Extraction and analysis of discrete synoptic pathology report data using R

Alexander Boag

1Department of Pathology and Molecular Medicine, Queen’s University, Kingston, Ontario, K7L 3N6, Canada

E-mail: *Dr. Alexander Boag - boaga@queensu.ca
*Corresponding author

Received: 18 August 2015 Accepted: 08 October 2015 Published: 27 November 2015

Abstract

**Background:** Synoptic pathology reports can serve as a rich source of cancer information, particularly when the content is available as discrete electronic data fields. Our institution generates such reports as part of a province wide program in Ontario but the resulting data is not easily extracted and analyzed at the local level. **Methods:** A low cost system was developed using the open sourced and freely available R scripting/data analysis environment to parse synoptic report results into a dataframe and perform basic summary statistics. **Results:** As a pilot project text reports from 427 prostate needle biopsies were successfully read into R and the data elements split out and converted into appropriated data classes for analysis. **Conclusion:** This approach provides a simple solution at minimal cost that can make discrete synoptic report data readily available for quality assurance and research activities.

**Key words:** Cancer, quality, R, synoptic

INTRODUCTION

The use of synoptic pathology reports has become well-established since first being described over 20 years ago. In comparison to the historical paragraph format, a synoptic structure offers the potential for improved report completeness and standardization as well as compliance with centrally developed classification and reporting systems, such as the College of American Pathologists (CAP) Cancer Protocol Templates. In Ontario, province-wide CAP-compliant electronic synoptic reporting of most cancer resection specimens has been in place for several years. Pathologists use various software systems which typically employ drop-down lists or check boxes to complete discrete data fields which are then transferred via interfaces to both a patient report and to the central provincial cancer registry operated by Cancer Care Ontario (CCO). CCO has used this data to track both compliances with the reporting program and quality indicators such as colon cancer lymph node recovery rates at a provincial level. In the future, one could envisage a wealth of population-based quality indicator metrics that could be derived from such information.

One somewhat unexpected phenomenon that we have encountered at our institution is that while our discrete synoptic report data are available at a provincial level, it is not readily accessed through our own hospital laboratory information systems (LISs). At Kingston

© 2015 Journal of Pathology Informatics | Published by Wolters Kluwer - Medknow

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as the author is credited and the new creations are licensed under the identical terms.

For reprints contact: reprints@medknow.com

This article may be cited as:
Boag A. Extraction and analysis of discrete synoptic pathology report data using R. J Pathol Inform 2015;6:62. Available FREE in open access from: [http://www.jpathinformatics.org/text.asp?2015/6/1/62/170649](http://www.jpathinformatics.org/text.asp?2015/6/1/62/170649)
General Hospital, mTuitive xPert® Cancer Reporting version 3 software (mTuitive Corporation, Centerville MA, USA) is used by pathologists to complete synoptic cancer templates with the data transferred via electronic interface from mTuitive to CCO and also to our Sunquest CoPath pathology system. However, once this process is complete, the discrete pathology data elements are not easily searched or extracted at the local level for academic or quality assurance (QA) purposes.

The objective of this project was to demonstrate that synoptic pathology report data elements suitable for QA purposes could be extracted into data analysis software at minimal cost. As the basis of this proof of principle pilot project, prostate cancer needle biopsy synoptic reports were selected due to their relative frequency and simplicity.

**METHODS**

**Software and Equipment**

R was selected as the software tool for both synoptic report extraction and for data analysis. R is an open source scripting, data analysis, and graphical environment which is available without cost for most operating systems (Institute for Statistics and Mathematics, Vienna, Austria www.r-project.org). R version 3.1.2 was installed on a standard Lenovo ThinkCenter workstation running Windows 7 Professional SP1 (Microsoft Corporation, Redmond, WA, USA) equipped with 4 GB of RAM. doPDF version 8.0 (available without cost at www.doPDF.com, Softland SRL, Romania) was used as a virtual printer driver to create a PDF file from the CoPath synoptic report extract output. Adobe Reader XI (version 11.0.06, www.adobe.com, Adobe Systems Incorporated, San Jose, CA, USA) was used to view the PDF files and Microsoft Office Word 2003 SP3 (Microsoft Corporation, Redmond, WA, USA) was used to create text files.

**Approach**

An overview of the approach is shown in Figure 1. For patient report creation, the mTuitive software is launched from CoPath, and the resulting synoptic report is transferred back to the patient report in CoPath. The synoptic report data elements for cases of interest were extracted by running the default “CoPath InfoMaker Wizard for Natural Language Search” for the time span and specimen types desired and specifying “mTuitive Synoptic Report” as the text type to print. This extract generates a listing of all cases containing the desired synoptic report with common case information (accession number, medical record number, accession and sign off dates) which is present in all reports. This is followed by the synoptic data elements on separate lines, each prefixed by uniform descriptors, as shown in Figure 2, which vary between types of synoptic reports and potentially within a report type due to the optional nature of some CAP-compliant report data elements. The print output was routed to the doPDF virtual printer, and the PDF file was converted to text by opening in Adobe Reader, using the “copy file to clipboard” menu command, pasting the file into MS Word, and saving the result as a .txt file. This produced a single long text file including all the synoptic report data elements for all cases over the desired time span.

A schematic of the R script is shown in Figure 3. The synoptic report text file was first imported as a master data list of text lines using the R readLines function. Two rounds of data extraction were then performed. In the first round, lines of text containing common demographic data (accession number, etc.) were selected by matching text patterns using the R grep function and then split into individual strings with the strsplit function. As illustrated by

---

**Figure 1**: Flowchart showing approach used to generate synoptic reports and extract discrete data elements in R

**Figure 2**: CoPath InfoMaker output listing of a prostate needle biopsy synoptic report prior to import into R (patient identifiers changed)
the synoptic report in Figure 2, lines found to contain the text “gender” after splitting would yield gender and age as the second and fourth data elements, respectively. In this round of extraction, the specifications were hard coded for simplicity and to allow easier handling of data complexities such as variable structures of pathologists’ names. Using the R sapply function, the extraction was applied to the entire master data list without looping to create a separate list for each of the six demographic data elements. These lists were then used to define and populate the main data frame.

In the second round of data extraction, the interpretative data (tumor type, grade, etc.) were captured and appended to the newly created main data frame. To allow the same code to be adapted for multiple synoptic report templates, a simple delimited text file was read into a secondary data frame specifying appropriate data variable names, search strings, data element positions, and class types (factor vs. numeric). For example, the file shown in Figure 4 for the prostate biopsy template indicates that the predominant Gleason grade is to be extracted by searching for lines containing the text “predominant,” selecting element 5 from the split string, storing it temporarily in a variable called “gg1,” and converting the string to a numerical data class. For each case, a loop cycled through each data specification in the secondary data frame and appended the result extracted from the master text list to the main data frame. The final output was a single R data frame that for prostate needle biopsies contained 17 data elements for each case [Figure 5] which were then analyzed within R.

Pilot Project
The synoptic data for 427 prostate needle biopsy cases of prostate cancer were extracted from CoPath by running the InfoMaker Wizard for the prostate needle biopsy part types over a 5 years time span (February 2010–February 2015). The text output was imported into R as described above and basic summary statistics applied using methods available in the base R package and the R “psych” library. Processing times in R were measured using the proc.time command.

RESULTS
The CoPath InfoMaker Wizard Natural Language Search applied over a 5 years period yielded 427 synoptic reports for prostate cancer in needle biopsies specimens with a run time of 45 s. The search output file sizes, when saved as PDF and then converted to .txt, were 3.89 MB and 535 KB, respectively. Processing time in R to read the data, parse into a data frame, and correct the data classes was 3.22 s total elapsed time comprised 2.02 s user (i.e., code) time and 1.20 s system (i.e., CPU) time.

Summary statistics for the resulting numerical data (age, the number of cores positive for cancer, total cores, and percent of tissue involved by cancer) are shown in Table 1. Selected categorical data which showed the greatest variation between pathologists is illustrated in Tables 2 and 3. The variation in primary Gleason grade diagnostic rates [Table 2] and diagnosis of extraprostatic spread [Table 3] did not reach statistical significance when tested using Pearson’s Chi-squared test ($P = 0.061$ and $P = 0.43$, respectively).

DISCUSSION
A synoptic structure has become the standard for cancer resection pathology reports. When implemented such that results are available as discrete data elements, the value of the report is leveraged and facilitates QA and research. There are increasing expectations on the part of hospitals, regulatory agencies, and professional associations that pathologists will practice within the
Unfortunately, the resulting policies and practice guidelines are often not accompanied by funding to support the ongoing costs of data acquisition and analysis required for such activity. It is, therefore, necessary for pathology groups to seek out low-cost processes that can be used by support staff, who may have limited technical expertise.

In this paper, the R environment forms the cornerstone of a low-cost synoptic pathology report data analysis system. R is freely available, is widely supported through multiple online user groups, can be expanded with multiple add-on libraries and has a simple but powerful scripting language. R is commonly used for educational and research purposes by colleges and universities such that many recent graduates will have a working knowledge of this system. Using a relatively basic R script, data elements from a CoPath synoptic report extract were parsed from a text file into a data frame to become available for a wealth of statistical exploration.

A discrete synoptic structure has been previously shown to improve pathology report completeness while maintaining workflow and report format acceptable to pathologists and clinical physicians, respectively.[2,4,6] However, there has been very little published previously documenting the actual extraction and use of resulting data from these reports. Some LIS’s offer integrated synoptic reporting and the ability to access this synoptic data using their own query tools. This approach offers a simple solution but may be expensive and can limit flexibility and lack more sophisticated statistical methods. A recently described alternative is a web-based stand-alone reporting system with a back-end database which would provide easily accessible synoptic data.[7] For the LIS, which lacks the ability to directly query synoptic reports one is left with the extraction of data from text, PDF, or similar files into a separate database for analysis. Commercial parsing tools (e.g., Datawatch Monarch, http://www.datawatch.com) are able to import complex text data into most databases but can be costly. Alternatively, a number of open source systems offer both natural language processing (NLP) and analysis capabilities including R, Python Natural Language Toolkit (www.nltk.org), and tools from the Stanford NLP Group (www.nlp.stanford.edu/software). The choice will depend largely on the local expertise, the complexity of the extraction process, and the type of statistical analysis required.

To illustrate the potential of this process, the CAP-compliant synoptic pathology reports from 427 cases of prostate cancer in needle biopsies were extracted from CoPath into R and analyzed using basic summary statistics. The data elements available included basic patient demographics, specimen dates, pathologist name, multiple online user groups, can be expanded with multiple add-on libraries and has a simple but powerful scripting language. R is commonly used for educational and research purposes by colleges and universities such that many recent graduates will have a working knowledge of this system. Using a relatively basic R script, data elements from a CoPath synoptic report extract were parsed from a text file into a data frame to become available for a wealth of statistical exploration.

A discrete synoptic structure has been previously shown to improve pathology report completeness while maintaining workflow and report format acceptable to pathologists and clinical physicians, respectively.[2,4,6] However, there has been very little published previously documenting the actual extraction and use of resulting data from these reports. Some LIS’s offer integrated synoptic reporting and the ability to access this synoptic data using their own query tools. This approach offers a simple solution but may be expensive and can limit flexibility and lack more sophisticated statistical methods. A recently described alternative is a web-based stand-alone reporting system with a back-end database which would provide easily accessible synoptic data.[7] For the LIS, which lacks the ability to directly query synoptic reports one is left with the extraction of data from text, PDF, or similar files into a separate database for analysis. Commercial parsing tools (e.g., Datawatch Monarch, http://www.datawatch.com) are able to import complex text data into most databases but can be costly. Alternatively, a number of open source systems offer both natural language processing (NLP) and analysis capabilities including R, Python Natural Language Toolkit (www.nltk.org), and tools from the Stanford NLP Group (www.nlp.stanford.edu/software). The choice will depend largely on the local expertise, the complexity of the extraction process, and the type of statistical analysis required.

To illustrate the potential of this process, the CAP-compliant synoptic pathology reports from 427 cases of prostate cancer in needle biopsies were extracted from CoPath into R and analyzed using basic summary statistics. The data elements available included basic patient demographics, specimen dates, pathologist name,
primary and secondary Gleason grade, Gleason grade sum, the number of cores positive, the total number of cores, percent tissue involved by tumor and, when present, extraprostatic extension, seminal vesicle invasion, perineural spread, and optionally lymphatic-vascular invasion.

Data extraction from CoPath was possible in less than a minute using a default query tool (“InfoMaker Wizard for Natural Language Search”) and the results converted successfully to a .txt file using readily available software without added expense. Processing time to import, parse, and analyze the data using R was just over 3 s.

In this instance, the lack of statistically significant differences between the reporting pathologists constitutes a reassuring QA finding. However, the analysis illustrates the ability to identify trends in variation that might be potential targets of more detailed QA activities. For example, a pathology group might decide that the diagnosis of extraprostatic extension in a needle biopsy should be the subject of review by a second pathologist prior to sign out. Given the ease and speed with which this approach can be used once established, departments could monitor diagnostic rates on a regular basis to detect evidence of criteria drift or variant practice by a new group member.

Ideally, a single pathology LIS would offer the complete functionality of this process, but at present that will often not be the case and changing an LIS for QA purposes alone would likely be cost prohibitive. The use of an in-house supported add-on data analysis tool such as an R not only saves up front costs but also provides flexibility and allows for data analysis methods to be changed relatively easily as potential QA targets develop over time.

Financial Support and Sponsorship
Nil.

Conflicts of Interest
There are no conflicts of interest.

REFERENCES
1. Markel SF, Hirsch SD. Synoptic surgical pathology reporting. Hum Pathol 1991;22:807-10.
2. Hassell LA, Parwani AV, Weiss L, Jones MA, Ye J. Challenges and opportunities in the adoption of College of American Pathologists checklists in electronic format: Perspectives and experience of Reporting Pathology Protocols Project (RPP2) participant laboratories. Arch Pathol Lab Med 2010;134:1152-9.
3. Amin W, Sirintrapun SJ, Parwani AV. Utility and applications of synoptic reporting in pathology. Open Access Bioinformatics 2010;2:105-12.
4. Lankshear S, Srigley J, McGowan T, Yurcan M, Sawka C. Standardized synoptic cancer reporting – So what and who cares? A population-based satisfaction survey of 970 pathologists, surgeons and oncologists. Arch Pathol Lab Med 2013;137:1599-602.
5. Srigley J, Lankshear S, Brierley J, McGowan T, Divaris D, Yurcan M, et al. Closing the quality loop: Facilitating improvement in oncology practice through timely access to clinical performance indicators. J Oncol Pract 2013;9:e255-61.
6. Messenger D, McLeod R, Kirsch R. What impact has the introduction of a synoptic report had on reporting outcomes for specialist gastrointestinal and nongastrointestinal pathologists? Arch Pathol Lab Med 2011;135:1471-5.
7. Baskovich BW, Allan RW. Web-based synoptic reporting for cancer checklists. J Pathol Inform 2011;2:16.