Data Quality Audit on Melanoma Tumor Depth with Rule-Based Natural Language Processing

Glenn Abastillas1, Clara Lam1, Valentina Petković2, Peggy Adam1, Linda Coyle1, Jennifer Stevens1, Xiao-Cheng Wu3, Mei-Chin Hsieh3, Serban Negoița2
Information Management Systems Inc.1, National Cancer Institute2, Louisiana Tumor Registry3

BACKGROUND
- Breslow’s tumor depth is the most important determinant of prognosis for localized melanoma.
- Greatest measured thickness
- Tumor depth value
- Develop and test an algorithm to

METHODS
- There are three main areas of activity that inform the methods implemented to gather and analyze the data using natural language processing (NLP):
  1. Algorithm Development
  2. Testing
  3. Analysis

RESULTS TESTING
Once the algorithm is complete, it is run on all the data and an analytic data set is created for registry level testing. At this stage, a consolidation algorithm is developed to consolidate the machine values, which are compared to reconstituted values by the registry.

ANALYSIS
- Provide description of errors in melanoma tumor depth values among cases with supporting source documentation.
- Describe errors by demographic factors among the following cases:
  1. Cases with supporting text documentation in abstracts.
  2. Cases with supporting electronic pathology (e-path) reports.
  3. Cases with both abstract and e-path test data.
- Compare pre-error and post-error data for:
  1. Depth category levels.
  2. Stage.

DEPLOYMENT
- In the deployment phase, two products are made available to the registry (Error Correction) and extramural organizations (Algorithm Release).

Algorithm Release
- The query file written in I2E and algorithm are released for use by the NCI, IMS, registry, and extramural researchers and organizations.

IMPLICATIONS
- Provide registries with a set of flagged cases with high likelihood of inaccurate depth measurement values for review instead of having CTAs manually extract measurement values.
- Provide registries with a method for automatic error correction.
- Disseminate algorithm logic and query algorithm files to cancer surveillance research community.
- Provide evidence-based input for training materials for registrars.