

NATURAL LANGUAGE PROCESSING OF PATHOLOGY REPORTS IDENTIFIES SOFT TISSUE SARCOMAS MISSED BY INTERNATIONAL CLASSIFICATION OF DISEASES CODES IN THE VA NATIONAL DATABASE

Sarah E. Lindsay, Cecelia Jane Madison, Yee-Cheen Doung, James B Hayden, Kenneth Robert Gundle

INTRODUCTION:

Soft tissue sarcoma (STS) comprises less than 1% of all adult malignancies, and includes a diverse group of more than fifty distinct diagnoses. In databases used to identify cancers, STS may be misclassified as a different form of cancer. The rarity of these tumors would make large-scale databases particularly useful for the investigation of clinical outcomes, but relying on International Classification of Diseases (ICD) codes may undercount sarcoma diagnoses.¹ The national VA database provides a unique opportunity for STS investigation, due to the availability of all clinical results and reports.

The purpose of this study was to utilize natural language processing (NLP) to analyze pathology reports to identify patients with STS. To test the feasibility of this methodology, we trialed application of NLP to pathology reports to identify myxofibrosarcoma and dedifferentiated liposarcoma. Our hypothesis was that utilizing NLP to analyze pathology reports would identify patients who lack International Classification of Diseases (ICD) codes for STS. Further, we hypothesized that that incidence of patients lacking ICD codes for STS would increase in the ICD-10 era. In addition to assessing this novel method of cohort creation, we aimed to determine myxofibrosarcoma and dedifferentiated liposarcoma overall survival in the VA population.

METHODS:

Through the national VA corporate data warehouse, all surgical pathology notes were identified from 2003 through 2022. These reports were retrospectively searched for the terms, "myxofibrosarcoma" and "dedifferentiated liposarcoma." Patient gender, age at diagnosis, location, and oncologic outcomes were abstracted from the data warehouse to assess survival outcomes and to generate Kaplan Meier curves. A Cox proportional hazards regression analysis was conducted relative hazard ratios. Diagnosis codes were abstracted and the difference in appropriate ICD-9 and ICD-10 coding was compared by Chi-square test.

RESULTS:

In searching 10,684,177 pathology reports, we identified 379 that included the diagnosis "myxofibrosarcoma" and 583 with "dedifferentiated liposarcoma." Overall, the population was predominately male in both groups (936 of 962, $p=0.76$). Patients with myxofibrosarcoma had a mean age of 69 (range 30-97, SD 11.8), slightly older than those with dedifferentiated liposarcoma who had a mean age of 67 (range 29-94, SD 11.4, $p=0.004$). Myxofibrosarcoma patients were significantly more likely to be referred to both medical oncology ($p=0.001$) and radiation oncology ($p=0.0006$) than dedifferentiated liposarcoma patients were.

In patients treated before the VA adopted ICD-10 in October 2010, 82/188 (44%) of myxofibrosarcoma patients and 51/201 (25%) dedifferentiated liposarcoma patients had ICD-9 codes for STS. In the ICD-10 era, both myxofibrosarcoma patients (126/180, 70%, $p<0.00001$) and dedifferentiated liposarcoma patients (165/370, 46%, $p<0.0001$) were more likely to have STS ICD codes compared to the ICD-9 era. Overall, for myxofibrosarcoma, 264 had an STS diagnosis and 155 (30%) lacked an STS ID code. In the dedifferentiated liposarcoma group, 265 had an ICD for soft tissue sarcoma, and 318 (55%) had no ICD for STS.

In a Cox proportional hazards regression analysis, myxofibrosarcoma had a hazard ratio of 0.65 (0.50-0.85) for overall survival compared to dedifferentiated liposarcoma ($p=0.001$). A Kaplan Meier curve confirmed worse 5-year survival probability for dedifferentiated liposarcoma patients compared to myxofibrosarcoma patients ($p=0.0026$).

DISCUSSION AND CONCLUSION:

NLP is a powerful tool that can be utilized for cohort creation in database research. In the national VA data warehouse, NLP of pathology reports provided a sensitive way to identify patients with STS and captured patients that would have been missed had ICD-10 codes been used alone. This cohort confirms decreased overall survival at 5 years of dedifferentiated liposarcoma compared to myxofibrosarcoma.