TITLE

Increasing the Efficiency on Producing Radiology Reports for Breast Cancer Diagnosis by means of Structured Reports: A comparative Study

AUTHORS

J. Damian Segrelles Quilis †*,
Rosana Medina García ‡,
Ignacio Blanquer Espert † ⍰
Luis Martí Bonmatí § ⍰

† Instituto de Instrumentación para Imagen Molecular (I3M). Centro mixto CSIC – Universitat Politècnica de València – CIEMAT, camino de Vera s/n, 46022 Valencia, Spain

‡ Radiology Department, Dr. Peset University Hospital, Valencia, Spain

§ Medical Imaging Department, La Fe University and Polytechnic Hospital, Valencia, Spain

⍰ GIBI2[30] – Biomedical Imaging Research Group, La Fe University and Polytechnic Hospital, Valencia, Spain

* Corresponding author:

Full institutional mailing address: Institute for Molecular Imaging Technologies (I3M), Universitat Politècnica de València (UPVLC), Camino de Vera S/N 46022 Valencia, Spain
Tel: +34-963877007, ext 88254

Email: dquilis@dsic.upv.es

STRUCTURED ABSTACT

Background. Radiology reports are commonly written on free-text using voice recognition devices. Structured reports (SR) have a high potential but they are usually considered more difficult to fill-in so their adoption in clinical practice leads to a lower efficiency. However, some studies have demonstrated that in some cases, producing SRs may require shorter time than plain-text ones. This work focuses on the definition and demonstration of a methodology to evaluate the productivity of software tools for producing radiology reports. A set of SRs for breast cancer diagnosis based on BI-RADS have been developed using this method. An analysis of their efficiency with respect to free-text reports has been performed.

Material and Methods. The methodology proposed compares the Elapsed Time (ET) on a set of radiological reports. Free-text reports are produced with the speech recognition devices used in the clinical practice. Structured reports are generated using a web application generated with TRENCADIS framework. A team of six radiologists with three different levels of experience in the breast cancer diagnosis was recruited. These radiologists performed the evaluation, each one introducing 50 reports for mammography, 50 for ultrasound scan and 50 for MRI using both approaches. Also, the Relative Efficiency (REF) was computed for each report, dividing the ET of both methods. We applied the T-Student (T-S) test to compare the ETs and the ANOVA test to compare the REFs. Both tests were computed using the SPSS software.

Results. The study produced three DICOM-SR templates for Breast Cancer Diagnosis on mammography, ultrasound and MRI, using RADLEX terms based on BIRADs 5th edition. The T-S test on radiologists with high or intermediate profile, showed that the difference between the ET was only statistically significant for mammography and ultrasound. The ANOVA test performed grouping the REF by modalities, indicated that there were no significant differences between mammograms and ultrasound scans, but both have significant statistical differences with MRI. The ANOVA test of the REF for each modality, indicated that there were only

significant differences in Mammography (ANOVA p =0.024) and Ultrasound (ANOVA p =0.008). The ANOVA test for each radiologist profile, indicated that there were significant differences on the high profile (ANOVA p =0.028) and medium (ANOVA p =0.045).

Conclusions. In this work, we have defined and demonstrated a methodology to evaluate the productivity of software tools for producing radiology reports in Breast Cancer. We have evaluated that adopting Structured Reporting in mammography and ultrasound studies in breast cancer diagnosis improves the performance in producing reports.

KEYWORDS: Structured Reporting, DICOM-SR, BI-RADS, Breast Cancer

1. INTRODUCTION AND BACKGROUNG

Today, radiology exams are key in the diagnose of oncological diseases. The findings spot on radiology reports support crucial treatment decisions for the outcome of patient's health. Radiology reports are managed by means of Radiology Information Systems (RIS) [1]. The images from a radiology exam as well as other similar objects concerning the exams are held in Picture Archiving and Communication System (PACS) [2], using Digital Imaging and Communications in Medicine (DICOM) [3] standard to distribute, store, exchange and display DICOM objects.

Radiology reports are typically documents written in free text, which can be transcribed directly from a keyboard, a dictation or a voice recognition system. Those free-text documents are then stored in the RIS system [4, 5, 6]. Nevertheless, the use Structured Reports (SR) with standardized terminologies and definitions of the terms is increasing. The adoption of SRs in a hospital depends on the outcome of the analysis of pros (i.e. the increase of productivity) and cons (i.e. point-and-click Reporting is Impractical for complex cases) of their use [7]. Studying the benefits at several levels (user's profile and complexity of the case for example) will lead to a better understanding of the benefits.

In this sense, moving from free-text radiological reports to structured standardized data descriptions fits into European research data trends on the urgent need to improve the reusing of scientific data [8].

DICOM Structured Reporting (DICOM-SR) [9, 10] allows to create SRs templates as a tree structure, where each node from the tree is a report field coded using a standardized terminology. DICOM-SR templates enable developers to define additional characteristics for each report field, such as multiplicity (a field can have multiple values), mandatory (the field is obligatory) and conditional fields (the availability of the field depends on other field's value). Several studies have suggested that DICOM-SR is convenient and useful [11]. In the scope of this paper, we use DICOM-SR templates to code and validate the SRs written by a radiologist.

In our work, we use DICOM-SR as the format to code the reports for its storage in the PACS, despite that they are translated into XML files within the application. We also leverage DICOM-SR to code the new templates for interoperability purposes.

In breast cancer, the Breast Imaging Reporting and Data System (BI-RADS) [12] terminology standardizes the reporting and data collection of Mammography, Ultrasound and MRI for breast cancer diagnosis. All the terms defined in BI-RADs are coded in the RADLEX [13][14] lexicon which was built by the Radiological Society of North America (RSNA), representing a unified language of radiology terms for standardized indexing and retrieval of imaging information resources. Furthermore, we consider that initiatives such as the Radiology Reporting Initiative[1] led by the RSNA, which promote the creation of libraries of clear and consistent SR templates, are very important to improve quality and interoperability of clinical practice. In the knowledge of the authors there is no implementation of the whole set of term and procedures from BI-RADS for breast cancer diagnosis rather than the one presented in this article.

Recent studies have concluded that SR reports have better content and greater clarity than conventional reports [15, 16]. However, despite the existence of BI-RADS and standards as DICOM-SR to define and code SRs, free-text is still the most common used method for breast cancer diagnosis reports. However, the use BI-RADS is quite extended, but the BI-RADS terms are entered in the free-text report documents due to the perceived great difficulty and lower efficiency of structured reports in clinical practice. Although, some studies show that the use of structured reports may lead to lower productivity with respect to conventional reports [17, 18], other studies have demonstrated that the use of structured reports in given areas related to medical imaging, such as plain chest films [19] require a shorter time to produce the reports than free-text methods.

[1] Radiology Reporting Initiative. https://www.rsna.org/reporting_initiative.aspx

The contributions of this paper are first, a description of a methodology to evaluate the productivity of software tools to generate radiological reports considering the user's profile and different modalities. Also, we contribute with three DICOM-SR templates (for mammography, ultrasound and MRI) for the diagnosis of breast cancer, based on BI-RADS. Second, we analyse the conditions were we expect a shorter Elapsed Time (ET) on the writing of structured report by means of web interfaces (Method 2) generated from these templates, with respect to free-text reports filled-in with conventional speech recognition (Method 1). In this work, the structured report web interfaces are provided by TRENCADIS [20], an application that automatically generates web interfaces from DICOM-SR templates, implementing the integrity restrictions defined in the templates [20]. Furthermore, the influence of the report method in the elapsed time (ET) of the generation process of the reports in each modality and radiologists experience will be also evaluated.

2. MATERIAL AND METHODS

Figure 1 shows the phases of the experiment and the dependences of their activities, which are described below.

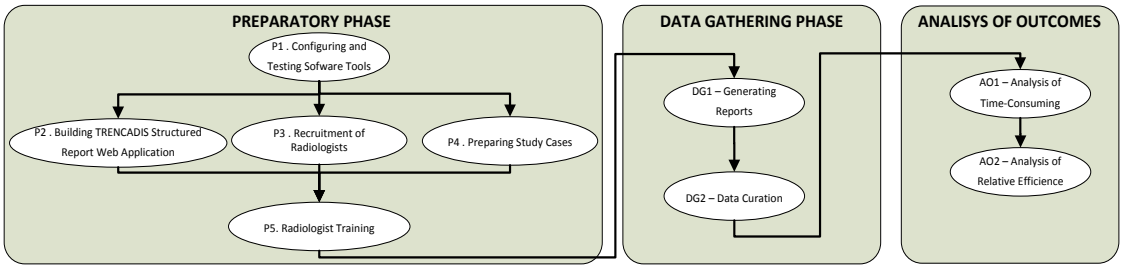

Figure 1. Time dependences of scheduled activities in the study. P, Preparatory phase; DG, Data Gathering phase; and AO, Analyses of Outcomes.

2.1. Preparatory Phase

This phase is composed of five activities which are outlined below.

2.1.1. Configuration and Testing of Software Tools

The target of this activity was to configure and test the software for performing the study. We based on the workstations used in clinical routine for viewing and processing the images. These workstations are provided with a speech recognition software for the transcription of the free-text reports. The Structured Reports were completed using TRENCADIS web application.

- **Free-text reports with speech recognition (Method 1).** The Orion Clinic [21], current system used in all the public hospitals in our region, was used for the generation of the radiological reports. It consists on a text processor that incorporate a speech recognition system.
- **Structured Reports with TRENCADIS Web Interface (Method 2).** The radiologists used a Google Chrome web browser to access the web application built by the TRENCADIS framework [20] for generating DICOM-SR structured reports. It was setup in the same workstations where ORION clinic was installed, setting up a direct access pointing out to the web application.

2.1.2. Building TRENCADIS Structured Report Web Application

To compare the increased efficiency of structure reports, radiologists were provided with a web application for creating structured reports. The application fulfils the integrity restrictions and relations defined in the DICOM-SR templates [9], such as cardinality, mandatory fields and data type (e.g. Text, Numeric, Code).

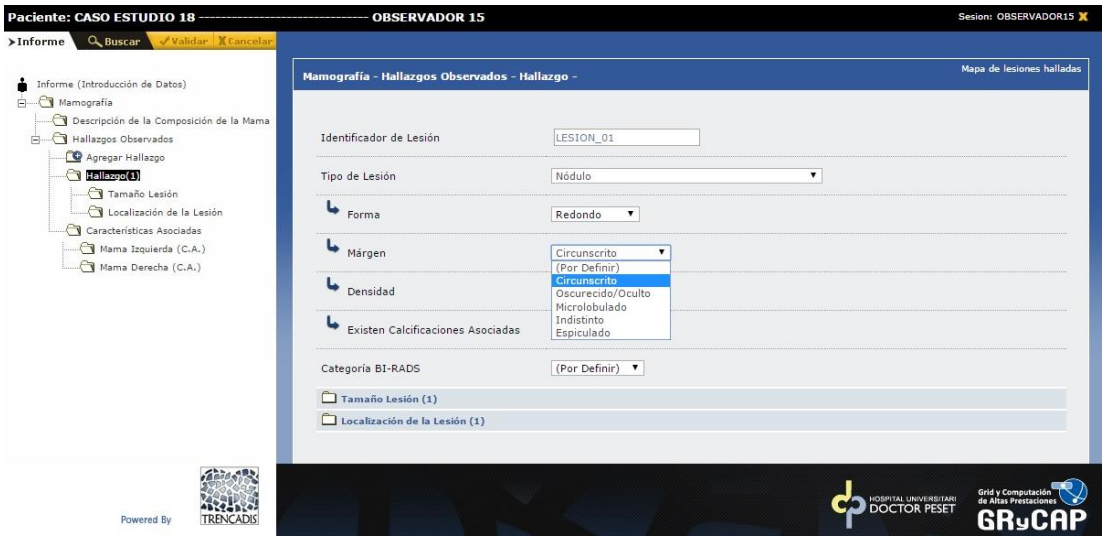

Figure 2. Screenshot of the TRENCADIS web interface (in Spanish) generated from the Mammography DICOM-SR template.

For this purpose, TRENCADIS [20] framework can automatically generate web interfaces from DICOM-SR templates. These TRENCADIS web interfaces will show up warnings to the user if an integrity restriction is not fulfilled, forcing the user to correct it.

DICOM-SR templates are defined before generating the TRENCADIS web interfaces (see Figure 2), following an XML schema (see Figure 3) that it is generated through a graphical tool (see Figure 4).

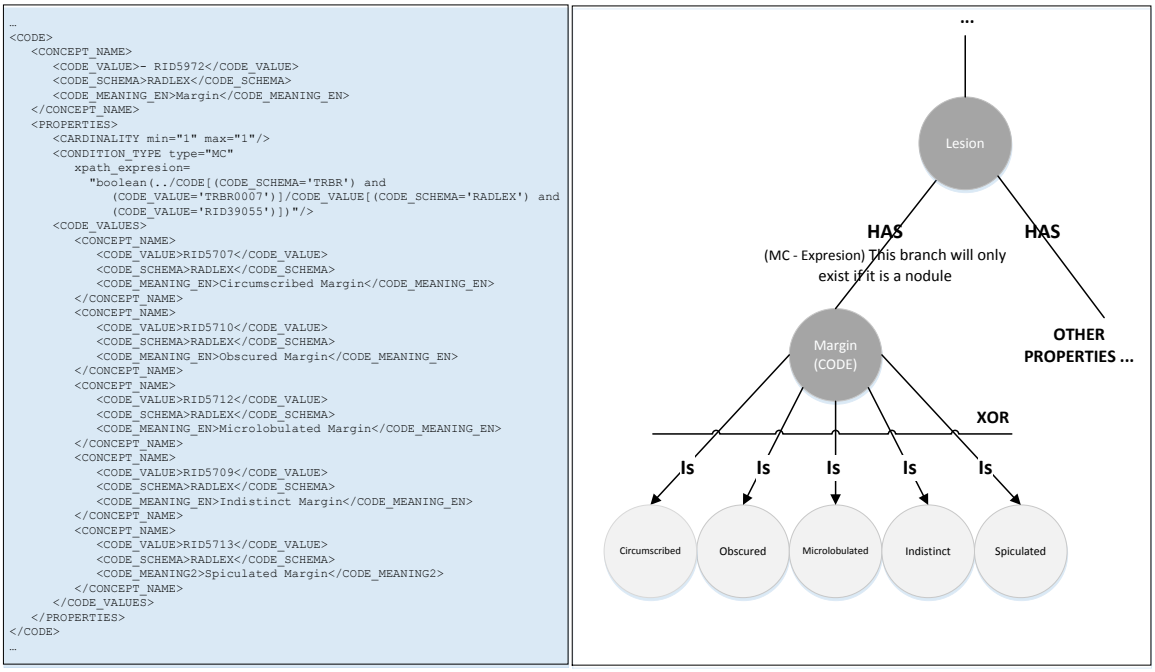

Figure 3. The left side shows a fragment of an XML document describing the field "Density of a Breast Mass", which is defined in the Mammography DICOM-SR template. The right side shows a schematic overview of the structure and semantic restrictions of the DICOM-SR template part that represents the same information as the XML document.

Three DICOM-SR templates were defined in the study (mammograms, ultrasound and MRI), following the structure and terms recommended by 5th edition BI-RADS standard. In the templates, the BI-RADS terms were coded in RADLEX lexicon. However, to define the structure of the reports recommended by BI-RADS, new terms not present in RADLEX were added. Also, other terms specific to the study such as the starting and ending time of reports needed for

computing the Elapsed Time, were included. All the terms that are missing in RADLEX were defined using our own terminology named TRENCADIS BREAST (TRBR).

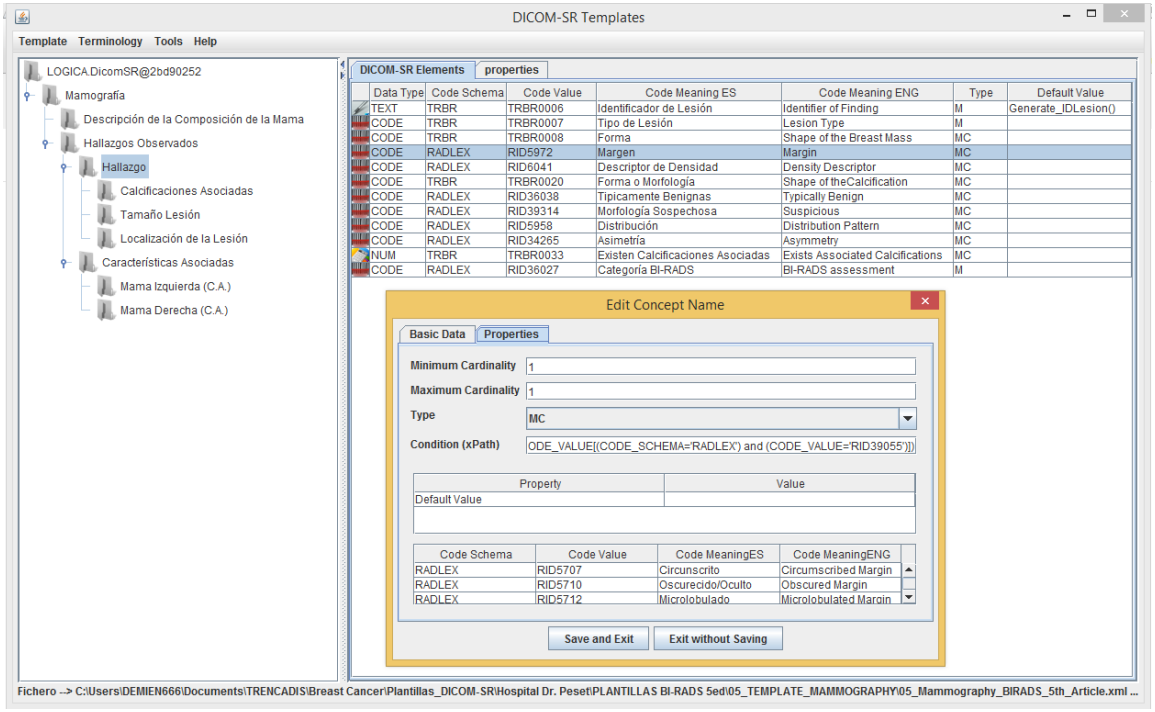

Figure 4. Screenshot of the TRENCADIS framework to create/update a Mammography DICOM-SR template. The templates supports two languages (Spanish and English).

2.1.3. Recruitment of Radiologists

In order to have significant results, a team of radiologist were recruited to perform the evaluation of the tools. Radiologists had different levels of experience in the breast cancer imaging and they were asked to complete the reports using the two exposed methods.

The team was recruited on a voluntary basis. Two members from each one of the three following profiles were randomly chosen:

- **High experience**. Radiologists with more of 8 years of experience in breast cancer diagnosis.
- **Intermediate experience**. Resident physicians who had been trained for four years in general radiology and during one-month per year in breast cancer diagnosis.

- **Low experience**. First-year resident physicians who had been trained for only one month in breast cancer diagnosis.

2.1.4. Preparing Study Cases

The target of this activity was to select the set of radiology exams of mammogram, ultrasound and MRI studies to be reported by the selected radiologists. The reports were performed following the two methods exposed at section 2.1.1.

For this purpose, a set of retrospective radiology exams (mammogram, ultrasound and MRI) from different patients were gathered. Those radiology exams were selected from the available studies that fulfil the inclusion criteria, starting on a fixed date in time (backwards) and selecting them consecutively, without taking into account the difficulty of the cases.

The inclusion criteria were: Studies with only one lesion (either benign or malign), from the first diagnosis of a patient and therefore without contextual alterations (previous treatments, post-operative status etc…). All studies must comprise a mammography, ultrasound or a MRI (from clinical routine). We defined a threshold of 50 cases by modality as a reasonable number for obtaining statistically significant results.

For each study, an annex was prepared with the following information:

- PACs Identifier of the radiology exam to facilitate the radiologist to download the images to the workstation where the radiology exam will be examined.
- Precise indications about the location of the lesion in the radiology exam images. This way, the time required for searching the lesions in the radiology exams is not included in the ET, which only comprises the actual reporting time using each one of the methods.

Another set of 30 radiology exams (10 mammograms, 10 ultrasounds and 10 MRI) was gathered for training the radiologists with the Free-Text and Structure-Report tools, before the actual experiment was carried out.

2.1.5. Radiologists' training

The first target of this activity was to make the radiologists familiar with the tools for the two reporting methods to avoid a bias due to their inexperience in using the tools.

On one hand, all selected radiologists had already acquired the needed skills for using Orion Clinic (Method 1) as it is the tool used in the hospital for daily clinical work, including the speech-transcribed reporting. On the other hand, it was necessary to perform a tutorial for providing the radiologists with the skills on the TRENCADIS web interfaces (Method 2).

After a few sessions with a training data set of radiology exams, the efficiency of TRENCADIS web interfaces was evaluated [20]. The ET got stable quickly, concluding that the web interfaces are intuitive and easy to learn.

The second target of the tutorial was to explain the strict protocol defined in the experiment for generating the reports and to enable an accurate timing of the reporting process. This protocol should reduce undetermined bias in the acquisition of the ET of the reporting process. The protocol defined for the radiologists was the following:

1) Each radiologist had to choose a PACS workstation, a study for reporting and the method for reporting, following the predefined schema that will be described in the next section.

2) Each radiologist had to read the PACS identifier of the radiology exam from the associated annex that was prepared, and to download all the images from the PACS server on the workstation. The radiologists could not open the viewer of the workstation at this step.

3) The radiologist had to open the Orion Clinic application or the TRENCADIS web interface from a Google Chrome browser (depending on the method chosen), open the image in the workstation viewer and start the reporting process. The radiologists had to locate the lesion following the indications of the annex, minimizing the searching time.

4) Each radiologist had to send the free-text reports to the RIS and the SRs to the TRENCADIS database when finished. At this moment, the reporting process will end, registering the ET.

5) Each radiologist had to close explicitly the reporting application (either Orion Clinic or Google Chrome browser).

Any alteration in the order, or the non-fulfilment of any of the above mentioned rules will invalidate the report to be considered for the study in this paper.

2.2. Data gathering phase

The radiologists produced the reports between 04:00 p.m and 07:00 p.m, after the end of the working day schedule (from 8:00 a.m. to 15:00 p.m), Monday to Friday.

2.2.1. Generation of the Reports

The target of this activity was to measure the elapsed time of the reporting of the 150 radiology exams by each one of the radiologists and using both methods and tools.

The generation of the reports was organized by blocks that indicated the radiology exams to be reported and the method to be used (Table 1). For example, Block B1 indicates that the radiologist must generate the reports of the radiology exams of mammography, ultrasound and MRI numbered from 1 to 15, and must use the Structured Report method with the Web Interface (Method 2). This block comprises a total of 45 structured reports.

Table 1. Schedule for methods and the number of radiology exams involved in each block defined in the study.

ID Block	Radiology Exams	Method
B1	1-15	2
B2	16-30	1
B3	31-45	2
B4	46-50	1
B5	1-15	1
B6	16-30	2

B7	31-45	1
B8	46-50	2

This schedule forces each radiologist to perform two cycles that were separated by one month (Table 2). Each cycle will take four weeks and for each week the radiologist had to report all radiology exams belonging to a given block.

The blocks were distributed avoiding a given block to be processed in the same week by two different radiologists. This will minimize potential biases as a result of radiologists exchanging information as the workstations are located in the same room.

Also, the order of the blocks assigned in the first cycle and the second cycle for the same radiologist are related, as they report the same radiology exams using different method. For example, in the first week from the first cycle, radiologist 1 will deal with block B1 and in the first week of the second cycle block B5 will be analysed. B1 and B5 deal with the same radiology exams but using different methods. This is to avoid that the radiologists recall what they have included in the report in the first cycle.

Table 2. Time schedule of reporting Blocks.

	1ª Cycle (1 Month)					**1ª Cycle (1 Month)**			
	Week 1	**Week 2**	**Week 3**	**Week 4**		**Week 1**	**Week 2**	**Week 3**	**Week 4**
Radiologist 1	B1	B2	B3	B4		B5	B6	B7	B8
Radiologist 2	B2	B3	B4	B5	1 Month	B6	B7	B8	B1
Radiologist 3	B3	B4	B5	B6		B7	B8	B1	B2
Radiologist 4	B4	B5	B6	B7		B8	B1	B2	B3
Radiologist 5	B5	B6	B7	B8		B1	B2	B3	B4
Radiologist 6	B6	B7	B8	B1		B2	B3	B4	B5

2.2.2. Data Curation

The results of the ET obtained are the source for evaluating the efficiency on producing the reports. Those studies that had extremely long or extremely short values of ET (with respect to

three standard deviations with respect to the mean value) were discarded. Furthermore, if a report were invalid, the corresponding one on the complementary method was also invalidated.

2.3. Analysis of Results

Finally, results were analysed carefully in the last phase. In this phase two types of analysis were performed. Horizontal analysis compared the ET in the two reporting methods grouping by modalities and profiles, and vertical analysis compared the Efficiencies between modalities and profiles of radiologists.

2.3.1. Analysis of Elapsed Time

A horizontal analysis was performed to compute whether there are statistical differences between the ET in the two reporting methods for each one of the modalities and profiles of radiologists. For this, we use the t-Student (T-S) [22] test for paired samples with a 95% confidence interval.

The T-S requires the two paired distributions to be are normal. We used the Kolmogorov-Smirnov (K-S) test as the number of values in of both distributions were larger than 30 [23]. Initially both distributions were not normal, therefore they had to be transformed using a logarithmic function. The logarithmic distributions fulfilled the K-S test and the T-S could be applied on them, maintaining the same statistical behaviour that the origin distributions [24].

2.3.2. Analysis of the Relative Efficiencies

Three vertical analysis were performed to evaluate the differences in the Relative Efficiency (REF) of both modality and radiologist profile on a set of independent samples. The REF compares the two reporting methods for each study. The REF was computed for each report and radiologist using the formula shown in figure 5, where Relative Efficiency (REF) is the ratio between the Elapsed Time (ET) of the Method 1 and Method 2 for generating a given report by the same radiologist.

$$Relative\ Efficience\ (REF) = \frac{Elapsed\ Time\ \ for\ Reporting\ using\ Method\ 1}{Elapsed\ Time\ for\ Reporting\ using\ Method\ 2}$$

Figure 5. Relative Efficiency (REF) between the Elapsed Time (ET) of the Method 1 and Method 2 for generating a given report by a same radiologist.

REF could depend on the modality and on the radiologist profile. To evaluate if a specific radiologist profile is more sensitive to the use of structured reports (e.g if skilled radiologists could speed-up more if they use structured reports or if the efficiency of a specific modality is significantly smaller than in other), several vertical analysis were performed.

The first vertical analysis evaluated if there were significantly statistical differences among the REFs of the three modality groups. For this, we used the one-way ANOVA with the Post Hoc LSD [25] and a level of significance of 5%. As this method requires distributions to be homoscedastic, we performed the Levenne (LV) [26] test to validate that the variances are equal with a level of significance of 5%.

The second vertical analysis searched for significant statistical differences between the REFs resulting from each radiologist profile (Higher, Intermediate and Low), inside each one of the modalities. In this analysis we can identify if within a given modality, the REF depends also in the radiologist's profile. For this, we used the one-way ANOVA with the Tamhane's T2 (T2T) Post Hoc test [27] with a level of significance of 5%. In this case the distributions were not homoscedastic, so we were unable to use the Post Hoc LSD as in the previous experiment.

The third vertical analysis evaluated the significant statistical differences among the REFs of each modality (Mammography, Ultrasound, MRI) for each radiologist profile (High, Medium or Low). This way we can identify if within a specific expertise group there are differences among the REFs for the different modalities. Second analysis would show, if Structured Reporting in a specific modality is equally beneficial in time for all the profiles. This third analysis will show if there are modalities that are more or less effective given a specific expertise group. For this, the one-way ANOVA with the Post Hoc LSD with a level of significance of 5% was used. In this case

the nine distributions were homoscedasticity so the LV test with a level of significance of 5% was used.

The statistical analysis (T-S, K-S, LV, ANOVA, LSD and Tamhane's T2) was performed using IBM SPSS [28].

3. RESULTS

3.1. DICOM-SR templates based on BIRADS 5th Edition

A preliminary version of DICOM-SR templates (mammography, ultrasound and MRI) used in this study was presented in [29]. These templated were based on the 4th Edition of BIRADs. We updated these templates using BIRADs 5th Edition, replacing the deprecated terms by the new items and restructuring the whole templates according to the recommendations of the 5th edition of BI-RADS. Figure 6 shows the DICOM-SR data type for each field. Basically, the next five types were considered: a) TEXT to represent text values; b) DATE to represent date values; c) NUM to represent numerical values including booleans (0 is false and 1 is true); d) CODE to represent the values coded using a standard terminology; and e) CONTAINER to represent the sections of the report. The templates also indicate the requirement for each field, which are: a) Mandatory (M) if the field is required in the report; b) Mandatory Conditional (MC) if the field is required only when a given condition is met; and c) User (U) if the field is optional, letting the user decide to include the value or not.

The DICOM-SR templates are included in the supplementary material as annexes. Annex A includes the mammography DICOM-SR template, ANNEX B includes the ultrasound template and ANNEX C so does the MRI template. We outlined in the annexes the RADLEX, SNOMED-CT and TRBR terminology used to code the CONCEPT NAMES for all fields and the CODE VALUES in the CODE data types. The cardinalities of the fields, the restrictions associated to each field and the default values are also described.

Figure 6. Basic tree of DICOM-SR templates for the three modalities: A) mammography, B) ultrasound and c) MRI.

3.2. Radiology reports

Table 3. Number of valid reports by modality and radiologist's profile.

Profile of Radiologist	Modality	Nº Valid Reports
High	Mammography	84
	Ultrasound	85
	MRI	66
Intermediate	Mammography	79
	Ultrasound	80
	MRI	67
Low	Mammography	74
	Ultrasound	74
	MRI	78

After performing the Data Curation, we ended up with two series of data corresponding to the ET for each one of the methods. Table 3 shows the number of valid reports by modality and radiologist's profile.

3.3. Measurements of the Elapsed Times (ETs)

Table 4 shows the averages of the ETs for each one of the reporting methods by modality. Ultrasound reports took the shortest time and MRI reports the longest, for both methods.

For each modality, a horizontal study was performed showing that the ET average of M2 was lower than M1. To verify if there was a significant statistical difference, the T-S for pair samples was computed. Before computing T-S, data were transformed using the logarithmic function to guarantee the normality of the distributions, which was tested drawing a Normal Q-Q plot (Figure 7) and computing the K-S test (Table 4). All transformed distributions were close to normal (K-S $p > 0.05$). The T-S indicated that the difference between the ET of M2 and M1 was statistical significant in Mammography (T-S $p=0,000 < 0.05$) and Ultrasound (T-S $p=0,000 < 0.05$). However the differences of the reporting methods were not statistical significant in MR Imaging (T-S $p=0,097 > 0.05$).

Table 4. Verifying the differences of the Elapsed Time (ET) for the two reporting methods. The columns show: Average of ET of free-text (M1). Average of ET of Structured Reports (M2). Difference of ETs between the average values of M2 and M1. Significance probability for Kolmogorov-Smirnov (K-S) test, using the logarithmic data distribution from M1. Significance probability for K-S test, using the logarithmic data distribution from M2. Significance probability of t-Student (T-S) test for paired samples. Number (N) of valid reports.

Modality	Avg. ET M2	Avg. ET M1	Difference	p (K-S) M1	p (K-S) M2	p (T-S)	N
Mammography	0:03:35	0:04:18	0:00:44	0,063	0,063	0,000	237
Ultrasound	0:03:21	0:03:57	0:00:36	0,059	0,200	0,000	237
MRI	0:06:11	0:06:21	0:00:10	0,085	0,053	0,097	211

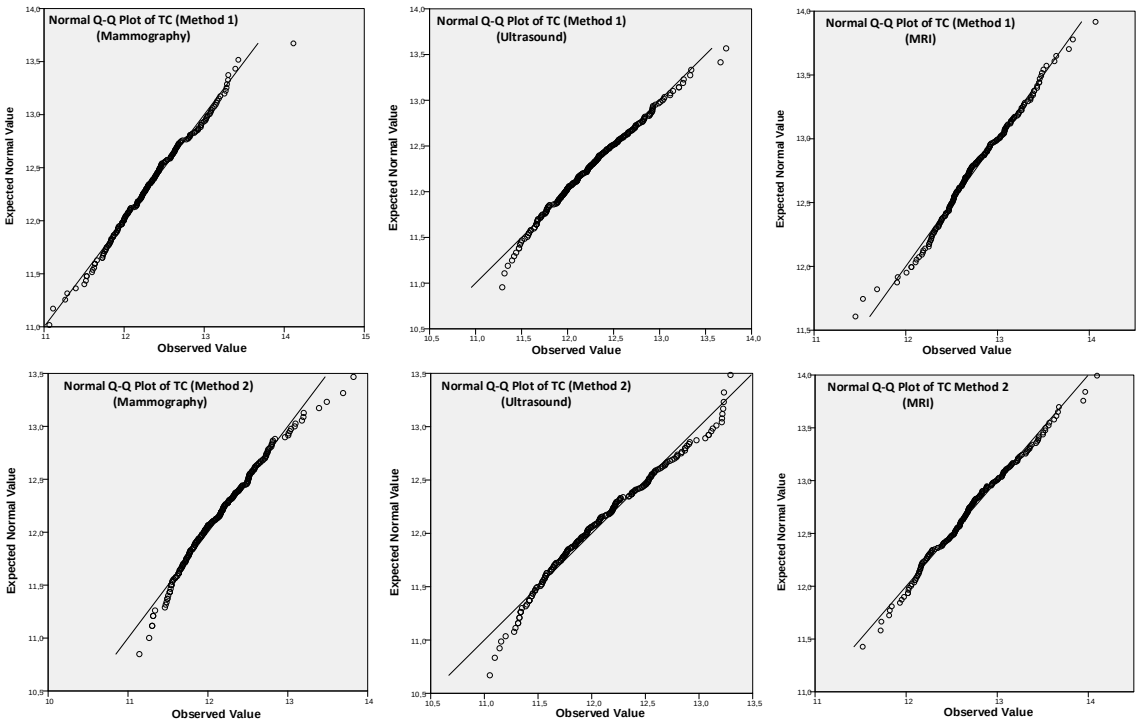

Figure 7. Normal Q-Q plot of logarithmic data distributions of M1 and M2 in the three modalities.

We wanted to analyse if the radiologist's profile affected the difference between the two reporting methods. Therefore, we grouped the values in the horizontal study in the three radiologist's profile (High, Intermediate and Low). In the three cases, the ultrasound was the fastest and the MRI the lowest as happened before when the profiles were not considered (Table 5).

The T-S (Table 5) test was computed by modality for each profile. This will show us if there are statistically significant differences in the ET between both reporting methods for each profile. All data were transformed using the logarithmic function and the K-S test was computed to validate the normality of each distribution (Table 3). We could verify the normality of all the distributions. The results for the high and intermediate profiles indicated that the difference between the ET for M2 and M1 was statistically significant in both mammography and ultrasound but not in MRIs. However, the differences between both methods for all the modalities in the case of low profile were not significant statistical.

Table 5. Analysing the Elapsed Time (ET) for each radiologist's profile. The columns indicate: Average of ET in free-text (M1). Average of ET in Structured reports (M2). Difference of ETs. Significance probability for K-S test, using the logarithmic data distribution from M1. Significance probability for K-S test, using the logarithmic data distribution from M2. Significance probability of T-S test for paired samples. Number (N) of valid reports.

	Modality	Avg. ET M2	Avg. ET M1	Difference	p (K-S) M1	p (K-S) M2	p (T-S)	N
High	Mammography	0:02:54	0:03:42	0:00:47	0,163	0,200	0,000	84
	Ultrasound	0:02:43	0:03:20	0:00:37	0,200	0,200	0,000	85
	MRI	0:05:49	0:06:06	0:00:17	0,200	0,200	0,157	66
Intermediate	Mammography	0:03:40	0:04:43	0:01:03	0,200	0,200	0,000	79
	Ultrasound	0:03:14	0:04:09	0:00:55	0,051	0,078	0,000	80
	MRI	0:06:11	0:06:13	0:00:02	0,200	0,200	0,383	67
Low	Mammography	0:04:15	0:04:33	0:00:19	0,200	0,200	0,192	74
	Ultrasound	0:04:14	0:04:27	0:00:13	0,070	0,051	0,469	72
	MRI	0:06:29	0:06:41	0:00:12	0,182	0,200	0,536	78

3.2.1. Measurements of Relative Efficiencies (REFs)

The REF shows the conditions where Structured Reporting leaded to higher productivity. Therefore, we wanted to deeply compare the Relative Efficiency (REFs) obtained in several vertical analyses.

Table 6. Results of the Relative Efficiency (REF) by modality. Columns indicate: Average of REF. Significance probability for K-S test, using the logarithmic data distribution from REFs. Number (N) of valid reports. Significance probability of Levenne Test (LV). Significance probability of ANOVA test. Significance probability of LSD pot hoc ANOVA test.

Modality	Avg. REF	p (K-S)	N	P (LV)	p (ANOVA)	Multiple Comparations	p (LSD)
Mammography	1.35	0.2	237	0.297	0.003	Mammography - Ultrasound	0.950
Ultrasound	1.35	0.2	237			Mammography - MRI	0.003
MRI	1.16	0.2	211			Ultrasound-MRI	0.003

We first analyse how modalities contribute to the REF (Table 6) without considering the radiologist profiles.

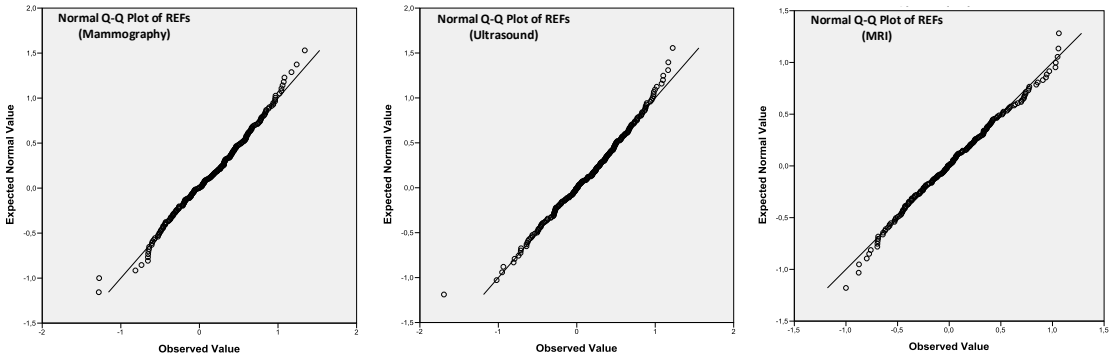

Figure 8. Normal Q-Q plot of logarithmic data distributions of REFs in the three modalities.

For this statistical analysis we compute an ANOVA test, grouping the data by modalities. The normality of the distributions was tested and data was transformed using a logarithmic function and Q-Q graphics (Figure 8). We computed the K-S test (Table 6) for each distribution. The requirement of homoscedasticity was validated using LV test (Table 6) which was fulfilled (LV $p=0.297 > 0.05$) in all distributions too.

Table 7. Analysis of REF by Radiologists' profile within each modality. Columns show: Average of REF. Significance probability for K-S test, using the logarithmic data distribution from REFs. Number (N) of valid reports. Significance probability for LV test. Significance probability for ANOVA test. Significance probability of Tamhane's T2 (T2T) pot hoc ANOVA test.

Modality	Group Experience	Avg. REF	p (K-S)	N	p (LV)	P (ANOVA)	Multiple Comparations	P (T2T)
Mammography	High	1.35	0.200	84	0.000	0.024	High- Intermediate	0.965
	Intermediate	1.53	0.200	79			High-Low	0.033
	Low	1.15	0.200	74			Intermediate -Low	0.052
Ultrasound	High	1.36	0.200	85	0.000	0.008	High- Intermediate	0.974
	Intermediate	1.53	0.050	80			High-Low	0.010
	Low	1.13	0.092	72			Intermediate -Low	0.023
MRI	High	1.17	0.200	66	0.078	0.844	High- Intermediate	--
	Intermediate	1.19	0.200	67			High-Low	--
	Low	1.13	0.200	78			Intermediate -Low	--

The ANOVA test indicated that there were significant statistical differences (ANOVA p =0.003 < 0.05). The post hoc LSD test (see Table 6) was computed to detect where were these differences are. The results indicated that they are no differences between mammograms and ultrasound scans (LSD p=0.950 > 0.05), but both have significant statistical differences with MRI (LSD p < 0.05).

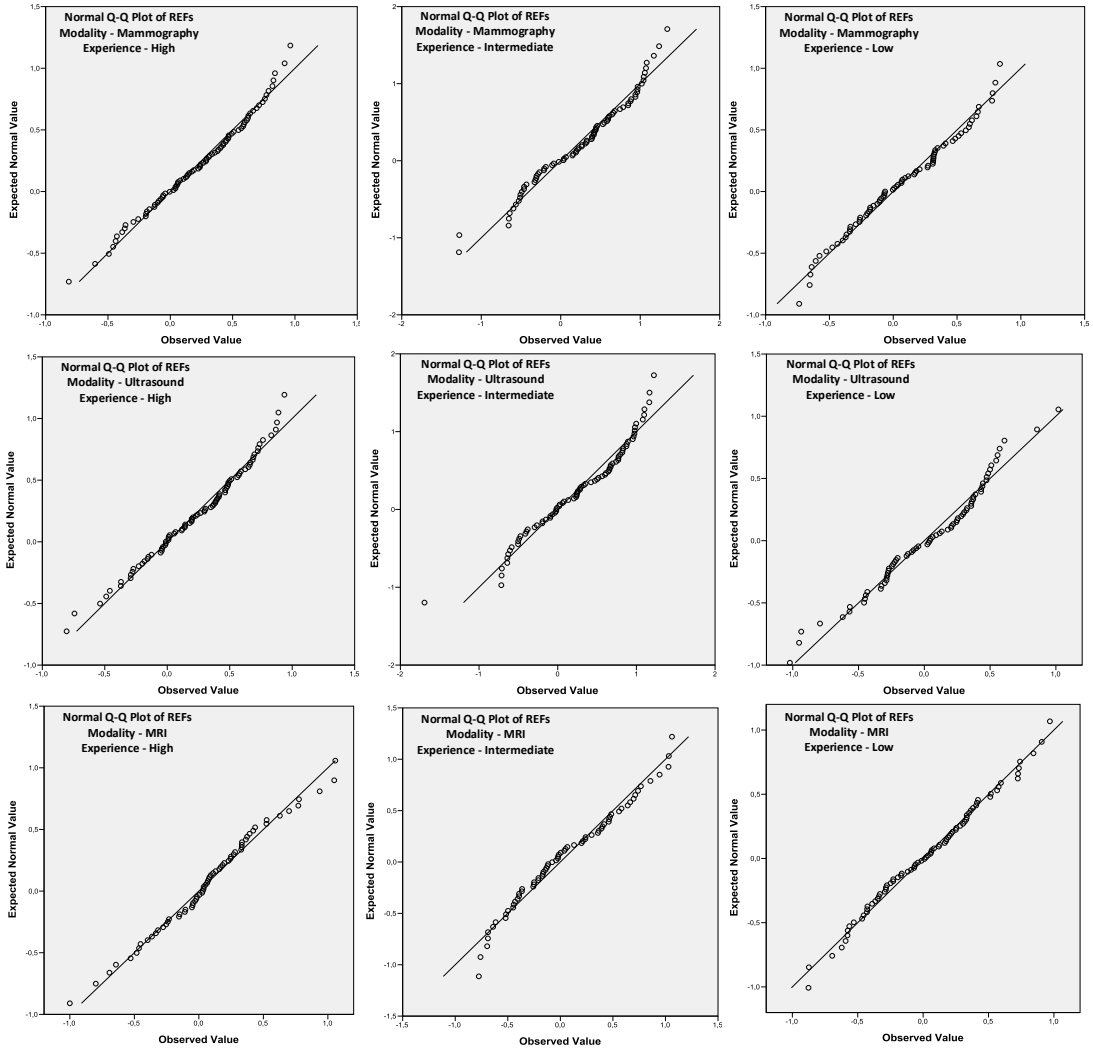

Figure 9. Normal Q-Q plot of logarithmic data distributions of REFs in the three modalities by experience.

The second vertical analysis analysed if the radiologist's profile affected the efficiency in each modality. This way, we can analyse if the reduced REF of MRI and the higher REF for the other two modalities were the same for all profiles. For this analysis, an ANOVA test was

computed for each modality, grouping by profile. The normality of the distributions was obtained by using a logarithmic function and Q-Q graphics (Figure 9. . K-S test (Table 7) for each distribution were computed. As the homoscedasticity (see Table 7) was not fulfilled (LV $p > 0.05$) for Mammograms and Ultrasound, we applied the post hoc T2T instead of the LSD test.

The ANOVA test indicated that there were significant differences in Mammograms (ANOVA $p = 0.024 < 0.05$) and Ultrasounds (ANOVA $p = 0.008 < 0.05$) between the radiologist profile groups. We confirmed that in MRI there were no significant differences (ANOVA $p = 0.078 > 0.844$) for any group. The results of the post hoc T2T test (Table 7) indicated that in Mammograms there were no differences between High-Intermediate profiles (T2T $p=0.965 > 0.05$) and Intermediate-Low profiles (T2T $p=0.052 > 0.05$, but there were differences between High-Low profiles (T2T $p=0.033 < 0.05$). These values indicate that the differences between the extremes (High and Low) were significant, and the intermediate profile was closer to be the same group of High Profile (T2T p=0.0.0965) than Low Profile (T2T $p = 0.052$). We can state then that the group of intermediate experience have the highest REF. In Ultrasound there were no significant differences between High-Intermediate profiles (T2T $p=0.974 > 0.05$) but there were differences between High-Low profiles (T2T $p=0.010 < 0.05$) and Intermediate-Low profiles (T2T $p=0.023 < 0.05$). In this case, we observe the same tendency but showing a higher benefit for a smaller profile experience.

In the third vertical analysis performed searched for significant differences between modalities for each radiologist profile. In this case, we want to reassure that, as previous results show, Mammography and Ultrasound in high and intermediate profiles experiment showed the highest improvement with Structured Reporting. For this analysis, we computed an ANOVA test for each radiologist profile, grouped by modality. The normality of the distributions was obtained with a transformation of the data using a logarithmic function and Q-Q graphics (Figure

9), and verified with the K-S test (Table 8) for each distribution. The homoscedasticity was verified using the LV test (Table 8) for all cases (LV $p > 0.05$) so the LSD could be used.

Table 8. Studying the effect of the modalities per experience group in the Relative Efficiency (REF). The columns show: Average of REF. Significance probability for K-S test, using the logarithmic data distribution from REFs. Number (N) of valid reports. Significance probability for LV test. Significance probability for ANOVA test. Significance probability of LSD pot hoc ANOVA test.

Experience	Group Modality	Avg. REF	p (K-S)	N	p (LV)	p (ANOVA)	Multiple Comparations	P (LSD)
High	Mammography	1.35	0.200	84	0.975	0.028	Mammo – Ultrasound	0.914
	Ultrasound	1.36	0.200	85			Mammo -MRI	0.021
	MRI	1.17	0.200	66			Ultrasound -MRI	0.016
Medium	Mammography	1.53	0.200	79	0.199	0.045	Mammo – Ultrasound	0.970
	Ultrasound	1.53	0.050	80			Mammo -MRI	0.030
	MRI	1.19	0.200	67			Ultrasound -MRI	0.027
Low	High	1.15	0.200	74	0.925	0.886	Mammo – Ultrasound	--
	Ultrasound	1.13	0.092	72			Mammo -MRI	--
	MRI	1.13	0.200	78			Ultrasound -MRI	--

The ANOVA test indicated that there were significant differences on the High profile (ANOVA $p = 0.028 < 0.05$) and medium (ANOVA $p = 0.045 < 0.05$) between some modality groups (Mammography or Ultrasounds and MRI). For the Low Profile radiologists there were not significant differences (ANOVA $p = 0.886 > 0.05$) among the modalities (the efficiency is similar in all the cases).

4. DISCUSSION

The results of Table 4 show that ultrasound reports took the shortest time and MRI reports the longest, for both methods. This may be caused due to the higher volume of MRI (256 images) with respect to mammography (between 2-4 images) and ultrasound radiology exams (between 2-4 images). Furthermore, studies reported fair to moderate variability in inter-observer mammographic interpretation [30, 31, 32, 33, 34] and fair to substantial in sonographic features

[35, 36, 37, 38] according kappa statistics [39]. One can argue that the ultrasound scan was quicker than mammography as its interpretation is less subjective, and therefore the radiologist doubt less to report the terms, being the reporting quicker.

Furthermore, Table 4 shows that the difference between the ET of M2 and M1 was statistically significant in Mammography and Ultrasound but were not statistically significant in MR Imaging. There are studies in other areas (CT of chest, head, abdomen or pelvis) concluding that structured reports do not affect radiologist time [40]. However, we think that the differences in the significance in mammography and ultrasound could be produced by the subjectivity of the BI-RADS terms. When a radiologist evaluates a subjective term tends to extend the report with clarifying notes when use the free-text (M1) penalizing the ET and even more if he or she uses the voice recognition dictation, as these systems result in higher physician time than conventional transcription [40, 41]. However, if they use the structured reporting (M2) they have to choose among a set of BI-RAD terms, which could lead to a reduced ET. As has been indicated above, the variability of Mammography is fair to moderated, in Echography is fair to substantial. However, in MRIs, previous studies reported moderate to substantial variability in inter-observer interpretation [42, 43], improving the objectivity regarding Mammograms and Ultrasound.

Table 5 shows if the radiologist's profile affected the difference between the ET. Always, the ultrasound was the fastest and the MRI the lowest as happened before when the profiles were not considered. This fact could be explained by the same reasons exposed above relative to the volume of data managed and the subjectivity of BI-RADs terms. As one can expect, we observed that in both methods and for the three modalities, the higher the experience, the shorter the ETs. Also, the results of Table 5 show that the differences between both methods for all the modalities in the case of low profile were not statistically significant. Therefore, we understand that the inexperience led to longer times, regardless of the reporting method.

Due to the results of ETs, we expected that the average REF was always greater than 1 (Table 6 and Table 7). Table 6 shows that the averages of REFs in mammograms and ultrasound scans were equals and the MRI was lower than the other modalities. This fact could be justified by the same reasons exposed at previous vertical analysis of ETs relative to the subjectivity of the terms. Due to the better objectivity of BI-RADS terms in MRI, the differences on the ETs between the two methods are similar to the differences observed in the other modalities. Furthermore, the results indicated that they are no differences between mammograms and ultrasound scans, but both have significant statistical differences with MRI. This indicates that the time gain in Structured Reporting is similar for both Mammography and Ultrasound, but the effect on Structured Reports is smaller. These results are consistent with the findings we had in the horizontal study.

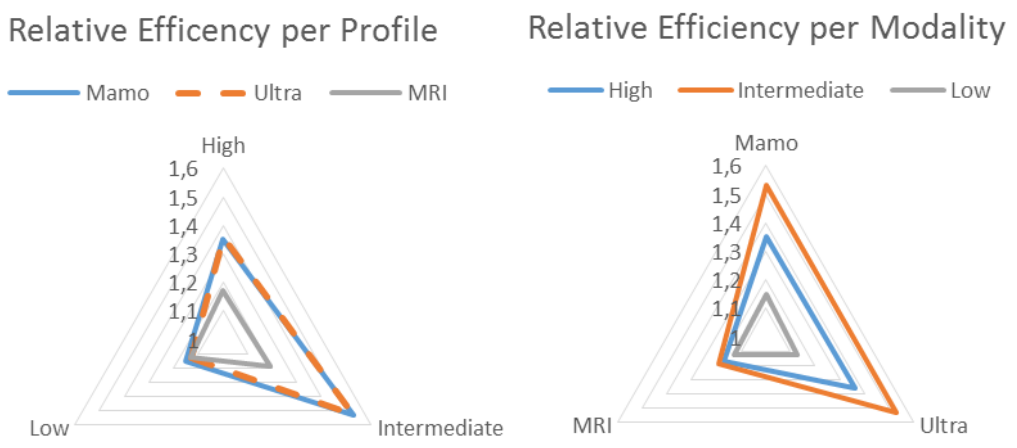

Figure 10. Radial graphs showing the REFs per profile (left) and REFs per Modality (right).

The results of Table 7 indicate that the differences between the extremes (High and Low) were significant (see figure 10), and the intermediate profile was closer to be the same group of High Profile. We think that is because the intermediates profiles have had a good formation during four years in breast cancer diagnosis, instead, in the case of the low profile, radiologists only have one year of training and the expertise is poor yet.

The results of Table 8 indicated that the REF in High and Intermediate profiles is significantly similar for Mammograms and Ultrasounds, but there were differences between Mammograms-MRI and Ultrasound-MRI. These values show that the complexity between Mammograms and Ultrasound are similar due to the variability of terms.

5. CONCLUSIONS AND FUTURE WORKS

In this work we have defined and demonstrated a methodology to evaluate the productivity of software tools for producing radiology reports. This methodology can be applied to many other research works that aim at comparing the introduction of new approaches in radiological departments.

The methodology proposed is used to compare the Elapsed Time (ET) on a set of radiological reports for breast cancer diagnosis based on BI-RADS using two different methods (Structured Report with TRENCADIS Web Interface and plain-text with speech recognition). The impact in the ET, of both the modality used (mammography, ultrasound or MRI) and the experience of the radiologists (low, medium and high) was lower than the required for free-text, with statistically significant differences. Both the modality and the radiologist's profile affect the ET of the reports, in particular, in all cases where the radiologist's profile is low, the differences on the ET were not statistically significant. The same effect is observed for the MRI studies. However, there were significant differences in all cases where the radiologist's profile is medium or high.

The Relative Efficiency (REF) (speed up in time using the structured reporting) between the groups of modalities and profiles of radiologists is always higher than one, the Structured Reporting taking shorter time in all the conditions explored in the article. More precisely, the REF is significantly higher when the profile of the radiologist is intermediate or high.

Therefore, adopting Structured Reporting in mammography and ultrasound studies in breast cancer diagnosis improves the performance of radiology departments.

As a future work we planned to study the possibility to propose the inclusion of the BI-RADS missing terms in RADLEX lexicon. Also, the next step will be to perform a new experiment to assess and compare the quality of radiological reports both in free-text and SR in terms of correctness and completeness with respect to a gold standard.

ACKNOWLEDGMENT

The authors want to thank the INDIGO-Datacloud project. INDIGO - DataCloud receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement RIA 653549.

REFERENCES

[1] J. Oakley, A Primer for Radiographers, Radiologists and Health Care Professionals. Digital Imaging, Cambridge University Press, 2003.
[2] O Ratib, Imaging informatics: From image management to image navigation, Yearb Med Inform 2009; 167–172
[3] M. Mustra, K. Delac, M. Grgic. Overview of the DICOM Standard. ELMAR, 2008. 50th International Symposium. Zadar, Croatia. pp. 39–44. ISBN 978-1-4244-3364-3.
[4] A. Zafar, J.M. Overhage, C.J. McDonald, Continuous speech recognition for clinicians, J. Am. Med. Inform. Assoc. 6 (3) (1999) 195/204.
[5] R.G. Zick, J. Olsen, Voice recognition software versus a traditional transcription service for physician charting in the ED, Am. J. Emerg. Med. 19 (4) (2001) 295/298.
[6] S.M. Borowitz, Computer-based speech recognition as an alternative to medical transcription, J. Am. Med. Inform. Assoc. 8 (1) (2001) 101/102.
[7] D.L. Weiss, C.P. Langlotz, Structured reporting: patient care enhancement or productivity nightmare? Radiology. 2008 Dec;249(3):739-47. doi: 10.1148/radiol.2493080988.
[8] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*.
[9] D.A. Clunie, DICOM Structured Reporting. PixelMed, Bangor (2000).
[10] Association National Electrical Manufactures: Digital Imaging and Communication in Medicine (DICOM) 3.0, supplement 23 DICOM Structured Reporting.
[11] N. Rita, Benefits of the DICOM Structured Report. J Digit Imaging 2006; 19: 295–306.
[12] BI-RADS American College of Radiology. BI-RADS Committee. (1998). Breast imaging reporting and data system. American College of Radiology (Ed.). American College of Radiology.
[13] RadLex, Radiology Society of North America (RSNA). Available at http://www.rsna.org/radlex.aspx. Accessed July 2016.
[14] D. Marwede, T. Schulz, T. Kahn, Indexing Thoracic CT Reports Using a Preliminary Version of a Standardized Radiological Lexicon (RadLex). J Digit Imaging 2007; 28: 1–8.
[15] Sahni VA, Silveira PC, Sainani NI, Khorasani R. Impact of a Structured Report Template on the Quality of MRI Reports for Rectal Cancer Staging. AJR Am J Roentgenol. 2015 Sep;205(3):584-8. doi: 10.2214/AJR.14.14053.

[16]Schwartz LH, Panicek DM, Berk AR, Li Y, Hricak H. Improving communication of diagnostic radiology findings through structured reporting. Radiology. 2011 Jul;260(1):174-81. doi: 10.1148/radiol.11101913. Epub 2011 Apr 25.
[17]SB. Johnson, S. Bakken, D. Dine, S. Hyun, E. Mendonça, F. Morrison, T. Bright, T. Van Vleck, P. Wrenn J,Stetson, An electronic health record based on structured narrative, J Am Med Inform Assoc 2008; 15: 54–64.
[18]BT. O'Connell, C. Cho, N. Shah, KM. Brown, RN. Shiffman, Take note(s): differential HER satisfaction with two implementations under one roof. J Am Med Inform Assoc 2004; 11: 43–49.
[19]Y. Hasegawa , Y. Matsumura, N. Mihara, Y. Kawakami, K. Sasai, H. Takeda, H. Nakamura, Development of a system that generates structured reports for chest x-ray radiography, Methods Inf Med, 2010;49(4):360-70. doi: 10.3414/ME09-01-0014.
[20]C. Maestre, D. Segrelles, E. Torres, I. Blanquer, R. Medina, V. Hernández, L. Martí, Assessing the usability of a science gateway for medical knowledge bases with TRENCADIS, J Grid Computing 2012; 10:665–688.
[21]D.L. Moody, The method evaluation model: a theoretical model for validating information systems design methods, In: Proceedings of the 11th European Conference on Information Systems. Naples: ECIS; 2003. p. 1327-1336.
[22]R. Walpole, R. Myers, K. Ye, Probability and Statistics for Engineers and Scientists. Pearson Education, 2002.
[23]Daniel, W. Wayne, "Kolmogorov–Smirnov one-sample test". Applied Nonparametric Statistics (2nd ed.), 1990, Boston: PWS-Kent. pp. 319–330. ISBN 0-534-91976-6
[24]M.S. Bartlett, D.G. Kendall. The Statistical Analysis of Variance Heterogeneity and Logarithmic Transformation, Journal of the Royal Statistical Society, Ser. B. 8, 128‑138, 1946.
[25]A.J. Hayter, "The Maximum Familywise Error Rate of Fisher's Least Significant Difference Test". Journal of the American Statistical Association 81 (396): 1000-1004, 1986, doi:10.2307/2289074.
[26]H. Levene, "Robust tests for equality of variances", In Ingram Olkin, Harold Hotelling, et alia. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, 1960, Stanford University Press. pp. 278–292.
[27]D. Cramer, D. Howitt, Tamhane's T2 multiple comparison test, In (Eds.), The SAGE dictionary of statistics. (p. 169). London, England: SAGE Publications, Ltd, 2004, doi: http://dx.doi.org/10.4135/9780857020123.n590
[28]IBM Corp. Released 2015. IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp
[29]R. Medina, E. Torres, D. Segrelles, I. Blanquer, L. Martí, D. Almenar, A Systematic Approach for Using DICOM Structured Reports in Clinical Processes: Focus on Breast Cancer, J. Digit. Imaging 2015; 28:132-145
[30]J.G. Elmore, S.L. Jackson, L. Abraham, D.L. Miglioretti, P.A. Carneya et al., Variability in Interpretive Performance at Screening Mammography and Radiologists' Characteristics Associated with Accuracy. 2009. Radiology.
DOI: http://dx.doi.org/10.1148/radiol.2533082308
[31]WA. Berg, C. Campassi, P. Langenberg, MJ. Sexton, Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment. AJR Am J Roentgenol 2000;174:1769–77
[32]K. Kerlikowske, D. Grady, J. Barclay, SD. Frankel, SH. Ominsky, EA. Sickles, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J Natl Cancer Inst 1998;90:1801–9
[33]JG. Elmore, CK. Wells, CH. Lee, DH. Howard, AR. Feinstein. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331:1493–9

[34]E. Lazarus, MB. Mainiero, B. Schepps, SL. Koelliker, LS. Livingston. BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. Radiology 2006;239:385–91

[35]N. Abdullah, B. Mesurolle, M. El-Khoury, E. Kao, Breast Imaging Reporting and Data System Lexicon for US: Interobserver Agreement for Assessment of Breast Masses. Radiology, ,2009, 252, 665-672.

[36]M.J. Calas, R.M. Almeida, B. Gutfilen, W.C. Pereira, Intra-Observer Interpretation of Breast Ultrasonography Following the BI-RADS Classification. European Journal of Radiology, 2009, 74, 525-528. http://dx.doi.org/10.1016/j.ejrad.2009.04.015

[37]C.S. Park, J.H. Lee, H.W. Yim, B.J. Kang, H.S. Kim, J.I. Jung, N.Y. Jung, S.H. Kim, Observer Agreement Using the ACR Breast Imaging Reporting and Data System (BI-RADS)-Ultrasound. Korean Journal of Radiology, 2007, 8, 397-402.

[38]H.J. Lee, E.K. Kim, M.J. Kim, J.H. Youk, J.Y. Lee, D.R. Kang, K.K. Oh, Observer Variability of Breast Imaging Reporting and Data System (BI-RADS) for Breast Ultrasound. European Journal of Radiology, ,2008, 65, 293-298.http://dx.doi.org/10.1016/j.ejrad.2007.04.008

[39]J. Cohen, Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 1968;70:213-20.

[40]Pezzullo, J.A., Tung, G.A., Rogg, J.M. et al. Voice Recognition Dictation: Radiologist as Transcriptionist. J Digit Imaging (2008) 21: 384. doi:10.1007/s10278-007-9039-2

[41]Bhan, Sasha N, MD; Coblentz, Craig L, MD, FRCPC; Norman, Geoffrey R, PhD; Ali, Sammy H, MD, BHSc. Effect of Voice Recognition on Radiologist Reporting Time . Canadian Association of Radiologists Journal 59.4 (Oct 2008): 203-9.

[42]Hanna, T. N., Shekhani, H., Maddu, K., Zhang, C., Chen, Z., & Johnson, J. O. (2016). Structured report compliance: effect on audio dictation time, report length, and total radiologist study time. Emergency radiology, 23(5), 449-453

[43]E.K. Mona, E. Khoury, L. Lalonde, J. David, M. Labelle, B. Mesurolle, I.E Trop, Breast imaging reporting and data system (BI-RADS) lexicon for breast MRI: Interobserver variability in the description and assignment of BI-RADS category. DOI: http://dx.doi.org/10.1016/j.ejrad.2014.10.003