal. Darker color indicates larger number of patients.

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 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)

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.

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.