OPTICAL SPECTROSCOPY & VISUAL ASSESSMENT FOR GRADING ERYTHEMA

INVESTIGATION OF OPTICAL SPECTROSCOPY AND VISUAL ASSESSMENT FOR GRADING ERYTHEMA

By LILIAN DOERWALD-MUNOZ, B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Radiation Sciences-Radiobiology

McMaster University MASTER OF SCIENCE (2019) Hamilton, Ontario (Radiation Sciences-Radiobiology)

TITLE: Investigation of optical spectroscopy and visual assessment for grading erythema

AUTHOR: Lilian Doerwald-Munoz, B.Sc. (McMaster University)

SUPERVISOR: Professor Thomas J. Farrell

NUMBER OF PAGES: xii, 154.

ABSTRACT

Erythema is a well-documented early indicator of tissue injury resulting from exposure to high doses of ionizing radiation. Close monitoring of radiation-induced injury to the skin can help identify opportunities for early interventions that may prevent or reduce more severe reactions. The gold standard for monitoring erythema is visual assessment (VA) by a trained clinician. This method has been criticized for being subjective and designed with very broad category descriptors.

This work introduces a newly developed VA scale called the clinician erythema assessment for radiation therapy (CEA-RT).The reliability and accuracy of the CEA-RT scale was tested among 20 radiation therapists who trained to use the scale on digital images of radiation-induced erythema. CEA-RT demonstrated to be highly reliable when therapist's grades were compared to themselves, but moderately accurate when therapist's grades were compared to each. A follow-up study with real patients and fewer but more extensively trained raters was proposed to demonstrate the grading accuracy of the CEA-RT scale.

As an alternatively to VA, spectroscopy has the ability to monitor erythema by measuring the change in concentration of hemoglobin (Hb) within the vessels of the skin. These changes in Hb concentration are linked to the degree of erythema. This thesis also investigated the use of hyperspectral imaging (HSI) and diffuse reflectance spectroscopy (DRS) as potential technological alternatives for evaluating erythema.

In a second study, Erythema was artificially induced in three volunteers who participated in a pilot study designed to assess the ability of an experimental HSI camera to detect skin changes. The data extracted from the hyperspectral images was found to effectively yield spectral profiles and Dawson's erythema indices (EI) in agreement with the erythema grades assigned by the gold standard therefore showing HSI to be a viable alternative of assessing erythema.

Finally, a third study compared DRS measurements to VA using the CEA-RT scale. The DRS system was previously used to measure in vivo erythema but did not compare spectral measurements to an accepted standard. Ten patient volunteers received daily DRS and VA evaluations for a period of 2 to 4 weeks. The results demonstrated that the Dawson's EI calculated from the spectral data correlated well with the gold standard (VA grades) and that DRS is able to detect changes in the skin throughout the course of radiation treatments.

ACKNOWLEDGMENT

I would first like to thank my principal thesis advisor Dr. Thomas Farrell of the Medical Physics Faculty at McMaster University. His office door was always open whenever I had a mental block and needed some words of encouragement. He had a special talent for perceiving what my difficulties were even when I could not clearly express these difficulties myself. He consistently provided thought provoking feedback yet allowed this paper to be my own work.

I am also very grateful to Dr. Joseph Hayward of the Medical Physics Faculty at McMaster University. He was instrumental in helping me complete the validation study for the Clinician Erythema Assessment tool and to Dr. Qiyin Fang of the Department of Engineering Physics at McMaster University for welcoming me into the Biophotonics lab at McMaster University and providing me with several educational opportunities.

My sincere thanks go to Dr. Raimond Wong of the Radiation Oncology department at Juravinski Cancer Centre who granted me the opportunity to invite patients under his care to participate in my research. Thanks to Dr. D. Ian Hodson who helped me appreciate the art of good photography and post-processing techniques and thanks to Young Lee for her valuable assistance with the CRIE study. Without their valuable support, it would not have been possible to conduct this research.

I will forever be thankful to Ramy Abdlaty for being the best research partner anyone could ever ask for. Ramy, always so kind, reliable and hard working. His enthusiasm for research was contagious. I cannot thank enough Randy and Bud of the machine shop at the Juravinski Cancer Centre for engineering a superb adjustable cart for the experimental camera system.

Finally, I must express my very profound gratitude to my wonderful spouse Fabian, my loving son Martin and daughter Sophia, my extended family, my dearest friend Emilia Timotin and my colleagues at the Juravinski Cancer Centre for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Author

Lilian Doerwald-Munoz

[Title Here, up to 12 Words, on One to Two Lines]

Table of Content

List of Figures and Tables

1. Introduction

1.1. Clinical Motivation and Background

Ionizing radiation is radiation in the form of photons or charged particles with enough energy to remove a bound electron from an atom or molecule, causing the atom or molecule to become charged or ionized. Radiation therapy and interventional radiology (also call interventional fluoroscopy) procedures are associated with skin exposure to significant doses of ionizing radiation. One of the most common effects of exposure to ionizing radiation is dermatitis, which ranges from mild inflammation and redness (erythema) to severe desquamation, ulceration and skin necrosis.

The risk factors for radiation-induced skin injury are many and can be grouped into three main categories: factors associated with patients' medical history, factors associated with the area and type of skin exposed and technical factors attributed to the radiation exposure itself. Much effort has been directed at understanding the harmful effects of radiation exposure and to applying new technologies to help maximize the therapeutic benefit. However, to date radiation induced normal tissue toxicity is still a major limiting factor, particularly in radiation therapy and interventional radiology.

The gold standard for assessing erythema is visual inspection by a trained clinician. Howeever, this method has been criticized for being subjective and for relying on the skill and experience of the assessing clinician. Furthermore, there are several grading systems used in clinical trials for visual assessment of radiation dermatitis. The absence of a global grading scale for skin reactions makes it difficult to compare

outcomes from different studies. This has contributed to the lack of reliable evidence-based knowledge on the optimal management of skin toxicities.

A more systematic and quantifiable approach for monitoring early skin toxicities prior to their full development may offer new opportunities for non-expert clinicians to appropriately evaluate reactions. Additionally, an objective and uniform measure for radiation induced skin reactions would simplify comparing research results evaluating the efficacy of novel therapeutic agents for the management of skin toxicities.

Technology capable of detecting changes in the optical properties of skin such as diffuse reflectance spectroscopy (DRS) and hyperspectral imaging (HSI) have the potential to provide quantitative and objective assessment of structural and functional changes in tissue such as skin erythema. Furthermore, in the future, objective methods may prove to be helpful as a prediction test to adjust radiation doses based on individual patient sensitivity. The purpose of this work is to explore the feasibility of using DRS and HSI technology as an objective alternative to visual assessment for grading erythema in the clinical setting.

1.2. Historical Background

The discovery of x-rays in 1895 and radioactivity in 1896, led to widespread use of x-rays and radioactive isotopes for their many perceived medicinal benefits. At the time, scientists and clinicians were unaware of the serious biological damage resulting from unregulated use and exposure to ionizing radiation. There were no instruments to measure the strength of the radiation fields. Instead, the calibration of x-ray tubes was

based on the amount of skin reddening (erythema) produced when the operator placed a hand directly in the x-ray beam. [1] In fact skin erythema dose (SED) was an early unit of radioactive dose equal to the dose that slightly reddens or browns the skin of 80% of all persons within 3 weeks after exposure ($\approx$ 600 Roentgens). [2, 3] Early scientists Pierre and Marie Curie shared with Henri Becquerel the 1903 Nobel Prize for Physics for their work with ionizing radiation. [4] By then the dangers of radiation were becoming apparent as Pierre Curie warns in his Nobel Prize lecture:

"If one leaves a small glass ampulla with several centigrams of radium salt in one's pocket for a few hours, one will feel absolutely nothing. But 15 days afterwards redness will appear on the epidermis, and then a sore, which will be very difficult to heal. A more prolonged action could lead to paralysis and death." [1]

History tells us that pioneer scientists in the radiation field such as Henri Becquerel and Marie Curie met early deaths attributed to radiation overexposure. Many less known scientists, laboratory assistants, patients and factory workers also suffered the effects of radiation. Permanent hair loss, skin scaring and disfiguration, amputation of fingers and limbs were common occurrences. [5] Those who were exposed and lived long enough, endured numerous painful surgeries, only to later die from infection or metastatic cancers induced by radiation exposure. [6]

More than a century later x-rays and radioactive isotopes continue to be used therapeutically in medicine. Furthermore, the use of ionizing radiation for the purpose of diagnostic imaging and interventional procedures is increasing exponentially over the

years. [7,8] Consequently, concern for appropriate monitoring and safeguarding from excessive exposure in the general population is still an actively researched topic. [9-12]

1.3.Radiation Therapy

Radiation therapy is the use of ionizing radiation in the treatment of patients with malignant and occasionally benign neoplasias. The aim of radiation therapy is to deliver a precisely measured dose of irradiation to a defined tumour volume with minimal damage to surrounding healthy tissue. The endpoint is to eradicate the tumor or prolong survival with a high quality of life. [13]

Generating a therapeutic effect within a given tissue is a complex balance of technical and biological factors. Important technical factors determining the effect on irradiated tissue include the energy of the radiation beam, the delivery technique, the total radiation dose and the size and number of fractions (or partial treatments). For example, unwanted dose to normal tissue can be controlled by careful selection of the radiation delivery technique. Sophisticated dynamic-beam techniques such as IMRT (intensity modulated radiation therapy) and VMAT (volumetric modulated arc therapy) are able to deliver highly conformal doses to tumour volumes within the body and very low doses to normal structures found in close proximity to the tumour. [14] To accomplish such conformity the radiation beam is divided into hundreds of smaller beamlets with varying intensity. Each of these radiation beamlets are optimized and delivered under computer control.

Electrons and low energy x-rays

High-energy radiation in the form of megavoltage x-rays (≥1 MV) is typically used in IMRT and VMAT. The advantage of megavoltage over kilovoltage x-rays (<500 kVp) is that with megavoltage irradiation, the dose absorbed at a depth within the body is considerably greater than the dose absorbed at the surface of the body. Although, low skin dose is normally not the main goal of radiation therapy careful selection of treatment technique is critical for optimal tumour dose delivery without exceeding normal tissue tolerance.

The relative low surface dose compared to the maximum dose at a depth is referred to as the skin sparing effect of radiation. Under the surface of the skin, the dose rises gradually until it reaches a maximum value at a characteristic depth and then, decreases almost exponentially with depth until it is completely absorbed or exits the body. [15] The slow build-up of absorbed dose is responsible for the skin sparing effect for high-energy photon beams.

The low dose region between the skin surface and the depth of maximum dose built-up results from Compton electrons released within the tissue by photon interactions (Compton effect) traveling a certain distance within the tissue before they deposit all their energy. [15] Surface dose sparing is an important feature of megavoltage beams when trying to limit skin toxicities.

Kilovoltage x-rays (which include superficial x-rays 20 to 150 kVp and orthovoltage x-rays 150-500 kVp) and low energy electron beams (4-12 MeV) do not exhibit the same skin sparing effect seen with megavoltage. In general, most of the energy

is absorbed within a depth of less than 5cm, and the surface dose ranges from 80-100% of the maximum absorbed dose. [15]

Patients with early stage skin cancers (i.e. limited lesion size and depth of invasion with no nodal involvement) who are not suitable for surgical resection may be treated successfully with radiation therapy. Local control at 5 years for basal cell carcinoma (BCC) and squamous cell carcinoma (SCC) is 95% and 85% respectively for lesions less than 5cm when treated with radiation therapy alone to a dose of 40 to 60 Gy. [13] Most early skin cancers that are less the 2cm in size are managed with orthovoltage RT with beam energies of 100 to 250 kVp. The advantage of this technique, compared with electron beam, is that the maximum absorbed dose is at the skin surface. Larger skin lesions are treated with electrons and may require a 1-1.5 cm of bolus (tissue equivalent) thickness over the lesion for maximum dose at the skin surface. [13]

Often the skin is not be the primary target, yet it is the outer protective layer of the body and therfore frequently injured as an ''innocent bystander''. This is true in radiation therapy when irradiating non-skin cancers, but also for invasive interventional procedures. Serious radiation induced skin injuries have been reported after unexpectedly high doses of kilovoltage irradiation exposure during fluoroscopic imaging, including cardiac catheterization and vascular embolization. [16, 17, 18]

Dose, fractionation & cell damage

In radiation therapy, a fractionated dose is a therapeutic dose of ionizing radiation divided into smaller doses (fractions). Each fractional dose is given at predetermined regular intervals over a set number of days adding up to the total

prescribed dose. Fractionation is commonly used in radiation therapy to exploit the difference in tissue sensitivity between normal cells and cancer cells. The biological factors determining the effect of ionizing radiation are governed by the four Rs of radiobiology; that is, repair of sublethal cell damage, repopulation of depleted cells, redistribution of cells through the sensitive stages of the cell cycle, and tissue reoxygenation to fixate cell damage by way of reactive intermediates. [19, 20]

The biologic effect of ionizing radiation is mainly a consequence of DNA damage, which may be caused by direct ionization within the DNA molecule or indirectly from interaction of DNA with highly reactive chemical radicals formed by local ionizations in water. The common forms of DNA damage are base damage, cross-links, single and double-strand breaks, and complex combinations of all of these. [19, 20]

Normal tissue is able to repair most damage (nonlethal damage) within hours, whereas tumour cells are rapidly proliferating cells and therefore more likely to sustain lethal damage during the sensitive stages of the cell cycle (Mitosis and Gap 2 phase). [19] Fractionated doses take advantage of tumour biology to induce a higher rate of progressive reduction in the number of surviving viable tumour cells with each successive fractional dose compared to normal cell. [19, 20] The success of the treatment greatly depends on the total dose delivered to the tumour. For most cancers, the dose required to control the tumour ranges from 45-80 Gy and the fractional dose is 1.8-2 Gy, depending on the type tumour cell and physiology. [13]. In general, since normal cells repair a significant proportion of radiation-induced DNA damage within a few hours, fractionated irradiation will also result in the reduction of normal tissue toxicities [21].

1.4. Erythema reaction

Despite efforts to minimize skin toxicities, radiation damage to the skin represents one of the most common side effects hindering the application of radiation therapy. Skin changes after radiation exposure follow a predictable course determined by the radiation dose, fractionation, and the biological response to tissue injury. A continuous state of skin renewal that occurs approximately every 26-28 days for a middle-aged adult makes the skin more vulnerable than other organs to ionizing radiation injury. Skin injuries ranging from erythema to full chronic wounds occur in about 95% of patients who undergo radiation therapy. [22] The predominant cell type found in the epidermis (outermost layer of the skin) is the keratinocyte, representing 90% of all cells. The primary function of keratinocytes is the formation of a protective barrier against damage by pathogens, UV radiation and water loss. [22, 23] Unlike other forms of mechanical or chemical damage to the skin, radiation therapy-induced injury is repetitive and accumulates throughout the course of treatment, leading to a delayed disruption of the epidermal barrier. This causes a reaction that occurs slowly, and may not be symptomatic until weeks after the initiation of treatment.

Depending on the extent of exposure, the clinical symptoms of erythema reaction can occur and resolve within hours. Initially, when the epidermis is damaged, keratinocytes react by producing proinflammatory mediator proteins (i.e. cytokines), which trigger vasodilation and increased permeability of blood vessels enhancing the recruitment of white blood cells and other inflammatory cells to begin the healing

process. Histologically, the vasodilation and transient increase in capillaries permeability will result in mild erythema and edema. [23, 24, 25]

The threshold dose for initial erythema is 2 Gy or greater (Table 1-1); therefore, erythema it is often seen in regimens with high superficial doses aimed at treating skin cancer, metastasis invading the skin surface, or hypofractionated treatment (involving fractions 6 Gy or larger). [16] Kilovoltage (20-500 kVp) irradiation, used for superficial radiation therapy and complex diagnostic interventions, is likely to exceed the erythema threshold due to the lack of skin sparing effect compared to megavoltage (1 MV or higher) irradiation.

Most early skin changes after radiation therapy are superficial and reversible; yet, acute radiation dermatitis remains a serious side effect, which may limit radiological interventions or the delivery of therapeutic doses of radiation to the tumour. Erythema can occur hours after radiation exposure and then fade within hours to days. [24, 25] A second phase of erythema is apparent 10 to 14 days after exposure. It is characterized by intact reddened skin, blanching to pressure, and most likely an inflammatory response mediated by cytokines. [20, 22, 23] Initial changes include mild generalized erythema, dry desquamation, pruritus, epilation, and scaling. Skin dryness and hair loss are secondary to injury to sebaceous glands and hair follicles. [19] More severe changes consisting of persistent tenderness or inflammation may progress to blisters and focal breakdown of the epidermis resulting in moist desquamation, particularly in skin folds. [21, 23] Severe reactions usually occurs after 4 to 5 weeks of therapy, with radiation doses to the skin of 40 Gy or greater (Table 2-1). Moist desquamation is characterized by

the loss of epidermal layer, serous exudates, crusting and considerable pain. Acute skin reactions peak 1 to 2 weeks after the last treatment and then gradually heal. [21, 23]

The care of radiation dermatitis is generally not different from generic wound care. The standard of care for managing radiation dermatitis involves the use of moisturizing creams, avoidance of trauma to the skin or exposure to extreme temperatures, the use of protective soft non-adherent and absorbent dressings, as well as application of topical antibiotic and steroid creams. [22, 27, 28] If the patient is in a lot of discomfort, radiation therapy treatments may be modified to mitigate skin tissue injury. The dose to the skin can be reduced by removing dose build-up devices (i.e. bolus) and, if necessary, treatments can be temporarily halted so the skin has a chance to heal.

Table 1-1.Dose-dependent acute cutaneous findings after local radiation exposure

Observed acute skin reaction	Dose threshold (Gy)	Onset of findings
Temporary erythema	2	Hours
Faint erythema and epilation	6–10	7–10 days
Defined erythema & hyperpigmentation	12–20	2–3 weeks
Dry desquamation	20–25	3–4 weeks
Moist desquamation	30–40	4 weeks or more
Ulceration	>40	6 weeks or more

Note: Adapted from Ryan JL. Ionizing Radiation: The Good, the Bad, and the Ugly. The Journal of investigative dermatology. 2012;132(302):985-993. doi:10.1038/jid.2011.411. [16]

1.5. Visual assessment of erythema

Most grading systems used in clinical trials for evaluating radiation-induced skin toxicity are based on visual assessment (VA) by a trained clinician. Well-known examples of VA grading tools are the Radiation Therapy Oncology Group (RTOG) [29] scale, the Common Terminology Criteria for Adverse Events (CTCAE) [30, 31] scale and the Radiation-induced Skin Reaction Assessment Scale (RISRAS) [32]. Although often criticized for their subjective nature, visual grading tools for skin reactions are widely accepted throughout the oncology community as the gold standard for classification and grading of the severity of adverse events. [33, 34] One difficulty with using visual grading scales is that the degree of skin toxicity has been categorized using a numerical analog scale with discrete grades for which the boundaries between grades are broad and uncertain (e.g., faint, brisk, mostly confined, and life threatening) (Table 1-2).

Table 1-2. Common Terminology Criteria for Adverse Events (CT-CAEV 5.0). Grading scale for dermatitis associated with radiation or chemoradiation

Grade	Description
1	Faint erythema or dry desquamation
2	Moderate to brisk erythema; patchy moist desquamation, mostly confined to skin folds and creases; moderate edema
3	Moist desquamation other than skin folds and creases; bleeding induced by minor trauma or abrasion
4	Life-threatening consequences; skin necrosis or ulceration of full thickness dermis; spontaneous bleeding from involved site; skin graft indicated
5	Death

Note: Common Terminology Criteria for Adverse Events (CTCAE) Version 5.0 Published: November 27, 2017 [30]

In order to characterize early radiation induced erythema in more detail, it is necessary to consider a grading scale that is more responsive to subtle changes in the amount of skin redness. Dermatology is the branch of medicine concerned with diseases of the skin, and as such offers several tools for facial erythema assessments. The Clinician Erythema Assessment (CEA) is a 5-point (0-4) visual assessment scale (Table 1-3) used by dermatologists that has been validated for grading erythema linked to rosacea. [36] Currently, there is no literature supporting a similar visual assessment tool for grading early radiation induced erythema. In chapter 3, a modified version of the CEA is investigated as a possible standardized measure for grading radiation-induced erythema.

Table 1-3. Clinician Erythema Assessment (CEA) scale

Erythema Grade	Description
0 = Clear	Clear skin with no signs of erythema
1 = Almost clear	Almost clear; slight redness
2 = Mild	Mild erythema, definite redness
3 = Moderate	Moderate erythema; marked redness
4 = Severe	Severe erythema; fiery redness

Note: The CEA visual assessment scale was validated for grading rosacea. [36]

2. Optical Methods for Assessing Erythema

Table 2-1	*List of Acronyms and Abbreviations*
DHb	Deoxy-hemoglobin
DRS	Diffuse Reflectance Spectroscopy
EI	Erythema Index
Hb	Hemoglobin
HSI	Hyperspectral Imaging
LIR	Logarithmic Inverse Reflectance
OHb	Oxy-hemoglobin
ROI	Region Of Interest
RGB	Red-Green-Blue
S/N	Signal to Noise ratio

2.1. Optical properties of skin

Knowledge of the optical properties of the skin is essential to understanding the application of spectroscopy for evaluating erythema changes. Depending on the wavelengths of the incident electromagnetic radiation, such as light or x-rays, the interaction with biological tissue will result in photons being reflected, absorbed, scattered or transmitted through the medium. The likelihood of each type of tissue-photon interaction is partially due to particles known as chromophores, which are embedded within tissue and give the tissue its optical properties. Melanin, oxy-hemoglobin (OHb), deoxy-hemoglobin (DHb), betacarotene and bilirubin are typical examples of protein-based chromophores found in skin.

A chromophore absorbs electromagnetic energy with distinctive efficiency depending on the wavelength of the incident radiation. Therefore, presence of these

chromophores will vary the intensity of radiation at specific wavelengths across the electromagnetic spectrum resulting in characteristic absorbance spectra. [37] The main chromophores of skin that absorb photons within in the visible region of the electromagnetic spectrum 450-800nm, are OHb, DHb, and melanin [38, 39].

Relative absorbance spectra for these three most abundant skin chromophores are plotted in logarithmic scale (Fig. 2-1) to show that Hb absorbance decreases by three orders of magnitude going towards longer wavelengths, whereas melanin absorbance decreases only by one order of magnitude. Melanin absorption generally affects the reflectance spectra over a broad spectral range; it has no characteristic maximum in the visible region but demonstrates a uniform increase towards shorter wavelengths. OHb absorption shows two characteristic peaks at approximately 542 and 577 nm and then decreases sharply in the red region (>600nm). [38, 39] DHb on the other hand, has a notable absorption peak at 555 nm in the green region and an absorption decline at approximately 736 nm. [38, 39]
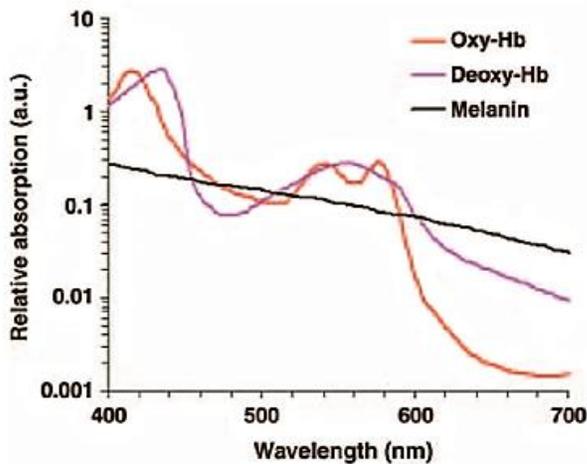
Figure 2-1. Absorbance spectra for typical skin chromophores OHb, DHb and melanin plotted in logarithmic scale. Obtained from Stamatas et al.

2.1.1. DRS Instrumentation

Reflection is the return of radiation by a surface, without a change in wavelength of the radiation. In reflectance spectroscopy experiments, the source light collected by the spectrometer has been affected by the sample being measured. Generally, the light from the source passes through, is absorbed or is reflected off a sample and then enters the spectrometer where it is measured. The characteristic spectrum of the light source is then compared to the spectrum of the light reflected off the sample. By comparing both spectra a determination can be made about how the sample affects the characteristics of the incident light and in turn derive optical properties of the sample; in this case in vivo human skin.

Diffuse reflectance measurements presented in this work were acquired using the same system described by Glennie *et al*. [40] The main components of this DRS system include a light source, a spectroscope, a light probe (in this case an integrating sphere), a processor and optical fibres (Fig 2-2A). The port opening of the integrating sphere was 15.2 mm in diameter, and was fabricated from a cube of Spectralon® measuring 52.6 mm x 52.6 mm x 40.3 cm. The optic fiber carrying the light source signal was connected to the side of the sphere, so that the light entering the sphere first scattered off the sphere wall to produce a diffuse illuminating source for the sampling port. A Blu-Loop Multi-LED light source with four different LED bulbs (blue, cyan, white 4500K and white 6500K) was selected for this study. The Blu-Loop Multi-LED source emitted a well-balanced white light spectrum over the visible range (Fig 2-2B). The light reflected from the sample and integrating sphere was directed by optical fibres to the USB4000

Fiber Optic Spectrometer. The spectrometer was controlled by a laptop computer on which spectral processing software OceanView 1.5.2 was installed. The spectrometer measured the signal intensity in counts, which is proportional to the number of photons collected by the detector. The processing software converted the spectroscope signal into reflectance data in tabular and graphical form. The main components for the DRS system are listed in Table 2-2.
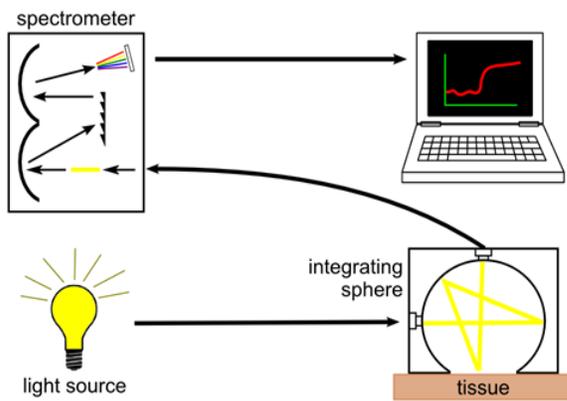
A) DRS system set-up   *B) Blu-Loop Light source spectrum*
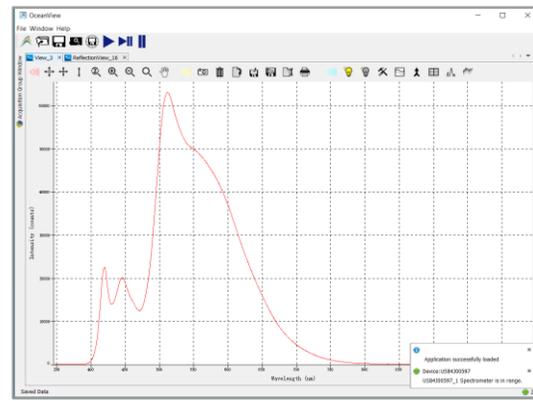
Figure 2-2. *A) Diagram of the main components of the diffuse reflectance spectroscopy (DRS) system. [40] B) Intensity vs wavelength graph of the spectral distribution for the Blu-Loop light source (400-700 nm).*

Table 2-2. *The main components of the DRS system*

Probe	In-house single integrating sphere with a 15.2 mm sampling port set in a cube of Spectralon® measuring 52.6 mm x 52.6 mm x 40.3 cm. [40]
Software	OceanView® Software 1.5.2 for processing reflectance data loaded on a Dell Inspiron laptop with Windows 8 operating system.
Spectroscope	OceanOptics USB4000 spectrometer (200 nm to 1100 nm range, optical resolution ~0.3 (FWHM), pixel size 8µmx200µm).
Light source	OceanOptics Blu-Loop Multi-LED light source (white light 400-700 nm range).
Fiber optic	Thor Labs optical fiber leads (2x) 400µm core size, NA= 1.22.

2.1.2. DRS measurements

Data acquisition parameters for the DRS system were selected according to the recommendations published on the OceanOptics website. [41] The vendor provides users with instructions for setting up and performing experiments with the USB4000 Spectrometer that have been tested in the field and produce accurate and reproducible results. As recommended by OceanOptics, the Blu-Loop light source was warmed up for 15-20 minutes to ensure it had reached thermal equilibrium and measurements were stable. OceanView Software 1.5.2 for Windows 8 by OceanOptics was used for obtaining reflectance data. The integration time was initially set to 10 ms and then adjusted using the diffuse white standard (SRS-99%) reference intensity spectrum peaks to 80% - 90% range of the maximum intensity value as recommended in the user manual. [42] Reducing the integration time reduces the intensity of the spectrum peaks; this way the spectrometer signal does not become saturated and the measurements benefit from the full range of the spectrometer's capabilities improving the signal to noise (S/N) ratio.

In order to correct for undesirable sources of signal that could compromise useful data, a background spectrum was acquired. These extra sources of signal other than the light source can include ambient light, stray reflections, thermal noise, hot pixels, and instrument response. The background spectrum was used to measure and subtract the undesirable signal from the spectral data. The background measurements were obtained by blocking all the light from the Blu-Loop source at its origin with a small piece of thick black cardboard and aiming the opening of the integrating sphere towards a dark corner of the room. In addition to the background spectrum, a white reference spectrum was also

obtained before each set of measurements. The white reference spectrum establishes the characteristics of the light source spectrum with no sample present. The reference spectrum was obtained by placing a diffuse white standard (SRS-99%) over the sampling port of the unobstructed light source.

The number of scan averages was set to 10 to strike a balance between a high scan average for a better S/N ratio and a short acquisition time for more accurate results. The S/N improves with the square root of the number of scans averaged, in this case by a factor of 3.2. The boxcar value was set to 5. The boxcar value is similar to a moving average calculation applied to the wavelength; hence, a boxcar value of 5 averaged an additional 5 pixels on each side of the midpoint (11 pixels in total) and assigned that average to the center pixel of the detector. This feature smoothed the spectral curve without affecting resolution.

The OceanView Software provides a reflection spectroscopy mode designed for automatic processing for transmission and reflection spectroscopy since the calculations necessary to compute both properties are identical (Eq.1). OceanView calculates the reflection of a sample using the following equation:

$$\% \, T_\lambda = \frac{S_\lambda - D_\lambda}{W_\lambda - D_\lambda} \, x \, 100\%$$

$T_\lambda = \textit{Transmition or Reflectance intensity at wavelength } \lambda$

$S_\lambda = \textit{Sample intensity at wavelength } \lambda$

$D_\lambda = \textit{Dark} - \textit{Background intensity at wavelength } \lambda$

$W_\lambda = \textit{White} - \textit{Reference intensity at wavelength } \lambda$

(1)

2.1.3. Calculation of Erythema Index (EI)

The Dawson Erythema Index is a well-established model for analyzing spectral reflectance.[43] Dawson *et al* describes a double peak for hemoglobin between 510 and 600 nm as a unique characteristic of the logarithmic inverse reflectance (LIR) spectrum of skin. Furthermore Dawson *et al* considered that the contribution of hemoglobin to the spectrum of skin could be quantified by establishing a baseline created by joining the points on the spectrum at 510 and 610 nm, and that the area under the curve and above that baseline was a function of the hemoglobin content of the skin (Fig 2-3A&B).
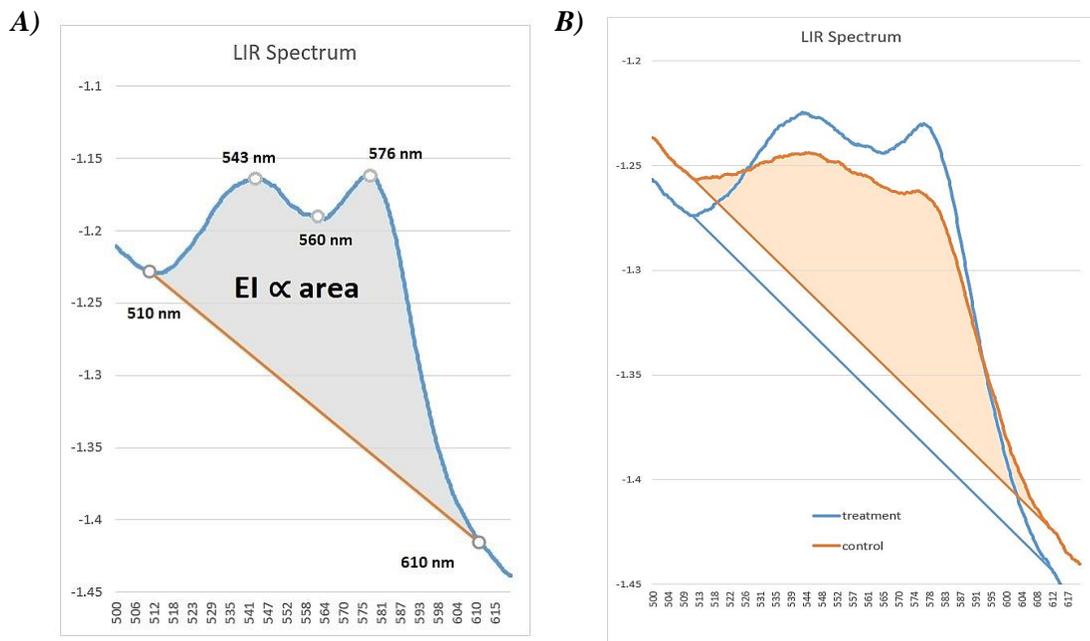
Figure 2-3. *A) Diagram illustrating the Dawson's EI defined as a parameter proportional to the area under the LIR curve between 510 nm and 610 nm. B) LIR spectra for normal skin (control) and skin treated with radiation therapy (treatment). Dawson et al attributes the change in spectrum to variation in hemoglobin concentration. In this case, the hemoglobin variation is likely a result of inflammation caused by radiation-induced skin toxicity.*

Dawson *et al* proposed that a parameter proportional to this area referred to as erythema index (EI) could be used as a response to changes in the spectrum of skin resulting from variations in hemoglobin. The EI is calculated using equation (2) where p, q, r, s and t are the logarithm on base 10 of the inverse reflectance values at wavelengths 510, 543, 560, 576 and 610 nm respectively [43].

Dawson's equation for erythema index (EI):

$$EI = 100\,[r + 1.5(q + s) - 2(p + t)] \qquad\qquad (2)$$

For ease of formulating the EI calculation using Microsoft Excel 2013 the original equation (2) was expanded and rearranged as the sum of five factors, one for each LIR value corresponding to each of the wavelengths of Dawson's equation (3).

$$EI = -200 * p + 150 * q + 100 * r + 150 * s - 200 * t \qquad\qquad (3)$$

2.2. Hyperspectral Imaging (HSI)

Hyperspectral imaging (HSI) is a promising modality for monitoring optical properties of skin. The DRS system described earlier computes a single combined spectral profile for any given area measured by the sampling port, whereas HSI adds spatial resolution and provides more detailed spectral data in one modality. Technological advancements in sensor optics fabrication, lower costs, smaller component size and faster

acquisition times have made HSI a viable option for clinical applications. [44] A hyperspectral camera effectively divides the spectrum data into a number of thin image slices depending on the system's design. The acquired data are organized in a 3-dimensional structure referred to as a data cube. Figure 2-4 shows an example of a data cube composed of a stack of monochromatic images sorted by wavelength. Each image in the data cube was built from a distinct spectral band.

The RGB images are limited to the three major visible bands (red, green, and blue) and thus offer less information for differentiating between elements with close spectral features. [45] In contrast, the set of hyperspectral images or data cube reveals spectral structures of the specimen that may not be evident by visual assessment or by RGB (red-green-blue) camera photography.
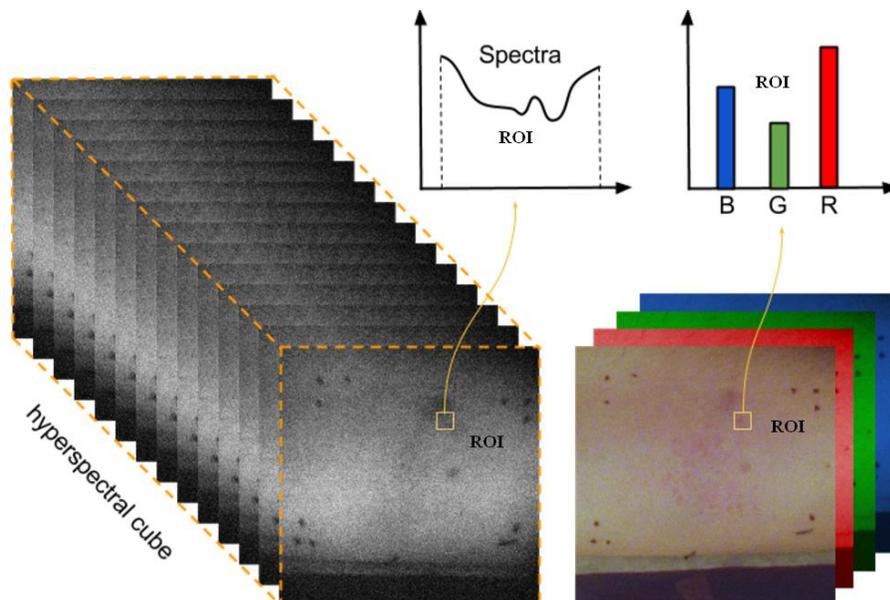
Figure 2-4. *A representation of staked hyperspectral images or data cube and corresponding RGB image data. The image shows artificially induced erythema on the inside of the forearm of a volunteer.[45] The intensity graphs for the square ROI is not a true representaion of the spectral data.*

HSI is capable of overcoming the individual limitations of DRS (lack of spatial resolution) and RGB imaging (poor spectral resolution) and in this way, HSI may provide new clinically relevant information. Furthermore, additional features of HSI are that it requires no skin-contact and is capable of studying considerably large regions of interest. [44]

An exploratory study was performed by Abdlaty *et al* to examine the feasibility of HSI and data analysis for quantifying erythema. A brief description of the HSI system and results are reviewed in the next section. For more detailed information refer to the doctoral thesis by Ramy Abdlaty. [44]

2.2.1. HSI study on artificially induced erythema

The exploratory study was designed to examine the feasibility of using HSI for detecting erythema as a more objective modality compared to the gold standard (visual assessment). The objective set for the study was to compare erythema grades assigned by a clinical specialist by visual inspection using the CEA-RT scale to the Dawson erythema index (EI) calculated from the spectral information obtained using the HSI system.

HSI instrumentation

The HSI system developed for experimental measurements on paatints used an acousto-optical tunable filter (AOTF) technology. An AOTF consists of a crystal in which radio frequency acoustic waves are used to separate wavelengths of light from a broadband source. The system's spectral range was 450-800 nm. Figure 2-5 shows a schematic drawing for the AOTF-based hyperspectral imaging (HSI) system.
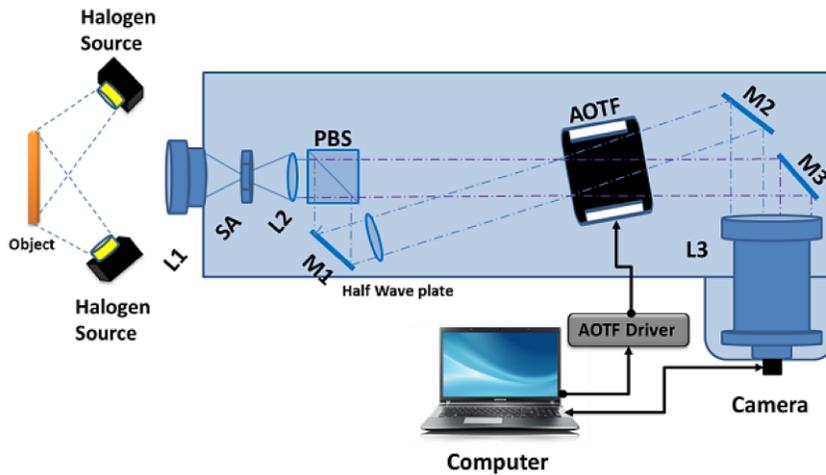
Figure 2-5*. A schematic drawing for the AOTF-based hyperspectral imaging (HSI) system showing lenses L1 -L3; SA: square field aperture; M1- M3: flat mirrors; PBS: polarizing beam splitter; AOTF: acousto-optic tunable filter; camera for capturing hyperspectral images, and a computer for system control and data storage. All the components were fixed to a metal platform and protected within a metal housing to ensure portability.[44]*

The system's image detector was a monochromatic complementary metal oxide semiconductor (CMOS) with an active area of 11.24 mm2 (2048×2048 pixels). The CMOS sensor has the advantages of being compact and low cost with respect to comparable charge-coupled device (CCD) sensors. The major components for the HSI system are listed in Table 2-3.

Table 2-3. The main components of the HSI system

Light source	Three halogen lights totaling 550 W were used for subject illumination.
Acquisition Software	C-language based interface was developed for HSI image acquisition. [44] It functioned to synchronize the AOTF and the camera operation simultaneously. The workstation was an Intel i7 dual-core processor ASUS laptop operated by Windows 7 Enterprise.
Filter	AOTF (acousto-optic tunable filter) highest spectral resolution 1.5: 4 nm varying inversely and nonlinearly within visible to near-infrared (VIS-NIR) range of 450: 800 nm.
Filter driver	8-channel digital radio frequency (RF) synthesizer (MSD040-150-0.2ADM-A5H-8 x 1, Gooch & Housego, FL, USA). The RF signal used range was 65-135 MHz.
Sensor	CMOS (MQ042RG-CM, Ximea, Munster, Germany) monochromatic 11.27mm x 11.27 mm sensor format is one-inch2 with a resolution of 2048 x 2048 pixels and pixel pitch of 5.5 µm.
Zoom Lens	Spectral camera used a Canon EF-S 18-55 mm f/3.5-5.6 STM Zoom Lens
RGB Camera	16-megapixel compact digital color video camera, HYUNDAI, HDMI-768, China.
White Reference	A white standard calibration surface with 99% reflection, SRS-99, Labsphere, NH, USA.
Image processing software	Open source Image-J freeware was used to process and contour regions graded for erythema. Image-J bundled with 64-bit Java 1.8.0_112 (70 MB). Downloaded from https://imagej.nih.gov/ij/download.html.

Study Design

Erythema was artificially induced on the forearm of 3 healthy volunteers. Two males volunteers and one female volunteer, all from different backgrounds; Asian, North African, and Caucasian respectively. The volunteers' age ranges from 35 to 50 years. Participants were informed of the details of the study including the induction of transient erythema in the inner aspect of the forearm. The study received approval from the local research ethics board and all participants signed informed consent. Figure 2-6 illustrates the HSI system setup used to carry out the clinical research on the volunteers.
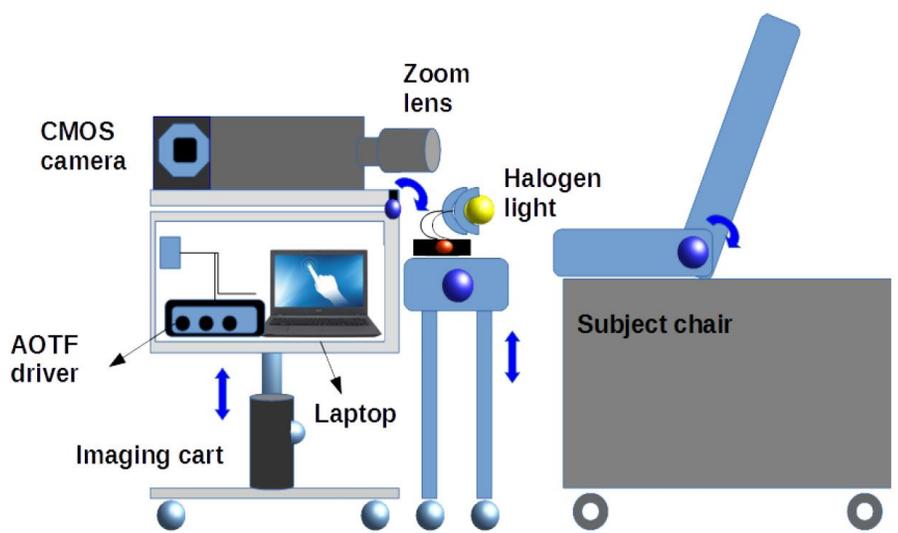
Figure 2-6. *Drawing showing the setup used to acquire HSI. The hyperspectral camera was mounted on a cart, which enclosed the control unit (laptop and AOTF driver). Two commercial halogen lights illuminated the target bilaterally and a 3rd halogen light was placed next to the HSI camera (not shown). The subject was positioned on a clinical chair in front of the camera. The diagram was obtained from Abdlaty et al. [44]*

Clinical Data Collection

The inner aspect of the forearm was selected as the region of interest (ROI) for this experiment because it was a convenient area of even skin that was easy to reposition and image repeatedly. The ROI was a rectangular area delineated with a marker of a colour other than red so it would not interfere with erythema assessments (Fig.2-7).

Baseline images of the ROI were taken using the HSI system and RGB camera before inducing erythema. The RGB camera was oriented in the same direction and under same lighting condition as the HSI camera. A similar orientation between the HSI and the RGB cameras would facilitate feature comparison between the images of both modalities. Following baseline imaging, erythema was artificially induced in the ROI. Volunteers were asked to use a plastic ruler to strike the skin in rapid succession for

approximately 3 minutes until the skin would become bright red. Immediately after

erythema was induced, the ROI was imaged with both technologies at regular intervals

until the erythema reaction resolved.

Erythema was assessed by visual inspection using the CEA-RT (Table 3-1)

grading system. A radiation clinician trained to use the CEA-RT rated erythema within

the ROI after each successive cycle of HSI and RGB images were obtained. The

reliability of the CEA-RT for grading erythema reactions was established in a prospective

study among radiation therapists using RGB images. The results of that work is presented

in chapter 3 of this thesis.

A white reference data cube was collected right after the volunteer was imaged

and a dark data cube was collected with the lens cap on and the room lights off. Eighty-

nine narrow band images (spectral range of 450-850nm) were captured in each data cube.

The integration time for each image was one second, and thus acquisition time for each

data cube was 89 seconds.

Following data acquisition, the clinician annotated the RGB images using

Image-J software. Erythema regions of equivalent grade corresponding to the CEA-RT

scores assigned during visual inspection were delineated within the ROIs. These

annotated regions of equivalent erythema were used for analyzing the data cube (Fig.2-7).
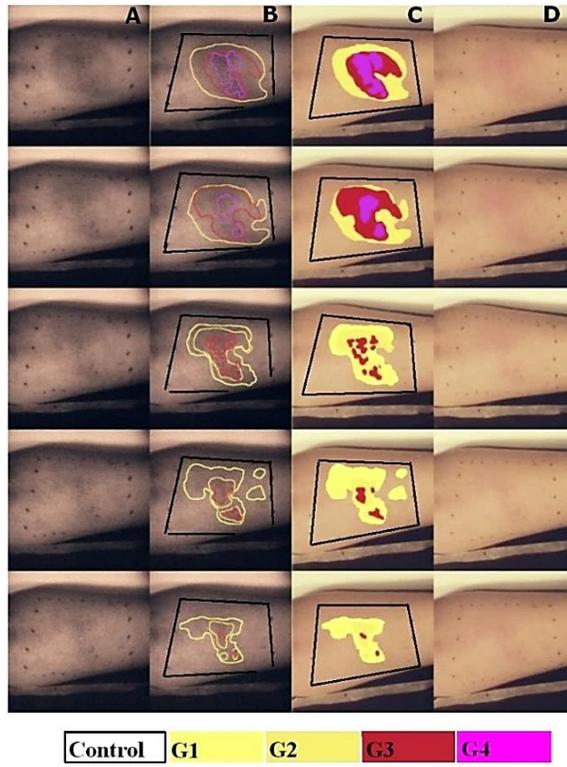
Figure 2-7. Skin annotations for each distinct erythema grade are overlaid on the RGB and HSI images for one volunteer. The image sequence from top to bottom is based on the time of acquisition. The top row is the initial induction of erythema and the bottom row is the final resolution. Columns A and B are the hyperspectral images captured at 540nm (absorption peak for OHb. Columns C and D are the corresponding RGB images. Erythema contours are displayed in columns B and C. Erythema was graded using the CEA-RT scale (Table 3-1). The regions are colour coded based on grade. Minimal erythema (G1)–yellow, mild erythema (G2)–orange, moderate erythema (G3)–red and severe erythema (G4)–magenta. The region within the ROI borders (black) but outside the annotated regions is the control region (G0). [44]

Calculation of Erythema Index

In order to analyze the data cubes, the raw data were converted to reflectance spectra. Similar to DRS, reflectance change was computed using the equation (1) for Dawson's EI. For each data cube, the dark spectra was subtracted from the ROI spectra and divided by the white reference spectra minus the dark spectra. The relative HSI reflectance was further corrected for the patterns of uneven illumination and fixed pattern noise of the sensor ("hot pixels"). The noise correction was accomplished by subtracting a low output image slice corresponding to 446 nm wavelength.

Following image processing, the mean of the pixel intensities (relative reflectance) was determined using Image-J software for each region of equivalent erythema grade contoured by the clinician. The region of skin outside the erythema

contours but inside the ROI borders was considered normal skin and assigned as the control or G0 (no erythema). The overall reflectance spectrum for each erythema region corresponding to all five CEA-RT grades is shown in Figure 2-8. There is a noticeable decrease in reflectance for the curve representing severe erythema (G4) compared to normal skin (G0) within the characteristic range for OHb (542 nm, 577nm) and DHb (555 nm) absorption peaks. [44]

Results
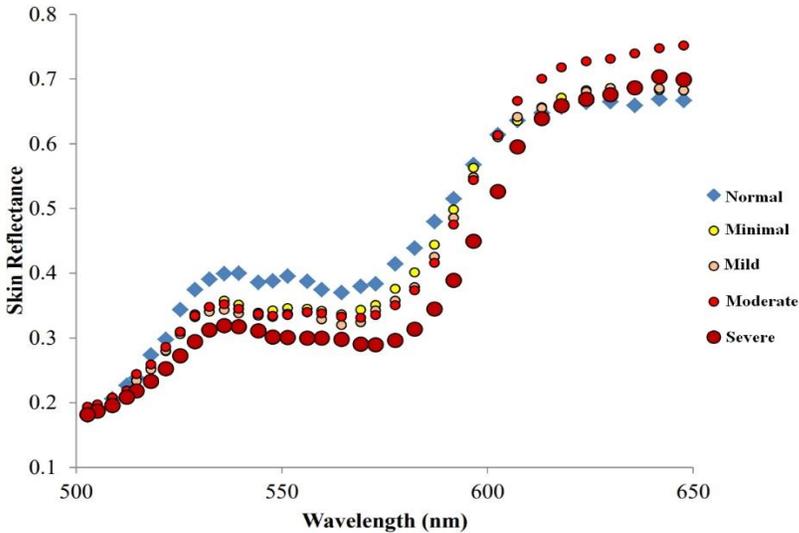
Figure 2-8. Reflectance data for the visible spectral range (500 to 650nm) of the skin obtained by HSI of the same volunteer as in Fig. 2-6. There is a noticeable decrease in reflectance for the curve representing severe erythema compared to normal skin within the characteristic range for OHb (542 nm, 577 nm) and DHb (555 nm) absorption peaks. [44]

The Dawson erythema index equation (2) was used to calculate the EI for each of the different grade regions contoured by the clinician. The image slices resulting from the spectral bands closest to the wavelengths necessary to compute the Dawson EI (510, 543, 560, 576, 610 nm) were selected for calculation. A single EI for each erythema grade

was computed from the average of all the regions rated with the same grade by the clinician. The relative EI was determined by subtracting the EI for the control region (normal skin) from the EI computed for each erythema grade (Fig. 2-9)

Relative Dawson Erythema Index

CEA-RT Visual assessment grades	G1-Minimal	G2-Mild	G3-Moderate	G4-Severe
Dawson EI	7	7	9.5	8.9
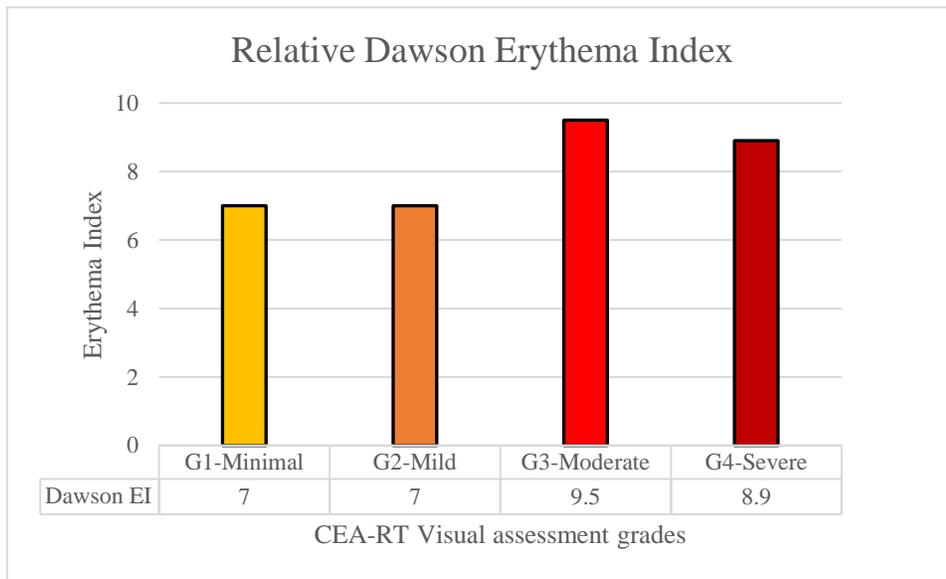
Figure 2-9. The relative Dawson EI was computed by averaging the EI for all the regions with the same erythema grade assigned by visual inspection within the ROI.

Conclusions

The reflectance data calculated from the HSI data showed a decrease in reflectance within the 500-600 nm wavelength range as the erythema grade increases from minimal to severe (G1 to G4). The relative erythema indices calculated using the reflectance spectra for the HSI data showed a subtle increase between mild and moderate erythema grades. These results are consistent with the known optical properties of skin and expected changes in concentration of OHb and DHb as a response to artificially induced erythema.

3. Reliability of the Clinician Erythema Assessment-RT scale

3.1. Study Rational

Skin erythema is a common side effect of radiation therapy. Although there are a number of visual assessment scales validated for radiation dermatitis, the reliability of visual assessment scores for early radiation induced erythema has not been well investigated. This study was designed to determine the reliability of a modified version of the CEA developed for use in the field radiation therapy. The revised assessment referred to as Clinician Erythema Assessment-Radiation Therapy (CEA-RT) is 5-point visual assessment tool intended for grading erythema linked to early radiation dermatitis by radiation clinicians such as radiation therapists and radiation oncologists. The language used to describe the different grades in the CEA-RT scale is more consistent with the language used in the field of radiation oncology (Table 3-1).

Table 3-1. Clinician Erythema Assessment—Radiation Therapy (CEA-RT) scale

Erythema Grade	Description
0 = Normal skin	No erythema or observable difference relative to normal skin.
1 = Minimal erythema	Slight redness but distinguishable, very faint diffuse appearance.
2 = Mild erythema	Definite redness but not very marked, faint erythema.
3 = Moderate erythema	Marked redness, noticeable bright erythema.
4 = Severe erythema	Fiery redness or brilliant red, very bright erythema.

Note: The CEA-RT is a visual assessment scale adapted from the CEA for use in RT to grade radiation induced skin erythema.

The CEA-RT scale range is from 0-4, where 0 is normal skin, 1 is minimal erythema, 2 is mild, 3 is moderate and 4 is severe erythema. In addition, to revising the

grade descriptions, a set of reference photographs representing each erythema grade,

simulating 2 different shades of Caucasian skin, was introduced as a visual aide (Fig.3-1).

Grade 0 Normal Skin	Grade 1 Minimal Erythema	Grade 2 Mild Erythema	Grade 3 Moderate Erythema	Grade 4 Severe Erythema
No erythema or observable difference relative to normal skin.	Slight redness but distinguishable, very faint erythema diffuse appearance.	Definite redness but not very marked, faint erythema.	Marked redness, noticeable bright erythema.	Fiery redness, or brilliant red, very bright erythema.

Figure 3-1. Clinician Erythema Assessment-Radiation Therapy (CEA-RT) scale developed for use in the field radiation therapy. The scale includes a set of reference photographs representing each grade of erythema, simulating 2 different shades of Caucasian skin.

3.2. Method

The purpose of this single centre study was to determine whether the CEA-RT, a

5-point visual assessment scale revised for RT was a reliable method for grading the

severity of radiation-induced erythema reactions. Radiation therapists with a minimum of

4 years of work experience at Juravinski Cancer Centre (JCC) were approached to

participate in this study. Therapists were asked to use the CEA-RT visual grading system

to evaluate the severity of erythema reactions on digital images of the skin of Caucasian

patients treated with radiation therapy. The images used for this study were obtained from consenting patients participating in a related study (HiREB15-017). [47]

The study was divided into 3 parts, one training session and two grading sessions. The purpose of the training session was to familiarize therapists on the use of the CEA-RT for grading erythema on digital images of patient's skin. The training material was introduced to therapists in the form of a slide presentation (Appendix A). The training material included theoretical concepts of colour, lighting and digital image assessment (saturation and value), as well as a discussion time to examine pertinent features to help identify erythema on digital images. During the training session, therapists observed sample photographs of skin exhibiting various CEA-RT grades of erythema. At the end of the training session therapist participated in a collaborative exercise grading images as a group. The exercise was intended to help the group achieve grading conformity.

After completing the training session, therapists were asked to use the CEA-RT scale to individually score digital images of patient's skin during two separate assessment sessions. The two assessment sessions were a repeat of each other, that is; the same images were displayed in the same order for both sessions. To reduce recall bias both grading sessions were at least 3 hours apart. The photographs used for the study were originally taken as a time series to document the progression of erythema during radiation therapy. For the purpose of this study, the photographs were shown to the therapists in a random order to avoid possible association of treatment dose with an expected progression of erythema reaction.

All radiation therapists viewed the same digital images for both sessions. All images were displayed on the same monitors and in the same room to maintain identical viewing conditions. Therapists were provided with the CEA-RT scale and a scoring form to record the grades. Two sets of images were examined at each session. One set had 60 images that included the full treatment region or field of view (FOV) of 6 different patients treated with radiation. Each FOV image was marked with a small 2cm circular area of skin referred to as the region of interest (ROI). The second set had 98 images from 9 different patients that only contained the picture of the 2 cm circular ROI without showing the remaining FOV.

Therapists were asked to only assign an erythema grade to the area of skin within the ROI in both sets of images. The specific area of interest (the 2 cm circular ROI) was marked on each image so that all the therapists evaluated the same patch of skin on each image (Fig.3-2). The irradiated area (treatment field) was outlined on the FOV image set so that the therapists had the opportunity to compare the skin in the untreated area to the skin within the ROI when grading erythema (Fig.3-2).

To evaluate the reliability of the CEA-RT scale for grading radiation-induced erythema, the uniformity of erythema grades within each individual therapist (intra-rater variability) and amongst the therapists (inter-rater variability) were measured. Rater's grade correctness was evaluated by comparing the scores of each therapist to a standard score as well as and the scores for all therapists together to the standard.

The gold standard was selected by convention to be the grades assigned by the instructor who trained the therapists to use the CEA-RT scale in the study. The instructor

was an experienced radiation therapist who had the skills and knowledge necessary to grade radiation-induced erythema on real patients and on photographs of erythema reactions using the CEA-RT scale. Prior to initiation, the study protocol was reviewed and granted approval by the local institutional research ethics board. All therapists provided written informed consent before entering the study.
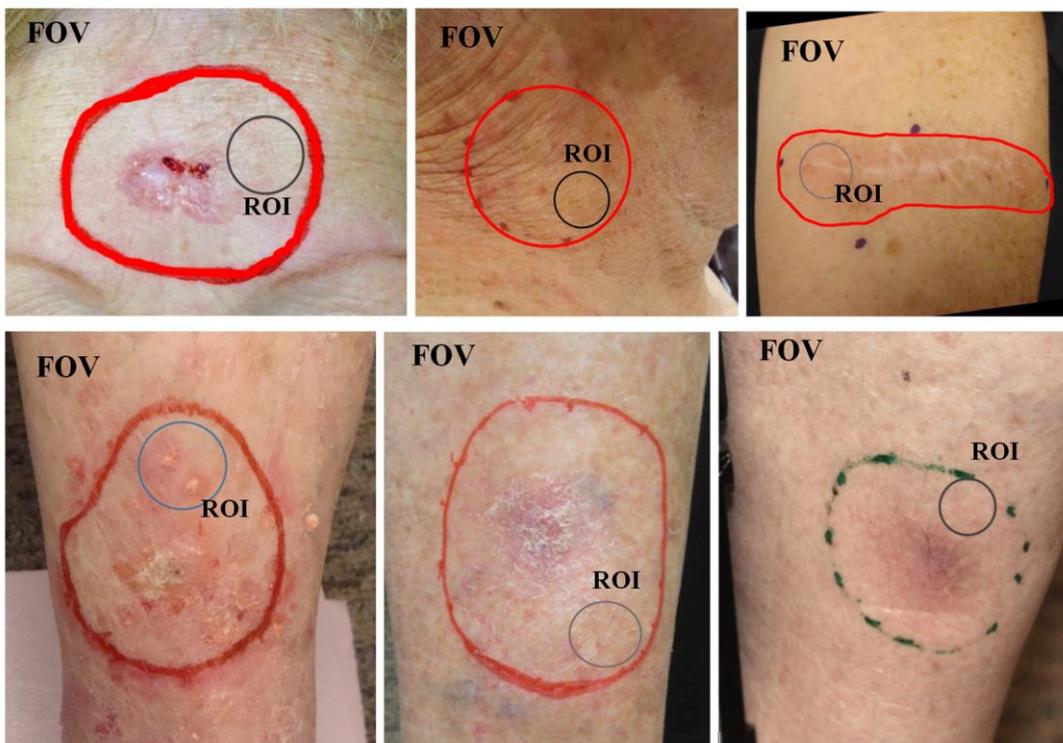
Figure 3-2. *Shown here are 6 sample images, which fully enclosed the treatment region or field of view (FOV). Inside the FOV is a small 2cm wide circular area of skin selected for grading referred to as the region of interest (ROI).*

3.3. Statistical analysis

In this study, we want to assess the consistency and correctness of the therapists' grades when assessing the degree of erythema on two sets of images (FOV and ROI)

using the new CEA-RT scoring system. The scores are nominal and ordinal data; comprised of 5 ranked categories of erythema ranging from 0 to 4 and described as normal, minimal, mild, moderate, or severe. Since individual ratings by appraisers are an example of attribute data, a statistical analysis of for attribute agreement was performed using Minitab®18 Statistical Software. Attribute Agreement Analysis assesses whether appraisers are consistent with themselves (intra-rater variability), with one another (inter-rater variability), and with known standard (degree of correctness). The tests used by Minitab® 18 for attribute agreement were: percent of absolute (100%) agreement assessment, Fleiss' kappa statistical significance test for agreement by grade and Kendall's Correlation Coefficient for assessment of association. [48]

3.3.1. Absolute (100%) Agreement Assessment

Since each therapist provided two ratings for each image, the consistency within each therapist's grades can be evaluated in terms of number times of grades matched perfectly between session. The number of grades matched is the number of times the appraiser agreed with him or herself across the two trials. There is no widely adopted standard for determining "good" agreement but one can be selected based on the domain of the data. For this study, a match of less than 50% was considered a "poor" match, between 50% and 70% was considered a "good" and a match of 70% or higher was considered an "excellent" match.

3.3.2. Fleiss' Kappa (κ) Statistical Significance Test for Agreement

Fleiss' kappa statistics was used to evaluate statistical significance of agreement of scores with $\alpha = 0.05$. The hypothesis for κ to test for agreement were:

H_0: the agreement between ratings is due to chance.

H_1: the agreement between ratings is not due to chance.

The hypothesis was tested for within appraiser, each appraiser vs the standard, between appraisers, and all appraisers vs the standard. The validity of the hypothesis is tested using the p-value. If p-value is ≤ 0.05 then the H_0 is rejected in favour of H_1 concluding that the agreement between the ratings not due to chance. Kappa values range from -1 to $+1$. The higher the value of κ, the stronger the agreement. When:

$\kappa = 1$, perfect agreement exists.

$\kappa = 0$, agreement is the same as would be expected by chance.

$\kappa < 0$, agreement is weaker than expected by chance; this rarely occurs.

There is no single accepted criteria for reliability but Landis and Koch [49] proposed the following weighted κ statistic convention: less than 0, "poor"; 0 to 0.2, "slight"; 0.2 to 0.4, "fair"; 0.4 to 0.6, "moderate"; 0.6 to 0.8, "substantial"; and 0.8 to 1, "almost perfect". The Automotive Industry Action Group (AIAG) (2010) suggests that a kappa value of 0.75 or greater indicates good agreement. [48] For this study, the Landis and Koch criteria was used since it offered a more detailed evaluation of agreement between grades. [49]

3.3.3. Kendall's Coefficients

Ordinal ratings can be examined for degree of association using Kendall's coefficients. Kendell's coefficient of concordance (KCC) is used for assessing intra-rater

consistency between trials and when there is a known standard for each trial, Kendall's correlation coefficient is used. The advantage determining the KCC is that it accounts for the order of the scores, whereas κ statistics do not account for order, all disagreements are considered of equal value. More specifically, Kendall's coefficients considers the consequences of misclassifying an item by more than one degree as more serious than misclassifying the item by only one degree.

Kendall's coefficient values can range from −1 to 1. A positive value indicates positive association, either variables similarly increase or decrease. A negative value indicates negative association, one variable increases while the other variable decreases. There is no single standard for evaluating degree of association but the higher the KCC magnitude, the stronger the association. For this study, a KCC less than 0.70 was considered "low" association, from 0.70 to 0.80 considered "medium", from 0.80 to 0.90 considered "high" and between 0.90 and 1 was considered "very high" association. The hypothesis for Kendall's correlation coefficient test for degree of association with α = 0.05 was as follows:

H_0: there is no association among multiple ratings.
H_1: the ratings within appraiser are associated.

The hypothesis is tested using the p-value for within appraiser, each appraiser vs the standard, between appraisers, and all appraisers vs the standard. If p-value is ≤ 0.05 then the H_0 is rejected in favour of H_1 concluding that the ratings within appraisers (in this case) are associated (i.e. not by chance). [48]

3.3.4. Attribute Data Considerations

When using the Attribute Agreement Analysis test, appraisers should evaluate samples in a random order to ensure that the data collection order does not influence the results. The test should have at least 50 samples for an adequate estimates of agreement, samples should be selected from the entire range of process variation, appraisers should rate each sample at least twice and include at least 3 to 5 appraisers in the study. [48]

3.3.5. Chi-square Goodness-of-Fit test for appraiser group

The Chi-square Goodness-of-Fit test was used to evaluate the similarities between the group of therapists participating in the study and the total population of therapists at the JCC. The groups were compared based on the proportions representing gender and years of experience. The test was run using Minitab$^{®}$18. A p-value of 0.05 was selected as the threshold for rejecting or affirming the null hypothesis (H_0). The H_0 states that these data comes from a specified distribution and therefore there is no statistically significant differences between the observed and expected values for the sample population. A χ^2 less than the tabular χ^2 critical value and a p-value greater than 0.05 is required to reject the H_0 and conclude that the sample is not statistically different to the total population of therapists. A disadvantage of the χ^2 test is that it requires a sufficient sample size in order for the χ^2 approximation to be valid, the expected values should not be less then 5 for reliable results. In the case of small sample sizes, categories can be combined to meet constraints.

3.4.Results

Table 3-2. CEA-RT study acronyms (also referred to as CRIE study)	
FOV	Field of view. Image or set of images that fully shows the treatment area.
ΔFOV	Grade difference for FOV image scores between sessions 1 &2
ΔFOV$_{Ave}$	Average ΔFOV for all images in the set
FOV$_1$	First grading session for the set of FOV images
FOV$_2$	Second grading session for the set of FOV images
ROI	Region of interest. Circular area of skin (2 cm) selected for grading. Set of images that only include the ROI.
ΔROI	Grade difference for ROI image scores between sessions 1 &2
ΔROI$_{Ave}$	Average ΔROI for all images in the set
ROI$_1$	First grading session for the set of ROI images
ROI$_2$	Second grading session for the set of ROI images
YOE	Years of experience

3.4.1. Demographic information for radiation therapists

Twenty therapist with 4 to 35 years of experience (YOE) consented to participate in this study; 40% (8/20) had between 0-10 YOE, 30% (6/20) had 11 to 20 years of experience, 15% (3/20) had 21 to 30 YOE, and 15% (3/20) had 31 to 40 YOE. There were 80% (16/20) female and 20% (4/20) male therapists. Out of 90 therapists working at JCC in August 2018, 76% were female and 24% were male. The years of experience of all therapists working at JCC varied from 1year to 35 years. The demographic information for therapists participating in the study and all therapists working at JCC is listed in Table 3-3. The distribution for YOE and gender is compared graphically in figure 3-3.

Table 3-3. *Demographic information of the sample of therapist that participated in the study and total population of therapists at the JCC.*

Category	Participating therapists	All therapists at JCC
0-10 yrs. exp.	40% (8/20)	46% (41/90)
11-20 yrs. exp.	30% (6/20)	28% (25/90)
21-30 yrs. exp.	15% (3/20)	20% (18/90)
31-40 yrs. exp.	15% (3/20)	7% (6/90)
Males	20% (4/20)	24% (22/90)
Females	80% (16/20)	76% (68/90)
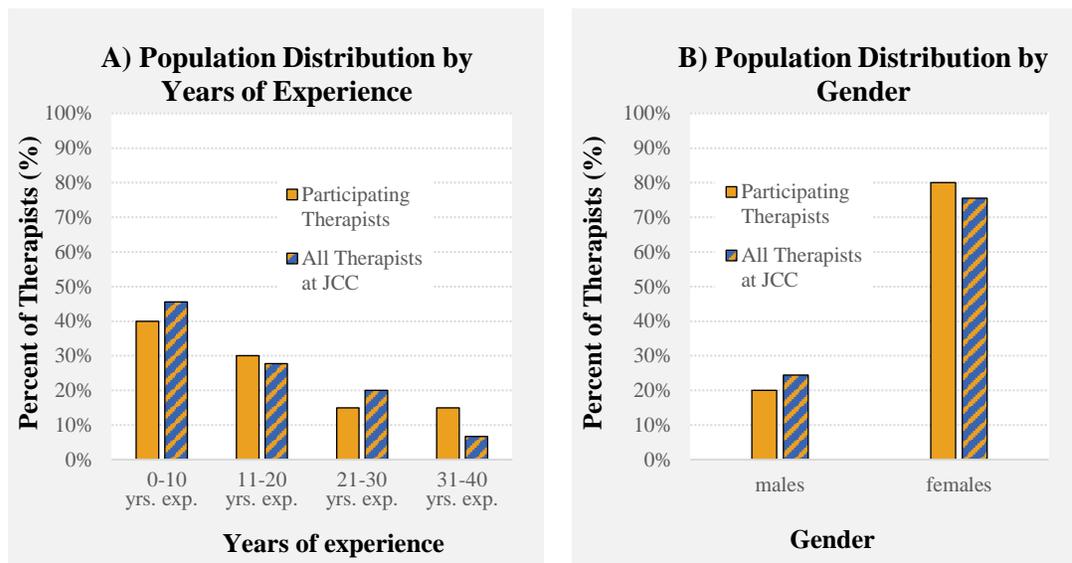
Figure 3-3. *A) Graphical comparison of the years of experience for the sample of participating therapists and the overall population of therapists at the Juravinski Cancer Centre. B) Distribution for both groups of therapists by gender.*

The statistical difference between the sample of therapists participating in this study and the overall population of therapists at JCC were compared using the Chi-Square

Goodness-of-Fit test. This test determines whether the sample of therapists fits the distribution of the total population of therapists and can therefore, be considered a good representation of the total population based in this case, on two nominal categories of gender and YOE.

The observed values (number of participating therapist for each category) versus the expected values (expected or a sample of 20 based on the existing distribution of all therapists at JCC) for YOE and gender was obtained using the χ2 Goodness-of-Fit test run by the statistical analysis software Minitab®18 (Fig. 3-4).
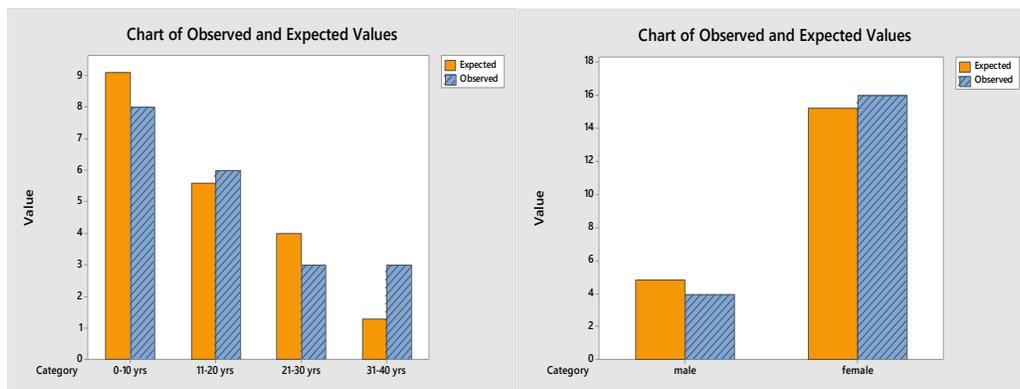
Figure 3-4. *Chart representing the observed values versus expected values for therapists' years of experience and gender (obtained using χ^2 Goodness-of-fit test in Minitab18).*

The test yields a $\chi^2 = 2.63$ and p-value $= 0.451$ when comparing YOE between both groups (Table 3-4). A comparison between the participating therapists and the whole population of therapists at JCC by gender results in a $\chi^2 = 0.175$ and p-value $= 0.675$ (Table 3-4).

Table 3-4. *Chi-Square Goodness-of-Fit test results for years of experience and gender categories of the group of therapist. The level of significance was α = 0.05.*

	N	DF	χ^2 Critical value	χ^2	p-Value
YOE	20	3	7.82	2.635	0.451
Gender	20	1	3.84	0.175	0.675

3.4.2. Demographic information for patients and corresponding images

Therapists were asked to use the CEA-RT Scale to grade erythema in photographs of 9 patients, ages 56 to 88 years old treated at the JCC between 2015 and 2017. Patient P06 was not included in this study because there were no pictures availabe at the time. All patients were Caucasian (4 male and 5 females) and had a diagnosis of early stage squamous cell carcinoma (SCC) or basal cell carcinoma (BCC) of the skin. Patients were treated with a curative course of electrons (6 to 12 MeV) or orthovoltage x-rays (65 to 130 kVp) (Table 3-5) and had not received any radiation therapy to the same area in the past. The treatment dose varied from 3160 cGy to 5000 cGy and was delivered in 10, 15 or 20 fractions. The treated area included various anatomical sites such as extremities, neck, face, forehead or ear. (Table 3-5).

As previously mention in the section describing the study design, two different sets of clinical images were used in this study. One set of 60 photographs fully include the treatment region referred to as the field of view (FOV) and the other set of 98 photographs only contained the small circular area of skin (ROI) without the remaining FOV.

Table 3-5. *Patient Demographics for CRIE Study*

Pt	Gender	Age	Histology	Radiation	Energy	Dose (cGy)	Frac	Site
P01	Male	84	SCC	Electrons	6 MeV	3160*	10	Rt neck
P02	Female	82	SCC	Electrons	6 MeV	5000	20	Lt leg
P03	Female	88	SCC	Electrons	9 MeV	5000	20	Lt leg
P04	Female	56	BCC	X-rays	65 kVp	5000	20	Lt forearm
P05	Male	68	BCC	X-rays	65 kVp	4250	10	Lt cheek
P06	Male	76	BCC	Electrons	6 MeV	4740	15	Lt chest
P07	Male	85	BCC	Electrons	12 MeV	4250	10	Rt ear
P08	Male	75	SCC	X-rays	130 kVp	4250	10	Lt cheek
P09	Female	79	BCC	X-rays	65 kVp	4725	15	Forehead
P10	Female	68	SCC	Electrons	6 MeV	4720	15	Lt leg

*Note: *P01 did not complete total treatment prescription of 4740 in 15 fractions.*

Each FOV grading session resulted in 1200 assessments (60 images x 20 therapists) and for each ROI grading session there were 1960 assessments (98 images x 20 therapists). All therapists graded all images in two different sessions. Altogether, there were 2400 assessments for FOV images and 3920 assessments for ROI images.

The 60 FOV images were selected from six patients out of the nine patients listed on table 3-5. There were 10 images for each patient. Patients #5, #7 and #8 were included in the ROI image-set but not included in the FOV image set because the images were of inferior quality (i.e. too dark, pixelated, blurry, etc.). Image quality was relevant for the FOV image-set since poor quality images would hinder the therapists' ability to interpret the images clinically and erythema features with clarity. The ROI image-set was made up of 10 images for each of the patients except for patient #6 and #1 for which there were 8 and 20 images respectively.
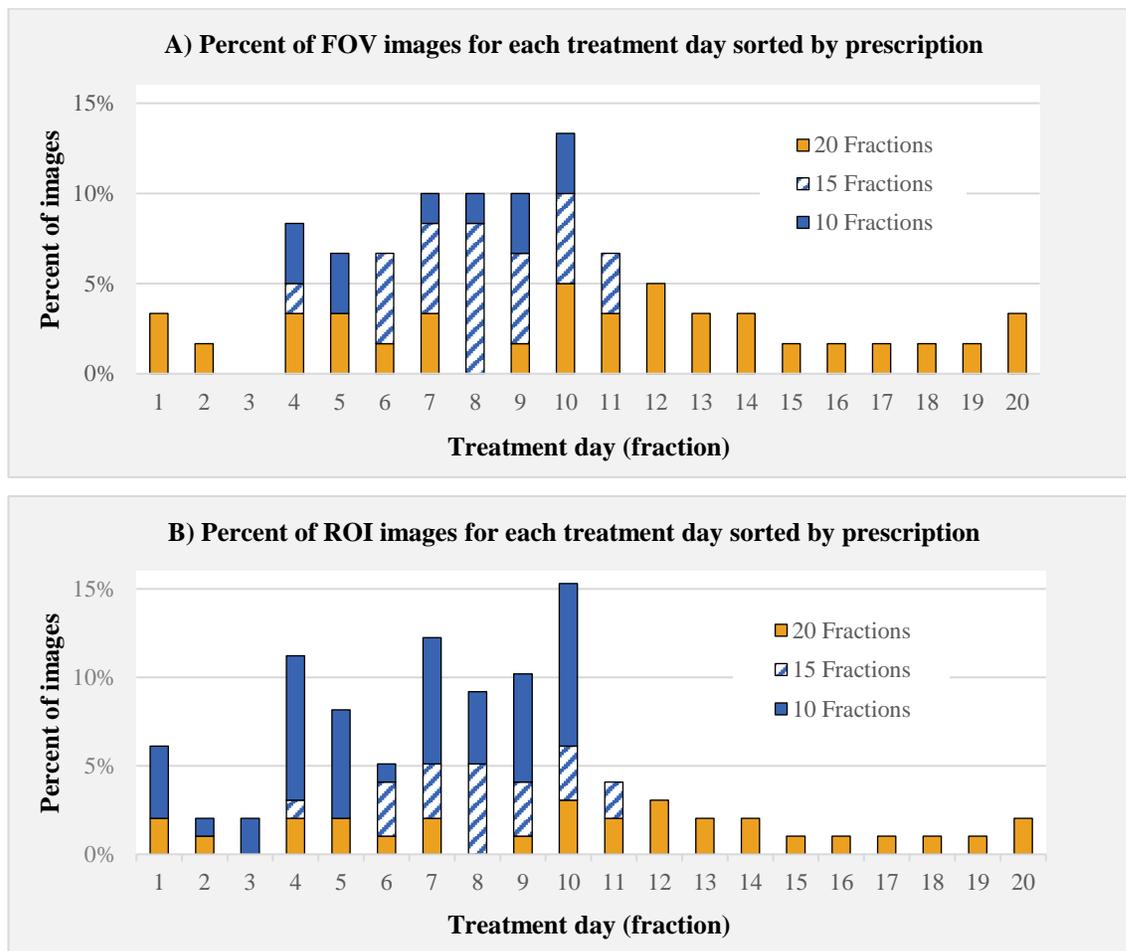
Figure 3-5. Distribution of the percentage of images in the FOV image-sets (A) and ROI image-sets (B) representing each of the three different treament fractionation schema of 10, 15 and 20 days listed in Table 3-5.

The FOV image-set had at least one image representing each treatment day, from the 1st to the 20th day of treatment (fraction) with the exception of the 3rd day for which there was no image available (Fig. 3-5). The ROI image set had a least one image for each of the treatment days without exception. The five most frequent FOV images for a specific treatment day in increasing frequency were for fractions 4, 7, 8, 9, and 10.

Similarly, the five most frequent ROI images for a specific treatment day were for fractions 8, 9, 4, 7 and 10 in increasing frequency (Fig. 3-5).

3.4.3. Standard erythema grades

The distribution of the standard grades assigned by the trainer (gold standard) to the FOV images was such that 32% of the FOV images were grade 0, 42% were grade 1, 25% were grade 2, 2% were grade 3 erythema and 0% were grade 4. For the ROI images, the distribution of the standard grades was 26% for grade 0, 26% for grade 1, 31% for grade 2, 18% for grade 3 and 0% for grade 4 erythema (Fig. 3-6 and Table 3-6).

Figure 3-6. *Percentage of all the standard grades assigned by the gold standard (trainer) to the FOV and ROI images grouped by CEA-RT grade. Higher grades (3 and 4) were were less represented in both image-sets.*

Table 3-6. *Distribution of standard erythema grades assigned by the trainer to the FOV and ROI image-set.*

	Grade 0	Grade 1	Grade 2	Grade 3	Grade 4
FOV	33% (19/60)	42% (25/60)	25% (15/60)	2% (1/60)	0% (0/60)
ROI	26% (25/98)	26% (25/98)	31% (30/98)	18% (18/98)	0% (0/98)

3.4.4. Intra-Rater Variation—Within Appraiser Agreement

Example of within ΔGrade for a Single Therapist

The following is an example of the scores assigned by a single therapist's to consecutive grading sessions for the same set of images (FOV) and of the corresponding grade differences (ΔGrade) for each image. The actual grades assigned to sessions 1 and 2 by therapist #7 to all 69 FOV images is illustrated in Figure 3-7A.

To calculate ΔGrade for each image, the grade for the second session was subtracted from the grade assigned during the first session to the same image. A negative ΔGrade value indicates the score decreased from session 1 to session 2; a positive value indicates the grade increased. A zero value for ΔGrade indicates the therapist had 100% agreement (Table 3.7) between scores for each session. The results of the score variation for therapist #7 for the FOV image-set (ΔFOV) was: 17% (10/60) of the scores decreased by 1 point, 62% (37/60) of the scores matched 100%, and 21% (13/60) increased by 1 point. The difference in scores between sessions for therapist #7 is illustrated in Figure 3-7B for the FOV images (ΔFOV).

Table 3-7. *Change in erythema grades between session 1 and 2 (ΔFOV) assigned by therapist #7 to the FOV image-set.*

	Δ Grade = -1	Δ Grade = 0	Δ Grade = 1
ΔFOV	17% (10/60)	62% (37/60)	21% (13/60)
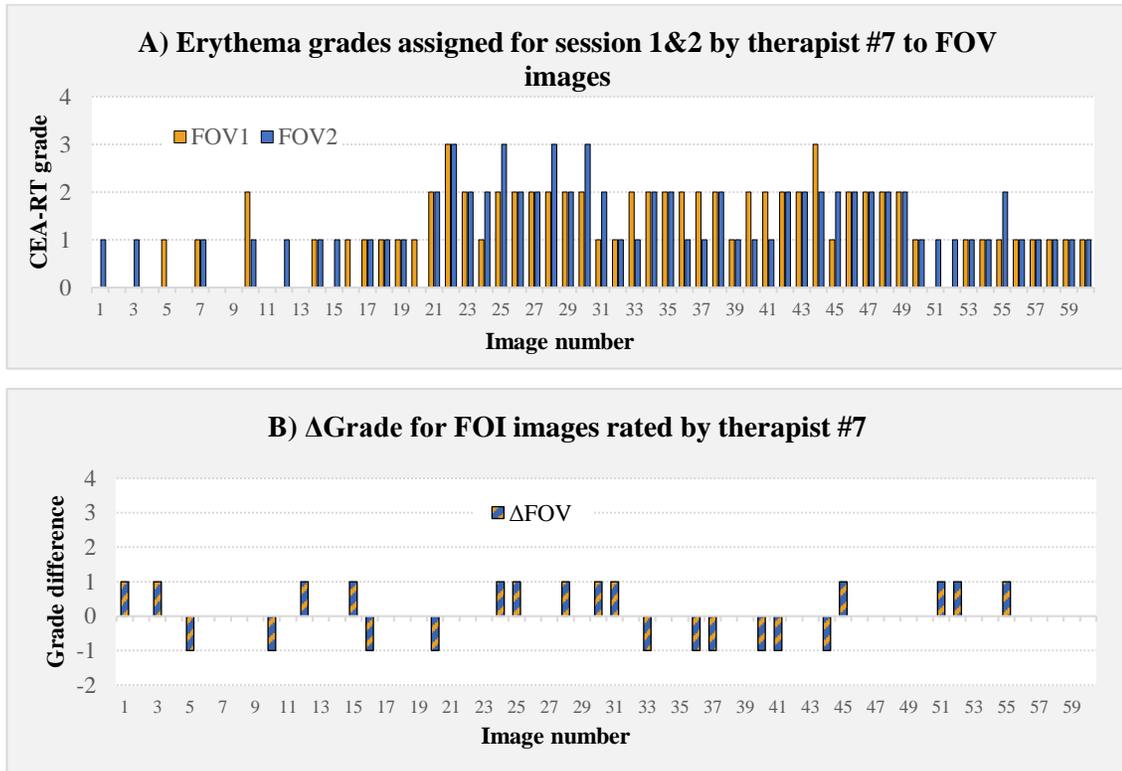
Figure 3-7. *A) Example of the distribution of grades assigned by a therapist (#7) to the FOV image-set for each of the rating session. B) Distribution of the change in erythema grades between rating sessions (ΔFOV) for the same therapist (#7) and FOV image-set.*

Within Appraiser Correlation for YOE and average absolute grade difference

The absolute grade differences between session 1 and 2 for each therapist and image-set (ΔFOV_{Ave} and ΔROI_{Ave}) was calculated as the absolute value of the grade difference between repeat assessments. The average absolute grade difference between trials is a general indicator of how much variation was present among each grades assigned by therapists. The smaller the value of absolute average grade difference the more consistent the therapists were in grading erythema (Table 3-8). A scatter plotted of the average absolute grade difference against the therapists' YOE is seen in Figure 3-8. The resulting scatter plot evaluates the correlation between ΔFOV_{Ave} and ΔROI_{Ave} set

against YOE. The coefficient of determination (R^2 value) between the average grade

difference and the YOE for each set of images were calculated and determined to be

$R^2_{FOV} = 0.12$ and $R^2_{ROI} = 0.041$respectively (Fig 3-8).

Table 3-8. *Distribution of absolute grade differences between session 1 and 2 for each therapist and image-set (ABS ΔFOV$_{Ave}$ and ABS ΔROI$_{Ave}$)*

YOE	4	5	5	6	8	8	9	10	11	14	16	17	18	20	21	26	30	31	32	35
ABS ΔROI$_{Ave}$	0.40	0.45	0.55	0.37	0.22	0.18	0.38	0.24	0.33	0.22	0.33	0.54	0.12	0.69	0.21	0.24	0.39	0.49	0.42	0.54
ABS ΔFOV$_{Ave}$	0.42	0.22	0.25	0.43	0.27	0.30	0.38	0.45	0.60	0.45	0.23	0.42	0.15	0.38	0.37	0.25	0.57	0.43	0.40	0.62
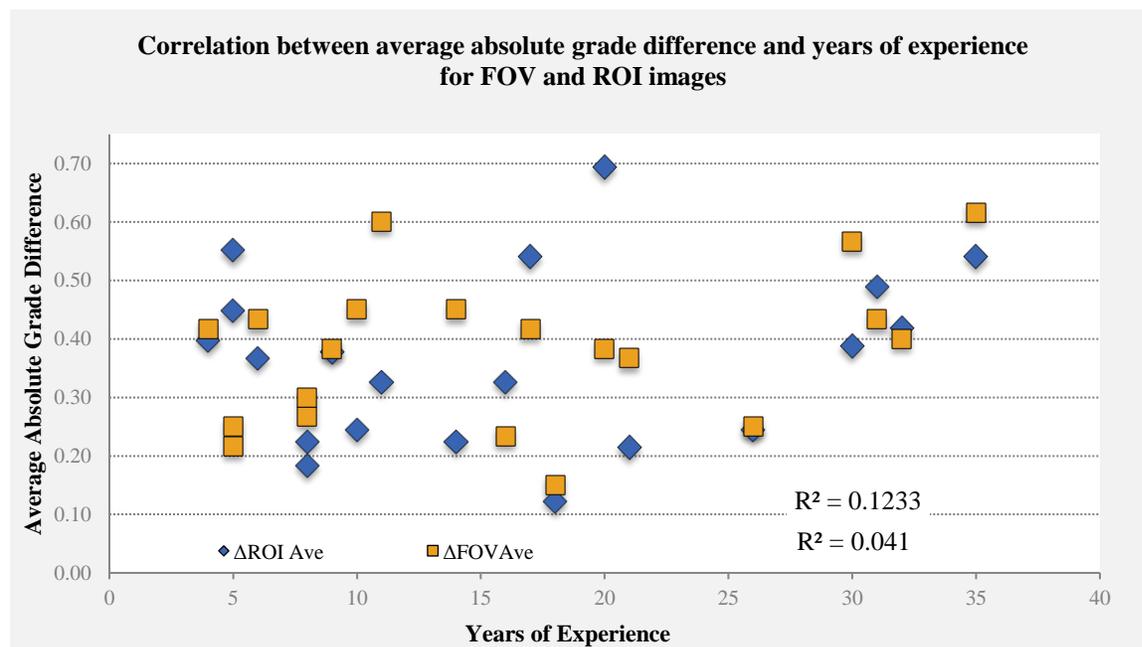
Figure 3-8. *A scatter plot for the average of all absolute grade differences between sessions (ΔFOV$_{Ave}$ and ΔROI$_{Ave}$) for each therapist and for each of the two image sets against years of experience. The value for R^2 values are shown for each image-set. Low R^2 values indicate no correlation between variables.*

Within Appraiser ΔGrade variation

Intra-rater variations were reviewed in terms of the number of points the

therapists' grades differed between sessions for all images (ΔFOV & ΔROI). As mention

previously to obtain the ΔGrade for each image, the grade for the second session was

subtracted from the grade assigned during the first session. A negative ΔGrade value

indicates the score decreased; a positive value indicates the grade increased and ΔGrade =

0 indicates no change. In total, each scoring session involved 1200 FOV image

assessments, as there were 60 images, graded by 20 participants. The grade difference for

FOV images was such that 63% (758/1200) of the images were graded with 100%

concordance between sessions (ΔGrade = 0) and 36% (429/1200) were rated with a grade

difference of ± 1 grade. (Table 3-9).

Table 3-9. *Summary of ΔGrade within appraisers. Percentage of images enclosed in each grade difference (ΔGrade) category.*

	ΔGrade = 1	**ΔGrade = 2**	**ΔGrade = 3**	**ΔGrade = 4**
FOV	16% (194/1200)	1% (8/1200)	0% (0)	0% (0)
ROI	21% (411/1960)	1% (15/1960)	0% (3/1960)	0% (0)
	ΔGrade = -1	**ΔGrade = -2**	**ΔGrade = -3**	**ΔGrade = -4**
FOV	20% (235/1200)	0% (5/1200)	0% (0)	0% (0)
ROI	11% (222/1960)	1% (20/1960)	0% (2/1960)	0% (0)
	ΔGrade = 0	**ΔGrade (– 1 & 1)***	**ΔGrade (– 1 to 1)†**	
FOV	63% (758/1200)	36% (429/1200)	99% (1187/1200)	
ROI	66% (1287/1960)	32% (633/1960)	98% (1920/1960)	

Note: Summary of the grade differences between sessions 1 and 2 for all therapists grouped by the number of points the grades differed between sessions. *This category includes ΔGrade = 1 and ΔGrade = -1. †This category includes ΔGrade = 1, ΔGrade = 0 and ΔGrade = -1.

Similarly, the grade difference (ΔGrade) for ROI images was such that 66% (1287/1960) of the images were graded with 100% concordance and 32% (633/1960) were graded with a difference of ± 1 grade (Table 3-9).

In summary, when the ΔGrade for the range from -1 to +1 was combines for all therapists, it was found that 99% (1187/1200) of the FOV images and 98% (1920/1960) of the ROI images were graded within a plus or minus one-grade range difference between sessions (Table 3-9). Of note is that a ΔGrade of ± 2, ± 3, and ± 4 had a maximum combined incidence of 3% (Table 3-9). Histogram plots for the ΔGrade data for both sets of images are displayed in Figure 3-9. The histogram plot for all images together is a narrow and tall bell-shaped distribution centered on zero. The majority of the values are 0, 1 and -1, beyond these values, frequency drops sharply (Figure 3-9A). To evaluate whether within appraiser grade variation showed a trend with respect to the images' standard grade, the ΔGrade data were grouped by the standard grade. Plot A represents all images by grade difference, plot B to E corresponds to ΔGrade for the group of images with the same standard grade from 0 to 3 respectively. There were no images with a standard grade 4 (§2.4.3). All histograms have a similar narrow and tall bell-shaped distribution centered on zero with the majority of the values between 1 and -1(Figure 2-9B-E). There is a sight higher incidence in the range of 6%–12% of ΔGrade = -1 for FOV images compared to ROI images for all standard grades. Variation for within appraiser grade difference did not show any visual trends with respect to the images' standard grade (Figure 2-9).
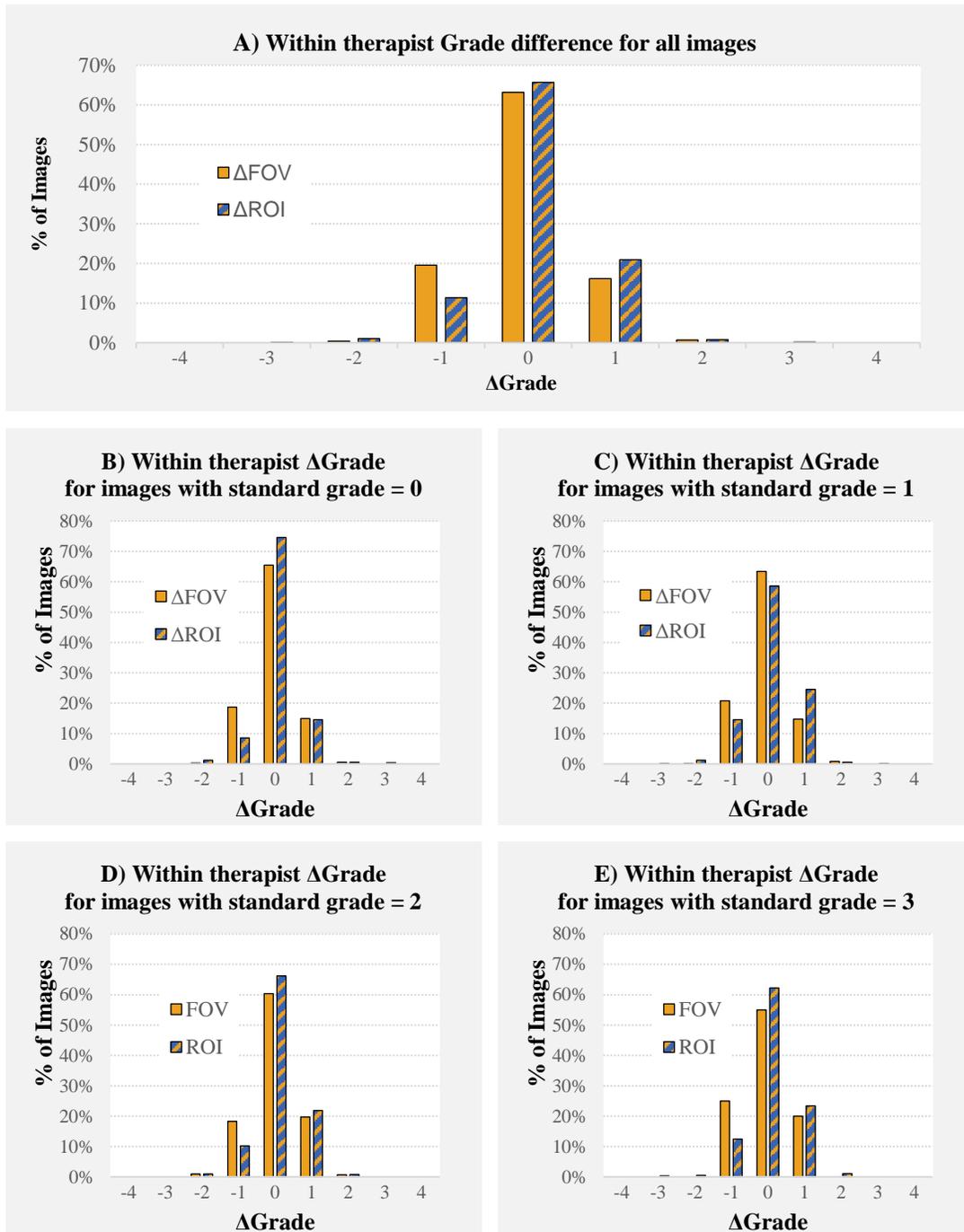
***Figure 3-9.** Histogram plots of the grade difference between the therapists' scores for session 1 and 2 for all images. Plot A represents all images by grade difference, plot B to E corresponds to ΔGrade for the group of images with the same standard grade from 1 to 3 (there was no standard grade 4).*

Within Appraiser 100% Attribute Agreement

Agreement within appraiser was evaluated for each therapist. The range for 100% attribute match between sessions was 45% to 85% for FOV images and 47% to 88% for ROI images (Table 3-10).

Table 3-10. *Assessment Agreement—Within Appraisers for FOV and ROI Images. Data are ordered in decreasing number of grades matched.*

A) FOV Images					B) ROI Images					
Appr aiser	#	Match	%	95% CI	Appr aiser	#	Match	%	95% CI	Agree ment*
20	60	51	85.0	(73.43, 92.90)	20	98	86	87.8	(79.59, 93.51)	E
16	60	47	78.3	(65.80, 87.93)	18	98	80	81.6	(72.53, 88.74)	E
2	60	46	76.7	(63.96, 86.62)	14	98	77	78.6	(69.13, 86.22)	E
3	60	45	75.0	(62.14, 85.28)	6	98	76	77.6	(68.01, 85.36)	E
19	60	45	75.0	(62.14, 85.28)	9	98	76	77.6	(68.01, 85.36)	E
9	60	44	73.3	(60.34, 83.93)	3	98	74	75.5	(65.79, 83.64)	E
18	60	42	70.0	(56.79, 81.15)	8	98	74	75.5	(65.79, 83.64)	E
14	60	38	63.3	(49.90, 75.41)	2	98	66	67.4	(57.13, 76.48)	G
5	60	37	61.7	(48.21, 73.93)	4	98	66	67.4	(57.13, 76.48)	G
7	60	37	61.7	(48.21, 73.93)	1	98	64	65.3	(55.02, 74.64)	G
11	60	37	61.7	(48.21, 73.93)	7	98	62	63.3	(52.93, 72.78)	G
10	60	36	60.0	(46.54, 72.44)	10	98	61	62.2	(51.88, 71.84)	G
15	60	35	58.3	(44.88, 70.93)	11	98	60	61.2	(50.85, 70.90)	G
1	60	34	56.7	(43.24, 69.41)	13	98	60	61.2	(50.85, 70.90)	G
6	60	34	56.7	(43.24, 69.41)	12	98	56	57.1	(46.75, 67.10)	G
12	60	34	56.7	(43.24, 69.41)	16	98	56	57.1	(46.75, 67.10)	G
8	60	33	55.0	(41.61, 67.88)	17	98	53	54.1	(43.71, 64.20)	G
13	60	29	48.3	(35.23, 61.61)	15	98	47	48.0	(37.76, 58.29)	P
4	60	27	45.0	(32.12, 58.39)	19	98	47	48.0	(37.76, 58.29)	P
17	60	27	45.0	(32.12, 58.39)	5	98	46	46.9	(36.78, 57.29)	P

Agreement key; E: excellent (70% ≤ AA), G: good (50% ≤ AA < 70%), P: poor (AA < 50%)

A graph demonstrating the 95% confidence intervals for therapist's score agreement between sessions is shown in figure 3-10. A visual evaluation of within

appraiser consistency indicates the majority of the therapists matched ratings 45% to 75% of the time for both sets of images (Fig.3-10).
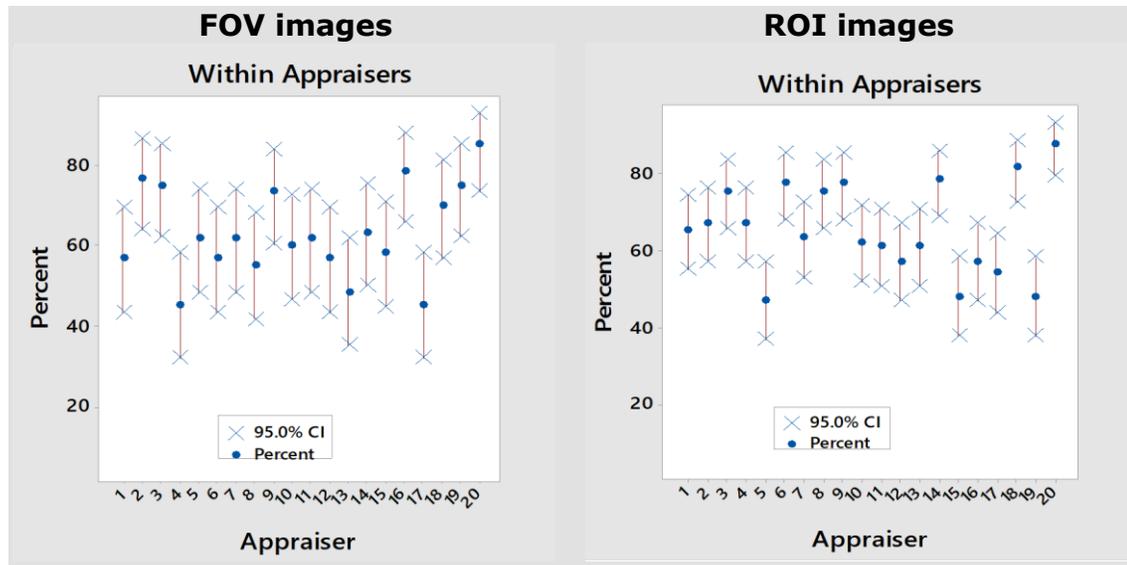
Figure 3-10. *Graph for agreement assessment for intra-rater variation (within appraiser) for ROI and FOV images. Graph obtained using Minitab18.*

The scores for both set of images (FOV and ROI) yielded the same results when tabulated according to the match criteria adopted for 100% agreement assessment (§2.3.1). Thirty-five percent (7/20) of the therapists matched grades for 70% or more of the images (considered excellent match), 50% (10/20) matched at least 50% but less than 70% of the grades (considered a good match) and 15% (3/20) of the therapists matched less than 50% of the grades (poor match). Overall, 85% of the therapists graded images with good or better agreement between trials (Table 3-11A).

Table 3-11. *Summary of Attribute Agreement Analysis–Intra-Rater Variation (within therapists)*

A) 100% Agreement Assessment (AA)					Overall
Poor	*Good*	*Excellent*			*Good or better*
AA < 50%	*50% ≤ AA < 70%*	*70% ≤ AA*			*50% ≤ AA*
FOV 15% (3/20)	50% (10/20)	35% (7/20)			85% (17/20)
ROI 15% (3/20)	50% (10/20)	35% (7/20)			85% (17/20)

B) Fleiss' Kappa test for overall agreement for each therapist (κ)					Moderate or better
Slight	*Fair*	*Moderate*	*Substantial*	*Almost perfect*	
κ < 0.20	*0.20 ≤ κ < 0.40*	*0.40 ≤ κ < 0.60*	*0.60 ≤ κ < 0.80*	*0.80 ≤ κ ≤ 1*	*0.40 ≤ κ*
FOV 0% (0/20)	40% (8/20)	50% (10/20)	10% (2/20)	0% (0/20)	60% (12/20)
ROI 0% (0/20)	15% (3/20)	50% (10/20)	30% (6/20)	5% (1/20)	85% (17/20)

C) Kendall's Correlation Coefficient (KCC)					Medium or better
Low	*Medium*	*High*	*Very High*		
KCC < 0.70	*0.70 ≤ KCC < 0.80*	*0.80 ≤ KCC < 0.90*	*0.90 ≤ KCC ≤ 1*		*0.70 ≤ KCC, p≤0.002*
FOV 0% (0/20)	0% (0/20)	80% (16/20)	20% (4/20)		100% (20/20)
ROI 0% (0/20)	5% (1/20)	10% (2/20)	85% (17/20)		95% (19/20)

Within Appraiser Fleiss' Kappa Statistics

Fleiss' kappa statistics was used to evaluate the statistical significance of the degree of agreement between grades. Although arbitrary in nature, the Landis and Koch criteria was applied selected to weigh the resulting κ values (§2.3.2) [49]. Fleiss' statistics calculates agreement for each response (grade 0 to 4) for individual appraisers and summarizes these results into an overall κ values and a corresponding p-value (Appendix B). The overall range for Fleiss' kappa for individual therapists is listed in Table 3-12 and Table 3-13. The range for kappa was between 0.21 and 0.78 (Table 3-12.) for FOV images (p=0) and between 0.26 and 0.83 for ROI images with a p-value less than 0.01 (Table 3-13).

Table 3-12. *Fleiss' Kappa Statistics—Within Appraisers for FOV Images. Data are ordered in decreasing Kappa coefficient.*

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)	Agreement
20	Overall	0.77	0.088	8.8	0	Substantial
16	Overall	0.63	0.089	7.1	0	Substantial
2	Overall	0.58	0.102	5.6	0	Moderate
19	Overall	0.58	0.099	6.0	0	Moderate
9	Overall	0.58	0.095	6.1	0	Moderate
18	Overall	0.56	0.085	6.6	0	Moderate
3	Overall	0.56	0.096	5.9	0	Moderate
14	Overall	0.47	0.084	5.6	0	Moderate
5	Overall	0.45	0.081	5.6	0	Moderate
7	Overall	0.43	0.086	5.0	0	Moderate
15	Overall	0.42	0.078	5.4	0	Moderate
11	Overall	0.42	0.087	4.8	0	Moderate
1	Overall	0.38	0.080	4.7	0	Fair
10	Overall	0.38	0.092	4.1	0	Fair
6	Overall	0.37	0.084	4.4	0	Fair
12	Overall	0.34	0.089	3.8	0	Fair
8	Overall	0.30	0.088	3.4	0	Fair
17	Overall	0.27	0.071	3.8	0	Fair
13	Overall	0.22	0.088	2.5	0.01	Fair
4	Overall	0.21	0.086	2.5	0.01	Fair

Note: Data were rounded to two significant digits.

Table 3-13. *Fleiss' Kappa Statistics—Within Appraisers for ROI Images. Data are ordered in decreasing Kappa coefficient.*

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)	Agreement
20	Overall	0.83	0.059	14.1	0	Almost perf
18	Overall	0.73	0.069	10.7	0	Substantial
14	Overall	0.71	0.059	12.0	0	Substantial
9	Overall	0.68	0.064	10.6	0	Substantial
8	Overall	0.67	0.059	11.2	0	Substantial
6	Overall	0.66	0.067	9.8	0	Substantial
3	Overall	0.65	0.064	10.0	0	Substantial
4	Overall	0.56	0.058	9.6	0	Moderate
2	Overall	0.53	0.065	8.1	0	Moderate
1	Overall	0.51	0.062	8.2	0	Moderate
10	Overall	0.51	0.055	9.2	0	Moderate
7	Overall	0.48	0.062	7.8	0	Moderate
12	Overall	0.44	0.054	8.1	0	Moderate
11	Overall	0.43	0.066	6.5	0	Moderate
13	Overall	0.43	0.067	6.3	0	Moderate
16	Overall	0.41	0.059	6.9	0	Moderate
17	Overall	0.41	0.052	7.8	0	Moderate
15	Overall	0.32	0.054	5.9	0	Fair
5	Overall	0.27	0.058	4.7	0	Fair
19	Overall	0.26	0.061	4.3	0	Fair

Note: Data were rounded to two significant digits.

When applying the Landis and Koch criteria, the overall Fleiss's κ test showed statistically significant substantial agreement ($0.60 \leqslant \kappa < 0.80$) for 10% (2/20) of the therapists' for FOV scores and 30% (6/20) for the ROI scores. No therapist had almost perfect agreement ($0.80 \leqslant \kappa \leqslant 1$) for FOV images and 5% (1/20) for ROI images. Half of the therapist (10/20) had moderate ($0.40 \leqslant \kappa < 0.60$) agreement with both image-sets (Table 3-11B). There was fair agreement for 40% (8/20) of the FOV images and 15% (3/20) of the ROI images. In summary, 60% (12/20) of the therapists demonstrated moderate or better agreement with themselves for FOV images and 85% (17/20) of the therapists for ROI images.

Within Appraiser Kendall's Correlation Coefficient

Kendall's correlation coefficient (KCC) measures agreement taking into consideration the order of each category (§2.3.3). The KCC for Intra-rater variation ranges from 0.82 to 0.96 for the FOV images with a p-value of 0.0015 or lower (Table 3-14). The range for KCC for the RI images is 0.76 to 0.97 with a p-value of 0.0007 or lower (Table 3-15). Kendall's test demonstrates statistically significant high correlation ($0.80 \leqslant KCC < 0.90$) between scores for 80% (16/20) of therapists for FOV images and 10% (2/20) for ROI images. There was very high ($0.90 \leq KCC \leqslant 1$) correlation for 20% (4/20) of therapist for FOV images and 85% (17/20) for ROI images (Table 3-11C). Overall, 100% of the FOV images and 95% of the ROI images had high or very high statistically significant correlation.

Table 3-14. *Kendall's Coefficient of Concordance—Within Appraisers FOV Images. Data are ordered in decreasing concordance coefficient.*

Appraiser	Coef	Chi - Sq	DF	P	Assessment
20	0.96	113.1	59	0	Very High
16	0.92	109.0	59	0.0001	Very High
18	0.92	108.8	59	0.0001	Very High
5	0.92	108.1	59	0.0001	Very High
15	0.89	105.5	59	0.0002	High
9	0.89	105.3	59	0.0002	High
14	0.88	104.8	59	0.0002	High
1	0.88	104.6	59	0.0002	High
4	0.88	104.3	59	0.0003	High
19	0.88	103.8	59	0.0003	High
6	0.88	103.5	59	0.0003	High
3	0.87	102.5	59	0.0004	High
7	0.86	102.0	59	0.0004	High
8	0.86	101.8	59	0.0005	High
17	0.86	101.7	59	0.0005	High
10	0.86	101.6	59	0.0005	High
11	0.85	100.6	59	0.0006	High
13	0.84	99.4	59	0.0008	High
2	0.84	99.1	59	0.0008	High
12	0.82	96.4	59	0.0015	High

Note: Data were rounded to two significant digits.

Table 3-15. *Kendall's Coefficient of Concordance—Within Appraisers ROI Images. Data are ordered in decreasing concordance coefficient.*

Appraiser	Coef	Chi - Sq	DF	P	Agreement
20	0.97	188.9	97	0	Very High
14	0.96	185.4	97	0	Very High
9	0.95	183.8	97	0	Very High
18	0.95	183.7	97	0	Very High
8	0.95	183.5	97	0	Very High
6	0.95	182.8	97	0	Very High
4	0.94	181.9	97	0	Very High
3	0.93	180.2	97	0	Very High
12	0.93	179.7	97	0	Very High
16	0.93	179.7	97	0	Very High
7	0.92	178.9	97	0	Very High
10	0.92	178.7	97	0	Very High
13	0.92	177.8	97	0	Very High
2	0.92	177.7	97	0	Very High
17	0.91	175.6	97	0	Very High
1	0.90	175.3	97	0	Very High
15	0.90	174.7	97	0	Very High
11	0.86	167.3	97	0	Very High
19	0.86	166.7	97	0	Very High
5	0.77	147.6	97	0.001	Medium

Note: Data were rounded to two significant digits.

3.4.5. Inter-Rater Variation—Between All Appraisers Agreement

Between Appraiser Fleiss' Kappa Statistics

Between all appraisers or also between all therapists, inter-rater agreement for Fleiss' κ statistics test resulted in a statistically significant (p = 0) moderate agreement for grade = 0 for FOV images (κ = 0.44) and ROI images (κ = 0.46). There was slight to fair ($\kappa \approx$ 0.077–0.28 range) for grades 1 to 4 for both FOV and ROI images with the exception of grade = 4 for the ROI images, for which agreement was not statistically significant (p = 0.0949) (Table 3-16). Fleiss' κ statistics test resulted in a statistically significant (p = 0) fair agreement (0.2 to 0.4) for FOV images (κ = 0.27) and ROI images (κ = 0.26).

Table 3-16. *A) Fleiss' Kappa Statistics for inter-rater variability for FOV images*

Response		Kappa	SE Kappa	Z	P(vs > 0)	Agreement
	0	0.44	0.005	95.5	0	Moderate
	1	0.19	0.005	40.7	0	Slight
	2	0.23	0.005	50.1	0	Fair
	3	0.21	0.005	46.1	0	Fair
	4	0.01	0.005	1.3	0.095	Slight
Overall		0.27	0.003	90.1	0	Fair

B) Fleiss' Kappa Statistics for inter-rater variability for ROI images

Response		Kappa	SE Kappa	Z	P(vs > 0)	Agreement
	0	0.46	0.004	128.4	0.00	Moderate
	1	0.22	0.004	61.4	0.00	Fair
	2	0.20	0.004	54.2	0.00	Slight
	3	0.28	0.004	78.6	0.00	Fair
	4	0.08	0.004	21.4	0.00	Slight
Overall		0.27	0.0024	133.8	0.00	Fair

Between Appraiser Kendall's Correlation Coefficient

When considering order between categories, the Kendall's test for assessing inter-rater agreement among all therapists resulted in statistically significant (p=0) borderline low to medium (KCC = 0.69) and medium (KCC = 0.77) correlation (§2.3.3) between all therapists for FOV and ROI images respectively (Table 3-17).

Table 3-17. *Kendall's Correlation Coefficient (KCC) for inter-rater variability*

	Coef	**Chi-Sq**	**DF**	**P-value**	**Agreement**
FOV images	0.690	1629.39	59	0	Low-Med
ROI images	0.774279	3004.2	97	0	Medium

3.4.6. Rater Correctness—Each Appraiser vs the Standard

Appraiser vs the Standard: 100% Attribute Agreement

Agreement between each therapist and the standard was assessed using 100% Agreement Assessment (AA) to determine the correctness of each appraisers' ratings. A graph demonstrating the 95% confidence intervals for therapist's score agreement with the standard is shown in figure 2-11. A visual evaluation of the Appraiser vs Standard graph indicates that the majority of the therapists' ratings were correct 20% to 65% of the time (Fig.2-11).
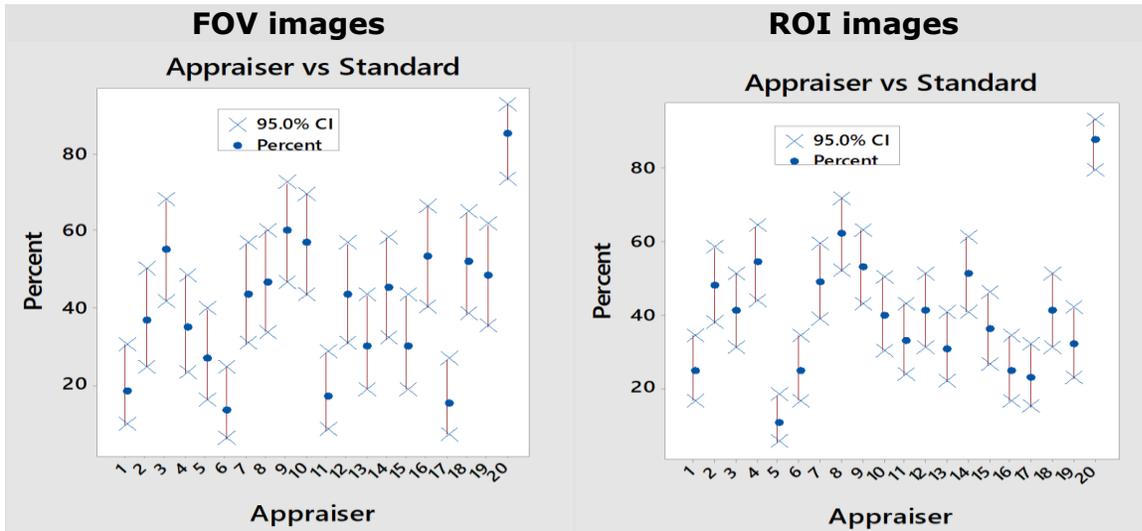
Figure 3-11. *Appraiser correctness. Graph for assessment for agreement between each therapists (appraiser) and the standard for ROI and FOV images. Graph obtained using Minitab18.*

The percent of grades matched between each therapist and the standard ranged from 13% to 85% for FOV images and 10% to 88% for ROI images (Table 3-18.). Attribute agreement assessment demonstrated that for FOV images, 70% (14/20) of the therapists matched less than 50% of the grades (poor match) and 75% (15/20) for ROI images. A good match was seen for 25% (5/20) of the therapists for FOV images and 20% (4/20) for the ROI images. There was an excellent match for one therapist (5%) for both, FOV and ROI images. Overall, 70% to 75% of the therapists had a poor match and 25% to 30% of the therapists had a good to excellent match when assessing 100% agreement with the standard (Table 3-19A).

Table 3-18. *Assessment Agreement—Appraiser vs Standard for FOV and ROI Images. Data are ordered in decreasing number of grades matched.*

A) FOV Images					B) ROI Images					Assessments	
Appraiser	#	Match	%	95% CI	Appraiser	#	Match	%	95% CI	FOV	ROI
20	60	51	85.0	(73.43, 92.90)	20	98	86	87.8	(79.59, 93.51)	E	E
9	60	36	60.0	(46.54, 72.44)	8	98	61	62.2	(51.88, 71.84)	G	G
10	60	34	56.7	(43.24, 69.41)	4	98	53	54.1	(43.71, 64.20)	G	G
3	60	33	55.0	(41.61, 67.88)	9	98	52	53.1	(42.71, 63.22)	G	G
16	60	32	53.3	(40.00, 66.33)	14	98	50	51.0	(40.72, 61.26)	G	G
18	60	31	51.7	(38.39, 64.77)	7	98	48	49.0	(38.74, 59.28)	G	P
19	60	29	48.3	(35.23, 61.61)	2	98	47	48.0	(37.76, 58.29)	P	P
8	60	28	46.7	(33.67, 60.00)	3	98	40	40.8	(30.99, 51.21)	P	P
14	60	27	45.0	(32.12, 58.39)	12	98	40	40.8	(30.99, 51.21)	P	P
7	60	26	43.3	(30.59, 56.76)	18	98	40	40.8	(30.99, 51.21)	P	P
12	60	26	43.3	(30.59, 56.76)	10	98	39	39.8	(30.04, 50.18)	P	P
2	60	22	36.7	(24.59, 50.10)	15	98	35	35.7	(26.29, 46.03)	P	P
4	60	21	35.0	(23.13, 48.40)	11	98	32	32.7	(23.52, 42.87)	P	P
13	60	18	30.0	(18.85, 43.21)	19	98	31	31.6	(22.61, 41.80)	P	P
15	60	18	30.0	(18.85, 43.21)	13	98	30	30.6	(21.70, 40.74)	P	P
5	60	16	26.7	(16.07, 39.66)	1	98	24	24.5	(16.36, 34.21)	P	P
1	60	11	18.3	(9.52, 30.44)	6	98	24	24.5	(16.36, 34.21)	P	P
11	60	10	16.7	(8.29, 28.52)	16	98	24	24.5	(16.36, 34.21)	P	P
17	60	9	15.0	(7.10, 26.57)	17	98	22	22.5	(14.64, 31.99)	P	P
6	60	8	13.3	(5.94, 24.59)	5	98	10	10.2	(5.00, 17.97)	P	P

*Agreement key; E: excellent, G: good, P: poor.

Table 3-19. *Summary of Attribute Agreement Analysis–Each Appraiser vs Standard*

A) 100% Agreement Assessment (AA)					Overall
Poor	*Good*	*Excellent*			*Good or better*
AA < 50%	*50% ≤ AA < 70%*	*70% ≤ AA*			*50% ≤ AA*
FOV 70% (14/20)	25% (5/20)	5% (1/20)	-	-	30% (6/20)
ROI 75% (15/20)	20% (4/20)	5% (1/20)	-	-	25% (5/20)

B) Fleiss' Kappa test for overall agreement (*κ*)					
Slight	*Fair*	*Moderate*	*Substantial*	*Almost perfect*	*Moderate or better*
κ ≤<0.20	*0.20 ≤ κ < 0.40*	*0.40 ≤ κ < 0.60*	*0.60 ≤ κ < 0.80*	*0.80 ≤ κ ≤ 1*	*0.40 ≤ κ, p=0*
FOV 30% (6/20)	15% (3/20)	45% (9/20)	5% (1/20)	5% (1/20)	55% (11/20)
ROI 15% (3/20)	35% (7/20)	40% (8/20)	5% (1/20)	5% (1/20)	50% (10/20)

C) Kendall's Correlation Coefficient (KCC)					
Low	*Medium*	*High*	*Very High*		*Medium or better*
KCC < 0.70	*0.70 ≤ KCC < 0.80*	*0.80 ≤ KCC< 0.90*	*0.90 ≤ KCC ≤ 1*		*0.70 ≤ KCC, p=0*
FOV 40% (8/20)	55% (11/20)	0% (0/20)	5% (1/20)	-	60% (12/20)
ROI 30% (6/20)	50% (10/20)	15% (3/20)	5% (1/20)	-	70% (14/20)

Appraiser vs the Standard: Fleiss' Kappa Statistics

As mentioned earlier , Fleiss' statistics calculates agreement for each response (grade 0 to 4) for individual appraisers and summarizes these results into an overall κ values and a corresponding p-value ([Appendix B](#)) ([§2.4.4](#)). The range for overall Fleiss' kappa was between 0.094 and 0.89 for FOV images (Table 3-20) and between 0.01 and 0.92 for ROI images (Table 3-21) with a p-value equal or less than 0.01. The Fleiss's κ test showed statistically significant substantial agreement ($0.60 \leqslant \kappa < 0.80$) and almost perfect agreement ($0.80 \leqslant \kappa \leqslant 1$) for 5% (1/20) of the therapists in each category for both types of images. There was moderate agreement with the standard for 45% (9/20) of the therapists for FOV images and 40% (8/20) of the therapists for ROI images. Overall, at least 55% (11/20) of the therapists had statistically significant moderate or better agreement with the standard for FOV images and 50% (10/20) for ROI images (Table 3-19B).

Table 3-20. *Fleiss' Kappa Statistics–Each Appraiser vs Standard FOV Images. Data are ordered in decreasing Kappa coefficient.*

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)	Agreement
20	Overall	0.89	0.063	14.1	0	Almost Perf
10	Overall	0.63	0.064	9.8	0	Substantial
9	Overall	0.59	0.065	9.1	0	Moderate
8	Overall	0.50	0.063	8.0	0	Moderate
18	Overall	0.49	0.062	8.0	0	Moderate
3	Overall	0.46	0.065	7.0	0	Moderate
14	Overall	0.45	0.061	7.3	0	Moderate
7	Overall	0.43	0.061	7.0	0	Moderate
12	Overall	0.43	0.062	6.8	0	Moderate
16	Overall	0.43	0.062	6.9	0	Moderate
4	Overall	0.40	0.061	6.6	0	Moderate
19	Overall	0.40	0.064	6.2	0	Fair
13	Overall	0.29	0.063	4.6	0	Fair
15	Overall	0.26	0.058	4.5	0	Fair
2	Overall	0.18	0.065	2.7	0.003	Slight
5	Overall	0.18	0.057	3.1	0.001	Slight
17	Overall	0.09	0.054	1.7	0.04	Slight
1	Overall	0.09	0.057	1.6	0.06*	-
11	Overall	0.05	0.059	0.8	0.22*	-
6	Overall	-0.05	0.057	-0.8	0.80*	-

*Note: *p > 0.05 not statistically significant. Data rounded to two significant digits.*

Table 3-21. *Fleiss' Kappa Statistics–Each Appraiser vs Standard ROI Images. Data are ordered in decreasing Kappa coefficient.*

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)	Agreement
20	Overall	0.92	0.042	22.0	0	Almost Perf
8	Overall	0.66	0.042	15.7	0	Substantial
4	Overall	0.57	0.042	13.7	0	Moderate
7	Overall	0.52	0.043	12.3	0	Moderate
2	Overall	0.51	0.043	11.8	0	Moderate
9	Overall	0.50	0.043	11.7	0	Moderate
14	Overall	0.47	0.041	11.5	0	Moderate
12	Overall	0.45	0.040	11.3	0	Moderate
10	Overall	0.43	0.040	10.7	0	Moderate
15	Overall	0.41	0.040	10.4	0	Moderate
3	Overall	0.35	0.043	8.3	0	Fair
19	Overall	0.35	0.042	8.1	0	Fair
11	Overall	0.30	0.042	7.2	0	Fair
18	Overall	0.29	0.044	6.7	0	Fair
13	Overall	0.28	0.043	6.4	0	Fair
16	Overall	0.26	0.041	6.4	0	Fair
17	Overall	0.20	0.038	5.3	0	Fair
1	Overall	0.13	0.041	3.1	0.001	Slight
6	Overall	0.10	0.042	2.3	0.01	Slight
5	Overall	0.05	0.039	1.2	0.12*	-

*Note: *p > 0.05 not statistically significant. Data rounded to two significant digits.*

Appraiser vs the Standard: Kendall's Correlation Coefficient

        The KCC for each appraiser vs standard grades ranged from 0.53 to 0.94 for

FOV images (Table 3-22) and 0.55 to 0.96 for ROI images with a p-value = 0 (Table 3-

23). Kendall's test demonstrated statistically significant very high correlation ($0.90 \leq$

$KCC \leqslant 1$) with standard grades for 5% (1/20) of the therapists for FOV and ROI. There

was high ($0.80 \leqslant KCC < 0.90$) correlation for 15% (3/20) of therapist for ROI images and

medium correlation ($0.70 \leq KCC < 0.80$) for 55% (11/20) of the therapists for FOV

images and 50% (10/20) for the ROI images. Overall, therapists had at least medium

statistically significant correlation with the standard in 60% (12/20) of the therapists for

the FOV images and 70% (14/20) of the ROI images (Table 3-19C).

Table 3-22. *Kendall's Correlation Coefficient—Each Appraiser vs Standard FOV Images. Data are ordered in decreasing correlation coefficient.*

Appraiser	Coef	SE Coef	Z	P	Assessment
20	0.94	0.063	15.0	0	Very High
10	0.78	0.063	12.5	0	Medium
9	0.76	0.063	12.1	0	Medium
17	0.75	0.063	11.9	0	Medium
6	0.72	0.063	11.5	0	Medium
1	0.72	0.063	11.4	0	Medium
7	0.71	0.063	11.4	0	Medium
16	0.71	0.063	11.4	0	Medium
8	0.70	0.063	11.2	0	Medium
4	0.70	0.063	11.1	0	Medium
14	0.70	0.063	11.1	0	Medium
15	0.70	0.063	11.1	0	Medium
19	0.68	0.063	10.9	0	Low
18	0.66	0.063	10.6	0	Low
12	0.61	0.063	10.4	0	Low
11	0.64	0.063	10.3	0	Low
5	0.63	0.063	10.1	0	Low
3	0.63	0.063	10.0	0	Low
13	0.59	0.063	9.3	0	Low
2	0.53	0.063	8.4	0	Low

Note: Values are rounded to two significant digits.

Table 3-23 *Kendall's Correlation Coefficient—Each Appraiser vs Standard ROI Images. Data are ordered in decreasing correlation coefficient.*

Appraiser	Coef	SE Coef	Z	P	Agreement
20	0.96	0.048	19.8	0	Very High
13	0.82	0.048	16.9	0	High
8	0.82	0.048	16.9	0	High
9	0.81	0.048	16.7	0	High
14	0.79	0.048	16.2	0	Medium
2	0.78	0.048	16.2	0	Medium
12	0.75	0.048	15.5	0	Medium
17	0.75	0.048	15.5	0	Medium
4	0.75	0.048	15.5	0	Medium
7	0.75	0.048	15.4	0	Assessment
10	0.74	0.048	15.3	0	Medium
16	0.74	0.048	15.2	0	Medium
3	0.74	0.048	15.2	0	Medium
6	0.72	0.048	15.0	0	Medium
15	0.69	0.048	14.3	0	Low
1	0.68	0.048	14.1	0	Low
18	0.67	0.048	13.7	0	Low
19	0.65	0.048	13.4	0	Low
11	0.63	0.048	13.0	0	Low
5	0.55	0.048	11.4	0	Low

Note: Values are rounded to two significant digits.

Appraiser vs the Standard: ΔGrade Assessment

A comparison between the therapists' grades and the standard grades determines the correctness of the therapists' responses. The grade difference (ΔGrade) between the standard and the appraiser was calculated by subtracting the standard grade from the therapist's response. A positive ΔGrade value indicates the therapist's response was greater than the standard grade, a zero value indicates the therapist response was correct, and a negative value indicates the therapist's response was a lower grade than the standard.

The majority of the therapists (80%) graded images with good consistency between sessions (Table 3-10). Since therapists were consistent raters, it was reasonable to group together all the grades for both rating sessions. The ΔGrade between the standard and the appraiser had 2400 values for FOV images and 3920 values for ROI images (§2.4.2). A summary of ΔGrade between appraisers and the standard is listed in Table 3-24. The ΔGrade for FOV images was such that 57% (1360/2400) and 55% (2142/3920) of the ROI images were graded with 100% concordance and 40% of the FOV and ROI were graded with a difference of ± 1 grade (Table 3-24). In summary, when combining the ΔGrade for the range from -1 to +1 for all therapists, it was found that 97% (2330/2400) of the FOV images and 95% (2775/3920) of the ROI images were graded within plus or minus one-grade range difference with the standard (Table 3-24). Histogram plots for the data corresponding to ΔGrade between the therapist and the standard data for both sets of images are displayed in Figure 3-12.
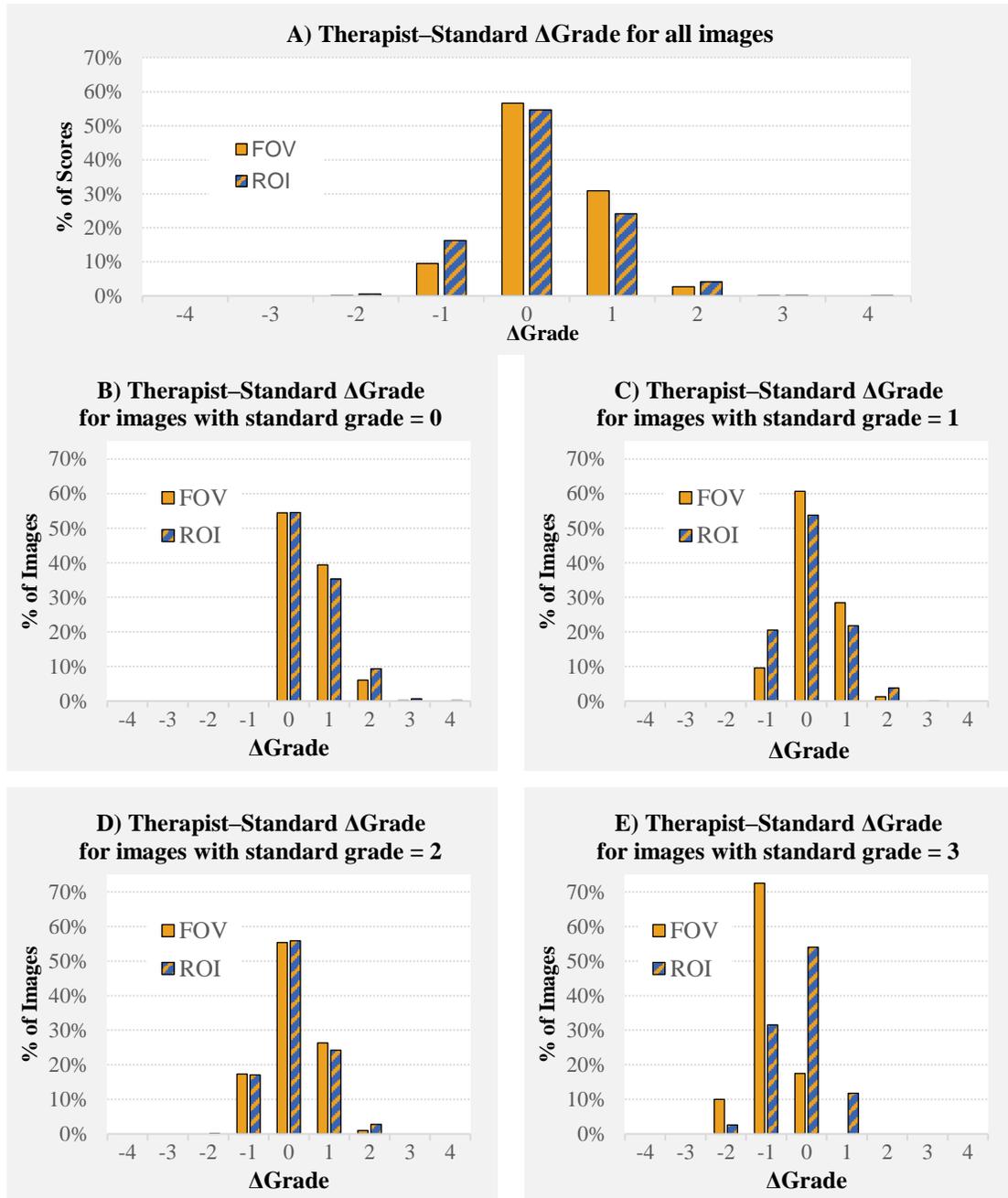
Figure 3-12. *Histogram plots of the grade difference between the therapists' scores and the standard for all images. Plot A represents all images by grade difference (ΔGrade), plot B to E corresponds to ΔGrade for the group of images with the same standard grades from 1 to 3 (there was no standard grade 4).*

Table 3-24. *Summary of ΔGrade between appraisers and the standard. Percentage of images enclosed in each grade difference (ΔGrade) category.*

	ΔGrade = 1	**ΔGrade = 2**	**ΔGrade = 3**	**ΔGrade = 4**
FOV	31% (741/2400)	3% (65/2400)	0% (1/2400)	0% (0)
ROI	24% (947/3920)	4% (164/3920)	0% (8/3920)	0% (2/3920)
	ΔGrade = -1	**ΔGrade = -2**	**ΔGrade = -3**	**ΔGrade = -4**
FOV	10% (229/2400)	0% (4/2400)	0% (0)	0% (0)
ROI	16% (637/3920)	1% (20/3920)	0% (0)	0% (0)
	ΔGrade = 0	**ΔGrade (– 1 & 1)***	**ΔGrade (– 1 to 1)†**	
FOV	57% (1360/2400)	40% (970/2400)	97% (2330/2400)	
ROI	55% (2142/3920)	40% (1584/3920)	95% (2775/3920)	

*Note: The grade differences between appraisers and the standard is grouped by the number of points the grades differed. *This category includes ΔGrade = 1 and ΔGrade = -1. †This category includes ΔGrade = 1, ΔGrade = 0 and ΔGrade = -1.*

The histogram distributions for the ΔGrade for the FOV and ROI images together is centered on zero and declines rapidly bilaterally (Fig. 3-12A). The majority of the values are within 0, 1 and -1; beyond these values, frequency is very low. Only 8% of all images have a ΔGrade of more than one grade (Table 3-24). For images with standard grade = 0 (Fig. 3-12B), there were no negative values since therapist cannot rated an image with a grade less than zero. The histogram distributions for standard grade =1 (Fig. 3-12C), and standard grade =2 (Fig. 3-12D), have a similar shape to the distribution of the ΔGrade for all images together (Fig. 3-12A).

In the plot for standard grade =3 (Fig. 3-12E), the ΔGrade for FOV images show higher values for ΔGrade = -1 as opposed to for ΔGrade = 0. The different distribution for standard grade = 3 may not be statistically relevant since there was only 1 out of 60 FOV images with a standard grade of 3 (Table 3-6).

3.4.7. Rater Correctness—All Appraisers vs the Standard

All Appraisers vs the Standard: Fleiss' Kappa Statistics

Kappa statistics test for all appraisers or all therapist vs the standard resulted in a statistically significant (p = 0) moderate agreement (κ = 0.49 and κ = 0.47) for grade 0 for FOV and ROI images respectively. Therapist's agreement with the standard was also moderate for grade 3 erythema (κ = 0.42) for the ROI images. Grade agreement for grade 3 for the FOV images was not statistically significant (p = 0.36). Agreement for grade 4 scores could not be computed since there were no standard grade 4 scores (§2.4.3). Agreement was fair (κ range ≈ 0.31–0.35) for grade 1 and grade 2 for both sets of images (Table 3-25). The overall, Fleiss κ agreement was fair for all images, since κ = 0.36 and κ = 0.39 for FOV and ROI images respectively with a p-value of zero (Table 3-25).

Table 3-25. *A) Fleiss' Kappa Statistics for all appraisers vs standard for FOV images*

Response		Kappa	SE Kappa	Z	P(vs > 0)	Agreement
	0	0.49	0.020	24.1	0	Moderate
	1	0.31	0.020	15.0	0	Fair
	2	0.34	0.020	16.8	0	Fair
	3	0.01	0.020	0.3	0.36	Slight
	4	*	*	*	*	–
Overall		0.358406	0.014	26.2	0	Fair

B) Fleiss' Kappa Statistics for all appraisers vs standard for ROI images

Response		Kappa	SE Kappa	Z	P(vs > 0)	Agreement
	0	0.47	0.016	29.5	0	Moderate
	1	0.32	0.016	20.0	0	Fair
	2	0.35	0.016	22.0	0	Fair
	3	0.42	0.016	26.5	0	Moderate
	4	*	*	*	*	–
Overall		0.39	0.009	41.7	0	Fair

*Note: * When all sample standards and responses of a trial(s) equal the value or none of them equals the value, kappa cannot be computed.*

All Appraisers vs the Standard: Kendall's Correlation Coefficient

Kendall's test for assessing all therapists' agreement with the standard resulted in statistically significant borderline medium correlation (KCC = 0.69) and medium correlation (KCC = 0.74) for FOV and ROI images respectively (p = 0) (Table 3-25).

Table 3-26. Kendall's Correlation Coefficient (KCC)—All Appraisers vs Standard

	Coef	**SE Coef**	**Z**	**P-value**	**Assessment**
FOV images	0.70	0.014	49.6	0	Low-Med
ROI images	0.74	0.011	68.3	0	Medium

3.5. Discussion

3.5.1. Therapist Population

The group of therapists at the JCC was compared to the group of therapist who consented to the study to determine whether both populations were statistically similar on the basis of their years of experience and gender. The χ^2 Goodness-of-Fit test was used to determine whether the sample population proportions in each category were consistent with the specified values in each category. The χ^2 test recommends that the expected counts (n) for each category should be at least n=5 for reliable results. For this population, the expected values were less than 5 for the following 3 categories: 21-30 yrs. (n=4.0), 31-40 yrs. (n=1.3) and male (n=4.8), (Fig. 3-4 and Appendix B) and therefore, the p-value for the test may not be reliable. If we accept the low numbers, since these categories are known to be less populated in the field of radiation therapy, we find that the calculated χ^2 value was smaller than the tabulated critical χ^2 value and the p-value was much greater

than 0.05 for both YOE and gender. Therefore, it is not possible to reject the H_0 and we can conclude that both populations of therapists are not statistically different (Table 3-4). Hence, the group of therapists participating in this study were a good representation of the total therapist population at JCC.

3.5.2. Patients and Associated Test Images

Two sets of randomly ordered images were assessed by more than 3 raters (20 therapists participated) in 2 separate but identical sessions. The number of images rated was greater than 50 (FOV and ROI had 60 and 98 images respectively). These conditions satisfied data requirement for reliability of results when using the Attribute Agreement assessment tests. Both image sets included a similar number of photographs for each of the treatment fractions representing the initial, the middle and the end of the three radiation prescriptions of 10, 15 and 20 fractions (Fig. 2-5). The three most common treatment prescriptions were relatively equally represented in the FOV and ROI images, as was the treatment modality of orthovoltage and electrons (Table 3-5).

3.5.3. Standard Erythema Grades

The proportion of the standard erythema grades from 0 to 3 was fairly evenly distributed for both, FOV and ROI images with exception of grade 3 erythema. There were only 3% of the FOV images and 18% of the ROI images that had a standard grade of 3. There were no images with a standard grade 4 erythema score (Fig. 3-6 and Table 3-6). Although equal representation in all grades is desirable for establishing the CEA-RT validity and accuracy of the statistical test results, it is highly unlikely that therapists

would have difficulties identifying grade 4 erythema correctly, since it has a characteristic and noticeable very bright or fiery red appearance. Therefore, the lack of grade 4 images should not impact the reliability of the CEA-RT other than there is the possibility that some therapist may have assigned a grade 4 incorrectly as a result of expecting to see at least a few examples included in the image-sets.

3.5.4. Intra-Rater Variation—Within Appraiser Agreement

Correlation between YOE and Average Absolute Grade Difference

The average of all the absolute grade differences between trials for each therapist and image-set (ΔFOV_{Ave} and ΔROI_{Ave}) was plotted against the therapists' years of experience (YOE) (Fig. 3-8) to evaluate any possible correlation between grade difference and YOE. The coefficient of determination was found to be $R^2_{FOV} = 0.12$ and $R^2_{ROI} = 0.041$ (Fig 3-8). For strong correlation R^2 values are close to 1 or 100% [50], in this case the R^2 values were low and close to zero indicating that there was no correlation between the therapists' average grade difference and YOE. Therefore, YOE is not a meaningful factor influencing intra-rater consistency. This may be interpreted as a direct result of training prior to participating in the grading sessions. If all therapists were similarly proficient, it would be expected that the average grade difference would not be associated with the therapists' YOE.

Agreement within raters

Agreement within raters was evaluated by comparing the grade difference (ΔGrade) between trials for all images. Result demonstrates a high intra-rater agreement

72

since therapists graded at least 98% of the images within a one-grade difference between trials (Table 3-9).

Attribute agreement assessment within appraiser yielded similar results for both, the FOV and ROI images. It showed that overall there was high intra-rater agreement since 85% of the therapists had at least a good match for FOV and ROI images (Table 3-11A).

Fleiss' kappa statistics was used to evaluate the statistical significance of agreement between scores using the criteria for reliability by Landis and Koch [49].

The range for Fleiss' kappa was between 0.834 and 0.205 for all images with a p-value less than 0.01. Therefore, the null hypothesis is rejected and the conclusion is that the agreement within appraiser is statistically significant and ranges from fair (greater than 0.20) to almost perfect (equal or greater than 0.80). Overall, at least 60% of the therapists had statistically significant moderate agreement for FOV images and 85% for ROI images (Table 3-11B). More therapists graded ROI images with moderate agreement compared to the FOV images.

Kendall's coefficient of concordance measures the associations between ratings and thus, accounts for the significance of order of the ratings. For within appraiser assessment, the null hypotheses tested is that there is no association among multiple ratings made by an appraiser. The KCC ranged between 0.761 and 0.974 for all therapists' grades and all images with a p-value ≤ 0.002. Since the p-value is less than 0.05, the null hypothesis was rejected concluding that there is a statistically significant association within therapists' responses (Table 3-14 and Table 3-15). Overall, when

considering order between grades, intra-rater agreement was high to very high for 100% of the FOV images and 95% of the ROI images (Table 3-11C).

Intra-rater variation evaluation using Attribute Agreement statistical analysis found to have statistically significant good or better concordance for various tests such as percent agreement analyses as well as Kappa and Kendall's metrics, which means that therapist where able to reproduce their scores consistently

3.5.5. Inter-Rater Variation—Between All Appraisers Agreement

For all grade categories together, Fleiss' κ statistics test resulted in a statistically significant (p = 0) fair agreement for all images (Table 3-16). For grades 1 to 4 kappa values were lower, corresponding to slight to fair agreement between appraisers. Concordance for grade 0 was the highest with moderate agreement for all images. Higher agreement between therapists for grade 0 may be because therapists are trained to assess patients for radiation therapy side effects. During treatments, therapists monitor patients daily for signs and symptoms of erythema. Hence, a higher agreement between therapists in identifying normal skin (Grade 0) is expected since therapists are used to distinguishing erythema from normal skin. On the other hand, assessing the degree of erythema using a predefined scale is not a standardized practice among therapists at JCC; therefore, a lower degree of agreement can be expected when comparing erythema assessments for grades from 1 to 4. Overall, therapists had moderate agreement for grade 0 and weak agreement for all other grades regardless of the image-set.

Kendall's coefficient of concordance was 0.69 for FOV images and 0.77 for ROI images with a p-value = 0. KCC values corresponds to a statistically relevant medium

level of correlation for ROI images and borderline medium to low correlation among therapists scores for the FOV images. Overall, between all appraisers agreement is fair for absolute agreement (Fleiss' Kappa) and improves to medium when order is considered (Kendall's coefficient).

3.5.6. Rater Correctness—Each Appraisers vs Standard

The attribute agreement range for grades matched between each therapist and the standard (correctness of each appraiser's ratings) was broad since it ranged from 10% to 88% for all images (Fig. 3-18). More specifically only 25% to 30% of the therapists achieved a good or excellent match by being correct for 50% or more of the ROI and FOV images respectively (Table 3-19A). A wide match range, from poor to excellent is not surprising when comparing therapists' rates to the standard. Agreement within appraiser was only slightly better compared to between appraiser since it ranged from 45% to 86% (§2.4.5). Therefore, one can expect that raters' agreement with a standard score, which were not assigned by the raters themselves, would likely result in less agreement and more variation.

The total range for Fleiss' kappa was between 0.09 and 0.92 for all images with the majority of the p-values being less than 0.05 (Table 3-20 & 3-21). At least 50% to 55% of the therapists had statistically significant moderate or better agreement with the standard for ROI and FOV images respectively.

Kendall's correlation coefficient for each appraiser vs standard grades ranged from 0.53 to 0.96 with a p-value = 0 (Table 3-22 & 3-23). Overall, Kendall's test demonstrated therapists had at least medium statistically significant correlation with the

standard in 60% of the therapists for the FOV images and 70% for the ROI images (Table 3-19C).

3.5.7. Rater Correctness—All Appraisers vs Standard

Kappa statistics test for all appraisers vs the standard resulted in a statistically significant (p = 0) agreement for all grades except for grade 3 for FOV images and for all grade 4 scores. This is not unexpected since only 2% of the FOV images were grade 3 and there were no images with standard grade 4 (Fig. 3-6). Agreement was moderate for grade 0 for all images and for grade 3 erythema for the ROI images. Agreement was fair for grade 1 and grade 2 for all images (Table 3-25). When considering all grades together, the overall Fleiss' κ agreement was fair for all images. Therapists were more concordant with the standard when rating normal skin (grade 0) and advanced erythema (grade 3) but less concordant when rating early erythema reactions (grade 1&2) (Table 3-25). This may be explained due to the fact that subtle changes in the amount of skin redness may be more challenging to evaluate compared to assessing normal skin or skin with marked redness.

Kendall's test for assessing all therapists' agreement with the standard resulted in borderline-medium to medium correlation (p = 0) for FOV and ROI images respectively (Table 3-26). Kappa and Kendall's test demonstrate there is statistically significant agreement with the standard for all therapists as a group, and the degree of grade concordance is fair to moderate.

3.6. Conclusions for CRIE Study

The reliability of the CEA-RT scale as a visual assessment tool for radiation-induced erythema was evaluated by determining intra-rater and inter-rater agreement.

The grade difference between sessions was within plus or minus one-grade for at least 98% of the images, more specifically 100% agreement assessment demonstrated that 85% of the therapists graded images with a "good" or better agreement between trials. Fleiss' kappa statistics demonstrated that 60% of the therapists had a "moderate" or better agreement with themselves for FOV images and 85% of the therapists for ROI images.

When considering order between categories, Kendall's test showed that there was a high to very high statistically significant correlation in least 95% of the images. These results all show that therapist where able to reproduce their scores consistently between trials.

Inter-rater agreement between all therapists was less consistent than when comparing therapist to themselves. Fleiss' κ statistics test resulted in a statistically significant overall "fair" agreement for all images and the Kendall's test resulted in a "medium" association for all therapists. Therapist were consistent raters but only moderately consistent with compared to each other.

Although the appraisers' ratings were consistent, therapists to therapists comparisons do not indicate whether the ratings were correct. Rater correctness was assessed by comparing therapists' grades to a know standard. Matched attribute agreement showed that 25% to 30% of the therapists had a "good" to "excellent" match with the standard. Kappa statistics determined that at least 50% of the therapists had

statistically significant "moderate" or better agreement with the standard for all images and Kendall's coefficients showed that therapists had at least medium statistically significant correlation with the standard for least 60% of the therapists for all images. Overall, although some therapist had a high degree of accuracy, the majority of the therapists were only moderately correct.

Rater correctness for all therapists together resulted in a statistical significant Kappa test that showed fair agreement for all images and a statistically significant Kendall's test with borderline-medium to medium correlation. Both intra-rater and inter-rater analysis demonstrated that therapists are able to grade erythema consistently, but better with themselves that with each other. In addition, some therapists had a high level of correctness but the majority showed only medium concordance with the standard. Therefore, one can conclude that the therapists can use the CEA-RT grading system reliably but with only moderate accuracy. Since years of experience did not show to be a factor for grade difference, it is likely that therapists' accuracy can be improved with additional training sessions.

Although numerous steps were taken to avoid any bias and errors in this study, this study had some limitations. The gap of time between two assessment sessions varied from 3 hours to one week among therapists. The variation of the time interval between the assessment sessions may have influenced their performance in grading the erythema. The shortest time interval of 3 hours between 2 sessions may have been inadequate to eliminate the recall bias as a result of this therapists may have been able to recall the grades that they assigned to erythema images previously.

In addition, the therapist with 5, 11, 20, 30 and 35 YOE had a larger average absolute grade difference ($\Delta ROI_{Ave} \geq 0.55$ or $\Delta FOV_{Ave} \geq 0.55$) compared to the rest of the therapists (Table 3-8). Within their data, a few grade differences were more than 2 or 3 scores apart between the two sessions and the rest of their scorings were within the range of ± 1 difference. These unexpected large grade difference in just a few of their assessments may have been caused by human error due to fatigue as they might have lost track while grading 158 (60+98) slides of erythema images during each session. Another limitation was that the study was conducted using digital photographs. Photographs are affected by light reflection, hue intensity, brightness and saturation that individual raters may interpreted differently. Although great effort was made to standardize viewing conditions, grading real patients may be more familiar to therapists and result in less variation stemming from image feature interpretation and viewing fatigue.

Finally, the results of this study may not be applicable to other studies. In other studies, investigators validate assessment tools in vivo, using real patients whereas this study tests the validity of the CEA-RT tool based on photographs. In addition, only photographs of Caucasian patients were available for this study, which limits the application of the results of this study to other skin types. Future research might focus on creating additional reference photographs for different skin types to accompany the CEA-RT scale and consider additional training and involving in vivo patients instead of digital photographs.

4. Experimental application of DRS and CEA-RT for grading in-vivo radiation induced erythema

Acronyms & Abbreviations

Table 4-1. *List of Acronyms and Abbreviations*

DTE	Day-to-erythema: the day erythema was first detected
DTE_{EI}	Day-to-erythema: the day erythema was first detected by EI
DTE_{VA}	Day-to-erythema: the day erythema was first detected by VA
EI	Erythema index calculated using Dawson's equation
VA	Visual assessment of erythema performed by a clinician
TRT ROI	Treatment region of interest
COA	Control A–skin next to the TRT ROI
COB	Control B–skin of the inner forearm

4.1. Study Rational

In this chapter, the results of an in-vivo radiation induced erythema study are presented with the purpose of demonstrating a clinical application for the DRS system described in chapter 2.

Radiation toxicities such as acute soft tissue inflammation is observed in up to 95% of patients. Of these patients, 87% will experience moderate to severe skin reactions including erythema, dry and moist desquamation during or towards the end of the treatment resulting in significant discomfort. [22, 23]

Most grading systems used for determining radiation-induced skin reactions are based on visual assessment (VA) by a trained clinician. Although often criticized for its subjective nature, VA is the standard for assessing and grading the severity of erythema

and other adverse events. [29-32] The absence of a universal visual assessment scale makes it difficult to compare results from different studies and hinders the progress of evidence-based knowledge.

In radiation therapy, erythema is a visual manifestation of epithelial tissue damage and activation of the inflammatory response resulting in vasodilation and changes in the blood concentration within the dermis. DRS is a suitable technique for detecting these physiological changes in skin and therefore, DRS has the potential to provide a quantitative measurement of radiation induced erythema. [40]

This study compared two approaches for evaluating erythema, DRS measurements described in chapter 2 and visual inspection using the Clinician Erythema Assessment for Radiation Therapy (CEA-RT) validated in chapter 3.

Erythema resulting from skin toxicity induced by radiation therapy was measured using DRS and by visual inspection using the CEA-RT grading tool. The primary objectives of this study was to compare the sensitivity of DRS and VA in detecting erythema, and to determine whether DRS was statistically correlated with the known standard for erythema assessments (VA).

The potential implication of this research is the addition of a practical instrument for objectively monitoring erythema changes to complement a clinician's assessment. Although this study may not result in direct changes to clinical practice, it may provide additional evidence to support using DRS as a valuable grading tool for investigators who do not have clinical expertise in assessing erythema. If DRS is determined to be a reliable evaluation device for erythema, its objective quality will facilitate comparing results

across controlled experiments and clinical trials. This would in turn help increase the

body of evidence for effective therapeutic interventions and optimal management of

radiation induced skin toxicities.

4.2. Study design

The study design was a prospective single institution observational study

comparing DRS measurements to visual grading of radiation induced skin erythema. In

the process of conducting this research no experimental interventions or therapeutic

changes were made to the participants' treatments. All patients were treated at JCC during

May 2015 and June 2016. All patients received the standard of care treatment as

prescribed by the radiation oncologist in accordance with current clinical best practices.

The study received formal approval from the local ethics board (HiREB application 15-

017). All participants in this study gave informed consent in accordance with the HiREB

guidelines. [47] For inclusion in this study, patients were required to meet the following

criteria:

- Be at least 18 years old
- Be a patient receiving radical radiation therapy for early stage skin cancer at the
  JCC
- Have not had any previous radiation treatment resulting in an overlap of the
  original field with the current study field.
- Be able to understand and comply with the protocol
- Agree to attend daily study sessions
- Be able to give informed consent.

Only patients with early stage skin cancer who were offered radiation therapy were included in this study. The reason being that radiation for early skin cancers (i.e. limited size and depth of invasion with no nodal involvement) are treated with single beam electrons or ortho-voltage (low energy photons). Single beam irradiation results in higher uniform doses to the skin compared to radiation treatments delivered by more sophisticated multi-beam techniques such as IMRT (intensity modulated radiation therapy) and VMAT (volumetric modulated arc therapy).

The more complex IMRT and VMAT approaches have a skin dose spearing effect. Therefore, these techniques are better suited for treating advanced tumours that are more extensive or deeper within the body where dose to skin and other critical structures are essential treatment limitations.

Higher uniform doses to the skin are well suited for VA and obtaining DRS measurements since they elicit the prompt development of skin toxicities. This facilitates the study by increasing the likelihood of visible skin changes and variations in the optical properties of the skin.

Furthermore, radiation therapy prescriptions for early skin cancers are more condensed resulting in a higher doses per fraction, hence a higher chance of developing erythema changes early on in the course of treatment, which assists data collection.

4.2.1. Study Population

Ten patients were included in this study, 5 males and 5 females. The ages ranged from 56 to 88 years old (72yrs ± 16yrs). Four patients were treated with ortho-voltage radiation and 6 were treated with electrons. The anatomical location of the skin cancer

and the total prescription dose varied between patients. See table 4-2 for a detailed description of the patients' demographic information.

Table 4-2. Patient Demographics

Pt	Gender	Age	Histology	Radiation	Energy	Dose (cGy)	Frac	Site
P01	Male	84	SCC	Electrons	6 MeV	3160*	10	Rt neck
P02	Female	82	SCC	Electrons	6 MeV	5000	20	Lt leg
P03	Female	88	SCC	Electrons	9 MeV	5000	20	Lt leg
P04	Female	56	BCC	X-rays	65 kVp	5000	20	Lt forearm
P05	Male	68	BCC	X-rays	65 kVp	4250	10	Lt cheek
P06	Male	76	BCC	Electrons	6 MeV	4740	15	Lt chest
P07	Male	85	BCC	Electrons	12 MeV	4250	10	Rt ear
P08	Male	75	SCC	X-rays	130 kVp	4250	10	Lt cheek
P09	Female	79	BCC	X-rays	65 kVp	4725	15	Forehead
P10	Female	68	SCC	Electrons	6 MeV	4720	15	Lt leg

*P01 did not complete the full radiation prescription of 4740 cGy in 15 fractions.

4.2.2. Study protocol and visual assessments

Patients participating in this study were asked to attend a daily 15-20 minute assessment session before each treatment. During these sessions, an experienced therapist trained to use the CEA-RT 5-point grading scale inspected the patients' skin within the treatment field and rated erythema. The same therapist rated all cases.

Daily digital RGB (red-green-blue) images of patients' skin within the treatment area were obtained to record the appearance of the treatment field.

During the daily visual assessment, the therapist reviewed with the patient any new or recurring side effects such as itching, burning, dryness, scabs, blisters or accidental trauma (i.e. scratches or bruises).

In agreement with the treating radiation oncologist, patients were instructed not to expose the treatment area to the sun or use topical creams not indicated by the doctor. Sun exposure and topical creams have the potential of intensifying skin reactions and consequently influencing erythema reactions, VA scores and DRS measurements.

4.2.3. Selection of Region of Interest and Data Acquisition

Two to three areas of skin referred to as regions of interest (ROIs), were selected on each volunteer. At least one ROI was chosen within the treatment field to monitor radiation- induced erythema, and one or two ROIs were chosen outside the treatment field to monitor day-to-day variations of normal skin. The ROI selected within the treatment field (TRT ROI) was always located least 3mm from the margin of the treatment and would exclude the cancer lesion as much as possible (Fig.4-1A). Choosing the TRT in this way reduced DRS data contamination from the pigment of the coloured ink markers used to outline the field borders. A set distance from the field border also ensures a homogeneous skin dose distribution since the radiation dose drops quickly close to the field edge (known as penumbra).

Excluding the cancer lesion from the TRT ROI minimized the influence of changes related to the therapeutic response of the cancer lesion to radiation such as tissue discolouration, tissue breakdown, bleeding and crusting. These tissue changes would mask the erythema features relevant for VA and DRS measurements (Fig.4-1A).

Figure 4-1A&B
Image of the ROIs selected on volunteer P03. The ROI were circular to simulate the sampling port of the single integrating sphere. Image A) TRT and COA ROIs. The treatment ROI (TRT) was located within the treatment field at least 3mm away from the field margin and the cancer lesion. The control ROI labeled COA was located 8 cm directly above the TRT site. Image B) COB ROI. Image of the inner aspect of the forearm for the same participant showing the location of the control ROI (COB). Note: The diagrams are not to scale.

A) TRT and COA ROIs
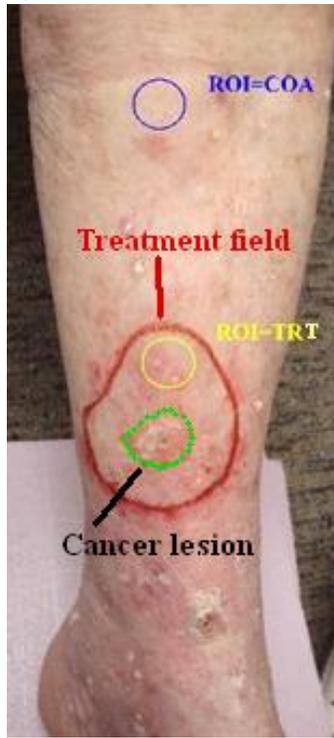
B) COB ROI
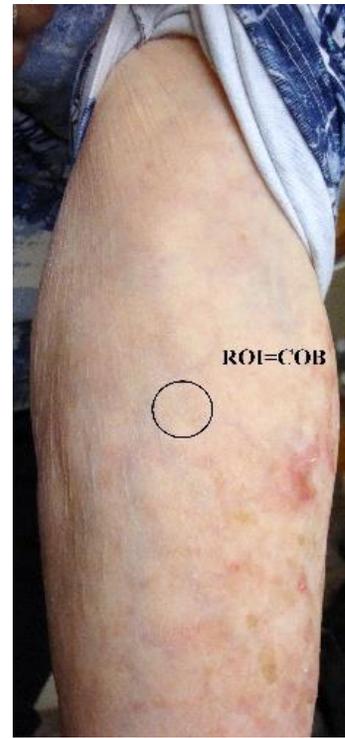
The control ROIs were selected outside the radiation field to avoid including any skin at risk of developing radiation induced erythema. The controls were used to establish a DRS baseline for normal skin and to calculate EI thresholds to evaluate time to development of radiation induced erythema.

When possible two control sites were chosen on normal tissue. One control ROI labeled COA was selected close to the treatment area but sufficiently far away so radiation exposure was negligible (i.e. 5-10 cm away from the outside margin of the treatment field).

A second control ROI site labeled COB was selected from the skin in the inner aspect of the arm 5cm above or below the elbow, whichever position was more comfortable for the volunteer. COA was chosen closer to the treatment site so that its

inherent skin properties would better resemble the skin within the TRT ROI. COB the was selected on the skin of the inner arm because the skin in this area tends to have fewer blemishes from sun exposure and is more protected from injury, therefore there is a greater chance of obtaining more consistent baseline measurements (Fig.4-1B).

To overcome the challenges of reproducing daily measurements, on the first day of treatment a transparent template was made for all the ROIs in each patient. The purpose for the template was to serve as a means of mapping the position of ROIs on the patient skin relative to surrounding noticeable skin features. This template was used every day to accurately reproduce the anatomy within the ROI and the position of the integrating sphere's port over the ROI for measurements. The templates were also useful for ensuring the ink marks on the patients' skin were correctly positioned before the daily RGB images were obtained.

All DRS measurements were collected with the room lights off using only the dim light from laptop's screen. All DRS readings were repeated 3 times for each ROI. The probe was physically removed and repositioned on the ROI before repeating each the measurement. Repositioning the probe before each measurement was adopted to help account for the random sampling error associated with the DRS measurements.

4.3. Data Analysis and Statistical Tests

4.3.1. Data Fine-Tuning

The spectroscope processing software generated data files containing the acquisition parameter information and a table listing the wavelength (nm) and the corresponding reflectance figures (counts). At times inadvertently the experimental data

were acquired using the system's default parameters (boxcar = 0 and averaged scans = 0) instead of the recommended parameters (boxcar = 5 and averaged scans = 10) which optimized the data quality. To compensate for such discrepancies a moving average of 10 points was applied to the raw reflectance figures. The moving average was a good approximation of the protocol parameters; it reduced the noise and smoothed the data.

The sensitivity range of the system's spectrometer was wider than the spectral range of the light source use (Table 2-2). This resulted in a very low S/N ratio affecting the reflectance data above 620nm and below 500nm. Therefore, these data were discarded and only the reflectance between 500nm to 600nm was kept.

The wavelength range of the remaining reflectance data were suitable for calculating the EI using the Dawson EI equation (2) discussed in section 2.1.3.

Ten patients were involved in this study and resulted in 10 data sets of 10 to 25 daily consecutive DRS measurements that did not include weekends or statutory holidays.

The EI was calculated for each of the 3 repeat measurements and then averaged to determine the daily EI for each ROI.

To test the effect of how the EI value depended on the wavelength bandwidth (0.206 nm) given by the spectrometer, the EI was calculated for 4 patients using the logarithmic inverse of reflectance for 3 ranges. EI was calculated for a single (1) bandwidth, for 11 bandwidths (1–5 to 1+5 range) and for 21 bandwidths (1–10 to 1+10 range). The difference between the EI calculated from LIR averaged over 1, 11 or 21 bandwidths values was less than 0.35% and ranged from 0.08% to 0.35% (Table 4-3). Since the percent difference between EI calculated for 1 bandwidth and the EI calculated

for a range of bandwidths was very small, to simplify calculations, an average of one

wavelength band value was used for computing the EI. The data points with the closest

wavelength values to the wavelengths proposed by Dawson's equation were selected to

calculate the erythema indices. The wavelengths used were 510.088 nm, 543.143 nm,

560.058nm, 576.062 nm and 610.038 nm.

Table 4-3. Percent difference between the EI calculated for 1 bandwidth and the EI calculated for a range of bandwidths (11 & 21) for 4 patients.

Patient	P01	P02	P03	P04
ΔEI(1&11)%	0.19%	0.21%	0.08%	0.08%
ΔEI(1&21)%	0.31%	0.35%	0.19%	0.19%

4.3.2. Selection of Confidence Intervals for EI values

EI is an indirect measure of the Hb present in skin (discussed in chapter 2),

although normal skin is expected to vary its Hb concentration over time; it is also

expected to fluctuate within a normal physiological range. The natural range variation of

EI values of normal skin was considered the EI baseline for normal skin.

The control ROIs (COA and COB) were purposely selected on normal skin

outside the area of skin exposed to radiation. The COA and COB measurements served to

monitor normal daily changes in the skin spectra, therefore the corresponding variations

in the EI were used to determine normal random variations in the DRS data not associated

with erythema. Since the daily EIs for the control sites were expected to fluctuate around

an average baseline EI in normal conditions, the set of EI values could be represented by

a two-tailed distribution. The standard deviation (σ) for the EIs for each of the control

ROI's may be calculated using z = 1.96 which is the critical number for a 95% CI for a two-sided test.

As opposed to normal skin, skin exposed to radiation will react progressively as radiation damage to the skin develops. Inflammation leading to vasodilation, increased blood circulation and interstitial Hb concentration are associated with erythema resulting from radiation induced damage to the skin. The severity of erythema reactions are expected to increase as the accumulative irradiation dose increases. Therefore, the EI values in the treated skin are also expected to increase over time during radiation therapy treatments. Since the EI for the TRT ROIs is expected to increase in value with repeated radiation exposure, these data best resembles a single sided distribution. Therefore, a one-tailed test using the critical number z = 1.65 is appropriate to calculate a 95% CI.

The errors associated with the TRT EI values were calculated using the standard deviation for the 3 repeated daily DRS measurements obtained from the TRT ROI.

The figures for the EI for repeat measurements were assumed to be normally distributed and therefore the error for TRT EI was defined as 1.96σ. The most representative σ value for the set of daily TRT EI figures was selected to calculate a single error for the set of EI values for each case (Table 4-5).

4.3.3. Baseline Erythema and EI Threshold–Time Sequence Plots

To determine the point at which TRT EI change was more likely attributed to radiation-induced erythema and not to normal random daily variations within the skin, an EI threshold needed to be established. Since the TRT ROI and the control ROI were assumed to be of similar skin type and biological optical properties, the mean EI

calculated for the control (μCOA or μCOB) was presumed to be the baseline EI for TRT

ROI. Therefore, at the point where the EI for the TRT ROI surpasses this baseline EI in a

consistent manner, subsequent changes in EI for TRT ROI beyond this point may be

attributed to radiation-induced erythema. The control was chosen as a surrogate for

calculating the EI baseline due to the limitations on obtaining repeat measurements for

the TRT ROI before the start of irradiation.

The EI threshold for the TRT ROIs was defined as equal to the mean (μ) EI for

the control plus 1.65σ, where 1.65 is the critical number for a one sided test with 95%

confidence interval and σ is the standard deviations for the EI values of the control. More

specifically, the EI threshold was calculated as μCOA+1.65σ or μCOB+1.65σ depending

on the most appropriate control ROI.

Time sequence plots for EI and VA against treatment day were used to find

trends and patterns in the data. The trends is a general direction of the data over a period

of time (i.e. increasing, decreasing, or lateral move), and the pattern is a recognizable

repeated form followed by the data (i.e. wave pattern). [51]

4.3.4. Day-to-Erythema (DTE)–Scatterplots

Another aspect of evaluating the sensitivity of the EI and VA as different

approaches for assessing erythema was to establish a common end-point. Day-to-

erythema (DTE) was chosen as a convenient arbitrarily end objective. DTE is represents

the first day erythema was detected using EI measurements or VA grades. The criteria to

determine DTE using EI was defined as the first day that the EI value rose above the EI

threshold for least two treatment days in a row. For VA, the criteria to determine DTE

was defined as the first treatment day after the start of radiation, that the VA grade increased by one grade for at least two consecutive treatment days.

Radiation-induced erythema is a response to accumulative tissue injury resulting from irradiation. The daily radiation dose is a constant; therefore, it is more representative to measure erythema progression in terms of dose, fractions or treatment days rather than to count the number of days since the start of treatment. Hence, DTE was measured in "Treatment Days". Choosing the number of days since the start of treatment to track DTE would inflate the delay in erythema reactions because it included non-treatment days such as weekends, missed treatments or statutory holidays. Note that by using this definition, since measurements were taken before each treatment, on day 1 (treatment day 1) the dose is zero and therefore EI for TRT ROI for day 1 would the EI baseline for normal skin within the treatment field.

A scatterplot comparing DTE for EI (DTE_{EI}) versus DTE for VA (DTE_{VA}) was used to decide if one approach for assessing erythema was more sensitive compared to the other. The treatment day that erythema was first picked-up by VA (the standard) was plotted against the treatment day erythema was first picked-up by EI (experimental data). For this scatter plot, points on the 1:1 trendline indicate that both modalities have similar sensitivity; whereas, data points below the 1:1 trendline point towards EI having a greater sensitivity compared to VA for detecting early erythema. Conversely, data points above the 1:1 trendline point to VA as a more sensitive tool for assessing early erythema compared to EI.

4.3.5. The Spearman's Rank-Order Correlation between EI and VA

The Spearman rank-order correlation coefficient (rho) is a nonparametric measure of the strength and direction of association that exists between two variables measured on at least an ordinal scale. [52] This test is used either for ordinal variables or for continuous data that has failed the assumptions necessary for conducting other tests (such as Pearson's correlation). There are two assumptions that are required for Spearman's correlation to give a valid result; the two variables should be measured on an ordinal, interval or ratio scale and there must be a monotonic relationship between the two variables.

A monotonic relationship exists when the variables increase in value either together, or as one variable value increases, the other variable value decreases. [53] When the correlation coefficient rho is equal to –1 or +1, each of the variables is a perfect monotone function of the other. [54] For this study, an absolute value for rho of 0.65 or greater (rho ≥ 0.65 or rho ≤ – 0.65) was arbitrarily considered a strong monotonic association. To check for a monotonic relationship between your two variables, one variable was plotted against the other using a scatterplot. When the data of the scatterplot was in the shape of a "∩" or "∪" it was considered non-monotonic relationships and therefore Spearman's correlation would not yield accurate results. [53]

4.4. Results

4.4.1. Trends and Patterns for EI and VA Time-Sequence Plots

A visual assesment of the ten time-sequence plots ([Appendix C]) suggested the EI values for patients P01, P02, P04, P06, P07 and P09 (60%) increased with time, and

for patients P03, P05, P08 and P10 (40%) EI decreased with time. On the other hand, the observed time-sequence trends for VA data suggested that one patient had no change in VA grade (P01), one patient had only a one-grade increase in VA score (P10), whereas the remaining 8 patients had a solid increase in VA grades during treatment. The initial VA grade for patients P01, P02, P04 and P09 (40%) was zero and for patients P03, P05, P06, P07, P08 and P10 was G1 (60%) (Table 4-4).

Four examples of erythema time-sequences for EI and VA data were selected to illustrate the trends and patterns observed in more detail (Fig. 4-2A&B and Fig. 4-3 A&B). Case P02 and case P06 are examples of the TRT EI values that increased over time with subsequent radiation treatments. Unfortunately, there was no EI data recorded for the TRT ROI for the first 2 days of treatment for P02, but from days 3 to 6 the EI fluctuated close to the threshold (long dashed line in Fig. 4-2A). This is the similar case for P06, from days 1 to 6 the EI closely followed the respective EI threshold with exception of day 1 when the TRT EI was below threshold (Fig. 4-2B).
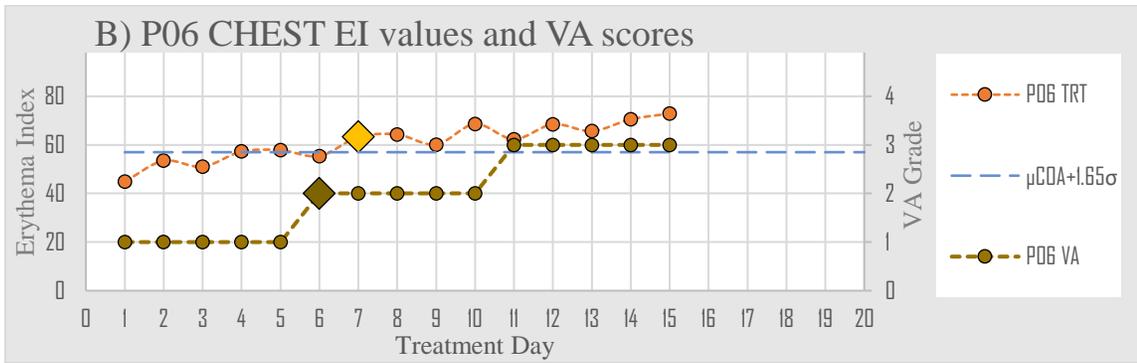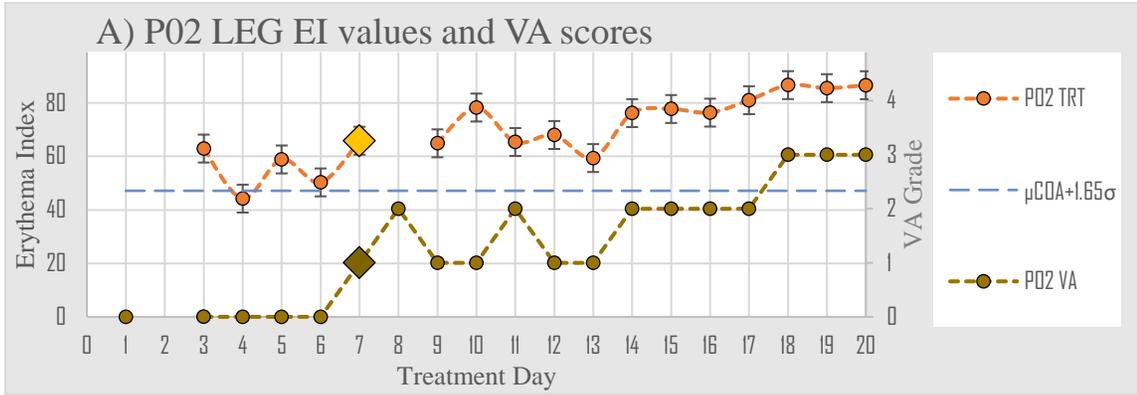
Figure 4-2A&B. *P02 and P06 are examples of two patients showing an increasing trend for EI and VA grades over time as would be expected with progressive development of the skin reaction. The DTE_{EI} and DTE_{VA} are indicated with a large diamond marker. The long dashed line represents the patients' skin EI threshold calculated from the control site.*

Starting on the 7th day of treatment and continuing for the remainder of the treatments, the TRT EI for P02 and P06 were consistently greater than the EI threshold. Day 7 was the first treatment day that erythema was identified for P02 and P06, and it continued to increase thereafter. Hence, Day 7 was identified as the day-to-erythema for the TRT EI (DTE_{EI}) for both cases (Fig. 4-2A&B). Similarly, the VA grades for P02 and P06 showed an overall increasing trend from the start of treatment to its completion. The

95

VA grade on the first day of treatment (baseline) was zero for P02 and was G1 for P06. The DTE by VA (DTE$_{VA}$) (§4.4.3) for P02 was day 7 and for P06 day 6 (Fig. 4-2A&B).

Patients P02 and P06 demonstrated a similar EI pattern from week to week. On the first week of treatment, the EI remained close to the threshold, on the 2nd week it surpassed the threshold. On the last day of week 2 (day 10) the EI had climbed considerably above the threshold, but them dropped close to threshold on the first day of week 3 to only to rise again towards the end of the week and continued to rise during week 4 until the end of the treatment.

The cases P03 and P08 are examples of two patients that demonstrated an overall decreasing trend for EI over time for the TRT ROIs (Fig. 4-3A&B). Initially, the EI in P03 increases until day 10 when it drops below the baseline EI. Then the EI increases slightly and drops again on day 15, then continues at the level of the EI threshold until the completion of treatment (day 20). The EI values for Patient P08 do not fluctuate over time as in P03, but rather decline steady since the first day of treatment until the end. Of note is that for both cases, P03 and P08, the baseline erythema indices for the TRT ROIs are greater than the EI thresholds calculated from the respective control ROIs. Using the convention previously defined (§4.4.3), the DTE for P03 was found to be day 1. The day-to-erythema for P08 was day 2. There was no data available to calculate the EI for TRT ROI on day 1. The VA grades started at grade 1 for both P03 and P08. The VA grades then increase overtime as treatment progressed. DTE$_{VA}$ for P03 and P08 was day 6 (Fig. 4-3A&B).
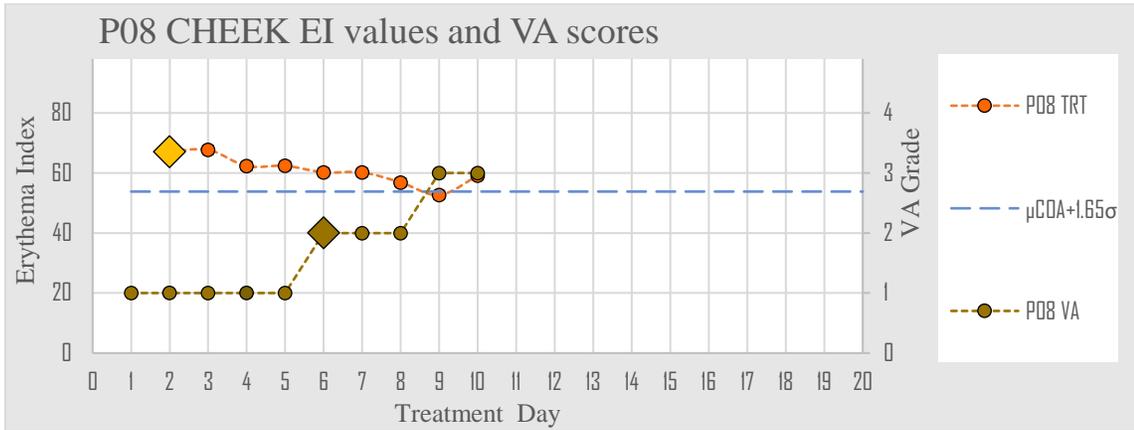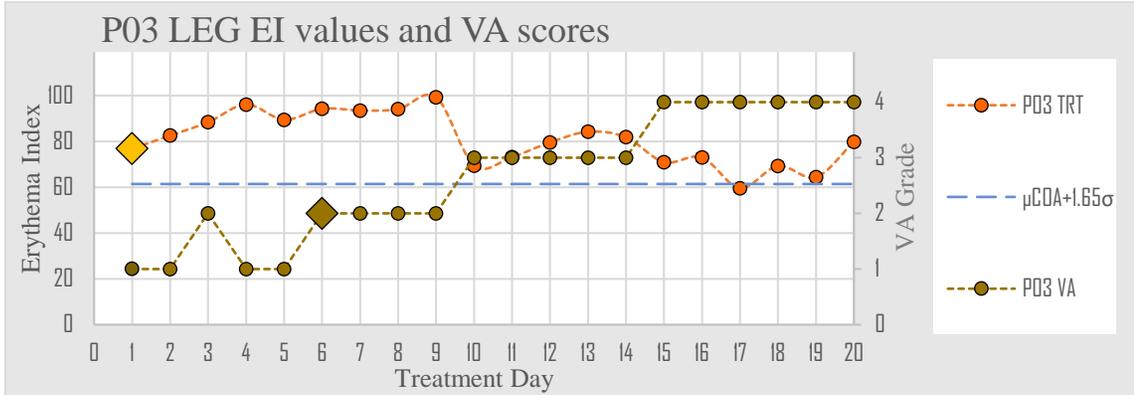
Figure 4-3A&B. *P03 and P08 are examples of two patients showing a decreasing trend for EI and an increasing trend for VA grades over time. These seemingly opposed results may be explained by the gradual development of skin desquamation within the TRT ROIs. The DTE$_{EI}$ and DTE$_{VA}$ are indicated with a large diamond marker. The long dashed line represents the EI threshold calculated from the control site. Both cases have a baseline TRT EI that is greater than the EI threshold, which may be explained by the presence of pretreatment skin reddening in the TRT ROIs and is supported by a VA grade of 1 on day1.*

Patients P03 and P08 showed a different week to week pattern for EI compared to patients P02 and P06. For these cases, the EI started and remained above threshold week 1 and remained above the erythema threshold during week 2. On the 3rd week of treatment, the EI declined and this decline continued throughout the 4th week until the end of the radiation treatment.

4.4.2. Assessment of Clinical Pictures

Daily RGB images of the treatment field were taken for each patient. A photographic time-sequence of the treatment field for patients P02 and P03 is shown in figure 4-4. The treatment day and VA grade is identified in each image. The images show progressive skin reactions within the TRT ROI (yellow circle) as well as in the entire field. The images in the top row correspond to P02 and are an example of increasing EI and increasing VA grade over time. The degree of skin reddening within the ROI intensifies with the number of treatment days and correlates with increasing VA grades (from G0 (interpolated value) to G2 and then G3).

The bottom row correspond to images of P03 and are an example of decreasing EI and increasing VA grade over time. Clinical assessment of the TRT ROIs for P03 showed evidence of initial minimal erythema progression to dry scaly and later to pale skin plaques with moist desquamation (Fig. 4-4). On the first day of treatment, P03 exhibited a few dotted areas of minimal erythema (G1) within the TRT ROI. On day15 of treatment, the skin was dry and scaly (referred to as dry desquamation) with diffuse G4 (very bright) erythema confined to trt area. On the last day of treatment (day 20) there was widespread G4 erythema extending beyond treated area. This was coupled with the loss of skin integrity with pale-white skin plaques and generalized moist desquamation with patches of dry skin was over the entire field including the TRT ROI (Fig. 4-4).
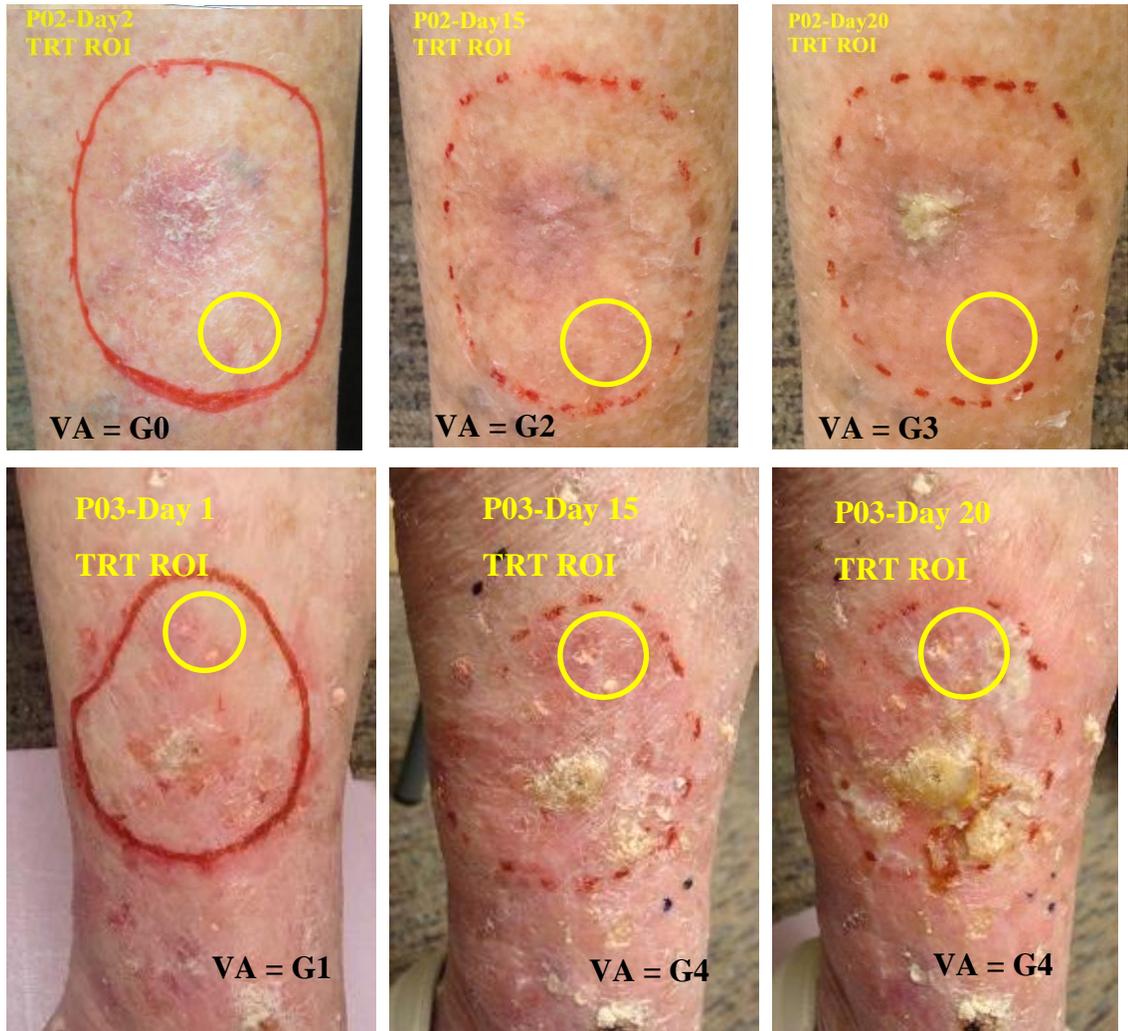
Figure 4-4. *A photographic time-sequence of the treatment field for patients P02 and P03. The treatment day and VA grade is identified in each image. The images show progressive skin reactions within the TRT ROI (yellow circle) as well as in the entire field.*
The images in the top row are an example of increasing EI and increasing VA grade over time. The degree of skin reddening within the ROI intensifies with the number of treatment days. The bottom row is an example of decreasing EI and increasing VA over time. At baseline, (day1) clinical assessment reveals spotty G1 erythema; on treatment day 15 and day 20 there is an increase in size of the area with redness yet the increase in intensity is less obvious. At this point, the skin developed dry and later moist desquamation (last day) which masks erythema and may explain why the clinician assigned a high VA grade. The decline in EI over time may be the result of the loss of skin surface tissue integrity.

4.4.3. Day-to-Erythema (DTE)

The day erythema was first detected by EI (DTE$_{EI}$) was plotted against the day erythema was first detected by VA (DTE$_{VA}$) using a scatterplot (Fig. 4-5). Nine out of the 10 patients in the study (90%) were included in this analysis. Patient P01 was excluded because he did not develop visible skin reddening during treatment and therefore VA grade was zero throughout treatment.

The 1:1 trendline is marked as a dotted blue line on the DTE scatter plot (Fig.4-5). Points on this trendline indicate that DTE$_{EI}$ and DTE$_{VA}$ are the same, hence VA and EI have the same sensitivity for detecting erythema. Out of 9 patients, 3 (33%) patients (P02, P06 and P07) had similar DTE$_{EI}$ and DTE$_{VA}$, that is, the number of days to initial detection of erythema was the same or there was less than one day difference when comparing EI to VA. In 6 out of 9 (67%) patients (P03, P04, P05, P08, P09 and P10), the DTE$_{EI}$ was less than DTE$_{VA}$ suggesting that the EI method had better sensitivity for detecting erythema.

When comparing the EI baseline for the TRT ROI to the EI threshold, 6/10 cases (60%) had a baseline EI that was below EI threshold and 4/10 (44%) had a baseline EI that was above the EI threshold. See table 4-4 for a list of all the patients and corresponding data for baseline erythema for TRT ROI, DTE$_{EI}$ and DTE$_{VA}$ and table 4-6 for a list of the EI thresholds.
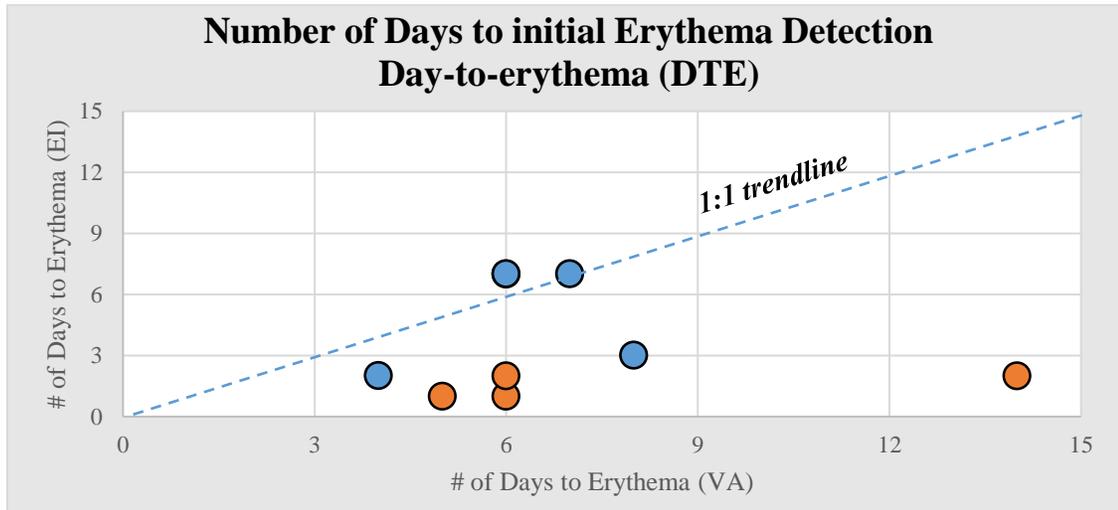
Figure 4-5. *Scatterplot for day-to-erythema (DTE). The 1:1 trendline indicates when the DTE was the same for EI and VA. One case showed slightly more sensitivity for VA, 2 cases had the same sensitivity for EI and VA and 6 cases showed greater sensitivity for EI. Five cases had a baseline EI similar to or below threshold (blue markers) and 4 cases had a baseline EI greater than the threshold (orange markers).*

Table 4-4. *Data for baseline VA grade the TRT ROI and DTE for EI and VA*

Patient	Site	Baseline VA (TRT ROI)	DTE $_{VA}$	DTE $_{EI}$
P01∗	neck	0	-	5
P02	leg	0	7	7
P03	leg	1	6	1
P04	forearm	0	4	2
P05	cheek	1	5	1
P06	chest	1	6	7
P07	ear	1	6	7
P08	cheek	1	6	2
P09	forehead	0	8	3
P10	leg	1	14	2

Note: ∗ *Patient P01 did not develop visible erythema during the course of treatment.*

4.4.4. Spearman's Correlation Analysis for EI and VA

The Spearman's rank-order correlation is the nonparametric test used to measure the statistical dependence between the two ranked variables, in this case TRT EI and VA grades. The results of this test are listed in table 4-5. The scatterplots evaluating monotonicity for the two variables are found in Appendix C.

Table 4-5. Spearman's rank-order correlation results comparing TRT EI and VA grades.

Patient	Site	Spearman rho coefficient	P-Value	Correlation assessment	Correlation direction
P01*	neck	NA	NA	NA	NA
P02	leg	0.88	0	Strong	(+)
P03	leg	-0.74	0	Strong	(−)
P04	forearm	0.68	0	Strong	(+)
P05	cheek	-0.80	0.01	Strong	(−)
P06	chest	0.83	0	Strong	(+)
P07	ear	0.84	0	Strong	(+)
P08	cheek	-0.90	0	Strong	(−)
P09	forehead	0.57	0.04	Weak	(+)
P10	leg	-0.02	0.94	Not significant	(−)

*Note: *Rho was not computed for P01 since the VA grade was zero for the duration of therapy and did not follow a monotonic relation.*

The data for case P01 was not included in the analysis because the VA grades assigned by the clinician were zero for the entire treatment. Of the remaining 9 cases, 8 cases had a statistically significant correlation between TRT EI and VA scores. The p-value for case P10 was not statistically significant (p-value = 0.94). Of these 8 cases, 5 (62.5%) had a positive association and 3 (37.5%) had a negative association between TRT EI and VA.

A strong positive monotonic correlation (rho ≥ 0.65) was found for patients P02,

P04, P06, and P07. Case P09 only showed a weak association (rho = 0.565). A strong

negative monotonic correlation (rho ≤ –0.65) was seen for cases P03, P05 and P08 (Table

4-5). All patients with a negative monotonic association excluding P10, had a baseline

TRT EI that was greater than the erythema threshold. The reverse was also true. Patients

with a positive rho had a baseline TRT EI that was at threshold or below threshold (Table

4-5 and Table 4-6).

Table 4-6. *Baseline, threshold and standard deviation values EI for all the ROIs.*

Case	TRT Baseline EI	*TRT σ*	EI threshold *(μEI + 1.65σ)*	TRT EI Baseline vs EI threshold	COA μEI	*COA σ*	*COB μEI*	*COB σ*
P01	40.7	1.8	49.5*	Below	74.4	11.0	39.8	5.83
P02	62.9	2.7	47.2	At threshold[≠]	43.0	2.5	45.0	3.85
P03	76.8	0.9	61.3	Above	51.8	5.7	46.1	4.31
P04	46.1	1.6	43.2	At threshold[≠]	35.4	4.7	NA	NA
P05	76.0	1.5	45.0	Above	41.3	2.3	36.0	4.39
P06	44.9	0.8	57.0	Below	50.8	3.8	44.2	3.45
P07	55.3	1.5	57.2	At threshold[≠]	45.5	7.1	35.6	4.08
P08	67.0	0.7	53.6	Above	47.7	3.6	59.0	5.40
P09	50.9	2.8	48.5	At threshold[≠]	38.9	5.8	NA	NA
P10	55.3	1.5	52.9	Above	50.5	1.5	NA	NA

*Note: *The EI threshold was calculated using COB data due to the large σ for COA values. All other EI thresholds were calculated using COA data. [≠]The "At threshold" label is an estimate derived from the time-sequence plots for the first 3 days of treatment that concludes the EI Baseline for TRT is similar to EI threshold.*

4.5.Discussion

This study compared two approaches for evaluating radiation-induced erythema,

the experimental approach, diffuse reflectance spectroscopy (DRS) and the standard,

visual inspection. Visual inspection by a trained clinician was achieved using the

Clinician Erythema Assessment for Radiation Therapy (CEA-RT), its validity was studied in chapter 3. Although visual assessment is the standard clinical method used by clinicians for evaluating erythema, it is considered by many to be subjective and dependent on the observer's skills.

The experimental approach on the other hand may offer a more objective assessment of erythema. DRS has been demonstrated to detect changes in skin reddening by exploiting the optical properties of skin. [40] More specifically, DRS detects changes in concentration of hemoglobin within the skin, which are linked to changes in erythema. DRS can be used to measure erythema using the EI (Dawson *et. al.*) as a proxy and therefore has the potential for being an objective continuous scale for erythema values.

This study evaluated DRS by comparing association between EI and VA based on the time-sequence trends for EI vs VA plots, the date to first detection of erythema (DTE) and the correlation between EI and VA grades using Spearman's test.

4.5.1. Erythema time-sequence plots for EI vs VA
The values for EI for the TRT ROIs increased over time for only 60% of the cases, 40% had decreasing EI values. This is not the trend expected based on the known tissues changes that would result from a radiobiological induced response.

The accumulation of radiation dose triggers the development of inflammation resulting in an increase of superficial hemoglobin circulating (erythema) in the exposed tissues. Therefore, since DRS is known to detect changes in hemoglobin concentration, one would assume the EI calculated form the tissue's spectral profile would also increase with dose. As mentioned previously, 40% of the cases showed a decrease in EI over time

yet visual assessments, performed by a trained clinician, showed that only 1 out of 10 patients (10%) had no change in VA grade (P01), and the remaining 90% (9/10) of the patients had at least a one-grade increase in VA grade. At first glance, these results seem to suggest VA may be more effective at detecting erythema changes compared to DRS or EI to be more specific. Althuogh VA grades increase over time for most case, this may not be purely due to the increase in erythema intensity of the TRT ROI, but rather the clinician's perception of the progressive intensity of the skin toxicity.

Analysis of the pattern for 4 cases (P02, P03, P06 and P08) shows EI lingered around the erythema threshold during the first days of treatment (P02, P06) and then increase for week 2, dropped in the beginning if week 3 and the continued to increase until the end of week 3 (P02, P03, P06) (See Fig. 4-2A&B and Fig. 4-3A&B). Radiation therapy is fractionated into small daily treatments and usually patients do not have treatment during the weekends. This fractionation and treatment breaks help reduce the acute side effects of therapy by allowing the normal cells to repopulate and repair tissue damage. The drop in EI value between week 2 and 3 may be explained as a successful reduction of radiation induced inflammation and erythema reaction. With add treatments there is added insult to the skin which manifests as progression of skin reaction and damage to the extent that a week end break is not enough time to repair the damage accumulated.

On the last week of treatment the pattern changed; P02 and P06 experienced an increase in EI and P03 and P08 experienced a decrease in EI down to the level of the EI threshold. These observations are well correlated with the clinician's observations and are

evident in the daily RGB images. The TRT ROI for patient P02 exhibited a progressive increase in the intensity of erythema from no erythema (G0) to bright erythema (G3) with the skin integrity remaining intact. For this patient DRS seems to successfully measure erythema as the EI follows the VA grades (Fig. 4-4). For patient P03 the VA grades assigned by the clinician start at G1 (minimal erythema) and increase to G4 (very bright erythema), yet the EI values decrease overtime and return to close to the threshold. DRS in this case does not appear to be able to the measure erythema progression at the end of the treatment. Since this patient developed skin desquamation and moist desquamation, the skin was flacky, with areas of fluid discharge and white plaques. It is plausible that the loss of normal skin architecture due to tissue breakdown may limit the abilty of the DRS system to obtain a reliable spectral profile. Therefore, the development of desquamation may be a limitation for measuring erythema with DRS.

4.5.2. DTE for EI (DRS) compared to VA (standard)

The treatment day on which for the first time the EI was greater than the EI threshold for 2 days in a row, was defined as the Day-to-Erythema (DTE) (§4.4.3). The scatterplot for DTE_{EI} versus DTE_{VA} indicated that Day-to-Erythema varied between cases. One case (P02) showed slightly more sensitivity for VA by 1 day and 2 cases had the same sensitivity for EI and VA (P06 and P07). Six cases (P03, P04, P05, P08, P09 and P10) showed greater sensitivity for EI; out of these cases, 4 cases (P03,P05, P08 and P10) had a baseline EI greater than the EI threshold. If the Baseline EI is greater than the EI threshold this suggests that there is likely preexcisting skin redness in the TRT ROI compared to the control ROI. This is consistent with the the clinical observations, since

the 4 cases mentioned previously all have and initial VA grade =1 corresponding to minimal erythema. If the skin of the TRT ROI is visably more red than that of the control ROI, it is not valid to assume that the skin of the control ROI had similar optical properties as the TRT ROI. If this was the case than the control was not an appropriet choice to calculate an EI threshold for the TRT ROI. For these 4 cases (P03,P05, P08 and P10), DTE may be overstimating the sensitivity of EI or DRS since the TRT EI is above threshold since the very beginning of treatment. A more accurate approach may be to not rely on a convenient control site but to establish a threshold by measuring EI for TRT ROI several times before the start of radiation, since the EI for day 1 would also the EI baseline for normal skin within the treatment field (§4.4.3).

4.5.3. Spearman's Correlation Analysis for EI and VA

The Spearman's rank-order correlation test measures of the strength and direction of the relationship between two ranked variables; in this case, the variables compared were TRT EI and VA grades. This test calculates a coefficient of determination rho, which assesses how well the relationship between two variables can be described using a monotonic function. Nine out of 10 patients were included in this analysis. P01 was excluded because the VA scores were null. Another patient, P10 had non-statistically significant correlation which may be related to the fact that the VA scores for P10 only had 2 grade categories which is not enough to establish monotonicity. A statistically significant positive correlation (rho >0) was found for 63% (5/8) of the patients which indicated that increasing VA grades correlated with increasing EI value (Table 4-5). A significant negative correlation (rho <0) was found for 37% (3/8) of the patients which

indicated that decreasing EI values correlated with increasing VA grades (Table 4-5). The data collected did not show any examples of a persistent decreasing trend for VA grades. The rho coefficients are consistent with the trends observed in the time-sequence plots. Patients with increasing EI over time had a positive rho, and patients with decreasing EI over time had a negative rho coefficient.

Patients with a positive rho had a baseline TRT EI that was at threshold or below threshold. Similarly, patients with a negative rho had a baseline TRT EI that was above threshold (Table 4-5 and Table 4-6). Althuogh it is not possible to determine within his study whether these findings are significant, patients with an initially high EI at baseline may be more prone to developing severe skin reaction such as dry and moist desquamation early on since they already may be showing signs of inflammation. Likewise, treated skin that has an EI at baseline that is similar or lower than the threshold may be more resistant or may require more time to progress to desquamation.

4.5.4. Challenges with data collection and sources of error

Running repeat experiments on real patients over several days can be challenging work for investigators and participants. During clinic time, patients may often become anxious or impatient, especially if they have multiple appointments and they are tired, or in discomfort, or have a family member or friend that is waiting for them. Sometimes patients arrived late or too early and the investigators would need to rush to set-up the research equipment. At times patients complained they were feeling very hot or cold; these various states could affect the patients' skin in the form of pallor or flushed skin, which could potentially affect the assessment of erythema. Also, DRS

measurements were obtainied by holding the probe against the patients skin which could induce temporary blanching of the skin during measurements and lower the EI values.

Patients liked to engage in conversation during the study session, which made it more likely for the investigator to be distracted resulting in errors or missing data while obtaining measurements. The DRS system was an experimental prototype assembled with components designed for research applications. Although the system performed well, it was very hands on and required the operator's full attention and operation with both hands simultaneously. Adding a digital device to the DRS system able to compute and display the EI in real time, without the need of post-acquisition processing, could help reduce user errors and missing measurements.

The acquisition protocol involved several sequential steps to ensure accurate measurements; this, coupled with the need for a low lighting environment made the system not ideally suited for a busy clinical environment.

The integrating sphere was set back within a Spectralon cube with an external square surface measuring approximately 5x5cm (Table 2-2). The cube probe obstructed the users'view of ROI and made accurate placement of the port over the ROI challenging. The size and flat shape of the Spectralon cube made it difficult to ensure a good seal around the port for uneven surfaces such as the ear and face. Therefore, loss of signal due to a poor skin-port contact and variation in the positioning of the probe need to be considered.

Reproducing the daily position of the ROIs was important for accurate results. A plastic template was used to map the position of the ROIs on the patients' skin each day.

Patients were asked not to wash the marks off but often the marks would fade from one day to the next. Although templates were useful for reproducing skin marks, they also were not free from variation since a flat plastic sheet does not conform well to the curved shape of the patients' anatomy. In some cases, the DRS data were inadvertently acquired using the software's default parameters. The default parameters deviated from the study protocol which were chosen to optimize data quality. Data collected using the default settings had a higher level of noise. To compensate for the increased noise, the data were made smooth using a moving average of 10 wavebands resulting in some loss of data.

4.5.5. *Limitations of the data/ study*

As many other studies, this study has some limitations due to constraints on the research design or methodology. These limitations can influence the findings and therefore need to be acknowledged for accurate interpretation of the results.

Sample size of the study only involved 10 subjects. Although these patients had various demographic characteristics that represent well the skin cancer patient population at JCC, the sample population was too small to be able to confidently generalize observations.

All the study participants were Caucasian. Melanin concentration is known to be low in this group, therefore the contribution of melanin to the EI was not considered. Other ethnic groups such as Asians, Indigenous people, Latinos, and African Americans have higher melanin concentration in the skin which should not be readily ignored. Hence, the assumptions made in this study precludes its results from being generalized to other ethnic groups with various skin pigmentation.

It was not possible to ensure visual assessment data did not suffer from clinician recall bias. The same clinician assessed all cases so that the VA data would have the same standard applied to all assessments. The clinician may have been aware of the number of treatment days each patient received but they were less likely to know the energy or the daily dose prescribed to each patient. Therefore, radiation toxicities such as erythema were harder to anticipate.

The clinician graded the treatment area for erythema by observing the entire treatment field. In some instances, the clinician may have underestimated or overestimated the erythema grade for the TRT ROI based on the global assessment of the fields. The use of a device to obscure the field outside the TRT ROI such as a neutral grey cardboard with a cut-out to expose only the ROI may have helped the rater be more focused on the ROI and be less influenced by the surrounding skin reaction.

Another limitation of the study is that the clinician was only instructed to grade the treatment field. There was no real time visual assessment of the control ROIs. It was assumed that the control regions would be representative of the baseline erythema for the patient's normal skin yet it was not assessed by the standard, visual assessment, hence there is no certainty on whether the controls identified the true erythema thresholds which inturn would influence the accuarcy of DTE_{EI}. Althuogh not ideal, the control ROIs, could be visually assessed retrospectivelly using the RGB images obtained of the treatment field.

4.5.6. Conclusions for DRS vs VA Study

The present study was designed to evaluate the correlation between diffuse reflectance spectroscopy measurements and visual inspection for assessing erythema. Ten patients treated with radiation therapy for skin cancer volunteered for this study. The treatment area was assessed daily using the DRS system, digital RGB pictures and visual inspections by a trained clinician. Dawson's EI was calculated from the DRS spectral profiles as a measure of erythema. DRS was compared to the standard measure for erythema, visual inspection. A trained clinician used the CEA-RT 5-point scale to grade erythema reactions. The daily EI were compared to the clinician's grades for accuracy. The data were analyzed using time-sequence plots for EI vs VA and the Spearman's correlation test. A strong statistically significant monotonic correlation ($|rho| \geq 0.65$) was found for 8/10 (80%) patients. Only one patient showed a weak association (rho = 0.565). Both interpretations demonstrated that the EI correlate well with the gold standard (VA grades) and that DRS is able to detect changes in the skin throughout the course of radiation treatments. DRS may offer more subtle details for day-to-day changes compared to VA. The EI values varied from week to week in a manner that was consistent with the radiobiological response of normal tissue. One of the limitations of this spectroscopy system is its dependence on the integrity of the skin architecture. DRS was not able to register erythema changes once the skin reaction had progressed to desquamation. Intact smooth skin is better suited for assessing changes in the optical properties of skin.

DRS measurements were successfully obtained in a clinical setting using the study system, but some opportunities to further improve the methodology and design were identified during the course of this study. Preexisting erythema in the regions of

interest was not fully addressed. Erythema at baseline needs to be accounted for to accurately determine the first day-to-erythema. Therefore, daily VA grades should be obtained for all the ROIs, and EI measurements of the treatment ROI should be repeated several times before the start of treatment.

The potential implication of this research is the addition of a practical tool for objectively monitoring erythema changes to aid the clinician. The objective quality of DRS may also facilitate knowledge transfer and comparing multiple research trials.

5. Thesis Conclusions

Toxicity to normal tissues is still a major limiting factor for medical practices involving high doses of ionizing radiation, particularly in radiation therapy and interventional fluoroscopic radiology. One of the most common effects of exposure to ionizing radiation is dermatitis, which can vary from mild inflammation and redness (erythema) to severe desquamation, ulceration and skin necrosis.

As opposed to unexposed skin, skin exposed to the radiation field will react progressively as the radiation dose accumulates and damage to the skin deepens. Inflammation is a natural defense mechanism evolved to protect the organism. When inflammation occurs in the skin, it leads to swelling and vasodilation, triggering an increase in blood circulation and interstitial Hb concentration. This is known as erythema and is an important indicator for the onset of tissue toxicity. Consequently, standardized strategies to assess for tissue toxicities are key to ensuring patient safety.

The standard for assessing erythema is visual inspection by a trained clinician. However, this method is prone to bias and reliant on the skills of the assessing clinician. Furthermore, there are several grading systems used in clinical trials for visual assessment of radiation dermatitis. The absence of a global grading scale for skin reactions makes it difficult to link results from different studies and has contributed to the lack of reliable evidence-based knowledge on the optimal management of skin toxicities. A more systematic and quantifiable approach for monitoring early skin toxicities prior to their full development may offer new opportunities for non-expert clinicians to suitably evaluate

reactions, and offer an objective and uniform measure for grading radiation induced skin reactions and for timing of therapeutic interventions.

Technology capable of detecting changes in the optical properties of skin such as hyperspectral imaging (HSI) and diffuse reflectance spectroscopy (DRS) have the potential to provide a quantitative and objective assessment of structural and functional changes in tissue such as skin erythema. This work explored the feasibility of using HSI and DRS technology as an alternative to standard VA for grading erythema in the clinical setting.

The artificially induced erythema pilot discussed in chapter 2 demonstrated that the spectra obtained using an experimental HSI system showed a decrease in reflectance within the 500-600 nm wavelength range when eryhtema was artificially induced. This was consistent with the known optical properties of skin and expected changes in concentration of OHb and DHb. The Dawson EI was then calculated from HSI reflectance data and found to be consistent with the increase in erythema grade. In this experiment, HSI system proved to be capable of detecting erythema changes and was well optimized for obtaining experimental data. For a more extensive clinical application, this system could be improved by adding features that would increase the camera's portability, image reproducibility and patient comfort. A more portable camera would facilitate bringing the technology to the patient rather than requiring a dedicated clinic space and additional patient appointments. A support stand that could hold the camera in a variety of reproducible and stable positions would improve the image perspective and data collection. Adding dual hyperspectral-RGB imaging to the HSI system would facilitate documentation of the object's features as well as image-to-image registration. Lastly, the halogen lights used for the pilot study became very hot during the imaging sessions; therefore, a cooler source such as broad-spectrum LED lights would be a better option.

Another accomplishment of this work was to develop the CEA-RT scale since there was no literature evidence for a validated tool for grading early radiation induced erythema. The CRIE study assessed the validity of the CEA-RT scale (chapter 3) by determining intra-rater and inter-rater agreement among 20 trained radiation therapists. The results showed that 85% of the therapists matched their own grades between repeat trials with at least "good" agreement (100% attribute agreement was $\geq 50\%$) but only 20% of the therapists matched grades when compared to the standard. When order between grade categories was considered, 95% of the therapists had "high" or greater Kendall's correlation (KCC ≥ 0.80) with themselves, but only 60% had at best "medium" Kendall's correlation (KCC ≥ 0.70) when compared with the standard. The high agreement between repeat scores shows that the therapists were consistent and reliable when grading erythema. Therapists performed less consistently when compared to the gold standard therefore, more work needs to be done to confirm the validity of the new CEA-RT scale.

A follow-up study designed to further evaluate the accuracy of the CEA-RT scale is recommended before it is introduced for clinical practice. Based on lessons learned from the CRIE study, the new study should include fewer raters (preferably 3-5 individuals) that are thoroughly trained to use the CEA-RT scale. Assessments should be performed on real patients to avoid uncertainty resulting from RGB images interpretation and to better reflect the new scale's intended clinical use.

The final accomplishment of the thesis work was to compare the DRS system described by Glennie *et al* to the gold standard. This had not been investigated before.

The DRS system was previously used to measure in vivo erythema in the radiation-exposed skin of patients treated for a head and neck cancer [55] but, it did not methodically compare daily DRS measurements to an accepted standard.

The DRS versus VA study (chapter 4) involved 10 patient volunteers diagnosed with early skin cancer and treated with radiation therapy. The system was used to obtain daily reflectance measurements of the patients' skin. Similarly, the patient's skin was rated daily using the CEA-RT 5-point grading scale by a trained clinician considered to be the gold standard. The Dawson EI calculated form the daily spectral data were compared to the corresponding daily VA grades.

Spearman's correlation showed a strong statistically significant monotonic correlation ($|rho| \geq 0.65$) for 80% of the patients. The results demonstrated that the EI correlated well with the gold standard (VA grades) and that DRS was able to detect changes in the skin throughout the course of radiation treatments.

One of the limitations of this spectroscopy system was its dependence on the integrity of the skin since DRS was not able to register erythema changes once the skin reaction had progressed to desquamation. This limits the scope of DRS.

DRS appeared to detect subtle changes in the skin, which were not captured by the clinician's assessments. The DRS measurements varied daily from week to week in a manner that was consistent with the radiobiological response of normal tissue. The VA grades did not follow the same week to week pattern as the EI values, instead VA grades increased steadily over time for most cases. DRS may have a greater potential for predicting acute reactions based on changes that are not distinguishable to the naked eye

compared to the standard VA. Further improvements were suggested to optimize the DRS system for clinical practice. The most significant improvement recommended was a more compact probe that would help avoid obscuring the region of interest during the measurement process.

References

[1] Reed AB, The history of radiation use in medicine, *Journal of Vascular Surgery*, Vol53 (1) Supplement, 2011, Pages 3S-5S, ISSN 0741-5214, URL:https://doi.org/10.1016/j.jvs.2010.07.024

[2] Skin erythema dose. (n.d.) McGraw-Hill Dictionary of Scientific & Technical Terms, 6E. (2003). URL:https://encyclopedia2.thefreedictionary.com/skin+erythema+dose. Access date August 31 2018.

[3] Alan Jennings, W. (2007). Evolution over the past century of quantities and units in radiation dosimetry. *Journal of Radiological Protection*, 27(1), 5-16. URL:https://doi.org/10.1088/0952-4746/27/1/R01

[4] Article title: Henri Becquerel biography. Website name: The biography.com Website URL: https://www.biography.com/people/henri-becquerel-40055 Access date: July 4, 2018 Publisher A&E Television networks. Last updated: April 1, 2014.

[5] Inkret W C, Meinhold C B, and Taschner J C. Radiation and Risk–A Hard Look at the Data. *Los Alamos Science*, Number 23 1995.

[6] Sansare K, Khanna V, Karjodkar F. Early victims of X-rays: a tribute and current perception. *Dentomaxillofacial Radiology*. 2011;40(2):123-125. URL:https://doi.org/10.1259/dmfr/73488299

[7] Einstein AJ, Moser KW, Thompson RC, Cerqueira MD, Henzlova MJ. Radiation dose to patients from cardiac diagnostic imaging. *Circulation*. 2007;116:1290–1305.

[8] Nguyen PK, Wu JC. Radiation exposure from imaging tests: is there an increased cancer risk? *Expert review of cardiovascular therapy*. 2011;9(2):177-183. URL:https://doi.org/10.1586/erc.10.184.

[9] Troyer J L, Jones A E, Shapiro N I, Mitchell A M, Hewer I, Kline J A.Cost-effectiveness of quantitative pretest probability intended to reduce unnecessary medical radiation exposure in emergency department patients with chest pain and dyspnea. *Acad Emerg Med*. 2015 May; 22(5):525-35.

[10] Kline J A, Jones A E, Shapiro N I, Hernandez J, Hogg M M, Troyer J, Nelson R D. Multicenter, randomized trial of quantitative pretest probability to reduce unnecessary medical radiation exposure in emergency department patients with chest pain and dyspnea. *Circ Cardiovasc Imaging*. 2014 Jan;7(1):66-73.

[11] Miller D G, Vakkalanka P, Moubarek M L, Lee S, Mohr N M. *Diagnosis. West J Emerg Med*. 2018 Mar;19(2):319-326.

[12] Linder F, Mani K, Juhlin C, Eklöf H. Routine whole body CT of high energy trauma patients leads to excessive radiation exposure. *Scand J Trauma Resusc Emerg Med*. Jan 27;24:7, 2016. URL:https://doi.org/10.1186/s13049-016-0199-2

[13] Halperin E.C., Brady L. W., Perez C. A., Wazer D.E., Principles and Practice of Radiation Oncology, Lippincott Williams and Wilkins, Philadelphia, 6th Edition (2013).

[14] Chiao-Ling Tsai, Jian-Kuen Wu, Hsiao-Ling Chao, Yi-Chun Tsai, Jason Chia-Hsien Cheng, Treatment and Dosimetric Advantages Between VMAT, IMRT, and Helical TomoTherapy in Prostate Cancer, *Medical Dosimetry*, Vol 36 (3), 264-271, 2011. ISSN 0958-3947. URL:https://doi.org/10.1016/j.meddos.2010.05.001

[15] Khan F.M., The Physics of Radiation Therapy, Lippincott Company, Williams and Wilkins, Baltimore, MD (2003).

[16] Ryan JL. Ionizing Radiation: The Good, the Bad, and the Ugly. *The Journal of Investigative Dermatology*. 2012;132(3 0 2):985-993. URL:https://doi.org/10.1038/jid.2011.411.

[17] Dehen L, Vilmer C, Humiliere C, Corcos T, Pentousis D, Ollivaud L, *et al*. Chronic radiodermatitis following cardiac catheterization: a report of two cases and a brief review of the literature. *Heart* 1999;81:308-12.

[18] Dunsakul V, Iamsumang W, Suchonwanit P. Radiation-induced alopecia after endovascular embolization under fluoroscopy. *Case Reports in Dermatological Medicine*, 2016 URL:https://doi.org/10.1155/2016/8202469

[19] Hall EJ., Giaccia AJ., Radiobiology for the Radiologist, J.B. Lippincott Company, Philadelphia, (2011)

[20] Hall E.J., Cox J.D. Physical and biological basis of radiation therapy. In: Cox J D, Ang K K, editors. Radiation oncology. St Louis MO: Mosby; 2003. pp. 3-62.

[21] Bentzen S. M. Preventing or reducing late side effects of radiation therapy: radiobiology meets molecular pathology. *Nat. Rev. Cancer* 6, 702–713, 2006.

[22] Tortora G.J., Grabowski S.R. Principles of Anatomy and Physiology. Harper Collins College Publisher 8th edition (1996). Chapter 5. The Integumentary System. p. 122.
[23] Bryant R., Nix D. Acute and Chronic Wounds, Current Management Concepts. 5th Edition. St. Louis, Missouri: Elsevier; 2016.)

[24] Hymes SR, Strom EA, Fife C, Radiation dermatitis: clinical presentation, pathophysiology, and treatment, *J. Am. Acad.Dermatol.* 54 (Jan. 2006) 28 –46.

[25] Schmuth M, Sztankay A, Weinlich G, Linder DM, Wimmer MA, Fritsch PO, *et al*. Permeability barrier function of skin exposed to ionizing radiation. *Arch Dermatol* 2001;137:1019-23

[26] Brown KR, Rzucidlo E. Acute and chronic radiation injury.*J Vasc Surg*. 2011;53:15S–21S.[PubMed]

[27] Bolderston A, Lloyd NS, Wong RK, Holden L, Robb-Blenderman L, Supportive Care Guidelines Group of Cancer Care Ontario Program in Evidence-Based C The prevention and management of acute skin reactions related to radiation therapy: a systematic review and practice guideline. *Support Care Cancer*. 2006;14(8):802–817. URL:https://doi.org/10.1007/s00520-006-0063

[28] Mendelsohn FA, Divino CM, Reis ED, Kerstein MD. Wound care after radiation therapy. *Adv Skin Wound Care* 2002;15:216-24.

[29] Cox JD, Stetz J, Pajak TF. Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC). *Int J Radiat Oncol Biol Phys*. 1995;31:1341–6. 3. Cancer Therapy Evaluation Program (CTEP) of the National Cancer Institute (NCI).

[30] Common Terminology Criteria for Adverse Events (CTCAE): version 4.0. US Department of Health and Human Services, National Institutes of Health, National Cancer Institute, Bethesda: Maryland; 2009. v4.02: 2009.

[31] Common Terminology Criteria for Adverse Events (CTCAE) version 5.0. URL: https://ctep.cancer.gov/protocolDevelopment/electronic_applications/docs/CTCAE_v5_Quick_Reference_8.5x11.pdf. Accessed date: July 22 2018.

[32] Noble-Adams, R. (1999). Radiation induced skin reactions. Development of a measurement tool. *Br J Nurs* 8(18): 1208-11, 1999 Oct.

[33] Heggie S, Bryant GP, Tripcony L, Keller J, Rose P, Glendenning M, *et al*. A phase III study on the efficacy of topical aloe vera gel on irradiated breast tissue. *Cancer Nursing* 2002; / 25: / 442-51.

[34] Olsen D, Raub W, Jr Bradley C, Johnson M, Macias J, Love V, *et al*. The effect of aloe vera gel/mild soap versus mild soap alone in preventing skin reactions in patients undergoing radiation therapy. *Oncol Nurs Forum* 2001; / 28: / 543-7.

[35] Ho AY, Olm-Shipman M, Zhang Z, Siu CT, Wilgucki M, Phung A, *et al*. A Randomized Trial of Mometasone Furoate 0.1% to Reduce High-Grade Acute Radiation

Dermatitis in Breast Cancer Patients Receiving Postmastectomy Radiation. *Int J Radiat Oncol Biol Phys*. 1;101(2):325-333, 2018 Jun.

[36] Tan J, Liu H, Leyden JJ, Leoni MJ. Reliability of Clinician Erythema Assessment grading scale. *J Am Acad Dermatol*. 2014 Oct; 71(4):760-3. URL:https://doi.org/10.1016/j.jaad.2014.05.044.Epub 2014 Jul 4

[37] T. Lister, P.A. Wright, P.H. Chappell, Optical properties of human skin, *Journal of Biomedical Optics* 17(9), September 2012.

[38] Stamatas G. N., Zmudzka B. Z., Kollias N., and Beer, J. Z. Non-Invasive Measurements of Skin Pigmentation In Situ. *Pigment Cell Res* 17: 618–626, 2004.

[39] Zonios G. & Dimou A., Light scattering spectroscopy of human skin in vivo., *Opt. Express*, vol. 17, no. 3, pp. 1256–1267, 2009.

[40] Glennie D. L., Hayward J. E., McKee D. E., Farrell T. J., Inexpensive diffuse reflectance spectroscopy system for measuring changes in tissue optical properties, *Journal of Biomedical Optics* 19(10), 105005, 7 October 2014.

 [41] Website name: www.oceanoptics.com. Website URL: https://oceanoptics.com/support/technical-documents/ Access date: Sept 18, 2018

[42] Website name: www.oceanoptics.com. Website URL:http://www.manualsdir.com/manuals/292616/ocean-optics-bluloop-multi-led-light-source.html. Access date: Sept 18, 2018

[43] Dawson, J.B., Barker, D.J., Ellis, D.J., Grassam, E., Cotteril, J.A., Fisher, G.W., and Feather, J.W. A theoretical and experimental study of light absorption and scattering by in vivo skin. In: *Physics in Medicine and Biology* 25.4 (1980-07), 695– 709.

[44] Abdlaty RM. Hyperspectral imaging and data analysis of skin erythema post radiation therapy treatment [PhD Thesis]. Hamilton, Ontario. School of Biomedical Engineering, McMaster University. September 2016. URL:https://macsphere.mcmaster.ca/handle/11375/20765.

[45] Madooei A., Abdlaty R., Doerwald-Munoz L., Hayward J., Drew M. S., Fang Q., Zerubia J., Hyperspectral image processing for detection and grading of skin erythema. *Proceedings* Volume 10133, Medical Imaging 2017: Image Processing; 1013322, 2017. URL:https://doi.org/10.1117/12.2254132

[46] Lu G. & Fei B. Medical hyperspectral imaging: a review. *J. of Biomedical Optics*, 19(1), 010901, 2014. URL: https://doi.org/10.1117/1.JBO.19.1.010901

[47] https://www.hireb.ca/guidelines. Access date: Nov 10, 2018

[48] Website name: www.minitab.com. Website URL: http://support.minitab.com/en-us/minitab/18/ . Access date: Sept 4, 2018

[49] Landis J. R., & Koch G. G. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159, 1977. URL: https://doi.org/10.2307/2529310

[50] De Veaux R.D.,Velleman P.F., Bock D.E., Stats Data and Models. 3rd Edition. Addison-Wesley, Boston. Chapter 8 (2012).

[51] Article title: What are the differences between patterns and trends? URL: https://www.investopedia.com/ask/answers/010715/what-are-differences-between-patterns-and-trends.asp. Access date: Mar 8, 2019 Investopedia. Last updated: Mar 30, 2018.

[52] Website name: www.minitab.com. Website URL: https://support.minitab.com/en-us/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/correlation-and-covariance/a-comparison-of-the-pearson-and-spearman-correlation-methods/. Access date: Jan 12, 2019

[53] Website name: www.minitab.com. Website URL: https://support.minitab.com/en-us/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/basics/linear-nonlinear-and-monotonic-relationships/. Access date: Jan 12, 2019

[54] Website name: www.minitab.com. Website URL: https://support.minitab.com/en-us/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/correlation/interpret-the-results/key-results/. Access date: Jan 12, 2019

[55] Glennie DL. Integrating sphere-based spectrally constrained total diffuse reflectance for in vivo quantification of hemoglobin in skin [PhD]. Hamilton, Ontario. Medical Physics, McMaster University. January 2015. URL: https://macsphere.mcmaster.ca/handle/11375/16596.

Appendix A

CRIE Study Presentation for Participants

Colour Temperatures in the Kelvin Scale

North Light (Blue Sky)

Overcast Daylight

Noon Daylight, Direct Sun
Electronic Flash Bulbs

Household Light Bulbs
Early Sunrise
Tungsten Light
Candlelight

9

Stanley Kubrick film 'Barry Lyndon' filmed and released in 1975

Director used natural lighting to illuminate scenes. Taken with candle light only 1000-2000K. Note all colours look warm.

10

Scene filmed outside on overcast day typical 7000-8000K range
Note colour green is bright but reds are dull

11

12

Paint Colour Terminology

<u>Hue</u> is a technical word to define a <u>basic or "pure" colour</u>.

- <u>Tint</u> is the amount of <u>white</u> added to a basic colour, a <u>tint is lighter</u> than the original colour.
- <u>Shade</u> is the amount of <u>black</u> added to a basic colour, a <u>shade is darker</u> than the original colour.
- <u>Tone</u> is the amount of <u>gray</u> added to a basic colour, a <u>tone is softer</u> than the original colour.

13

14

Colour Wheels and Charts

15

Image Colour Terminology

- <u>Hue</u> defines a basic or "pure" colour.
- <u>Saturation</u> defines the intensity of a colour, ranges from 100% hue saturation (pure color) to 0% saturation (neutral gray). A pure color is fully saturated and is perceived as a vivid color or image. A desaturated image is said to be dull, less colorful or washed out, but can also make the impression of being softer.
- <u>Value</u> is defined as the relative brightness (degree of whiteness) or darkness (degree of blackness) of a color.

16

Same image modified to show the effect of increasing <u>hue saturation</u> starting from neutral gray on the top left to highly saturated on bottom right. Brightness and colour temperature are constant

17

VALUE: Same image modified to show the effect of increasing <u>brightness</u> starting at black (-100%) on the top left to white (100%) brightness on bottom right. Saturation and colour temperature are constant

18

19

CEA-RT scale: Diagram shows skin images of progressing skin redness proposed as a 5-point grading scale for assessing radiation induced erythema

Grade 1 Grade 2 Grade 0 Grade 3 Grade 4

20

Grade 0 Normal Skin	Grade 1 Minimal Erythema	Grade 2 Mild Erythema	Grade 3 Moderate Erythema	Grade 4 Severe Erythema
No erythema or observable difference relative to normal skin.	Slight redness but distinguishable, very faint erythema diffuse appearance.	Definite redness but not very marked, faint erythema.	Marked redness, noticeable bright erythema.	Fiery redness, brilliant red, very bright erythema.

21

22

23

24

Appendix B

Statistical Analysis for CEA-RT study

Chi-Square Goodness-of-Fit Test for Observed Counts in Fequency_yr

Using category names in experience

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
0-10 yrs	8	0.455	9.1	0.13297
11-20 yrs	6	0.280	5.6	0.02857
21-30 yrs	3	0.200	4.0	0.25000
31-40 yrs	3	0.065	1.3	2.22308

2 (50.00%) of the expected counts are less than 5.

Chi-Square Test

N	DF	Chi-Sq	P-Value
20	3	2.63462	0.451
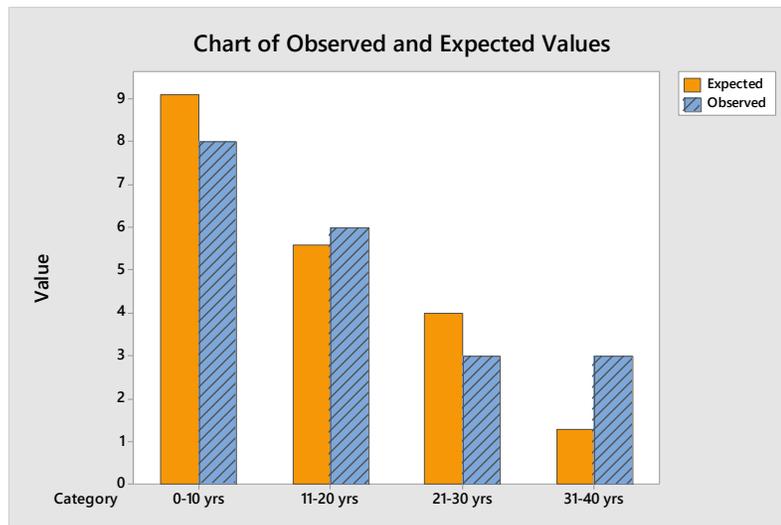
Chi-Square Goodness-of-Fit Test for Observed Counts in Frequency_mf

Using category names in gender

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
male	4	0.24	4.8	0.133333
female	16	0.76	15.2	0.042105

1 (50.00%) of the expected counts are less than 5.

Chi-Square Test

N	DF	Chi-Sq	P-Value
20	1	0.175439	0.675

Chart of Observed and Expected Values
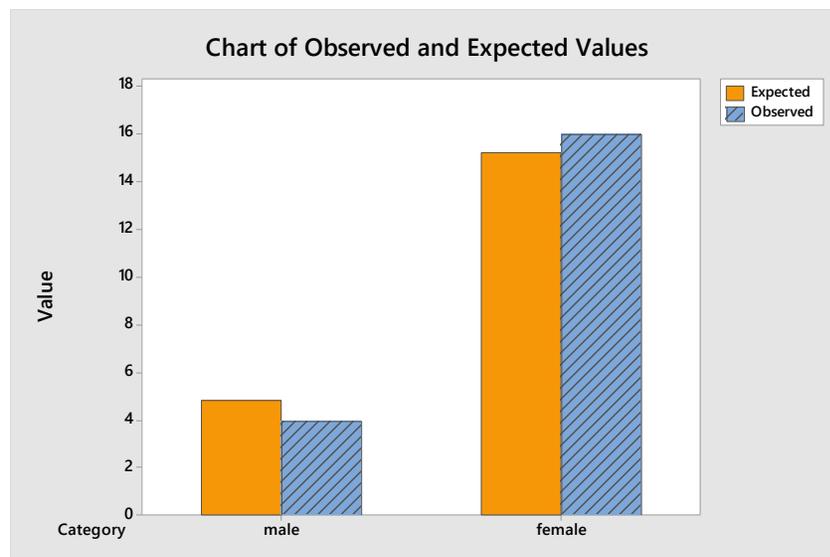
Attribute Agreement Analysis for response score
Within Appraisers-FOV-Assessment Agreement

Appraiser	# Inspected	# Matched	Percent	95% CI
1	60	34	56.67	(43.24, 69.41)
2	60	46	76.67	(63.96, 86.62)
3	60	45	75.00	(62.14, 85.28)
4	60	27	45.00	(32.12, 58.39)
5	60	37	61.67	(48.21, 73.93)
6	60	34	56.67	(43.24, 69.41)
7	60	37	61.67	(48.21, 73.93)
8	60	33	55.00	(41.61, 67.88)
9	60	44	73.33	(60.34, 83.93)
10	60	36	60.00	(46.54, 72.44)
11	60	37	61.67	(48.21, 73.93)
12	60	34	56.67	(43.24, 69.41)
13	60	29	48.33	(35.23, 61.61)
14	60	38	63.33	(49.90, 75.41)
15	60	35	58.33	(44.88, 70.93)
16	60	47	78.33	(65.80, 87.93)
17	60	27	45.00	(32.12, 58.39)
18	60	42	70.00	(56.79, 81.15)
19	60	45	75.00	(62.14, 85.28)
20	60	51	85.00	(73.43, 92.90)

 # Matched: Appraiser agrees with him/herself across trials.

Within Appraisers-FOV-Fleiss' Kappa Statistics

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)
1	0	0.223580	0.129099	1.73184	0.0417
	1	0.236111	0.129099	1.82891	0.0337
	2	0.335721	0.129099	2.60049	0.0047
	3	0.777365	0.129099	6.02145	0.0000
	4	*	*	*	*
	Overall	0.380584	0.080263	4.74171	0.0000
2	0	0.741193	0.129099	5.74126	0.0000
	1	0.524887	0.129099	4.06576	0.0000
	2	0.582146	0.129099	4.50929	0.0000
	3	*	*	*	*
	4	*	*	*	*
	Overall	0.584467	0.101682	5.74800	0.0000

3	0	0.744318	0.129099	5.76546	0.0000
	1	0.520000	0.129099	4.02790	0.0000
	2	0.383210	0.129099	2.96833	0.0015
	3	-0.008403	0.129099	-0.06509	0.5259
	4	*	*	*	*
	Overall	0.560225	0.095581	5.86123	0.0000
4	0	0.590754	0.129099	4.57596	0.0000
	1	0.027778	0.129099	0.21517	0.4148
	2	0.152941	0.129099	1.18468	0.1181
	3	-0.071429	0.129099	-0.55328	0.7100
	4	*	*	*	*
	Overall	0.205298	0.082644	2.48414	0.0065
5	0	0.223580	0.129099	1.73184	0.0417
	1	0.395556	0.129099	3.06396	0.0011
	2	0.487179	0.129099	3.77368	0.0001
	3	0.640000	0.129099	4.95742	0.0000
	4	*	*	*	*
	Overall	0.450309	0.080951	5.56276	0.0000
6	0	0.544880	0.129099	4.22062	0.0000
	1	0.383825	0.129099	2.97310	0.0015
	2	0.236111	0.129099	1.82891	0.0337
	3	0.474313	0.129099	3.67401	0.0001
	4	*	*	*	*
	Overall	0.368037	0.083863	4.38855	0.0000
7	0	0.515912	0.129099	3.99624	0.0000
	1	0.344639	0.129099	2.66956	0.0038
	2	0.487179	0.129099	3.77368	0.0001
	3	0.298246	0.129099	2.31020	0.0104
	4	*	*	*	*
	Overall	0.429162	0.086001	4.99022	0.0000
8	0	0.492022	0.129099	3.81119	0.0001
	1	0.199110	0.129099	1.54230	0.0615
	2	0.160000	0.129099	1.23935	0.1076
	3	0.373913	0.129099	2.89632	0.0019
	4	*	*	*	*
	Overall	0.298853	0.088287	3.38501	0.0004
9	0	0.818786	0.129099	6.34229	0.0000
	1	0.461279	0.129099	3.57306	0.0002
	2	0.408337	0.129099	3.16297	0.0008
	3	*	*	*	*
	4	*	*	*	*

	Overall	0.578300	0.094540	6.11699	0.0000
10	0	0.487179	0.129099	3.77368	0.0001
	1	0.259259	0.129099	2.00821	0.0223
	2	0.443414	0.129099	3.43467	0.0003
	3	-0.016949	0.129099	-0.13129	0.5522
	4	*	*	*	*
	Overall	0.376083	0.091696	4.10142	0.0000
11	0	0.196429	0.129099	1.52153	0.0641
	1	0.497608	0.129099	3.85445	0.0001
	2	0.420290	0.129099	3.25555	0.0006
	3	0.383210	0.129099	2.96833	0.0015
	4	*	*	*	*
	Overall	0.419558	0.086653	4.84181	0.0000
12	0	0.480519	0.129099	3.72209	0.0001
	1	0.177143	0.129099	1.37214	0.0850
	2	0.395556	0.129099	3.06396	0.0011
	3	0.482759	0.129099	3.73943	0.0001
	4	*	*	*	*
	Overall	0.337439	0.089314	3.77811	0.0001
13	0	0.482610	0.129099	3.73828	0.0001
	1	0.040000	0.129099	0.30984	0.3783
	2	0.134199	0.129099	1.03950	0.1493
	3	-0.043478	0.129099	-0.33678	0.6319
	4	*	*	*	*
	Overall	0.217501	0.088014	2.47120	0.0067
14	0	0.671683	0.129099	5.20283	0.0000
	1	0.282297	0.129099	2.18666	0.0144
	2	0.444444	0.129099	3.44265	0.0003
	3	0.649123	0.129099	5.02808	0.0000
	4	*	*	*	*
	Overall	0.470730	0.084018	5.60273	0.0000
15	0	0.607201	0.129099	4.70336	0.0000
	1	0.229782	0.129099	1.77988	0.0375
	2	0.413919	0.129099	3.20621	0.0007
	3	0.568656	0.129099	4.40479	0.0000
	4	-0.008403	0.129099	-0.06509	0.5259
	Overall	0.420066	0.078042	5.38253	0.0000
16	0	0.396118	0.129099	3.06832	0.0011
	1	0.657143	0.129099	5.09021	0.0000
	2	0.733333	0.129099	5.68038	0.0000
	3	0.544880	0.129099	4.22062	0.0000

	4	*	*	*	*
	Overall	0.627774	0.089100	7.04571	0.0000
17	0	0.400000	0.129099	3.09839	0.0010
	1	0.126984	0.129099	0.98361	0.1627
	2	0.261286	0.129099	2.02391	0.0215
	3	0.444211	0.129099	3.44084	0.0003
	4	-0.043478	0.129099	-0.33678	0.6319
	Overall	0.273261	0.071434	3.82538	0.0001
18	0	0.768935	0.129099	5.95614	0.0000
	1	0.537778	0.129099	4.16561	0.0000
	2	0.488636	0.129099	3.78496	0.0001
	3	-0.043478	0.129099	-0.33678	0.6319
	4	*	*	*	*
	Overall	0.562842	0.085206	6.60569	0.0000
19	0	0.629630	0.129099	4.87709	0.0000
	1	0.600000	0.129099	4.64758	0.0000
	2	0.608889	0.129099	4.71643	0.0000
	3	-0.025641	0.129099	-0.19861	0.5787
	4	*	*	*	*
	Overall	0.582463	0.097565	5.97003	0.0000
20	0	0.960925	0.129099	7.44329	0.0000
	1	0.794286	0.129099	6.15251	0.0000
	2	0.644444	0.129099	4.99185	0.0000
	3	-0.025641	0.129099	-0.19861	0.5787
	4	*	*	*	*
	Overall	0.775514	0.087971	8.81558	0.0000

* When no or all responses across trials equal the value, kappa cannot be computed.

Within Appraisers-FOV-Kendall's Coefficient of Concordance

Appraiser	Coef	Chi - Sq	DF	P
1	0.886574	104.616	59	0.0002
2	0.839623	99.075	59	0.0008
3	0.868249	102.453	59	0.0004
4	0.883694	104.276	59	0.0003
5	0.916165	108.107	59	0.0001
6	0.877513	103.547	59	0.0003
7	0.864153	101.970	59	0.0004
8	0.862749	101.804	59	0.0005
9	0.892294	105.291	59	0.0002
10	0.861293	101.633	59	0.0005

11	0.852848	100.636	59	0.0006
12	0.816617	96.361	59	0.0015
13	0.842556	99.422	59	0.0008
14	0.888026	104.787	59	0.0002
15	0.894432	105.543	59	0.0002
16	0.924000	109.032	59	0.0001
17	0.861574	101.666	59	0.0005
18	0.921911	108.786	59	0.0001
19	0.879592	103.792	59	0.0003
20	0.958645	113.120	59	0.0000

Each Appraiser vs Standard-FOV-Assessment Agreement

Appraiser	# Inspected	# Matched	Percent	95% CI
1	60	11	18.33	(9.52, 30.44)
2	60	22	36.67	(24.59, 50.10)
3	60	33	55.00	(41.61, 67.88)
4	60	21	35.00	(23.13, 48.40)
5	60	16	26.67	(16.07, 39.66)
6	60	8	13.33	(5.94, 24.59)
7	60	26	43.33	(30.59, 56.76)
8	60	28	46.67	(33.67, 60.00)
9	60	36	60.00	(46.54, 72.44)
10	60	34	56.67	(43.24, 69.41)
11	60	10	16.67	(8.29, 28.52)
12	60	26	43.33	(30.59, 56.76)
13	60	18	30.00	(18.85, 43.21)
14	60	27	45.00	(32.12, 58.39)
15	60	18	30.00	(18.85, 43.21)
16	60	32	53.33	(40.00, 66.33)
17	60	9	15.00	(7.10, 26.57)
18	60	31	51.67	(38.39, 64.77)
19	60	29	48.33	(35.23, 61.61)
20	60	51	85.00	(73.43, 92.90)

 # Matched: Appraiser's assessment across trials agrees with the known standard.

Each Appraiser vs Standard-FOV-Fleiss' Kappa Statistics

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)
1	0	0.259164	0.0912871	2.8390	0.0023
	1	0.101587	0.0912871	1.1128	0.1329
	2	0.020998	0.0912871	0.2300	0.4090
	3	-0.111492	0.0912871	-1.2213	0.8890
	4	*	*	*	*
	Overall	0.088655	0.0570713	1.5534	0.0602
2	0	0.324783	0.0912871	3.5578	0.0002
	1	-0.001503	0.0912871	-0.0165	0.5066
	2	0.289193	0.0912871	3.1679	0.0008
	3	-0.008403	0.0912871	-0.0921	0.5367
	4	*	*	*	*
	Overall	0.180742	0.0650729	2.7775	0.0027
3	0	0.636071	0.0912871	6.9678	0.0000
	1	0.366491	0.0912871	4.0147	0.0000
	2	0.393286	0.0912871	4.3082	0.0000
	3	-0.012676	0.0912871	-0.1389	0.5552
	4	*	*	*	*
	Overall	0.455081	0.0649915	7.0022	0.0000
4	0	0.557495	0.0912871	6.1071	0.0000
	1	0.331459	0.0912871	3.6310	0.0001
	2	0.409116	0.0912871	4.4816	0.0000
	3	-0.044742	0.0912871	-0.4901	0.6880
	4	*	*	*	*
	Overall	0.401942	0.0612838	6.5587	0.0000
5	0	0.287157	0.0912871	3.1457	0.0008
	1	0.252328	0.0912871	2.7641	0.0029
	2	0.048955	0.0912871	0.5363	0.2959
	3	-0.015019	0.0912871	-0.1645	0.5653
	4	*	*	*	*
	Overall	0.175583	0.0573522	3.0615	0.0011
6	0	0.097537	0.0912871	1.0685	0.1427
	1	-0.091949	0.0912871	-1.0073	0.8431
	2	-0.139601	0.0912871	-1.5293	0.9369
	3	0.039580	0.0912871	0.4336	0.3323
	4	*	*	*	*
	Overall	-0.047167	0.0566285	-0.8329	0.7976
7	0	0.579451	0.0912871	6.3476	0.0000
	1	0.363344	0.0912871	3.9802	0.0000

	2	0.443146	0.0912871	4.8544	0.0000
	3	-0.034560	0.0912871	-0.3786	0.6475
	4	*	*	*	*
	Overall	0.430471	0.0613178	7.0203	0.0000
8	0	0.628762	0.0912871	6.8877	0.0000
	1	0.461462	0.0912871	5.0551	0.0000
	2	0.490165	0.0912871	5.3695	0.0000
	3	-0.030214	0.0912871	-0.3310	0.6297
	4	*	*	*	*
	Overall	0.501550	0.0629802	7.9636	0.0000
9	0	0.757830	0.0912871	8.3016	0.0000
	1	0.487847	0.0912871	5.3441	0.0000
	2	0.535734	0.0912871	5.8687	0.0000
	3	-0.008403	0.0912871	-0.0921	0.5367
	4	*	*	*	*
	Overall	0.588513	0.0648870	9.0698	0.0000
10	0	0.738118	0.0912871	8.0857	0.0000
	1	0.553009	0.0912871	6.0579	0.0000
	2	0.658630	0.0912871	7.2149	0.0000
	3	-0.016949	0.0912871	-0.1857	0.5736
	4	*	*	*	*
	Overall	0.627112	0.0638693	9.8187	0.0000
11	0	0.180958	0.0912871	1.9823	0.0237
	1	0.052223	0.0912871	0.5721	0.2836
	2	-0.059558	0.0912871	-0.6524	0.7429
	3	0.006064	0.0912871	0.0664	0.4735
	4	*	*	*	*
	Overall	0.045515	0.0585518	0.7773	0.2185
12	0	0.527864	0.0912871	5.7825	0.0000
	1	0.348364	0.0912871	3.8161	0.0001
	2	0.475870	0.0912871	5.2129	0.0000
	3	-0.025641	0.0912871	-0.2809	0.6106
	4	*	*	*	*
	Overall	0.425770	0.0622731	6.8371	0.0000
13	0	0.405864	0.0912871	4.4460	0.0000
	1	0.148090	0.0912871	1.6222	0.0524
	2	0.360295	0.0912871	3.9468	0.0000
	3	0.144921	0.0912871	1.5875	0.0562
	4	*	*	*	*
	Overall	0.290850	0.0629543	4.6200	0.0000
14	0	0.631818	0.0912871	6.9212	0.0000

	1	0.337375	0.0912871	3.6958	0.0001
	2	0.458614	0.0912871	5.0239	0.0000
	3	-0.034483	0.0912871	-0.3777	0.6472
	4	*	*	*	*
	Overall	0.446096	0.0611338	7.2970	0.0000
15	0	0.528135	0.0912871	5.7854	0.0000
	1	0.209874	0.0912871	2.2991	0.0108
	2	0.161364	0.0912871	1.7677	0.0386
	3	-0.066688	0.0912871	-0.7305	0.7675
	4	*	*	*	*
	Overall	0.260697	0.0581135	4.4860	0.0000
16	0	0.350108	0.0912871	3.8352	0.0001
	1	0.432703	0.0912871	4.7400	0.0000
	2	0.555556	0.0912871	6.0858	0.0000
	3	-0.039136	0.0912871	-0.4287	0.6659
	4	*	*	*	*
	Overall	0.425673	0.0618179	6.8859	0.0000
17	0	0.497939	0.0912871	5.4547	0.0000
	1	0.101587	0.0912871	1.1128	0.1329
	2	-0.188472	0.0912871	-2.0646	0.9805
	3	-0.126785	0.0912871	-1.3889	0.9176
	4	*	*	*	*
	Overall	0.093851	0.0539202	1.7406	0.0409
18	0	0.576587	0.0912871	6.3162	0.0000
	1	0.459908	0.0912871	5.0380	0.0000
	2	0.516604	0.0912871	5.6591	0.0000
	3	-0.030517	0.0912871	-0.3343	0.6309
	4	*	*	*	*
	Overall	0.494774	0.0618377	8.0012	0.0000
19	0	0.291718	0.0912871	3.1956	0.0007
	1	0.292993	0.0912871	3.2096	0.0007
	2	0.593487	0.0912871	6.5013	0.0000
	3	0.320585	0.0912871	3.5118	0.0002
	4	*	*	*	*
	Overall	0.395246	0.0635063	6.2237	0.0000
20	0	0.980462	0.0912871	10.7404	0.0000
	1	0.897143	0.0912871	9.8277	0.0000
	2	0.821605	0.0912871	9.0002	0.0000
	3	0.237178	0.0912871	2.5982	0.0047
	4	*	*	*	*
	Overall	0.887172	0.0627750	14.1326	0.0000

* When all sample standards and responses of a trial(s) equal the value or none of them equals the value, kappa cannot be computed.

Each Appraiser vs Standard-FOV-Kendall's Correlation Coefficient

Appraiser	Coef	SE Coef	Z	P
1	0.717266	0.0626372	11.4466	0.0000
2	0.527993	0.0626372	8.4249	0.0000
3	0.626935	0.0626372	10.0045	0.0000
4	0.695824	0.0626372	11.1043	0.0000
5	0.631113	0.0626372	10.0712	0.0000
6	0.720265	0.0626372	11.4945	0.0000
7	0.713285	0.0626372	11.3831	0.0000
8	0.699444	0.0626372	11.1621	0.0000
9	0.756079	0.0626372	12.0663	0.0000
10	0.783602	0.0626372	12.5057	0.0000
11	0.642315	0.0626372	10.2500	0.0000
12	0.651107	0.0626372	10.3904	0.0000
13	0.585063	0.0626372	9.3360	0.0000
14	0.695364	0.0626372	11.0969	0.0000
15	0.694713	0.0626372	11.0866	0.0000
16	0.712437	0.0626372	11.3695	0.0000
17	0.745057	0.0626372	11.8903	0.0000
18	0.661562	0.0626372	10.5573	0.0000
19	0.682390	0.0626372	10.8898	0.0000
20	0.941209	0.0626372	15.0218	0.0000

Between Appraisers-FOV-Assessment Agreement

# Inspected	# Matched	Percent	95% CI
60	0	0.00	(0.00, 4.87)

 # Matched: All appraisers' assessments agree with each other.

Between Appraisers-FOV-Fleiss' Kappa Statistics

Response	Kappa	SE Kappa	Z	P(vs > 0)
0	0.441627	0.0046225	95.5386	0.0000
1	0.188189	0.0046225	40.7114	0.0000
2	0.231628	0.0046225	50.1089	0.0000
3	0.212889	0.0046225	46.0549	0.0000
4	0.006062	0.0046225	1.3114	0.0949

Overall	0.264610	0.0029366	90.1090	0.0000

Between Appraisers-FOV-Kendall's Coefficient of Concordance

Coef	Chi - Sq	DF	P
0.690417	1629.39	59	0.0000

All Appraisers vs Standard-FOV-Assessment Agreement

# Inspected	# Matched	Percent	95% CI
60	0	0.00	(0.00, 4.87)

Matched: All appraisers' assessments agree with the known standard.

All Appraisers vs Standard-FOV-Fleiss' Kappa Statistics

Response	Kappa	SE Kappa	Z	P(vs > 0)
0	0.491891	0.0204124	24.0976	0.0000
1	0.305217	0.0204124	14.9525	0.0000
2	0.342249	0.0204124	16.7667	0.0000
3	0.007131	0.0204124	0.3493	0.3634
4	*	*	*	*
Overall	0.358406	0.0136834	26.1929	0.0000

* When all sample standards and responses of a trial(s) equal the value or none of them
  equals the value, kappa cannot be computed.

All Appraisers vs Standard-FOV-Kendall's Correlation Coefficient

Coef	SE Coef	Z	P
0.694151	0.0140061	49.5596	0.0000
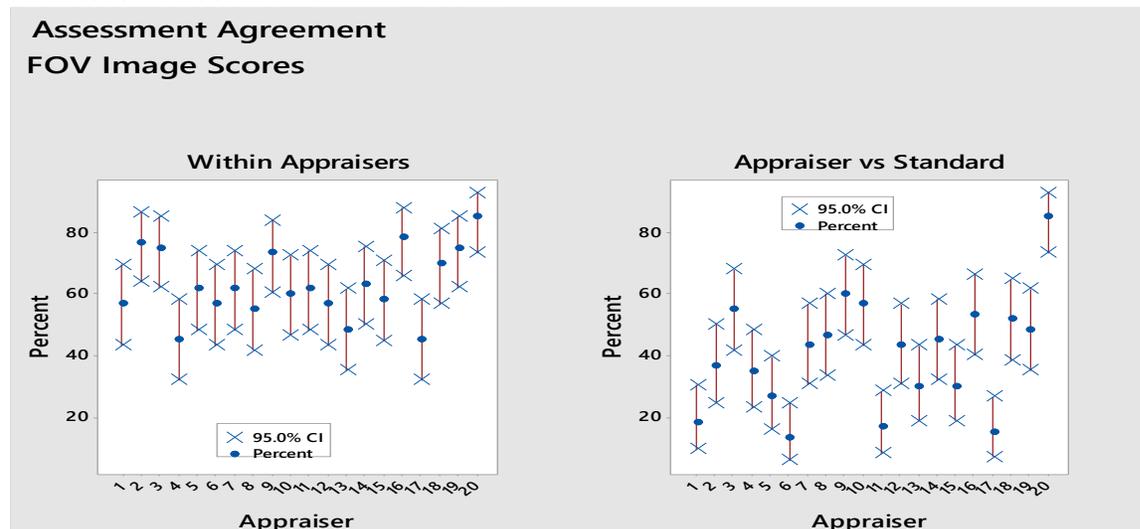
Attribute Agreement Analysis for ROI scores
Within Appraisers-ROI-Assessment Agreement

Appraiser	# Inspected	# Matched	Percent	95% CI
1	98	64	65.31	(55.02, 74.64)
2	98	66	67.35	(57.13, 76.48)
3	98	74	75.51	(65.79, 83.64)
4	98	66	67.35	(57.13, 76.48)
5	98	46	46.94	(36.78, 57.29)
6	98	76	77.55	(68.01, 85.36)
7	98	62	63.27	(52.93, 72.78)
8	98	74	75.51	(65.79, 83.64)
9	98	76	77.55	(68.01, 85.36)
10	98	61	62.24	(51.88, 71.84)
11	98	60	61.22	(50.85, 70.90)
12	98	56	57.14	(46.75, 67.10)
13	98	60	61.22	(50.85, 70.90)
14	98	77	78.57	(69.13, 86.22)
15	98	47	47.96	(37.76, 58.29)
16	98	56	57.14	(46.75, 67.10)
17	98	53	54.08	(43.71, 64.20)
18	98	80	81.63	(72.53, 88.74)
19	98	47	47.96	(37.76, 58.29)
20	98	86	87.76	(79.59, 93.51)

 # Matched: Appraiser agrees with him/herself across trials.

Within Appraisers-ROI-Fleiss' Kappa Statistics

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)
1	0	0.423287	0.101015	4.1903	0.0000
	1	0.393018	0.101015	3.8907	0.0000
	2	0.403718	0.101015	3.9966	0.0000
	3	0.759804	0.101015	7.5217	0.0000
	4	0.489583	0.101015	4.8466	0.0000
	Overall	0.511544	0.062195	8.2249	0.0000
2	0	0.572985	0.101015	5.6723	0.0000
	1	0.497940	0.101015	4.9294	0.0000
	2	0.560932	0.101015	5.5529	0.0000
	3	0.420309	0.101015	4.1608	0.0000
	4	*	*	*	*
	Overall	0.525890	0.064593	8.1415	0.0000

3	0	0.755102	0.101015	7.4751	0.0000
	1	0.619828	0.101015	6.1360	0.0000
	2	0.642149	0.101015	6.3570	0.0000
	3	0.467391	0.101015	4.6269	0.0000
	4	-0.005128	0.101015	-0.0508	0.5202
	Overall	0.645703	0.064466	10.0162	0.0000
4	0	0.696970	0.101015	6.8996	0.0000
	1	0.386541	0.101015	3.8266	0.0001
	2	0.490783	0.101015	4.8585	0.0000
	3	0.748718	0.101015	7.4119	0.0000
	4	-0.005128	0.101015	-0.0508	0.5202
	Overall	0.562225	0.058349	9.6355	0.0000
5	0	0.312281	0.101015	3.0914	0.0010
	1	0.217139	0.101015	2.1496	0.0158
	2	0.355263	0.101015	3.5169	0.0002
	3	0.248219	0.101015	2.4572	0.0070
	4	0.212851	0.101015	2.1071	0.0176
	Overall	0.273246	0.057920	4.7176	0.0000
6	0	0.367742	0.101015	3.6405	0.0001
	1	0.566888	0.101015	5.6119	0.0000
	2	0.615686	0.101015	6.0950	0.0000
	3	0.814764	0.101015	8.0658	0.0000
	4	*	*	*	*
	Overall	0.660820	0.067246	9.8269	0.0000
7	0	0.757576	0.101015	7.4996	0.0000
	1	0.393789	0.101015	3.8983	0.0000
	2	0.410625	0.101015	4.0650	0.0000
	3	0.385580	0.101015	3.8170	0.0001
	4	*	*	*	*
	Overall	0.484173	0.061880	7.8244	0.0000
8	0	0.733267	0.101015	7.2590	0.0000
	1	0.559220	0.101015	5.5360	0.0000
	2	0.677808	0.101015	6.7100	0.0000
	3	0.750000	0.101015	7.4246	0.0000
	4	-0.005128	0.101015	-0.0508	0.5202
	Overall	0.667397	0.059455	11.2252	0.0000
9	0	0.804098	0.101015	7.9602	0.0000
	1	0.664498	0.101015	6.5782	0.0000
	2	0.650000	0.101015	6.4347	0.0000
	3	0.423287	0.101015	4.1903	0.0000
	4	*	*	*	*

	Overall	0.680427	0.064268	10.5873	0.0000
10	0	0.622532	0.101015	6.1628	0.0000
	1	0.300357	0.101015	2.9734	0.0015
	2	0.448198	0.101015	4.4369	0.0000
	3	0.542971	0.101015	5.3751	0.0000
	4	0.711057	0.101015	7.0391	0.0000
	Overall	0.505557	0.054743	9.2351	0.0000
11	0	0.312281	0.101015	3.0914	0.0010
	1	0.555556	0.101015	5.4997	0.0000
	2	0.330468	0.101015	3.2715	0.0005
	3	0.414005	0.101015	4.0984	0.0000
	4	0.489583	0.101015	4.8466	0.0000
	Overall	0.430538	0.066238	6.4999	0.0000
12	0	0.673551	0.101015	6.6678	0.0000
	1	0.437663	0.101015	4.3326	0.0000
	2	0.393789	0.101015	3.8983	0.0000
	3	0.291566	0.101015	2.8864	0.0019
	4	0.350276	0.101015	3.4676	0.0003
	Overall	0.440951	0.054145	8.1440	0.0000
13	0	0.636364	0.101015	6.2997	0.0000
	1	0.306630	0.101015	3.0355	0.0012
	2	0.352321	0.101015	3.4878	0.0002
	3	0.184789	0.101015	1.8293	0.0337
	4	*	*	*	*
	Overall	0.426283	0.067204	6.3432	0.0000
14	0	0.699321	0.101015	6.9229	0.0000
	1	0.671438	0.101015	6.6469	0.0000
	2	0.711765	0.101015	7.0461	0.0000
	3	0.790598	0.101015	7.8265	0.0000
	4	-0.010309	0.101015	-0.1021	0.5406
	Overall	0.709671	0.059086	12.0108	0.0000
15	0	0.756522	0.101015	7.4892	0.0000
	1	0.319728	0.101015	3.1651	0.0008
	2	0.094867	0.101015	0.9391	0.1738
	3	0.282658	0.101015	2.7982	0.0026
	4	0.133170	0.101015	1.3183	0.0937
	Overall	0.323589	0.054439	5.9441	0.0000
16	0	-0.076923	0.101015	-0.7615	0.7768
	1	0.466667	0.101015	4.6198	0.0000
	2	0.368150	0.101015	3.6445	0.0001
	3	0.503378	0.101015	4.9832	0.0000

	4	0.650624	0.101015	6.4408	0.0000
	Overall	0.410104	0.059426	6.9011	0.0000
17	0	0.583333	0.101015	5.7747	0.0000
	1	0.475688	0.101015	4.7091	0.0000
	2	0.161926	0.101015	1.6030	0.0545
	3	0.443182	0.101015	4.3873	0.0000
	4	0.370281	0.101015	3.6656	0.0001
	Overall	0.409362	0.052235	7.8370	0.0000
18	0	0.846635	0.101015	8.3813	0.0000
	1	0.709232	0.101015	7.0210	0.0000
	2	0.729779	0.101015	7.2244	0.0000
	3	-0.026178	0.101015	-0.2591	0.6022
	4	*	*	*	*
	Overall	0.729552	0.068350	10.6738	0.0000
19	0	0.580180	0.101015	5.7435	0.0000
	1	0.057692	0.101015	0.5711	0.2840
	2	0.226783	0.101015	2.2450	0.0124
	3	0.145349	0.101015	1.4389	0.0751
	4	-0.010309	0.101015	-0.1021	0.5406
	Overall	0.263321	0.060606	4.3448	0.0000
20	0	0.946301	0.101015	9.3679	0.0000
	1	0.863946	0.101015	8.5526	0.0000
	2	0.764083	0.101015	7.5640	0.0000
	3	0.756522	0.101015	7.4892	0.0000
	4	*	*	*	*
	Overall	0.834634	0.059109	14.1203	0.0000

* When no or all responses across trials equal the value, kappa cannot be computed.

Within Appraisers-ROI-Kendall's Coefficient of Concordance

Appraiser	Coef	Chi - Sq	DF	P
1	0.903697	175.317	97	0.0000
2	0.915654	177.637	97	0.0000
3	0.928810	180.189	97	0.0000
4	0.937809	181.935	97	0.0000
5	0.761112	147.656	97	0.0007
6	0.942366	182.819	97	0.0000
7	0.921986	178.865	97	0.0000
8	0.945770	183.479	97	0.0000
9	0.947312	183.778	97	0.0000
10	0.921020	178.678	97	0.0000

11	0.862476	167.320	97	0.0000
12	0.926361	179.714	97	0.0000
13	0.916377	177.777	97	0.0000
14	0.955888	185.442	97	0.0000
15	0.900299	174.658	97	0.0000
16	0.926284	179.699	97	0.0000
17	0.905111	175.591	97	0.0000
18	0.946962	183.711	97	0.0000
19	0.859083	166.662	97	0.0000
20	0.973913	188.939	97	0.0000

Each Appraiser vs Standard-ROI-Assessment Agreement

Appraiser	# Inspected	# Matched	Percent	95% CI
1	98	24	24.49	(16.36, 34.21)
2	98	47	47.96	(37.76, 58.29)
3	98	40	40.82	(30.99, 51.21)
4	98	53	54.08	(43.71, 64.20)
5	98	10	10.20	(5.00, 17.97)
6	98	24	24.49	(16.36, 34.21)
7	98	48	48.98	(38.74, 59.28)
8	98	61	62.24	(51.88, 71.84)
9	98	52	53.06	(42.71, 63.22)
10	98	39	39.80	(30.04, 50.18)
11	98	32	32.65	(23.52, 42.87)
12	98	40	40.82	(30.99, 51.21)
13	98	30	30.61	(21.70, 40.74)
14	98	50	51.02	(40.72, 61.26)
15	98	35	35.71	(26.29, 46.03)
16	98	24	24.49	(16.36, 34.21)
17	98	22	22.45	(14.64, 31.99)
18	98	40	40.82	(30.99, 51.21)
19	98	31	31.63	(22.61, 41.80)
20	98	86	87.76	(79.59, 93.51)

 # Matched: Appraiser's assessment across trials agrees with the known standard.

Each Appraiser vs Standard-ROI-Fleiss' Kappa Statistics

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)
1	0	0.073817	0.0714286	1.0334	0.1507
	1	0.001158	0.0714286	0.0162	0.4935
	2	0.117757	0.0714286	1.6486	0.0496
	3	0.310089	0.0714286	4.3413	0.0000

	4	-0.010309	0.0714286	-0.1443	0.5574
	Overall	0.129904	0.0414733	3.1322	0.0009
2	0	0.484241	0.0714286	6.7794	0.0000
	1	0.442562	0.0714286	6.1959	0.0000
	2	0.582110	0.0714286	8.1495	0.0000
	3	0.497836	0.0714286	6.9697	0.0000
	4	*	*	*	*
	Overall	0.506551	0.0430397	11.7694	0.0000
3	0	0.473393	0.0714286	6.6275	0.0000
	1	0.217956	0.0714286	3.0514	0.0011
	2	0.354670	0.0714286	4.9654	0.0000
	3	0.430233	0.0714286	6.0233	0.0000
	4	*	*	*	*
	Overall	0.353774	0.0428592	8.2543	0.0000
4	0	0.601726	0.0714286	8.4242	0.0000
	1	0.557752	0.0714286	7.8085	0.0000
	2	0.504333	0.0714286	7.0607	0.0000
	3	0.641400	0.0714286	8.9796	0.0000
	4	*	*	*	*
	Overall	0.568036	0.0414649	13.6992	0.0000
5	0	0.081084	0.0714286	1.1352	0.1282
	1	0.061229	0.0714286	0.8572	0.1957
	2	-0.037829	0.0714286	-0.5296	0.7018
	3	0.166467	0.0714286	2.3305	0.0099
	4	-0.083402	0.0714286	-1.1676	0.8785
	Overall	0.045294	0.0386029	1.1733	0.1203
6	0	0.196304	0.0714286	2.7483	0.0030
	1	0.099819	0.0714286	1.3975	0.0811
	2	-0.179291	0.0714286	-2.5101	0.9940
	3	0.307145	0.0714286	4.3000	0.0000
	4	*	*	*	*
	Overall	0.099464	0.0424377	2.3438	0.0095
7	0	0.603387	0.0714286	8.4474	0.0000
	1	0.403363	0.0714286	5.6471	0.0000
	2	0.601680	0.0714286	8.4235	0.0000
	3	0.439256	0.0714286	6.1496	0.0000
	4	*	*	*	*
	Overall	0.523672	0.0426306	12.2840	0.0000
8	0	0.759240	0.0714286	10.6294	0.0000
	1	0.487073	0.0714286	6.8190	0.0000
	2	0.671106	0.0714286	9.3955	0.0000

	3	0.736613	0.0714286	10.3126	0.0000
	4	*	*	*	*
	Overall	0.655535	0.0417967	15.6839	0.0000
9	0	0.745206	0.0714286	10.4329	0.0000
	1	0.411112	0.0714286	5.7556	0.0000
	2	0.387806	0.0714286	5.4293	0.0000
	3	0.433333	0.0714286	6.0667	0.0000
	4	*	*	*	*
	Overall	0.501543	0.0429373	11.6808	0.0000
10	0	0.524804	0.0714286	7.3473	0.0000
	1	0.183518	0.0714286	2.5692	0.0051
	2	0.539906	0.0714286	7.5587	0.0000
	3	0.503800	0.0714286	7.0532	0.0000
	4	-0.028878	0.0714286	-0.4043	0.6570
	Overall	0.426169	0.0399352	10.6715	0.0000
11	0	0.040613	0.0714286	0.5686	0.2848
	1	0.253986	0.0714286	3.5558	0.0002
	2	0.348136	0.0714286	4.8739	0.0000
	3	0.530734	0.0714286	7.4303	0.0000
	4	-0.010336	0.0714286	-0.1447	0.5575
	Overall	0.304137	0.0422943	7.1910	0.0000
12	0	0.497092	0.0714286	6.9593	0.0000
	1	0.422200	0.0714286	5.9108	0.0000
	2	0.550370	0.0714286	7.7052	0.0000
	3	0.387118	0.0714286	5.4197	0.0000
	4	-0.040146	0.0714286	-0.5620	0.7130
	Overall	0.448164	0.0397595	11.2719	0.0000
13	0	0.644779	0.0714286	9.0269	0.0000
	1	0.063632	0.0714286	0.8909	0.1865
	2	0.068395	0.0714286	0.9575	0.1691
	3	0.302382	0.0714286	4.2333	0.0000
	4	*	*	*	*
	Overall	0.275812	0.0433727	6.3591	0.0000
14	0	0.596063	0.0714286	8.3449	0.0000
	1	0.466394	0.0714286	6.5295	0.0000
	2	0.399384	0.0714286	5.5914	0.0000
	3	0.472488	0.0714286	6.6148	0.0000
	4	*	*	*	*
	Overall	0.473154	0.0412895	11.4594	0.0000
15	0	0.594384	0.0714286	8.3214	0.0000
	1	0.446187	0.0714286	6.2466	0.0000

	2	0.264722	0.0714286	3.7061	0.0001
	3	0.457052	0.0714286	6.3987	0.0000
	4	-0.029049	0.0714286	-0.4067	0.6579
	Overall	0.414780	0.0397971	10.4224	0.0000
16	0	0.122801	0.0714286	1.7192	0.0428
	1	0.208033	0.0714286	2.9125	0.0018
	2	0.294743	0.0714286	4.1264	0.0000
	3	0.403134	0.0714286	5.6439	0.0000
	4	-0.023561	0.0714286	-0.3299	0.6292
	Overall	0.259868	0.0408615	6.3597	0.0000
17	0	0.583024	0.0714286	8.1623	0.0000
	1	0.356105	0.0714286	4.9855	0.0000
	2	-0.089526	0.0714286	-1.2534	0.8950
	3	0.110750	0.0714286	1.5505	0.0605
	4	-0.084064	0.0714286	-1.1769	0.8804
	Overall	0.200401	0.0377583	5.3075	0.0000
18	0	0.424303	0.0714286	5.9402	0.0000
	1	0.167548	0.0714286	2.3457	0.0095
	2	0.373465	0.0714286	5.2285	0.0000
	3	0.079999	0.0714286	1.1200	0.1314
	4	*	*	*	*
	Overall	0.290835	0.0436658	6.6605	0.0000
19	0	0.406319	0.0714286	5.6885	0.0000
	1	0.208569	0.0714286	2.9200	0.0018
	2	0.393172	0.0714286	5.5044	0.0000
	3	0.370281	0.0714286	5.1839	0.0000
	4	*	*	*	*
	Overall	0.345561	0.0424131	8.1475	0.0000
20	0	0.973151	0.0714286	13.6241	0.0000
	1	0.932696	0.0714286	13.0577	0.0000
	2	0.880996	0.0714286	12.3339	0.0000
	3	0.879771	0.0714286	12.3168	0.0000
	4	*	*	*	*
	Overall	0.917421	0.0417442	21.9772	0.0000

*When all sample standards and responses of a trial(s) equal the value or none of them
  equals the value, kappa cannot be computed.*

Each Appraiser vs Standard-ROI-Kendall's Correlation Coefficient

Appraiser	Coef	SE Coef	Z	P
1	0.681570	0.0484705	14.0594	0.0000
2	0.783093	0.0484705	16.1539	0.0000
3	0.737160	0.0484705	15.2062	0.0000

4	0.751928	0.0484705	15.5109	0.0000
5	0.553036	0.0484705	11.4076	0.0000
6	0.723734	0.0484705	14.9292	0.0000
7	0.746464	0.0484705	15.3982	0.0000
8	0.820150	0.0484705	16.9184	0.0000
9	0.811120	0.0484705	16.7321	0.0000
10	0.743011	0.0484705	15.3270	0.0000
11	0.630624	0.0484705	13.0083	0.0000
12	0.752875	0.0484705	15.5305	0.0000
13	0.820660	0.0484705	16.9289	0.0000
14	0.786062	0.0484705	16.2151	0.0000
15	0.693819	0.0484705	14.3121	0.0000
16	0.737840	0.0484705	15.2203	0.0000
17	0.752437	0.0484705	15.5214	0.0000
18	0.666544	0.0484705	13.7494	0.0000
19	0.649328	0.0484705	13.3942	0.0000
20	0.958217	0.0484705	19.7669	0.0000

Between Appraisers-ROI-Assessment Agreement

# Inspected	# Matched	Percent	95% CI
98	0	0.00	(0.00, 3.01)

Matched: All appraisers' assessments agree with each other.

Between Appraisers-ROI-Fleiss' Kappa Statistics

Response	Kappa	SE Kappa	Z	P(vs > 0)
0	0.464350	0.0036169	128.382	0.0000
1	0.222237	0.0036169	61.444	0.0000
2	0.196105	0.0036169	54.219	0.0000
3	0.284331	0.0036169	78.611	0.0000
4	0.077479	0.0036169	21.421	0.0000
Overall	0.271364	0.0020272	133.860	0.0000

Between Appraisers-ROI-Kendall's Coefficient of Concordance

Coef	Chi - Sq	DF	P
0.774279	3004.20	97	0.0000

All Appraisers vs Standard-ROI-Assessment Agreement

# Inspected	# Matched	Percent	95% CI
98	0	0.00	(0.00, 3.01)

Matched: All appraisers' assessments agree with the known standard.

All Appraisers vs Standard-ROI-Fleiss' Kappa Statistics

Response	Kappa	SE Kappa	Z	P(vs > 0)
0	0.471287	0.0159719	29.5072	0.0000
1	0.319545	0.0159719	20.0067	0.0000
2	0.351305	0.0159719	21.9952	0.0000
3	0.422994	0.0159719	26.4836	0.0000
4	*	*	*	*
Overall	0.387004	0.0092878	41.6678	0.0000

* When all sample standards and responses of a trial(s) equal the value or none of them
   equals the value, kappa cannot be computed.

All Appraisers vs Standard-ROI-Kendall's Correlation Coefficient

Coef	SE Coef	Z	P
0.739984	0.0108383	68.2741	0.0000
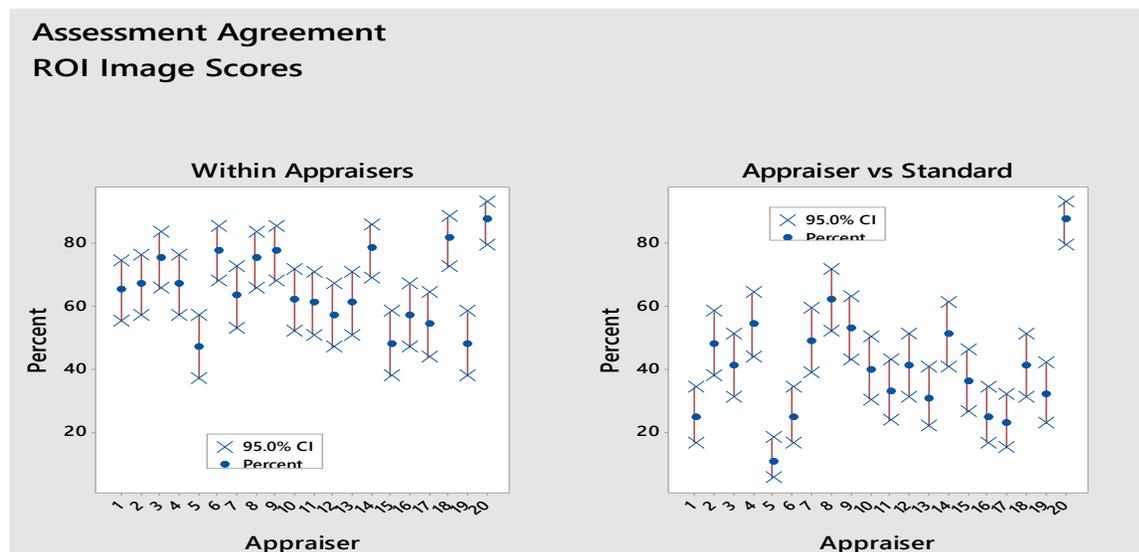
Appendix C

Time sequence plots for EI and VA

P06 CHEST EI values and VA scores

P07 EAR EI values and VA scores

P08 CHEEK EI values and VA scores

P09 FOREHEAD EI values and VA scores

P10 LEG EI values and VA scores
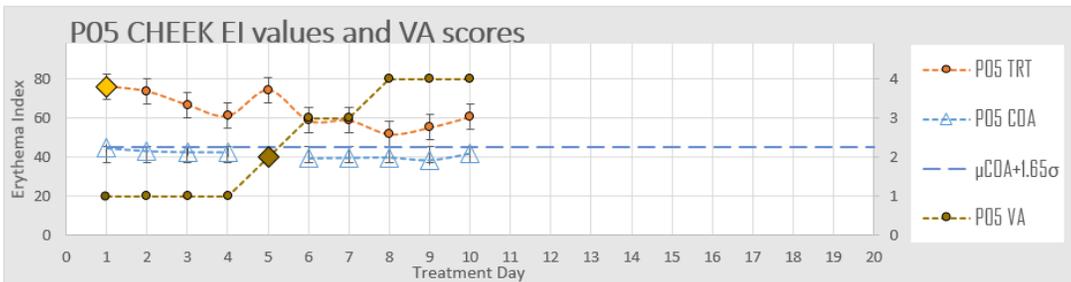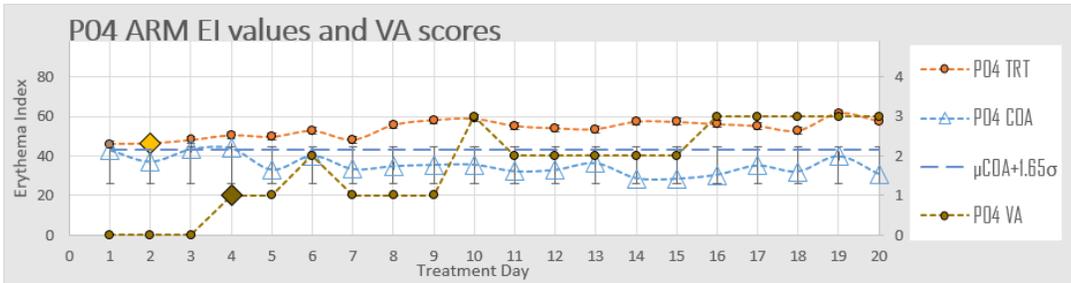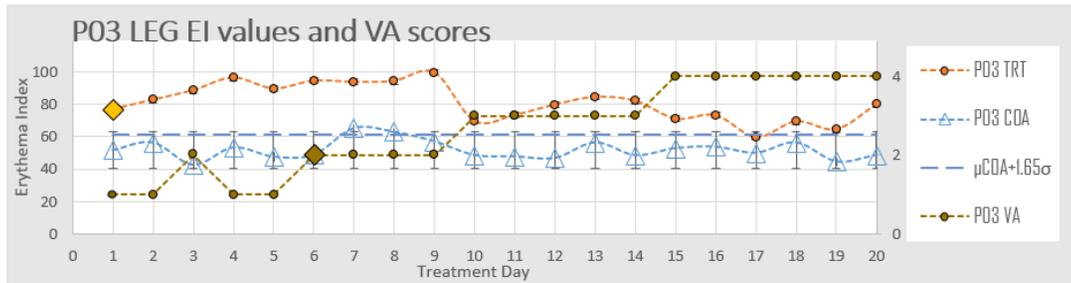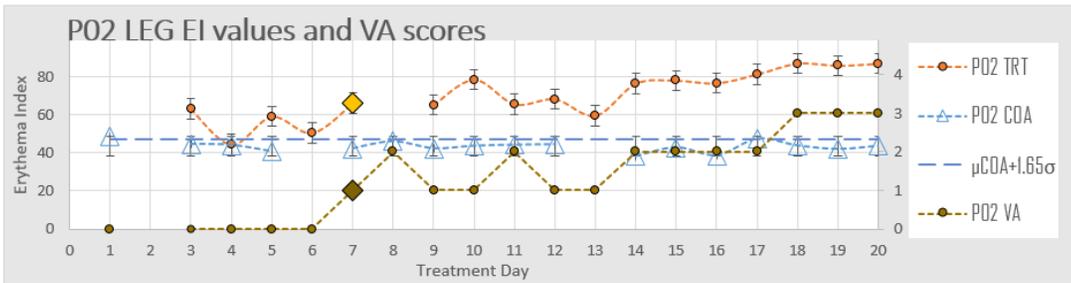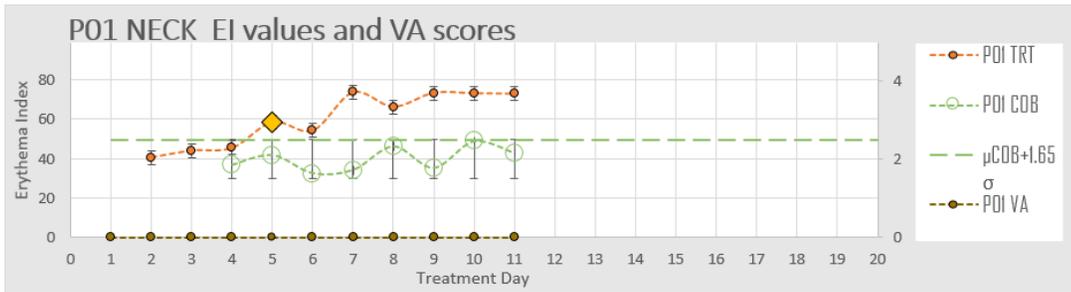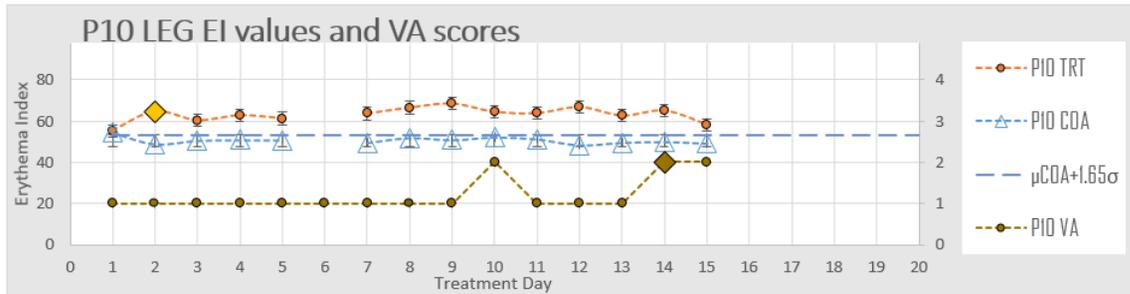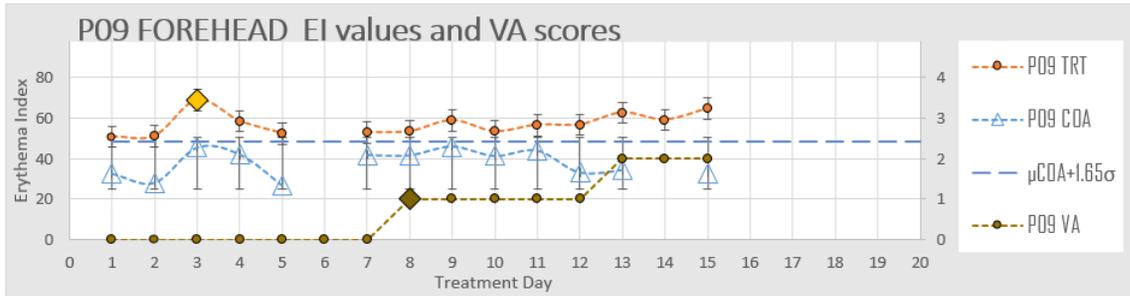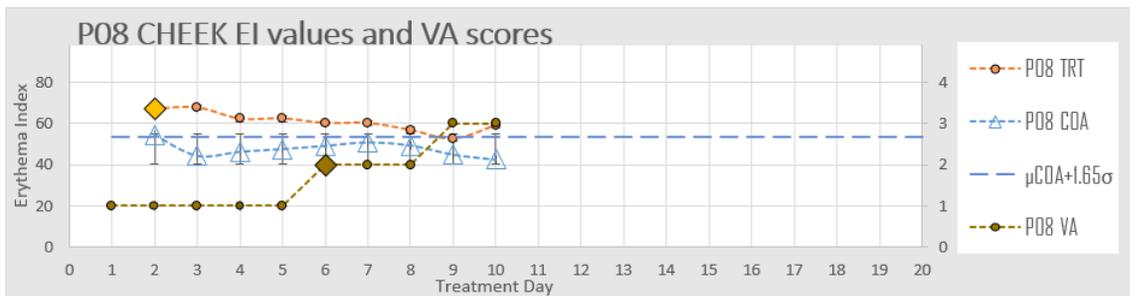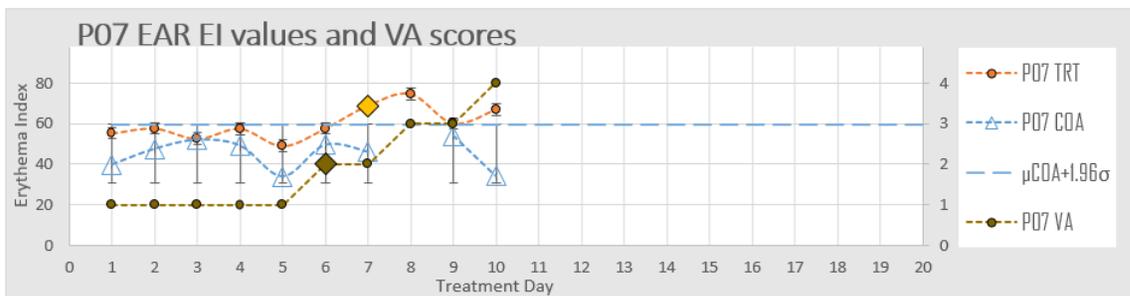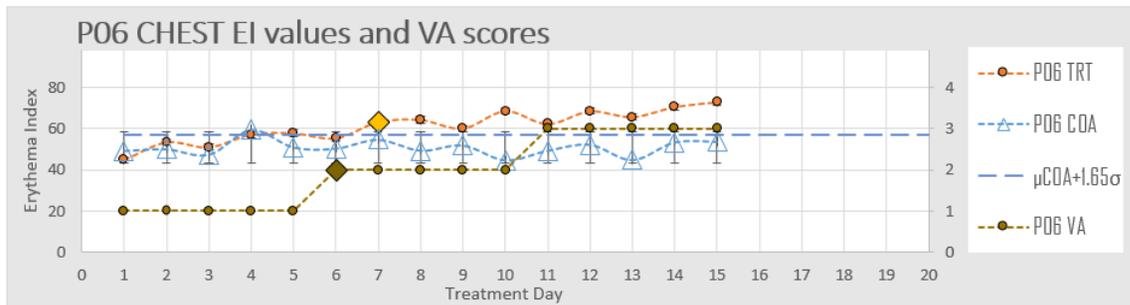
Scatterplots for EI and VA (Minitab 17)

Test for monotonic relationship between your two ranked variables.
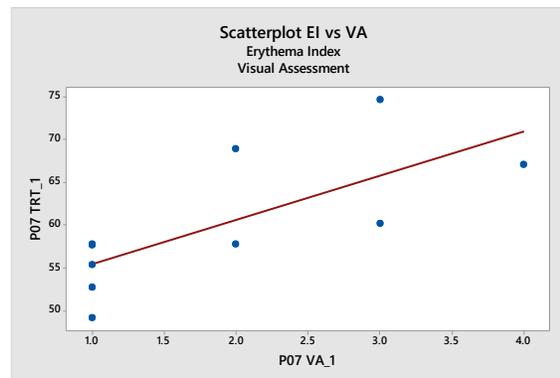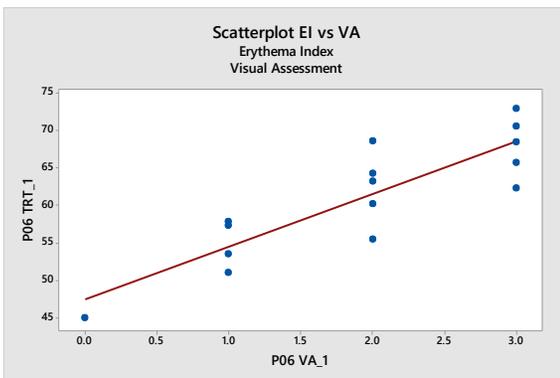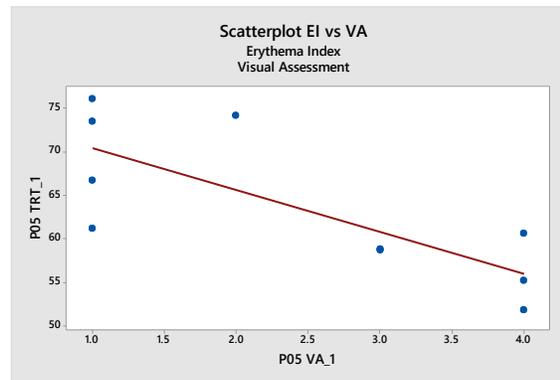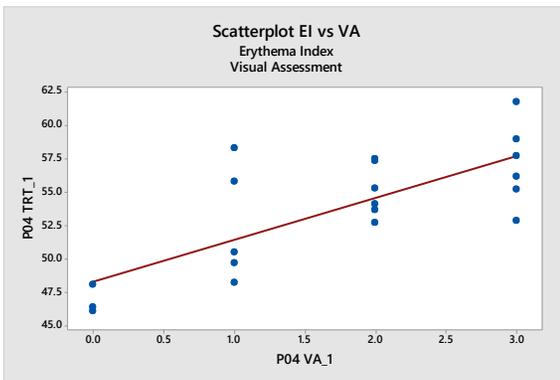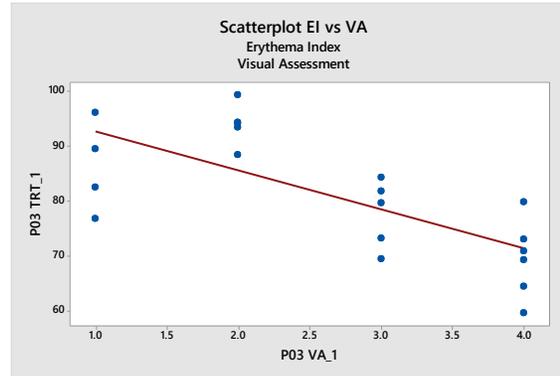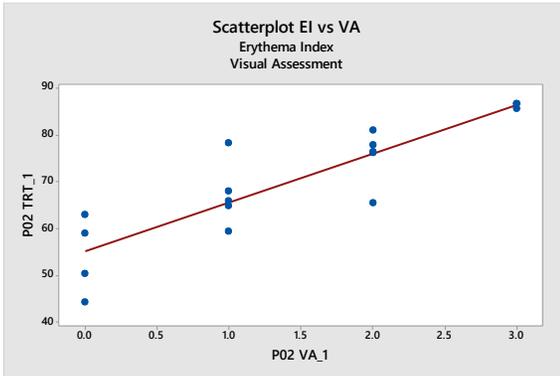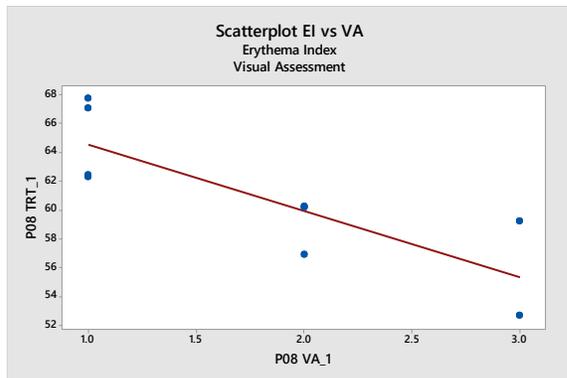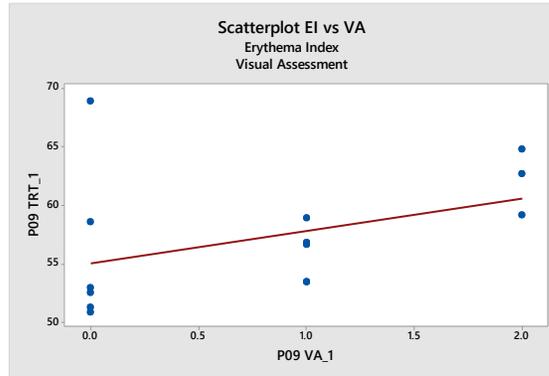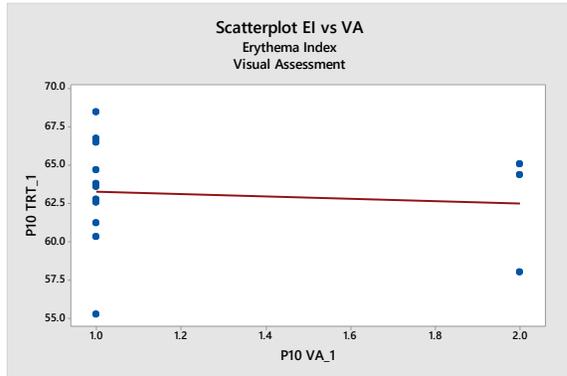
Spearman Rho Calculations (Minitab 17)

```
Spearman rho for P02 VA_1 and P02 TRT_1 = 0.879, P-Value = 0.000

Spearman rho for P03 VA_1 and P03 TRT_1 = -0.737, P-Value = 0.000

Spearman rho for P04 VA_1 and P04 TRT_1 = 0.677, P-Value = 0.001

Spearman rho for P05 VA_1 and P05 TRT_1 = -0.795, P-Value = 0.006

Spearman rho for P06 VA_1 and P06 TRT_1 = 0.827, P-Value = 0.000

Spearman rho for P07 VA_1 and P07 TRT_1 = 0.840, P-Value = 0.002

Spearman rho for P08 VA_1 and P08 TRT_1 = -0.891, P-Value = 0.001

Spearman rho for P09 VA_1 and P09 TRT_1 = 0.565, P-Value = 0.035

Spearman rho for P10 VA_1 and P10 TRT_1 = -0.022, P-Value = 0.942
```