In Pathology TNM codes are regularly used to encode cancer stages of cancer tumors. It is defined during microscopic examinations of tumor samples by pathologists. Pathological records are imported to clinical information systems to enable data search and analysis.

Unfortunately, this information is not standardized due to different styles and structures of pathological reports, e.g. some entries contain more than one TNM code without clear delimiters, making automatic separation difficult. Thus data must be cleaned and normalized before it can be imported to outside analytical systems.

We tested our program on a set of 1600 pathological records. Each of the records could contain from 1 to 5 TNMs. The TNMs were automatically separated and their properties were parsed to be represented in an object oriented manner. A table with results in a printed form was produced and directed to a pathologist for assessment.

In most cases the pathologist confirmed correctness of the extracted information. Not covered entries were immoderately erroneous or ambiguous and hard for the pathologist to derive encoded information. This fact states a promising performance of the applied approach. We also prepared TNMs to import to the analytical system to provide a search on samples using attributes of TNMs as search criterion in ontology based data warehouse I2B2.

DISCLOSURE STATEMENT OF FINANCIAL INTEREST: I. Shumeiko DO NOT have a financial interest or affiliation with one or more organizations that could be perceived as a real or apparent conflict of interest in the context of the subject of this presentation.