12-2012

Cross Practitioner Inter-rater Variability: Grading of Adverse Skin Reactions after Radiation Therapy

Carleen A. Evans
St. John Fisher College

Cross Practitioner Inter-rater Variability: Grading of Adverse Skin Reactions after Radiation Therapy

Abstract

This study estimated the inter-rater variability for adverse event grades for skin reactions related to radiation therapy among nurses and physicians. Currently, there appears to be a gap in literature on the reliability of severity grades across grading systems. Physicians and nurses conducted a retrospective review of photographs of skin reactions of patients with a diagnosis of breast cancer who received cancer therapy. Rater participants rated the condition(s) seen in each photograph on a scale of Grade 0 to Grade 5. This inter-rater reliability study used quantitative methods to estimate initer-rater agreement and inter-rater reliability of severity grades by comparing physicians' and nurses' scores of adverse events due to skin reactions after radiation therapy. Fleiss' Kappa and intraclass correlation (ICC) was used to estimate inter-rater agreement and inter-rater reliability across practitioners respectively. Fair agreement was seen (k = >0.2) among nurses and physicians. The findings also showed ICC values above 0.9 across all practitioner groups. The study underscores the value of objectivity in the use of adverse events severity scales.

Document Type

Dissertation

Degree Name

Doctor of Education (EdD)

Department

Executive Leadership

First Supervisor

Richard E. Maurer

Second Supervisor

Ajay Kapur

Subject Categories

Education

Cross Practitioner Inter-rater Variability:

Grading of Adverse Skin Reactions after Radiation Therapy

By

Carleen A. Evans

Submitted in partial fulfillment

of the requirement for the degree

Ed.D. in Executive Leadership

Supervised by

Richard E. Maurer, Ph.D

Committee Member

Ajay Kapur, Ph.D.

Ralph C. Wilson, Jr. School of Education

St. John Fisher College

December 2012

Dedication

I would like to take this opportunity to thank to everyone who supported me on this journey. In particular, I would like to express my immense gratitude to Dr. Richard Maurer, my committee chair, for his invaluable and unwavering support throughout the entire process and especially during the final weeks. Without his guidance this dissertation would not have been possible. I am profoundly indebted to Dr. Ajay Kapur of the Department of Radiation Medicine, North Shore Long Island Jewish Hospital, for generously giving his time and intellectual support every step of the way. As a member of my committee, Dr. Kapur opened doors that would have otherwise remained closed to me. I consider it an honor to work with him. I would also like to thank Danielle Pearson for her patience, advice, and for always being present even though she is miles away. Thank you to Dr. Jerry Willis, my advisor, for listening and offering his qualitative views. In addition, thank you to Dr. Claudia Edwards and Dr. Michael Robinson for their continuous support.

Over the past two years, I received encouragement and support from a number of people: Dr. Isaac Dapkins who encouraged me from the very start; Irma Rivera who has been a colleague and friend; Melissa Janjic, for always being there when I needed help; my friend, Wilford Beckles for his pearls of wisdom throughout this journey; and to my dissertation teammates for their support as we all worked toward our goals.

Finally to my children, Ryann and Ethan, for making every grueling day feel like Christmas! Your undying love and laughter have sustained me throughout this journey and will continue to do so for the rest of my life.

Biographical Sketch

Carleen Evans is currently the Manager of Nursing Quality at The Mount Sinai Hospital. Ms. Evans earned a Bachelor of Science in 2003 from York College (CUNY) and later attended Pace University from 2004 to 2005 where she received a Bachelor of Science degree in Nursing. She attended Bellevue University from 2007 to 2008 and graduated with a Master’s in Healthcare Administration. Ms. Evans came to St. John Fisher College in the summer of 2010 and began doctoral studies in the Ed.D. Program in Executive Leadership. She pursued her research in the reliability of adverse events grade for skin conditions after cancer therapy under the direction of Dr. Ajay Kapur and Dr. Richard Maurer and received the Ed.D. degree in 2013.

Abstract

This study estimated the inter-rater variability for adverse event grades for skin reactions related to radiation therapy among nurses and physicians. Currently, there appears to be a gap in literature on the reliability of severity grades across grading systems. Physicians and nurses conducted a retrospective review of photographs of skin reactions of patients with a diagnosis of breast cancer who received cancer therapy. Rater participants rated the condition(s) seen in each photograph on a scale of Grade 0 to Grade 5. This inter-rater reliability study used quantitative methods to estimate initer-rater agreement and inter-rater reliability of severity grades by comparing physicians' and nurses' scores of adverse events due to skin reactions after radiation therapy. Fleiss' Kappa and intraclass correlation (ICC) was used to estimate inter-rater agreement and inter-rater reliability across practitioners respectively. Fair agreement was seen (k = $\geq$0.2) among nurses and physicians. The findings also showed ICC values above 0.9 across all practitioner groups. The study underscores the value of objectivity in the use of adverse events severity scales.

Table of Contents

List of Tables

Item **Title** **Page**

Chapter 1: Introduction

Introduction

This study tested the reliability of severity grades for skin reactions after cancer therapy. Oncology nurses and physicians use scales such the Common Toxicity Criteria for Adverse Events (CTCAE) and the Radiation Therapy Oncology Group (RTOG) in the course of their care to evaluate adverse events related to radiotherapy and other treatment modalities. In some instances, practitioners modified these scales to meet specific needs of their treatment populations. Chapter 1 provides a brief description of the purpose of such scales and includes a background on these commonly used scales. This chapter also demonstrates that there is limited research on the reliability of grading (i.e., scoring). In addition, Chapter 1 discusses the importance of this study and presents the research questions related to the study. Lastly, this chapter presents a brief overview of the theoretical framework that guides the study.

Modern cancer therapies such as surgery, chemotherapy, radiation therapy, immunotherapy, bone marrow and peripheral blood and stem cell transplants, and targeted cancer therapies create undue burden on the human body (NCI, n.d.a). Cancer patients may experience acute and long-term adverse effects related to therapy from any of the various modalities. Toxicity or adverse events grading systems were established to monitor the severity of reactions resulting from these therapies. The main grading systems in use are:

- Common Terminology Criteria for Adverse Events (CTCAE)
- Radiation Therapy Oncology Group and European Organization for Research and Treatment of Cancer (RTOG/EORTC) scale
- Late Effects of Normal Tissue working group SOMA scale (LENT SOMA). SOMA represents the Subjective, Objective, Medical management, and Analytical evaluation of injury findings for adverse events (Van der Laan et al., 2008)

The severity weight ranges from no effect to death and may vary depending on which scales is being used.

Common Toxicity Criteria for Adverse Events scale. The scale was developed by the National Cancer Institute (NCI) and has been used by oncology practitioners in clinical trials worldwide in their assessments of adverse events related to cancer therapy. The CTCAE is used to evaluate "new cancer therapies, treatment modalities, and supportive measures and to standardize reporting of AEs across groups and modalities" (NCI, n.d. a). The first scale, Common Toxicity Criteria (CTC) was developed in 1982 for the assessment of the acute or short term adverse effects of chemotherapy and was updated in 1997 to form CTC, version 2.0 (Trotti et al., 2003). Nevertheless, CTC version 2.0 did not meet the needs of the practitioners. Considerations for the effects of radiotherapy were not included in the original CTC and CTC version 2.0. In 2003, the CTC version 2.0 was again updated and renamed the Common Terminology Criteria for Adverse Events. Late effects of therapy, and surgical and pediatric criteria were included in this version. The effects of radiotherapy were also systematically included in this and subsequent versions.

Severity grading systems arose from the need to develop a common language in the field of clinical trials and to enhance the reporting and comparison of results from the international community (World Health Organization, 1979). As a result of the explosion of cancer research and literature on cancer treatment, it became necessary to define acceptable standards for evaluating the data (World Health Organization, 1979). In 1977 and 1979, the European Organization for Research on the Treatment of Cancer (EORTC), The National Cancer Institute from the United States, the International Union Against Cancer, and Council for Mutual Economic Assistance, and members of other organizations met in Turin and Brussels, respectively, to determine the nature and details of the criteria (World Health Organization, 1979). The main result of the World Health Organization meetings was the development of standard criteria for the evaluation of cancer therapy. These criteria included data related to individual patient, data related to the tumor, and data from laboratory and radiological studies, nutritional status, socioeconomic status, and certain specific behavior that may influence the outcome of antitumor therapy (World Health Organization, 1979). These criteria were necessary to collect meaningful and comparative data that would be used in the assignment of severity grades based on patient report and symptom presentation. The idea was to gather detailed information on each modality used in therapy.

Healthcare providers are required to assign the most appropriate grade based on their interpretations of the adverse events being reported by the patient and observed symptoms in the appropriate body systems category. The categories are based on each anatomical or pathophysiological system such as endocrine, allergy/immunology, dermatology, general cardiac, and pain. Grading of adverse events is obtained from patients' reports to caregivers of the symptoms of their adverse event(s) and physical assessments by clinicians. Adverse events can be classified as acute and chronic, or late (World Health Organization, 1979). Acute is classified as occurring as early as one day after treatment while chronic or late effects occurs beyond 90 days after cancer therapy. Appropriate reporting of toxic effects allows for better comparison of toxicity and uniform treatment cases within a therapeutic program (World Health Organization, 1979). The development of the CTC v1.0 began the formidable grading task of significant toxic events. Since its development, the CTC has undergone several revisions and is currently in its fourth version, CTCAE version 4.0. The grades within the scale relate to the severity of the adverse reactions and are outlined in the NCI 2012 guidelines. Table 1.1 shows the criteria outlined in the NCI 2012 guidelines. Appendix A is an excerpt of one category, the Skin / Dermatology Disorders, from the CTCAE scale.

Table 1.1

Guidelines for CTCAE Severity Grades

Grades	Criteria
0	No adverse events or within normal limits
1	Mild; asymptomatic or mild symptoms; clinical or diagnostic observations only; intervention not indicated
2	Moderate; minimal, local, or noninvasive intervention such as packing or cautery indicated, limiting age-appropriate instrumental activities of daily living (ADL)
3	Severe; or medically significant but not immediately life-threatening; hospitalization or prolongation of hospitalization indicated; disabling; limiting self-care ADL
4	Life-threatening consequences; urgent interventions indicated
5	Death related to adverse event

Radiation Therapy Oncology Group (RTOG) scale. The RTOG scale is one of several well-known severity grading systems used to evaluate cancer therapy induced adverse events. Wells and McBride (2003) argued that the RTOG is presumably the most popular grading system used in practice and research. Table 1.2 outlines the criteria used to grade adverse events from radiation therapy and is a replication of the actual scale (see Appendix B).

Table 1.2

RTOG Acute Radiation Morbidity Scoring Criteria for Skin Reactions

Grades	Descriptions
0	No change over baseline
1	Follicular, faint or dull erythema/epilation/dry desquamation/ decreased sweating
2	Tender or bright erythema, patchy moist desquamation/ moderate erythema
3	Confluent, moist desquamation other than skin folds, pitting edema
4	Ulceration, hemorrhage, necrosis

The developers of the scale cautioned that any toxicity that caused death must be scored as Grade 5. While the scale distinguishes between the faint, dull, or bright erythema and between patchy and confluent moist desquamation, dry desquamation and faint erythema are given equal weights and may not reflect the actual experience of the adverse event from the perspective of the patient (Wells & McBride, 2003).

Inter-rater reliability of grading systems. Inter-rater reliability assessments across grading systems often show significant variability. Wells and McBride (2003) purported that the RTOG grading system in its original form does not always capture the essence of the skin reaction, which causes researchers and practitioners to modify the scale. Langendijk et al. (2008) stated that the RTOG scales were not validated for inter-rater reliability. However, Rosewall et al. (2009) evaluated the inter-rater reliability of RTOG grades for patients undergoing prostate radiotherapy and found good agreement among observer groups. In this study, kappa was 0.756. Rosewall and colleagues (2009) reported that the RTOG grading system has shown high inter-rater reliability compared to other grading systems.

In 1998, Postma et al. (1998) evaluated inter-rater agreement across toxicity grading systems with a primary focus on the differences in the peripheral neurotoxicity sections and how raters interpreted these scales. Postma et al. (1998) examined four grading systems – World Health Organization (WHO), Eastern Cooperative Oncology Group (ECOG), Ajani, and the National Cancer Institute of Canada – Common Toxixity Criteria (NCIC-CTC). Percentage agreement and intraclass correlations (ICC) were used to estimate reliability of grades. The results of this study are presented Table 1.3.

Table 1.3

Inter-rater Reliability across Grading Systems

Grading System	Inter-rater Agreement	ICC (Grade 0 – 4)
WHO	83.3% (31/37	0.55
ECOG	75.6% (28/37)	0.75
CTC	45.9% (17/37	0.58
Ajani	567% (21/37)	0.37

Note. Adapted from Postma et al. (1998)

Postma et al. (1998) found significant variability in scores across scales, which demonstrated the differences in how scales are interpreted by practitioners. Severity grades for the various scales are not interchangeable. Van der Laan et al. (2008) cited the differences across scales as an important factor in the interpretation, which underscores the value of specifying which scale was used in the evaluation of severity of adverse events. For example, a Grade 2 score for rectal toxicity is described as five or more stools per day while a grade 2 on the CTCAE v. 3.0 is defined as an increase in the stool frequency with at least four stools reference to baseline (Van der Laan et al., 2008).

Problem Statement

Researchers have seemingly agreed on the need for a standardized validated scale to assess adverse effects related to cancer therapy. Current research suggested that there is limited evidence of a comprehensive validation of the CTCAE scale (Davidson et al., 2007; Palazzi et al., 2008; Trotti & Bentzen, 2004; Watkins Bruner, 2007;) and hinted to the fact that the scales do not always meet the needs of their users and patient populations (Parulekar et al., 1998; Van der Laan et al., 2008). Researchers pointed to the need to examine the toxicity reporting methods and recommended a closer look at the assessment of toxicity, data display and analysis, and reporting methods. In the case of the CTCAE, Trotti et al. (2007) stated that the CTCAE did not go through a rigorous validation process. Brundage, Pater, and Zee (1993) discussed the fact that the validity and reliability of toxicity criteria grading scales were underdeveloped. Other experts purported that "a proposed set of toxicity criteria should undergo a systematic and scientific study of their feasibility, reliability, validity, responsiveness, and specificity for treatment before it is applied in clinical practice or in clinical research" (Bentzen et al., 2003). Despite the calls and inquiries on the validation of these grading systems, researchers have yet to develop a system that could be referred to as the gold standard for all grading of observed adverse events.

Validation of an instrument or scale consists of tests of validity and reliability. There are three major types of validity, namely, criterion-related, content validity, and construct validity (Carmine & Zeller, 1979). Criterion-related validity is used to predict how well an individual or group performs a behavior. Content validity shows the degree to which a measure reflects a domain of content. Construct validity reflects the extent to which a theoretical concept is in agreement with the measurement device or scale.

With respect to the validation of these scales, it appears that validity was conferred using content validity. Decades of use by the oncology community and enhancements made over time conferred content validity (Trotti et al., 2007). One could say that the very evolution of the CTCAE scale served as the content validation study. The clinical observations and updates or modification of original scales helped to fill in missing information or gaps in the scales. Works in health measurement scales such as Streiner and Norman (2008) suggested that content validity is an expression of the extent to which the scale covers all the important domains. Over the past thirty years of use, experts in oncology have judged all the domains of this scale valid as evidenced by their use in clinical trials, clinical observations of its usefulness, and subsequent revisions and updates over time. These activities by experts in the field conferred content validity. This may also be true for the RTOG/EORTC scale as indicated by Langendijk (2007).

Whereas content validity was established over the lifetime of the scale, reliability appears to be the segment of the validation process that has remained unsettled.

Reliability assesses that a test or scale is measuring an item in a reproducible manner (Streiner & Norman, 2008). The concept of reliability is predicated on the notion of repeatability and reproducibility. High reliability theorists posit that reliability is the ability to repeatedly reproduce an item or service of the same minimum standard on a consistent basis (Weick, Sutcliffe, & Obstfeld, 1999). Streiner and Norman (2008) viewed reliability as an index of the degree to which ratings obtained under different circumstances gives more or less the same results. An important distinction to make is that in this study, reliability has been defined the ability to yield more or less the same scores in the assessment of adverse events versus the reliability of the scale itself (Streiner & Norman, 2008). Reliability can be estimated using either the test-retest method where the same test is given to the persons in a given time period, split-halves method where a test is given in one sitting and the scores are split in half and the halves are correlated, or the alternative form method in which two tests that are intended to measure the same thing are administered to the same group of persons.

The test of reliability of the grading of adverse events using the CTCAE was never performed on any of the items in the scale (Trotti et al., 2007). Clearly, the literature has shown that the reliability of grading using the CTCAE is warranted. Researchers such as Brundage et al. (1993), Atkinson, (2011), and Langendijk et al. (2008) demonstrated the need for additional studies on the reliability of these scales. The current study sought to address this issue by conducting a retrospective study of the assessment of skin toxicity related to cancer therapy using quantitative methods to evaluate for reliability of grades.

Theoretical Rationale

Within the literature, there have been arguments that adverse events scores are not reliable. For example, in a 2004 study, Kaba, Fukuda, Yamamoto, and Ohashi evaluated the reliability of CTCAE, version 2.0 and demonstrated variability in toxicity assessments using common criteria. Kaba et al. (2004) highlighted the fact that there was much variability among clinicians within the same institution. To date, there have been few studies of this kind; a fact that supports the need for additional studies aimed at validating the CTCAE. Despite the efforts of combining acute and late effects scales into a more comprehensive scale and the benefits that would be realized from such a move, the lack of rigorous testing of reliability presents a major flaw in the CTCAE. Given the paucity of studies on the reliability of grading on such widely accepted scales, this dissertation study attempted to add to existing knowledge on the reliability of grading adverse events grades.

The benefits of conducting a study on the reliability of severity grades based on the scale are twofold: improvement in the treatment planning and quality of care for patients and improved quality of life for patients. Researchers from the H. Lee Moffit Cancer Center and the Department of Radiation Oncology at Columbia Presbyterian confirmed the fact that the need for additional knowledge about the side effects of cancer therapy is even more critical today since more than 500 new anticancer agents will be combined with traditional treatments as well as in more complex treatment modalities (Trotti & Chin, 2002). Therefore, having an instrument that is valid and whose grades are reliable is essential.

Statement of Purpose

The purpose of this study was to estimate the reliability of the grades given for skin reactions related to cancer therapy. Skin reactions were chosen because they are among the most common adverse events and are easier to measure in a retrospective study. Photographs were also chosen, to the extent possible, to minimize subjectivity in assessments. Inasmuch as toxicity assessments have been conducted in urban hospital in which the dissertation study took place, researchers have claimed that the process is still underdeveloped (Davidson et al., 2007; Kaba et al., 2004; Palazzi et al., 2008; Trotti et al., 2007; Watkins Brunner, 2007). Using a system approach, this study attempted to estimate the reliability of adverse events grades. To this end, the current retrospective study aimed to estimate the inter-rater agreement and inter-rater reliability of adverse events grades of skin reactions due to cancer therapy.

Research Questions

This study sought clarity by estimating the reliability of grades using the following questions:

1. What is the inter-rater agreement among rater participants?
2. What is the intraclass agreement between the following pairs of rater participants?
    a. physician – physician
    b. nurse – nurse
    c. physician – nurse
3. What factors contribute to the degree of variability identified in the study?

Potential Significance of the Study

Glatthorn and Joyner (2005) stated that a professionally significant study should make an important contribution to a field of study. Contributions may be judged from the viewpoints of (a) testing of a theory, (b) adding to the development of a theory, (c) extending existing knowledge, (d) changing existing beliefs, (e) suggesting a relationship between a phenomena, (f) extending a research methodology, or (g) providing greater insights into about a previously studied phenomenon. This study attempted to demonstrate professional significance in two areas: (a) extending existing knowledge and (b) confirming or changing prevailing beliefs about the grading systems for adverse events scales.

This study contributed to professional knowledge by demonstrating the degree of reliability of scores. The concept of reliability has been of special value to patients undergoing therapy and to physicians, nurses, health physicists, and technicians involved the treatment planning and implementation of said plan. The findings of this study may have substantive significance to the grading of adverse effects related to cancer therapy and could potentially change how grading is performed. A high degree of inter-rater reliability, kappa of >0.8 (Altman, 1991), would have served to reassure the oncology community that adverse events grades are indeed reliable. This would mean that the grades given by treatment professional can be reproduced repeatedly. Practitioners will have given, in most cases, the same numeric grade for the reported adverse events regardless of who performed the toxicity assessment. Furthermore, this study allowed for more accurate comparison of adverse event data. A high degree of reliability would have indicated that practitioners had a better understanding of the grading/scoring and less information would have been left for subjective interpretation. The study served to inform users and the developers of the need to improve this aspect of the scale. Overall, this study attempted to improve the quality of reported adverse events data related to cancer therapy.

Definitions of Terms

This section addresses some of the terminology used throughout the dissertation and provides an operational definition of major terms such as reliability and validity. In addition to providing the actual meaning of the terms used in this dissertation, the definitions also help to clarify the context in which certain terms are used.

Adverse events. An intended medical occurrence that happens during treatment with a medication or other therapy (National Cancer Institute, n.d.a). Adverse events may be classified as mild, moderate, or severe. The term adverse effect is often used interchangeably with adverse events.

Common Terminology Criteria for Adverse Events (CTCAE). An instrument used in the documentation of adverse events identified through clinical observations and laboratory findings (National Cancer Institute, 2011)

Grade or grading. Relates to the severity of the reported adverse event (National Cancer Institute, 2011).

Inter-rater agreement. Refers the extent to which rater participants agree on a set of judgments on the same items (Vogt, 2005).

Intraclass correlation (ICC). A multipurpose statistical procedure used to assess inter-rater reliability (Huck, 2008). ICC assesses the consistency of measurements made by different raters evaluating the same subjects.

Rater participant. Refers to the physicians and nurses who participated in the study. The term was coined to provide a clear distinction between raters and the patients' records under review in the retrospective study.

Toxicity or severity assessment. Refers to an evaluation of the degree to which something is harmful.

Chapter Summary

The issue of the reliability of grades using these scales has been widely studied, however, there has yet to be a study that demonstrates a high degree of inter-rater agreement across among practitioners and across disciplines. While there have been a few studies, commentaries, and editorials highlighting the need for a comprehensive validation of the scale, the depth of study required to make judgments of reliability is absent. Due to the paucity of studies of the test of reliability of grades, the focus of this dissertation study was, therefore, on the reliability of grades, and adds to the existing body of knowledge on the grading system of adverse effects. Additional knowledge on the subject matter in terms of a theoretical framework and the state of current literature on the reliability of the various grading systems is required to help close the gaps in everyday practice. Chapter 2 of this dissertation will present a review of the literature followed by an explanation of the research methodology in Chapter 3, a presentation of the study results in Chapter 4 and a discussion of those results in Chapter 5.

Chapter 2: Review of the Literature

Introduction and Purpose

The purpose of this chapter is to provide a review of existing literature on the estimate of reliability of adverse events grades and provides the theoretical framework that will guide the study. This chapter begins by discussing major works on the reliability of the adverse events grades specifically highlighting the extent to which the topic has been researched and gaps in the existing literature. The summary of the theoretical frameworks provides a historical background of General Systems Theory and High Reliability Theory and discusses their appropriateness and applicability to this study.

Literature Review

Searches for peer reviewed scholarly articles were conducted using the following main sets of keywords: "inter-rater agreement of toxicity scales, inter-rater agreement of RTOG scales, reliability of RTOG, reliability of CTCAE, validity of the CTCAE, validation of the CTCAE, and "reliability and validity of the CTCAE, and cross practitioner agreement of reliability grades". Searches were conducted using Elton Bryson Stephens Company (EBSCOhost), Proquest, Google Scholar, MEDLINE, and journal searches of a local medical library. The results of these searches yielded few studies on the reliability and validity of adverse events grades, which demonstrated that there is a gap in existing literature. In the studies found, researchers often cited an absence of a rigorous validation of the CTCAE scale (Bentzen, et al., 2003; Trotti, 2002; Trotti & Bentzen, 2004; Trotti et al., 2000) and of the RTOG (Langendijk et al. 2008).

To date, Brundage, et al. (1993) conducted a study that addressed the reliability of severity grades using scales that measure toxicity related to cancer therapy. In 2004, Kaba, Fukuda, and Yamamoto conducted research on the reliability of grading using the Common Toxicity Criteria, version 2.0. Other studies involving the grading systems were specialty based such as head and neck (Paccagnella et al., 2010), or adverse events such as oral mucositis **(**Parulekar et al., 1998)**.**

Due to the limited studies on inter-rater reliability of the grades, several experts in the field of radiation oncology have expressed concerns about the reliability of grading systems such as the CTCAE scale. Researchers contended that the "accuracy or validity of the grading criteria is not completely certain" (Trotti et al., 2000, p. 16). Accordingly, there has been no evidence of a rigorous systematic evaluation of validity or inter-rater reliability of the scale. Researchers further noted that variability in scoring is evident especially in categories that are considered more qualitative or subjective. Whereas these researchers purported that the quantitative-based criteria are more straightforward to apply, Brundage et al. (1993) demonstrated some variability among grades that are supposed to represent objective data. Kaba et al. (2004) also demonstrated variability in assessments and assignment of grades within study participants at a single institution.

Although scales such as the CTCAE scale have been in existence for decades, the instruments have not gained full acceptance by all radiation oncology practitioners perhaps due to the fact that studies on their reliability and validity appear to be absent. In terms of the validation process, one may only deem an instrument to be validated when both aspects of the validation process are satisfied, namely reliability and validity. The fact that an instrument is valid does not make it reliable or vice versa (Carmine & Zeller, 1979). In the case of the CTCAE, the dynamic nature of the instrument ensures that the instrument continually measures what it purports to measure. Its content validity and tests of reliability were perhaps not done on a continuous basis. In 2011, the Cancer Therapy Evaluation Program (CTEP) indicated that there is no evidence of a comprehensive validation for this instrument (Jackson, personal communication, 2011). Studies conducted by Brundage et al. (1993) and Kaba et al. (2004), in addition to editorials and commentaries, call for an in-depth review of the reliability of the toxicity scale and point to a gap in literature and a need to evaluate the reliability of grades.

Previous studies were evaluated to determine if new insights were reached on the topic (Galvan, 2009). The study conducted by Brundage et al. (1993) demonstrated inter-rater variability in the grading of adverse events experienced by subjects. Brundage and colleagues examined the reliability of the National Cancer Institute of Canada Clinical Trials Group (NCIC-CTG) and the World Health Organization (WHO) standard toxicity scale. In the study, Brundage et al. (1993) used simulated patients to present assigned scenarios to seven data managers who were used as raters. In random sequence, each rater assigned severity grades to 12 different scenarios on the two scales. The study focused on five categories within the scale: blood/bone marrow, flu-like symptoms, gastrointestinal, genitourinary, and neurologic. These categories involved quantitative (laboratory results) and qualitative (lethargy) criteria. Data managers/raters interviewed patients in a clinic setting and laboratory data were presented in written form to the data manager for interpretation and scoring. Researchers used proportion of agreement and kappa coefficient to estimate the degree of reliability of grading.

The objectives of the Brundage et al. (1993) study were (a) to determine the inter-rater and intra-rater reliability using both scales, (b) to determine conformity of these ratings to pre-established grades, (c) to identify areas of poor reliability in each scale. Statistical analysis of inter-rater scoring revealed perfect agreement for the NCIC-CTG white blood cells (BL_WBC) and WHO leucocytes categories. There was imperfect agreement in other quantitative-based categories. Sixty to eighty percent agreement was found in only 8 of the 18 clinical categories. There was a 60% agreement for lethargy, nausea, and WHO hemorrhage. The results showed higher intra-rater agreement. Most noteworthy is the fact that 65% (17/49) of grade 4 toxicity and 21% (15/70) were inappropriately scored. Inconsistencies in grading were attributed to random errors in coding adverse effects and the ambiguity in operational toxicity criteria. Researchers also condensed grade levels to form two broad categories of low grade versus high grade. Kappa statistics revealed only moderate agreement in approximately 17% (3/18) and poor in 11% (2/18) of clinical categories. Agreement within the broad groupings ranged from 0.04 to 0.82. While these findings were not exhaustive of all the results, they represented evidence of the variability of grading using toxicity scales.

Given the purpose of the dissertation study to estimate the reliability of grading of adverse events related to cancer therapy, the Brundage et al. (1993) study can be generalized to each criteria and type of cancer. Additionally, the Brundage study appeared to be easily reproducible. It has been consistently cited during the past 17 years. As of January 2012, the study was cited 79 times in scholarly articles ranging from cancer therapy to HIV-AIDS related studies. The high number of citations signals the strength of the study. The study is also noteworthy because of the importance of accurate recording of adverse events related to cancer therapy and for the purposes of comparing treatment and clinical outcomes. Brundage and colleagues provided a comprehensive review of the reliability of two widely used scales and provided the foundation for future studies on the reliability of grading using the CTCAE scale. The dissertation study attempted to estimate the degree to which the grading of adverse events yields the same results on repeated trials, hence the reliability of grades.

While the Kaba et al. (2004) study dealt directly with the National Cancer Institute’s Common Criteria, information herein was limited to the abstract. In the abstract, Kaba et al. (2004) highlighted the degree of inconsistency in the grading of adverse events. The study was based on a document review of 17 patient records, which is a small sample size. Five clinical research coordinators evaluated eight adverse events:

- Diarrhea (0.59 (95% CI, 0.35 – 0.82)
- nausea (0.47,0.23 – 0.71)
- vomiting (0.71, 0.49 – 0.92)
- stomatitis/pharyngitis 0.59 (0.35 – 0.82)
- infection (0.82, 0.64 – 1.01)
- febrile neutrapenia (0.88, 0.7 – 1.04)
- Infection by unknown source (0.82, 0.64 – 1.01)
- Sensory neuropathy (0.65, 0.42 – 0.87)

In the study, raters assigned grades based on consensus. Although the grades were indicative of fair to moderate agreement and in some instances very good agreement (Altman & Koch, 1991), Kaba et al. (2004) reported significant variability of grading, which were related to variations in clinical assessments and misunderstandings of severity or toxicity criteria. Kaba and colleagues suggested training and education on toxicity evaluations using common criteria, including interpretation of criteria.

Despite the Palazzi et al. (2008) claim that the CTCAE, version 3.0 proved reliable in the evaluation of head and neck cancer treatment**,** a more recent study reported imperfect agreement among raters (Atkinson et al., 2011). Palazzi et al. (2008) appeared to be the only researchers to claim that the scale is indeed reliable. However, the research team stated that original instrument, the CTCAE version 3.0, was simplified to enhance the recording of data. In this case, researchers may have inadvertently attributed a system failure in the CTCAE grading system to the actual CTCAE scale. As mentioned earlier, Carmine and Zeller (1979) stated that when assessing reliability, one is evaluating the grading or scoring on a particular test or activity, not the actual instrument. The research may be somewhat premature because not all components of the grading system were considered. Adjustments in the scale were directly related to the content validity of the scale.

Atkinson et al. (2011) evaluated the reliability of adverse events reporting for 393 cancer patients. Two clinicians using the CTCAE scale independently evaluated adverse events. Intraclass correlation coefficients were used to estimate reliability of grades. Atkinson et al. (2011) found moderate inter-rater reliability in the following seven criteria:

- Constipation: 0.50
- Diarrhea: 0.58
- Dyspnea: 0.69
- Fatigue: 0.50

- Nausea: 0.52
- Neuropathy: 0.71
- Vomiting: 0.46

Furthermore, researchers noted that the scores were less than desired. The study was one of a few that discussed the practical implications of lower than desired inter-rater reliability values. Atkinson et al. (2011) stated that a two-point difference in grades, in this case the CTCAE, is sufficient evidence to change the course of treatment. Atkinson et al. (2011) explored possible reasons for the lower levels of agreement and cited the fact that clinicians' own interpretations and experiences may have inadvertently influenced their ratings.

Theoretical Frameworks

General Systems Theory and High Reliability Theory were used as the frameworks for the dissertation study. General Systems Theory was used to examine components and subsystems involved in the grading of adverse events. The complimentary High Reliability Theory was used to guide the process of repeatability and reproducibility of scoring.

General Systems Theory. In explaining the intricacies of systems, General Systems Theory seeks to explore the interactions and dependencies and the actual links within the system to gain a broad view, or more aptly, a big picture view of the system in question (Covington, 1998). A General Systems Theory exploration includes an examination of the connection of the system with the environment. General Systems Theory was predicated on the notion of change (Coleman & Palmer, 1973). Whenever there is a change in the organization or the environment, further changes will occur to maintain equilibrium, and if homeostasis is not achieved, the system will wear out (Coleman & Palmer, 1973). Although some of these citations are dated, they are still meaningful today.

Applications of General System Theory. As a grand theory, General Systems Theory has been useful in answering questions and solving problems related to complex systems or organizations. Drack and Apfalter (2007) claimed that general systems theory was very influential in the scientific world and is still widely known. Although Coleman and Palmer (1973) did not cite specific studies, they alluded to the fact that General Systems Theory has been used to solve problems related to organizational design, leadership, and decision-making. Disciplines that have used system theory include, but are not limited to, computer security (Bell & Padula, 1973), career counseling (Patton & McMahon, 2006), vocational rehabilitation for persons with disabilities (Patton & McMahon, 2006) criminal justice (Bernard III, & Pare, 2005), military (Kuah, 2005; Schilling & Paparone, 2005), human resource (Mayrhofer, 2004), and neuroscience (Stephan, 2005). The wide use and popularity of General Systems Theory can be seen in the 314 papers published between 1970 and 2004 (Drack & Apfalter, 2007). Approximately 43% of these papers were published between 1995 and 2004 (Drack & Apfalter, 2007) highlighting the renewed interest in General Systems Theory.

More recently, Rivera and Karsh (2008) demonstrated the use of General System Theory by applying a systematic approach to patient safety for radiotherapy. Rivera and Karsh (2008) described the healthcare system as consisting of inputs, transformations, outputs, boundaries, environment, and feed processes. This description treated the healthcare system as a whole, bringing together all facets of providing quality care for patients. Rivera and Karsh (2008) discussed bringing together healthcare providers, equipment used in the planning and treatment processes, staff knowledge, policies and procedures, and environmental conditions in rooms, system boundaries such as different work shifts, and hierarchical boundaries (e.g., radiation medicine unit within a radiation medicine department or within a hospital). Rivera and Karsh (2008) explained the interactions and dependencies as part of the transformations component of the system. The authors discussed communication among healthcare providers and communication between healthcare providers and patients/family members. Another important feature of a system is the feed process and includes the processes of feedback, feed forward, or feed within (Rivera & Karsh, 2008). Information gathered in any of these processes is used to guide change(s) within the system.

In the dissertation study, General Systems Theory was used to examine the reliability of the grading and the degree to which internal and external factors influence the assignment of a given score to a reported case of an untoward event related to treatment. This theory was applicable because of the following:

- The complex and open nature of clinical settings
- The various systems involved in the final assignment of severity grades/scores
- The ability to predict what will happen

The basic components in the radiation medicine setting studied by the dissertation study consisted of the examiner, the examined (the photographs), and the examination. Brundage and colleagues (1993) used a similar configuration. The system was not inanimate in nature due to the required interactions, relationships, and dependencies between members of the radiation medicine healthcare team, radiation medicine caregivers, (i.e., physician to physician, nurse to nurse, and nurse to physician). Adding to the complexity of individualizing care for each radiotherapy patient is the caregiver's perception or expected reactions to the dose of the agent being used. These factors may influence how caregivers perceive the evidence being presented to them. Conversely, the examination refers to the act of evaluating a patient and takes into account the physical environment such as lighting, noise, room design and layout, and interruptions external to the examination process. Treatment modalities have changed over the years, which add to the complexities of providing care to patients. From a theoretical standpoint, Brundage et al. (1993) and Rivera and Karsh (2008) have used the same approach to improve patient safety for cancer patients.

Criticisms of General System Theory. While critics of General Systems Theory understand its unifying goal and seeing the overall picture of a system or an organization, they argued that General Systems Theory remains abstract. Mayrhofer (2004) claimed that General Systems Theory is not effective in concrete theoretical and practical analysis and suggested the need to supplement general systems theory with a more targeted theoretical concept. Mayrhofer (2004) suggested coupling General Systems Theory with a complimentary theory to make General Systems Theory more worthwhile for understanding changes in an organization and in the development of subsequent concrete actions to re-establish system equilibrium.

Another criticism of General Systems Theory has been its broad scope and the ease with which it can be applied across disciplines. Critics claimed that the theory attempts to answer too much and as a result, its explanatory value has declined (Covington, 1998). Even so, General Systems Theory has been embraced by the research community and has stood the test of time as evidenced by its past and current use (Covington, 1998).

Unfortunately, lack of consensus in the nomenclature of the theory may have created a flaw in General Systems Theory. A review of literature revealed the use of varying nomenclature and the inconsistent use of terms to describe the theory. For example, General Systems Theory has been termed system theory, system thinking theory, and social system theory. Troncale (2009) cited a need for a consensus of terms in describing the concepts. To further complicate matters, Troncale (2009) indicated a failure on the part of practitioners of General Systems Theory and the reductionist view to see the usefulness of both viewpoints in arriving at 'systemness". If the premise of General Systems Theory is "wholeness', then it is very probable that no one viewpoint is superior to the other.

Despite the numerous publications cited earlier, Troncale (2009) described the development of the field as sluggish and asserted that there is a need for long-term lineage of papers and investigators. The ad hoc "stop in and step out" (Troncale, 2009) method of research compared to the devoted lifelong research in fields such as chemistry and biology may have thwarted overall growth and knowledge base in General Systems Theory. In his study, Troncale (2009) identified 30 additional obstacles. One critically important criticism is that General Systems Theory needs to be more user-friendly. Troncale (2009) argued that making General Systems Theory more user friendly would remove the intimidating factor and open the field to others who might not have considered using General Systems Theory.

High Reliability Theory. For this study, High Reliability Theory (HRT) was used as a complimentary theory to address some of the research findings. HRT is based on the belief that organizations can have error free performance, despite the hazardous nature of their operations, if operations follow certain prescribed steps (Weick et al., 1999). According to HRT, "humans who operate and manage complex systems are themselves not sufficiently complex to sense and anticipate the problems that the system generates" (Ruchlin, Dubbs, & Callahan, 2004, p.52). High reliability is built on the lack of unwanted, unanticipated, and unexplained deviation in performance (Weick et al., 1999).

History of High Reliability Theory. In the face of catastrophic tragedies such as Chernobyl, Exxon Valdez, Bhopal, and Challenger, the Berkeley Group, consisting of researchers from the University of California, Berkeley (Morone and Woodhouse, Wildavsky, Roberts, LaPorte, Consolini, and Rochlin,) along with researchers from the University of Michigan (Weick, Sutcliffe, & Obstfeld) began exploring safety and reliability in high hazard industries. Most of the research conducted by the Berkeley Group involved nuclear power plant, nuclear submarine, and air traffic control. Critics of HRT questioned the merits of the studies due to the fact that research was conducted in ideal conditions as opposed to the worst of times such as combat and viewed this as a major flaw of HRT (Clarke, 1993).

Contributions of HRT. Regardless of this criticism, HRT has remained a strong contender in other high hazard fields (Christiansen, 2007). HRT has also taken root in the medical field (Beyea, 2005; van Stralen, 2008; van Stralen, Calderon, Lewis, & Roberts, 2008). In 2008, the Agency for Healthcare Research and Quality (AHQR), the research arm of the U.S. Department of Health and Human Services, developed a guidebook for hospital leaders interested in using the concepts of high reliability to improve quality of care and patient safety. Perhaps at this point, it is important to reiterate the meaning of reliability, which according to Carmines and Zeller (1979) is the tendency to consistently obtain repeated measurements of the same phenomenon.

Gaps in the Literature

The paucity of studies on the reliability of grades of adverse events experienced by cancer patients in clinical trials points to a significant gap in literature. To date, Brundage et al. (1993) provided one of the best source of information on reliability of grades. The discussion in this chapter clearly demonstrated that the questions about a comprehensive evaluation of the scale remains unanswered. A significant number of studies on the reliability of grades found inconsistencies in grades. Kaba et al. (2004), Brundage et al. (1993) and Atkinson et al. (2011) appeared to be the only studies that attempted to explore possible underlying reasons for the inconsistencies. However, researchers have yet to present an approach grounded in a theoretical perspective to resolve the issue of reliability of the grades.

Given the applicability of General Systems Theory and the fact that the AHRQ is advocating for high reliability systems in healthcare, researchers should start using theoretical approaches to examine the issue of validation of common health assessment scales. Additional research on the reliability of grades of adverse events related to cancer therapy using these theoretical approaches is needed to improve patient safety and overall health outcome of patients. Research should focus on practitioners' use of preferred scales as well as the reliability of grades from modified internal scales used in determining which grade to assign to the adverse event presented.

Chapter Summary

This chapter presented a brief summary of the General Systems Theory and aspects of High Reliability Theory and described how the theories were used as the guiding frameworks for the dissertation study. Use of General Systems Theory as a guiding framework was appropriate because of the complexities of assigning severity grade. On the other hand, High Reliability Theory was used to reliably obtain more accurate grading. Given the fact that this study used a quantitative methods approach, the theories were appropriate since the goals could be combined to advance the process of achieving a higher degree of inter-rater agreement. Since the majority of studies cited in this chapter showed an underlying theme of limited evidence on the reliability of grades, use of General Systems Theory and High Reliability Theory in a quantitative format provided the rigorous evaluation of reliability that is warranted.

The next chapter presents the methodology used to estimate the reliability of grades across oncology practitioners. The discussion demonstrates the effective use of General Systems Theory and High Reliability Theory by examining how raters' assigned grades followed by a statistical analysis similar to sensing making to identify subtle or obvious variations among rater participants.

Chapter 3: Research Design Methodology

Introduction

This chapter outlines the methodology used in this retrospective study to estimate the reliability of the adverse event grades for skin reactions related to cancer therapy. This chapter begins with a summary of quantitative research and presents the rationale for its selections. The chapter is organized in terms of the study objectives, research context, research participants, data collection instruments, and analysis of the data.

Summary of quantitative research. This study used a quantitative research design. Quantitative research is a reliable way of acquiring knowledge about specific observations or measurable aspects of behaviors. The distinguishing features of quantitative research that made this method suitable for the dissertation study are (a) focus on a small number of concept, (b) use of structured methodology and use of formal data collection tools, (c) emphasis on objectivity in the collection and analysis of data, (d) uses of statistical techniques to analyze data, (e) the researcher is not actively involved in data collection but collects data from a distance, (f) involves deductive reasoning (Brink, 2006). The study primarily focused on the concept of the reliability of severity grades as assigned by physician and nurses and used structured questions and formalized database method of data collection. Furthermore, data was analyzed using statistical techniques and used deductive reasoning in the interpretation of the findings.

The current study focused on the review of photographs of skin reactions and how practitioners rated subsequent adverse events. The main objective of the study was to estimate the inter-rater agreement and inter-rater reliability of grades among practitioners in one facility. Therefore, the study attempted to estimate agreement and inter-rater reliability as follows:

- Physician to physician
- Registered nurse to registered nurse
- Physician to registered nurse

Research Context

The study was conducted at an urban hospital in a major metropolitan area that provides radiation medicine services to oncology patients and participates in clinical trials at five centers. The clinical trials team was comprised of radiation oncologists, physicists, dosimetrists, radiation therapists, nurses, administrative, and information technology staff members from all sites of the Radiation Medicine department. The facility conducts approximately 2,800 consultations and treats 2,100 patients annually. On average 165 patients are treated daily. Intensity modulated radiation therapy (IMRT) is the most commonly used treatment field accounting for approximately 75% of treatment fields. This facility was important to the study because of its participation in cancer clinical trials and the frequent evaluation of patients who experience adverse events related to therapy. The facility also was an appropriate research site because it is representative of various treatment settings such as community hospitals, private practice, and is designated as a teaching facility. This facility used an internal departmental scale to grade adverse events related to cancer therapy. Appendix C is an excerpt of the internal departmental scale for skin reactions.

Research Participants

The study population was comprised of six physicians and five registered nurses. Criteria for selection of participants included (a) licensed independent practitioner (LIP) or registered nurse (RN) working in the field of oncology, and (b) previous experience in the grading of adverse events. Rater participants were recruited from the sponsoring institutions for the following reasons: (a) their expertise in radiation oncology, (b) familiarity with the adverse events scale used in the practice, (c) easier access to information, and (d) rater participants already bounded by the sponsoring institution's confidentiality statements. The rater participants were purposively selected because of close involvement with the process and the richness of experiences that enable the rater participant to elaborate or help explain the results of the quantitative phase.

Instruments Used in Data Collection

The data collection instrument consisted of an MS Access database form or template and related tables. The form displayed the photographs, associated clinical trials protocol, and questions designed to answer the research questions. Participants were asked to review 30 records each consisting of one to three photographs. The database contained photographs of only patients with a diagnosis of breast cancer. All individually identifiable patient information was removed from each case and assigned a unique identifier.

Participants evaluated the quality of the photograph to verify if, in their opinion, the quality was sufficient to render a grade. If a participant determined that the photograph was not sufficient to render a grade, he/she indicated the reason(s). Next, participants identified what criteria, standards, or guidelines assisted them in assigning grades for adverse events. Participants then indicated the chosen grade level based on their findings from the photographs. Within the MS Access form, participants were provided with a pre-labeled (Grade 0 – 5) drop down menu. The test took no more than 1 hour to complete. Participants were able to complete the test in one sitting or over shorter periods of time until completion.

Accessing tests materials. Test materials in the form of an MS Access database form was placed on the departmental shared drive in individual folders with each participant's name. Participants logged into their departmental shared drive to access the material. Participants only had access to their folder. The researcher had access to all folders and collected the data after each test. Scores were placed in pre-labeled de-identified Excel worksheet for further analysis. The worksheets were labeled MD1,…MD6 and RN1,…RN5. The researcher maintained a log of the participants' initials (first and last name) corresponding to the physician participant number or nurse participant number.

Methodological precautions. To ensure integrity of the data, rater participants were asked not to discuss the cases under review. Rater participants provided data independently and did not discuss scores with each other until after the study. Additionally, all individually identifiable patient and rater participant information was removed from all research documents to maintain confidentiality and anonymity.

Training of rater participants. Written instructions were provided in each participant's folder on the shared drive along with a copy of the test. The instructions included the purpose of the study, how to access the test, and contact information for the researcher.

Procedures. Rater participants reviewed the photograph(s) in each case and were asked to rate the quality of the photographs, assign a severity grade, and indicate what standards/references/guidelines were used to determine the assigned grade.

- Quality of photograph: Is the photograph of sufficient quality to render a severity grade?
- Severity Grade: Select the appropriate grade level for the adverse event e=seen in the photograph
- Standards/references/guidelines used in this study: Indicate standards/references/guidelines used, if any, to assign grade

Variables. There were two main variables in this study, the standards/guidelines/references and the severity grades. Standards/guidelines/references were defined as severity scales such as the CTCAE, RTOG, modified versions of these scales, or any other clinical guidelines. In this case, the independent variable, that probably caused or affected the outcome, was the standard/guidelines/references that influenced the participants' decision to assign a particular grade. The dependent variable was the assigned severity grade. The severity grade was a scaled representation of the adverse event with each increasing level representing a worsening in condition.

Data Analysis

The major analysis for this study was the estimation of inter-rater agreement and inter-rater reliability for all participants and across practitioners for each test. The analysis evaluated whether there was a difference in scores or whether the score was consistent. Inter-rater agreement for the quality of the photographs was also calculated to determine if the photographs were of sufficient quality to render severity grades. The primary endpoints, the estimation of the agreement and reliability of scores, were tested using kappa statistics and intraclass correlations. Cohen's kappa or kappa coefficient was used to determine the degree of agreement among raters (Brundage et al., 1993; Gross, 1986). However, Cohen's kappa deals mainly with ratings that involve two raters. Barnhart and Williamson (2002) stated that Cohen's Kappa appears to be the most popular index for evaluating agreement while accounting for rating by chance. Over the years, researchers have used different variations of kappa to estimate reliability based on the number of raters. In 1986, Gross estimated inter-rater reliability for multiple observers when the sample or population is small. In the dissertation study, general agreement or consensus was determined using Fleiss' method. Fleiss' kappa is a measure of inter-rater agreement, or whether judges make "exactly the same judgments" about the same cases; in other words, it assesses data for assignment of the same values for each case (Tinsley & Weiss, 1975, p. 359). Fleiss' kappa values were calculated in Microsoft Excel.

Alternatively, the intraclass correlation coefficient is a measure of inter-rater reliability or the degree to which ratings made by "different judges are proportional when expressed as deviations from their means" (Tinsley & Weiss, 1975, p. 359). Intraclass correlation (ICC) is a unitless measure or an index of reliability and varies between 0 and 1 (Weir, 2005). The use of ICC as a measure of inter-rater reliability has been considered as the "best measure of inter-rater reliability available for ordinal level measurement" (Tinsley & Weiss, 1975, p. 373). Furthermore, while an ICC of 0 indicates no reliability, an ICC of 1 reflects perfect reliability. ICC is generally reported based on the model and the type of analysis used and is denoted with two numbers in parentheses following the letters ICC (Huck, 2008). The three general models for ICC are Model 1, Model 2, and Model 3. Weir (2005) described the models as follows:

- Model 1: Subjects are rated by a different set of raters who were randomly selected from a population
- Model 2: Subjects are rated by the same group of raters and were also randomly selected
- Model 3: Subjects are rated by the same group of raters, however, raters were not randomly selected. Researchers typically use this model when the results of the study will not be generalized.

The method of analysis can be either a one-way or a two-way analysis of variance (ANOVA) (Huck, 2008). In a one-way analysis, the focus is on the reliability of a single rater and is distinguished as the second number in the ICC notation (e.g., ICC (3, 1)). A two-way analysis is interested in the reliability of the mean scores of a group of raters. When the focus is on the reliability of the mean for a group of raters, ICC is denoted as ICC (3, k), where k is the number of scores that were averaged to provide the mean. The ICC was used to determine the general agreement or consistency of single raters and the average agreement across raters. While there are six forms of the ICC that stem from these three models for reliability studies, the two-way random effects model was of interest in this particular study. A two-way random effects model was used given that a sample of raters was selected from a larger population of interest and each judge rated the same cases. ICC was calculated across scoring schema using SPSS software. SPSS is a collection of quantitative analysis software capable of performing statistical analyses and graphically displaying data.

Missing data. The statistical analysis was based on the following:

- Number of things to be rated: N = 30
- Number of raters: n = 11
- Number of ratings/categories possible (Grades 0-4): k = 5

While there were 11 rater participants and 30 cases to be rated, there were times when the data was missing. If a rater participant did not rate a particular case, the rater participant was omitted from the analysis, which resulted in a smaller pool of rater participants. Consideration was always given for all raters, raters with complete data sets, and completed cases.

Interpretation of Kappa. Inter-rater reliability scores were compared using the Altman (1991) guideline, which is an extension of Landis and Koch's (1977) guideline for interpreting the values of kappa. The Landis and Koch guidelines for interpretation of Kappa are presented in Table 3.1.

Table 3.1

Interpretation of Kappa Scores

Value of k	Strength of Agreement
≤ 0.20	Poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Good
0.81-1.00	Very good

Interpretation of ICC. Interpretation of ICC values was made on the scale suggested by Cicchetti and Sparrow (1981), closely resembling those developed by Fleiss (1981) and Landis and Koch (1977): $\text{ICC} < 0.40$ is "poor," $0.41 \leq \text{ICC} \leq 0.59$ is "fair," $0.60 \leq \text{ICC} \leq 0.74$ is "good," and $0.75 \leq \text{ICC} \leq 1.00$ is "excellent" (Cicchetti, 1994).

Chapter Summary

This chapter provided a general perspective of the methods that were used for data collection and analysis in this study estimating the reliability of scores related to adverse events after cancer therapy. The instrument used to collect the data was described and the overview of the data collection plan was outlined. Additionally, the data analysis section explained the statistical method used to estimate inter-rater agreement and inter-rater reliability, as appropriate, for picture quality and for rater participants. The next chapter presents the results of the study, specifically as it relates to inter-rater agreement for scores assigned by rater participants and ICC for group means across physicians and nurses.

Chapter 4: Results

Research Questions

This study was conducted to estimate the reliability of grades related to adverse events after cancer therapy. The chapter is organized in terms of the research questions and hypothesis from Chapter 1.

The research questions are:

1. What is the inter-rater agreement among rater participants?
2. What is the intraclass agreement between the following pairs of rater participants?
    a. physician – physician
    b. nurse – nurse
    c. physician – nurse
3. What factors contribute to the degree of variability identified in the study?

Data Analysis and Findings

The data analysis section summarizes the results of the research questions. The first subsection provides a report on the inter-rater agreement for the quality of the photographs. This included indices of inter-rater agreement, i.e., Fleiss' Kappa, for (a) all photographs with a severity grade regardless of whether or not the rater participants deemed the quality sufficient to render a grade, and (b) inter-rater agreement and inter-rater reliability for only those photographs with a positive rating for picture quality. ICC was not calculated for picture quality given that this is a strictly nominal rating, *yes* or *no.*

The next subsection provides the findings of Fleiss' Kappa and ICC across the two healthcare disciplines: (a) physician – physician (b) nurse – nurse, and (c) agreement between physicians and nurses.

Picture quality. Although there were a total of 11 raters, only 7 of those raters had complete data for picture quality, and only 18 cases had ratings by all 11 raters. As such, Fleiss' kappa was calculated in three ways:

1) all data were included in the calculation, and consideration of missing values was accounted for in the calculations of P_i and in the calculation of total number of ratings;
2) only data from the seven raters who rated every case for sufficient quality were included in the calculation; and
3) only data from the 18 cases that had ratings from all 11 raters were included in the calculation. This allowed for the researcher to evaluate agreement by assessing the whole data set, as well as taking into consideration the limitations faced by using Fleiss' kappa when data are missing.

Evaluation of the full data set resulted in Fleiss' kappa = 0.052 indicating poor agreement among all raters across all cases. Evaluation of data from only those seven raters who rated pictures for each case resulted in Fleiss' kappa = 0.118. This indicates that even among only those raters who gave a rating stating whether the picture was of sufficient quality or not to render a grade, agreement was poor. Evaluation of data from only those cases with quality ratings from all 11 raters resulted in Fleiss' kappa = 0.059. Again, this indicates that even when considering only cases where judgments of picture quality were made, there was poor inter-rater agreement among raters. It was evident that regardless of considerations for missing data, there was poor inter-rater agreement among raters in regard to picture quality. Interpretation of inter-rater agreement (κ = 0.059) was based on the Altman (1991) guidelines in which κ ≤ 0.20 is indicative of poor agreement.

Raters suggested the following reasons for stating a picture was not of sufficient quality to render a grade: (a) area obscured by crème such as silvadene or by the patient's finger as she supports the breast, (b) difficulty assessing edema and tenderness, (c) unclear if desquamation was dry or moist, (d) flash artifact that made the area difficult to visualize, (e) picture too dark making underarm and under breast difficult to see, or (f) unclear about varying levels/shades of erythema.

Ratings for photographs deemed of sufficient quality. Kappa value for inter-rater agreement for only those cases in which the physicians and nurses agreed that the quality of the photographs were sufficient to render a severity grades increased from 0.2051, which considered both *yes* and *no,* to 0.227, which accounted for all cases with yes for quality of photograph. κ = 0.227 indicated fair inter-rater agreement. For cases that a specific rater participant judged as insufficient, yet assigned at grade, κ = 0.2104 (from 0.1928 initial estimate), which is indicative of fair inter-rater agreement based on the Altman, (1991) guidelines. Controlling for only rater participants with full sets of data, κ = 0.2238. Again, this represented fair agreement as determined by Altman (1991). For only cases with full sets of grades, κ = 0.2376. The initial kappa value for this category regardless of the quality of the photographs was 0.2216. Again, there was no change in the interpretation of kappa value for this analysis. In sum, there was only a minor increase in κ- values for the quality of the photograph, which does not wholly explain the lower levels of inter-rater agreement.

Research question 1. Research question 1 was, "What is the inter-rater agreement among rater participants?" This research question was assessed by examining agreement of both photograph quality and grade rendered by participants. First, agreement on whether the photograph for each record was of sufficient quality to render a grade was assessed using Fleiss' kappa. Second, agreement on grade assigned was evaluated through use of both Fleiss' kappa and ICC.

Grades between 0 and 5 were assigned by raters with grades representative of skin and subcutaneous tissue disorder descriptions. Any grade of 5 was removed from the all data sets since grade 5 represents death and there were no reported deaths. Grade agreement was evaluated using both Fleiss' kappa and ICC; Fleiss' kappa was implemented due to the descriptive, ordinal nature of what the grades represented, while ICC was calculated given that data could be treated as ordinal ratings. Once again, missing data was an issue; only 8 raters provided a grade for every case, and only 22 cases had ratings from all 11 raters. As such, kappa was calculated three times in the same manner as was used for assessment of picture quality agreement.

Fleiss' kappa using all data from all raters was calculated as 0.193, indicating poor agreement. Agreement was slightly improved to fair agreement through considerations made for missing data. When data were used only from the eight raters with complete data sets, Fleiss' kappa = 0.205. When only the 22 cases with ratings from all 11 raters were considered, Fleiss' kappa = 0.222. This indicates that when assessed using Fleiss' kappa, inter-rater agreement was fair. In contrast, when considering inter-rater reliability through the use of a two-way random consistency ICC calculation, excellent agreement was found, ICC (2, 11) = 0.956, *p* < .001, 95% CI (0.521, 0.809). It is, however, important to note that only the 22 complete cases were included in this calculation. Inter-rater agreement ($\kappa = 0.193_{\text{all raters}}$) is indicative of poor agreement on the Altman (1991) interpretative guidelines. Inter-rater agreement ($\kappa = 0.205_{\text{8 raters}}$ and $\kappa = 0.222_{\text{11 raters}}$ was fair as indicated by Altman's guidelines where fair is $0.21 \leq \kappa \leq 0.40$. For ICC values, the magnitude of inter-rater agreement was compared to Cicchetti and Sparrow's (1981) interpretive guidelines, in which an ICC of 0.956 is considered excellent; 0.75 ≤ ICC ≤ 1.00 is considered "excellent" agreement.

Research question 2. Research question 2 was, "What is the inter-class agreement between the following pairs of rater participants? (a) physician-physician, (b) nurse-nurse, (c) physician-nurse". The overall measurement of inter-rater reliability was fair to fair to excellent. This research question was assessed using both Fleiss' kappa and ICC. Participants were split into two groups based on participant occupation. Fleiss' kappa and ICC were calculated for physicians and nurses separately; these values were considered separately in consideration of parts a. and b. and in contrast in consideration of part c. For simplicity and consistency between Fleiss' kappa and ICC values, only those 22 cases with ratings from all participants were included in the calculations presented.

Physician-physician. Inter-rater agreement among physicians ranged from fair to excellent depending on the statistical method used to calculate inter-rater reliability. When evaluating only the six physician participants, Fleiss' kappa for grades = 0.257. This indicates fair agreement on grades assigned by physicians. In contrast, use of ICC to evaluate inter-rater reliability indicated excellent reliability among physicians, ICC (2, 6) = 0.958, *p* < .001, 95% CI (0.924, 0.980).

Nurse-nurse. The evaluation of scoring for the group of nurses revealed a similar pattern of inter-rater reliability in which kappa was on the lower spectrum and ICC displayed almost perfect agreement. When evaluating grades assigned by the five participants who were nurses, Fleiss' kappa = 0.134. This indicates there was poor agreement on grades among those participants who were nurses. Evaluation of inter-rater reliability using the ICC indicated excellent reliability, however, ICC (2, 5) = 0.859, *p* < .001, 95% CI (0.737, 0.934).

Physician-nurse. When comparing the value of Fleiss' kappa found for physicians only with that found for nurses only, it is evident that physicians had better agreement on grades assigned than nurses. Physicians were found to demonstrate fair agreement, while nurses demonstrated poor agreement. Although both groups of participants showed excellent inter-rater reliability through evaluation of ICC, the resulting coefficient for physicians was higher than that for nurses, again indicating more reliable responses from physicians.

Research question 3. Research question 3 was, "What factors contribute to the degree of variability identified in the study?" While not done explicitly, variation in scoring was attributed to the use of different scales, which inherently led to different interpretations of the same photograph. Four participants indicated using the RTOG (Appendix B) to assist in the assignment of a severity grade. One participant reported using the CTCAE (Appendix A) while only one reported using an internal departmental scale (Appendix C). The remaining five participants did not indicate which scale, or more aptly, standards/guidelines/references were used in the assignment of grades. This research question was partially assessed through the examination of agreement among those participants who indicated the used of RTOG criteria for assigning grades. These participants displayed poor agreement, Fleiss' kappa = 0.154. Again, however, excellent inter-rater reliability was found among these participants, ICC (2, 4) = 0.921, $p < .001$, 95% CI (0.861, 0.959).

Summary of Results

The objective of this study was to examine the cross practitioner variability of adverse events grades for skin reactions related to radiation therapy by examining (a) inter-rater agreement for all participants, (b) intraclass correlation for physicians and nurses, and (c) to determine factors that may have contributed to any significant variability in grades. Fleiss' kappa and intraclass correlation coefficients were used to analyze adverse event grades. In order to present an accurate representation of inter-rater agreement across practitioners, missing values were excluded from the data analysis. Exclusion of missing values highlighted some issues in data collection instrument design and collection procedures. Using a benchmark of 0.8 (Altman, 1991), inter-rater agreement, based on Fleiss' kappa, was not met. Inter-rater reliability based on intraclass correlation coefficients exceeded expectations for each group. The findings indicated a significant difference in inter-rater agreement and inter-rater reliability. The next chapter discusses the findings and explores possible implications for various stakeholders. Additionally, recommendations for future study are presented.

Chapter 5: Discussion

Introduction

The aim of this study is to estimate the degree of inter-rater agreement and inter-rater reliability among physicians and nurses required to evaluate patients who are underwent radiotherapy and assign severity grades to adverse events related skin reactions after therapy. This chapter discusses possible implications of the findings, limitations of the study, and makes recommendations for future research.

Using Fleiss' kappa and intraclass correlation coefficients (ICC), this study estimates inter-rater agreement and inter-rater reliability as measured by the quality of photographs used in this study, the adverse events grades assigned by practitioners, and the grading system indicated by the participants, if any. The study finds fair agreement for grades assigned by six physicians and five nurses based on the Altman (1991) interpretive guidelines of where $0.21 \leq \kappa \leq 0.40$ is indicative of fair agreement. The study reveals excellent inter-rater reliability between each group as evidenced by correlations that are ≥ 0.859 (all raters – 0.956, physician to physician – 0.958, nurse to nurse – 0.859).

The fair agreement found by interpreting Fleiss' kappa indicates that raters did not often give the same exact grades. This is countered by the excellent reliability found through interpretation of ICC. This implies that while raters assigned grades in similar manners, (i.e., raters agreed on direction of severity, the raters may have had different definitions of grade classifications). While different raters might agree that Picture A shows a higher grade than Picture B, others might rate Picture A as grade 3 and Picture B as grade 2, or A is grade 2 and B is grade 1, or A is grade 4 and B is grade 3. This would result in high reliability as assessed by ICC, but low agreement as assessed by Fleiss' kappa. The fact that the significance test associated with the ICC is significant at $p < .001$ indicates that the ICC was significantly different from 0, or that there was not greater than chance reliability. In other words, the raters involved in this study are found to be significantly reliable in how they assigned grades.

Even though this is a small study, the results confirms the findings from previous studies and demonstrates that additional work is needed to achieve generally acceptable levels of inter-rater agreement (> 0.8, Altman, 1991). This study confirms the assertions made by Trotti & Bentzen (2004) about the differences in grades across grading systems. Although Brundage et al. (1993) evaluated a different component of the scale, the underlying issue in this particular study points to the fact that the grades are not reliable. Kaba et al. (2004) found variability in toxicity assessments and grades. In 1998, Postma et al. demonstrated significant variability across grading systems. In contrast, Parelukar (1998) found that the CTCAE scales was reliable; however, Parklukar noted that the scale was somewhat modified. On the other hand, the dissertation study demonstrated excellent inter-rater reliability for both physicians and nurses.

Implications of Findings

This study has implications for physician and nursing practices, education, and research. Changes in the practice environment and education, as well as additional research might improve inter-rater agreement and inter-rater reliability. This viewpoint takes into account a systems thinking approach since all three areas are interconnected. Each area has a direct impact on outcomes in other areas and on subsequent patient outcome.

Implications for practice. Physicians and nurses routinely assess patients and play a vital role in the delivery of comprehensive care to oncology patients. More objective assessments are needed in the evaluation of adverse events related to cancer therapy. Being able to differentiate between varying degrees of adverse events will contribute to increased patient outcomes. For example, being able to distinguish between mild and moderate conditions may change the course of treatment and significantly optimize patient outcome. The findings from this study suggest that practitioners who are required to use these scales to grade adverse events must be trained in the proper use and interpretation. Notwithstanding the importance of subjective assessments in patient care, the findings of this study suggest that the move toward a system that yields highly reliable grades means that practitioners rely on more objective measurements. Although care must be individualized to each patient, understanding the concept of reliability and designing work practices to ensure the repeatability of steps that almost guarantees an inter-rater agreement of $\geq$ 0.8 is essential. This may mean increased reliance on the scale of choice.

Implications for research. This current study is among the few studies on inter-rater reliability in which nurses are rater participants. The findings of this present study contribute to the body of research by documenting the nurses' assessments of the reliability of grades and evaluating how their assessments compared to a physician group. While these assessments were limited by a small sample size of only five nurses, the inter-rater reliability was slightly lower than that of the physicians. This is a noteworthy finding considering the differences in medical and nursing education. However, with a small sample size of both nurses and physicians and the recurring missing data in this study, conclusions drawn from the results of this study should be used with caution. Since nurses routinely assign severity grades, research should seek to include nurses in studies aimed at testing inter-rater agreement.

Implications for education. The low to fair inter-rater agreement findings show that a critical element in the education process of practitioners in this specialized area of practice has not been accomplished. The findings highlight concerns about the training needs of physicians and nurses as they begin and continue to work with particular grading systems.

Limitations

While the expectation of this study was to assess the reliability of the grading of adverse events related to cancer therapy, there are a few limitations that must be considered. Limitations are factors or boundaries that may impact the study but are outside the span of control of the researcher, yet these factors are not enough to discontinue research activities (Cottrell & McKenzie, 2011). The first limitation that must be acknowledged that the study was conducted in one institution only even though the institution has characteristics representative of national or statewide areas: community hospitals, private practice, medical schools, and diversity in patient population. This affected the sample size by limiting the number of participants in the study. Additionally, the number of raters in this study was small and represented a purposive sample, which significantly limits the study's generalizability.

Another limitation of this study is the decision to examine only skin reactions to cancer therapy. Although the adverse events included in the study are limited to one category, skin reactions are among the most common symptoms experienced by patients. The fact that these skin conditions were presented only in photographs as opposed to evaluating and grading in situ is another limitation of the study. Factors such as moisture, or the amount thereof, were not easily discernible in photographs. Also, the findings derived from palpation are not always captured in still photographs. There is no substitute for evaluating patients' in real time and subtle characteristics of conditions may be missed by the raters.

The final limitation of this study is related to the design of the database. The data collection tool did not contain any forced fields, which gave participants the option to only partially answer the questions for each record. The effect of this limitation is evident in the analysis of Research question 3 that attempted to elicit factors that may have contributed to any significant variability. To some degree, this limited the depth of comparison that could be made.

Recommendations

Anecdotally, a rating of fair is generally perceived as acceptable; however, for this study and based on the works of previous interpretive guidelines from Altman (1991), Landis and Koch (1977), and Cicchetti and Sparrow (1981), inter-rater agreement ≥ 0.8 is generally regarded as acceptable. These findings suggest ways to improve the reliability of grades across grading systems. Interventions geared toward the elimination or reduction of subjectivity in the grading process is most desirable. To improve the inter-rater reliability of grades, one strategy is to train practitioners on the scale of choice regardless of previous experience. Given the comprehensive details of many of these scales, it is advisable to have ongoing training on each anatomical system with each scale, including the related grading criteria. Infiltration of previous experience with other scale(s) into settings where the scale of choice is otherwise also might be eliminated or reduced. Strategies might also seek to undo these personal mental models of the presentation of each grade level of adverse events. This might be accomplished by embedding the scale in the electronic medical record and allowing practitioners to make grade selections within the actual scale.

Another effective strategy would be to design an inter-rater agreement performance improvement project. Such a strategy should lead practitioners to establish an inter-rater agreement target such as $\geq$ 0.80 and use tools such as Plan-Do-Study-Act (PDSA) to increase levels of inter-rater agreement. This strategy might also be grounded in high reliability theory whereby the goal and objectives would be to use the same steps and procedures to evaluate and assign severity grades. In addition to these strategies, a random perspective peer review might be a viable option. In this case, a second practitioner, preferably of the same healthcare discipline, evaluates the same patient during the same visit.

Consideration should be given to additional studies in this area, specifically test/retest method of reliability. This type of reliability testing will estimate inter-rater reliability over two time periods. Furthermore, testing the reliability of grades for specific scales such as the CTCAE or the RTOG for skin reactions is recommended.

The benefits of increasing inter-rater agreement include improved communication between practitioners. These improvements might be evident in the commonly occurring handoff process between practitioners as patients are transferred to different levels of care. One of the overarching goals of the first scales was to create a common language for practitioners. Given the substantial deviations from generally acceptable levels of inter-rater agreement, one might argue that while there is a common language for describing adverse events related to cancer therapy, the subsequent translations and interpretations are amiss.

Conclusions

This study tested the cross practitioner inter-rater agreement and inter-rater reliability of grades for adverse events related to skin reactions for patients undergoing cancer therapy. Researchers assert that the grades across the adverse events grading systems are not reliable. Researchers such as Langedijk (2008) purport that the RTOG does not reliable measure these adverse events. Atkinson (2011) also made a similar claim. Significant variability was demonstrated across four grading systems for the same adverse event (Postma, 1998). If practitioners can reliably scores these adverse events to the extent that they are $\geq$0.8, as determined by Altman's (1991) or Cicchetti and Sparrow's (1977) interpretative guidelines, it stands to reason that patients' outcomes and communications between practitioners should improve. Descriptive statistics show that there is a significant difference in how each discipline assigns severity grades. The results show the proclivity of nurses in this research setting to assign higher severity grades for skins reactions observed in the photographs. While this study establishes that kappa is fair and intraclass correlation (ICC) is excellent using Fleiss' kappa and ICC respectively, grades 2 and 3 are the most common grade of the skin reactions in photographs.

Claims that the scale, or more specifically the grades on these scales, is not reliable should not be ignored. In fact, the issue at hand is not merely one of unreliable grades but how to increase the inter-rater reliability. Understanding what actions contribute to increased inter-rater reliability is worth examining. To date, the literature shows that this is an issue that has not been adequately addressed. This dissertation study attempts to understand the variability by examining grades across grading systems. Unfortunately, this portion is not fully developed due to limited data for the different grading systems that were identified. As a result, inter-rater agreement and inter-rater reliability was estimated for only the RTOG scale. In the dissertation study, there is poor inter-rater agreement and excellent inter-rater reliability for grades on the RTOG. One may conclude from this that determining how to make the grades from these grading systems more reliable may lead to significant improvements in inter-rater agreement and inter-rater reliability.

References

Agency for Healthcare Research and Quality. (2008). *Becoming a high reliability organization: Operational advice for hospital leaders* (Publication No. 08-0022).

Altman, D. G. (1991). *Practical statistics for medical research*. London: Chapman & Hall.

Atkinson, T. M., Yuelin Li, Y., Coffey, C. W., Sit, L., Shaw, M., Lavene, D., Bennett, A. V., Fruscione, M., Rogak, L., Hay, J., Go¨nen, M., Schrag, D., & Basch, E. (2011). *Reliability of adverse symptom event reporting by clinicians. Quality of Life Research*. doi: 10.1007/s11136-011-0031-4.

Barnhart, H. X. & Williamson, J. M. (2002). Weighted least-squares approach for comparing correlated kappa. *Biometrics*, *58*(4), 1012-1019.

Bell, D. E. & La Padula, L. J. (1973). *Secure computer systems: Mathematical foundations. Electronic Systems Division.* Airforce Systems Command. United States Airforce. Report No.: ESD-T-73-287, vol.1.

Bentzen, S., M., Dorr, W., Anscher, M., S., Denham, J., W., Hauer-Jensen, M., Marks, L. B., Williams, J. (2003). Normal tissue effects: Reporting and analysis. *Seminars in Radiation Oncology, 13*(3), 189-202. Doi: 10.1016/S1053-4296(03)00036-5.

Bernard, T. J., Paoline III, E. A., & Pare, P. (2005). General system theory and criminal Justice. *Journal of Criminal Justice*, *33*(3), 203-211.

Beyea, S. (2005, June). High reliability theory and highly reliable organizations. *AORN Journal, 81*(6).

Brink, H. (2006). *Fundamentals of Research Methodology for Health Care Professionals. (2nd ed.).* Cape Town, South Africa: Juta & Co. (Pty) Ltd.

Brundage, M. D., Pater, J. L., & Zee, B. (1993). Assessing the reliability of two toxicity scales: Implications for interpreting toxicity data, *Journal of the National Cancer Institute*, *85*(14), 1138-1148. doi: 1093/jnci/85.141138.

Carmines, E. G. & Zeller, R. A. (1979). *Reliability and validity assessment.* Thousand Oaks, CA: Sage Publishers.

Clarke, L. (1993). Drs. Pangloss and Strangelove meet organizational theory: High reliability organizations and nuclear weapons accidents. *Sociological Forum*, *8*(4), 674-689.

Coleman, C. J. & Palmer, D. D. (1973). Organizational application of system theory. *Business Horizons*, *16*(6), 77-84.

Cottrell, R. R. & McKenzie, J. F. (2011). *Health promotion and education research methods: Using the five-chapter thesis/dissertation model*. Sudbury, MA: Jones and Bartlett Publishers.

Covington, W. G. (1998). *Creativity and general systems theory*. Retrieved from http://www.bookpump.com/upb/pdf-b/112872Xb.pdf

Davidson, S. E., Trotti, A., Ozlem, A., Seong, J., Lau, F. N., Da Motta, N. W., Jeremic, B. (2007). *International Journal of Radiation Oncology Biology*, Physics, *69*(4), 1218-1221. doi: 10.1016/ijrobp.2007.04.054.

Drack, M. & Apfalter, M. (2007). Is Paul A. Weiss' and Ludwig von Bertalannffy's system thinking still valid today? *Systems Research and Behavioral Science*, *24*, 537-546. doi: 10.1002/sres.885.

Galvan, J. L. (2009). *Writing literature reviews*. Glendale, CA: Pyrczak Publishing.

Glatthorn, A. A. & Joyner, R. L. (2005). *Writing the winning thesis or dissertation.* Thousand Oaks, CA: Corwin Press Publisher.

Gross, T. (1986). The kappa coefficient of agreement for multiple observers when the number of subjects is small. *Biometrics, 42*(4), 883-893.

Huck, S. W. (2008). *Reading Statistics and Research.* Boston, MA: Pearson Education, Inc.

Kaba, H., Fukuda, H., Yamamoto, S., & Ohashi, Y. (2004). Reliability of the National Cancer Institute –Common Toxicity Criteria version 2.0. (Abstract). *Gan To Kagaku Ryoho, 31*(8), 1187-1192.

Kuah, A. (2005). *Reconceptualising the military-industry complex: a general systems approach.* Institute of Defence and Strategic Studies, Singapore. Retrieved from http://www.isn.ethz.ch/isn/Digital-library/Publications/Detail/?ots591=0c54e3b3-1e9c-be1e-2c24-a6a8c7060233&lng=en&id=27040

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Langendijk, J. A., Doornaert, P., Verdonck-de Leeuw, I. M., Leemans, C. R., Aaronson, N. K., & Slotman, B. J. (2008). Impact of late treatment-related toxicity on quality of life among patients with head and neck cancer treated with radiotherapy. *Journal of Clinical Oncology, 26*(22), doi: 10.1200/JCO.2007.14.6647.

Marais, K., Dulac, N., & Levenson, N. (2004*). Beyond normal accidents and high reliability organizations: The need for an alternative approach to safety in complex systems.* Retrieved from http://esd.mit.edu/symposium/pdfs/papers/marais-b.pdf

Mayrhofer, W. (2004). Social system theory as a theoretical framework for human resource management – benediction or curse? *Management Review*.

National Cancer Institute. (n.d.a). *Types of treatment*. Retrieved from http://www.cancer.gov/cancertopics/treatment/types-of-treatment

National Cancer Institute. (n.d.b). *Common Terminology Criteria for Adverse Events - Instructions and Guidelines*. Retrieved from https://webapps.ctep.nci.nih.gov/webobjs/ctc/webhelp/welcome_to_ctcae.htm

National Cancer Institute. (2011). *NCI guidelines for investigators: Adverse events reporting requirements for DCTD (CTEP and CIP) and DCP INDs and IDESs.* Retrieved from http://ctep.cancer.gov/protocolDevelopment/electronic_applications/docs/aeguidelines.pdf

Paccagnella, A., Ghi, M. G., Loreggian, L., Buffoli, A., Koussis, H., Mione, C. A., Bonetti, A., Campostrini, F., Gardani, G., Ardizzoia, A., Dondi, D., Guaraldi, M., Tornio, R., & Gava, A. (2010). Concomitant chemoradiotherapy versus induction docetaxel, cisplatin and 5 fluorouracil (TPF) followed by concomitant chemoradiotherapy in locally advanced head and neck cancer: a phase II randomized study. *Annals of Oncology, 21*, 1515–1522. doi:10.1093/annonc/mdp573

Palazzi, M., Tomatis, S., Orlandi, E., Guzzo, M., Sangalli, C., Potepan, P., Fantini, S., Bergamini, C., Gavazzi, C., Licitra, L., Scaramellini, G., Cantu, G., & Olmi, P. (2008). Effects of treatment intensification on acute local toxicity during radiotherapy for head and neck cancer: Prospective observational study validating CTCAE, version 3.0 scoring system. *International Journal of Radiation Oncology Biology and Physics*, *70*(2), 330-337. doi: 10.1016/j.ijrobp.2007.06.022.

Parulekar, W., Mackenzie, R., & Bjarnason, G. (1998). Scoring oral mucositis. *Oral Oncology, 34*(1), 63-71. doi:10.1016/S1368-8375(97)00065-1.

Patton, W. & McMahon, M. (2006). The systems theory framework of career development and counseling: Connecting theory and practice. *International Journal for the Advancement of Counselling 28*(2), 153-166.

Postma, T. J., Heimans, J. J., Muller, M. J., Ossenkoppele, G. J., Vermorken, J. B., & Aaronson, N. K. (1998). Pitfalls in grading severity of chemotherapy-indiced peripheral neuropathy. *Annals of Oncology, 9*, 739-744.

Rivera, A. J. & Karsh, B. T. (2008). Human factors and systems engineering approach to patient safety for radiotherapy. *International Journal of Radiation Oncology Biology Physics, 71*(1), supplement, 174-177. doi: 10.1016/j.jrobp.2007.06.088

Rosewall, T., Yan, J., Bayley, A. J., Kelly, V. Pellizzari, A., Chung, P., & Catton, C. N. (2009). Inter-professional variability in the assignment and recording of acute toxicity grade using te RTOG system during radiotherapy. *Radiotherapy and Oncology, 90*, 395-399.

Ruchlin, H. S., Dubbs, N. L., & Callahan, M. A. (2004, January/February). The role of leadership in instilling a culture of safety: Lesson learned from the literature. *Journal of Healthcare Management, 49*(1), 47-59.

Schilling, M. & Paparone, C. (2005). Modularity: An application of general systems theory to military force development. *Defense Acquisition Review Journal*, 278-293. Retrieved from http://www.dtic.mil/cgibin/GetTRDoc?AD=ADA441764&Location=U2&doc=GetTRDoc.pdf

Streiner, D. L. & Norman, G. R. (2008). *Health Measurement scales: A practical guide to their development and use. (4th ed.)* Oxford: Oxford University Press.

Troncale, R. (2009). The future of General Systems Research: Obstacles, potentials, cases studies. *Systems Research and Behavioral Science*, *264*, 511-522. doi: 10.1002/sres.993.

Trotti, A. (2002). The evolution an application of toxicity criteria. Seminars in Radiation *Oncology, 12*(1), Suppl 1, 1-3. doi: 10.1053/srao.2002.31353.

Trotti, A. & Bentzen, S. M. (2004). The need for adverse effects reporting standards in oncology clinical trials. *Journal of Clinical Oncology, 22*(1), 19-22. Doi: 10.1200/JCO.2004.10.911.

Trotti, A., Byhardt, R., Stetz, J., Gwede, C., Corn, B., Fu, K., Gunderson, L., McCormick, B., Morris, M., Rich, T., Shipley, W., & Curran, W. (2000). Common Toxicity Criteria: Version 2.0. An improved reference for grading the acute effects of cancer treatment: Impact on radiotherapy. *International Journal of Radiation Oncology Biology, Physics, 47*(1), 13-47. PHS0360-3016(99)00559-3.

Trotti, A. & Chin, L. J. (2002). Adverse effects: A pandora's box for oncology. *International Journal of Radiation Oncology Biology, Physics, 54*(3), 642-646.

Trotti, A., Colevas, D., Sester, A., Rusch, V., Jaquwa, D., Budach, V., Langer, C., Murphy, B., Cumberlin, R., Coleman, C. N., & Rubin, P. (2003). CTCAE v3.0: Development of a comprehensive grading system for the adverse effects of cancer treatment. *Seminars in Radiation Oncology, 13*(3), 17-181. doi: 10.1016/S1053-4296(03)00031-6.

Van der Laan,, H. P., van den Bergh, A., Schilstra, C., Vlasman, R., Meertens, H., & Langendijk, J. A. (2008). Grading-system-dependent volume effects for late radiatation-induced rectal toxicity after curative radiotherapy for prostate cancer. *International Journal of Radiation Oncology Biology Physics, 70*(4), 1138-1145. doi: 10.1016/j.ijrobp.2007.07.2363

Van Stralen, D. (2008). High reliability organizations: Changing the culture of care in two medical units. *Design Issues*, *24*, (1).

Van Stralen, D., Calderon, R. M., Lewis, J. F., & Roberts, K. H. (2008). Changing a pediatric sub-acute facility to increase safety and reliability. *Patient Safety and Health Care Management. Advances in Health Care Management*, *7*, 259-282. doi: 10.1016/S1474-8231(08)07012-2.

Vogt, W. P. (2005). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences.* Thousand Oaks, CA: Sage Publications, Inc.

Wade, M. (2005). *Theories used in IS research: General systems theory.* Retrieved from http://www.istheory.yorku.ca/generalsystemstheory.htm

Watkins Bruner, D. (2007). Should patient-reported outcomes be mandatory for toxicity reporting in cancer clinical trials. *Journal of Clinical Oncology*, *25*(34), 5345-5347.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (1999). Organizing for high reliability: Processes of collective mindfulness. In R.S. Sutton and B.M. Staw (Eds.). *Research in Organizational Behavior*, pp. 81-123, Stanford, CA: Jai Press,.

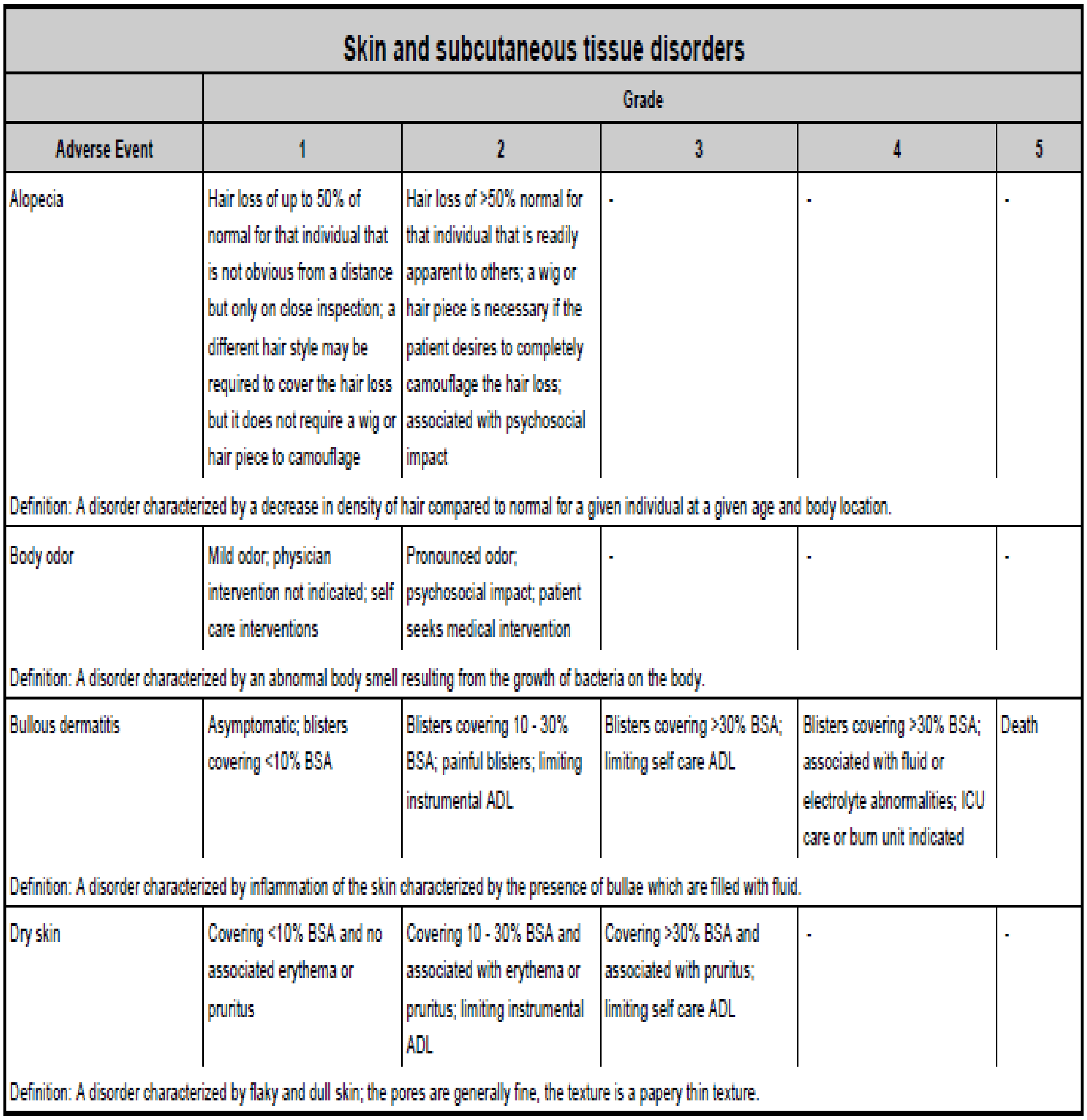
Weir, J. P. (2005). Quantifying test-retest reliability using intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research, 19*(1), 231-240.

Wells, M. & MacBride, S. (2003). Chapter 8: Radiation skin reactions. In Faithful, S. & MacBride, S. (eds.). *Supportive Care in Radiotherapy.* Retrieved from http://www.elsevier.ca/samplechapters/9780443064869/9780443064869.pdf

World Health Organization. (1979). *Handbook for Reporting Results of Cancer treatment.* Geneva, Switzerland. WHO Offset Publication No. 48. Retrieved from http://whqlibdoc.who.int/offset/WHO_OFFSET_48.pdf

Appendix A

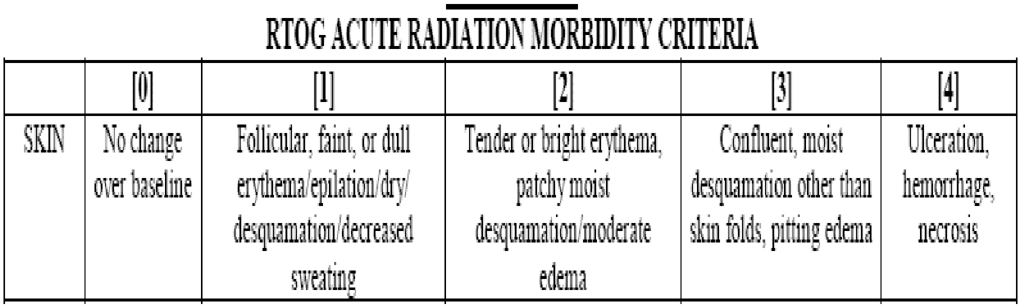
Excerpt of CTCAE scale for skin reactions



Skin and subcutaneous tissue disorders					
	Grade				
Adverse Event	1	2	3	4	5
Alopecia	Hair loss of up to 50% of normal for that individual that is not obvious from a distance but only on close inspection; a different hair style may be required to cover the hair loss but it does not require a wig or hair piece to camouflage	Hair loss of >50% normal for that individual that is readily apparent to others; a wig or hair piece is necessary if the patient desires to completely camouflage the hair loss; associated with psychosocial impact	-	-	-
Definition: A disorder characterized by a decrease in density of hair compared to normal for a given individual at a given age and body location.					
Body odor	Mild odor; physician intervention not indicated; self care interventions	Pronounced odor; psychosocial impact; patient seeks medical intervention	-	-	-
Definition: A disorder characterized by an abnormal body smell resulting from the growth of bacteria on the body.					
Bullous dermatitis	Asymptomatic; blisters covering <10% BSA	Blisters covering 10 - 30% BSA; painful blisters; limiting instrumental ADL	Blisters covering >30% BSA; limiting self care ADL	Blisters covering >30% BSA; associated with fluid or electrolyte abnormalities; ICU care or burn unit indicated	Death
Definition: A disorder characterized by inflammation of the skin characterized by the presence of bullae which are filled with fluid.					
Dry skin	Covering <10% BSA and no associated erythema or pruritus	Covering 10 - 30% BSA and associated with erythema or pruritus; limiting instrumental ADL	Covering >30% BSA and associated with pruritus; limiting self care ADL	-	-
Definition: A disorder characterized by flaky and dull skin; the pores are generally fine, the texture is a papery thin texture.					

Source: Common Terminology Criteria for Adverse Events Scale v. 4.0

Appendix B

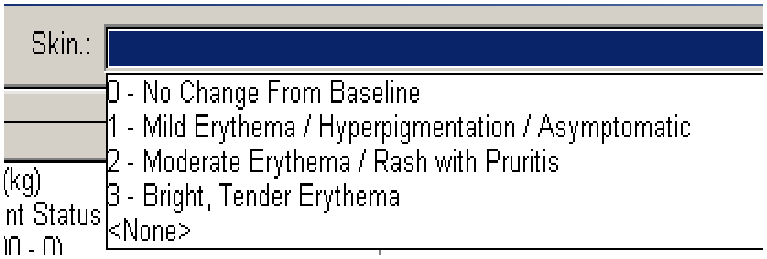


RTOG ACUTE RADIATION MORBIDITY CRITERIA

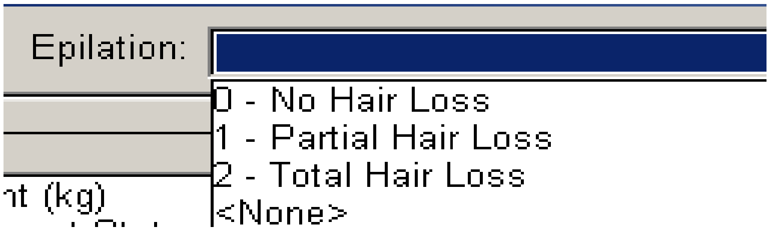
	[0]	[1]	[2]	[3]	[4]
SKIN	No change over baseline	Follicular, faint, or dull erythema/epilation/dry/ desquamation/decreased sweating	Tender or bright erythema, patchy moist desquamation/moderate edema	Confluent, moist desquamation other than skin folds, pitting edema	Ulceration, hemorrhage, necrosis

Appendix C

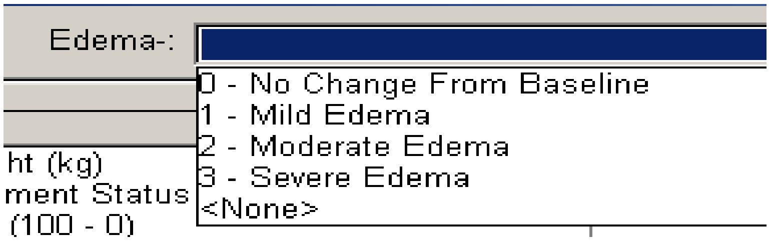
Internal Department Scale


Skin.:

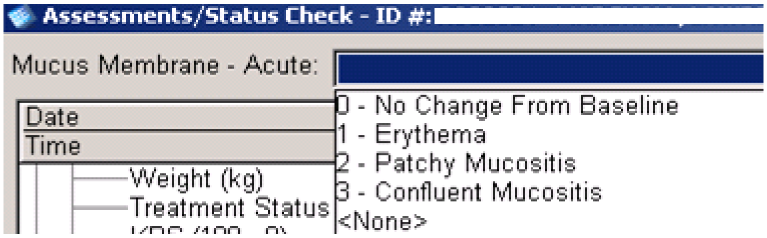
0 - No Change From Baseline
1 - Mild Erythema / Hyperpigmentation / Asymptomatic
2 - Moderate Erythema / Rash with Pruritis
3 - Bright, Tender Erythema
<None>


Epilation:

0 - No Hair Loss
1 - Partial Hair Loss
2 - Total Hair Loss
<None>

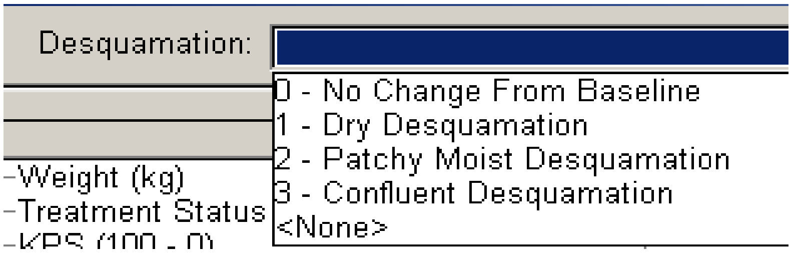

Edema-:

0 - No Change From Baseline
1 - Mild Edema
2 - Moderate Edema
3 - Severe Edema
<None>


Assessments/Status Check - ID #: [redacted]

Mucus Membrane - Acute:

0 - No Change From Baseline
1 - Erythema
2 - Patchy Mucositis
3 - Confluent Mucositis
<None>



Desquamation:

0 - No Change From Baseline
1 - Dry Desquamation
2 - Patchy Moist Desquamation
3 - Confluent Desquamation
<None>

-Weight (kg)
-Treatment Status
-KPS (100 - 0)

Appendix D

NCIC acute toxicity criteria for radiation dermatitis

Rating Symptom

0 None

1 Faint erythema or dry desquamation.

2 Moderate to brisk erythema. Patchy moist desquamation less than 1.5 cm mostly confined to skin folds and creases. Moderate edema.

3 Confluent moist desquamation greater than 1.5 cm not confined to skin. Pitting edema.

4 Skin necrosis or ulceration of full-thickness dermis may include bleeding—not induced by minor trauma or abrasion.