Title: Evaluation of an automated tool to identify positive cases from unstructured, free-text pathology reports in a Swiss Cancer Registry.

Pablo Iriarte[1]; Rafael Blanc Moya[2]; Nadia Elia[3]

[1]Institute for Social and Preventive Medicine, University of Lausanne, Switzerland.

[2] Vaud Cancer Registry, Institute for Social and Preventive Medicine, University of Lausanne, Switzerland

[3]Institute of Global Health, University of Geneva, Geneva, Switzerland

Objectives: The Vaud Cancer Registry receives about 150,000 pathology reports per year which need to be reviewed manually according to whether they describe a pathology requiring registration in the database as "positive reports", or discarded as "negative reports". This study examines the performance of a text mining automated tool (AT) created to scan these free-text medical reports for terms relevant to cancer.

Methods: We developed a custom-made list of 155 keywords including all terms likely to report a positive case in a pathology report, based on existing medical classifications, similar lists and on our working experience within the Vaud Cancer Registry. In order to identify the presence of the different keywords from the free-text of pathology reports in PDF format, we designed and launched an automated search script using Python Software (version 2.7). Additionally, for those reports provided as PDF files in image format, we used optical character recognition software (Adobe Acrobat Pro, version 9). The performance of the AT was evaluated by computing its sensitivity (Se), specificity (Sp), positive predictive value (PPV) and negative predictive value (NPV) based on a sample of 2302 pathology reports, and using the manual review performed by trained specialists as the gold standard.

Results: Of 2302 pathology reports, 742 were positive, 1560 were negative. The AT correctly identified 709 of the 742 positive cases (Se: 95.6%, 95%CI (93.8% to 96.9%)) and 1324 of the 1560 negative cases (Sp: 84.9%, 95%CI (83.0% to 86.6%)). For a prevalence of positive cases of 32%, the PPV was of 75%, 95%CI (72.1% to 77.8%) and the NPV was 97.6%, 95%CI (96.6% to 98.3%). The AT generated 236 false-positive, and 33 false-negative cases.

Conclusion: The AT is a promising tool that could improve greatly the efficiency of tumor registry human resources. Its sensitivity needs to be further improved by adding extra keywords, in order to avoid missing any positive case.