Investigation of Terminology Coverage in Radiology Reporting Templates and Free-text Reports

Yi Hong[*], Jin Zhang[**]

ARTICLE INFO

Article history:
Received 6 March 2014
Revised 7 December 2014
Accepted 30 December 2014

Keywords:
Radiology Reports,
Reporting Templates,
Terminology, RadLex,
Wilcoxon Signed-rank Test

ABSTRACT

The Radiological Society of North America (RSNA) is improving reporting practices by developing an online library of clear and consistent report templates. To compare term occurrences in free-text radiology reports and RSNA reporting templates, the Wilcoxon signed-rank test method was applied to investigate how much of the content of conventional narrative reports is covered by the terms included in the RSNA reporting templates. The results show that the RSNA reporting templates cover most terms that appear in actual radiology reports. The Wilcoxon test may be helpful in evaluatingexisting templates and guiding the enhancement of reporting templates.

1. Introduction

Radiology reports are variable in form, content, and quality. Sistrom and Langlotz (2005) proposed a framework for conceptualizing the reporting process and how it might be improved. This consists of standard language, a structured format, and consistent content. To efficiently represent and exchange radiological knowledge, the Radiological Society of North America (RSNA) has developed a large, freely accessible online library of radiology reporting templates (http://www.radreport.org), which is integrated with the Extensible Markup Language (XML) documents, metadata, standardized biomedical ontologies, and healthcare industry standards. Radiologists and other users can browse, retrieve, and download templates in text or XML-encoded format (Kahn et al., 2009).

As of December 2014, 268 reporting templates have been released to the RSNA reporting template library. These templates are intended to serve as examples of “best practice” to guide radiologists in formulating standardized radiology reports (Langlotz, 2006). Each reporting template has associated metadata, including information about the template's title, creator, subject, description, and date.

The elements of the reporting templates have been mapped to corresponding terms in standardized biomedical ontologies such as the RadLex® radiology lexicon and the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT®) by employing a semi-automated mapping tool called RadMap (Hong et al., 2012).

[*] Department of Production Engineering, MedeAnalytics, Emeryville, CA, USA (yi.hong@medeanalytics.com)
[**] School of Information Studies, University of Wisconsin Milwaukee, Milwaukee, WI, USA (jzhang@uwm.edu)
International Journal of Knowledge Content Development & Technology, 5(1): 5-14, 2015.
http://dx.doi.org/10.5865/IJKCT.2015.5.1.005

There have been quite a few studies accompanied on structured radiology reports (Mamlin, Heinze, and McDonald, 2003; Langlotz, 2009; Taira, Soderland, and Jakobovits, 2001; Bozkurt and Kahn, 2012) and RadLex® development (Hazen, et al., 2011). But only a small number of studies have been conducted to evaluate the coverage of biomedical terminologies for radiological reports, and most measures were developed through traditional data analysis approaches (Langlotz and Caldwell, 2002; Woods and Eng, 2013; Heilbrun, 2013). Few studies have explored terminology coverage in both reporting templates and free-text radiology reports using statistical methods. Ourresearch aimed at addressing this gap by examining how many terms in free-text reports are covered in the RSNA reporting templates and providing first-hand information to guide development of radiology reporting templates from users' perspective.

2. Methods

To compare term occurrences in free-text radiology reports and RSNA reporting templates, the Wilcoxon signed-rank test method, a non-parametric statistical hypothesis test (Woolson, 2007), was applied to investigate how much of the content of conventional narrative reports is covered by the terms included in the RSNA reporting templates.

Research data were collected from the RSNA reporting template library and a sample of 8,275 consecutive, de-identified free-text radiology reports from a Radiology Department of an academic medical center. While more than 285,000 imaging procedures are performed each year for inpatients and outpatients, the Radiology Department offers the complete spectrum of state-of-the-art imaging techniques and imaging-guided, minimally invasive procedures. All of the free-text reports were created by voice dictation, and were transcribed either manually or by a speech recognition system. The report text represented final, approved report content, and consisted of the procedure name, narrative ("findings") section, report impression, and other information. The study protocol contains one week's de-identified radiology reports that were approved by the appropriate Institutional Review Board (IRB), and the study was performed in compliance with the Health Insurance Portability and Accountability Act (HIPAA). Among the templates and free-text reports, computed tomography (CT), diagnostic radiology (DX), magnetic resonance imaging (MR), nuclear medicine (NM), and ultrasound (US) reporting templates and corresponding free-text reports were chosen as samples to carry out the analysis. These radiology examinations were selected to conduct this analysis because they are the most common exams and the most frequently performed at the institution. They have been used in previous studies, and so we may compare the test results of this study with that of previous studies and verify the reliability and consistency of the results From the sample data sets, 99 (CT, 35; DX, 17; MR, 33; NM, 2; US, 12) reporting templates and 6,410 corresponding free-text reports (CT, 1817; DX, 2932; MR, 550; NM, 646; US, 465) were selected to accomplish the study.

In this study, a set of the terms extracted from the free-text reports was compared with another set of the terms extracted from the RSNA reporting templates. Each term was associated with a raw frequency score that indicated how many times this term appeared in the sample data set. These raw frequency scores in each set were normalized prior to comparison analysis so that the negative impact of the different sample sizes of free-text reports and RSNA reporting templates on the comparison results was minimized.

2.1. Statements of the hypotheses

Since the RSNA reporting templates have been designed to represent the information in free-text reports with a consistent format, terms in the RSNA reporting templates are supposed to cover the content of the corresponding free-text reports. We presumed that there was no significant difference between the RSNA reporting templates and the free-text radiology reports in terms of terminology coverage.The following hypothesesas shown in Table 1 were tested with paired samples to prove our assumption. For simplicity, the alternative hypotheses for all hypotheses are omitted.

Table 1. Statements of hypotheses for Wilcoxon statistical study

Ha1: There is no significant difference between the CT Head exam reporting templates and the free-text CT Headexam reports in terms of general terminology coverage	Hb1: There is no significant difference between the CT Head exam reporting templates and the free-text CT Headexam reports in terms of RadLex® vocabulary coverage
Ha2: There is no significant difference between the DX Chest exam reporting templates and the free-text DX Chest exam reports in terms of general terminology coverage	Hb2: There is no significant difference between the DX Chest exam reporting templates and the free-text DX Chest exam reports in terms of RadLex® vocabulary coverage
Ha3: There is no significant difference between the MR Spine exam reporting templates and the free-text MR Spine exam reports in terms of general terminology coverage	Hb3: There is no significant difference between the MR Spine exam reporting templates and the free-text MR Spine exam reports in terms of RadLex® vocabulary coverage
Ha4: There is no significant difference between the NM Bone Scan reporting templates and the free-text NM Bone Scan reports in terms of general terminology coverage	Hb4: There is no significant difference between the NM Bone Scan reporting templates and the free-text NM Bone Scan reports in terms of RadLex® vocabulary coverage
Ha5: There is no significant difference between the US Abdomen exam reporting templates and the free-text US Abdomen exam reports in terms of general terminology coverage	Hb5: There is no significant difference between the US Abdomen exam reporting templates and the free-text US Abdomen exam reports in terms of RadLex® vocabulary coverage

In the Wilcoxon test, the independent variable is the term list. The dependent variable is the ranking score of a term on a term list. The ranking score of a term was measured by its position on a term list ranked by term frequency. The significance level α was set to 0.01 for all the proposed hypothesis tests. Thus if the generated p-value is equal to or smaller than the significance level (α), the null hypothesis is rejected. Otherwise, the null hypothesis is accepted. This judgment rule was used to examine all results.

In order to test the above hypotheses, the terms extracted from the RSNA reporting templates were collected as a baseline data set while the terms from free-text radiology reports were extracted and collected as a corresponding compared data set. With the extracted terms, the following term lists were formed:

1. All the meaningful terms and their frequencies in reporting templates and free-text radiology reports of CT Head, DX Chest, MR Spine, NM Bone Scan, and US Abdomen exams;

2. The RadLex® terms and their frequencies in reporting templates and free-text radiology reports of CT Head, DX Chest, MR Spine, NM Bone Scan, and US Abdomen exams.

2.2. Process of paired samples

For the reporting templates, the terms were extracted from the XML stream of CT Head, DX Chest, MR Spine, NM Bone Scan, and US Abdomen reporting templates and their frequencies in each specialty were calculated respectively. The extracted terms and their frequencies formed five term lists of CT Head, DX Chest, MR Spine, NM Bone Scan, and US Abdomen exams. All the terms in each term list were sorted by their frequencies in descending order.

For the full-text reports, all terms were parsed and extracted from the CT Head, DX Chest, MR Spine, NM Bone Scan, and US Abdomen reports using Apache Lucene program (The Apache Software Foundation, 2012). Apache Lucene is a full-featured open source text search engine, offering fielded searching and multiple-index searching with powerful query types such as phrase queries, wildcard queries, proximity queries, range queries, and more. Meaningless words, pronouns, numerals, and adverbs such as “a”, “the”, “of”, “with”, “and”, “about”, “also”, “along”, “been”, “could”, “from”, and “have” were identified as stop words and removed from the extracted term set. Stop words are words which are filtered out prior to, or after, processing of natural language data. After removing stop words, the frequency of each remaining term was calculated and all the terms were sorted by their frequencies in descending order.

The template term list and full-text term list in each exam type were paired to create matched samples. For each original term list, a ranked term list was formed after the terms were sorted by their frequencies in descending order. Each term on the ranked list obtained a ranking score based on its position on the list. The ranking scores were assigned to the terms in the following way: The first term received a ranking score 1, the second term received a ranking score 2, the third term received a ranking score 3, and so on. If a group of terms on a term list had the same frequency, the average ranking score of these terms was assigned to each of the terms. For instance, for a group of terms {T1, T2, T3, ⋯, Tn} that have the same frequency *f*, the ranking score *S* for these keywords is defined as:

$$S = \frac{i + (i+1) + (i+2) + ... + (i+n)}{n} = \frac{(2i+n)}{2} \qquad (1)$$

The strategy to handle the terms with the same frequency may avoid unnecessary bias in the later data process. Otherwise, ranking scores for the terms with the same frequency might be imprecise as each of the terms with the same frequency received a different ranking score without the process. Through the ranking process, the term sets extracted from reporting templates and free-text reports came up a pair of ranked term list where each term had a ranking score. It is apparent that ranking scores are ordinal data, which is appropriate for Wilcoxon statistical analysis to test the difference between the paired samples.

The ranking scores on the term list from the free-text reports were normalized by converting the absolute ranking scores to the relative scores on an abridged ranked term list prior to comparison analysis so that the negative impact of the term size difference of the templates and free-text reports was eliminated. The absolute ranking score of a term from the free-text reports is defined by its position on the original ranked term list where all terms are included. The relative ranking score of a term refers to its position on an abridged ranked term list where only those terms that appear on the term list of reporting templates are kept. More precisely, the terms that did not appear on the term list of reporting templates were excluded on the abridged ranked list, while the terms that appeared in the reporting templates but did notshow up in the free-text reports were added to the bottom of the abridged ranked term list with the frequencies of 0, which ranking scores were the lowest on the abridged ranked list. Adding these terms to the abridged ranked list is critical because it makes the abridged free-text ranked term list and the template ranked term list matched pairs. These pairs were used for the later Wilcoxon test analysis.

Notice that although the ranking scores of the terms on the abridged ranked term list are different from those of the terms on the original list, the ranking relationships among the terms on the abridged ranked list are the same as those among the terms on the original list. That is, the normalization does not change the ranking nature among the involved terms. Without the normalization, a ranking score of a term involved in the comparison would be affected by the number of the terms that were not involved on the original term list in the comparison, and the comparison results between the terms in the reporting templates and those in the free-text reports would be indistinct. With the normalization, the comparison results would be more reasonable.

The abridged ranked term list of free-text reports and the ranked term list of the reporting templates were combined into a new table that was used as an integrated ranking table for the Wilcoxon statistical analysis. The data are ranked from the smallest absolute value to the largest absolute value. In the case of a tie, ranks of the tied terms are added together and divided by the number of ties. Assuming there are 20 terms of the frequency 0 on the free-text term list, the ranks corresponding to these terms are 61-80. The sum of these ranks is 141. After dividing by the number of ties, a mean rank of 70.5 is assigned to all of these 20 terms.

The relationships among the involved lists are illustrated in Figure 1.

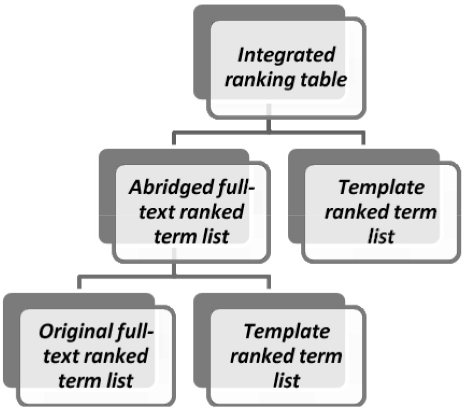
Integrated ranking table
Abridged full-text ranked term list
Template ranked term list
Original full-text ranked term list
Template ranked term list

Fig. 1. Relationships among the involved term lists

If the number of a sample is larger than 20, the values for the Wilcoxon statistic tend to form a normal distribution and the normal distribution table should be used for decision making on whether a proposed hypothesis is accepted or not(Gravetter and Wallnau, 2013). As a result, in addition to the resultant p-values, the Z-scores are provided in the results of the Wilcoxon statistical tests.

Both Z-score and p-value can be used to judge a hypothesis and the testing results from the two methods are the same. After the data sets were converted into ranked lists by the normalized ranking scores, the Wilcoxon signed-ranks test incorporated with SPSS (IBM SPSS Statistics 20.0) was applied to the data sets respectively to evaluate the difference between two ranked term lists. Consequently, the proposed hypotheses were tested. The significant level was set to 0.01 for this study. In other words, if the generated *p*-value from the Wilcoxon test is larger than 0.01, then the corresponding hypothesis is accepted. Otherwise, the hypothesis is rejected.

3. Results

The results of the Wilcoxon tests were obtained using the SPSS statistical program. SPSS produces a rank table reporting the sample size (*N*), the mean rank, and the sum of the ranks for both positive and negative ranks. The Test Statistics table reports the outcome of the test, including the *z*-score approximation and the level of significance (*p* value).

Table 2. Summary of the Wilcoxon test results

Null hypothese	Sample size (N)	*P* value (2-tailed)	*Z* score
Ha1: CT_Head	183	.000	-4.778
Hb1: CT_Head_RadLex	89	.028	-2.204
Ha2: DX_Chest	159	.000	-3.805
Hb2: DX_Chest_RadLex	74	.093	-1.680
Ha3: MR_Spine	162	.761	-.304
Hb3: MR_Spine_RadLex	103	.969	-.038
Ha4: NM_Bone	70	.531	-.627
Hb4: NM_Bone_RadLex	53	.599	-.526
Ha5: US_Abdomen	78	.746	-.324
Hb5: US_Abdomen_RadLex	76	.799	-.254

Of 10 hypotheses, 8 were accepted and 2 were rejected, which means that most free-text reports' terms are represented in the reporting templates. In other words, among 10 paired samples, 8 show little difference between the reporting templates and the free-text radiology reports in terms of terminology coverage. All the hypotheses related to RadLex® coverage were accepted, which means most RadLex® terms appearing in the reporting templates also show up in their corresponding free-text radiology reports.

Table 3. Summary of the proposed hypothesis test results

Comparison categories	Hypotheses	Test results (α =0.01)	
Hypotheses related to general terminology coverage	Ha1	Rejected	
	Ha2	Rejected	
	Ha3		Accepted
	Ha4		Accepted
	Ha5		Accepted
Hypotheses related to RadLex® vocabulary coverage	Hb1		Accepted
	Hb2		Accepted
	Hb3		Accepted
	Hb4		Accepted
	Hb5		Accepted
	Total	2	8

From Table 2 and Table 3, we may tell that the terms in CT Head and DX Chest reporting templates matches to relatively small percentages of the terms that appear in free-text radiology reports, while MR Spine, NM Bone Scan, and US Abdomen reporting templates capture more terms in the free-text reports. In other words, terms that frequently appear in the reporting templates of CT Head and DX Chest exams occur less frequently within the actual free-text reports while terms that appear in the reporting templates of MR Spine, NM Bone Scan, and US Abdomen exams occur more frequently within the corresponding free-text reports. Meanwhile, the results show that RadLex® terms appearing in the reporting templates better cover the terms that appear frequently in free-text radiology reports.

The results verify our findings in previous studies, which found an overall rate of 67%reporting elements derived from the RSNA reporting templates were matched to RadLex® terms (Hong et al., 2012), and suggest that the concepts that appear in the reporting templates occur frequently within free-text clinical reports and the reporting templates provide useful coverage of the "domain of discourse" in radiology reports (Hong and Kahn, 2013).

4. Discussion

The Wilcoxon signed rank test is a frequently used nonparametric test for paired data based on independent units of analysis (Rosner, Glynn, and Lee, 2006). It is used to compare two related samples to assess whether their population means differ. Prior to this study, no studies have applied the Wilcoxon test method to exam term occurrences in both reporting templates and free-text reports. The Wilcoxon signedrank test was chosen for the hypothesis test because it is good for the paired t-test when the population cannot be assumed to be normally distributed and is usually used to examine the difference between two treatments in a population with a repeated-measure design (Rey and Neuhäuser, 2011), which deals with the data sets of the free-text reports and RSNA reporting templates very well.

The Wilcoxon signed rank test calculates the difference between each set of pairs. Forming the matched pairs is a vital step that significantly influences the test results. The results of a Wilcoxon test only make sense when the paired samples are matched. In our study, a set of the terms extracted from the RSNA reporting templates of CT Head, DX Chest, MR Spine, NM Bone Scan, and US Abdomen exams was compared with its matching set of the terms extracted from the free-text reports with the ranking scores. That is, a ranked term list of CT Head reporting templates was paired with a ranked term list of full-text CT Head reports, a ranked term list of DX Chest reporting templates was paired with a ranked term list of full-text DX Chest reports, and so on.

Since the terms extracted from the full-text reports are much more than the terms extracted from the reporting templates and the term occurrences in the free-text reports are much higher than that in the templates, the term frequencies were converted to the ranking scores and all the terms appearing in reporting templates with their ranking scores in the templates and free-text reports were chosen to create the ranked term lists so that the paired term lists would be matched with each other. After the normalization process, the ranking scores might be the same for a term with a frequency of 1 in a template ranked term list and the same term with a frequency of 300 in a matched free-text ranked term list. The terms that appear in free-text but do not appear in templates were excluded in the test as the main purpose of this study is to find out whether the terms in the templates occur in the free-text reports and how frequently they are used in the actual reports.

The Wilcoxon study has some limitations. First, the paired samples were identified and selected from only one common exam in each of the five specialties (CT, DX, MR, NM, and US). If more exams and more specialties were involved, the comparison results would be more sound and robust. Second, the sample size of each set of pairs is different, which might affect the reliability of test results.

Our future research includes, but is not limited to, increasing sample size of investigated data with more types of radiological exams, making paired samples more equivalent, further investigating the reporting terms that were excluded from the Wilcoxon test, finding out the frequencies of these terms in the free-text reports, and incorporated most frequently used ones into RSNA reporting templates.

To maintain their utility as "best practices" for the radiology community, both RadLex® and RSNA reporting templates should continue to grow (Shore, Rubin, and Kahn, 2012). The use of standardized biomedical terminologies in reporting templates can reduce communication errors.The unmatched terms (the terms occurring in the free-text reports but not in the templates) from the reporting templates provide a "bottom-up" inventory of entities used in radiology reports, which most frequently occurred ones might be served as candidate terms for inclusion in the templates and therefore enhancestandardized terminology coverage of the templates.

5. Conclusion

The Wilcoxon test results show that the RSNA reporting templates cover most terms that appear in actual free-text radiology reports. The comparison analysis of terminology coverage between free-text reports and reporting templates filled a research gap using statistical methods to explore term occurrences in radiology reporting. By identifying terms that occur frequently in radiology reports, the Wilcoxon test may be helpful in evaluating existing templates, and to guide the development of both RadLex® and reporting templates.

Acknowledgments

This research was supported in part by the National Institute of Biomedical Imaging and Bioengineering (NIBIB). We thank the RSNA Radiology Informatics Committee for leading and supporting the radiology reporting initiative, and we acknowledge the many RSNA volunteers who helped develop the reporting templates.

References

Bozkurt, S., & Kahn Jr, C. E. (2012). An open-standards grammar for outline-style radiology report templates. *Journal of digital imaging*, 25(3), 359-3.

Gravetter, F., & Wallnau, L. (2013). *Essentials of statistics for the behavioral sciences*. Cengage Learning.

Hazen, R., Van Esbroeck, A. P., Mongkolwat, P., & Channin, D. S. (2011). Automatic extraction of concepts to extend RadLex. *Journal of digital imaging*, 24(1), 165-169.

Heilbrun, M. E. (2013). Evaluating RadLex and Real World Radiology Reporting: Are We There Yet?. *Academic radiology*, 20(11), 1327-1328.

Hong, Y., & Kahn Jr, C. E. (2013). Content analysis of reporting templates and free-text radiology reports. *Journal of digital imaging*, 26(5), 843-849.

Hong, Y., Zeng, M. L., Zhang, J., Dimitroff, A., & Kahn, Jr, C. E. (2013). Application of standardized biomedical terminologies in radiology reporting templates. *Information Services and Use*, 33(3), 309-323.

Hong, Y., Zhang, J., Heilbrun, M. E., & Kahn, C. E. (2012). Analysis of RadLex® Coverage and Term Co-occurrence in Radiology Reporting Templates. *The Journal of Digital Imaging*, 25(1), 56-62.

Kahn Jr, C. E., Langlotz, C. P., Burnside, E. S., Carrino, J. A., Channin, D. S., Hovsepian, D. M., & Rubin, D. L. (2009). Toward Best Practices in Radiology Reporting 1. *Radiology*, 252(3), 852-856.

Langlotz, C. P. (2006). RadLex: A new method for indexing online educational materials 1. *Radiographics*, 26(6), 1595-1597.

Langlotz, C. P. (2009). Structured Radiology Reporting: Are We There Yet? 1. *Radiology*, 253(1), 23-25.

Langlotz, C. P., & Caldwell, S. A. (2002). The completeness of existing lexicons for representing radiology report information. *Journal of digital imaging*, 15, 201-205.

Mamlin, B. W., Heinze, D. T., & McDonald, C. J. (2003). Automated extraction and normalization of findings from cancer-related free-text radiology reports. In *AMIA Annual Symposium Proceedings* (Vol. 2003, p. 420). American Medical Informatics Association.

Rey, D., & Neuhäuser, M. (2011). Wilcoxon-Signed-Rank Test. In *International Encyclopedia of Statistical Science* (pp. 1658-1659). Springer Berlin Heidelberg.

Rosner, B., Glynn, R. J., & Lee, M. L. T. (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*, 62(1), 185-192.

Shore, M. W., Rubin, D. L., & Kahn Jr, C. E. (2012). Integration of imaging signs into RadLex. *Journal of digital imaging*, 25(1), 50-55.

Sistrom, C. L., & Langlotz, C. P. (2005). A framework for improving radiology reporting. *Journal of the American College of Radiology*, 2(2), 159-167.

Taira, R. K., Soderland, S. G., & Jakobovits, R. M. (2001). Automatic Structuring of Radiology Free-Text Reports 1. *Radiographics*, 21(1), 237-245.

The Apache Software Foundation. (2012). Apache Lucene Core. Available at <http://lucene.apache.org/core> Retrieved 2014.12.15.

Woods, R. W., & Eng, J. (2013). Evaluating the Completeness of RadLex in the Chest Radiography Domain. *Academic radiology*, 20(11), 1329-1333.

Woolson, R. F. (2007). Wilcoxon Signed-Rank Test. *Wiley Encyclopedia of Clinical Trials*.