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 of 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

Context

The Vaud Cancer Registry receives about 150’000 pathology reports per year which need to be reviewed manually by trained specialists according to whether they describe a pathology requiring registration in the database as "positive reports", or discarded as "negative reports".

Objectives

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 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). The performance of the AT was evaluated by computing its sensitivity, specificity, positive predictive value, and negative predictive value based on a sample of 2’302 pathology reports, and using the manual review performed by trained specialists as the gold standard.

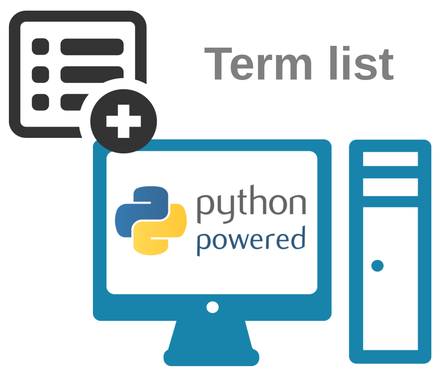
Results

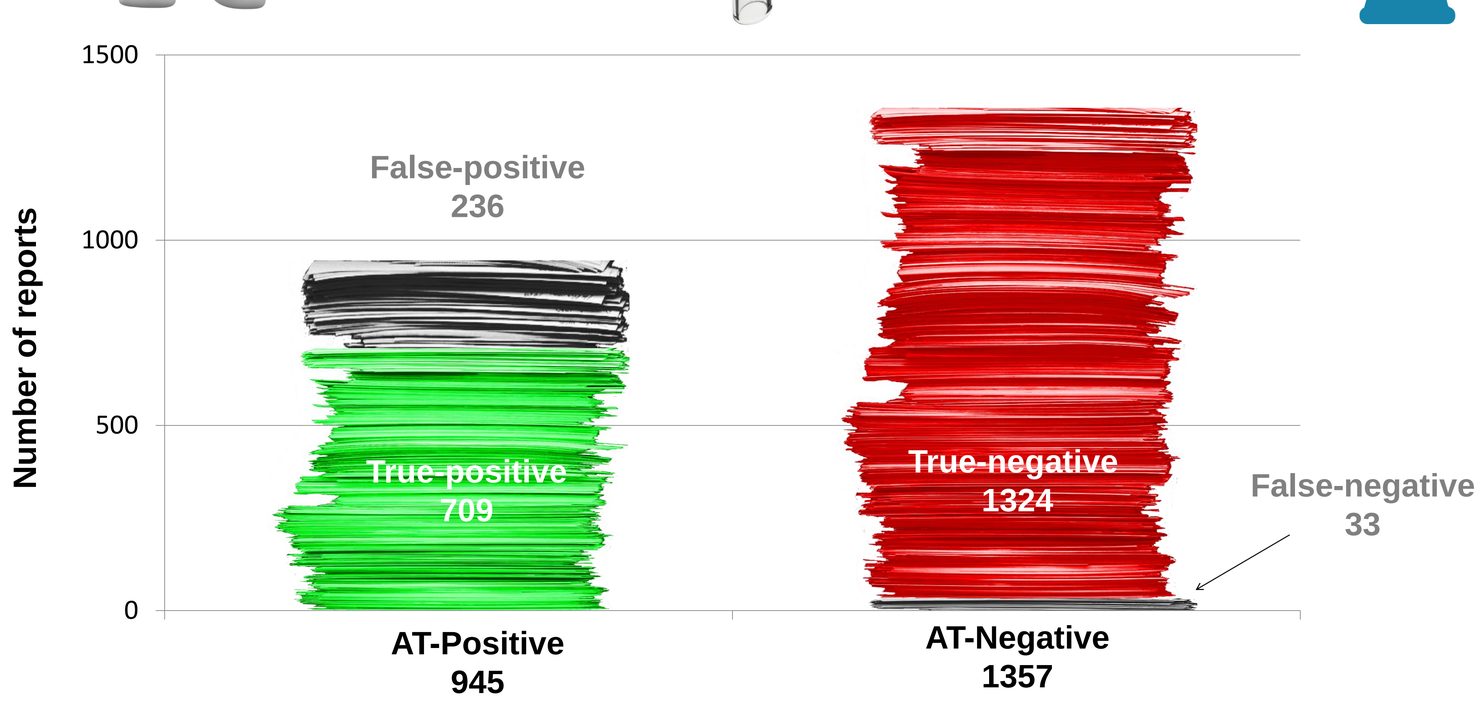
The trained specialists identified, among the 2’302 pathology reports:

- **742 positive reports**
- **1560 negative reports**

The AT generated:

- **236 false-positive**
- **33 false-negative** (1.4%)





Performance of the AT:

- **Sensitivity:** **95.6%**, 95%CI (93.8% to 96.9%)
- **Specificity:** **84.9%**, 95%CI (83.0% to 86.6%)

For an estimated prevalence of positive cases of 32%:

- **Positive Predictive Value :** **75.0%**, 95%CI (72.1% to 77.8%)
- **Negative Predictive Value :** **97.6%**, 95%CI (96.6% to 98.3%)

Conclusions

The AT is a promising tool that could greatly improve 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.

Source code

The AT code is available as open source project. Comments and collaboration are welcome : **https://github.com/pablogit/tdm**

Bibliography

- D’Avolio LW, Nguyen TM, Farwell WR, Chen Y, Fitzmeyer F, Harris OM, et al. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). J Am Med Inform Assoc. 2010 Jul–Aug;17(4):375–82.
- Hou JK, Chang M, Nguyen T, Kramer JR, Richardson P, Sansgiry S, et al. Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing. Dig Dis Sci. 2013 Apr;58(4):936–41.
- Contiero P, Tittarelli A, Maghini A, Fabiano S, Frassoldi E, Costa E, et al. Comparison with manual registration reveals satisfactory completeness and efficiency of a computerized cancer registration system. Journal of Biomedical Informatics. 2008;41(1):24–32.
- Luo Y, Sohani AR, Hochberg EP, Szolovits P. Automatic lymphoma classification with sentence subgraph mining from pathology reports. J Am Med Inform Assoc. 2014 Sep–Oct;21(5):824–32.
- Al-Haddad MA, Friedlin J, Kesterson J, Waters JA, Aguilar-Saavedra JR, Schmidt CM. Natural language processing for the development of a clinical registry: a validation study in intraductal papillary mucinous neoplasms. HPB (Oxford). 2010 Dec;12(10):688–95.
- Nguyen A, Moore J, Zuccon G, Lawley M, Colquist S. Classification of pathology reports for cancer registry notifications. Stud Health Technol Inform. 2012;178:150–6.
- Hanauer DA, Miela G, Chinnaiyan AM, Chang AE, Blayney DW. The registry case finding engine: an automated tool to identify cancer cases from unstructured, free-text pathology reports and clinical notes. J Am Coll Surg. 2007 Nov;205(5):690–7.
- Patrick J, Asgari P, Li M, Nguyen D. Using NLP to identify cancer cases in imaging reports drawn from radiology information systems. Stud Health Technol Inform. 2013;188:91–4.
- Spasic I, Livsey J, Keane JA, Nenadic G. Text mining of cancer-related information: Review of current status and future directions. International Journal of Medical Informatics. 2014 Sep;83(9):605–23.
- Python 2.7 : https://www.python.org/download/releases/2.7/
- Python pdfminer library : http://www.unixuser.org/~euske/python/pdfminer/index.html
- Python PyPDF2 library : https://pypi.python.org/pypi/PyPDF2/1.26.0

UNIL | Université de Lausanne

IUMSP - Institut universitaire de médecine sociale et préventive

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Institut de santé globale