

Using LLMs for Cancer Diagnosis from Clinical Notes

ISAAC HANDS, University of Kentucky, USA

1 INTRODUCTION

Cancer is a devastating illness, the second leading cause of death in the U.S [1]. One of the best ways to improve survival of this disease is to diagnose it as early as possible[3]. The way cancer is diagnosed in 90% of cancer cases[2] is through a biopsy of tissue, which is then observed under a microscope by a pathologist, leading to a narrative description of the cancer cells in a pathology report. These reports consist of clinical narrative text dictated by a pathologist, stored in a patient's electronic medical record. The characterization of a cancer diagnosis occurs when a physician interprets a pathology report for communication to the patient and determines a treatment plan, along with data abstraction of the codes associated with the cancer diagnosis by specialized medical coders. This entire process of determining whether a patient has cancer and what type of cancer it is, relies on interpretation of narrative text. In fact, narrative unstructured text makes up about 80% of all healthcare data[4], making it difficult to analyze systematically.

Processing, analyzing, and interpreting unstructured text can be time consuming and cause delays in life-saving treatment for cancer patients, but may be an area where machine learning classifiers could provide efficiencies. In order to build classifiers, however, labeled training data needs to be readily available. When the data required is clinical text containing patient identifiers, it is often very difficult to access due to privacy concerns and lack of access to medical record systems. Moreover, clinical text uses a specialized vocabulary, which may prove problematic for natural language processing methods that have been developed around standard English vocabulary. In this report, we attempt to address

2 METHODS

2.1 Data Description

2.2 Data Processing

2.3 Models Utilized

3 RESULTS

4 DISCUSSION

REFERENCES

- [1] CDC1 2023. An Update on Cancer Deaths in the United States. Retrieved April 29, 2023 from <https://www.cdc.gov/cancer/dcpc/research/update-on-cancer-deaths/index.htm>
- [2] CDC2 2023. How CDC Speeds Up Cancer Data Reporting. Retrieved April 29, 2023 from <https://www.cdc.gov/cancer/dcpc/research/articles/cdc-cancer-reporting.htm>
- [3] CRUK 2023. Why is early cancer diagnosis important? Retrieved April 29, 2023 from <https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important>
- [4] Hyoun-Joong Kong. 2019. Managing Unstructured Big Data in Healthcare System. *Health Inform Res* 25, 1 (Jan 2019), 1–2. <https://doi.org/10.4258/hir.2019.25.1.1>