Extracting information from unstructured electronic textual sources using regular expression rapid development tools

G Napolitano[1,2] C Fox[1] M Domanski[3] E O'Callaghan[1] M O'Rorke[3] Úna McMenamin[3] R Middleton[1]

[1]Northern Ireland Cancer Registry, Centre for Public Health, Queen's University Belfast; [2]Centre for Statistical Science and Operational Research, School of Mathematics and Physics, Queen's University Belfast, [3]Centre for Public Health, Queen's University Belfast.

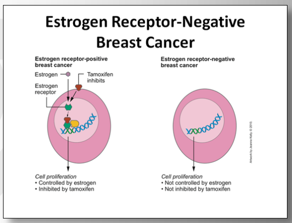
Background: Data items and prognostic factors, important in population-based research e.g. predicting survival, determining treatment effectiveness and outcomes for cancer patients, are often buried in free-text electronic annotations, comments and medical reports. In the past we have shown the use of short computer programs (Perl scripts) to automatically extract such information. Our objective is to enhance the rapid development of such applications. **Methods**: Previously developed regular expressions were tested and improved by means of a commercially available rapid development tool, without developing a full Perl (or other programming language) application. To assess the performance, we used the same test set of pathology reports used in a previously presented exercise, plus a set of annotations from the Northern Ireland Clinical Oncology Information System (COIS). **Results**: The time required for the design of a high precision and recall regular expression for hormone receptor status was one person-day, as opposed to around five person-days if a full Perl script and cycles of design-test-redesign were used. Precision and recall were also increased from ~90% to almost 100% for various data items. **Conclusion**: The tool proved dramatically effective in reducing the time spent developing regular expressions, cutting the workload by 80%. The tool also proved very valuable in allowing the testing of subtle changes in very complex regular expressions, which would be very unpractical to develop and evaluate in the full Perl setup. The increase in complexity of the regular expressions was responsible for the improved precision and recall. Although we tested an inexpensive commercial tool, free and open-source tools are also available with similar functionalities.

Past work in the NICR

- Developed various Perl scripts for the automatic extraction of information from free-text surgical pathology reports.
- Good results for Gleason score, Clark level, Breslow depth and hormone receptor status.
- See Poster 56 in NCIN 2011:
- Potential increase in staging completeness of up to 32% has been proved.

What are *receptors* and why are they important for breast cancer?

- The receptor status of a tumour is an indicator of what drugs may be effective in treatment.
- For instance, anti-oestrogens block oestrogen from activating genes for specific growth-promoting proteins and thus tumour growth.


Estrogen Receptor-Negative Breast Cancer
Estrogen receptor-positive breast cancer
Estrogen
Tamoxifen inhibits
Estrogen receptor
Cell proliferation
• Controlled by estrogen
• Inhibited by tamoxifen
Estrogen receptor-negative breast cancer
Cell proliferation
• Not controlled by estrogen
• Not inhibited by tamoxifen
Artwork by Jeanne Kelly © 2010.

- Other drugs target the progesterone receptor and the HER-2, although some mechanisms are not completely understood.

Receptor status mentions

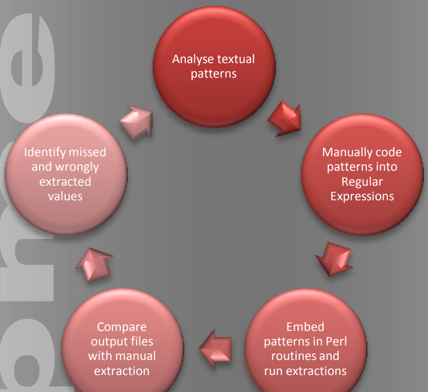
- In NI, the receptor status is mentioned with wide variation in wording and punctuation.

```
…ER/PR/HER-2 STATUS: ER
positive…
…PR weakly positive, score
3/8…
…ER STATUS – Negative…
…PR positive…
…both PR and ER are negative…
…Estrogen status +ve…
…Oestrogen score 5/8…
…negative for Her-2…
```

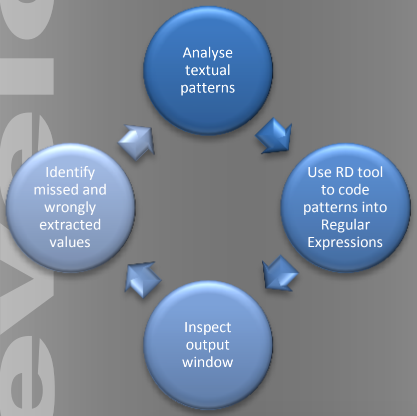
…and many more!

Technique and development process

- A number of reports for breast cancer patients were selected and inspected (O'Rorke and McMenamin) for manual extraction of receptor status.
- In the past, the following cycle would be repeated until acceptable level of performance was achieved:


Analyse textual patterns
Manually code patterns into Regular Expressions
Embed patterns in Perl routines and run extractions
Compare output files with manual extraction
Identify missed and wrongly extracted values

- This time, a Rapid Development tool for regular expressions (RegexBuddy) was used, which provided 'live' feedback on performance:


Analyse textual patterns
Use RD tool to code patterns into Regular Expressions
Inspect output window
Identify missed and wrongly extracted values

Evaluation

- A total of 230 documents were included in the evaluation set.
- The following performance indicators were calculated:

1. A measure of completeness:

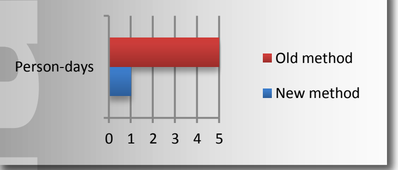
$$\text{Recall} = \frac{number\quad of\quad correctly\quad extracted\quad values}{total\quad number\quad of\quad values}$$

2. A measure of fidelity:

$$\text{Precision} = \frac{number\quad of\quad correctly\quad extracted\quad values}{total\quad number\quad of\quad extracted\quad values}$$

		Oestrogen	Progesterone	HER-2
Past results	*Recall*	88%	81%	85%
	Precision	97%	92%	85%
Present results	*Recall*	100%	100%	100%
	Precision	99%	92%	98%

3. Time spent in development and evaluation



Conclusions

- The regular expression Rapid Development tool proved dramatically effective
  - In helping to achieve higher precision and recall
  - In reducing the time spent developing regular expressions, cutting the workload by 80%.
- Subtle changes in very complex regular expressions were easily and quickly tested.
  - This allowed for the improved precision and recall.
- Free and open-source tools are available with similar functionalities.

Acknowledgements
The project was carried out at the Northern Ireland Cancer Registry, which is funded by the Public Health Agency, Northern Ireland.

Correspondence: g.napolitano@qub.ac.uk or scan: