HHS Public Access

Author manuscript

IEEE EMBS Int Conf Biomed Health Inform. Author manuscript; available in PMC 2018 April 24.

Published in final edited form as:

IEEE EMBS Int Conf Biomed Health Inform. 2017 February ; 2017: 349–352. doi:10.1109/BHI.2017.7897277.

Analyzing the Usage of Standards in Radiation Therapy Clinical Studies

Y. Zhen,
Department of Software and Information, Peking University 3rd Hospital, Beijing, China

Y. Jiang,
Department of Radiation Oncology, Peking University 3rd Hospital, Beijing, China

L. Yuan,
Department of Radiation Oncology, Duke University Medical Center, Durham, NC

J. Kirkpartrick,
Department of Radiation Oncology, Duke University Medical Center, Durham, NC

J. Wu, and
Department of Radiation Oncology, Duke University Medical Center, Durham, NC

Y. Ge*
Department of Software and Information, Peking University 3rd Hospital, Beijing, China

Abstract

Standards for scoring adverse effects after radiation therapy (RT) is crucial for integrated, consistent, and accurate analysis of toxicity results at large scale and across multiple studies. This project aims to investigate the usage of the three most commonly used standards in published RT clinical studies by developing a text-mining based analysis method. We develop and compare two text-mining methods, one based on regular expressions and one based on Naïve Bayes Classifier, to analyze published full articles in terms of their adoption of standards in RT. The full dataset includes published articles identified in MEDLINE between January 2010 and August 2015. A radiation oncology physician reviewed all the articles in the training/validation subset and produced the usage trending data manually as gold standard for validation. The regular-expression based method reported classifications and overall usage trends that are comparable to those of the domain expert. The CTCAE standard is becoming the overall most commonly used standards over time, but the pace of adoption seems very slow. Further examination of the results indicates that the usage vary by disease type. It suggests that further efforts are needed to improve and harmonize the standards for adverse effects scoring in RT research community.

I. Introduction

Radiation therapy (RT) is one of the most frequently used and effective treatments against cancer. Like any cancer treatment, the therapeutic benefit of radiation therapy is balanced

*Corresponding author: Y. Ge (phone and fax: +1-704-687-1951).

against potential adverse effects (or toxicity) in normal tissues. Therefore, accurate scoring and reporting of adverse effects in radiation therapy are critically important for RT quality improvement and treatment effectiveness research.

This paper focuses on a key aspect of data in radiation therapy, which is the scoring and reporting of adverse effects caused by radiation therapy. Unfortunately, a number of standards for scoring adverse effects have been proposed and used in RT clinical studies. Three of the most commonly used standards are the Common Terminology Criteria for Adverse Events (CTCAE)[1], the Radiation Therapy Oncology Group (RTOG)[2], and the Late Effect Normal Tissue Task Force (LENT)-subjective, objective, management, analytic (SOMA)[3]. These adverse effects scoring standards have been revised multiple times in recent years. In particular, CTCAE is strongly promoted as the comprehensive standard for adverse effects reporting in all cancer care [4, 5]. The Radiation Therapy Oncology Group (RTOG)/European Organization for Research and Treatment of Cancer (EORTC) Late Morbidity System have been updated for many versions and are still in use, containing criteria for grading late radiation morbidity, acute radiation morbidity and common toxicity criteria. To improve the RTOG/EORTC Late Effects System, the Late Effect Normal Tissue Task Force (LENT)-Subjective, Objective, Management, Analytic (SOMA) was published as a universal system for late effects of radiation therapy in 1995 [6].

A number of efforts have been reported to study and improve the scoring systems for grading tissue toxicities induced by radiation [7–11]. It is clear that the choice of standardized adverse effect scoring criteria has significant impact on the assessment and improvement of radiation treatment. However, there is a lack of studies that attempts to understand how the various standards have been used in RT clinical studies, and therefore lack of understanding as to how these standards should be adopted, harmonized, or improved. This project aims to investigate the usage of the above three adverse effect standards in published literature of RT clinical studies. We will address two specific questions: (1) the portion of clinical articles that use each standard by year and by cancer type; (2) the trend of usage in recent years.

In order to answer the above questions, we have developed a text mining based method for automatically categorizing articles based on grading criteria, and identifying cancer types of interest in the articles. While manual analysis by human experts is not prohibitive for individual questions, text mining techniques enable highly efficient and automatic investigation of multiple comprehensive questions using large-scale biomedical literature. Automated text analysis also allows continuous update of usage analysis as newly published articles are added. Numerous literatures exist for text mining techniques used in clinical medicine [12–15]. In this paper, we develop and compare two text mining approaches, one based on regular expression and one based on machine learning.

II. Methods and materials

A. Materials

We selected RT-related clinical articles in MEDLINE (pubmed.gov) that were published between January 2010 and December 2012 to train and validate models. A total of 668

articles were found using a search strategy that organized the search terms in following three groups: Group 1, terms related to radiation therapy, such as “radiotherapy”, “radiation therapy”, “chemo-radiotherapy”; Group 2, terms related to adverse effects, such as “toxicity”, “side effect”, “adverse effect”; and Group 3, terms related to standards in RT, such as “Common Toxicity Criteria”, “CTCAE”, “LENT-SOMA”, “RTOG”. The combination of search terms from each search group formed one complete search term. For instance, one such complete search term consisted of ‘radiation therapy’ from Group 1, ‘toxicity’ from Group 2, and ‘CTCAE’ from Group 3, respectively.

From the 668 results, we excluded 104 articles due to inadequate information, and 33 articles due to duplication, resulting in a total of 531 articles that were analyzed in our study. All selected full articles were downloaded and extracted into text files for analysis, including both abstracts and full texts, while excluding all figures, tables, and reference lists. The entire dataset for trends analysis includes additional published articles from January 2013 to August 2015, selected using the same search strategy. A total of 372 articles were found in MEDLINE. After excluding 38 articles due to lack of full texts, and 75 duplicated articles, a total of 259 articles were included in the full dataset amounting to a total of 790 articles.

B. Overview

The text mining methods presented in this paper focuses on two main tasks, categorizing articles and identifying cancer types. To understand the use of the three standards in RT clinical articles, we need to categorize articles based on the criteria used. Then we would like to identify the type of cancer the clinical articles addressed. Our approach consists of four basic steps: data preprocessing, feature extraction, classifier training, and cancer type identification.

To develop a gold standard for training and validation, a radiation oncology physician manually reviewed all 531 articles from 2010 to 2012 in the training/validation subset and labeled them according to the adverse effect scoring criteria used. This process generated 3 classes, one for each standard. After evaluating the two text mining methods using the training/validation dataset, we applied the more accurate method to analyze the entire set of full articles from 2010 to 2015 in terms of the overall usage trends over these years and also in term of usage of the three standards in different cancer types.

C. Data Preprocessing

Each article was first converted into a simple text document without figures, tables, or references. Second, we applied tokenizer to each document to remove numerals and punctuations transforming each document into a list of sentences, and each sentence tokenized into a list of words. Then, we removed stop words in each document based on the stop-words list. Finally, we applied lemmatization, a WordNet’s built-in function, to group different inflected forms of a word.

D. Feature Extraction

We extracted features for statistical analysis and classifier modeling. Major features include n-gram frequency, term frequency and inverse document frequency. Based on experimental

analysis, we use 4-gram frequency in modeling Naïve Bayes Classifier in the article categorization task. Term frequency is calculated by counting the occurrence of phrases after applying n-gram model to each document. To avoid bias caused by different length of documents, we use the total number of terms in a document to normalize raw term frequency in such document.

E. Classifier Modeling

The goal of document categorization task is to classify all documents based on grading criteria used. The task categorizes each document with one of the three criteria, 'CTCAE/CTC', 'RTOG', and 'LENT-SOMA'. We separately used two classification methods to categorize all documents, one is based on regular expression (RE), and the other is based on Naïve Bayes classifier. Both methods have been shown to work well in text mining tasks [16].

Regular Expression Based Classifier—The regular expressions enable a rule-based classifier. It assumes that the basic feature to differentiate documents is a specific pattern, such as particular characters, words, or patterns of characters. The patterns determine which document uses which criteria. If a document contains strings that match regular expression patterns for one of the three standard criteria, the document is categorized as a sample of that criterion. Some documents may be labeled with more than one criteria class if they contain strings that match more than one set of regular expression patterns.

We use an iterative process to discover regular expressions in three basic steps: 1) Extract text snippets from labeled object articles; 2) Extract keys from snippets; 3) Generate regular expressions. We called this process RE discovery process, which is used to train the regular expression based classifier. A snippet is defined as a sequence of characters that provide semantic information for our categorization task. Tokens refer to any words, numbers, or symbols in snippet. A phrase is defined as a sequence of consecutive tokens. A key is defined as an ordered list of phrases. The keys are critical source to generate general regular expressions. Currently, the last step, RE generation, is done manually. We use Python's 're' module to match articles with generated regular expressions. Following are examples of snippets, tokens, keys, and regular expression.

- *Snippets*: 'All symptoms were scored according to the Common Terminology Criteria for Adverse Effects V3.0'.
- *Tokens*: 'All', 'symptoms', 'were', 'scored', 'according', 'to', 'the', 'Common', 'Terminology', 'Criteria', 'for', 'Adverse Effects', 'V3.0'.
- *Phrases*: 'All symptoms', 'were', 'scored', 'according to', 'Common Terminology Criteria for Adverse Effects', 'V3.0'.
- *Keys*: ['were', 'scored', 'according to'] ['scored', 'according to', 'Common Terminology Criteria for Adverse Effects'].
- *Regular expressions*: [\s+\S]*\s+scored\s+[\s+\S]*\s+(ctcae|(common\s+terminology\s+criteria\s+for\s+adverse\s+effects))\s+(v|version)?\s+\d?

The algorithm for training regular expression based classifier involves the following main steps:

Step 1 Initialization: Select N labeled articles to generate initial regular expressions using the RE discovery process explained above. The N articles consist of 4 groups: 30% randomly from each of the three classes and 10 % randomly from the multiple criteria class. Set the initial regular expressions as current regular expressions $CRE\{r_1,r_2,\ldots, r_n\}$.

Step 2 Refinement: Select N new labeled articles to test current regular expression $CRE\{r_1,r_2,\ldots, r_n\}$ and get F-measure value f as an accuracy measure; refine current regular expressions by applying the regular expression discovery on the misclassified articles including false positive and false negative cases to get the new regular expressions CRE' $\{r_1,r_2,\ldots, r_n\}$; then use the new regular expressions $CRE'\{r_1,r_2,\ldots, r_n\}$ to get a new F-measure f'.

Step 3 Iteration: Iterate Step 2 by setting CRE to CRE' unless f' stops changing significantly, namely the change rate falls below ε, i.e. $(f' - f)/f \leq \varepsilon$.

Step 4 Testing: apply the final regular expressions to test the remaining labeled articles as a validation.

In the experiments reported here we empirically selected N of 10 and ε of 0.01.

Naïve Bayes Classifier—Naïve Bayes Classifier is a classical and effective model for text classification. In the article categorization task, we aim to compare the performance of Naïve Bayes Classifier with that of the regular expression based classifier. We use 5-fold cross-validation to partition training data and testing data. In each round, the training data is consisted of 424 randomly selected documents labeled by a domain expert. The rest of documents are testing data to validate the trained classifier model. After five rounds, each document has four candidate criteria labels. The final class for each document is the most voted candidate criteria.

E. Cancer Type Identification

Cancer type identification is straightforward for MEDLINE articles because most of them have already been tagged with MESH terms that indicate cancer types. For the few remaining articles, we use a dictionary based matching method to identify cancer types in the title and abstract. The look-up dictionary consists of cancer types from the domain expert's annotation, and terms under Neoplasms [C04] of PubMed MESH tree structures.

III. Results

A. Training and Validation of Classifiers

The training of the regular expression based classifier took 4 iterations to complete. We noticed that F-measure reaches a plateau of f = 86.7% in the 3rd iteration shown in. Thus, we used 30 randomly selected articles in total to learn the regular expressions, and we used the remaining 501 articles to test the resulting classifier.

The validation results show the regular expression based classifier to be reasonably accurate with a precision of 84.2% and recall of 85.1%. Compared to the regular expressions based classifier, the Naïve Bayes classifier is worse with a precision of 72.1% and recall of 73.8%, but is still comparable to reported text categorization results [17].

Based on the categorization results, Figure 1 presents the usage trends of adverse effect scoring criteria over the three years in comparison to results from the domain expert. As seen in the figures, the two classifiers show similar trends comparable to those of the domain expert. These results provide an indication that the classifiers have sufficient accuracy for detecting usage trends of adverse effect scoring criteria in clinical articles.

B. Usage Trends of Adverse Effect Scoring Criteria During 2010–2015

Using the more accurate regular expression based classifier, we analyzed the trends of the three standard adverse effect scoring criteria in RT clinical articles since 2010. The overall trends are shown in Figure 2. We observe that CTCAE and RTOG continue to be dominant standards in RT articles, each used by almost half of the articles while the LENT-SOMA criteria are used by a small percentage of articles. During this period, the usage of RTOG remains relatively stable, that of CTCAE trends slightly up, while that of LENT-SOMA trends slightly down.

Next, we analyzed the trends by cancer types. Figure 3 shows the overall usage of the three standard scoring criteria in major cancer types over the past five and half years. One interesting finding from this figure is the strong contrast between lung cancer studies that heavily favor CTCAE and the head and neck cancer studies that clearly favor the RTOG standard. We also notice that LENT-SOMA is not only rarely used, but also used only in select types of cancers, such as the prostate cancer and breast cancer studies. Furthermore, LENT-SOMA is especially not used in lung cancer studies.

IV. Conclusion

We analyzed the usage of the three most commonly used standards in radiation therapy by mining the full text of published literature during 2010–2015. We resorted to mining the full text because the abstract section of clinical articles normally lack details on which standardized criteria were used in scoring adverse effects or normal tissue toxicity after radiation therapy. With the large and growing number of clinical publications, manual analysis of literature is becoming increasingly difficult especially when new questions and more comprehensive analyses are needed. The text mining methods provide an important tool for understanding and improving the standards efficiently and continuously monitoring how standards are used for capturing and reporting adverse effects in practice. The accuracy of the classifiers can be further improved with expanded training dataset and elaborated NLP work.

From an informatics perspective, it is desirable that the research community adopts one standard for all clinical articles in radiation therapy and we believe that this standard should be CTCAE since it is based on the other two standards and is more up to date. We suggest that the CTCAE should be represented as a true ontology so that the relationship between

adverse events, their affected anatomy, the related synonyms, and severity are explicitly represented.

Acknowledgments

This work is supported in part by NIH grant #R21CA161389.

References

1. Trotti A, Colevas AD, Setser A, et al. Development of a comprehensive grading system for the adverse effects of cancer treatment. Semin Radiat Oncol. 2003; 1:176–181.
2. Cox JD, Stetz J, Pajak TF. Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC). Int J Radiat Oncol Biol Phys. 1995; 31:1341–1346. [PubMed: 7713792]
3. Rubin P. Late effects of normal tissues (LENT) consensus conference, including RTOG/EORTC SOMA scales. Int J Radiat Biol Phys. 1995; 31:1035–360.
4. Trotti A, Byhardt R, Stetz J, et al. Common toxicity criteria: version 2.0. an improved reference for grading the acute effects of cancer treatment: impact on radiotherapy. Int J Radiat Oncol Biol Phys. 2000; 47(1):13–47. [PubMed: 10758303]
5. Colevas AD, Setser A. The NCI Common Terminology Criteria for Adverse Effects (CTCAE) v 3.0 is the new standard for oncology clinical trials. ASCO Annual Meeting Proceedings. 2004; 22(Suppl 14):6098.
6. Jiang Y, Yuan L, Wu J, et al. Normal Tissue Toxicity Criteria in Radiation Therapy. Int J Radiat Oncol Biol Phys. 2013; 87(Suppl 2):621–622.
7. Zelefsky MJ, Fuks ZV, Hunt M, et al. High dose radiation delivered by intensity modulated conformal radiotherapy improved the outcome of localized prostate cancer. The journal of urology. 2001; 166(3):876–881. [PubMed: 11490237]
8. Kong FM, Hayman JA, Griffith KA, et al. Final toxicity results of a radiation dose escalation study in patients with non-small-cell lung cancer (NSCLC): Predictors for radiation pneumonitis and fibrosis. Int J Radiat Oncol Biol Phys. 2006; 65(4):1075–1086. [PubMed: 16647222]
9. Colen M, Skiadowski K, Wygoda A, et al. A comparison of two scoring systems for late radiation toxicity in patients after radiotherapy for head and neck cancer. Rep Pract Oncol Radiother. 2005; 10(4):179–192.
10. Denis F, Garaud P, Brardet E, et al. Late toxicity results of the Gortec 94-01 randomized trial comparing radiotherapy with concomitant radiochemotherapy for advanced-stage oropharynx carcinoma: comparison of LENT/SOMA, RTOG/EORTC, and NCI-CTC scoring systems. Int J Radiat Oncol Biol Phys. 2003; 55(1):93–98. [PubMed: 12504040]
11. Van der Laan HP, Van den Bergh A, Schilstra C, et al. Grading-system-dependent volume effects for late radiation-induced rectal toxicity after curative radiotherapy for prostate cancer. Int J Radiat Oncol Biol Phys. 2008; 70(4):1138–1145. [PubMed: 17931794]
12. Krauthammer M, Nenadic G. Term identification in the biomedical literature. J Biomed Inform. 2004; 37(6):512–526. [PubMed: 15542023]
13. Cohen A, Hersh W. A survey of current work in biomedical text mining. Briefings in Bioinformatics. 2005; 6(1):57–71. [PubMed: 15826357]
14. Harpaz R, DuMouchel W, Shah N, et al. Novel data mining methodologies for adverse drug effect discovery and analysis. Clin Pharmacol Ther. 2012; 91(6):1010–1021. [PubMed: 22549283]
15. Bui D, Zeng-Treitler Q. Learning regular expressions for clinical text classification. J Am Med Inform Assoc. 2014; 21(5):850–857. [PubMed: 24578357]
16. Sebastiani F. Machine learning in automated text categorization. ACM computing surveys (CSUR). 2002; 34(1):1–47.
17. Yang Y. An evaluation of statistical approaches to text categorization. Information retrieval. 1999; 1(1-2):69–90.

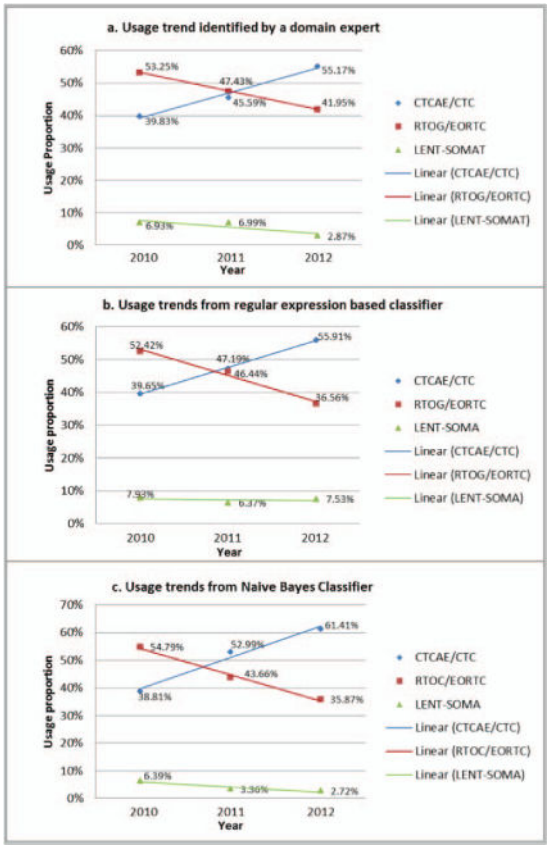

Figure 1.

Comparing usage trends generated by the two text mining methods with those by domain expert using data from 2010 to 2012. Each line shows the proportion of the articles that use a particular criteria.

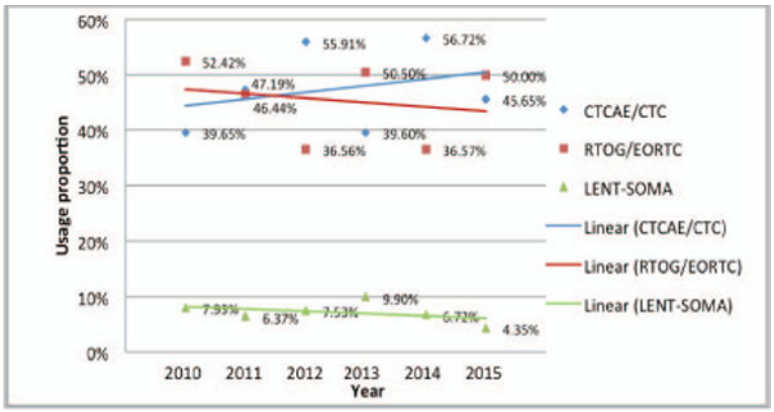

Figure 2.

Usage trends of the three standard adverse effect scoring criteria in RT clinical articles from 2010 to mid 2015 using the regular expression based classification method. Each line shows the proportion of the articles that use a particular criteria.

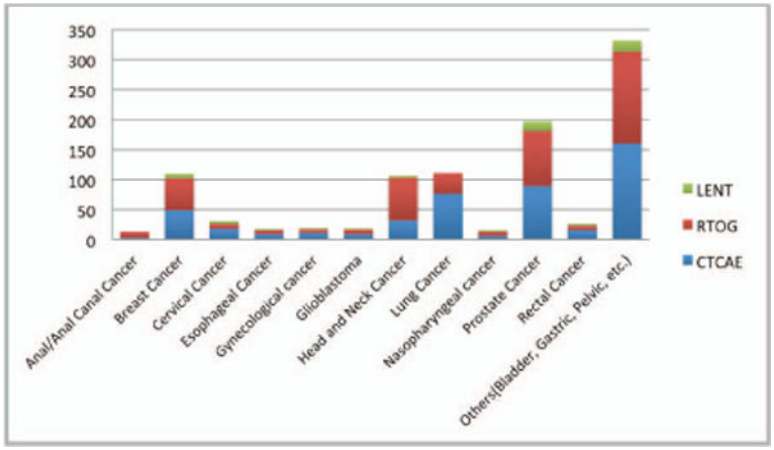

Figure 3.

Usage proportion of the three standard adverse effect scoring criteria by cancer types. The last category (Others) includes all other cancer types each with 10 or fewer articles.