UW Medicine
UNIVERSITY OF WASHINGTON MEDICAL CENTER

An Algorithm to Automatically Categorize Incident Learning Safety Reports Using Text Token Clustering

M. Havard[1], Q. Zhang[2], E. Ford[1]
1 University of Washington, Department of Radiation Oncology, Seattle, WA
2 University of Washington, School of Medicine, Seattle, WA

JULY 12–16 2020 VIRTUAL
Improving Health QUALITY. Increasing Global IMPACT
JOINT AAPM | COMP MEETING
EASTERN TIME [GMT-4]

INTRODUCTION

- Incident learning systems (ILS) are an important part of Quality & Safety programs in healthcare, with rising use in radiation oncology.[1]
- Incident learning systems are amassing reports both within institutions and nationally/internationally.
- It is valuable to be able to recognize trends in incident reporting both within and across institutions in review processes, so that these areas may be specifically addressed to minimize risk of error.
- As data repositories grow, the unstructured data must be analysed in a way that is scalable.

AIMS

- Develop a novel method for categorizing textual reports in ILS systems.
- Use an unsupervised clustering algorithm on narrative text.

METHODS

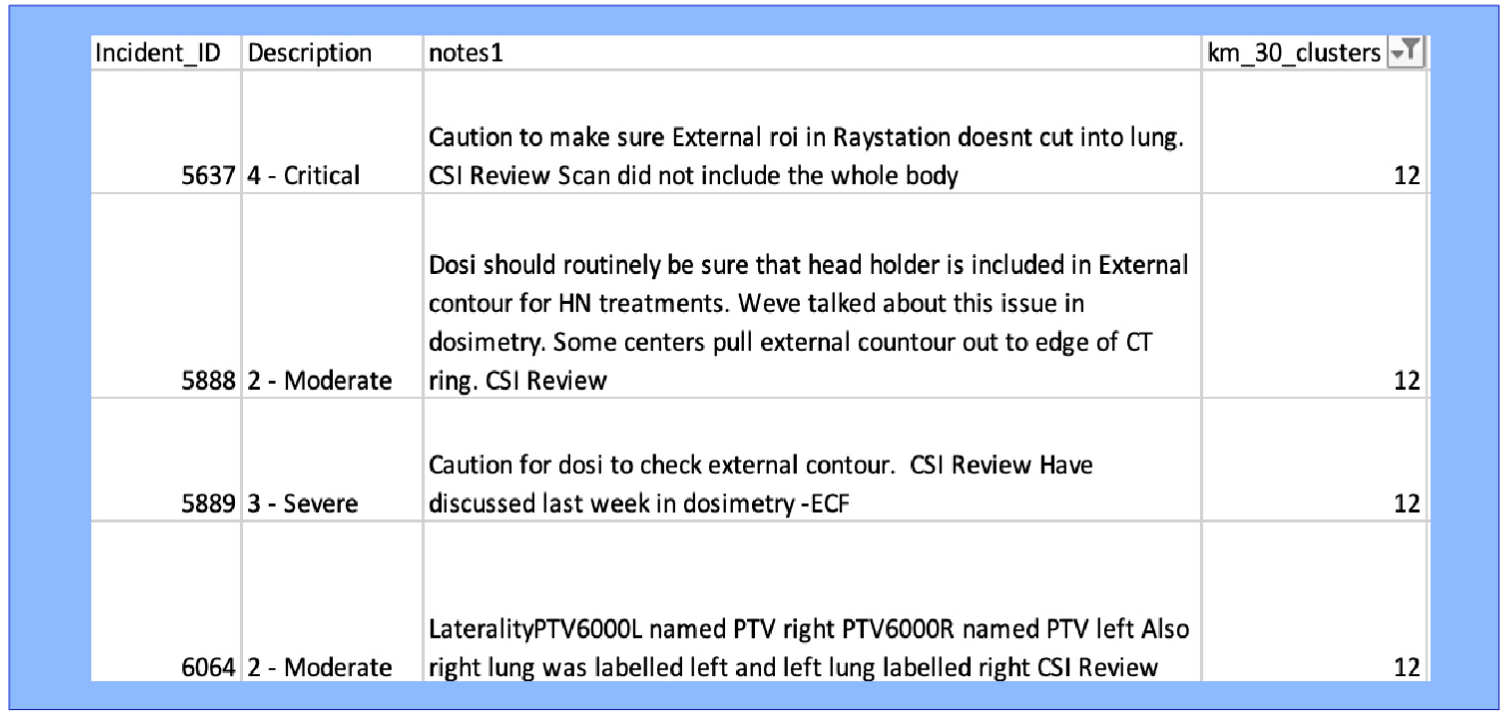
- Analyzed 6,430 reports of near-miss incident reports from a single institution
- Text of each report was tokenized (broken up into words) using the tidytext R package, which treats each report as a bag of words
- For each report and each token, we calculated the term frequency inverse document frequency (TF*IDF), which is a standard metric used in natural language processing that quantifies the frequency of occurrence within a report and the uniqueness among reports [2]
- We generated a numeric matrix of TF*IDF values, with rows corresponding to reports and columns corresponding to tokens
- We then applied K-means clustering to group the reports, using K = 30 clusters, 5 starting values, and a maximum of 100 iterations per starting value.

RESULTS

After obtaining an assignment of reports to clusters, to identify a theme amongst reports in a cluster, we:

- Read reports in the cluster, and
- Identified tokens with some of the highest TF*IDF values amongst reports in a cluster

Amongst high TF*IDF tokens, we found “junk tokens”, which are words commonly used in the reporting system that do not pertain to the event reported (e.g. numbers or “CSI Review” performed).



Incident_ID	Description	notes1	km_30_clusters
5637	4 - Critical	Caution to make sure External roi in Raystation doesnt cut into lung. CSI Review Scan did not include the whole body	12
5888	2 - Moderate	Dosi should routinely be sure that head holder is included in External contour for HN treatments. Weve talked about this issue in dosimetry. Some centers pull external countour out to edge of CT ring. CSI Review	12
5889	3 - Severe	Caution for dosi to check external contour. CSI Review Have discussed last week in dosimetry -ECF	12
6064	2 - Moderate	LateralityPTV6000L named PTV right PTV6000R named PTV left Also right lung was labelled left and left lung labelled right CSI Review	12

Figure 1: Example near-miss incident reports in cluster #12 (of 30)
All reports in this cluster are related to the region of interest for the external contour in the treatment planning system (RayStation v. 8b). Note that report #6064 should not be included in this cluster, but its inclusion is driven by “junk tokens” (Figure 2).

	Token Word			
Report #	“External”	“Dosimetry”	“Contour”	“Review”
5637	2.3	0.0	0.0	0.55
5888	1.9	0.96	0.85	0.45
5889	2.1	1.23	1.15	0.57
6064	0.0	0.0	0.0	0.61

Figure 2: Tokenization method and example TF*IDF (term frequency inverse document frequency) values.
“Review” is an example of a “junk token” that results in the inclusion of report 6064 in the cluster shown in Figure 1.

The K-means clustering iterative algorithm[3] provides a local optimum for the problem of finding the assignment of reports to clusters that minimizes the total within-cluster variation $\sum_{k=1}^{K} W(C_k)$, where k indexes K reports, within-cluster variation for the k-th cluster is

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \epsilon C_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 ,$$

$|C_k|$ is the number of reports in the k-th cluster C_k, i and i' index all pairs of reports in cluster C_k, $\sum_{j=1}^{p}\left(x_{ij} - x_{i'j}\right)^2$ is the squared Euclidean distance between reports i and i', j indexes p tokens and x_{ij} (or $x_{i\prime j}$) is the TF*IDF value for the j-th token in the i-th (or i'-th) report.

The K-mean clustering algorithm resulted in clusters that represented meaningful groupings (e.g. regarding handling of the patient support assembly). However, there were two large clusters that did not define particular, narrow topics.

Identification of tokens with high TF*IDF values within these clusters suggests that removing certain low-value “junk” tokens (e.g. numerical quantities) before clustering could help improve clustering.

CONCLUSIONS

- This method offers a novel means of categorizing and analysing safety incident reports.
- The K-mean clustering algorithm resulted in clusters that were largely representative of meaningful groupings.
- Initial analyses resulted in both large and small clusters driven by “junk tokens” rather than meaningful categorization. Further development is focusing on token refinement, and optimizing the starting values and number of iterations per starting value to improve clustering.
- The ability to categorize safety reports has many applications including the identification of high-priority quality gaps, quantifying trends over time and comparison across systems.
- This method requires little to no human intervention, and is potentially reproducible and scalable to large data sets and across systems.

ACKNOWLEDGEMENTS

The University of Washington, Department of Radiation Oncology Quality & Safety Review

REFERENCES

1 Ford EC, Evans SB. Incident learning in radiation oncology: A review. *Med Phys.* 2018;45(5):e100-e119. doi:10.1002/mp.12800

2. Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation.*

3. James, G., Witten, D., Hastie, T., Tibshirani, R., Casella, G., Fienberg, S., & Olkin, I. (2013). *An Introduction to Statistical Learning: With Applications in R* (Vol. 103, Springer Texts in Statistics). New York, NY: Springer New York.

CONTACT INFORMATION

Molly Havard, MD, MS
University of Washington, Department of Radiation Oncology
mhavard@uw.edu