The Journal of Open Source Software

EndoMineR for the extraction of endoscopic and associated pathology data from medical reports

Sebastian S Zeki[1]

1 Department of Gastroenterology, St Thomas' Hospital, Westminster Bridge Bridge Road, London SE1 7EH

DOI: 10.21105/joss.00701

Software

- Review
- Repository
- Archive

Submitted: 25 April 2018
Published: 27 April 2018

Licence
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Summary

Medical data is increasingly kept in an electronic format worldwide (Bretthauer M. 2016). This serves many purposes including more efficient storage, distribution and accessibility of patient-focussed data. As important is the ability to analyse healthcare data for to optimize resource deployment and usage. The tools for the analysis are often statistical and rely on the provision of 'clean' datasets before this can be done. 'Cleaning' a dataset is often the most difficult aspect of any data analysis and involves the provision of meaningful and well-formatted data so that the interpretation of the analysis is not subject to the doubts of the data quality.

The British Society of Gastroenterology recommends that all endoscopic data is kept in an electronic format particularly to facilitate audit and maintain standards through the Global Rating Scale (GRS) (Stebbing J. 2011). The endoscopic dataset is however only part of the patient's story as many aspects of a patient's gastroenterological care depend on the results of histopathological analysis of tissue taken during the examination. Pathology results are often available many days after the endoscopic result and usually stored in a separate data repository, although this may change with the arrival of an encompassing electronic patient record. Regardless of the method of storage, it is often difficult to associate the particular histopathological result with an endoscopic result. Further, even if the two data sets can be merged, a problem occurs in the isolation of various parts of each report such that each part can be individually analysed. Examples include the isolation of who the endoscopist was or the presence of dysplasia within a histopathology report. This is all the more difficult if the report is unstructured or partially structured free text.

However if this can be done then many downstream analyses which benefit individual patients as well as the department, can be automated and include more complex analyses to determine follow-up regimes or endoscopic –pathologic lesion recognition performance.

The EndoMineR package provides a comprehensive way to extract information from natural language endoscopy ann pathology reports as well as merging the two datasets so that pathology specimens are relevant to the endoscopy they came from. Furthermore the package also provides functions for the following types of analysis of endoscopic and pathological datasets:

- 1. Patient surveillance. Examples including determining when patients should return for surveillance and who is overdue.
- 2. Patient tracking. -Examples include determining the length of time since the last endoscopy, as well as aggregate functions such as finding how many endoscopies of a certain type have been done and predicting future burden.

- 3. Patient flow - determining the kinds of endoscopies an individual patient may get over time eg for ablation of Barrett's oesophagus.
- 4. Quality of endoscopy and pathology reporting- Determining whether endoscopy quality is being maintained using some of the Global Rating scale metrics. Also making sure the pathology reports are complete.
- 5. Diagnostic yield. Examples include determination of detection of dysplasia and cancer by endoscopist as a measure of lesion quality.

It is the purpose of the package to create a unified process for merging of endoscopy reports with their associated pathology reports and to allow the extraction and tidying of commonly need data. Furthermore the package has methods for the analysis of the data in areas that are commonly required for high quality endoscopic services. This includes methods to track patients who need endoscopic surveillance, methods to determine endoscopic quality and disease detection rates. Also included are methods to assess patient flow through different types of endoscopy and to predict future usage of certain endoscopic techniques.

The package is in the process of having each analysis function validated and functions some validation has been submitted in abstract form to gastroenterological societies.

References

Bretthauer M., Dekker E., Aabakken L. 2016. "Reporting systems in gastrointestinal endoscopy: Requirements and standards facilitating quality improvement: European Society of Gastrointestinal Endoscopy position statement." *United European Gastroenterology Journal* 4 (April):172–6. https://doi.org/10.1177/2050640616629079.

Stebbing J. 2011. "Quality assurance of endoscopy units." *Best Practice and Research Clinical Gastroenterol* 25 (June):361–70. https://doi.org/10.1016/j.bpg.2011.05.004.