

An Automatic Tool to Assign Receptor Status for Breast Cancers

G Napolitano^{1,2}, R Middleton, A Wilkinson¹, A Gavin¹

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



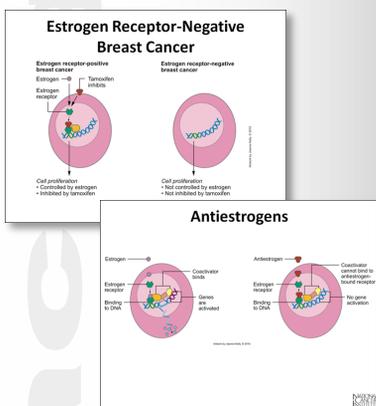
Objectives: The receptor status for breast cancer is an important prognostic factor for determining treatments and outcomes. Previously it was necessary to extract this information manually from histopathology reports. We planned to design and use a simple program in PERL to extract the oestrogen, progesterone and Her-2 receptor status from reports. **Methods:** 90 histopathology reports were analysed in their way to represent receptor status. Independently, a surgeon extracted receptor status manually from the same documents. A PERL script was designed to capture receptor status and tested on the reports, comparing the captured values with the surgeon's extracted values. **Results:** At the first run, the routine recall for oestrogen, progesterone and Her-2 receptor status, respectively, resulted 88%, 81% and 85%; precision was 97%, 92% and 85%; sensitivity was 97%, 95% and 71%; specificity was 100%, 83% and 88%. **Conclusion:** The PERL script can identify the status for all three receptors with an extremely high degree of accuracy. Further refinements in the script, already identified, will increase performance even further. This will make the extraction of these data items more economical and timely and with less need for manual intervention.

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 and Breslow depth.
- 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.
- Antiestrogens work by binding to estrogen receptors so that the estrogen molecules themselves cannot bind to those receptors.
- This also blocks estrogen from activating genes for specific growth-promoting proteins and thus tumour growth.



- Other drugs target the progesterone receptor and the HER-2.
- HER-2 positive cells are treated with trastuzumab, whose action is not completely understood.
- Possibly, by binding to HER-2 proteins it inhibits the sending of growth-promoting signals.

Technique and development process

- In NI, the receptor status is usually mentioned at the end of the report or in a supplementary report appended to the original one.
- Wide variation in wording and punctuation, sometimes with missing spaces between words.
- Some examples:

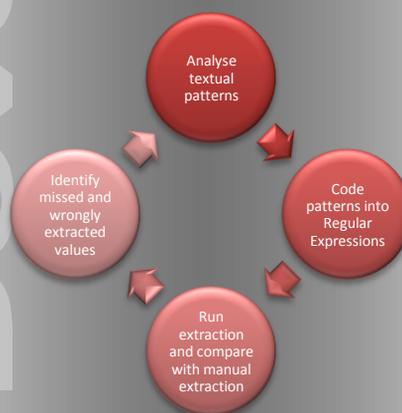
```
...ER/PR/HER-2 STATUS: ER
positive...
...PR weakly positive, score 3/8...
...ER STATUS - Negative...
...PR positive...
...both PR and ER are negative...
...Estrogen status +ve...
...Oestrogen score 5/8...
...negative for Her-2...
...and many more!
```

- A number of reports for breast cancer patients were selected and given to a surgeon (Wilkinson) for manual extraction of receptor status for

- Progesterone
- Oestrogen
- HER-2

- At the same time, all occurrences of relevant words ("PR", "ER", "positive"...) were automatically extracted from all reports, together with a 40+40 character window around the word.

- The following cycle was repeated until acceptable level of performance is achieved:



Evaluation

- A total of 90 surgical pathology reports were included in the training set.

- The following performance indicators were calculated:

1. A measure of completeness:

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

2. A measure of fidelity:

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

3. The ability to identify positive results:

$$\text{Sensitivity} = \frac{\text{number of true positive}}{\text{number of true positive} + \text{number of false negatives}}$$

4. The ability to identify negative results:

$$\text{Specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

- After only one evaluation cycle the following results were achieved on the training set:

	Oestrogen	Progesterone	HER-2
Recall	88%	81%	85%
Precision	97%	92%	85%
Sensitivity	97%	95%	71%
Specificity	100%	83%	88%

Conclusion

- The exercise confirmed that automatic extraction of valuable items is feasible, reliable and not overly complex.

Way Forward

- We are in the process of running more cycles to further refine the routine performance on the training set of reports.
- A test set of 2142 reports has been identified and will be used partially for further training and partially for independent testing.

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: