Using SAS Perl Regular Expressions to Extract Patient Health Questionnaire Scores from Free Text

Ron Johnson, Rob Penfold, David Carrell

GroupHealth® RESEARCH INSTITUTE

Introduction

Free text notes in the electronic medical record (EMR) contain valuable numeric information represented as attribute-value pairs such as "PHQ-9 score today is 4." Natural language processing (NLP) methods can be used to extract such information. An editorial in *JAMA* Aug 24, 2011 emphasizes the importance of leveraging NLP methods to more fully exploit clinical information contained in EMRs.

The Patient Health Questionnaire- 9 (PHQ-9) score is a measure of the severity of depression symptoms commonly used in clinical practice. About 15% of the time PHQ-9 scores are only recorded in free text notes. Researchers use the PHQ-9 to evaluate various mental health treatments and interventions.

The information extraction challenge is to identify *current* PHQ-9 scores in notes containing a mix of prior scores, scores from other instruments, and other numeric values such as dates and dosages. We implemented in SAS a set of Perl regular expressions (Regex) to extract current PHQ-9 scores. An iterative algorithm was used, starting with scores unambiguously described as the current score, and then extracting remaining scores based on less specific criteria.

Methods

Logic applied on free-text to extract Current PHQ-9 score

Example 1) Current score is described unambiguously:
"…PHQ9 score today is 4/27, Clinical Interpretation of phq 20-27 Severe 15-19 Moderately Severe 10-14…."

- **Perl Regex:** (phq9\s*score\s*today\s*is\:?|current\s*phq9\s*score\s*is\s*) (\d{1,2})(?!(\/[^2]))
- **In English:** if numeric value is preceded by word(s) referring to the present, extract it.

Example 2) Isolate Current score from past scores and scores on other instruments:
"…on 12/13/09 PHQ9 was 25/27, today is 6, scoring 22 on Burns Anxiety…"

- **Perl Regex:** (….phq9\s*was\s*\d{1,2})(?!(\/[^2])\s*\/?,?\s*today\s*is\s*)(\d{1,2})(?!(\/[^2]))
- **In English:** if a PHQ9 score from the past (identified by date format) is followed by another PHQ score surrounded by word(s) referring to the present, extract it.

Example 3) Identify Current score based on proximity to key words:
"…scoring a 12 on PHQ, which is lower than previous score of 22…"

- **Perl Regex:** ….call prxposn function used on regex such as: "/score\b|scores\b|scored\b|scoring\b/i" in order to identify position of words in relation to each other.
- **In English:** After excluding previous score, calculate number of characters between the numeric score and the terms "scoring" and "PHQ". If the score is within 20 characters of both terms extract it.

Results

Specificity: When a PHQ-9 score was extracted, 91% were correctly identified.

Sensitivity: When a PHQ-9 score was not extracted, 11% of the notes contained a valid score.

Error analyses: Most PHQ-9 scores not extracted were in mentions with challenging constructions not amenable to regular expression techniques.

Testing at other sites: in progress

Summary

- Regular expressions perform well for extracting PHQ-9 scores from text.
- Implemented in SAS these methods are easily transportable to other institutions.

Conclusion

- This work illustrates that information represented in clinical notes can be extracted using regular expressions.
- Other clinical information represented as attribute-value pairs, such as tumor sizes, percent stenosis measurements, and other clinical scale scores, may also be amenable to this approach.
- Future work should explore methods to flag for human review notes likely to contain PHQ-9 scores that cannot be extracted by regular expressions.

PRINTER SPECIFICATIONS from Group Health Research Institute
Project Name: IT Poster (Ron Johnson)
Accounting Unit: 10021200
Paper: poster stock
Size: Print 40h" x 52.5w" -- file built at 100%
Bleeds: none; Inks: CMYK;
Bindery: trim, deliver in mailing tube.

CONTACT INFORMATION
Designer: Michelle Kindall at Group Health Research Institute
Phone: 206-287-2011 Mobile: 206-697-0861; Email: kindall.m@ghc.org

DELIVERY INFORMATION
Due: by Tuesday, Nov. 1, 2011
Attn: Michelle Kindall/Ron Johnson