Title (max 120 characters):

Semi-automated extraction of information from pathology reports: proof of concept

Authors (max 10):

Joris Mattheijssens\*, Antoine Pironet\*, Kris Henau\*, Petra Denolf\*, Nancy Van Damme\*, Harlinde De Schutter\*

\*Belgian Cancer Registry, Brussels, Belgium

Body (max 2000 characters):

Objectives

The population-based Belgian Cancer Registry (BCR) collects data on all new cancer cases since 2004, and since 2008, also on all pathologically examined breast, colon and cervical specimens. Hospitals and laboratories send structured notifications which for the latter are generally accompanied by pathology reports. BCR explored the possibility to extract detailed information from these reports in a semi-automated way in two feasibility studies: (1) K-RAS mutation in colorectal cancer, and (2) human papillomavirus (HPV) genetic tests in cervical smears.

Methods

* Case study 1: A sentence-by-sentence approach was used to explore 16,109 colorectal cancer reports containing the search term “K-RAS”. Following extensive preprocessing, the 500 most frequent sentences were classified manually, and the remaining 10,609 automatically using the k-nearest-neighbours algorithm.
* Case study 2: A word-based approach was used to explore 163 cervical smear protocols. The most relevant words were selected using the mutual information criterion, and 13 classifiers were built from a training set of 72 texts. The classifiers were then tested on 91 different texts.

Results

* Case study 1: 46% of the reports contained a clear K-RAS test result, of which 32% were positive for K-RAS mutation. 44% of the reports mentioned the demand for a test, without stating a result. 77% sentences were correctly classified when testing with the leave-one-out method.
* Case study 2: The classifiers performed very well, with all sensitivities and specificities above 90%. The area under the receiver-operating characteristic (ROC) curve was > 0.92 for all classifiers.

Conclusions

Semi-automatic information extraction from pathology reports seems possible with acceptable accuracy, for different lesion types and text mining techniques. Relying on these encouraging results, the methodology will be further developed. If successful, the obtained information will substantially enrich the data available at BCR.