Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

Improving the quality of radiological findings through computer-based training

BACHELOR-ARBEIT

2020

Autorin
Lehmann, Manuela

betreuende Person
Dr. Michel, Stefan

Praxispartner
CASRA
Dr. Hardmeier, Diana

Improving the quality of radiological findings through computer-based training

Abstract

Imaging methods are well-known in the medical area and their interpretation is an important skill of a physician. However, it is not a main focus of the medical education in universities. This study investigates whether computer-based training can improve the detection performance of medical students for pleural effusions in x-ray images. The detection performance was improved significantly by the group who engaged in computer-based training for eight weeks. The participants mentioned their interests in expanding an online training tool for different imaging methods in the medical area. Improving the detection performance can have a positive outcome for patient and hospital and can be accomplished by computer-based training. This paper contains 49´182 characters (incl. spaces, without appendices).

Keywords: computer-based training (CBT); radiology; pleural effusion; signal detection theory (SDT)

Zusammenfassung

Bildgebende Verfahren werden im medizinischen Bereich oft verwendet wobei die Fähigkeit ein Bild richtig zu interpretieren zentrale Bedeutung besitzt. Trotzdem ist das Erlernen dieser Fähigkeit im Medizinstudium nicht von zentralem Wert. Die vorgelegte Studie untersucht ob anhand computer-basiertem Trainings Medizinstudierende ihre Erkennungsleistung eines Pleuraergusses in Röntgenbilder signifikant verbessern werden kann. Die Gruppe, welche ein achtwöchiges computer-basiertes Training absolviert hat, konnte ihre Erkennungsleistung signifikant verbessern. Die Studienteilnehmenden sind ebenfalls daran interessiert ein Online-Training für verschiedene Disziplinen von bildgebenden Verfahren im medizinischen Bereich zu verwenden. Eine verbesserte Erkennungsleitung kann sich positiv auf Patienten und Krankenhaus auswirken und kann durch computer-basiertes Training erreicht werden. Diese Arbeit beinhaltet 49´182 Zeichen (inkl. Leerzeichen, ohne Anhang)

Schlüsselwörter: Computer-basiertes Training (CBT), Radiologie, Pleuraerguss, Signal Detektion Theorie (SDT)

Table of Contents

1. Introduction

Imaging methods are widely spread in the medical area. Magnetic resonance imaging (MRI), computed tomography (CT), sonography (ultrasound) and radiography (x-ray) are the most well-known techniques and are used in different fields of medicine. It seems to be an important part of being a physician which makes it even more interesting that it is not a main focus of the medical education (Fabre et al., 2018; Salajegheh et al., 2016). For example, in Switzerland, a separate organization provides classes for medical students and assistant doctors to teach them the theory and practical use of sonography. They are independent of the university medical education and state that learning about sonography is not taught enough in the medical study even though sonography is gaining importance in clinical diagnostics (*Young Sonographers*, 2017). The importance of imaging methods can also be shown by the current COVID-19 disease where CT could play a major part in understanding and handling this new sickness (Lin et al., 2020).

A lack of knowledge in chest x-ray (CXR) reading among medical students and other medical staff members can be displayed by different studies (Eisen et al., 2006; Fabre et al., 2018). The lack of knowledge can lead to a high rate of incorrect statements in clinical radiology reports that can have adverse patient outcomes (Dikshit et al., 2005; Nyhsen et al., 2013). Therefore, the aim of several studies, including this presented study, is to increase the ability to read and interpret CXR images. Including E-learning as a teaching tool would therefore be an effective way to teach the necessary skills in order to increase accuracy of x-ray interpretation. Salajegheh et al. (2016) designed an experiment where the experimental group could benefit from an e-learning tool. Even though a workshop and lecture teaching on x-ray interpretation was provided for the whole sample, the experimental group improved their x-ray interpretation skills

significantly through the additional e-learning (Salajegheh et al., 2016, p. 3). Furthermore, the experimental group contained only medical students of the first-year Bachelor whereas the control group was in their second year. A recent study done by Sha et al. (2020) showed an increased skill of the participants through perceptual learning while perceptual learning was provided through training on the computer with CXR pictures.

This current study focuses on radiography, more specifically the interpretation of x-ray images. X-ray images are not only used in a medical setting, but are also well known in aviation security, for example screening the passenger bags to check for hidden prohibited items. Schwaninger and Hofer (2004) showed the positive effect of computer-based training (CBT) of the detection performance of screeners. They divided the screeners in four groups with equivalent detection ability and created the training using a Latin Square design. Between the four training blocks tests were done to measure the detecting performance. The detection performance was measured by d' based on the Signal Detection Theory (Green & Swets, 1966) which will be further discussed in the next section. The screeners were advised to perform two 20 minutes training each week over a six-month period. A relative increase of the detection performance of the screeners was shown. After an average of 28 training sessions an enhancement of 71% of the detection performance and for a subgroup after an average of 31 training session an increase of 84% could be seen.

Due to the success of CBT in aviation security, it will also be used for this study in the field of CXR images interpretation skills. In the previously discussed study by Schwaninger and Hofer (2004) the training software X-Ray Tutor was used. X-Ray Tutor distinguishes itself from other training softwares because of its adaption to each person that performs a training session. In order for the software to be able to adapt its

training to each individual person, it has to be defined what makes an x-ray image difficult to interpret. For x-ray images of passenger bags it is the effect of viewpoint, effect of superposition, and effect of bag complexity (Schwaninger, 2004). However, this presented study is the first which uses the X-Ray Tutor as a training software in the medical field. As the information to create an adaptive training system are not yet available, four levels of difficulty of x-ray images were created for this study (see 2.2.3).

To measure the detection performance the Signal Detection Theory (SDT) will be used for this study. The SDT is a method in the area of psychophysics and was developed by Green and Swets (1966). The theory differs between *noise* and *signal.* On a detection task, you search for a *signal* while the *signal* is surrounded by *noise*. In this study, *signal* is the pleural effusion (PE), which should be detected by the participants, while *noise* means everything else on the CXR. PE is an excessive accumulation of fluid between the lung and chest tissue which is not a specific disease itself but often refers to another underlying pathology (Karkhanis & Joshi, 2012). Four responses of the participants are possible where “correct rejection” and “hit” are correct detections and “false alarm” and “miss” are incorrect detected:

- Correct rejection: Image contains only *noise* and participant defines it as OK
- False alarm: Image contains only *noise* and participant defines it as NOT OK
- Miss: Image contains *noise* and *signal*. The participant defines it as OK
- Hit: Image contains *noise* and *signal*. The participant defines it as NOT OK

The sensitivity d' or A' defines the detection performance. It analyses the ratio between “hit” and “false alarm” under the conditions of statistical normal distribution and variance homogeneity. If the conditions cannot be met the non-parametric measure

A' can be used (Green & Swets, 1966; Schwaninger, 2005). This method of SDT is very complex therefore, Schwaninger (2005) provides an example. He states that even though two participants can have the same hit-rate, their detection performance is not the same. While participant B has a much higher false alarm-rate than participant A, he makes the work task less efficient (e.g. longer line on the passenger bag control point). To identify the real detection performance the ratio between "false alarm" and "hit" must be used which is provided by the measure of sensitivity. Another measure which has to be observed is the criterion. The criterion depends on cost/benefit estimation, work motivation and expected probability of occurrence of the *signal*. This measure can change very quickly while detection performance is a stable trait of a person and can be changed through specific training (Green & Swets, 1966; Schwaninger, 2005).

The success of CBT in aviation security and the need for further research in how to train the CXR interpretation skills led to this study. The expected changes of the CXR interpretation skills resulted in the three following hypotheses:

Detection performance	Detection performance will increase significantly through the CBT
Time of reaction	Participants will need significant less time to decide whether a PE can be detected or not through CBT
Confidence rating	Confidence about the given answer will increase significantly through CBT

2. Methods

This chapter describes the different methods of the presented study. First, the group of participants will be defined as well as the procedure. Second, the used materials will be explained. Additionally, A' as measure for detection performance will be further discussed.

2.1 Participants and procedure

In this study 29 students of medicine participated where 8 were male and 21 females between the age of 20 – 28 (M = 23,76, SD = 2,116) who took part voluntarily after receiving a description of the study (Appendix A). After a questionnaire about their demographics (see 2.2.1) and the first Medical X-Ray Competence Assessment Test (M-CAT) (see 2.2.2) an experimental and a control group were formed. These two groups were as similar as possible. A t-test was used to ensure no significant differences between the two groups while the following variables were regarded: age, sex, d', A', experience of PE and experience of x-ray interpretation (p > .419) (Appendix B). The experimental group (N = 14) performed two months of CBT with the Medical X-Ray Tutor training system (M-XRT) (see 2.2.3) whereas the control group (N = 15) did not engage in any kind of training until both groups started the second M-CAT.

2.2 Materials

In this chapter the used materials will be described. Surveys were used to collect demographic data and to ask about the experience of the experimental group after the training. M-CAT is the software which provides the test that calculates the detection performance of the participants. It was used in the beginning and in the end of the study. M-XRT is the training software where the experimental group completed their training sessions. M-CAT and M-XRT are developed by CASRA, the center for adaptive

security research and applications which was founded in 2008. They started their research with x-ray screening at the airport in Zurich and afterwards expand their research and applications not only to airports worldwide but also in different areas (e.g. cargo, mail, prison, etc.) (*CASRA*, n.d.).

2.2.1 Surveys

Before the first M-CAT and after finishing the M-XRT the participants are asked to take a survey. The first survey will collect the needed demographic data to create two statistically equivalent groups which will be defined as the experimental group and the control group. This survey will be sent to every participant who agreed to take part in this study. The second survey will only be sent to the experimental group that will have completed the M-XRT. This survey asks the participant about their opinion of the M-XRT, a general online-training tool for their study, and how x-ray image interpretation is taught now at their universities. Both surveys are created at Unipark Questback and can be found in Appendix C and Appendix D.

2.2.2 Medical X-Ray Competence Assessment Test (M-CAT)

X-Ray CAT was originally developed in the field of aviation security (*CASRA*, n.d.). This instrument measures the x-ray image interpretation competency of security screeners. The test contains several x-ray pictures of passenger bags where in some of them were prohibited items (i.e. knives, firearms, etc.). The screeners had to detect the forbidden items and the test measured how well they could find these particular items (Koller & Schwaninger, 2006).

For this medical study, the original X-Ray CAT was changed to a more fitted version for its use in the medical field. A University Hospital in Switzerland provided the x-ray pictures. The radiology department of the hospital divided the pictures into

whether they showed a PE or not and where the PE was located on the x-ray picture. For the M-CAT a total of 78 x-ray pictures were used. At the beginning of the test, an e-learning presentation (Appendix E) about PE was used to prepare all of the participants. Six of the before mentioned x-ray pictures (three with PE and three without) where used as trial runs before the original M-CAT started. These six images were the only ones that showed if the answer was correct or incorrect whereas in the official part of the M-CAT no feedback was given to the participants. Within the picture, the participants were able to zoom to help them detect the PE. The official test part contained two blocks of 36 pictures each in which 12 showed a PE and 24 did not. The x-ray pictures were shuffled therefore all levels of complexity were shown in M-CAT. For each picture the participants had to decide whether the x-ray picture contained a PE or not by clicking either one of two buttons ("pleural effusion visible" or "pleural effusion not visible") before the picture would disappear after two minutes. After each decision, the participant was asked to rate the confidence about their decision in a rating scale from "unsure" to "very sure" where one of five options could be chosen. A screenshot of the M-CAT with the confidence rating can be seen in Figure 1. The participants were asked to take a break of ten minutes between the two blocks of x-ray pictures. The M-CAT was planned to take around one hour and ended with a thank you note and the further plan of the study.

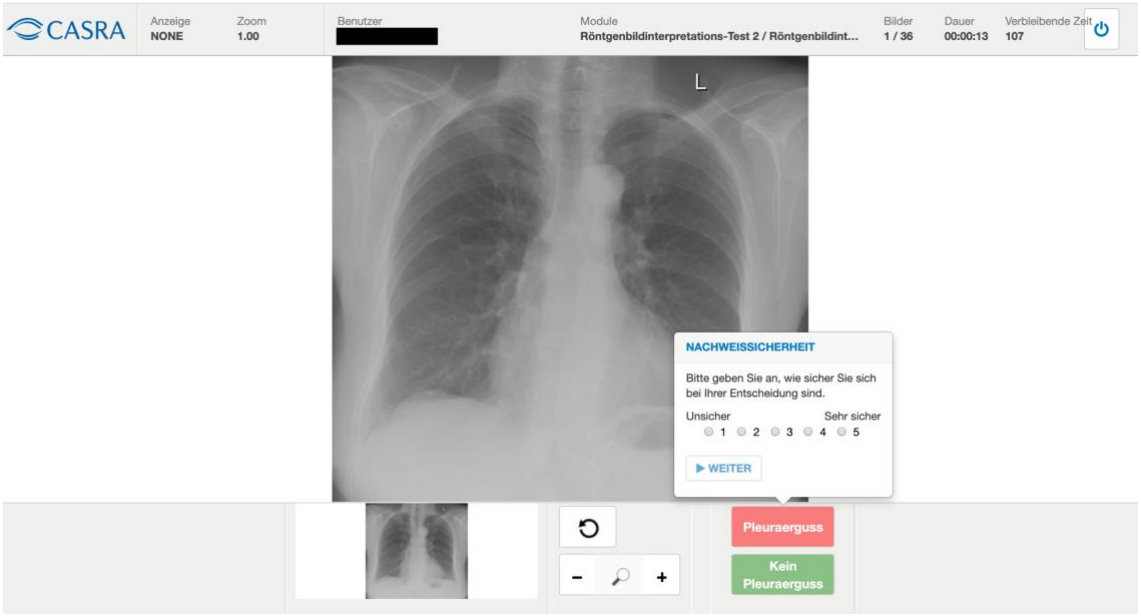

Figure 1: Screenshot of the M-CAT with confidence rating after deciding a pleural effusion is visible.

2.2.3 Medical X-Ray Tutor training system (M-XRT)

The X-Ray Tutor training system is also used to train security screeners in aviation security (*CASRA*, n.d.). The screeners had to decide for each picture if it contained harmful objects and through the feedback, they knew if they were correct or not. The procedure contained multiple sessions to train the screeners and to improve their ability to detect prohibited objects in passenger bags (Schwaninger, 2004).

Like in the case of M-CAT the X-Ray Tutor was designed a little differently in order to fit into the medical setting. The pictures were also provided by a University Hospital in Switzerland and were in the same condition as the pictures which were used for M-CAT. Like in the M-CAT the participants were able to zoom within the x-ray images. Before starting the training, the participants received instructions on how to interpret the feedback they received after each session of training (Appendix F). The M-XRT provided feedback after every training session and also after every picture if a PE is shown with its location or not, which can be seen in Figure 2.

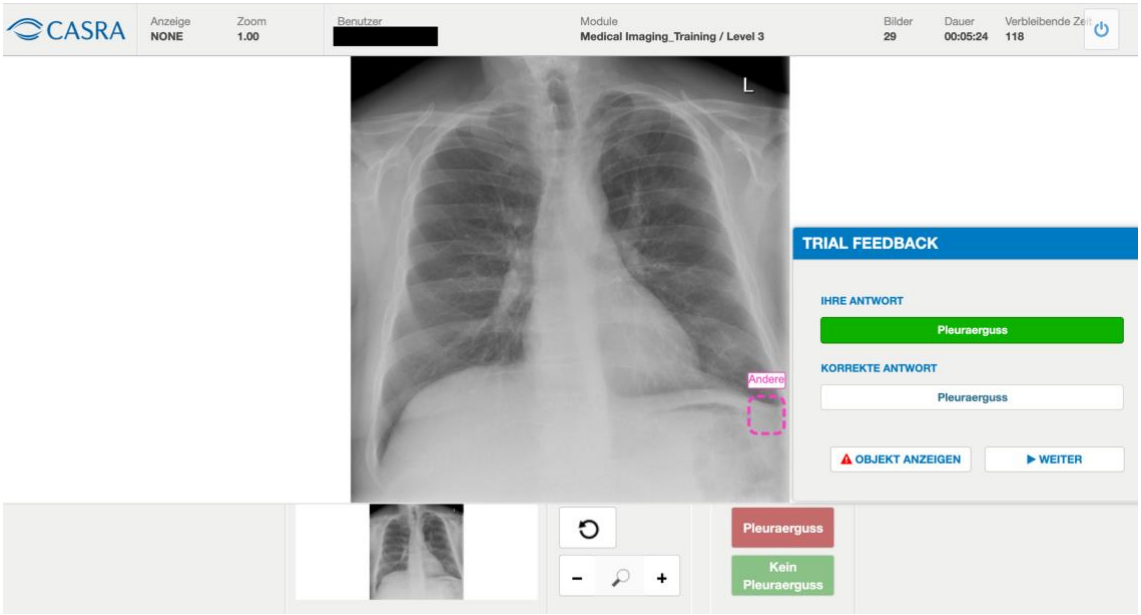

Figure 2: M-XRT image of level 3 where a pleural effusion got detected correctly.

The participants were asked to train each week twice for 20 minutes, which was recommended, or one time 40 minutes for an eight-week period. This equals a total training time of 5 hours and 20 minutes. The x-ray pictures in M-XRT were divided into three levels of complexity and a fourth level where the participant had also to mark on the picture where they assume the PE which can be seen in Figure 3. Level 1 to Level 3 contained 34 pictures with PE and 68 without. Level 4 included all the pictures of the training and some which were still unused to create an infinite learning space for fast learners. The following criteria had to be met to enter the next level:

- Level 1 to Level 2: Seen every picture twice and 75% right detections
- Level 2 to Level 3: Seen every picture twice and 80% right detections
- Level 3 to Level 4: Seen every picture twice and 85% right detections

After each training session, the participants were given individual feedback in order to see their progress.

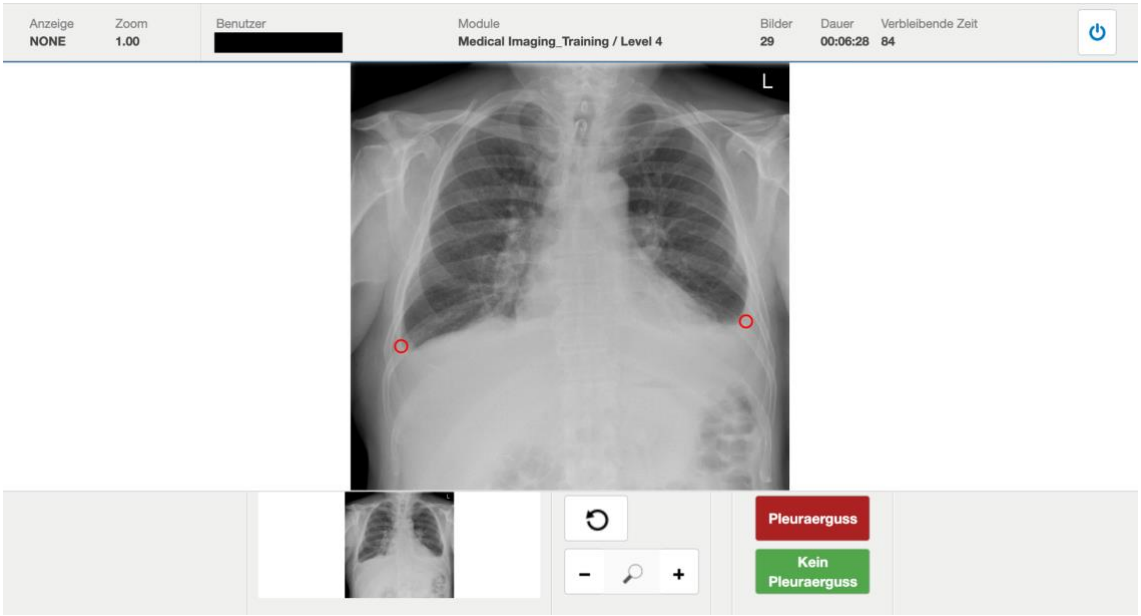

Figure 3: M-XRT image of level 4 with markings where pleural effusion is assumed.

2.3 Detection Performance A'

As it is explained in the introduction, d' and A' can be used as a measure for detection performance. Considering both hit-rate and false alarm-rate is crucial to investigate the real detection ability of a person, as Schwaninger (2005) shows in an example. A' does not require any prior assumptions about the underlying distributions which makes it a non-parametric measure and is calculated by the following formula (Grier, 1971):

$$A' = \frac{1}{2} + \frac{(H - F)(1 + H - F)}{4H\,(1 - F)}$$

In the formula H describes the hit-rate and F the rate of false alarm. Aaronson and Watts (1987) stated that a different formula must be used when $H < F$:

$$A' = \frac{1}{2} + \frac{(F - H)(1 + F - H)}{4F\,(1 - H)}$$

Even though A' is non-parametric and requires no prior assumptions about the underlying distributions, it is stated that “This does not mean that these measures are an accurate reflection of their theoretical origin (i.e., that A' reflects the area under a reasonable ROC curve) or that they are distribution-free measures” (Pastore et al., 2003, p. 565). However, A' is often used in a research and application context because it is easy to compute and interpret (Mendes et al., 2011). For this reason, it will also be used as a measure for detection performance in this study.

3. Results

This chapter is structured along the three hypotheses that are investigated in this study. It starts with the results of the detection performance, followed by the results of time of reaction, confidence rating and the survey.

3.1 Detection performance

As it is stated in chapter 2.3, A' will be used as the measure for the detection performance. The detection performance was measured in both experimental and control group two times in this study. First in the beginning of the study (pretest) and then again at the end (posttest) after the experimental group performed the training sessions. Both times the detection performance A' was calculated by the measure of hit, false alarm, correct rejection and miss with the M-CAT.

At the start, a Kolmogorov–Smirnov test was calculated to determine if the variables are normally distributed. A' in both pre- and posttest are normally distributed (Appendix G).

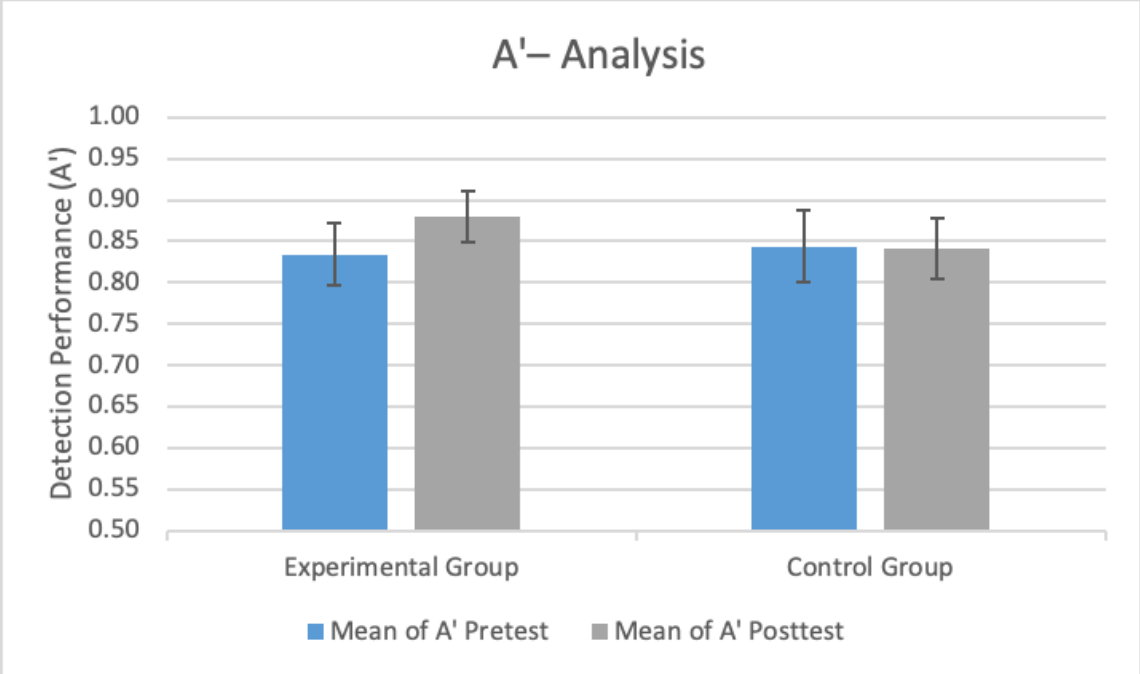

Figure 4: Comparison of Detection Performance A' between means of experimental and control group and standard deviation for both condition (pretest vs. posttest) and both groups.

Figure 4 shows the difference between the experimental and the control group considering the detection performance A'. A repeated measure ANOVA (Appendix H)

with the within-subjects factor *test date* (pretest vs. posttest) revealed a large main effect, $F(1,27) = 8.412$, $p = .007$, partial $\eta 2 = .238$. No significant main effect could be found for the between-subjects factor *groups* (experimental vs. control group), $F(1,27) = 1.630$, $p = .213$, partial $\eta 2 = .057$. The interaction between test date * group was revealed as significant ($F(1,27) = 10.772$, $p = .003$, partial $\eta 2 = .285$). The experimental group could increase their detection performance A' from pretest ($M = 0.834$, $SD = 0.037$) to posttest ($M = 0.880$, $SD = 0.031$). This difference was highly significant, $t(13) = -4.737$, $p < .001$, $n = 14$, 95% CI [-0.067, -0.025], $d = -1.343$, and can be seen in Table 1. Furthermore, a significant difference, $t(27) = 3.098$, $p = .005$, $n = 29$, 95% CI [0.013, 0.065], $d = 1.151$, is shown in Table 2 between the A' of the experimental group ($M = 0.880$, $SD = 0.031$) and the control group ($M = 0.841$, $SD = 0.037$) for the posttest. The difference of A' between the pretest ($M = 0.844$, $SD = 0.043$) and posttest ($M = 0.841$, $SD = 0.037$) for the control group was not significant ($t(14) = .255$, $p = .803$, $n = 15$, 95% CI [-0.021, 0.027], $d = 0.071$). In addition, no significant difference was found between the A' of the experimental ($M = 0.834$, $SD = 0.037$) and the control group ($M = 0.844$, $SD = 0.043$) for the pretest ($t(27) = -,633$, $p = .532$, 95% CI [-0.04, 0.021], $d = -0.235$). The non-significant t-tests can be found in Appendix I.

Table 1: T-Test of the experimental group between pre- and posttest for detection performance (A').

	A' of pretest		A' of posttest					
Variable	*M*	*SD*	*M*	*SD*	*t*(13)	*p (two-tailed)*	95% CI	Cohen's *d*
Pre- and posttest for experimental group	0.834	0.037	0.880	0.031	-4.737	< .001	[-0.067, -0.025]	-1.343

Table 2: T-Test of the posttest between the experimental and the control group for detection performance (A').

Variable	A' Experimental Group		A' Control Group					
	M	*SD*	*M*	*SD*	*t*(27)	*p (two-tailed)*	95% CI	Cohen's *d*
Group (experimental vs. control group) posttest	0.880	0.031	0.841	0.037	-3.098	.005	[0.013, 0.065]	1.151

As explained in the introduction, the Signal Detection Theory (Green & Swets, 1966) focuses on the relationship between hit- and false alarm-rate. Figure 5 shows the difference of those two rates for the experimental and the control group between the pre- and posttest.

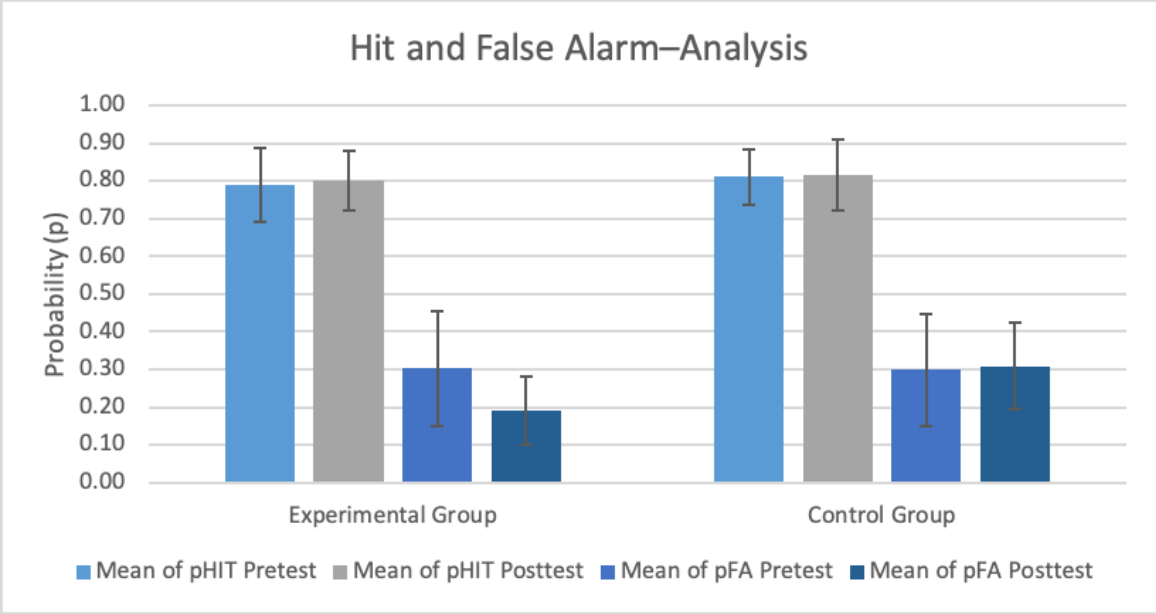

Figure 5: Comparison of the probability (p) of the hit-rate and false alarm-rate of the two groups in the pre- and posttest. Means and standard deviations are shown for both pretest and posttest and both groups.

A repeated measure ANOVA (Appendix J) of the hit-rate for the within-subjects factor *test date* (pretest vs. posttest) found no significant main effect, $F(1,27) = 0.134$, $p = .781$, partial $\eta 2 = .005$. The between-subjects factor *group* (experimental vs. control group) of the hit-rate did not find an significant main effect, $F(1,27) = 0.492$, $p = .489$, partial $\eta 2 = .018$. The interaction test date * group of the hit-rate was not found significant ($F(1,27) = 0.052$, $p = .822$, partial $\eta 2 = .002$). No significant main effect of the false alarm-rate could be found by a repeated measure ANOVA (Appendix J) for the within-subjects factor *test date* (pretest vs. posttest), $F(1,27) = 4.063$, $p = .054$, partial $\eta 2 = .131$. The between-subjects factor *group* (experimental vs. control group) of the false alarm-rate did not find a significant main effect, $F(1,27) = 1.947$, $p = 0.174$, partial $\eta 2 = .067$. A significant interaction was revealed for the false alarm-rate between test date * group, $F(1,27) = 5.762$, $p = .024$, partial $\eta 2 = .176$.

T-Tests of the hit-rate did not show any significant differences, all *p*-values > .492 (Appendix K). The false alarm-rate of the experimental group showed a difference between the pretest ($M = 0.302$, $SD = 0.153$) and posttest ($M = 0.190$, $SD = 0.091$) which was significant ($t(13) = 3.373$, $p = .005$, 95% IC [0.040, 0.183], $d = 0.817$) and can be seen in Table 3. Furthermore, a difference was found between the false alarm-rate of the experimental ($M = 0.190$, $SD = 0.091$) and the control group ($M = 0.308$, $SD = 0.114$) in the posttest. This difference was also significant ($t(27) = - 3.060$, $p = .005$, 95% IC [-0.197, -0.039], $d = -1.137$) and is shown in Table 4. The differences considering false alarm-rate between the pretest ($M = 0.299$, $SD = 0.149$) and posttest ($M = 0.308$, $SD = 0.114$) of the control group ($t(14) = - .257$, $p = .801$) along with the difference between the experimental ($M = 0.302$, $SD = 0.153$) and control group ($M = 0.299$, $SD = 0.149$) in the pretest ($t(27) = 0.062$, $p = .951$) was not significant (Appendix L).

Table 3: T-Test of the false alarm-rate between pre- and posttest of the experimental group.

	pFA of the pretest		pFA of the posttest					
Variable	*M*	*SD*	*M*	*SD*	*t*(13)	*p (two-tailed)*	95% CI	Cohen's *d*
Pre- and posttest for experimental group	0.302	0.153	0.190	0.091	3.373	.005	[0.040, 0.183]	0.817

Table 4: T-Test of the false alarm-rate between the experimental and the control group in the posttest.

	pFA Experimental Group		pFA Control Group					
Variable	*M*	*SD*	*M*	*SD*	*t*(27)	*p (two-tailed)*	95% CI	Cohen's *d*
Group (experimental vs. control group) posttest	0.190	0.091	0.308	0.114	-3.060	.005	[- 0.197, - 0.039]	-1.137

As stated in the introduction another measure which has to be observed along with the detection performance is the criterion. The Kolmogorov–Smirnov test could show that the criterion is normally distributed (Appendix M). A repeated measure ANOVA (Appendix N) for the within-subjects factor *test date* (pretest vs. posttest) did not find a significant main effect, $F(1,27) = 0.619$, $p = .438$, partial $\eta 2 = .022$. No significant main effect was revealed for the between-subjects factor *group* (experimental vs. control group), $F(1,27) = 1.585$, $p = .219$, partial $\eta 2 = .055$. The interaction between test date * group was not found significant ($F(1,27) = 2.17$, $p = .152$, partial $\eta 2 = .074$.

3.2 Time of Reaction

The hypothesis about the time of reaction was that it will take the participants less time to decide whether a PE was shown or not. In both M-CATs (pre- and posttest) for both groups (experimental and control group) the time of reaction was measured. For this analysis the mean of reaction time of each participant was calculated and used. The Kolmogorov–Smirnov test showed that the time of reaction for the pretest (experimental and control group) are not normally distributed while the time of reaction for the posttest in both group showed normal distribution (Appendix O). For this reason, a Mann-Whitney-U-Test for the independent samples was used while for the related samples a Wilcoxon-Test was used.

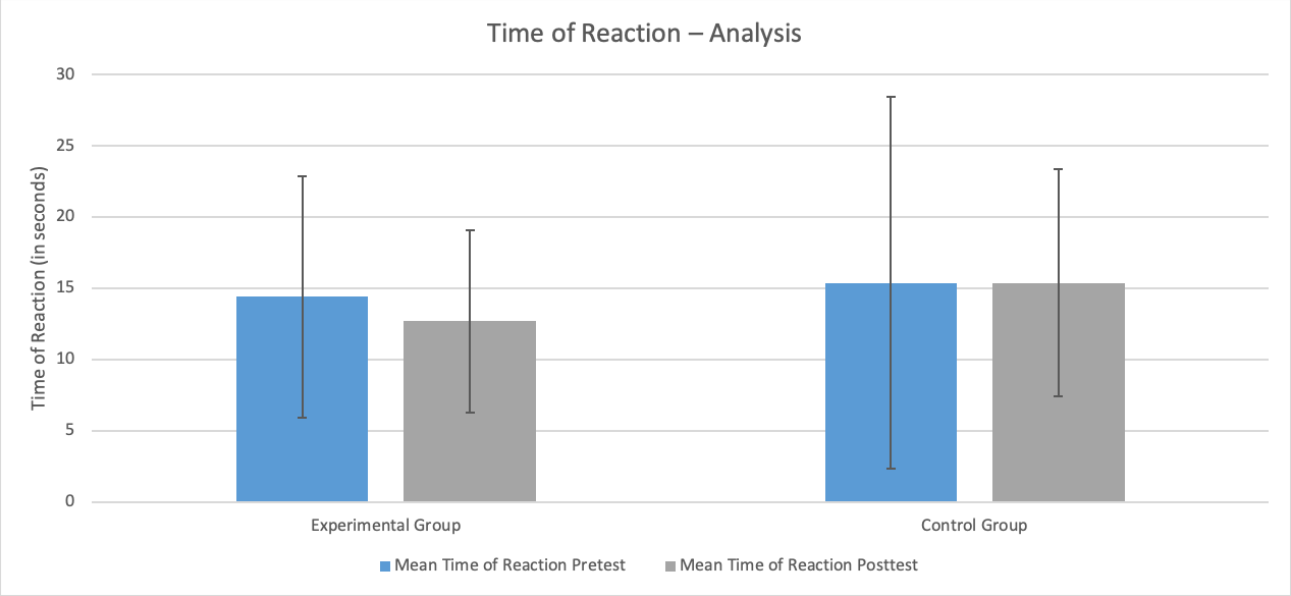

Figure 6: The means for time of reaction of experimental and control group and standard deviation for both condition (pretest vs. posttest) and both groups.

The differences between the means of the experimental and the control group considering the time of reaction in post- and pretest is illustrated in Figure 6. No significant difference could be found between the pretest (median = 12.025) and the posttest (median = 11.06) for the experimental group, z = -0.973, p = .331, n = 14, r =

.260. In addition, the difference between the pretest (median = 11.32) and the posttest (median = 12.24) for the control group was not revealed as significant, z = -0.398, p = .691, n = 15, r = .103. The Mann-Whitney-U-Test for the difference between the experimental group (median = 12.025) and the control group (median = 11.32) for the pretest was not found significant, U = 93.500, p = .621, r = .093. Furthermore, no significant difference between the experimental group (median = 11.06) and the control group (median = 12.24) was found, U = 82.000, p = .331, r = .186. All results for time of reaction can be found in Appendix P.

3.3 Confidence Rating

The confidence rating was measured in the pre- and posttest of both groups. As it is stated in the hypothesis the confidence rating should increase through CBT. Normal distribution was found for the control group for the pretest while the experimental group did not show that the confidence rating was normally distributed in the pretest. For the posttest of both groups the confidence rating was normally distributed. All Kolmogorov–Smirnov tests can be found in Appendix Q. Since the confidence rating is not normally distributed a Wilcoxon-Test was used for the related samples and the Mann-Whitney-U-Test for the independent samples.

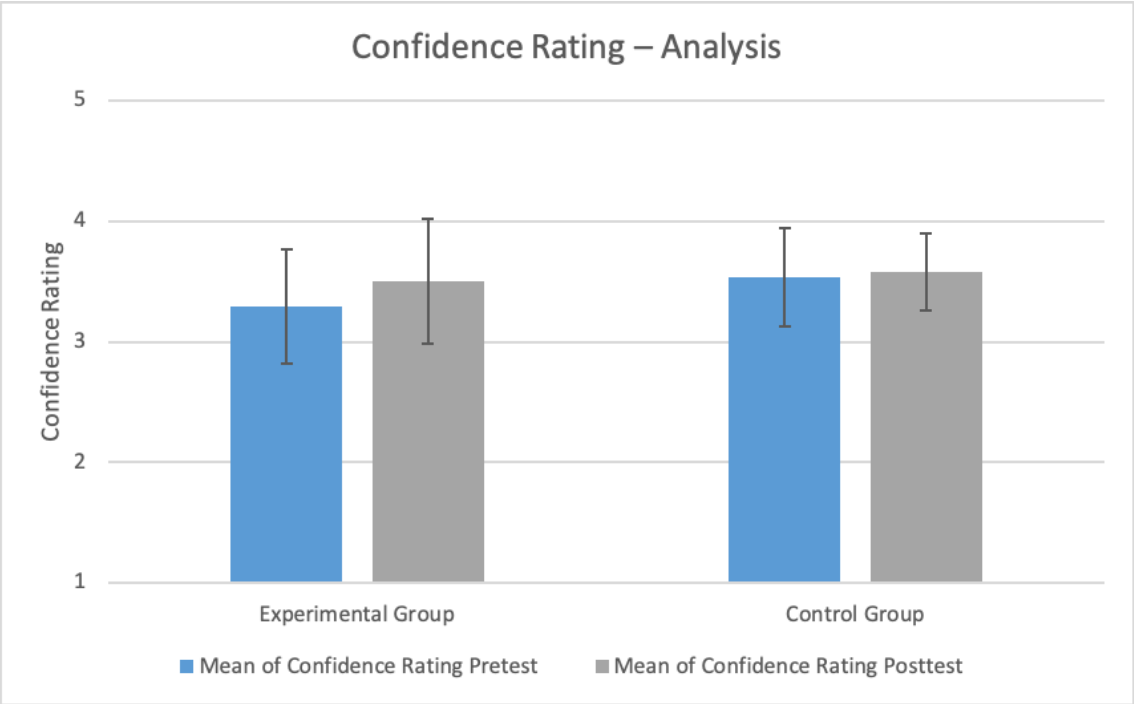

Figure 7: The means for the confidence rating of experimental and control group and standard deviation for both condition (pretest vs. posttest) and both groups. A ratio between 1 and 5 was possible for the participants to choose.

Figure 7 illustrates the means of both groups for the confidence rating of the pre- and the posttest. The Wilcoxon-Test could not find a significant difference between the pretest (median = 3.215) and the posttest (median = 3.51) for the experimental group, z = -1.538, p = .124, n = 14, r = .411. No significant difference could be revealed between the pretest (median = 3.4) and the posttest (median = 3.61) for the control group, z = -0.426, p = .670, n = 15, r = .110. For the pretest no significant difference between the experimental group (median = 3.215) and the control group (median = 3.4) could be found, U = 65.000, p = .085, r = .324. In addition, no significant difference between the experimental (median = 3.51) and the control group (median = 3.61) for the posttest was revealed, U = 91.000, p = .561, r = .113. All results of the analysis of the confidence rating can be found in Appendix R.

3.4 Survey

The survey contained questions about the training software, which was used in this study, the participants' opinion about an online training tool in general and how x-ray picture interpretation is currently taught at universities. Only the experimental group could participate in this survey because they are the only group who performed the training (N = 14).

The training in general was liked by 42% of the participants. 7% did not like the training at all while 50% showed a medium opinion about the training (Appendix S). Especially liked was the usability of the software (43%) and 29% liked the fourth level, where the participants not only had to decide whether a PE was visible but also had to mark the suspected area (Appendix T). When asked specifically about Level 4, 50% liked it, 43% had a medium opinion and 7 % did not like it at all (Figure 8).

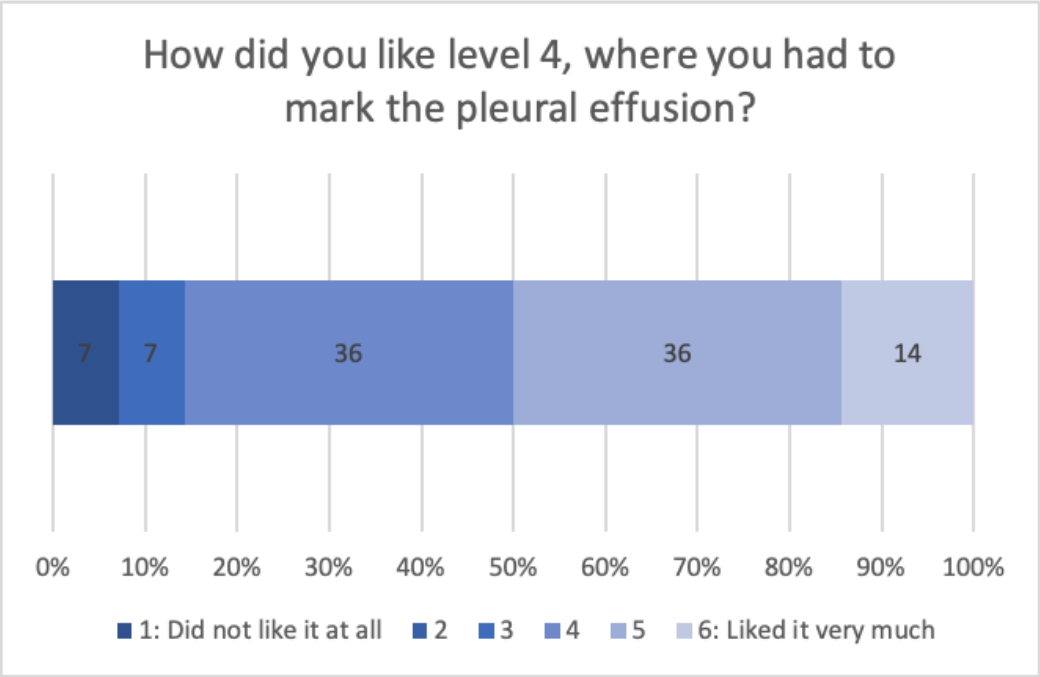

Figure 8: Illustration of the opinion about level 4. Opinion 2 was not chosen.

The comments concerning level 4 were mostly positive about the idea of marking the PE but there were some negative comments about its execution (i.e. “Die Idee war gut, die Umsetzung aufgrund der z.T. etwas eng gesetzten Markierung ziemlich frustrierend.”). All of the comments can be seen in Appendix U. 21% of the participants had a negative experience with the training software because there was no additional explanation about the PE only if it was visible or not (Appendix V). Missing explanation and information about the location of PE in the x-ray picture and PE in general was mentioned in different questions (Appendix W) (i.e. “Erklärung weshalb hier ein Pleuraerguss ist oder weshalb hier einer vermutet werden könnte bei schwierigeren Bildern.”). 93% of the participants think their progress of learning through the training was moderate while 7% feel like there was no progress at all (Appendix X). The differences of difficulty between the levels of the training were tangible for 71% of the participants while 28% felt a moderate increase of difficulty (Appendix Y). 21% were motivated to train regularly while 43% were moderately motivated and 35% were not motivated. Reasons for not being motivated were that the participants did not have an clear overview (14%) (i.e. “Motivation war da, Überblick nicht so.”) and that the increase of difficulty between level 2 and 3 was too strong (14%) (i.e. “…Generell ist der Sprung des Schwierigkeitslevels von Level 2 auf Level 3 meines Erachtens zu gross, während der Sprung von Level 1 auf Level 2 eventuell zu klein ist.”). It has to be said that 50% did not provide an explanation about their motivation (Appendix Z). To conclude the questions about the completed training the participants were asked if they had fun. 36% reported that they had fun with the training while 50% considered they had moderate fun and 14% did not have fun with the training (Appendix AA).

The second part of the survey asked about an online training tool in general. The participants found a lot of different areas of application for an online training tool (e.g. “Pleuraerguss, Pneumothorax, Herzinsuffizienz, Pneumonien, Frakturen, Bänderriss (MRI), Gehirnschäden (MRI) …”). A completed list with all the answers can be found in Appendix BB. 93% of the participants would recommend an online training to other radiologists which can be seen in Figure 9.

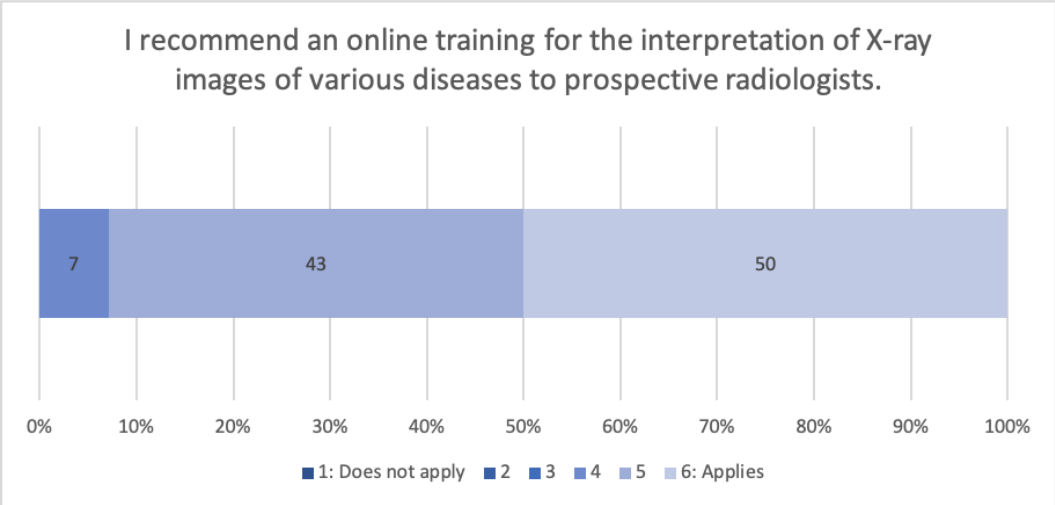

Figure 9: Illustration of the opinion about a recommendation of an online training for prospective radiologists. Option 1 (does not apply), 2 and 3 were not chosen.

36% of the participants think that every occupational group in the medical sector would benefit from an online training tool while other participants mentioned specific occupational groups like cardiologists or dermatologists (Appendix CC).

The last few questions asked about how x-ray picture interpretation is taught in universities at the moment. 57% of the participants think that x-ray picture interpretation is no enough taught in university while 36% have a medium opinion about it and 7% think it is taught enough. All the participants that explained their opinion described that they have too few lectures about this particular topic, 36% did not explain their opinion (Appendix DD). 72% answered that they would find it useful

if an online tool would be part of their curriculum while 21% find it moderately useful and 7% would say that an online training in the curriculum is not useful (Figure 10).

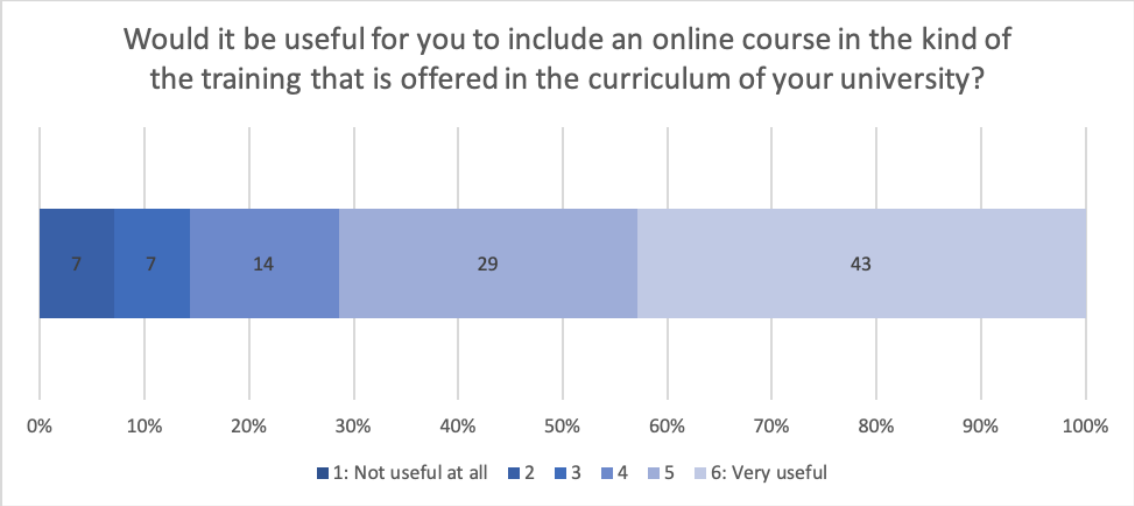

Figure 10: Illustration of the opinion about if an online course in the kind of the training as an addition to the curriculum would be useful. Option 1 (not useful at all) was not chosen.

The final question asked about if an online training was included in the curriculum and in which semester it should take place. 57% would start with an online training at the bachelor level while 42% would start in the master study (Appendix EE).

4. Discussion

Imaging methods are well-known in the area of medicine and are used on a regularly basis. An efficient interpretation of x-ray images is important for both the hospital and the patient (Dikshit et al., 2005; Nyhsen et al., 2013). The Covid-19 disease shows even more the importance of the interpretations skills of the physicians because of a recent study which shows that CT images could help understanding and handling the new disease (Lin et al., 2020). However, the interpretation of images from imaging methods are barely taught at the universities (Fabre et al., 2018; Salajegheh et al., 2016). This is consistent with the findings of this study where 57% of the participants stated that they

think that x-ray interpretation is not taught enough at their universities and 72% would find it useful if an online training tool would be included in the curriculum to learn more about this interpretational skill.

The improvement of x-ray interpretations skills is subject of many studies with different methods for improvement (Salajegheh et al., 2016; Sha et al., 2020). The interpretation of x-ray images is not only used in the medical area but also in aviation security. Schwaninger and Hofer (2004) where able to show that the detection performance improved through CBT. Due to the success of CBT in aviation security the presented study was created. The participants were divided into an experimental and a control group whereas both groups performed two M-CATs, once in the beginning of the study and once in the end. The experimental group engaged in eight weeks of training (40 minutes per week) with the M-XRT between these two tests. After the training the participants of the experimental group were asked to describe their opinion about the M-XRT, an online training tool in general and the teaching of x-ray interpretation at their universities. Three hypotheses (detection performance, time of reaction, confidence rating) were investigated in this study.

The detection performance (A') was calculated along the Signal Detection Theory (Green & Swets, 1966) while both softwares (M-CAT, M-XRT) measured the rates of hit, false alarm, correct rejection and miss. The experimental group increased their detection performance from the pretest to the posttest significantly whereas the control group did not show any improvement of their detection performance. Furthermore, the difference of the detection performance between the experimental and control group for the posttest was also found significant. Thus, the CBT had a positive influence on the detection performance on the experimental group. The ratio between the hit- and false alarm-rate states an important role in calculating the detection

performance (Green & Swets, 1966) which was shown by an example of Schwaninger (2005). The present study showed a significant interaction of the within-subjects factor test date with the between-subjects factor group for the false alarm-rate. The experimental group could decrease their false alarm-rate from the pretest to the posttest significantly. Furthermore, a significant difference of the false alarm-rate between the experimental and the control group for the posttest was found. CBT had an influence on the experimental group and could increase their ability to detect if an x-ray image without an PE was shown and classify it as OK. On the other hand, the hit-rate did not show any kind of improvement. One possible explanation for this finding could be that the hit-rate of both groups for the pretest was already quite high and could therefore only slightly improve whereas the false alarm-rate could decrease through CBT. The criterion which can change quickly while the detection performance is a stable trait of a person must be considered in this calculation (Green & Swets, 1966; Schwaninger, 2005). The results of the criterion are not significant which means that the actual detection performance was increased and that not just the criterion changed through as for example the participants deteced more images as NOT OK because they just expected more images with PE. The hypothesis about the detection performance that should increase significantly through CBT can be confirmed. A clear learning progress was shown through CBT, however, 93% of the participants of the experimental group considered their learning progress as moderate and 7% even stated that no progress was made. It seems likely the learning progress was subconsciously while in several questions of the survey the participants mentioned that the learning progress could be increased through more explanation about where the PE is shown in the x-ray picture and information about PE in general.

The hypothesis about the time of reaction stated that the experimental group will take less time to decide whether a PE was shown or not. The time of reaction should decrease through CBT. This hypothesis cannot be confirmed. None of the results about the time of reaction was found significant. However, the experimental group shows a slight decrease of time of reaction between the pre- and the posttest while the control group does not show these changes (Figure 6). Even though the difference was not significant a small influence of CBT on the time of reaction can be assumed. As the participants were able to complete both tests at home it is unclear to say if they needed every time of reaction which was measured by the M-CAT for deciding whether a PE was shown or not which can influence these findings. A study were the M-CATs could be completed in a specified office under the control of the study management could investigate more clearly how much time of reaction is needed for the participants. Another explanation could be that as it is stated in the survey, the experimental group of this study was not aware of their increase of detection performance. If more explanation and information was included in the training sessions which would make the increase of knowledge about interpretation skills more conscious for the participants and a significant decrease of time of reaction could be possible. For further research it should be discussed how important the time of reaction is for x-ray interpretation in general. This study did not investigate if participants that were faster than the others also show a better detection performance. The connection between time of reaction and detection performance should be investigated with the knowledge of expert radiologists to discuss its importance in actual practice.

The confidence rating was measured after every image in the two M-CATs where the participants were asked how confident they are about their opinion whether a PE was shown or not. An estimation between 1 and 5 was possible. The hypothesis that

the confidence rating would increase through CBT cannot be confirmed. None of the results for the confidence rating were found significant. A reason for this finding could be the subconscious learning progress that was made and described by the experimental group. Through the suggestions of the participants (e.g. more information about the PE and how to detect it in the x-ray image) the confidence could increase. Like the time of reaction, a slight increase of the confidence rating can be seen for the experimental group between pre-and posttest, even though the difference was not found significant (Figure 7). A connection between the time of reaction and the confidence rating is possible and should be investigated in a further study including their relationship with the detection performance. The confidence rating should be investigated in further research along with the detection performance because the physicians can detect pathologies better in x-ray images and are confident about their estimation it can have a direct influence whether the patient is considered in need for further medical examinations or not.

The M-XRT in general was liked by 42% of the experimental group while 21% were motivated to train regularly and 36% had fun with the training. The reasons for not being motivated were that the participants had no overview of the training and that the increase of difficulty between level 2 and level 3 was too steep. To improve the motivation and help the participants to get an overview of the training the M-XRT could include an agenda so the participants can see when they performed a training or work with training-reminders. Additionally, the x-ray images need to be sorted more sensibly according to their difficulty that levels could be created which show a steady increase of difficulty. For this reason, further research could focus on what makes an x-ray image in the medical setting difficult to interpret. With this information an adaptive training could be possible which may improve the detection performance even more. The level 4

where the participants had to mark where they suspect the PE was liked by 50%. To mark the specific area seems to be important to the participants and should be part of a further study with CBT and x-ray images under the condition of improving its execution. However, the experimental group sees great potential in an online training tool, 93% would recommend it to a prospective radiologist and 36% would recommend it to every occupational group in the medical sector. Furthermore, they mentioned a lot of different areas in medicine where such online training tool could improve the interpretation skills (e.g. ECG, MRI, CT, etc.). It appears like the students are interested in this kind of training and research should proceeded in this area.

In conclusion, CBT can be a tool to improve the interpretation skills of people who work in the medical sector. This study shows that the detection performance of medical students for PE in an x-ray image can be improved by CBT. A tool which can train the physicians and improves the detection performance can have a positive outcome for the hospitals and also the patients. Through a better detection performance, patients are less likely to be falsely defined as in the need for further medical examinations which can have a positive outcome for patients and hospitals, also seen from a financial point of view. Further research should focus on expanding the CBT to other pathologies which can be detected through different medical imaging methods. An important part of the CBT should be a feedback that contains information about the pathology and how it could be detected in the medical image. This can improve the learning progress and make the participants aware of their progress. For this reason, a strong cooperation between the software company, a hospital which provides the images and knowledge and the university with its students is recommended in order to combine the knowledge and progress of all three areas.

5. References

Aaronson, D., & Watts, B. (1987). *Extensions of Grief's Computational Formulas for A' and B" to Below-Chance Performance*. *102*(3), 439–442.

CASRA. (n.d.). CASRA. Retrieved May 21, 2020, from https://www.casra.ch

Dikshit, A., Wu, D., Wu, C., & Zhao, W. (2005). An online interactive simulation system for medical imaging education. *Computerized Medical Imaging and Graphics*, *29*(6), 395–404. https://doi.org/10.1016/j.compmedimag.2005.02.001

Eisen, L. A., Berger, J. S., Hegde, A., & Schneider, R. F. (2006). Competency in chest radiography: A comparison of medical students, residents, and fellows. *Journal of General Internal Medicine*, *21*(5), 460–465. https://doi.org/10.1111/j.1525-1497.2006.00427.x

Fabre, C., Proisy, M., Chapuis, C., Jouneau, S., Lentz, P.-A., Meunier, C., Mahé, G., & Lederlin, M. (2018). Radiology residents' skill level in chest x-ray reading. *Diagnostic and Interventional Imaging*, *99*(6), 361–370. https://doi.org/10.1016/j.diii.2018.04.007

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley & Sons.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin*, *75*(6), 424–429. https://doi.org/10.1037/h0031246

Karkhanis, V., & Joshi, J. (2012). Pleural effusion: Diagnosis, treatment, and management. *Open Access Emergency Medicine*, 31.

https://doi.org/10.2147/OAEM.S29942

Koller, S., & Schwaninger, A. (2006). *Assessing X-Ray Image Interpretation Competency of Airport Security Screeners*. https://doi.org/10.13140/RG.2.1.1037.2086

Lin, C., Ding, Y., Xie, B., Sun, Z., Li, X., Chen, Z., & Niu, M. (2020). Asymptomatic novel coronavirus pneumonia patient outside Wuhan: The value of CT images in the course of the disease. *Clinical Imaging*, *63*, 7–9. https://doi.org/10.1016/j.clinimag.2020.02.008

Mendes, M., Schwaninger, A., & Michel, S. (2011). Does the application of virtually merged images influence the effectiveness of computer-based training in x-ray screening? *2011 Carnahan Conference on Security Technology*, 1–8. https://doi.org/10.1109/CCST.2011.6095881

Nyhsen, C. M., Steinberg, L. J., & O'Connell, J. E. (2013). Undergraduate radiology teaching from the student's perspective. *Insights into Imaging*, *4*(1), 103–109. https://doi.org/10.1007/s13244-012-0206-8

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric"A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review*, *10*(3), 556–569. https://doi.org/10.3758/BF03196517

Salajegheh, A., Jahangiri, A., Dolan-Evans, E., & Pakneshan, S. (2016). A combination of traditional learning and e-learning can be more effective on radiological interpretation skills in medical students: A pre- and post-intervention study. *BMC Medical Education*, *16*(1), 46. https://doi.org/10.1186/s12909-016-0569-5

Schwaninger, A. (2004). Computer based training: The enhancement of human factors. *Aviation Security International, 2004*, 31–36.

Schwaninger, A. (2005). *Objekterkennung und Signaldetektion: Anwendungen in der Praxis*. https://doi.org/10.13140/RG.2.1.2151.3203

Schwaninger, A., & Hofer, F. (2004). *Evaluation of CBT for increasing threat detection performance in X-ray screening*. https://doi.org/10.13140/RG.2.1.4051.8649

Sha, L. Z., Toh, Y. N., Remington, R. W., & Jiang, Y. V. (2020). Perceptual learning in the identification of lung cancer in chest radiographs. *Cognitive Research: Principles and Implications, 5*(1), 4. https://doi.org/10.1186/s41235-020-0208-x

Young Sonographers. (2017, October 17). Young Sonographers. https://www.youngsonographers.ch

6. List of Figures

7. List of Tables