Automated Extraction of Free-Text from Pathology Reports

Anne-Marie Currie, PhD[1]*, Travis Fricke[1], Agnes Gawne[2], Ric Johnston[3], John Liu, PhD[1], and Barbara Stein, PhD[3]

[1]FineTooth, Austin, TX

[2]University of Washington, Seattle, WA

[3]Fred Hutchinson Cancer Research Center, Seattle, WA

Abstract. Manually populating a cancer registry from free-text pathology reports is labor intensive and costly. This poster describes a method of automated text extraction to improve the efficiency of this process and reduce cost. FineTooth, a software company, provides an automated service to the Fred Hutchinson Cancer Research Center (FHCRC) to help populate their breast and prostate cancer clinical research database by electronically abstracting over 80 data fields from pathology text reports.

Introduction. Pathology reports have been the subject of previous research in the area of automatic data retrieval and extraction. Xu *et al*. (2004) applied natural language processing (NLP) to surgical pathology reports to automatically abstract 13 types of findings. Our project, although employing different automation, expands the process to include more complex and comprehensive abstraction. In addition, we have supplied a secure web-based viewer to facilitate the review of abstracted data for QA/QC.

Methods. Using a targeted NLP process, document, patient and specimen level information were extracted from 5826 surgical breast and 2838 surgical prostate cancer pathology reports for FHCRC. Requirements for data abstraction were cooperatively developed between FineTooth and FHCRC and include specific rules and language patterns. Iterative testing of the abstraction was conducted as rules were added and implemented. The abstraction results were then made available to authorized FHCRC staff in a data file and through a web-based viewer. The viewer visually organizes the abstracted data according to document and specimen level information and includes a link to the source text document for easy comparison. The core focus of clinical data abstraction was the “Final Diagnosis” sections of the pathology reports, as well as specific marker data. To evaluate the effectiveness of the abstraction approach, a random sample of documents was selected and data elements that were automatically parsed were compared with the same data elements that had been manually reviewed.

Results and Discussion. Similar to previous research[1], we found that narrative data extraction is more challenging than data presented in table format. Most of the data in surgical pathology reports received from FHCRC are in narrative format. This required the generation of multiple controlled dictionaries for the specific contexts desired. However, the two biggest challenges faced were the correct abstraction of breast lymph node data and correctly associating individual specimen labels with the appropriate contexts when language patterns were non-standard. These challenges, which would be challenges to any automated extraction process, were also mentioned by Xu *et al*. (2004). We addressed these challenges by establishing rules to alert FHCRC where inconsistencies in the data were noted (e.g., out of sequence specimen labels, multiple specimen labels grouped together within a diagnosis section). We extracted the data from these fields and flagged it to allow authorized FHCRC staff to review it in a focused manner. Based on a random sampling of the documents, results were evaluated by domain experts who confirmed 90%-95% accuracy was obtained for most fields.

Conclusion. Document, patient and specimen level information were successfully extracted automatically from breast and prostate surgical cancer pathology reports using a targeted NLP process. By providing specific data elements from these reports, FHCRC is now able to populate their clinical research database more quickly and comprehensively than previous manual methods allowed. Another advantage of this automated method is the increase in specificity and consistency of the data extracted. Automated text extraction also offers cost benefits. The results of this project build on previous research and contribute to the discussion of the role and efficacy of automated pathology abstraction to improving human health and welfare.

REFERENCES

[1] Xu H, Anderson K, Grann VR, Friedman C. Facilitating cancer research using natural language processing of pathology reports. MEDINFO 2004; Proceedings of the 11th Triennial World Congress on Medical Informatics; 2004 Sep 7-11; San Francisco, CA. 2004. p. 565-72.