Abstract

We coded 700 radiology reports from 373 women using an unmodified deployment of the Cancer Text Information Extraction System (caTIES), a publicly-available tool using natural language processing techniques. We were moderately successfully using caTIES for case ascertainment, successfully identifying 9/11 of a random sample of cancer case (sensitivity 82%) and 5/100 controls (specificity 95%). We are currently developing a classification scheme to assess clinical risk of ovarian cancer and identifying required extensions to caTIES algorithms.

Introduction

Radiology reports contain rich and often complex information about disease state and risk stored as free text making it difficult to use for research. Use of natural language processing (NLP) methods for mining information from loosely-structured clinical text is of widespread interest but has not been explored extensively for radiological studies. The Cancer Text Information Extraction System (caTIES) uses NLP techniques to index clinical concepts from standardized medical thesauri found in clinical text. CaTIES uses publicly available tools and algorithms to index relevant clinical concepts including negated concepts. Originally developed for pathology reports, caTIES is gaining attention as a tool for mining other types of clinical text.

Methods

We deployed caTIES 2.3 without modification as downloaded from http://caties.cabig.upmc.edu. From a collection of 200,000 pelvic ultrasound examinations of 130,000 women during 1997-2006, we selected 700 reports, including 600 for 273 women later diagnosed with ovarian cancer, and 100 from women without cancer. CaTIES coded and indexed these reports using its built-in algorithms. An initial sample of 15 reports were selected at random for detailed review by a radiologist. Review of a larger sample continues. We summarized characteristics of the reports and the codes generated, and assessed the efficiency of caTIES for case ascertainment based on the initial random sample. We defined cancer as present based on a pathologically confirmed ovarian cancer diagnoses within 12 months of examination, and cancer to be absent when no cancer was diagnosed. CaTIES coding was considered positive if the concept of ovarian cancer was identified by any of a set of terms commonly used by radiologists. In a second, ongoing study we are developing an ordinal classification algorithm for assessing clinical risk of disease, assembling a list of concepts specific to radiological studies. We are also identifying constructions common in radiology reports that require extensions to caTIES algorithms.

Results

Report text ranged in length from 230 to 5100 characters (median 1435, IQR: 1015-2045) and contained from 27 to 744 words (median 217, IQR: 149-289). CaTIES generated a total of 55,476 codes for the 700 reports, of which 22,225 were classified as referencing a diagnosis (61%), organ (36%), or procedure (3%). Nine percent of all codes were negated as were 15% of all classified codes. CaTIES codes successfully identified 9/11 of the random sample of reports in which the radiologist indicated evidence of ovarian cancer (sensitivity 82%) and incorrectly identified 5/100 controls (specificity 95%). Initial review of CaTIES-coded concepts suggests caTIES will also be useful for establishing risk profiles, but algorithm extensions will be needed to achieve acceptable specificity. Proposed modifications include 1) mapping ovarian cancer risk-specific terms to thesaurus concepts (e.g., “solid and cystic mass” or “mass with mural nodules”), 2) associating measured values with their referents (e.g., “6mm cyst”), and 3) establishing temporality to distinguish historical from current descriptions.

Conclusions

CaTIES is useful for case ascertainment from radiological studies and, a useful platform for developing specialized coding schemes using non-pathology clinical text. Algorithm and domain extensions are promising areas for further work.