

**GEE MODELS FOR THE LONGITUDINAL ANALYSIS OF THE EFFECTS OF
OCCUPATIONAL RADIATION EXPOSURE ON LYMPHOCYTE COUNTS IN
RUSSIAN NUCLEAR WORKERS**

by

Adina Iulia Soaita

MD, Medical University Carol Davila, Bucharest, Romania, 1992

Submitted to the Graduate Faculty of
Graduate School of Public Health in partial fulfillment
of the requirements for the degree of
Doctor of Public Health

University of Pittsburgh

2006

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Adina Iulia Soaita

It was defended on

November 17, 2006

and approved by

Dissertation Adviser: Carol K. Redmond, ScD,
Distinguished Service Professor of Public Health, Department of Biostatistics,
Graduate School of Public Health, University of Pittsburgh

Dissertation Co-Adviser: Richard D. Day, PhD,
Associate Professor, Department of Biostatistics,
Graduate School of Public Health, University of Pittsburgh

Committee Member: Niel Wald, MD,
Emeritus Professor, Department of Environmental and Occupational Health,
Graduate School of Public Health, University of Pittsburgh

Committee Member: Ada O. Youk, PhD,
Assistant Professor, Department of Biostatistics,
Graduate School of Public Health, University of Pittsburgh

Committee Member: David M. Slaughter, PhD,
Research Professor, Nuclear Engineering Program
University of Utah

Copyright © by Adina Iulia Soaita

2006

Carol K. Redmond, ScD

**GEE MODELS FOR THE LONGITUDINAL ANALYSIS OF THE EFFECTS OF
OCCUPATIONAL RADIATION EXPOSURE ON LYMPHOCYTE COUNTS IN
RUSSIAN NUCLEAR WORKERS**

Adina Iulia Soaita, Dr.PH

University of Pittsburgh, 2006

The health effects of occupational radiation exposure have long been a source of scientific and administrative debates related to setting exposure standards. Relevant to this field are the effects of occupational long term radiation exposure on the lymphocyte counts which are especially sensitive to radiation. The trend of lymphocyte counts in radiation workers is of major importance since decreases in lymphocyte counts may be precursors of immunity disorders, cancer susceptibility or other chronic conditions. Another important question is whether the occupational radiation affects the lymphocyte counts similarly in males and females, given the relative lack of information on the effects and health implications of long term occupational radiation exposure on female subjects.

This dissertation presents a comprehensive statistical analysis of the relationship between dosimetric (yearly gamma exposure) and hematological (lymphocyte counts) data collected from a historical cohort (1948-1956) of highly exposed radiation workers at Mayak Plant Association located in Russia. The analysis controls for important covariates, such as the baseline lymphocyte counts, sex, work location related to Plutonium exposure lifestyle variables and the number of years from the first exposure. The analysis contrasts the most relevant radiation dose-response models by using marginal models and the GEE technique. STATA programming tools have been developed to check the assumptions required by the GEE technique, with special

attention to the missing data mechanisms and patterns in the framework of a longitudinal study with repeated measurements and unbalanced number of observations. The issue of non-linearity between the outcome variable and the explanatory covariates is addressed by the implementation of linear splines within GEE models.

Statistical analyses indicate: (a) that a linear radiation dose-response model is appropriate for the data, (b) a statistically significant negative relationship between the log-transformed lymphocyte counts and the log-transformed external gamma dose, (c) no statistically significant differences between males and females regarding the effect of occupational radiation exposure on the lymphocyte counts.

Public health significance of this research is:

- a) The linear radiation dose-response model is reasonable for regulatory purposes, and
- b) Males and females do not require differential regulatory standards for low dose occupational radiation exposure.

TABLE OF CONTENTS

AKNOWLEDGEMENTS	XVIII
1.0 INTRODUCTION.....	20
1.1 GENERAL FRAMEWORK.....	20
1.2 LITERATURE REVIEW	22
1.2.1 Radiation dose-response models.....	23
1.2.2 Radiation effects at chromosomal and cellular level	27
1.2.3 Radiosensitivity of the hematopoietic system.....	28
1.2.4 Mayak Product Association	32
1.2.5 Previous studies of long term occupational radiation exposure	35
1.2.6 Sex related differences in radiosensitivity	40
1.2.7 Review of repeated measures analysis techniques	41
1.2.7.1 Univariate analysis of variance (ANOVA)	42
1.2.7.2 Generalized Linear Models (GLM)	42
1.2.7.3 Linear mixed models	43
1.2.7.4 Marginal model and GEE (generalized estimating equations).....	44
1.2.7.5 Spline functions.....	48
1.2.7.6 Missing data	50
1.2.7.7 Methodological overview - summary	51

1.3	OBJECTIVES AND SPECIFIC AIMS	52
2.0	METHODOLOGY.....	54
2.1	BACKGROUND.....	54
2.2	MAYAK PA WORKERS' EARLY CLINICAL EFFECT DATABASE.....	55
2.2.1	Sampling procedures for MWECE database implementation	56
2.2.2	MWECE database format.....	56
2.2.3	Study cohort	58
2.3	STUDY DESIGN	60
2.3.1.1	Definition of study variables.....	61
2.3.1.2	Descriptive analysis	63
2.3.1.3	Statistical methods.....	64
2.3.1.4	Missing Data.....	72
3.0	RESULTS	75
3.1	DESCRIPTIVE ANALYSIS RESULTS.....	75
3.1.1	Study Cohort Structure.....	75
3.2	MISSING DATA ASSESSMENT RESULTS.....	104
3.2.1	Background	104
3.2.2	Descriptive analysis of the missing data patterns	106
3.2.3	Missing data patterns and mechanisms assessment	107
3.2.3.1	Drop-outs description.....	107
3.2.3.2	Drop-outs mechanism assessment.....	109
3.2.3.3	Mixed patterns missing data description.....	112
3.2.3.4	Mixed patterns missing data assessment	114

3.2.3.5	Missing data analysis conclusion.....	118
3.3	STATISTICAL MODELING RESULTS.....	119
3.3.1	Assessment of the coefficients in the models.....	124
3.3.2	Assessment of goodness of fit of the models	131
3.3.3	Numerical assessment of goodness of fit of the models	161
3.3.4	Predicting the lymphocyte counts using the five models.....	164
4.0	DISCUSSION	171
4.1	STRENGTHS OF THE ANALYSIS.....	185
4.2	LIMITATION OF THE ANALYSIS.....	186
5.0	CONCLUSIONS	189
5.1	PUBLIC HEALTH SIGNIFICANCE.....	190
	APPENDIX A - RADIATION PHYSICS	191
	APPENDIX B - MWECE DATABASE FORMAT	195
	APPENDIX C - MISSING DATA ASSESSMENT (STATA OUTPUT).....	198
	APPENDIX D - LYMPHOCYTE COUNTS RANGE IN HUMANS	203
	BIBLIOGRAPHY	204

LIST OF TABLES

Table 1: MWECE implementation: Number of Workers sampled in MWECE in Each Primary Diagnosis Category;.....	56
Table 2: Study Cohort Implementation: Frequency Distribution of Primary Diagnostic; Included Workers by Sex.....	59
Table 3: Study Cohort Implementation: Frequency Distribution of Primary Diagnostic; Excluded Workers by Sex.....	59
Table 4: Study Cohort Implementation; Frequency Distribution of Exclusion Criteria by Sex...	60
Table 5: Study cohort implementation; MWECE Frequency Distribution of Workers Included and Workers Not Included in the Study Cohort By Sex.....	60
Table 6: Absolute and Relative Frequency Distribution of Workers	75
Table 7: Age Distribution* at the Beginning of the Follow-up	76
Table 8: Absolute and Relative Frequency Distribution of Workers by Start of Employment Year and Sex.....	76
Table 9: Absolute and Relative Frequency Distribution of Workers by Year of First Non-Zero Non-Missing Exposure and Sex.....	77
Table 10: Absolute and Relative Frequency Distribution of Workers by Year of Blood Testing	77
Table 11: Departure from Normality of Yearly Median Lymphocyte Counts	79
Table 12: Males; Departure from Normality of Yearly Median Lymphocyte Counts	79

Table 13: Females; Departure from Normality of Yearly Median Lymphocyte Counts.....	79
Table 14: Distribution of Yearly Median Lymphocyte Counts (X1000/mm ³).....	81
Table 15: Males; Distribution of Yearly Median Lymphocyte Counts (X1000/mm ³).....	81
Table 16: Females; Distribution of Yearly Median Lymphocyte Counts (X1000/mm ³)	82
Table 17: Distribution of Yearly Log Transformed Median Lymphocyte Counts.....	84
Table 18: Males; Distribution of Yearly Log Transformed Median Lymphocyte Counts	84
Table 19: Females; Distribution of Yearly Log Transformed Median Lymphocyte Counts	85
Table 20: Distribution of Yearly Cumulative External Gamma Dose.....	88
Table 21: Males; Distribution of Yearly Cumulative External Gamma Dose.....	88
Table 22: Females; Distribution of Yearly Cumulative External Gamma Dose	88
Table 23: Distribution of Log Transformed Yearly Cumulative External Gamma Dose.....	89
Table 24: Males; Distribution of Log Transformed Yearly Cumulative External Gamma Dose.	90
Table 25: Females; Distribution of Log Transformed Yearly Cumulative External Gamma Dose	90
Table 26: Absolute and Relative Frequency Distribution of Workers by Smoking History	93
Table 27: Absolute and Relative Frequency Distribution of Workers by Alcohol Consumption	93
Table 28: Distribution of Two Groups of Lymphocyte Counts in the same 119 Workers:.....	95
Table 29: Distribution of Baseline* Lymphocyte Counts by Sex	97
Table 30: Distribution of Baseline* Lymphocyte Counts by Smoking History.....	97
Table 31: Males; Distribution of Baseline* Lymphocyte Counts by Smoking History	97
Table 32: Females; Distribution of Baseline* Lymphocyte Counts by Smoking History	98
Table 33: Distribution of Baseline* Lymphocyte Counts by Alcohol Consumption.....	98
Table 34: Males; Distribution of Baseline* Lymphocyte Counts by Alcohol Consumption.....	98

Table 35: Females; Distribution of Baseline* Lymphocyte Counts by Alcohol Consumption ...	99
Table 36: Frequency Distribution of Workers by Number of Changes of Work Location	99
Table 37: Absolute and Relative Frequency Distribution of Workers by Work Location	100
Table 38: Males; Study Cohort: Absolute and Relative Frequency Distribution of Workers by Work Location	100
Table 39: Females Absolute and Relative Frequency Distribution of Workers by Work Location	101
Table 40: Absolute and Relative Frequency Distribution of Workers by Work Location Related to Pu exposure and Years following the First External Gamma Exposure.....	101
Table 41: Males; Absolute and Relative Frequency Distribution of Workers by Work Location Related to Pu exposure and Years following the First External Gamma Exposure	102
Table 42: Females; Absolute and Relative Frequency Distribution of Workers by Work Location Related to Pu exposure and Years following the First External Gamma Exposure	102
Table 43: Distribution of Baseline* Lymphocyte Counts by Work Location Related to Pu exposure during the first year of follow-up	103
Table 44: Males; Distribution of Baseline* Lymphocyte Counts by Work Location Related to Pu exposure during the first year of follow-up	103
Table 45: Females; Distribution of Baseline* Lymphocyte Counts by Work Location Related to Pu exposure during the first year of follow-up;	103
Table 46: Absolute and Relative Frequency Distribution of the Missing Data Patterns	106
Table 47: Males; Absolute and Relative Frequency Distribution of the Missing Data Patterns	106
Table 48: Females; Absolute and Relative Frequency Distribution of the Missing Data Patterns	107

Table 49: Drop-Outs Only: Absolute and Relative Distribution of Drop-out Patterns by Year of Follow-up when the Drop-out Occurs	108
Table 50: Males; Drop-Outs Only: Absolute and Relative Distribution of Drop-out Patterns by Year of Follow-up when the Drop-out Occurs	108
Table 51: Females; Drop-Outs Only: Absolute and Relative Distribution of Drop-out Patterns by Year of Follow-up when the Drop-out Occurs	108
Table 52: Mixed Missing Data Patterns: Absolute and Relative Frequency Distribution of Observations by Missing Data Patterns and Sex	113
Table 53: Total; β -Coefficients, Standard Errors, and p-values calculated by using GEE Method for the Six Models.....	126
Table 54: Males: β -Coefficients. Semi-Robust Standard Errors, and p-values Calculated for the Six Models	128
Table 55: Females: Coefficients. Semi-Robust Standard Errors, and p-values Calculated for the Six Models	130
Table 56: Total; Goodness of Fit Assessment (Residuals distribution, $GEE-R^2$, QICu).....	162
Table 57: Males; Goodness of Fit Assessment (Residuals distribution, $GEE-R^2$, QICu)	163
Table 58: Females; Goodness of Fit Assessment (Residuals distribution, $GEE-R^2$, QICu)	163
Table 59: Expected Values of the Lymphocyte Counts.....	169
Table 60. Approximate LET* and RBE** for Different Types of Ionizing Radiations.....	192
Table 61: MWECE Database Format - Lifestyle Variables Definition and Coding	195
Table 62: MWECE Database Format - Hematological Variables Definition and Coding	196
Table 63: MWECE Database Format – Work Location Variables Definition and Coding.....	197

LIST OF FIGURES

Figure 1. Models of Radiation Dose Response Relationship:	25
Figure 2: MWECE Data Structure	57
Figure 3: Distribution of the Log-Transformed Total Number of Lymphocyte Counts Performed Yearly in Each Worker	63
Figure 4: Study Cohort: Yearly Mean of Log Transformed Lymphocyte Counts versus Log Transformed Median of Lymphocyte Counts; 10 Years of Follow-up Starting during the Year of First Non-Zero Non-Missing Exposure	80
Figure 5: Distribution of Yearly Median Lymphocyte Counts Distribution	82
Figure 6: Males; Distribution of Yearly Median Lymphocyte Counts Distribution	83
Figure 7: Females; Distribution of Yearly Median Lymphocyte Counts Distribution.....	83
Figure 8: Distribution of Log of Yearly Median Lymphocyte Counts Distribution.....	86
Figure 9: Distribution of Log of Yearly Median Lymphocyte Counts Distribution.....	86
Figure 10: Distribution of Log of Yearly Median Lymphocyte Counts Distribution.....	87
Figure 11: Median Lymphocyte Counts and Median Cumulative External Gamma Dose	91
Figure 12: Males; Median Lymphocyte Counts and Median Cumulative External Gamma Dose	92

Figure 13: Females; Median Lymphocyte Counts and Median Cumulative External Gamma Dose	92
Figure 14: Relative Distribution of Two Groups of Lymphocyte Counts Recorded in the same 119 Workers:.....	95
Figure 15: Distribution of Two Groups of Lymphocyte Counts Recorded	96
Figure 16: Workers from Study Cohort Who Drop Out: Histogram of the Relative Frequency of Drop-Out Patterns by Sex; Drop Out Patterns are Categorized According to the Number of Follow-up Year	109
Figure 17: Total, Model#1; Histogram of Residuals' Distribution.....	133
Figure 18: Total, Model#2 Histogram of Residuals' Distribution.....	133
Figure 19: total, Model#3; Histogram of Residuals' Distribution	134
Figure 20: Total, Model #4; Histogram of Residuals' Distribution.....	134
Figure 21: Total, Model#5; Histogram of Residuals' Distribution.....	135
Figure 22: Total; Model#6; Histogram of Residuals' Distribution	135
Figure 23: Males; Model#1; Histogram of Residuals' Distribution	136
Figure 24: Males; Model#3; Histogram of Residuals' Distribution	137
Figure 25: Males; Model#4; Histogram of Residuals' Distribution	137
Figure 26: Males; Model#5; Histogram of Residuals' Distribution	138
Figure 27: Males; Model #6; Histogram of Residuals' Distribution	138
Figure 28: Males; Model #6; Histogram of Residuals' Distribution	139
Figure 29 Females; Model#3; Histogram of Residuals' Distribution.....	140
Figure 30: Females; Model#4; Histogram of Residuals' Distribution.....	140
Figure 31: Females; Model#5; Histogram of Residuals' Distribution.....	141

Figure 32: Females; Model#6; Histogram of Residuals' Distribution.....	141
Figure 33: Total; Model#1; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years since the First External Gamma Exposure ..	142
Figure 34: Total; Model#2; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years since the First External Gamma Exposure ..	143
Figure 35: Total; Model #3; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure..	143
Figure 36: Total; Model#4; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure..	144
Figure 37: Total; Model#5; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure..	144
Figure 38: Total; Model#6; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure..	145
Figure 39: Males; Model#1; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure..	146
Figure 40: Males; Model#3; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure..	146
Figure 41: Males: Model#4; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure..	147
Figure 42: Males; Model#5; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative.....	147
Figure 43: Males; Model#6; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure..	148

Figure 44: Females; Model#1; Scatter Plot of Residuals versus the Log-Transformed Yearly .	149
Figure 45: Females; Model#3; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure..	149
Figure 46: Females; Model#4; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure..	150
Figure 47: Females; Model #5; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure..	150
Figure 48: Females; Model#6; Scatter Plot of Residuals versus the Log-Transformed YearlyCumulative External Gamma Exposure by Years Since the First External Gamma Exposure	151
Figure 49: Total; Model#1; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	152
Figure 50: Total; Model#2; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	152
Figure 51: Total; Model#3; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	153
Figure 52: Total; Model#4; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	153
Figure 53: Total; Model#5; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	154
Figure 54: Total; Model#6; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	154

Figure 55: Males; Model#1; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	155
Figure 56: Males; Model#3; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	156
Figure 57: Males; Model#4; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	156
Figure 58: Males; Model#5; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	157
Figure 59: Males; Model#6; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	157
Figure 60: Females; Model#1; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	158
Figure 61: Females: Model#3; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	159
Figure 62: Females; Model#4; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	159
Figure 63: Females; Model#5; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	160
Figure 64: Females: Model#6; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure.....	160

ACKNOWLEDGEMENTS

I would like to acknowledge my advisors Dr. Carol Redmond and Dr. Richard Day for their continuous support of my work on this dissertation. Their expertise and patience were very valuable for me while working on this doctoral thesis. They helped me so much that it is impossible for me to describe it in a few words. My gratitude for them is in my heart and will stay there forever.

I am very grateful for the wonderful help I received from the committee members: Dr. Niel Wald, Dr. Ada Youk, and Dr. Mike Slaughter during the years I had to work for my doctoral degree. They were very approachable and always willing to advice and support me. I also want to acknowledge Dr. Tamara Azizova from Southern Urals Biophysics Institute, Ozyorsk, Russia for the Mayak PA workers database that made possible the accomplishment of this dissertation research project. I also received invaluable financial support from Duquesne Light Company through the John Arthur fellowship awarded by the Department of Environment and Occupational Health chaired by Dr. Bruce Pitt. I would like to use this opportunity to express my gratitude to Dr. Bruce Pitt and the EOH Department. My work was also supported by the CDC/NIOSH grant R01-OH007866.

I feel I would never have finished the writing process of my dissertation without the lovely help I have received from my friend Dr. Alina Bodea in editing my thesis. She had a wonderful capability to understand what I was trying to write and to laugh with me and not at me

while reading over my draft. I also have to mention that I would never have been in the position of defending a doctoral dissertation without my friends, Ana Maria Iosif and Diana Luca who showed patience and love while tutoring me in statistical theory in preparation for the qualifying exams.

I would like to mention my Romanian friends who are here in the US or back home in Romania. I always felt they all supported me although they sometimes did not understand me. I am not mentioning their names but I strongly hope they feel my love and gratitude. I want to show from all my heart my warm feelings and gratitude to my American friends who made me and my family feel so welcome here in the United States. I want to say a special thank you to my office colleague Dr. Mike Kuniak who was always friendly and supportive to me. God bless all of you and may we be able to get together and enjoy each other in the future.

Last but not least, I would like to reinforce my gratitude and my love for my family, my husband Cornel and my daughter Ioana who showed me love, devotion and support although associated with criticism. I am very happy we were able to go together through these five years and to achieve so many things while being far away from home and facing so many difficulties. I would like to dedicate my work to my parents who always believed in me, in my skills and in my survivor abilities.

1.0 INTRODUCTION

1.1 GENERAL FRAMEWORK

The health effects of occupational radiation exposures, including long term exposure effects and risks at low doses have been a long-standing source of scientific, political, and administrative concern, especially with respect to setting health and safety standards. Particularly relevant to this area are the effects of occupational long term radiation exposure on the hematological system. Despite the critical importance of this issue, data on the overall and specific deterministic effects of occupational long term exposure to radiation¹ are limited. Most of the existing knowledge has been derived from: (a) studies on populations receiving instantaneous exposures from atomic bombs; (b) data about biological reactions after protracted high dose radiation therapy of cancer patients; and (c) studies on laboratory animals². However, each of these sources presents important limitations when used to estimate the health risks of radiation on humans in an occupational environment. The most salient in this regard is the difficulty of extrapolating findings regarding the effect of radiation exposure of differing doses and circumstances to the characteristics of long term occupational exposure. Occupational radiation exposure usually consists of long term relatively low doses applied to subpopulations with a good health status at the beginning of exposure. The complexities of this effect and, generally, the specific circumstances related to occupational exposure require a direct investigation of these

phenomena and special interpretation of the results. The major contribution of this research is that it focuses directly on evaluating in workers the effects on the hematological system of long term occupational radiation exposure.

It is worthwhile to mention that the majority of occupational studies focused on cancer incidence and cancer mortality as the main outcome while the effects of radiation on blood cell counts is were little analyzed. Moreover, there are no studies that use the lymphocyte counts as a biological marker of the occupational effect of radiation exposure on large cohorts of workers. This study is unique since it assesses the effects of long term occupational radiation exposure on the lymphocyte counts regardless the occurrence of a specific disease. The follow-up of the lymphocyte counts trend is very important since the lymphocyte counts drop may lead to immunity disorders, cancer susceptibility or other chronic conditions³⁻⁵.

One major limitation of the existing studies is the under-representation of female subjects in the analyses of occupational effect of radiation exposure. The number of female radiation workers available for epidemiologic studies has been small, even when females are included in relevant studies, their radiation doses are usually below recognized threshold levels.

Given this relative lack of information on the effects and health implications of long term occupational radiation exposure on female subjects, one objective of this research is to investigate the hematological effects of long term occupational radiation exposure and its possible differential effects in males and females.

It is acknowledged that longitudinal data on occupational cohorts requires specific and adapted statistical analysis techniques in order to obtain meaningful results. In most instances, the statistical analysis of occupational data collected longitudinally, encompassing a long period of time must accommodate a complex structure of the data and different patterns of missing data.

There are difficulties related to model building procedures and goodness of fit evaluation in longitudinal data. The current dissertation utilizes a longitudinal data analysis with repeated measures design based on a modern estimation technique known as the GEE (generalized estimating equations) to a subset of the cohort data collected at the Mayak PA nuclear plant located in Russia. Although the GEE technique is widely used in epidemiological and clinical trial studies, there are still methodological gaps regarding model building and model goodness of fit assessment. This study will attempt to fill the existing methodological gaps regarding the application of model building procedures, goodness of fit assessment and missing data testing in occupational cohort studies. Moreover, the use of linear splines appears not to have been previously incorporated into the GEE models for evaluating the linearity of effects at low doses in radiation exposed cohort studies.

1.2 LITERATURE REVIEW

In order to understand the relevance and specific contribution of the current study, critical considerations pertaining to the topic of the hematological effect of long term occupational radiation exposure include knowledge of: 1) radiation dose-response models, 2) the radiation effects at chromosomal and cellular level, 3) the sensitivity of the hematopoietic system, 4) details of the Mayak PA cohort data, and 5) the statistical techniques used for longitudinal data analysis. (In order to assist the reader in understanding the sections that follow a summary of key concepts, terms and definitions of radiation physics and dosimetry is provided in Appendix A.)

1.2.1 Radiation dose-response models

The study of radiation effects in humans has led to many conflicting findings and controversies. In order to respond to these scientific controversies, a number of different explanatory models and theories have been proposed that focus on the exploration of dose-response phenomena. Historically, it was considered that radiation exposure leads quasi-exclusively to cancer. However, the advent of cellular biology revealed the complexity of the mechanisms relating radiation exposure to health outcomes. Discussion of some of the current dose-response radiation models considered relevant for occupational exposure to radiation will follow.

Generally, the effects of occupational radiation exposure have been categorized as stochastic effects and deterministic effects.

The stochastic effect is related to cancer induction. The magnitude of a given stochastic effect is not dose dependent, although the probability of the effect occurring is often proportional to the dose, especially at low doses.

The deterministic effect model proposes the existence of a critical threshold of radiation exposure. For doses below a specific threshold the effects may be considered negligible, while for doses above the threshold the higher the dose, the higher the severity of the radiation effect³.

In addition to the two general models mentioned above, there are several theories describing the radiation dose-response relationship. Currently, the dose-response models referred to most frequently in the literature are the linear non-threshold model, and the non-linear threshold models. The discussion surrounding these different theories shows that the reaction of humans to long term radiation is not yet well-understood, especially in terms of long term low dose effects.

Radiation exposure may have different effects on various parts of the human body, from stimulatory beneficial effects, to damaging ones^{6,7}. Moreover some long term moderate dose exposures may influence the reaction to additional high rate of radiation exposure or even to some chemicals exposures^{6,8-10}.

The simplest theory regarding the radiation dose-response effect, largely accepted and considered a reference for setting protection regulations is the linear non-threshold model (LNT). It states that the effect of low-dose, especially tumor incidence, can be estimated by linear extrapolation. The implication is that there is no safe dose and that even low doses can produce harmful biological effects. The low dose radiation effect is as harmful per gray as the high dose effect¹¹. The dose-response relationship is similar at low radiation doses as well as at high radiation doses.

A more complex theory, the non-linear threshold dose-response curve is often discussed in the recent literature¹². According to this model, there are different effects for different dose ranges. At low radiation doses, there is low or no radiation effect. At high radiation doses, the higher radiation doses become, the stronger the effect. Beyond a saturation radiation level, additional radiation doses increase effects only weakly or not at all. Figure1 illustrates the radiation models described above.

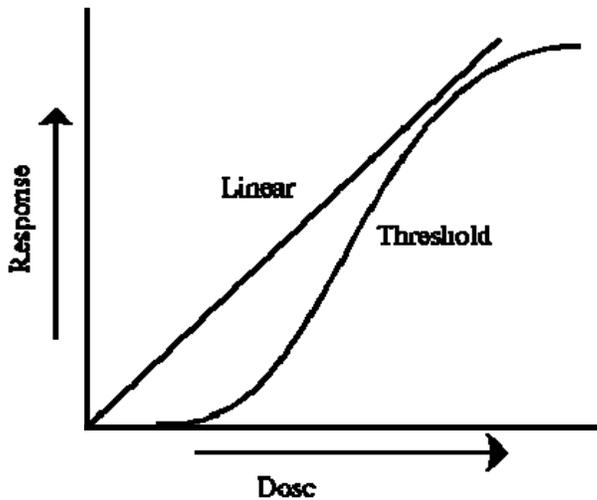


Figure 1. Models of Radiation Dose Response Relationship:
Linear Non-Threshold Model and Curvilinear Threshold Model.

Source: http://www.rstp.uwaterloo.ca/manual/bio_effects/response/dose_response.htm¹³

Since the non-linear radiation dose-response relationship is very complex, more sub-models are presented in order to assess and mathematically describe the radiation effects.

The *hormesis* phenomenon is described as a non-linear radiation dose-response model. Hormesis consists of a stimulatory response at low doses and an inhibitory response at higher doses, thereby assuming a beneficial low-dose effect¹.

The *adaptive response* has been also described as a non-linear radiation dose-response model. It is defined as an increased resistance to relatively high radiation dose after low dose pre-exposures. According to the adaptive response model, when cells are exposed to a low-adapting dose (e.g. 1 cGy=1 rad) and later on to a higher challenging dose (e.g. 1Gy=100 rads,) the exposure effect is less than the individual effect of a single challenging dose¹⁴⁻¹⁶. The adaptive response is also described as cross-adaptation phenomenon in which the exposure of cells to low doses of radioactive or chemical agents could lead to a decrease in cells' sensitivity to the same or other agents¹⁷.

¹ Doses of approximately 1cGy=0.01Gy are considered low doses and doses of approximately 1Gy are considered high doses.

The adaptive response and hormesis have sometimes been used interchangeably although they are not synonymous. The main difference between them is that the adaptive response needs large radiation doses for the challenge-dose, while hormesis as a special expression of the adaptive response, does not need any challenge-dose. The main common features are that they are both induced at low radiation priming doses, and the priming dose is observed as a hormetic dose^{15,18-20}.

There continues to be debate among scientists regarding the mechanisms and the appropriateness of different radiation dose-response models²¹. Many controversies around this topic have remained unresolved to date and it is suggested that a better understanding of the biological mechanisms is needed to reassess the use of these operational definitions²².

In these respects it is worthwhile to mention that the analysis performed in this dissertation focuses on the above described radiation dose-response models applied to a subset of Mayak PA occupational data. It is the first occupational study that tests the most popular theoretical radiation dose-response models using the valuable database of Mayak PA workers. Moreover, the analytic approaches employed offer the opportunity to contrast the above described theoretical radiation dose-response models applied to Mayak PA workers and to select that model which best explains the Mayak PA data. This dissertation research implements complex statistical models that correspond to the above described radiation dose-response models. The analysis of these complex statistical models provides results and scientific basis for a better understanding of occupational radiation effects.

1.2.2 Radiation effects at chromosomal and cellular level

In order to assess the effects and the health risks of occupational radiation exposure on humans it is important to understand the effects of radiation at the cellular level. The radiation response is assessed using endpoints considered most relevant for the health outcome of interest in this study, the lymphocyte counts. According to the literature reviewed, the main biological endpoints for the assessment of radiation effects are at the chromosomal and cellular level. At the chromosomal level, the radiation effects that have been described consist of: the formation of micronuclei (MN), dicentrics, rings, chromatid and isochromatid deletions such as breaks and gaps. At the cellular level the main radiation effects consist of cell killing²³.

It has been shown that low doses of radiation which do not yield detectable changes at the cellular level may produce alterations within cells that can influence the damage produced by subsequent exposures⁸. Human leucocytes exposed to low dose radiation (0.01Gy=1 rad) may become less sensitive to subsequent higher doses based on the DNA double strand break as the endpoint analyzed^{7,24,25}. This experiment is consistent with the adaptive response radiation dose-response model which was described above.

The sensitivity to further higher radiation dose in populations chronically exposed to radiation has been assessed by Russian authors who have concluded that the level of damaged lymphocytes assessed by micronuclei (MN) scoring is not higher in inhabitants of radiation contaminated areas than in inhabitants of non-contaminated areas²⁶. Interestingly however, they use a different definition of the adaptive response model than western authors. The Russian authors define the adaptive response model as an increase of the lymphocyte counts in people chronically exposed to radiation²⁶. They also have concluded that people living in contaminated

areas were less sensitive to further acute radiation exposure compared to those living in non-contaminated areas²⁶.

Another Russian source²⁷ describes a moderate increase in platelet and leucocytes counts during the first years of work in 432 individuals who worked at the first Russian plant treating nuclear fuel waste²⁷.

Thus, there are scientific disagreements surrounding the issues of chromosomal and cellular effects of radiation. However, it is notable that the hematopoietic system is useful for assessing both the chromosomal and cellular effects of radiation exposure.

Most of the existing studies on the cellular effects of radiation exposure are limited since the radiation exposure occurred “*in vitro*”, thereby being a short term radiation exposure applied to human cells outside the human body. This dissertation research it is a statistical analysis based on “*in vivo*” data. The endpoint of this analysis is the peripheral blood lymphocyte counts.

This dissertation presents the first study that assesses the occupational deterministic radiation effect on the lymphocyte counts in both sexes. It is notable that the analysis is based on the lymphocyte count data recorded in workers during their exposure to radiation doses that are much higher than those currently accepted in the United States.

1.2.3 Radiosensitivity of the hematopoietic system

In the literature on radiation effects in humans, most of the publications focus on the blood cells. Consistent with this orientation, the current study gives special attention to the hematopoietic system.

An important radiobiological principle stated by Bergonie and Tribondeau in 1906 is that cells²⁸ that normally divide often are more radiosensitive. One of the human organs well known

for its cell renewal capacities is the hematological system, comprising the central and peripheral components.

In adults the hematopoietic tissue is located primarily in the bone marrow, a tissue with some unique features that make it sensitive to radiation. The hematopoietic system has the following characteristics that explain its sensitivity to radiation:

- 1) It is distributed essentially throughout the body mass and therefore is an easy target to hit and
- 2) It provides a steady supply of mature blood cells which accomplish the missions of the hematopoietic system and therefore it is a place of continuous cell division.

However, the bone marrow's anatomic subunits act independently and, in the case of local damage, a less affected part of the bone marrow may continue to provide mature blood cells into the blood stream. This mechanism attempts to compensate to some extent for the high sensitivity of the bone marrow²⁸ to destructive influences.

The hematological system consists mainly of the hematopoietic tissues and the circulating blood cells. Taking into account the difficulties related to the investigation of the hematological reaction at a central level, the peripheral blood examination is more feasible and also informative for assessing radiation effects. The mature blood cells, part of the blood stream, react differently to radiation exposure since they vary in morphology and dynamics.

A review of the characteristics of the main cell lines which are components of the peripheral blood reveals the following:

The erythrocytes are resistant to radiation because of the lack of a nucleus and the resultant incapacity for division. Therefore the erythrocyte counts drop as a result of bone marrow damage rather than erythrocyte damage following high exposures.

The thrombocytes themselves, as well as the mature megakaryocytes, their precursors, are also radiation resistant. The platelet does not divide because of the absence of a nucleus. However, thrombocytopenia is a result of radiation exposure effect at the bone marrow level where the immature megakaryocytes have a high division capacity in order to provide platelets for the blood stream. One feature of induced radiation thrombocytopenia in humans is its late occurrence. Thrombocytopenia occurs 10 days or more from the radiation exposure. Late occurrence is due in part to the circulating platelets' survival time (9-10days). The thrombocyte counts usual recover in 6-8 weeks.

The mature neutrophile develops from the precursor pool of stem cells in the bone marrow via myeloblasts, promyelocytes, myelocytes and metamyelocytes. The total transit time in the bone marrow is about 10 days. A characteristic of the neutrophil is its nuclear lobulation. The nuclear lobulation purpose and mechanism are still not very well known⁵. Radiation effects on the neutrophiles result from disturbance at the bone marrow level. The sensitivity of the neutrophiles in the circulating blood can be excluded because there is no evidence for direct cell death in non-dividing granulocytic cells. Radiation effects on the neutrophiles can be categorized in a lag phase, depletion phase and recovery.

The lymphocytes. According to the literature reviewed, the lymphocytes are the most sensitive blood cells to radiation, given their morphological and functional characteristics. Their radio-sensitivity comes from their special structure and physiology. Unlike the red cells, they have a nucleus. Moreover, the nucleus does not become smaller over time. Therefore, lymphocytes never lose their ability to divide. In the blood stream there are several types of lymphocytes, and at least three maturation levels of the lymphocytes have been identified: virgin cells, memory cells and activated cells, depending upon contact with an antigen. Lymphocytes

have different life spans; whereas virgin cells and memory cells can live for months or even years, activated cells die by programming cell death within days⁴. There is no difference in radio-sensitivity between young and old lymphocytes²⁸.

Radiation damage occurs first in the nucleus and, as the radiation dose increases, the damage occurs in the cytoplasm. The first change, which occur immediately after exposure, is the appearance of vacuoles within the nucleus that may be followed by significant cell destruction. Inside the nucleus, damage occurs at the DNA level: in the form of single and double DNA breaks and base damage. The study of poly (ADP-ribose) polymerase as a mediator which is released in response to double strand breaks has led some scientists to believe that double strand breaks might be illustrative for radiation damage^{7,17}.

In view their radio-sensitivity, the lymphocytes have been described as a useful indicator of radiation injury. They have also demonstrated their special capacity of “remembering” by showing chromosomal aberrations and translocations many years after the acute exposure²⁹. Thus, due to their special features, the lymphocytes are an appropriate biological marker for the study of radiation response^{4,5,28}.

In summary, the lymphocytes in the peripheral blood cells as well as the lymphopoietic tissues are sensitive to radiation. Therefore a reduction in lymphocyte count reflects the damage at central and peripheral levels as a result of whole body irradiation. The peripheral blood lymphocyte counts provide valuable information about bone marrow functionality, which would otherwise examined by more invasive techniques such as bone marrow biopsy.

In terms of time, the lymphocyte counts are clearly affected shortly after radiation exposure. A reduction in lymphocyte counts occurs within hours after exposure to radiation, thereby making them a suitable marker for assessing a cause-effect relationship. The recovery of

the lymphocyte counts may take a few months. Dienstbier and Hempelmann have shown that the lymphocyte counts' decline after irradiation is a sensitive index of radiation injury²⁸.

The lymphocyte counts can be performed repeatedly in humans since peripheral blood testing is a minimal invasive laboratory procedure. This aspect makes it possible to follow-up the counts and to assess trends.

Thus, due to all these characteristics, the peripheral blood lymphocytes are considered the most appropriate indicator cells that can be used to evaluate radiation effects. In these respects this dissertation is among the first to use peripheral lymphocyte counts for a 10 year follow up of workers employed at Mayak. PA. The lymphocyte count data are associated with the external gamma doses data applying comprehensive statistical models in order to determine the effects of occupational radiation exposure.

This analysis is feasible since in Mayak PA workers there are data available from blood tests which allow estimation of the lymphocyte counts. Blood tests were performed at start of employment and subsequently about three times each year. Cumulative yearly radiation doses were recorded in each worker. The database structure and the selection of the data subset used in this dissertation are described in further detail in the methodology section.

1.2.4 Mayak Product Association

Peripheral blood determinations, including information on lymphocyte counts, have been recorded for workers at Mayak PA. There are many other important characteristics that make the Mayak PA population a valuable and unique source of occupational data.

The Mayak PA nuclear plant started to function about 60 years ago and included among employees thousands of males and females who were exposed for the first 10 years of operation to radiation doses much higher than those allowed currently in the US.

The facility started its operations on January 1, 1948. It was the first nuclear production site in the former Soviet Union and included a reactor plant, a radio-chemical plant and a plutonium processing plant. In the first decade of its existence (1948-1958), inexperience with production techniques, combined with an emphasis on urgent political priorities, resulted in at least 8000 workers receiving high levels (1-10 Gy=100-1000 rads) cumulative exposures of external gamma radiation. Some workers also received acute accidental gamma exposure or internal alpha radiation from inhaled plutonium aerosols.

Medical examinations were carried out on more than 95% of these workers as part of the Radiation Protection Program. During these examinations, routine peripheral blood counts were carried out and periodic bone marrow samples were drawn. In the case of workers having radiation-related diagnoses or very high exposures, additional clinical studies such as cytogenetics, pulmonary function tests and other procedures were carried out.

The study records contain important information on workers' medical conditions that are related to radiation exposures. The radiation related diagnoses of Mayak PA workers have been categorized as follows: Acute Radiation Sickness (ARS), Chronic Radiation Sickness (CRS) Plutonium Pneumosclerosis (PPn) or No Radiation Related Diagnosis. The diagnoses were established according to the diagnostic criteria of Guskova and Boysogolov³⁰ which take into account the occurrence, duration and level of radiation exposure along with the clinical and laboratory results.

The most important signs for Acute Radiation Sickness (ARS) are summarized as follows: nausea, vomiting, increased neutrophils and lymphopenia. These clinical signs are considered non-specific since they are observed in a large number of diseases.

The concept of Chronic Radiation Sickness (CRS) is characterized by a large number of clinical signs and laboratory results involving multiple target systems, including the cardiovascular, nervous and digestive systems³⁰. Chronic Radiation Sickness is described as a combination of various clinical signs including: lymphopenia, thrombocytopenia, neutropenia, and asthenia. Interestingly, Chronic Radiation Sickness (CRS) has been described only by Russian authors thereby being an almost unknown clinical entity in Western countries. Due to the large number of target organs involved, and to the large variety of clinical signs, Chronic Radiation Sickness consists of a group of highly non-specific characteristics.

Plutonium Pneumosclerosis was a diagnosis given to workers who were exposed to Plutonium who also had respiratory problems associated with specific X-ray images.

Workers were diagnosed as having no radiation related diseases when there was an absence of clinical signs associated with low dose radiation exposure or absence of radiation exposure.

The existing records on Mayak PA Workers, starting with the year 1948 and continuing through the present, served as source for the design and implementation of an electronic database referred to as the Mayak PA Workers' Early Clinical Effect (MWECE) that continues to be updated on an ongoing basis. A systematic sampling method was used in order to select a representative sample of the workers for the MWECE database. The sampling procedures used for the MWECE database implementation and the structure of MWECE database are described in detail in the methodology section. The electronic database contains selected demographic

characteristics, work history, occupational exposures and clinical information on a representative sample of 591 Mayak PA workers including 361 males and 230 females hired between 1948 and 1962³¹.

The large amount of data on males and females chronically exposed at relatively high radiation doses makes the MWECE database unique and very valuable for research on the health effects of chronic radiation exposure. The MWECE database is probably the most complete source of information on a large number of males and females working at a nuclear industrial facility, providing this study with the relevant and important data for exploring hematological responses in both sexes.

In this dissertation the statistical analysis is performed on a subset of the MWECE database. The process used to select this subset of data from the MWECE database is presented in the methodology section.

This dissertation represents the first study performed in a Western country which focuses on the comprehensive data recorded in Mayak PA workers and provides a detailed description of the available data.

1.2.5 Previous studies of long term occupational radiation exposure

Consistent with the focus of this research, the literature review focuses on the occupational studies of long term radiation exposure, as well as studies that have provided information on potential differences by sex in radiation sensitivity.

A study by Court-Brown and Doll in 1956, analyzed occupational radiation effects among British radiologists³². This study examines mortality from cancer and other causes over forty years of observation. It included an initial cohort of 1338 male radiologists that was expanded

later by adding 1352 male radiologists. The results show an increased risk of dying from cancer only in the radiologists registered before 1921. Those radiologists who registered after 1921 did not have increased cancer mortality. Regarding mortality from non-cancer causes, there was no evidence of any radiation effects even among early registered radiologists.

A limitation of this study is that it does not include any women in the analysis and there is no quantitative dose information available in the data base. Rough dose categories have been taken into account by Smith and Doll by subdividing the cohort into three subgroups according to their year of first registration. The categories were chosen in order to approximately correspond to high, medium and low levels of exposure.

A more informative and statistically powerful source of occupational data is the nuclear shipyard workers study (NSWS) which is the largest study on health effects of occupational low dose rate and the only radiation study where nuclear workers were compared to age matched and job matched unexposed workers as controls. The search for health risks was performed on civilian employees of eight shipyards that repaired the nuclear propelled U.S. Navy ships and submarines. The workers were followed up for about 20 years. The NSWS compared a high dose cohort of 27,872 nuclear workers exposed to more than 0.5 rads cumulatively to 32,510 unexposed controls shipyard workers who had the same ages and jobs. Dose assessment was considered very accurate because of a strict policy regarding the use of monitoring badges. The common shipyard doses were 0.5-22.5 mGy/y (0.05-2.25 rads/year) and the yearly median dose was 2.8 mGy/y (0.28 rads/y)³³.

It is noteworthy that these are low doses compared to the 5 rads/y which is the upper limit in the current US radiation safety regulations. Although there is a very large female cohort included in the study, women were excluded from the analysis. The outcomes of the analysis

were mortality from all causes and cancer mortality. The results of this study do not show any health risk which can be clearly associated with relatively low occupational radiation exposure.

Another important study evaluated radiation induced chromosome aberrations of the nuclear dockyard workers³⁴. It included ten years follow-up on 197 dockyard workers exposed to mixed neutron and gamma radiation during the refueling of nuclear reactors. Most of the recorded doses were estimated to be below 5 rads/y. The target cells were the lymphocyte counts in the peripheral blood. The primary endpoint was the frequency of chromosome aberrations. The lymphocytes were used as a “biological dosimeter” because of their radiation sensitivity. The types of chromosome damage evaluated were dicentrics, rings and acentric fragments. In this study dose measurements from film badges were available.

The results showed a significant increase in chromosome damage with increasing exposure although no biological consequences were detected. In terms of study limitations, no women were included in the analysis and no results regarding peripheral blood counts are presented although it is mentioned that dicentrics are genetic anomalies that may lead to cell death. It is also noteworthy that the sample size is small, compared to the previous referenced studies.

An occupational study that included a large number of women is the Chinese study of medical x-ray workers comparing 27,011 medical x-ray personnel to 27,782 other medical specialists. The study included workers employed between 1950 and 1980. The outcome that was analyzed was cancer risk. Unfortunately, no dosimetry was available for the Chinese medical workers and, therefore, the doses were reconstructed by physical and biological retrospective methods². The findings indicated a significantly higher cancer risk among diagnostic x-ray workers compared to the general population.

Another large occupational study of Canadian radiation workers based on the National Dose Registry of Canada, included 191,333 workers, of which 95,643 were women. Dosimetry information was available from personal dosimeters. The job locations included dental, medical, industrial and nuclear power facilities. Most women received low doses which did not allow a powerful stratification by dose and gender with a reasonable sample size³⁵. The results showed no increased standardized incidence ratio for all cancers, but there was an excess risk of several sites specific cancers, including melanoma and thyroid cancer.

A study of US female radiologists included about 70,000 women certified during 1926-1982 who have been followed up longitudinally for breast cancer mortality. Breast cancer mortality risk was highest among women who were first employed prior to 1940³⁶.

Focusing on the same job category, another study of US radiologists presents a fifty year follow-up of a cohort comprised of male radiologists who began the employment between 1920 and 1969³⁷. Females radiologists were excluded because of their small number. The results suggested an excess risk of mortality from all causes at all ages for radiologists registered in the thirties. For recent cohorts of radiologists the mortality risk at young ages for all causes except cancer seems to decrease while the cancer risk appears to persist.

A study of a Danish cohort of 4200 male and 3200 female radiotherapy workers who worked between 1954 and 1982 examined whether the health medical staff was affected by radiation. exposure Neither radiation doses and cancer risk nor the number of exposure years and cancer risk were found to be associated³⁸.

A study of male and female workers at the Hanford nuclear weapons facility during the period 1945-1981 showed mortality rates in workers significantly lower than mortality in the US population. No increase in cancer mortality as a consequence of radiation was reported³⁹.

Analyses of all cause mortality among 14,327 workers employed at the Sellafield plant of British Nuclear Fuels Ltd between 1947 and 1975 found lower all causes mortality among the workers than in the general population of England and Wales. Sellafield workers had also a lower cancer mortality than the general population⁴⁰.

All of the above described studies made an important contribution to a better understanding of the health effects of occupational radiation exposure on humans. However these studies have significant limitations which can be summarized as follows:

- most of the existing data refer only to the stochastic effects which consist of cancer incidence,

- in most of the studies the subjects are males, with females excluded from the analysis or being relatively few in number,

- dosimetry information is not always available; sometimes only dose ranges are reconstructed based upon the work environment. Rarely, doses are reconstructed by using information stored in radiosensitive film badges worn by the workers,

- when exposure measurements are available, the exposure doses are below 5 rads/year which is the radiation limit accepted currently in the US,

- none of the studies reviewed evaluated a decrease of the lymphocyte counts as an outcome of occupational radiation effects;

Therefore, this dissertation can be considered the first use of detailed occupational data to test the deterministic effect of occupational exposure to radiation on lymphocyte counts in both sexes. The high levels of occupational radiation exposure² that occurred in the Mayak PA workers are fortunately no longer observed, so this cohort provides a unique resource for

² refers to radiation doses much higher than 5 rads/year which is currently the standard accepted in the US

evaluating high dose occupational radiation exposure occurring over many years in healthy young male and female workers.

1.2.6 Sex related differences in radiosensitivity

Most studies of occupational radiation exposure which include women among have not indicated any differences in radiation sensitivity between males and females. It is worthwhile to point out, however, that many studies of occupational radiation exposure women did not include women probably due to the fact that the majority of workers employed in radioactive environment were male^{32-34,37}. However, there are a few occupational studies in which women were included. The majority of these studies focused exclusively on cancer incidence and mortality^{1,36,38}. It is also noteworthy that, when women were included in a radiation workers' cohort, they were likely to have received lower radiation doses than males and therefore a comparative dose-response analysis is not possible³⁵.

Interestingly, there are a few papers that suggest differences between males and females in terms of radiation sensitivity^{3,41}. The Staff Review of the National Academies Study of the Health Risks from Exposure to Low Levels of Ionizing Radiations (BEIR VII) suggests "potential influence of gender on radiation sensitivity"³. Data in BEIR VII suggests that females are more sensitive than males in terms of "life time attributable risk for cancer incidence".

Another study investigated the increase of a urinary marker of DNA damage as a result of indoor exposure to gamma radiation and radon. The urinary marker of DNA damage was more affected by radiation and radon in females than in males⁴¹.

These findings reinforce the impression that there exists the lack of definitive information regarding differences by sex in radiation sensitivity, especially regarding the deterministic effect

in an occupational setting. Obviously, the comparative analysis of the effect of long term radiation exposure in males and females is of critical interest, as an increasing number of women enter the work force. Methodological limitations of existing studies impacted considerably on the investigation of potential differences between sexes.

This dissertation study aims to fill these gaps using the best statistical approaches available to scientifically evaluate the effects of occupational radiation on lymphocyte counts in males and females.

1.2.7 Review of repeated measures analysis techniques

The literature review also considered methodological issues regarding statistical methods that can be used for occupational data analyses.

The appropriate analysis of occupational data involves sophisticated statistical techniques that are able to discern meaningful results. The occupational data analyzed are recorded longitudinally, and consist of repeated measurements over a variable number of follow-up years. In order to introduce the most suitable analytical approach for these occupational data, this section reviews statistical procedures commonly used for longitudinal data analysis.

Longitudinal data have important characteristics. Notably, they are clustered data, the clusters consisting of repeated measurements obtained from a single individual at different points in time. Observations within a cluster are positively correlated. Failure to take this correlation into account in the statistical analysis will lead to incorrect estimates of the sampling variability and incorrect inferences⁴². Longitudinal data have also a temporal order, the measurements being taken in an ordered time sequence.

The methods used for the analysis of repeated measurements data have evolved considerably in recent years as the computational software and computers performances have evolved. The faster the computers became, the more complex the calculations implemented in the statistical procedures, as illustrated in the overview of methods that follows below.

1.2.7.1 Univariate analysis of variance (ANOVA)

In the univariate analysis of variance (ANOVA) the response in the i^{th} individual is assumed to differ from the population mean, by an individual specific random effect b_i and a within subject measurement error, e_{ij} .

ANOVA can be applied to the repeated measures design. The repeated measures ANOVA model distinguishes two main sources of variation in the data: between subject and within subject variation. Repeated measures ANOVA models have been widely used because of the relatively simple computational formulas⁴². There are some limitations of these models, such as: the repeated measures has to be based on a set of observations common to all individuals, and the data must be complete and therefore unbalanced data can not be handled.

1.2.7.2 Generalized Linear Models (GLM)

GLM are likelihood based models³. In the classical linear models the observations are independent and identically distributed^{42,43}. When there are repeated measures on the same individual, the observations on the same individual are highly correlated which must be taken into account in the statistical procedures.

The GLM (generalized linear model) can be written mathematically as follows:

³ The parameters are calculated by using the maximum likelihood function

$$Y_{ij} = \beta_1 x_{ij1} + \dots + \beta_p x_{ijp} + \varepsilon_{ij}$$

$$\mathbf{Y} \sim MVN(X\beta, \sigma^2 V)$$

Where:

Y_{ij} =the outcome variable for the i^{th} individual at j^{th} time

β =the estimated coefficients

X =the covariates

The outcome variable Y is considered multivariate normally distributed with mean $x\beta$ and variance $\sigma^2 V$ where V is the variance covariance matrix and σ^2 is the estimated dispersion parameter.

The most popular estimation method used in the GLM (generalized linear model) is the maximum likelihood estimation procedure. Alternative estimation methods include weighted least squares estimation and restricted maximum likelihood estimation (REML).

As shown in the mathematical formula, the GLM (generalized linear model) assumptions require a multivariate normal distribution for a continuous outcome variable. In terms of missing data mechanisms GLM can deal with data missing completely at random (MCAR) or missing at random (MAR)⁴⁴. The missing data patterns and mechanisms are described in a separate section of this dissertation.

1.2.7.3 Linear mixed models

Linear mixed models account for sources of natural heterogeneity among individuals. Therefore individuals are assumed to have their own subject specific mean responses. A distinctive characteristic of the linear mixed model is that the mean response is modeled as a combination of fixed and random effects. Fixed effects are population characteristics shared

by all individuals. Random effects consist of subject specific characteristics. The interpretation of the results should be made at an individual level.

Linear mixed models deal with different patterns of unbalanced data. Thus this statistical procedure is more flexible than GLM (generalized linear models) in accommodating the missing data.

1.2.7.4 Marginal model and GEE (generalized estimating equations)

Marginal models or population-average models⁴⁵, are an extension of the general linear models. They are relevant when the main focus of a study is investigating the effect of covariates on the population mean and not necessarily at individual level. Marginal models are considered more flexible than classic generalized linear models since they can handle unbalanced longitudinal data with repeated measurements and therefore they can handle as well some patterns of missing data. Marginal models do not require precise specification of the outcome distribution.

The equation of the model can be mathematically written using the following notation:

$$\mu_i = E(y_i)$$

$$h(\mu_i) = X_i\beta$$

The elements of the notation have the following definitions:

Y_i = the set of measurements for the outcome for the i -th individual

$$i = 1, \dots, n$$

$$Y_i = [y_{i1} \dots y_{iT}]$$

T = number of repeated observations

β = the corresponding vector of linear models parameter⁴⁶

\mathbf{X} = (p x n) matrix of non-stochastic explanatory variables for the i -th individual

= columns consist of fixed or time dependent covariates

= rows consist of individuals

The ability of this modeling approach to accommodate time–dependent covariates allows a more accurate assessment of the predictor effects than forcing those covariates to be time independent⁴⁷.

$h_1(\cdot)$ = the known link function

It is possible to use different link functions. The link function converts the expected value μ that may be range restricted to the unrestricted linear predictor \mathbf{X} ⁴⁵. The link transformation function allows the dependent variable to be expressed as a vector of parameter estimates (β) in the form of an additive model^{48,49}.

Some examples of the available link functions⁵⁰ are:

- a. *the identity link* $g(a)=a$ for measured data
- b. *the log link* $g(a)=\log(a)$ for count data
- c. *the logit link* $g(a)=\log(a/(1-a))$ for binary data

There is an important difference between the link modifications and outcome variable transformation⁴⁹. This issue is explained as follows:

Let a be the outcome variable then, when the log link is used $\log(E(a))$ is modeled.

If a is transformed as $\log(a)$ and identity link is used, then $E(\log(a))$ is modeled

Therefore the log link is not similar with the log transformation of the variable combined with the identity link.

In marginal models it is useful to specify the distribution of the outcome variable so that the variance can be calculated as a function of the mean. The marginal models may still lead to consistent variance estimates even when there is some misspecification of the variance.

The variance-covariance matrix, part of the model used in the estimating equation, is:

$$V_i^1 = \psi_1 (\mathbf{A}_i^1)^{1/2} \mathbf{R}_{11} (\mathbf{A}_i^1)^{1/2}$$

ψ = the dispersion parameter

\mathbf{R}_{11} = the “working” correlation matrix

\mathbf{A} = the diagonal matrix $A_i^1 = \text{diag}[g_1(\mu_{i1}^1) \dots g_1(\mu_{i1}^T)]$

$g(\cdot)$ = a known variance function

In the marginal models, the estimation of the parameters is performed using the Generalized Estimating Equations method (GEE) which is an extension of simple linear regression that accounts for repeated measures and correlated responses⁵¹. The term generalized estimating equations indicates that an estimating equation is obtained by generalizing another estimating equation⁴⁵. It was introduced by Liang and Zeger⁵² as a method for calculating consistent estimates of the regression parameters and of their variances under weak assumptions about the joint distribution. The method for parameter estimation proposed by Liang and Zeger is the quasi-likelihood estimation method that reduces to maximum likelihood method if the outcome variables have a multivariate normal distribution⁵². In the quasi-likelihood approach a known transformation of the marginal expectation of the outcome is assumed to be a linear function of the covariates.

The correlation among repeated measurements is considered a nuisance parameter. This correlation must be taken into consideration in order to obtain correct parameter estimators⁴⁷. An important step in choosing a specific correlation structure is to find the simplest structure which fits the observed data well⁵³.

The correlation matrix structure may be the following:

a. Independent

The observations in an individual are uncorrelated with every other observation in that individual.

b. Exchangeable

Every observation in an individual is equally correlated with every other observation in that individual. The degree of correlation is measured by the intraclass correlation coefficient

c. Autoregressive

The observations taken closer in time are more correlated than the observations taken far apart in the same individual.

d. Unstructured

No assumption is made about the correlation coefficients between any two pairs of observations.

e. User fixed

Correlation coefficients are fixed by the user rather than being estimated from the data and the values are fixed prior to the analysis.

A useful feature of the GEE model is that the estimators are robust to departures from the true correlation patterns. A loss in estimator efficiency can occur but this loss decreases as the sample becomes larger⁴⁶.

Model diagnostics and goodness of fit assessment are more limited for GEE than for linear regression and therefore standard procedures have to be applied with caution. Modern model selection techniques are unfortunately not included in classical statistical packages and are

computationally intense^{54,55}. In most cases the goodness of fit tools for marginal models are programmed by the user.

Model diagnostics, which are an important part of the model building process, consist of:

a) *Residual analysis:*

Residuals are obtained according to the formula:

$$r_{ij} = y_{ij} - \mathbf{x}_{ij}' \hat{\beta}$$

The residual analysis will be used to assess the model fitting

b) *Outliers Analysis*

- Descriptive exploratory data analysis will identify the whiskers and outliers for both dependent and independent variables.

c) *Checking for normality in the distributions*

Although the GEE methods do not require a multivariate normal distribution of the continuous outcome variable, the normality assumptions has to hold for most of the panels, which usually consist of time points⁵⁶.

d) *Model Validity:*

Checking the model against the data completes the model fitting process and reveals any existing discrepancies. This is done by superimposing the response profile against the time points for different possible situations.

1.2.7.5 Spline functions

As mentioned previously, since the marginal models are an extension of the generalized linear models, a known transformation of the marginal expectation of the outcome is assumed to be a linear function of the parameters. Sometimes, based on the descriptive analysis of data, this

assumption does not hold, therefore some optimization techniques are recommended. One of the optimization methods applied in statistics is the linear spline implementation.

The splines are statistical tools used for covariate transformation when different slopes are assumed for different ranges of the covariate. A spline function is a piecewise polynomial function joined together with certain continuity conditions satisfied. The knots are the principal components of the spline. The knots are located where the slope is assumed to change^{57,58}.

The splines can be introduced using the following mathematical notation:

Given: $n+1$ distinct knots and x_i such that:

$x_0 < x_1 < \dots < x_{n-1} < x_n$ Let the spline function of degree n be:

$$S(x) = \left\{ \begin{array}{ll} S_0(x) & x=[x_0, x_1] \\ S_1(x) & x=[x_1, x_2] \\ \dots\dots\dots & \dots\dots\dots \\ S_{n-1}(x) & x=[x_{n-1}, x_n] \end{array} \right\}$$

Algebraically, each S_i is a linear function constructed such as⁵⁹,

$$S_i(x) = y_i + \frac{y_{i+1} - y_i}{x_{i+1} - x_i} (x - x_i)$$

The spline function must be continuous at each data point such that,

$$S_i(x_i) = S_{i+1}(x_i) \text{ where: } i=1, \dots, n-1$$

Splines are implemented according to the steps presented in the methodology section.

1.2.7.6 Missing data

Although the generalized estimating equation (GEE) technique is a flexible estimation method, there are strict assumptions about missing data. This overview of the missing data features aims to explain some important implications of the missing data on the statistical analysis.

The missing data patterns and mechanisms consist of two important and distinct features. The patterns describe which values are observed in the data matrix and which values are missing. The mechanisms describe the relationship between missingness and the values of variables in the data matrix⁶⁰. The missing data mechanisms have been extensively studied since Rubin elaborated his theory using missing data indicators and their distribution⁶¹. If complete data are defined as $Y = (y_{ij})$, the missing data indicator matrix is defined as $M = (M_{ij})$. The conditional distribution of M given Y is $f(M | Y, \phi)$; where ϕ is the unknown parameter. If the missingness does not depend on the values of the data, missing or observed, then the data are called missing completely at random where (MCAR),

$$f(M | Y, \phi) = f(M | \phi) \text{ for all } y, \phi$$

An assumption less restrictive than MCAR is referred to as missing at random (MAR) when the missingness depends only on the observed data values and not on the missing values:

$$f(M | Y, \phi) = f(M | Y_{obs}, \phi) \text{ for all } y_{obs}, \phi$$

When the distribution of M depends on the missing values the mechanism is referred to as not missing at random (NMAR).

Depending on the missing data patterns and mechanisms, the implications for longitudinal analysis differ:

-Data is missing completely at random (MCAR) implies that individuals with missing data are a random subset of the sample. In this case no bias will arise with almost any method of analysis⁶¹.

-Data is missing at random (MAR) implies that standard GEE methods based on all available observations yield biased estimates of mean response trends as opposed to likelihood based methods which lead to valid estimates if the model is correctly specified.

-Data is not missing at random (NMAR) implies that statistical methods can not be applied since they will lead to biased estimators of the parameters.

1.2.7.7 Methodological overview - summary

The literature review of statistical methodologies suggests a number of important aspects that inform the methodological choice of this study on long term occupational radiation exposure in Mayak PA workers. In this respect, the most relevant literature findings are:

- the marginal model and GEE statistical technique are considered the most flexible and appropriate methods for analysis of occupational data similar to the Mayak PA cohort;
- model building aspects and assessing models' goodness of fit require special attention;
- testing of missing data requires special considerations;
- nonlinear data lead to implementation of linear splines.

In order to be able to implement the statistical methodology mentioned above, a detailed descriptive analysis of the data used in this dissertation is required. The descriptive analysis is designed to provide information needed to assess if the assumptions required by the statistical methods hold. The data used in this study are very complex and include time dependent as well

as time independent variables. This dissertation provides a detailed description of the temporal trend of variables included in the subset of Mayak Pa workers data used in this study.

In addition to the detailed results of the descriptive analysis, this study implements a set of tools used in GEE model building, GEE models' goodness of fit assessment and testing of missing data mechanisms. Currently, the statistical software packages available commercially do not include any standard commands for marginal model building, goodness of fit assessment or for testing of missing data mechanisms.

1.3 OBJECTIVES AND SPECIFIC AIMS

The objective of this dissertation is to model the effects of occupational radiation exposure on lymphocyte counts in Mayak PA workers. The analysis will assess similarities and differences between male and female workers regarding the radiation effects on the lymphocyte counts. Statistical models appropriate for longitudinal data analysis will be fitted and assessed for goodness of fit, and several types of covariates will be tested. This objective is attainable given the detailed data available in the Mayak PA database which includes hematological and dosymetric longitudinal data on workers who received radiation exposure higher than currently accepted by international regulations, in addition to a high proportion of females in the work force.

The objective of this study entails accomplishing the following specific aims:

- To model the pattern of response of the lymphocyte counts to long term, relatively high, occupational exposure to radiation. The lymphocyte counts will be assessed over time while adjusting for the baseline count.

- To assess the goodness of fit of the models
- To test whether males and females show different patterns in lymphocyte counts response to similar long term, relatively high occupational radiation exposures.

This dissertation implements sophisticated statistical methods for the analysis of a subset of data recorded on Mayak PA workers. It is the first study that provides comprehensive analyses descriptively and analytically of the data available for Mayak PA male and female workers.

2.0 METHODOLOGY

2.1 BACKGROUND

The Mayak Product Association (Mayak PA) facility started its operations on January 1, 1948. It was the first nuclear production site in the former Soviet Union and included a reactor plant, a radio-chemical plant and a plutonium processing plant. In the first decade of its existence (1948-1958), inexperience with production techniques, combined with an emphasis on urgent political priorities, resulted in at least 8000 workers receiving high levels (1-10 Gy=100-1000 rads) cumulative exposures of external gamma radiation. Some workers also received acute accidental gamma exposure or internal alpha radiation from inhaled plutonium aerosols³¹.

Radiation doses were measured as part of the Radiation Protection Program that was initiated at the start-up of the Mayak facility. Systematic exposure measurements were carried out on all radiation workers according to an Operations Manual which required each worker to wear individual film-badge and ionizing dosimeters. Medical examinations were also carried out on more than 95% of these workers as part of the same Radiation Protection Program. The medical department of the facility routinely did pre-employment physical examinations on all newly hired workers at Mayak PA. Between 1948 and 1954 each worker underwent a scheduled medical examination every three months; during 1955-1960 every six months; and since 1960, every 12 months. During these examinations, routine peripheral blood counts were carried out

and periodic bone marrow samples were drawn. In the case of workers having radiation-related diagnoses or very high exposures, additional clinical work such as cytogenetics, pulmonary function tests and other procedures were carried out. After retirement, former Mayak workers were followed-up by telephone or mail by the same specialized medical hospital every 24-36 months and underwent physical medical examinations if they remained resident in Ozyorsk. The research staff has continued to collect and maintain these unique human data over the past 54 years in order to study deterministic and other health effects, including those involving hematopoietic, nervous, cardiovascular, respiratory, visual and cytogenetic systems. These detailed longitudinal data on human occupational (internal and external) exposures to ionizing radiation and their resulting clinical outcomes can be used for research and regulatory purposes.

2.2 MAYAK PA WORKERS' EARLY CLINICAL EFFECT DATABASE

The existing records on Mayak PA Workers, starting with the year 1948 and continuing through the present, served as the source for the design and implementation of an electronic database that continues to be updated on an ongoing basis. This database is the result of the long term American-Russian collaborative projects: NRC-Phase I (1996-1997), NIOSH ARS Grant (1998-2001), NRC Phase II (1999-2002) and current NIOSH grant. The electronic database contains selected demographic, work history, occupational exposures and clinical information on a representative sample of 591 Mayak PA workers, out of which 361 males and 230 females were hired between 1948 and 1962.

The MWECE database serves as the source for the data subset named the “study cohort” on which the current research is based.

2.2.1 Sampling procedures for MWECE database implementation

The MWECE cohort selection was based on the available medical records on Mayak PA workers hired between 1948 and 1960. A systematic sampling method was used in order to select representative samples of the workers for the MWECE database. The complete list of workers was initially sorted by start date of employment (1948-1960), then by primary diagnosis (Acute Radiation Sickness, Chronic Radiation Sickness, Plutonium Pneumosclerosis, No radiation Related Diagnosis) and sex. A systematic selection of workers was applied to this list within each one of these sets using a circular sampling technique, with a random starting point. This procedure ensured that workers of both genders were selected in a systematic fashion across the full 10-year period. The samples were checked to make sure that all the workers had been employed for at least 12 months during the critical sampling period, and resampling was carried out when necessary. This procedure resulted in a representative sample of workers with a known sampling proportion for each one of the individual strata (Table 1).

Table 1: MWECE implementation: Number of Workers sampled in MWECE in Each Primary Diagnosis Category; Percentage Sampled from Total Mayak Pa Workers in Each Primary Diagnosis Category

Primary Diagnosis	MWECE Number of workers	Sampling Percentage from Total Mayak PA Workers in Each Category (%)
Acute Radiation Syndrome	60	100
Chronic Radiation Syndrome	202	10
Plutonium Pneumosclerosis	120	100
No Radiation-Related Diagnostic	209	3.5
Total	591	7.3

2.2.2 MWECE database format

Figure 1 provides a summary of the MWECE database format. The database consists of a series of modular electronic data sets. Each worker is identified by a unique identification number in all data modules. This format permits data import into statistical packages (e.g., STATA) used for

complex statistical analyses. Files containing information of interest (e.g., hematological and dosimetry data) can be generated by merging data from different data modules.

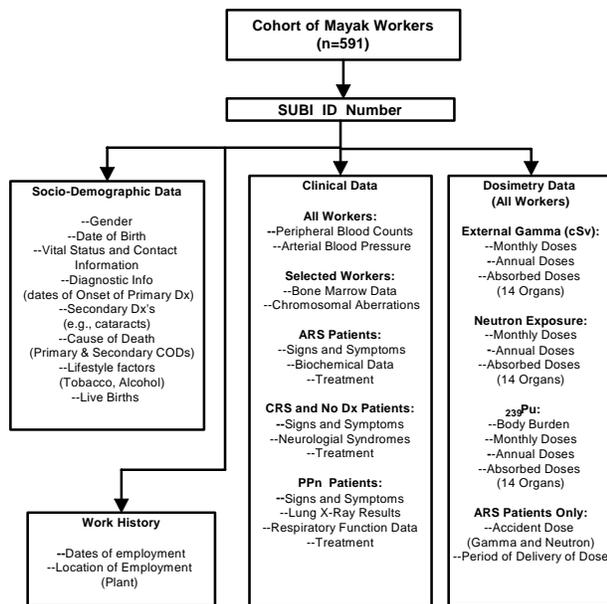


Figure 2: MWECE Data Structure
 Source: Current NIOSH grant proposal

Peripheral blood counts and radiation exposure recorded longitudinally are considered important variables in MWECE. Peripheral blood counts were recorded as absolute blood counts (e.g., erythrocytes, platelets) or as percentages from the total number of leucocytes (e.g., lymphocytes, neutrophils). The exact date is available for each blood test (Appendix Table 1). Whole body external gamma dose is recorded in rads, as cumulative monthly and yearly exposure data. MWECE exposure data show high long term radiation exposure¹. Plutonium exposure data are also available on a monthly and yearly basis in males and females.

Some life style variables such as smoking history and alcohol consumption are also included in the MWECE database. Smoking history and alcohol consumption were collected

retrospectively from medical records. Smoking history was categorized as never, former or current. Smoking history was also quantified as the estimated average number of cigarettes per day at start of employment, before 1954 and after 1954. Alcohol consumption has been categorized according to the Russian standards: no consumption, moderate consumption, domestic alcohol abuse, alcoholism. Drinking excessively at home was considered domestic alcohol abuse. Drinking at work was considered alcoholism. Alcohol consumption has been recorded in the MWECE database at start of employment, before 1954 and after 1954 (Appendix Table 2).

Work history data consists of employment dates and work locations. Information regarding work location consists of plants to which workers were assigned at specific dates. Work location is a time dependent variable since workers may have worked at different plants during employment. The Mayak PA plants are designated as A-reactor, B-biochemical C-plutonium plant D-other plants (Appendix B, Table 3).

Clinical data available on males and females chronically exposed at high radiation doses makes the MWECE database unique and very valuable. Overall, the MWECE database is probably the most complete source of information in the literature on a large number of males and females working at a nuclear industrial facility.

2.2.3 Study cohort

A subset of the MWECE database has been selected for this dissertation project. The project has the main goal of studying the hematological effect of long term occupational external radiation exposure. The data subset includes only workers with chronic external gamma exposures. Since acute and internal exposure effects are not the main interest of this dissertation, workers with

acute radiation sickness (involved in radiation accidents), as well as workers with plutonium pneumosclerosis, have been excluded from the analysis. After these exclusions, 411 out of 591 workers are included in the first step of subset selection (Tables 2, 3).

Table 2: Study Cohort Implementation: Frequency Distribution of Primary Diagnostic; Included Workers by Sex

Primary Diagnostic	Total		Males		Females	
	n	(%)	n	(%)	n	(%)
Chronic Radiation Sickness	202	(49.1)	141	(57.1)	61	(37.2)
No Radiation-Related Diagnosis	209	(50.9)	106	(42.9)	103	(62.8)
Total Included	411	(100)	247	(100)	164	(100)

Table 3: Study Cohort Implementation: Frequency Distribution of Primary Diagnostic; Excluded Workers by Sex

Primary Diagnostic	Total		Males		Females	
	n	(%)	n	(%)	n	(%)
Acute Radiation Sickness	60	(33.3)	50	(43.9)	10	(15.2)
Plutonium Pneumosclerosis	120	(66.7)	64	(56.1)	56	(84.8)
Total Excluded	180	(100)	114	(100)	66	(100)

The MWECE database includes workers exposed at radiation doses higher than those accepted in western countries. Before 1960 in Mayak PA radiation workers the mean yearly external gamma dose was much higher than the 5 rads/year cut-off considered acceptable currently in the US. After 1960 protective rules and regulations were implemented at Mayak and the average radiation doses decreased. The cohort for this study is represented only by workers hired during 1948 through 1960 and encompasses 1 to 10 years of follow-up from the first exposure to occupational external gamma radiation. The choice of time frame is determined by the interest of the study in the effect of the highest chronic occupational radiation exposure in recorded history.

In addition to the criteria described, workers included in the study have non-missing records of exposure and blood counts along with a non-missing start of employment date. Workers for whom at least one blood count was recorded are included. Likewise, workers for whom at least one external gamma exposure dose was recorded are included. Exclusion criteria

of the study cohort are summarized in Table 4. Table 5 presents the structure of the cohort study compared to MWECE.

Table 4: Study Cohort Implementation; Frequency Distribution of Exclusion Criteria by Sex

Exclusion Criteria	Total		Males		Females
	n	(%)	n	(%)	n (%)
Acute Exposures and Pneumosclerosis	180	(79.3)	114	(85.7)	66 (70.3)
Missing Hematological and Dosimetry Data	6	(2.2)	4	(3.0)	2 (1.0)
Missing Dosimetry Data	29	(12.8)	11	(8.3)	18 (19.1)
Missing Hematological Data	12	(5.3)	4	(3.0)	8 (8.6)
Missing start of employment	1	(0.4)	0	(0.0)	1 (1.0)
Total Excluded	228	(100)	133	(100)	95 (100)

Table 5: Study cohort implementation; MWECE Frequency Distribution of Workers Included and Workers Not Included in the Study Cohort By Sex

MWECE	Total		Males		Females
	n	(%)	n	(%)	n (%)
Included in Study Cohort	363	(61.4)	228	(63.1)	135 (58.7)
Not included in Study cohort	228	(38.6)	133	(36.9)	95 (41.3)
MWECE	591	(100)	361	(100)	230 (100)

2.3 STUDY DESIGN

The main objective of this study is the assessment of the effects of external gamma exposure on lymphocyte counts controlling for the baseline blood count, work location and lifestyle variables. Since multiple measures on the same subject are recorded over time, the study is designed as a repeated measures longitudinal data analysis. This specific design has the advantage of a powerful analysis as it takes into account both the variability between subjects and within the same subject. The design is very appropriate for making full use of the complexity of the study cohort.

The study involves a descriptive component and a statistical component. The statistical component consists of a modeling section and an analytical section. All components of this methodology are created and implemented to answer the following research questions:

1. Are lymphocyte counts affected by long term occupational exposure to gamma radiation in Mayak PA workers?
2. Are there differences between males and females in terms of inhibition of lymphocyte counts by occupational exposure to gamma radiation in Mayak PA workers?
3. Is there a differential effect on lymphocyte counts, at lower gamma doses (below 5 rads) compared to higher gamma doses? If yes, is the effect size similar in males and females?
4. Does the effect on the lymphocyte counts vary over time in workers exposed to occupational external gamma radiation? If yes, is the time effect similar in males and females?

The research questions stated above are addressed using the following strategies:

2.3.1.1 Definition of study variables

The outcome variables are the lymphocyte counts. They are time-dependent variables. In this analysis the blood counts themselves are used; they are calculated by multiplying the total number of leucocytes by the percentage of lymphocytes:

(Lymphocyte counts=lymphocyte percentage*leucocyte count/100).

The explanatory covariate is the cumulative yearly external gamma dose, defined as a time dependent variable.

The study is designed to control for the following variables: sex, the baseline blood count, work history, smoking history and alcohol consumption at the start of employment.

The first lymphocyte count performed in the year of the first external gamma exposure is considered the baseline count.

The work location related to Plutonium exposure is an adjustment variable generated by using the work location variable which is considered a time dependent variable. The work location variable consists of the plant in which an individual worked most of the current year. Work location related to Plutonium exposure variable is categorized as follows:

- a) the reactor and other work locations are characterized as work locations with low Plutonium exposure
- b) the radio-chemical plant is characterized as a work location with moderate Plutonium exposure
- c) the Plutonium plant is characterized as a work location with high Plutonium exposure

Lifestyle variables consist of smoking history and alcohol consumption at start of employment. Smoking history is collapsed as ever or never smoked. Alcohol consumption at start of work is collapsed as no consumption or any alcohol consumption.

Longitudinal studies also have a temporal variable. In this analysis the number of years from the first external gamma exposure is considered the temporal variable.

In the study cohort there is one record per year for external gamma dose and multiple records per year for lymphocyte counts in each worker. The average worker has 5 blood tests per year although the number varies considerably and one worker had a maximum of 173 recorded tests in one year (Fig. 2). In order to match the blood count records with the external gamma exposure, the median of the blood count is calculated by year for each worker. Thus, the multiple blood test values per year in each individual are replaced by the median value which is not unduly influenced by extreme values. The median eliminates the yearly variability of the blood counts. To determine if using the yearly median of blood counts changes drastically the nature of the data, it is necessary to compare the log-transformed median counts against the mean of log-

transformed counts. This comparison performed for lymphocytes shows a very good correlation between the log-transformed median counts and the mean of log-transformed counts. Therefore, the use of the yearly median lymphocyte counts for each individual is appropriate.



Figure 3: Distribution of the Log-Transformed Total Number of Lymphocyte Counts Performed Yearly in Each Worker

2.3.1.2 Descriptive analysis

Although a subset of the MWECE database, this study cohort is still a very complex database. The information on lymphocyte counts and external gamma exposure recorded for males and females allow complex and valuable statistical analyses.

A preliminary descriptive analysis is mandatory in order to design in detail the statistical analysis and the hypotheses to be tested. The descriptive data analysis involves summarizing the existing information to describe the pattern of lymphocyte counts over time in relationship to the pattern of exposure over time. Preliminary descriptive data analysis is designed as follows:

a) The median lymphocyte counts are calculated by year from the first external gamma exposure. This is based on the yearly median lymphocyte counts calculated for each worker.

b) The median yearly gamma exposures are calculated by year from the first external gamma exposure. This is based on the yearly cumulative external gamma exposures data of the study cohort.

c) The median yearly gamma exposures are plotted against the number of years from the first external gamma exposure.

d) The median lymphocyte counts are plotted against the number of years from the first external gamma exposure.

e) The plots are overlapped in order to assess the pattern of dose-response relationship.

The results of the descriptive analysis will be presented in the results section.

2.3.1.3 Statistical methods

Statistical methods consist of a modeling and an analytical component. The modeling component is concerned with model building and model diagnostics. These procedures lead to statistical models which are the basis for the analytical component.

Modeling section

1. Model building

The marginal model approach and the corresponding analytical technique of GEE are applied to the study cohort. The choice of this model is based on its flexibility and adequacy for the analysis of the study cohort. More specifically, the marginal model: accounts for repeated measures, time dependent variables and correlated responses; deals with unbalanced number of observations and missing data; deals with different outcome variable correlation structures, and

provides consistent estimates of the regression parameters and consistent estimates of the variances under weak assumptions about the joint outcome distribution.

The marginal model was introduced as an extension of general linear models (GLM). As opposed to general linear models which assume that all observations are independent, marginal models account for correlation between repeated measurements on the same subject. This approach, also known as the population averaged model, is recommended when the main focus of a study is investigation of the effect of covariates on the outcome mean rather than at an individual level⁴². The non-specification of multivariate distribution leads to a quasi-likelihood based method of estimation named generalized estimating equations (GEE) which reduces to the maximum likelihood method if the outcome variables have a multivariate normal distribution⁵². The specification of the multivariate distribution of the outcome is not required. Instead, the marginal distribution of the outcome at each time point must be specified⁵⁶. Therefore, the distribution of lymphocyte counts is assessed for each year since the start of work. If these blood counts are not normally distributed in the majority of the points in time, then a transformation is necessary. For the blood counts a log-normal transformation has been found to be approximately normally distributed. The log transformation may also lead to an invariant outcome variance across time, which is also a marginal model assumption.

Data in the literature suggest possible differences in radiation sensitivity between males and females^{3,41} and unlike most other occupational cohorts^{2,32,33,35,39,62}, this study cohort includes 135 women occupationally exposed to radiation. The large number of women included in the study cohort offers the opportunity to perform a statistical analysis by sex. The differences in radiation sensitivity between males and females are analyzed using the following approach:

- a) an interaction term between log-transformed yearly cumulative external gamma dose and sex is tested in a model including the main effects
- b) if the interaction term is statistically significant or at the borderline of statistical significance then the analyzed models are stratified by sex

The elements of the marginal model introduced in the study design section are mathematically defined using the following notation:

Let Y_{it} = the lymphocyte count for the i^{th} worker recorded at t years from the first external gamma exposure

Let X_{1it} = the yearly cumulative external gamma dose recorded for the i^{th} worker at t years from the first external gamma exposure

Let X_{2i} = the lymphocyte baseline count for the i^{th} worker

Let X_{3i} = smoking history in the i^{th} worker

Let X_{4i} = alcohol consumption in the i^{th} worker

Let X_{5it} = 1 if the i^{th} worker was employed the most part of the t^{th} year from the first external gamma exposure at the radio-chemical plant and X_{5it} = 0 otherwise

Let X_{6it} = 1 if if the i^{th} worker was employed the most part of the t^{th} year from the first external gamma exposure at the plutonium plant and X_{6it} = 0 otherwise

Let X_{7i} = 0 if the i^{th} worker is a male and X_{7i} = 1 if the i^{th} is a female

Let X_{8it} = the interaction term between ln(yearly cumulative external gamma dose) and work location related to Plutonium exposure

Let X_{9it} = the interaction term between ln(yearly cumulative external gamma dose) and sex recorded in each worker

Let X_{10i} = the number of years from the first external gamma exposure

Let $E(Y_{it}|X_{it})=\mu_{it}$ be the conditional expectation of the outcome variable given the covariates,

Let $g(\mu_{it})=X_{it}\beta$, where $g(\mu_{it})$ is the known link function.

Let $Var(Y_{it}) =\phi v(\mu_{it})$, where ϕ = estimated scale parameter and $v(\mu_{it})$ is the known variance function of the mean.

2. Model diagnostics

Model diagnostics are challenging in marginal models due to the correlated responses, and standard diagnostics procedures available in linear regression can not be applied to marginal models exactly because of the correlated responses. Model diagnostics procedures for marginal models-GEE are not included in statistical packages as standard tools. Thus, they have to be programmed by writing statistical code according to formulas available in the literature.

Model diagnostics have been performed through the following procedures implemented by code written in STATA9 statistical software.

- Residuals computation results in a variable which has specific values for each model

$$res_{it} = Y_{it} - \hat{Y}_{it}$$

where

y_{it} =the observed variable measured for i-th individual at time t

\hat{y}_{it} =the predicted variable for i-th individual at time t

- Residuals are plotted against a continuous covariate in order to check if there is a random distribution^{42,45}. A scatter with no obvious pattern supports a random residual distribution

and a good fit of the considered model. If there is a pattern, then the model has to be reassessed.

- Residuals random distribution around zero is tested using the Wald-Wolfowitz test. This is a non-parametric procedure used to test if the residuals have a random distribution in a repeated measures setting⁶³. The Wald-Wolfowitz test can be described using the following notation:

$$E(T) = \frac{2n_p n_n}{n_n + n_p} + 1$$

$$V(T) = \frac{2n_p n_n (2n_p n_n - n_n - n_p)}{(n_n + n_p)^2 (n_n + n_p - 1)}$$

$$W_z = \frac{T - E(T)}{\sqrt{V(T)}}$$

where

n_n = number of negative residuals

n_p = number of positive residuals

T = number of runs = how many times the sign of residual changes

H_0 = residuals are not randomly distributed around zero; under H_0 W_z has a standard normal distribution

- GEE- R^2 computation according to the formula:

$$R^2_m = 1 - \frac{\sum_{i=1}^T \sum_{j=1}^n (Y_{it} - \hat{Y}_{it})^2}{\sum_{i=1}^T \sum_{j=1}^n (Y_{it} - \bar{Y})^2}$$

where:

y_{it} = the observed variable measured for i-th individual at time t

\hat{y}_{it} = the predicted variable for i-th individual at time t

$$\bar{Y} = \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n Y_{it} = \text{the overall mean}$$

GEE- R^2 is interpreted as in multiple linear regression, namely the proportion of the outcome variation due to the model. If GEE- $R^2 = 1$ the model for which GEE- R^2 was calculated explains perfectly the outcome variation. Since it has a more complex formula than the classic R^2 , negative values are also possible. A negative GEE- R^2 means that the model with intercept only explains better the outcome variation than the full model and therefore the full model fits poorly.

- QICu adapted for longitudinal data is be calculated⁵⁵ according to the formula:

$$QICu = -2Q(g^{-1}(y)) + 2p$$

where $Q = -\frac{1}{2} \sum (y - \hat{y})^2$ is the value of quasi-likelihood function calculated for the independence model but with the regression coefficients fitted for the hypothesized correlation structure.

y = the observed values

\hat{y} = the predicted values

QICu combines in one statistical tool the predicted versus the observed values differentials and the penalty induced by the increase in the number of covariates. The QICu is a measure used to determine the best subset of covariates for a particular model. The model with the smallest QICu should be chosen.

Analytical Section

The statistical analysis of the marginal models is performed using the generalized estimating equation technique (GEE). The term indicates that an estimating equation is obtained by generalizing another estimating equation⁴⁵. This permits the calculation of robust estimates for the standard error of the regression coefficients, accounting for the correlation of the outcomes. It is a quasi-likelihood based procedure, introduced by Liang and Zeger⁵² for calculating the consistent estimates of the regression parameters and their variances under weak assumptions about the joint distribution.

The quasi-likelihood function is built based on classic likelihood function and permits relaxation of the assumption about the multivariate normal distribution of the outcome. The quasi-likelihood estimator is the solution for the score-like estimation system which reduces to the score equation when the outcome has a distribution from the exponential family (normal distribution). The solution can be obtained by an iteratively reweighted least squares⁶⁴ method. The coefficients are actually pooled coefficients of a between subject and within subject relationship⁵³.

A useful feature of the GEE model is that the estimators are robust even when there are departures from the true correlation patterns. A decrease of the estimator efficiency can occur but this is not significant as the sample size becomes larger⁴⁶. The correlation among repeated measures is considered a nuisance parameter. In this study equal correlation among repeated measures is assumed since preliminary results show no time related correlation patterns. The exchangeable correlation structure is used along with the robust variance estimator which adjusts for eventual correlation misspecifications.

In the quasi-likelihood estimation approach a known transformation of the marginal expectation of the outcome is assumed to be a linear function of the covariates. The transformation is the link function. In this study the link function is considered the identity: $g(a)=a$ since this is the most appropriate for the analysis goals.

The linearity between outcome and covariates becomes an important issue of this statistical analysis since there is ongoing debate in the literature about linear versus non-linear dose response patterns of radiation effects^{12,14-16,20,21}. This issue is addressed using optimization methods. Specifically, the linear splines are implemented for external gamma dose to identify and describe the non linear relationship between external gamma dose and lymphocyte, neutrophil and platelet counts⁶⁵. Linear splines are statistical tools used for covariate transformation when different slopes are assumed for different ranges of the covariates. The knots are components of the spline located in most of the cases where the slope is assumed to change^{57,58}. The linear splines are introduced using the following notation:

Let $X_{1_{it}}$ =the 1st spline for the external gamma dose recorded in the i^{th} worker at t years from the first external gamma exposure

Let $X_{2_{it}}$ =the 2nd spline for the external gamma dose recorded in the i^{th} worker at t years from the first external gamma exposure

Let $X_{3_{it}}$ =the 3rd spline for the external gamma dose recorded in the i^{th} worker at t years from the first external gamma exposure

The implemented splines have the following knots location:

1st knot: $\ln(\text{dose})=0$ It corresponds to a yearly cumulative external gamma dose of 1 rem.

2nd knot: $\ln(\text{dose})=1.609437$ It corresponds to a yearly cumulative external gamma dose of 5 rem.

It is important to note that the first knot corresponds to a value of the log-transformed external gamma dose where the residuals distribution changes. The second knot corresponds to a theoretical value of the log-transformed external gamma dose which consists of the upper limit of yearly occupational external gamma doses currently accepted in the US.

2.3.1.4 Missing Data

Another important issue related to the use of marginal models and generalized estimating equations refers to the missing data patterns and mechanisms. The missing data patterns and mechanisms consists of two important and distinct features: the patterns describe which values are observed in the data matrix and which values are missing; and the mechanisms describe the relationship between missingness and the values of variables in the data matrix⁶⁰. The missing data mechanisms have been extensively studied since Rubin elaborated his theory using missing data indicators and their distribution⁶¹. If complete data are defined as $Y = (y_{ij})$, the missing data indicator matrix is defined as $M = (M_{ij})$. The conditional distribution of M given Y is $f(M | Y, \phi)$; where ϕ is the unknown parameter. If the missingness does not depend on the values of the data, missing or observed then the data are called missing completely at random where (MCAR),

$$f(M | Y, \phi) = f(M | \phi) \text{ for all } y, \phi$$

An assumption less restrictive than MCAR is referred to missing at random (MAR) when the missingness depends only on the observed data values and not on the missing values:

$$f(M | Y, \phi) = f(M | Y_{obs}, \phi) \text{ for all } y_{obs}, \phi$$

When the distribution of M depends on the missing values the mechanism is referred to not missing at random (NMAR).

Missing data have three important consequences for longitudinal data analysis⁴²:

- They create unbalanced data across time since not all subjects have the same number of repeated measures;

- They cause loss of information thereby reducing the efficiency and precision in the estimated mean response over time;

- They introduce bias and thereby potentially misleading inferences.

In order for GEE method to be applied, it is very important to assess if the missing data mechanism is completely at random. In this respect, formal statistical tests for missing data mechanisms and pattern assessment are performed for the overall study cohort and after stratification by sex since separate models are fit in males and females^{43,44,66}. The missing data are analyzed in a separate chapter of this dissertation.

In the case of this study cohort, it is important to check if data are missing completely at random (MCAR). MCAR is one of the assumptions required by the GEE techniques.

In summary, in the study cohort the following assumptions required by the generalized estimating equation technique have to hold:

1. Missing data has to be completely at random (MCAR), this in order to obtain unbiased estimates of the parameters;

2. The model has to be correctly specified. The adjustment for correlation misspecifications is made by using a robust variance estimator;

3. The outcome variable must be univariate normally distributed across time;

4. The outcome variable must have constant variance across time;

5. The mean response measure must be directly related to a linear combination of the covariates.

The statistical procedures described above are considered the optimal approaches, based on what is currently available, for addressing the research questions proposed by this study.

3.0 RESULTS

3.1 DESCRIPTIVE ANALYSIS RESULTS

3.1.1 Study Cohort Structure

The study cohort includes 353 workers at start of the follow-up including 223 males and 130 females. The follow-up of the study cohort begins with the year of the first external gamma exposure and encompasses 10 years. The number of workers decreases during the follow-up interval of 10 years as some workers drop out. The number of workers varies across time, yet males and females are almost equally represented during each follow-up year (Table 6).

Table 6: Absolute and Relative Frequency Distribution of Workers by Year since the First Non-Zero Non-Missing External Gamma Exposure (Total, Males, Females)

Numbers of years since the first external gamma exposure	Total n (%)	Males n (%)	Females n (%)
0	353* (100.0)	223 (100.0)	130 (100.0)
1	343 (97.2)	217 (97.3)	126 (96.9)
2	328 (92.9)	208 (93.3)	120 (92.3)
3	308 (87.3)	197 (88.3)	111 (85.4)
4	299 (84.7)	188 (84.3)	111 (85.4)
5	278 (78.8)	177 (79.4)	101 (77.7)
6	263 (74.5)	167 (74.9)	96 (73.8)
7	239 (67.7)	151 (67.7)	88 (67.7)
8	234 (66.3)	148 (66.4)	86 (66.2)
9	223 (63.2)	141 (63.2)	82 (63.1)
Total	2,868 (100.0)	1,817 (63.4)	1,051 (36.6)

*initial size of the study cohort which decreases by year of follow-up due to drop outs

For the first follow-up year, the age distribution shows that the majority of workers (80%) are between 19 and 34 years old with small variation by sex, thereby indicating a relatively young population of males and females radiation workers (Table 7). Females are slightly younger than males.

Table 7: Age Distribution* at the Beginning of the Follow-up in the Year of the First Non-Zero Non-Missing External Gamma exposure by Sex

Sex	Min	10 th percentile	Median	90 th percentile	Max	Mean	Standard error of the mean	N
Males	17	19	24	36	48	26.4	0.33	223
Females	17	19	23.5	31.5	41	24.2	0.46	130
Total	17	19	24	34	48	25.6	0.42	353

- age is expressed in years

a. Study cohort across time units relevant for this study

The study cohort includes workers hired between 1947 and 1958. The peak of hiring was in 1948 and 1949 (Table 8), which are the first years of plant operation.

Table 8: Absolute and Relative Frequency Distribution of Workers by Start of Employment Year and Sex

Start of employment year	Total		Males		Females	
	n	(%)	n	(%)	n	(%)
1947	37	(100.0)	23	(61.1)	14	(38.9)
1948	86	(100.0)	54	(62.80)	32	(37.2)
1949	67	(100.0)	42	(62.7)	25	(37.3)
1950	40	(100.0)	26	(65.0)	14	(35.0)
1951	44	(100.0)	35	(79.6)	9	(20.5)
1952	23	(100.0)	11	(47.8)	12	(52.2)
1953	25	(100.0)	11	(44.0)	14	(56.0)
1954	12	(100.0)	5	(41.7)	7	(58.3)
1955	4	(100.0)	1	(25.0)	3	(75.0)
1956	2	(100.0)	2	(100.0)	0	(0.0)
1957	5	(100.0)	5	(100.0)	0	(0.0)
1958	8	(100.0)	8	(100.0)	0	(0.0)
Total	353	(100.0)	223	(63.2)	130	(36.8)

Most of the workers were first exposed to external gamma radiation in 1949 and 1950 which are the second and the third year of plant operation (Table 9).

Table 9: Absolute and Relative Frequency Distribution of Workers by Year of First Non-Zero Non-Missing Exposure and Sex

Year of first external gamma exposure occurrence	Total		Males		Females	
	n	(%)	n	(%)	n	(%)
1948	11	(100.0)	9	(81.9)	2	(18.2)
1949	98	(100.0)	58	(59.2)	40	(40.8)
1950	66	(100.0)	45	(68.2)	21	(31.8)
1951	57	(100.0)	43	(75.4)	14	(24.6)
1952	34	(100.0)	21	(61.8)	13	(38.2)
1953	37	(100.0)	16	(43.2)	21	(56.8)
1954	20	(100.0)	11	(55.0)	9	(45.0)
1955	9	(100.0)	2	(22.2)	7	(77.8)
1956	5	(100.0)	3	(60.0)	2	(40.0)
1957	5	(100.0)	5	(100.0)	0	(0.0)
1958	8	(100.0)	8	(100.0)	0	(0.0)
1959	3	(100.0)	2	(66.7)	1	(33.3)
Total	353	(100.0)	223	(63.2)	130	(36.8)

In the study cohort, the number of workers tested for lymphocyte count in each year is subject to variation. More than 100 workers were hematologically tested each year between 1950 and 1960 (Table 10).

Table 10: Absolute and Relative Frequency Distribution of Workers by Year of Blood Testing

Year when the blood test was performed	Total		Males		Females	
	n	(%)	n	(%)	n	(%)
1948	11	(100.0)	9	(81.8)	2	(18.2)
1949	109	(100.0)	67	(61.5)	42	(38.5)
1950	173	(100.0)	111	(64.2)	62	(35.8)
1951	226	(100.0)	150	(66.4)	76	(33.6)
1952	250	(100.0)	164	(65.6)	86	(34.4)
1953	282	(100.0)	177	(62.8)	105	(37.2)
1954	281	(100.0)	180	(64.1)	101	(35.9)
1955	275	(100.0)	168	(61.1)	107	(38.9)
1956	269	(100.0)	161	(59.9)	108	(40.1)
1957	251	(100.0)	154	(61.4)	97	(38.6)
1958	236	(100.0)	148	(62.7)	88	(37.3)
1959	161	(100.0)	106	(65.8)	55	(34.2)
1960	115	(100.0)	72	(62.6)	43	(37.4)
1961	81	(100.0)	49	(60.5)	32	(39.5)
1962	56	(100.0)	33	(58.9)	23	(41.1)
1963	34	(100.0)	22	(64.7)	12	(35.3)
1964	22	(100.0)	15	(68.2)	7	(31.8)
1965	16	(100.0)	14	(87.5)	2	(12.5)
1966	10	(100.0)	9	(90.0)	1	(10.0)
1967	7	(100.0)	6	(85.7)	1	(14.3)
1968	3	(100.0)	2	(66.7)	1	(33.3)
Total	2,868	(100.0)	1,817	(63.5)	1,051	(36.5)

b. Description of the outcome variable

Lymphocyte counts

The outcome variable of the analysis consists of the lymphocyte counts. As explained in the methodology section, the study cohort comprises one record per year for external gamma dose and multiple records per year for lymphocyte counts in each worker. In order to match the blood counts records with the external gamma exposure, the median of the blood counts is calculated by year in each worker. Thus, the multiple blood tests' values per year in each individual are replaced by one median which is not influenced by extreme values. The median eliminates the yearly variability of the blood counts.

According to the assumptions of the statistical techniques described in the methodology section, the specification of the multivariate distribution of the outcome is not required. Instead, the marginal distribution of the outcome at each time point must be specified.

The analysis of the yearly median lymphocyte counts during each year since the first non-zero non-missing external gamma exposure illustrates a non-normal distribution for most of the years considered. However, the log-transformation of the yearly median lymphocyte counts normalizes the distribution in more than half of the follow-up years. This result holds after stratification by sex (Tables 11-13).

Table 11: Departure from Normality of Yearly Median Lymphocyte Counts by Years following the First External Gamma Exposure

Numbers of years since the first external gamma exposure	Normal distribution of the median lymphocyte counts*	Normal distribution of the log-transformed median of the lymphocyte counts*
0	Non-normal	Normal
1	Non-normal	Normal
2	Non-normal	Normal
3	Non-normal	Non-normal
4	Non-normal	Normal
5	Non-normal	Non-normal
6	Non-normal	Normal
7	Non-normal	Non-normal
8	Non-normal	Normal
9	Non-normal	Non-normal

*Departure from normality was assessed using a statistical test based on skewness and curtosis
Data was considered normally distributed if $p > 0.05$

Table 12: Males; Departure from Normality of Yearly Median Lymphocyte Counts by Years following the First External Gamma Exposure

Numbers of years since the first external gamma exposure	Normal distribution of the median lymphocyte counts	Normal distribution of the log-transformed median of the lymphocyte counts
0	Non-normal	Non-normal
1	Non-normal	Normal
2	Non-normal	Normal
3	Non-normal	Non-normal
4	Non-normal	Normal
5	Non-normal	Normal
6	Non-normal	Normal
7	Non-normal	Normal
8	Non-normal	Normal
9	Non-normal	Non-normal

*Departure from normality was assessed using a statistical test based on skewness and curtosis
Data was considered normally distributed if $p > 0.05$

Table 13: Females; Departure from Normality of Yearly Median Lymphocyte Counts by Years following the First External Gamma Exposure

Numbers of years since the first external gamma exposure	Normal distribution of the median lymphocyte counts	Normal distribution of the log-transformed median of the lymphocyte counts
0	Non-normal	Normal
1	Non-normal	Non-normal
2	Non-normal	Non-normal
3	Non-normal	Non-normal
4	Non-normal	Normal
5	Non-normal	Non-normal
6	Normal	Normal
7	Non-normal	Non-normal
8	Non-normal	Normal
9	Normal	Non-normal

*Departure from normality was assessed using a statistical test based on skewness and curtosis
Data was considered normally distributed if $p > 0.05$

To determine if using the yearly median of blood counts changes drastically the nature of the data, it is necessary to compare the log-transformed median counts against the mean of log-transformed counts. This comparison performed for lymphocyte counts shows a good correlation between the log-transformed median counts and the mean of log-transformed counts (Fig 4).

Therefore, the yearly median lymphocyte counts are an appropriate way to summarize the multiple counts recorded during one year.

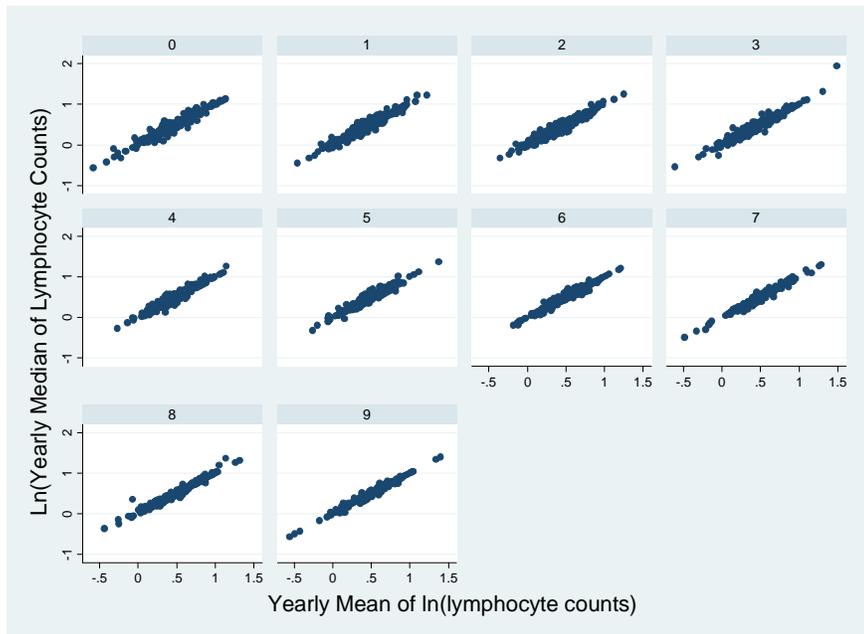


Figure 4: Study Cohort: Yearly Mean of Log Transformed Lymphocyte Counts versus Log Transformed Median of Lymphocyte Counts; 10 Years of Follow-up Starting during the Year of First Non-Zero Non-Missing Exposure

The distribution of the yearly median lymphocyte counts, which are the study outcomes, is described in each year for ten years starting with the year of the first non-zero non-missing external gamma exposure. Yearly median lymphocyte counts are described using the minimum,

median, maximum, mean and standard error of the mean. This descriptive analysis is given for the study cohort, overall and after stratification by sex (Table 14-16).

Table 14: Distribution of Yearly Median Lymphocyte Counts (X1000/mm³) by Year since the First Non-Zero Non-Missing External Gamma Exposure

Numbers of years since the first external gamma exposure	Min	Median	Max	Mean	Standard error of the mean	N*
0	.57	1.61	3.12	1.67	.02	347
1	.64	1.56	3.39	1.59	.02	342
2	.72	1.51	3.51	1.56	.02	328
3	.59	1.51	6.98	1.58	.03	308
4	.77	1.59	3.53	1.65	.02	298
5	.72	1.66	3.95	1.68	.02	278
6	.83	1.66	3.36	1.70	.03	261
7	.62	1.67	3.65	1.73	.03	239
8	.70	1.67	3.91	1.75	.03	234
9	.57	1.73	4.05	1.76	.03	223

*Number of workers who had at least one lymphocyte count in that specific year

Table 15: Males; Distribution of Yearly Median Lymphocyte Counts (X1000/mm³) by Year since the First Non-Zero Non-Missing External Gamma Exposure

Numbers of years since the first external gamma exposure	Min	Median	Max	Mean	Standard error of the mean	N*
0	.57	1.60	3.01	1.65	.03	221
1	.72	1.52	3.39	1.58	.03	217
2	.72	1.50	3.07	1.56	.03	208
3	.77	1.54	6.98	1.62	.04	197
4	.88	1.59	3.53	1.65	.03	188
5	.72	1.66	3.05	1.66	.03	177
6	.83	1.65	3.36	1.71	.03	166
7	.84	1.66	3.65	1.74	.04	151
8	.70	1.70	3.91	1.80	.04	148
9	.61	1.80	4.05	1.80	.04	141

*Number of workers who had at least one lymphocyte count in that specific year

Table 16: Females; Distribution of Yearly Median Lymphocyte Counts (X1000/mm³) by Year since the First Non-Zero Non-Missing External Gamma Exposure

Numbers of years since the first external gamma exposure	Min	Median	Max	Mean	Standard error of the mean	N*
0	.66	1.63	3.12	1.69	.04	126
1	.64	1.59	3.05	1.60	.03	125
2	.94	1.52	3.51	1.56	.04	120
3	.59	1.48	3.02	1.52	.04	111
4	.76	1.58	3.04	1.66	.04	110
5	.90	1.72	3.95	1.70	.04	101
6	.84	1.67	2.80	1.69	.04	95
7	.62	1.71	3.24	1.70	.05	88
8	.87	1.62	3.72	1.67	.05	86
9	.57	1.70	2.82	1.69	.05	82

*Number of workers who had at least one lymphocyte count in that specific year

The descriptive analysis shows that the yearly mean and median lymphocyte counts drop during the first years following the first external gamma exposure and recover afterwards. This result holds after stratification by sex. The distribution of lymphocyte counts is shown graphically using the histograms drawn as percentages overlapped with normal density plots for the overall study cohort and after stratification by sex (Fig 5-7).

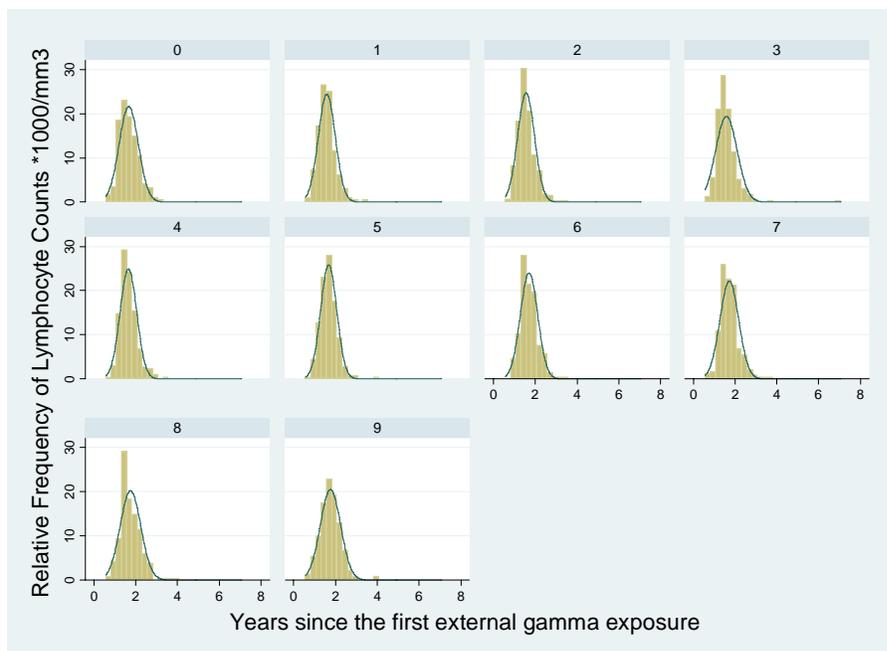


Figure 5: Distribution of Yearly Median Lymphocyte Counts Distribution by Years since the First Non-Zero Non-Missing Exposure (Width of bin=0.25)

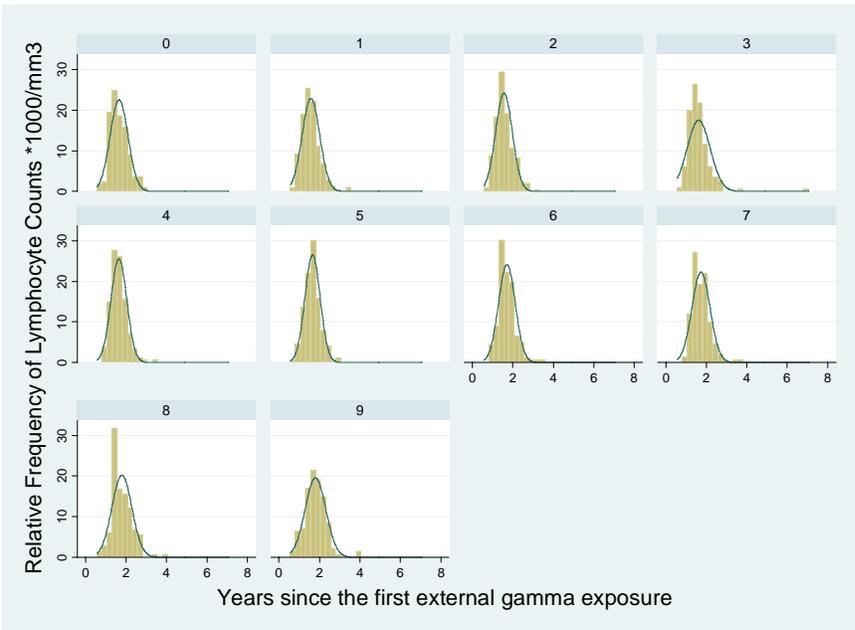


Figure 6: Males; Distribution of Yearly Median Lymphocyte Counts Distribution by Years since the First Non-Zero Non-Missing Exposure (Width of bin=0.25)

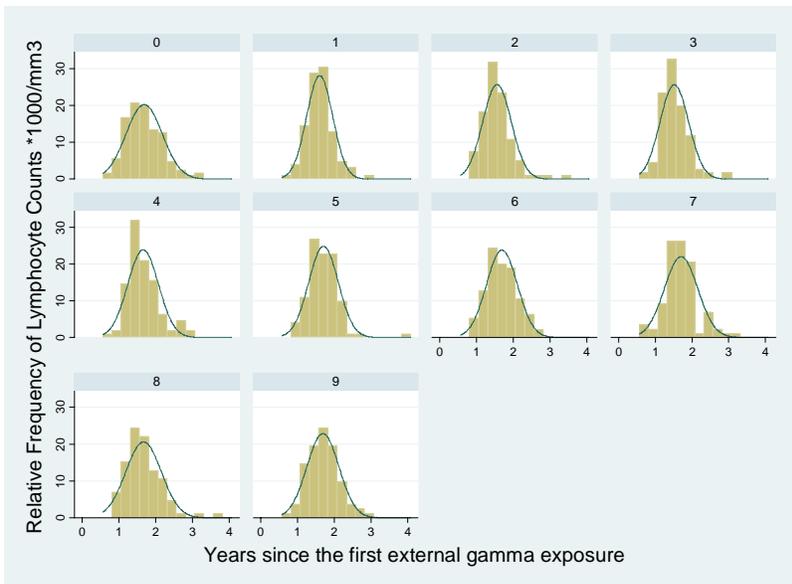


Figure 7: Females; Distribution of Yearly Median Lymphocyte Counts Distribution by Years since the First Non-Zero Non-Missing Exposure (Width of bin=0.25)

Log-transformed lymphocyte counts

In order to meet the assumptions required by the statistical technique, the log-transformation of the yearly median lymphocyte counts is an appropriate procedure. The distributions of the yearly log-transformed median lymphocyte counts are described in each year for ten years starting with the year of the first non-zero non-missing external gamma exposure. Log-transformed yearly median lymphocyte counts are described using the minimum, median, maximum, mean and standard error of the mean. This descriptive analysis is applied to the study cohort, overall, and after stratification by sex (Table 17-19).

Table 17: Distribution of Yearly Log Transformed Median Lymphocyte Counts (X1000/mm³) by Year since the First Non-Zero Non-Missing External Gamma Exposure

Numbers of years since the first external gamma exposure	Min	Median	Max	Mean	Standard error of the mean	N*
0	-.57	.48	1.14	.47	.02	347
1	-.45	.44	1.22	.43	.01	342
2	-.33	.41	1.26	.41	.01	328
3	-.53	.41	1.94	.42	.02	308
4	-.27	.46	1.26	.47	.01	298
5	-.33	.51	1.37	.49	.01	278
6	-.18	.51	1.21	.50	.02	261
7	-.48	.51	1.30	.51	.02	239
8	-.36	.52	1.36	.52	.02	234
9	-.56	.55	1.40	.53	.02	223

*Number of workers who had at least one lymphocyte count in that specific year

Table 18: Males; Distribution of Yearly Log Transformed Median Lymphocyte Counts (X1000/mm³) by Year since the First Non-Zero Non-Missing External Gamma Exposure

Numbers of years since the first external gamma exposure	Min	Median	Max	Mean	Standard error of the mean	N*
0	-0.57	0.47	1.10	0.47	0.02	221
1	-0.33	0.42	1.22	0.42	0.02	217
2	-0.33	0.40	1.12	0.41	0.02	208
3	-0.26	0.43	1.94	0.44	0.02	197
4	-0.13	0.46	1.26	0.47	0.02	188
5	-0.33	0.51	1.11	0.48	0.02	177
6	-0.18	0.50	1.21	0.51	0.02	166
7	-0.17	0.51	1.30	0.52	0.02	151
8	-0.36	0.53	1.36	0.55	0.02	148
9	-0.49	0.59	1.40	0.55	0.02	141

*Number of workers who had at least one lymphocyte count in that specific year

Table 19: Females; Distribution of Yearly Log Transformed Median Lymphocyte Counts (X1000/mm³) by Year since the First Non-Zero Non-Missing External Gamma Exposure

Numbers of years since the first external gamma exposure	Min	Median	Max	Mean	Standard error of the mean	N*
0	-0.42	0.49	1.14	0.48	0.03	126
1	-0.45	0.47	1.11	0.45	0.02	125
2	-0.07	0.42	1.26	0.42	0.02	120
3	-0.53	0.39	1.11	0.39	0.02	111
4	-0.27	0.46	1.11	0.48	0.02	110
5	-0.10	0.54	1.37	0.51	0.02	101
6	-0.17	0.51	1.03	0.49	0.03	95
7	-0.48	0.53	1.18	0.49	0.03	88
8	-0.14	0.48	1.31	0.48	0.03	86
9	-0.56	0.53	1.04	0.49	0.03	82

*Number of workers who had at least one lymphocyte count in that specific year

The log-transformed yearly median lymphocyte counts drop during the first years following the first external gamma exposure and recover afterwards. This result holds after stratification by sex. The distribution of the log-transformed lymphocyte counts is shown graphically using the histograms drawn as percentages overlapped with normal density plots for the overall study cohort and after stratification by sex (Fig 8-10).

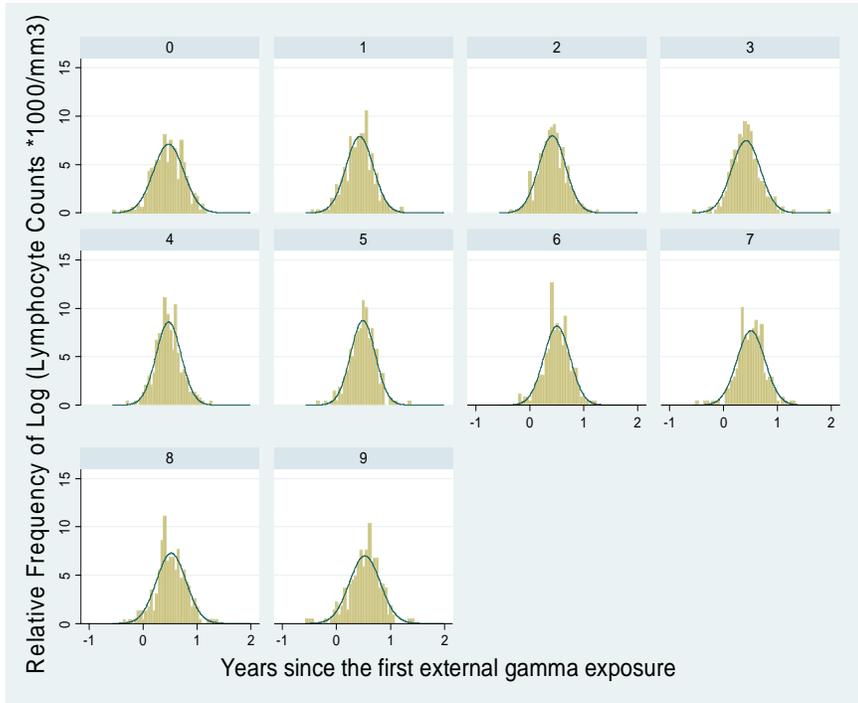


Figure 8: Distribution of Log of Yearly Median Lymphocyte Counts Distribution by Years since the First Non-Zero Non-Missing Exposure (Width of bin=0.05)

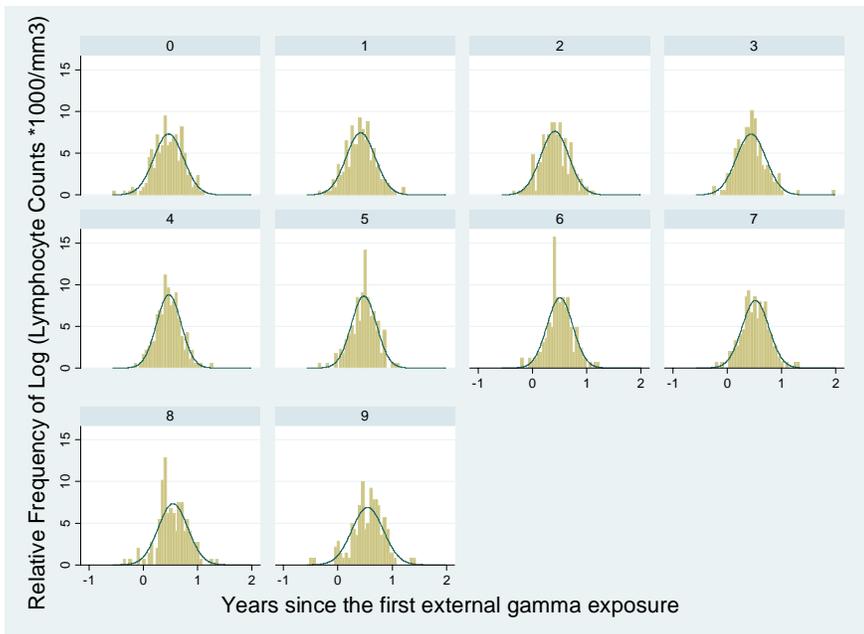


Figure 9: Distribution of Log of Yearly Median Lymphocyte Counts Distribution by Years since the First Non-Zero Non-Missing Exposure - Males (Width of bin=0.05)

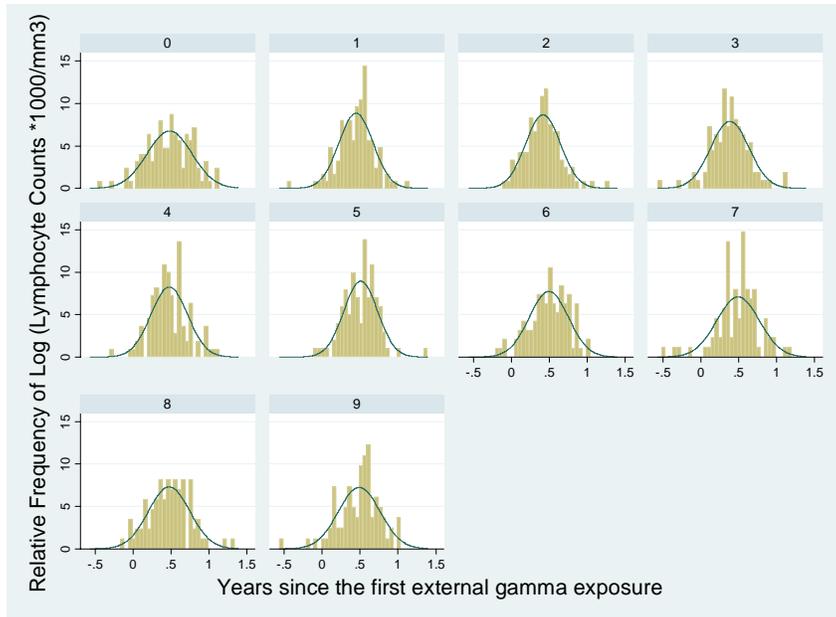


Figure 10: Distribution of Log of Yearly Median Lymphocyte Counts Distribution by Years since the First Non-Zero Non-Missing Exposure in Females (Width of bin=0.05)

c. The yearly cumulative external gamma dose

Yearly cumulative external gamma dose is the main covariate in the study. It is a continuous time dependent variable. The distribution of the yearly cumulative external gamma dose is described in each year for ten years starting with the year of the first non-zero non-missing external gamma exposure using the minimum, median, maximum, mean and standard error of the mean. This descriptive analysis is applied to the study cohort, overall and after stratification by sex (Table 20-22).

Table 20: Distribution of Yearly Cumulative External Gamma Dose (rads) by Year since the First Non-Zero Non-Missing External Gamma Exposure

Numbers of years since the first external gamma exposure	Min	Median	Max	Mean	Standard error of the mean	N*
0	.06	27.70	391.85	46.97	3.26	353
1	0	45.21	795.10	70.96	4.91	341
2	0	25.01	501.90	45.21	3.41	326
3	0	13.65	175.23	26.08	1.84	307
4	0	7.76	81.20	12.72	.84	294
5	0	3.88	92.36	7.915	.71	270
6	0	2.81	59.78	5.62	.53	250
7	0	2.30	92.46	4.67	.57	227
8	0	1.48	45.06	3.65	.40	217
9	0	1.25	34.02	2.96	.33	207

*Number of workers who have non-zero non-missing external gamma dose for that specific year

Table 21: Males; Distribution of Yearly Cumulative External Gamma Dose (rads) by Year since the First Non-Zero Non-Missing External Gamma Exposure

Numbers of years since the first external gamma exposure	Min	Median	Max	Mean	Standard error of the mean	N*
0	.06	30.20	391.85	53.06	4.59	223
1	0	50.02	795.10	80.02	6.89	216
2	0	27.00	501.90	44.20	4.13	207
3	0	15.19	157.34	25.05	2.11	196
4	0	8.15	81.20	14.14	1.13	185
5	0	5.40	92.36	9.49	.95	173
6	0	3.46	59.78	6.60	.73	162
7	0	2.72	92.46	5.64	.84	146
8	0	2.56	45.06	4.50	.55	138
9	0	2.30	34.02	3.79	.47	133

*Number of workers who have non-zero non-missing external gamma dose for that specific year

Table 22: Females; Distribution of Yearly Cumulative External Gamma Dose (rads) by Year since the First Non-Zero Non-Missing External Gamma Exposure

Numbers of years since the first external gamma exposure	Min	Median	Max	Mean	Standard error of the mean	N*
0	.56	17.63	252.62	36.54	3.90	130
1	0	28.49	321.40	55.29	5.94	125
2	0	19.99	326.30	46.98	5.97	119
3	0	10.65	175.23	27.89	3.45	111
4	0	5.36	74.28	10.29	1.19	109
5	0	2.24	78.11	5.11	.95	97
6	0	1.73	38.85	3.84	.63	88
7	0	1.15	20.95	2.94	.47	81
8	0	.52	20.68	2.16	.45	79
9	0	.52	12.75	1.46	.28	74

*Number of workers who have non-zero non-missing external gamma dose for that specific year

The yearly cumulative external gamma dose increases in the early years after first external gamma exposure occurred and drops afterwards. The results hold after stratification by sex. Also, during the ten years of follow-up females receive a lower yearly cumulative external gamma dose than males.

The log-transformed cumulative external gamma dose

The cumulative external gamma dose is used in the statistical analysis after log-transformation. The distribution of the log-transformed yearly cumulative external gamma dose is described in each year for ten years starting with the year of the first non-zero non-missing external gamma exposure. The log-transformed yearly cumulative external gamma exposure is described using the minimum, median, maximum, mean and standard error of the mean. This descriptive analysis is applied to the study cohort, overall and after stratification by sex (Table 23-25).

Table 23: Distribution of Log Transformed Yearly Cumulative External Gamma Dose (rads) by Year since the First Non-Zero Non-Missing External Gamma Exposure

Numbers of years since the first external gamma exposure	Min	Median	Max	Mean	Standard error of the mean	N*
0	-2.81	3.32	5.97	2.97	0.08	353
1	-0.24	3.88	6.68	3.52	0.08	327
2	-3.91	3.33	6.22	3.06	0.09	303
3	-2.81	2.88	5.17	2.68	0.09	269
4	-3.91	2.39	4.40	2.18	0.08	246
5	-3.51	1.81	4.53	1.69	0.09	211
6	-2.21	1.62	4.09	1.53	0.09	178
7	-2.30	1.43	4.53	1.33	0.09	159
8	-2.81	1.27	3.81	1.17	0.09	143
9	-2.30	1.03	3.53	0.95	0.09	138

*Number of workers who had non-zero non-missing external gamma dose in that specific year

Table 24: Males; Distribution of Log Transformed Yearly Cumulative External Gamma Dose (rads) by Year since the First Non-Zero Non-Missing External Gamma Exposure

Numbers of years since the first external gamma exposure	Min	Median	Max	Mean	Standard error of the mean	N*
0	-2.81	3.41	5.97	3.08	0.11	223
1	-0.20	3.99	6.68	3.66	0.10	208
2	-1.83	3.42	6.22	3.12	0.10	193
3	-2.30	2.96	5.06	2.73	0.10	170
4	-1.20	2.57	4.40	2.32	0.09	155
5	-1.90	2.09	4.53	1.94	0.10	137
6	-1.66	1.82	4.09	1.71	0.10	117
7	-1.27	1.62	4.53	1.58	0.10	103
8	-2.81	1.38	3.81	1.40	0.10	97
9	-0.97	1.27	3.53	1.24	0.10	94

*Number of workers who had non-zero non-missing external gamma dose in that specific year

Table 25: Females; Distribution of Log Transformed Yearly Cumulative External Gamma Dose (cGy) by Year since the First Non-Zero Non-Missing External Gamma Exposure

Numbers of years since the first external gamma exposure	Min	Median	Max	Mean	Standard error of the mean	N*
0	-0.58	2.87	5.53	2.78	0.13	130
1	-0.24	3.45	5.77	3.27	0.14	119
2	-3.91	3.20	5.79	2.96	0.16	110
3	-2.81	2.63	5.17	2.60	0.15	99
4	-3.91	2.17	4.31	1.93	0.13	91
5	-3.51	1.35	4.36	1.24	0.15	74
6	-2.21	1.47	3.66	1.18	0.15	61
7	-2.30	1.03	3.04	0.87	0.16	56
8	-1.35	0.46	3.03	0.68	0.17	46
9	-2.30	0.38	2.55	0.34	0.17	44

*Number of workers who had non-zero non-missing external gamma dose in that specific year

The log-transformed yearly cumulative external gamma dose increases in the early years after first external gamma exposure occurred and drops afterwards. The results hold after stratification by sex. Also, as previously mentioned, during the ten years of follow-up females receive a lower yearly cumulative external gamma dose than males.

d. The relationship between lymphocyte counts and external gamma dose

The relationship between the external gamma exposure and the lymphocyte count is illustrated by graphing them together by years since the first external gamma exposure (Fig 11-13). The graphs suggest an inverse dose-response relationship between the yearly median lymphocyte counts and the cumulative external gamma dose: the median lymphocyte counts decreases as the cumulative external gamma dose increases. When the gamma dose decreases the lymphocyte counts seem to recover.

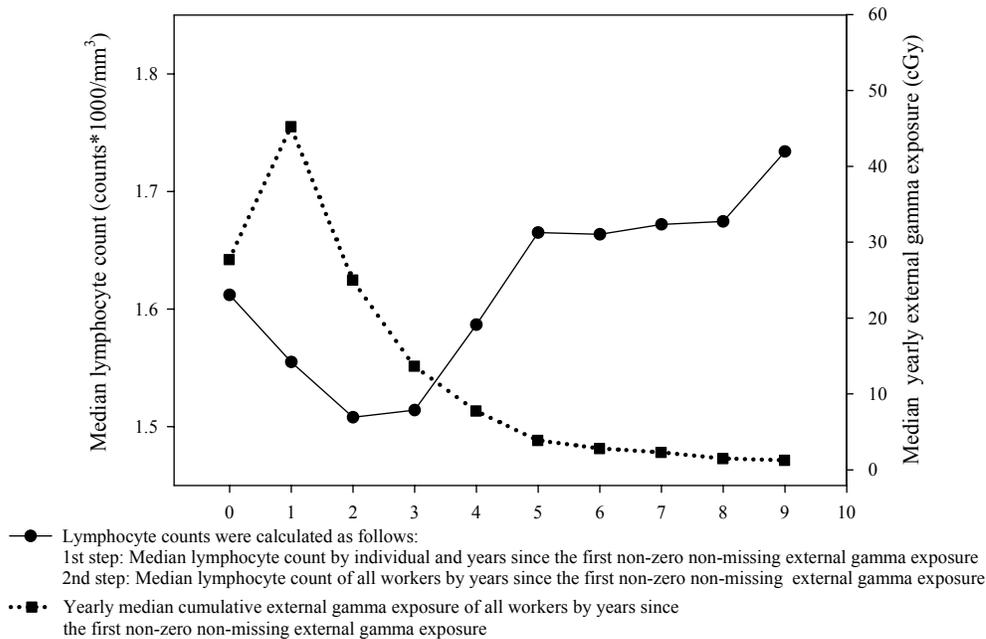


Figure 11: Median Lymphocyte Counts and Median Cumulative External Gamma Dose by Years since the First Non-Zero Non-Missing Gamma Exposure

Furthermore, the descriptive analysis shows differences between males and females. Based on the descriptive analysis females seem to have a greater decrease in the lymphocyte counts in response to a given dose of external gamma radiation.

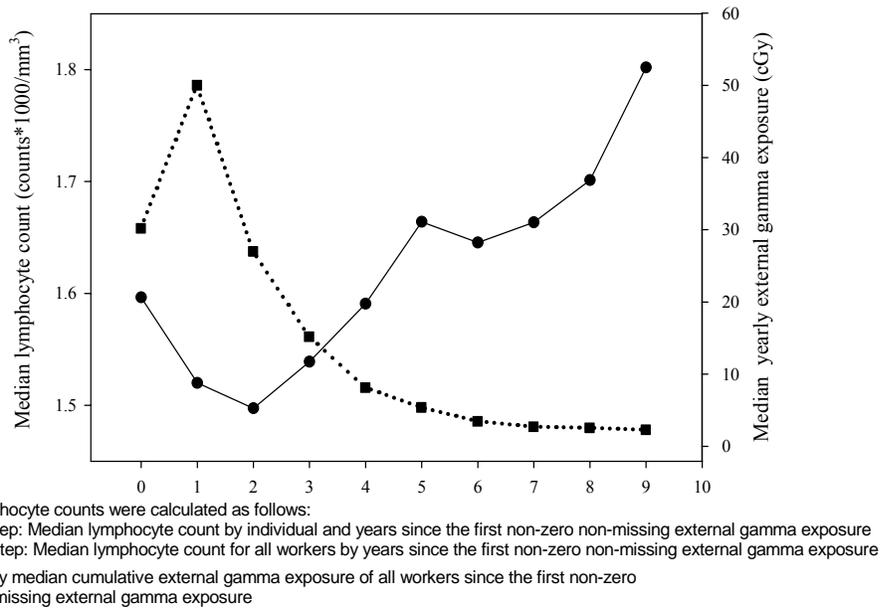


Figure 12: Males; Median Lymphocyte Counts and Median Cumulative External Gamma Dose by Years since the First Non-Zero Non-Missing Gamma Exposure

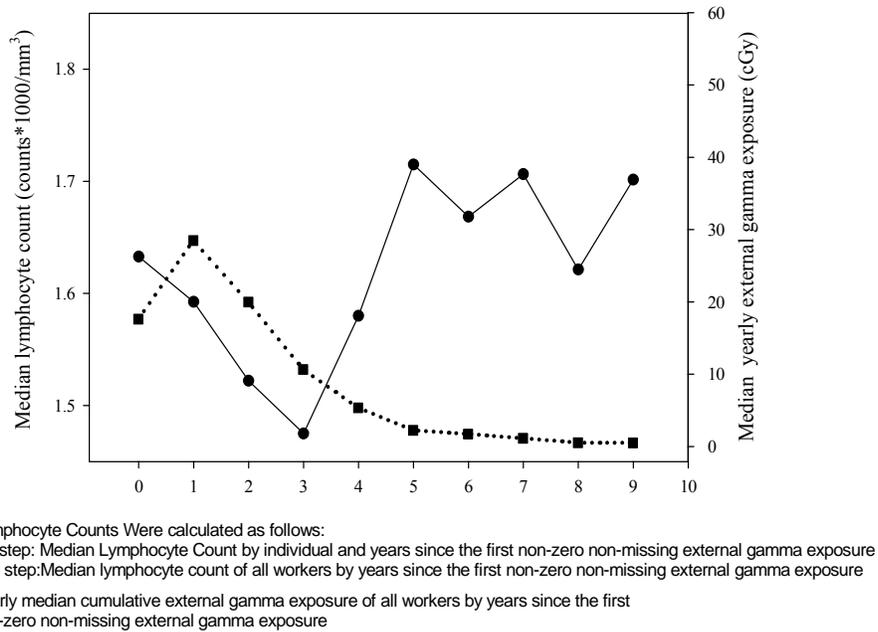


Figure 13: Females; Median Lymphocyte Counts and Median Cumulative External Gamma Dose by Years since the First Non-Zero Non-Missing Gamma Exposure

e. Descriptive analysis of lifestyle variables

The life style variables included in this analysis as time independent variables are smoking history and alcohol consumption at start of employment.

The descriptive analysis of smoking status shows that the majority of workers who had ever smoked are males (97%). It is worthwhile to mention the small number of women with positive smoking history (6 out of 127) and the small number of workers (8 out of 353) with unknown smoking history (Table 26).

Table 26: Absolute and Relative Frequency Distribution of Workers by Smoking History and Sex

Smoking history	Total		Males		Females	
	n	(%)	n	(%)	n	(%)
Ever smoked	176	(100.0)	170	(96.6)	6	(3.4)
Never smoked	169	(100.0)	48	(28.4)	121	(71.6)
Total	345*	(100.0)	218	(63.2)	127	(36.8)

* 8 workers have missing data about smoking history

The majority of workers known as alcohol consumers at start of employment are males (96%). There are a small number of women (7 out of 123) known as alcohol consumers at start of employment. There are 20 out of 353 workers with unknown alcohol consumption status at start of employment (Table 27).

Table 27: Absolute and Relative Frequency Distribution of Workers by Alcohol Consumption at Start of Employment and Sex

Alcohol consumption at start of employment	Total		Males		Females	
	n	(%)	n	(%)	n	(%)
Yes	177	(100.0)	170	(96.1)	7	(3.9)
No	156	(100.0)	40	(25.6)	116	(74.4)
Total	333*	(100.0)	210	(63.1)	123	(36.9)

*20 workers have missing data about alcohol consumption at start of employment

f. Selection of the baseline lymphocyte count used in the models

In order to perform a statistical analysis of the gamma radiation effect on the lymphocytes it is important to select a lymphocyte count baseline and to adjust for it in the analysis. A possible strategy is to choose the baseline lymphocyte count as the last lymphocyte count in the year previous to the first external gamma exposure. However, there are only 119 out of 353 workers who have records of lymphocyte counts in the year preceding the first external gamma exposure while as many as 221 workers have their first lymphocyte count recorded in the year of the first external gamma exposure. Thus, 221 workers do not have any lymphocyte count recorded in the year preceding the first external gamma exposure.

One possibility for selection of a baseline count in these 221 workers is to choose as baseline, the first count in the year of the first external gamma exposure. It is possible that the first lymphocyte count in the year of the first external gamma exposure is performed after the gamma exposure started and, therefore, it is already affected by radiation exposure, thereby, introducing bias in the baseline lymphocyte count. In order to check whether the first lymphocyte count performed during the first year of exposure differs from the last lymphocyte count in the year preceding the first external gamma exposure, an additional analysis was conducted. This analysis consists of comparing the last lymphocyte count in the year preceding the first exposure with the first lymphocyte count in the year when the first exposure occurred in the 119 workers for whom data are available to make the comparison. In these 119 workers the distribution of the last lymphocyte counts in the year preceding the first external gamma exposure and the distribution of the first counts in the year of the external gamma exposure is described and compared (Table 28, Fig 14, 15).

Table 28: Distribution of Two Groups of Lymphocyte Counts in the same 119 Workers:
 Last Lymphocyte Count in the Year Preceding the First External Gamma Exposure and the First
 Lymphocyte Count in the Year of the First external Gamma Exposure

	Min	Median	Max	Mean	Standard error of the mean	N
Last count in the year preceding the first exposure	.8	1.8	4.9	1.9	.064	119
First count in the year when the first exposure occurred	.72	1.7	4	1.8	.059	119
Total	.72	1.7	4.9	1.8	.044	238

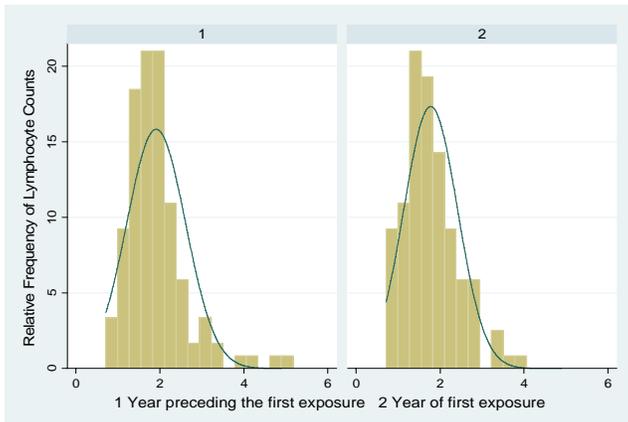


Figure 14: Relative Distribution of Two Groups of Lymphocyte Counts Recorded in the same 119 Workers:

Last Lymphocyte Count in the Year Preceding the First External Gamma Exposure (1) and the First Lymphocyte Count in the Year of the First external Gamma Exposure (2)

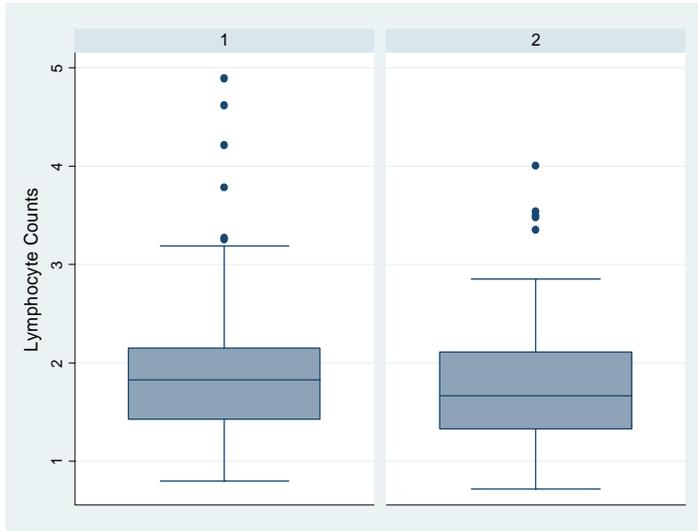


Figure 15: Distribution of Two Groups of Lymphocyte Counts Recorded in the same 119 Workers:

Last Lymphocyte Count in the Year Preceding the First External Gamma Exposure (1) and of the First Lymphocyte Count in the Year of the First external Gamma Exposure (2)

The table, histograms and the boxplots presented above indicate that the two lymphocyte count distributions can be considered similar despite small differences between them. Furthermore, the two distributions in the same 119 workers were compared using a non-parametric test for paired data since the counts were not normally distributed. The sign rank test showed that they can be considered similar ($p\text{-value}=0.08$).

Based on these findings, and in order to achieve consistency across all workers in selecting the baseline, the baseline lymphocyte count is chosen in each worker as the first count recorded in the year of the first external gamma exposure. Setting up the baseline lymphocyte count in the year of the first external gamma exposure may introduce a bias since this lymphocyte count could be already affected by the gamma exposure. However, the baseline is chosen as the first blood count in the year of the first gamma exposure and the statistical analysis adjusts for the baseline in the same way for all workers. Thus, the exposure effect on the chosen baseline count is expected to be minimal.

Baseline lymphocyte counts distribution by categorical covariates

The distribution of the baseline lymphocyte count is similar in males and females (Table 29).

Table 29: Distribution of Baseline* Lymphocyte Counts by Sex

Sex	Min	Median	Max	Mean	Standard Error of the Mean	N
Males	.63	1.6	4	1.7	.038	221
Females	.66	1.6	3.3	1.8	.052	126
Total	.63	1.6	4	1.8	.031	347

*The baseline count is the first non-missing count during the year of the first external gamma exposure

The distribution of the baseline lymphocyte count is also similar in workers with positive smoking history and in workers with negative smoking history. The similarity holds after stratification by sex. Six workers (3 males and 3 females) with unknown smoking history have higher values for the baseline lymphocyte counts. The number of missing smoking history data is small (6 out of 347) and therefore it is not expected to have a significant influence on the analysis (Table 30-32).

Table 30: Distribution of Baseline* Lymphocyte Counts by Smoking History

Smoking history	Min	Median	Max	Mean	Standard Error of the Mean	N
Ever smoked	.63	1.6	4	1.8	.044	176
Never smoked	.66	1.6	3.3	1.7	.045	165
Unknown	1.1	1.5	2.1	1.6	.18	6
Total	.63	1.6	4	1.8	.031	347

*The baseline count is the first non-missing count during the year of the first external gamma exposure

Table 31: Males; Distribution of Baseline* Lymphocyte Counts by Smoking History

Smoking history	Min	Median	Max	Mean	Standard Error of the Mean	N
Ever smoked	.63	1.6	4	1.8	.044	170
Never smoked	.69	1.7	2.8	1.7	.076	48
Unknown	1.1	1.1	1.7	1.3	.2	3
Total	.63	1.6	4	1.7	.038	221

*The baseline count is the first non-missing count during the year of the first external gamma exposure

Table 32: Females; Distribution of Baseline* Lymphocyte Counts by Smoking History

Smoking history	Min	Median	Max	Mean	Standard Error of the Mean	N
Ever smoked	.72	1.4	2.4	1.5	.25	6
Never smoked	.66	1.6	3.3	1.8	.054	117
Unknown	1.3	2	2.1	1.8	.23	3
Total	.66	1.6	3.3	1.8	.052	126

*The baseline count is the first non-missing count during the year of the first external gamma exposure

The distribution of the baseline lymphocyte counts do not seem to differ by alcohol consumption status at start of employment. After stratification by sex, there are some small differences in the baseline lymphocyte counts by alcohol consumption status (Table 33-35). As previously mentioned, the number of women who drank alcohol at start of employment is small relative to the total number of women (7 out of 126).

Table 33: Distribution of Baseline* Lymphocyte Counts by Alcohol Consumption at Start of Employment

Alcohol consumption at start of Employment	Min	Median	Max	Mean	Standard Error of the Mean	N
Yes	.63	1.6	4	1.7	.045	177
No	.66	1.7	3.3	1.8	.046	151
Unknown	.87	1.6	2.7	1.7	.11	19
Total	.63	1.6	4	1.8	.031	347

*the baseline count is considered the first non-missing count during the year of the first external gamma exposure

Table 34: Males; Distribution of Baseline* Lymphocyte Counts by Alcohol Consumption at Start of Employment

Alcohol consumption at start of Employment	Min	Median	Max	Mean	Standard Error of the Mean	N
Yes	.63	1.6	4	1.8	.046	170
No	.77	1.7	2.8	1.8	.078	39
Unknown	1.1	1.6	2.2	1.6	.094	12
Total	.63	1.6	4	1.7	.038	221

*The baseline count is the first non-missing count during the year of the first external gamma exposure

Table 35: Females; Distribution of Baseline* Lymphocyte Counts by Alcohol Consumption at Start of Employment

Alcohol consumption at start of Employment	Min	Median	Max	Mean	Standard Error of the Mean	N
Yes	.93	1.5	2.1	1.5	.15	7
No	.66	1.6	3.3	1.8	.056	112
Unknown	.87	1.6	2.7	1.7	.27	7
Total	.66	1.6	3.3	1.8	.052	126

*The baseline count is the first non-missing count during the year of the first external gamma exposure

g. Work location variable description

Work location is a categorical time dependent covariate used in this statistical analysis. It is defined as the plant where a worker spent the most time during a specific year. Since it is analyzed as a time dependent variable, it is interesting to describe the dynamic changes in workers assignments by plant during the follow-up. As shown in Table 36, most of the workers did not change work location across the ten years of follow-up (199 out of 352).

Table 36: Frequency Distribution of Workers by Number of Changes of Work Location and Sex; Work location represents the plant where the workers spent the most time during a specific year

Number of Work Location Changes	Total n (%)	Males n (%)	Females n (%)
0	199 (100.0)	131 (65.8)	68 (34.2)
1	128 (100.0)	77 (60.2)	51 (39.8)
2	23 (100.0)	12 (52.2)	11 (47.8)
3	2 (100.0)	2 (100.0)	0 (0.0)
Total	352*(100.0)	222 (63.1)	130 (36.9)

*1 worker (male) has no work location data

The frequency distribution of workers by year since the first external gamma exposure and work location across follow-up time shows a relatively stable percent of workers assigned to the reactor and plutonium plant. The percentage of workers assigned to the radio-chemical plant decreases across follow-up time while the percentage of workers assigned to other work locations or having missing observations regarding the work location increases across the follow-up time. After stratification by sex the described pattern still holds. However it is interesting to note that during the first years of follow-up the percentages of females working at the radio-

chemical and plutonium plant were higher than the percentages of males assigned at these hazardous work locations (Tables 37-39).

Table 37: Absolute and Relative Frequency Distribution of Workers by Work Location and Years following the First External Gamma Exposure

Work location represents the plant where the workers spent the most time during a specific year

Numbers of years since the first external gamma exposure	Total n (%)	Reactor n (%)	Radio-chemical n (%)	Plutonium n (%)	Others n (%)	Missing Data n (%)
0	353 (100.0)	89 (25.2)	203 (57.5)	32 (9.1)	27 (7.7)	2 (0.5)
1	343 (100.0)	85 (24.8)	214 (62.4)	31 (9.0)	12 (3.5)	1 (0.3)
2	328 (100.0)	82 (25.1)	191 (58.2)	29 (8.8)	25 (7.6)	1 (0.3)
3	308 (100.0)	79 (25.7)	162 (52.6)	27 (8.8)	36 (11.6)	4 (1.3)
4	299 (100.0)	79 (26.4)	134 (44.8)	23 (7.7)	57 (19.1)	6 (2.0)
5	278 (100.0)	73 (26.3)	98 (35.2)	25 (9.0)	73 (26.3)	9 (3.2)
6	263 (100.0)	68 (25.9)	84 (31.9)	25 (9.6)	74 (28.1)	12 (4.5)
7	239 (100.0)	62 (25.9)	70 (29.3)	22 (9.2)	73 (30.5)	12 (5.1)
8	234 (100.0)	63 (26.9)	63 (26.9)	20 (8.6)	72 (30.8)	16 (6.8)
9	223 (100.0)	55 (24.7)	59 (26.5)	19 (8.5)	75 (33.6)	15 (6.7)
Total	2,868 (100.0)	735 (25.6)	1,278 (44.6)	253 (8.8)	524 (18.3)	78 (2.7)

Table 38: Males; Study Cohort: Absolute and Relative Frequency Distribution of Workers by Work Location and Years following the First External Gamma Exposure

Work location represents the plant where the workers spent the most time during a specific year

Numbers of years since the first external gamma exposure	Total n (%)	Reactor n (%)	Radio-chemical n (%)	Plutonium n (%)	Others n (%)	Missing Data n (%)
0	223 (100.0)	65 (29.2)	119 (53.4)	19 (8.4)	19 (8.5)	1 (0.5)
1	217 (100.0)	61 (28.0)	129 (59.5)	18 (8.3)	8 (3.7)	1 (0.5)
2	208 (100.0)	62 (29.8)	113 (54.3)	17 (8.2)	15 (7.2)	1 (0.5)
3	197 (100.0)	58 (29.4)	95 (48.2)	15 (7.6)	25 (12.8)	4 (2.0)
4	188 (100.0)	56 (29.8)	81 (43.1)	13 (6.9)	34 (18.1)	4 (2.1)
5	177 (100.0)	53 (29.9)	66 (37.3)	14 (7.9)	40 (22.6)	4 (2.3)
6	167 (100.0)	49 (29.3)	55 (32.9)	14 (8.4)	45 (27.0)	4 (2.4)
7	151 (100.0)	44 (29.1)	46 (30.5)	12 (8.0)	44 (29.1)	5 (3.3)
8	148 (100.0)	43 (29.1)	45 (30.4)	10 (6.7)	41 (27.7)	9 (6.1)
9	141 (100.0)	38 (27.0)	42 (29.8)	10 (7.1)	44 (31.2)	7 (4.9)
Total	1,817 (100.0)	529 (29.1)	791 (43.6)	142 (7.8)	315 (17.3)	40 (2.2)

Mayak PA workers were not only exposed to gamma radiation but also to plutonium radiation which can also affect blood counts.

Table 39: Females Absolute and Relative Frequency Distribution of Workers by Work Location and Years following the First External Gamma Exposure

Work location represents the plant where the workers spent the most time during a specific year

Numbers of years since the first external gamma exposure	Total n (%)	Reactor n (%)	Radio-chemical n (%)	Plutonium n (%)	Others n (%)	Missing Data n (%)
0	130 (100.0)	24 (18.5)	84 (64.5)	13 (10.0)	8 (6.2)	1 (0.8)
1	126 (100.0)	24 (19.1)	85 (67.4)	13 (10.3)	4 (3.2)	0 (0.0)
2	120 (100.0)	20 (16.7)	78 (65.0)	12 (10.0)	10 (8.3)	0 (0.0)
3	111 (100.0)	21 (18.9)	67 (60.4)	12 (10.8)	11 (9.9)	0 (0.0)
4	111 (100.0)	23 (20.7)	53 (47.8)	10 (9.0)	23 (20.7)	2 (1.8)
5	101 (100.0)	20 (19.8)	32 (31.7)	11 (10.9)	33 (32.7)	5 (5.0)
6	96 (100.0)	19 (19.7)	29 (30.2)	11 (11.5)	29 (30.2)	8 (8.3)
7	88 (100.0)	18 (20.5)	24 (27.3)	10 (11.3)	29 (32.9)	7 (8.0)
8	86 (100.0)	20 (23.3)	18 (20.9)	10 (11.6)	31 (36.1)	7 (8.1)
9	82 (100.0)	17 (20.7)	17 (20.7)	9 (11.0)	31 (37.8)	8 (9.8)
Total	1,051 (100.0)	206 (19.6)	487 (46.3)	111 (10.6)	209 (19.9)	38 (3.6)

Therefore, it is important for this analysis to adjust for plutonium exposure. In order to do so, the Mayak PA plants are categorized according to the plutonium exposure in 3 groups as follows: category 0 meaning no plutonium or low plutonium exposure consists of two locations: the reactor and “others”. The radio-chemical plant is categorized as 1 meaning moderate plutonium exposure. The plutonium plant is categorized as 2 meaning high plutonium exposure. The frequency distribution of workers by years since the first external gamma exposure and work location categories related to plutonium (Pu) exposure is shown for the study cohort and after stratification by sex in Tables 40-42.

Table 40: Absolute and Relative Frequency Distribution of Workers by Work Location Related to Pu exposure and Years following the First External Gamma Exposure

Numbers of years since the first external gamma exposure	Total	Reactor and other locations combined (0 or low Pu exposure)	Radio-chemical (Moderate Pu exposure)	Plutonium (High Pu exposure)	Missing Data
0	353 (100.0)	116 (32.8)	203 (57.5)	32 (9.1)	2 (0.6)
1	343 (100.0)	97 (28.3)	214 (62.4)	31 (9.0)	1 (0.3)
2	328 (100.0)	107 (32.7)	191 (58.2)	29 (8.8)	1 (0.3)
3	308 (100.0)	115 (37.3)	162 (52.6)	27 (8.7)	4 (1.4)
4	299 (100.0)	136 (45.5)	134 (44.8)	23 (7.7)	6 (2.0)
5	278 (100.0)	146 (52.5)	98 (35.3)	25 (9.0)	9 (3.2)
6	263 (100.0)	142 (54.0)	84 (31.9)	25 (9.5)	12 (4.6)
7	239 (100.0)	135 (56.5)	70 (29.3)	22 (9.2)	12 (5.0)
8	234 (100.0)	135 (57.7)	63 (26.9)	20 (8.6)	16 (6.8)
9	223 (100.0)	130 (58.3)	59 (26.5)	19 (8.5)	15 (6.7)
Total	2,868 (100.0)	1,259 (43.9)	1,278 (44.6)	253 (8.8)	78 (2.7)

Table 41: Males; Absolute and Relative Frequency Distribution of Workers by Work Location Related to Pu exposure and Years following the First External Gamma Exposure

Numbers of years since the first external gamma exposure	Total	Reactor and other locations combined (0 or low Pu)	Radio-chemical (Moderate Pu)	Plutonium (High Pu)	Missing Data
0	223 (100.0)	84 (37.6)	119 (53.4)	19 (8.5)	1 (0.5)
1	217 (100.0)	69 (31.7)	129 (59.5)	18 (8.3)	1 (0.5)
2	208 (100.0)	77 (37.0)	113 (54.3)	17 (8.2)	1 (0.5)
3	197 (100.0)	83 (42.1)	95 (48.2)	15 (7.6)	4 (2.1)
4	188 (100.0)	90 (47.9)	81 (43.1)	13 (6.9)	4 (2.1)
5	177 (100.0)	93 (52.5)	66 (37.3)	14 (7.9)	4 (2.3)
6	167 (100.0)	94 (56.3)	55 (32.9)	14 (8.4)	4 (2.4)
7	151 (100.0)	88 (58.3)	46 (30.5)	12 (8.0)	5 (3.2)
8	148 (100.0)	84 (56.7)	45 (30.4)	10 (6.8)	9 (6.1)
9	141 (100.0)	82 (58.2)	42 (29.7)	10 (7.1)	7 (5.0)
Total	1,817(100.0)	844 (46.5)	791 (43.5)	142 (7.8)	40 (2.2)

Table 42: Females; Absolute and Relative Frequency Distribution of Workers by Work Location Related to Pu exposure and Years following the First External Gamma Exposure

Numbers of years since the first external gamma exposure	Total	Reactor and other locations combined (0 or low Pu)	Radio-chemical (Moderate Pu)	Plutonium (High Pu)	Missing Data
0	130 (100.0)	32 (24.6)	84 (64.6)	13 (10.0)	1 (0.8)
1	126 (100.0)	28 (22.2)	85 (67.5)	13 (10.3)	0 (0.0)
2	120 (100.0)	30 (25.0)	78 (65.0)	12 (10.0)	0 (0.0)
3	111 (100.0)	32 (28.8)	67 (60.4)	12 (10.8)	0 0.0
4	111 (100.0)	46 (41.4)	53 (47.8)	10 (9.0)	2 (1.8)
5	101 (100.0)	53 (52.5)	32 (31.6)	11 (10.9)	5 (5.0)
6	96 (100.0)	48 (50.0)	29 (30.2)	11 (11.5)	8 (8.3)
7	88 (100.0)	47 (53.4)	24 (27.2)	10 (11.4)	7 (8.0)
8	86 (100.0)	51 (59.3)	18 (20.9)	10 (11.7)	7 (8.1)
9	82 (100.0)	48 (58.5)	17 (20.7)	9 (11.0)	8 (9.8)
Total	1,051(100.0)	415 (39.4)	487 (46.3)	111 (10.6)	38 (3.6)

The frequency distribution of workers by years since the first external gamma exposure and work location categories related to plutonium exposure shows a relatively stable percent across the ten years of follow-up of workers assigned to the plutonium plant. The percentage of workers assigned to the radio-chemical plant decreases across follow-up time while the percentage of workers assigned to the reactor plant combined with other work locations or having missing observations regarding the work location increases across the follow-up time. After stratification by sex the described pattern still holds.

Baseline count by work location categories related to plutonium exposure

It is also important to describe the distribution of the baseline lymphocyte counts by work location categories related to plutonium exposure at the beginning of follow-up. The distribution of the first lymphocyte count during the year of the first external gamma exposure (the baseline count) is described by work location categories in the first year of follow-up which is the year of the first external gamma exposure (Tables 43-45). The distribution of the baseline lymphocyte counts appears similar in all three work location categories. The number of missing observations is small (2 out of 347).

Table 43: Distribution of Baseline* Lymphocyte Counts by Work Location Related to Pu exposure during the first year of follow-up

Work location categories	Min	Median	Max	Mean	Standard Error of the Mean	N
Reactor and other locations combined (0 or low Pu)	.66	1.6	4	1.8	.056	115
Biochemical (Moderate Pu)	.7	1.7	3.3	1.8	.04	198
Plutonium (High Pu)	.63	1.5	2.7	1.6	.089	32
Missing data	1.7	2.4	3.1	2.4	.71	2
Total	.63	1.6	4	1.8	.031	347

*the baseline count is considered the first non-missing count during the year of the first external gamma exposure

Table 44: Males; Distribution of Baseline* Lymphocyte Counts by Work Location Related to Pu exposure during the first year of follow-up

Work location categories	Min	Median	Max	Mean	Standard Error of the Mean	N
Reactor and other locations combined (0 or low Pu)	.69	1.7	4	1.8	.068	83
Biochemical (Moderate Pu)	.7	1.7	3.1	1.7	.05	118
Plutonium (High Pu)	.63	1.5	2.7	1.7	.11	19
Missing data	1.7	1.7	1.7	1.7	.	1
Total	.63	1.6	4	1.7	.038	221

*the baseline count is considered the first non-missing count during the year of the first external gamma exposure

Table 45: Females; Distribution of Baseline* Lymphocyte Counts by Work Location Related to Pu exposure during the first year of follow-up;

Work location categories	Min	Median	Max	Mean	Standard Error of the Mean	N
Reactor and other locations combined (0 or low Pu)	.66	1.5	2.9	1.7	.095	32
Biochemical (Moderate Pu)	.73	1.7	3.3	1.8	.067	80
Plutonium (High Pu)	.72	1.4	2.6	1.5	.15	13
Missing data	3.1	3.1	3.1	3.1	.	1
Total	.66	1.6	3.3	1.8	.052	126

*the baseline count is considered the first non-missing count during the year of the first external gamma exposure

The results of the descriptive analysis show that assumptions required by the statistical methods are met and the proposed statistical analysis is appropriate and doable. The results of the statistical analysis presented will address the questions explained in the methodology section.

3.2 MISSING DATA ASSESSMENT RESULTS

3.2.1 Background

Missing data always represents an important issue for longitudinal data statistical analysis due especially to the bias that missing data can introduce into the estimators⁴³. First, missing data have to be described and assessed in order to check whether the assumptions required by the statistical techniques hold. Second, if the required assumptions are not met additional procedures like imputation of missing values may be necessary.

The missing data are characterized by patterns and mechanisms:

The patterns describe which values are observed and which values are missing in the data matrix. The missing data patterns most often observed in longitudinal data analysis consist of drop-outs, late entries and gaps in information.

The drop-outs occur when the subjects leave the study prematurely and do not return.

The late entries occur when the subjects are not followed-up from the beginning of the study since they enter the study later on.

Gaps occur when the subjects have missing data at one or more points in time across the follow-up period but they return in the study.

The mechanisms describe the relationship between missingness and the values of variables in the data matrix⁶⁰. The missing data mechanisms have been studied more deeply after Rubin elaborated his theory using missing data indicators and their distribution⁶¹.

In order to categorize the missing data mechanisms it is necessary to introduce the following notation:

Let complete data be defined as: $Y = (y_{ij})$

Let the missing data indicator matrix be defined as: $M = (M_{ij})$

Let the conditional distribution of M given Y be defined as:

$f(M | Y, \phi)$, where ϕ is the unknown parameter.

The missing data mechanisms can be categorized as follows:

a. Missing completely at random (MCAR) if the missingness does not depend on the values of the data, missing or observed:

$$f(M | Y, \phi) = f(M | \phi) \text{ for all } y, \phi$$

b. Missing at random (MAR) if the missingness depends only on the observed data values and not on the missing ones:

$$f(M | Y, \phi) = f(M | Y_{obs}, \phi) \text{ for all } y_{obs}, \phi$$

c. Not missing at random (NMAR) if the distribution of M depends on the missing values

When data are not missing at random there are three important consequences for longitudinal data analysis⁴²:

-loss of information, meaning a reduction in efficiency or a drop in the precision with which changes in the mean response over time can be estimated

-under certain circumstances bias and thereby misleading inferences can occur

-unbalanced data across time since the subjects have different numbers of repeated measurements

As specified in the methodology section, the outcome variable for the analysis is the log-transformed yearly median lymphocyte count in each worker. The explanatory covariate is the log-transformed yearly cumulative external gamma dose received by each worker. The outcome variable and the explanatory covariate are both analyzed as time dependent continuous variables.

In this dissertation the missing data analysis consists of the description and assessment of the missing lymphocyte counts in the study cohort.

3.2.2 Descriptive analysis of the missing data patterns

The first step of missing data analysis is to assess descriptively the missing data patterns for the study cohort, overall and after stratification by sex. Missing data are assessed according to the existence of records in each year from the first external gamma exposure. The missing data patterns in the study cohort are described in Tables 46-48.

Table 46: Absolute and Relative Frequency Distribution of the Missing Data Patterns

Missing data patterns	Absolute frequency	Relative frequency (%)	Cumulative frequency (%)
No missing years	187	53.0	53.0
Dropouts only	113	32.0	85.0
Late entries only	0	0.0	85.0
Mixed patterns	53	15.0	100.0
Total	353	100.0	

Table 47: Males; Absolute and Relative Frequency Distribution of the Missing Data Patterns

Missing data patterns	Absolute frequency	Relative frequency (%)	Cumulative frequency (%)
No missing years	117	52.5	52.5
Dropouts only	72	32.3	84.8
Late entries only	0	0.0	84.8
Mixed patterns	34	15.2	100.0
Total	223	100.0	

Table 48: Females; Absolute and Relative Frequency Distribution of the Missing Data Patterns

Missing data patterns	Absolute frequency	Relative frequency (%)	Cumulative frequency (%)
No missing years	70	53.8	53.8
Dropouts only	41	31.5	85.3
Late entries only	0	0.0	85.3
Mixed patterns	19	14.7	100
Total	130	100.0	

Among the 353 subjects, 187 (53%) have data recorded for each year since the first external gamma exposure. The most of the missing data patterns consist of dropouts; 113 (32%) workers in the study cohort drop the analysis after a number of years since the first external gamma exposure occurred. Since the workers are lined-up according to their first non-zero non-missing external gamma exposure, there are data in all 353 workers in the first year of follow-up. All 353 workers have an external gamma dose recorded in the first year and therefore there are no late entries in the study cohort.

Stratification by sex shows that the relative frequency distribution of missing data patterns looks similar in males and females.

3.2.3 Missing data patterns and mechanisms assessment

3.2.3.1 Drop-outs description

The frequency distributions of drop-outs are analyzed in more detail in the study cohort and after stratification by sex since it was shown that they are the most representative missing data patterns. Tables 49-51 and Fig 16 presents the frequency distribution of workers by number of years from the first external gamma exposure when they drop-out. Tables 4-6 and Fig 1 includes workers who drop-out only.

Table 49: Drop-Outs Only: Absolute and Relative Distribution of Drop-out Patterns by Year of Follow-up when the Drop-out Occurs

Follow-up year when workers drop-out	Absolute frequency	Relative frequency (%)	Cumulative frequency (%)
1 st year	5	4.4	4.4
2 nd year	11	9.7	14.1
3 rd year	15	13.3	27.4
4 th year	15	13.3	40.7
5 th year	15	13.3	54.0
6 th year	12	10.6	64.6
7 th year	16	14.2	78.8
8 th year	12	10.6	89.4
9 th year	12	10.6	100
Total	113	100	

Table 50: Males; Drop-Outs Only: Absolute and Relative Distribution of Drop-out Patterns by Year of Follow-up when the Drop-out Occurs

Follow-up year when workers drop-out	Absolute frequency	Relative frequency (%)	Cumulative frequency (%)
1 st year	2	2.8	2.8
2 nd year	7	9.7	12.5
3 rd year	7	9.7	22.2
4 th year	13	18.1	40.3
5 th year	6	8.3	48.6
6 th year	10	13.9	62.5
7 th year	10	13.9	76.4
8 th year	9	12.5	88.9
9 th year	8	11.1	100
Total	72	100	

Table 51: Females; Drop-Outs Only: Absolute and Relative Distribution of Drop-out Patterns by Year of Follow-up when the Drop-out Occurs

Follow-up year when workers drop-out	Absolute frequency	Relative frequency (%)	Cumulative frequency (%)
1 st year	3	7.3	7.3
2 nd year	4	9.7	17.0
3 rd year	8	19.5	36.5
4 th year	2	4.9	41.4
5 th year	9	22.0	63.4
6 th year	2	4.9	68.3
7 th year	6	14.6	82.9
8 th year	3	7.3	90.2
9 th year	4	9.8	100
Total	41	100	

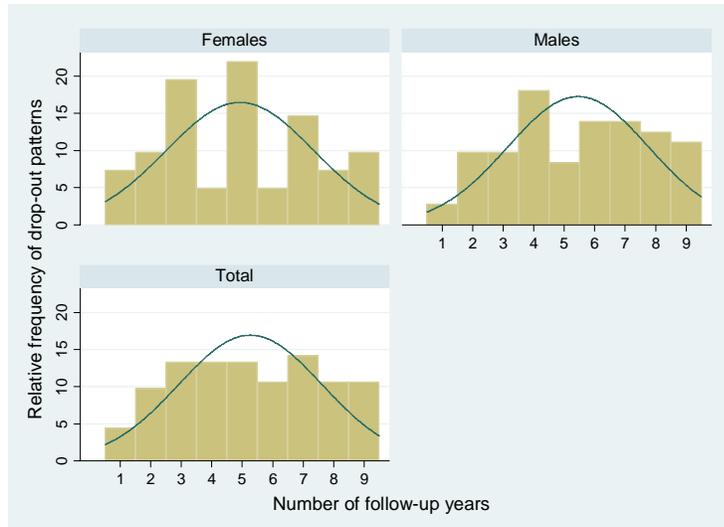


Figure 16: Workers from Study Cohort Who Drop Out: Histogram of the Relative Frequency of Drop-Out Patterns by Sex; Drop Out Patterns are Categorized According to the Number of Follow-up Year

Tables 49-51 and Fig 16 show that in the study cohort most of the drop-outs occur between the 3rd and the 5th years of follow-up and in the 7th year of follow-up. The distribution of drop-out patterns differ by sex as follows: between the 2nd and the 9th follow-up year, males seem to have a more stable drop-out rate than women

3.2.3.2 Drop-outs mechanism assessment

The drop-out is the missing data pattern which occurs most often in the study cohort. According to the assumptions required by the statistical techniques, it is of interest to assess if the drop-outs mechanism can be considered completely at random in the study cohort overall and after stratification by sex.

The data are considered complete if there are at least one measurement recorded on each subject at a common set of times t_j ($j=1, \dots, n$). Complete data consists of two-way array of

measurements y_{ij} ($i=1, \dots, m; j=1, \dots, n$) in which y_{ij} represents the j -th measurement performed on the i -th worker.

The drop-outs consist of premature termination of some of the measurement sequences. The drop-outs lead to an incomplete array of measurements

y'_{ij} where ($i=1, \dots, m; j=1, \dots, n_i$) in which y'_{ij} represents the j -th measurement performed on the i -th worker.

y'_{ij} corresponds to each measurement taken at time t_j , where $t_j < n$.

A strong and important hypothesis to be checked is whether the drop-out event is not predicted by the values of the outcome variables^{44,67}. This can be stated mathematically as follows: at each point in time j , it is tested if for each $j \leq n-1$ the subset of workers with $n_i = j$ represent a random sample from the group of workers with $n_i \geq j$.

Thus, it is assessed if the value of the measurement performed immediately before a drop-out can be considered a predictor for that drop-out event.

In order to test if the drop-out occurs completely at random, a separate analysis is needed. Only a subset of workers who drop-out between the first and the ninth year are kept in this analysis. Thus, workers with gaps in the recorded follow-up years are not included in this analysis. For each point in time $t=j < n$, a binary outcome variable is generated. The outcome variable is coded 1 if the worker drops out of the study at time $t=j+1$. The outcome variable is coded 0 if the workers drops out at time $t > j+1$ or never drops out. Using these outcome variables, logistic regression models are run at each point in time.

For example, during the first year of follow-up, workers who will drop out of the analysis during the second year of follow-up have the outcome variable coded as 1. Workers who will drop out later on or who will never drop out have the outcome variable coded as 0. Then, the

logistic regression model is run in workers included in the first year of follow-up having the outcome variable set-up as described and the covariate the $\ln(\text{yearly median lymphocyte counts})$ recorded in the first year of follow-up.

This analysis is run each year, until the 9th year of follow-up and a total of nine p-values are calculated. For the 10th year of follow-up, no p-value can be calculated since there is no “next year” anymore.

Thus, the goal of the drop-out analysis is to check whether the subjects who are about to drop out are a random sample from the group who will not drop out at the next point in time. Each run of the logistic regression models provides a p-value for the coefficient of the predictor.

The drop-outs can be considered to be completely at random if the distribution of the p-values are uniform $(0,1)$ ⁴⁴. Therefore, the distribution of logistic regression p-values is compared with a simulated uniform $(0,1)$ distributed variable. The comparison is performed using the Kolmogorov-Smirnov statistic implemented via Monte Carlo tests with 900 replications. The Kolmogorov-Smirnov statistic p-value=0.2544, which does not give reason to reject the null hypothesis. Therefore one can conclude that the p-value distribution is uniform $(0,1)$ and the drop-outs can be considered completely at random^{44,66}. The STATA program and the output illustrating this procedure are presented in Appendix C.

The same analysis is performed separately in males and females in order to check if after stratification by sex the drop-outs can still be considered completely at random.

Drop-out analysis in males

In males the Kolmogorov-Smirnov statistic implemented via Monte Carlo tests with 900 replications leads to a p-value=0.2589. Therefore in males, the drop-outs can be considered completely at random.

Drop-out analysis in females

In females the Kolmogorov-Smirnov statistic implemented via Monte Carlo tests with 900 replications leads to a p-value=0.2733. Therefore in females, the drop-outs can also be considered completely at random.

It can be concluded that in the study cohort, overall and after stratification by sex, the drop-out events can be considered completely at random and therefore regarding the drop-outs, the assumptions required by the statistical techniques described in the methodology section are met.

3.2.3.3 Mixed patterns missing data description

Mixed missing data patterns are defined as gaps in the workers' records for one or more follow-up years. The gaps may be associated with drop-outs. In order to describe and assess the mixed patterns missing data, a separate analysis is necessary⁶⁸. Workers having data in all follow-up years and drop-outs only are not included in this analysis. Mixed patterns missing data analysis is performed for the overall study cohort and after stratification by sex.

In the analysis of the mixed missing data patterns the focus is on the two main variables of interest: the outcome, which is the log-transformed yearly median lymphocyte count and the explanatory covariate which is the log-transformed yearly cumulative external gamma exposure. Both of these variables are analyzed as continuous time dependent variables with repeated measurements.

Four categories for the existing mixed missing data patterns are defined as follows:

-first pattern when both variables, the log-transformed yearly median lymphocyte count and the log-transformed yearly cumulative external gamma exposure, are present;

-second pattern when the log-transformed yearly median lymphocyte count is present and the log-transformed yearly cumulative external gamma exposure is missing;

-third pattern when the log-transformed yearly median lymphocyte count is missing and the log-transformed yearly cumulative external gamma exposure is present;

-fourth pattern when the log-transformed yearly median lymphocyte count is missing and the log-transformed yearly cumulative external gamma exposure is missing as well;

As previously shown, about 53 workers (15%) have mixed missing data patterns. After stratification by sex, 34 males and 19 females have mixed missing data patterns.

The frequency distribution of observations according to their mixed missing data pattern is described in Table 52. It is worthwhile to note that Table 52 refers to the number of observations and therefore some of the measurements may belong to the same workers.

Table 52: Mixed Missing Data Patterns: Absolute and Relative Frequency Distribution of Observations by Missing Data Patterns and Sex

Pattern	Mixed missing data patterns	Total* n (%)	Males n (%)	Females n (%)
1	ln(yearly median lymphocyte counts) present ln(yearly cumulative external gamma dose) present	299 (100.0)	200 (66.9)	99 (33.1)
2	ln(yearly median lymphocyte counts) present ln(yearly cumulative external gamma dose) missing	101 (100.0)	53 (52.5)	48 (47.5)
3	ln(yearly median lymphocyte counts) missing ln(yearly cumulative external gamma dose) present	2 (100.0)	0 (0.0)	2 (100.0)
4	ln(yearly median lymphocyte counts) missing ln(yearly cumulative external gamma dose) missing	1 (100.0)	1 (100.0)	0 (0.0)
Total		403 (100.0)	254 (63.0)	149 (37.0)

*These observations are based on 53 workers (34 males and 19 females) with mixed missing data patterns; Workers who have data in all years and workers who have drop-outs only are excluded

The descriptive analysis of the observations on workers with mixed patterns missing data shows that 299 out of 403 observations have information for both variables of interest (log-transformed lymphocyte counts and dose). The most often observed missing data pattern is pattern 2 (101 out of 402 observations) in which ln(yearly median lymphocyte counts) is present and ln(yearly cumulative external gamma dose) is missing.

Since the frequency distribution of the mixed missing data patterns looks skewed and there are few observations in pattern 3 and pattern 4, there is no examination of differences by sex.

3.2.3.4 Mixed patterns missing data assessment

The workers with mixed patterns missing data are analyzed in order to assess whether the data are missing completely at random. Only the subset of workers with mixed missing data patterns are analyzed using the statistical test proposed by Little⁶⁸. The statistical test introduced by Little consists of the analysis of the ln(yearly median lymphocyte counts) and ln(yearly cumulative external gamma exposure), as independent observations without taking into account the correlations between measurements.

In the first part of the analysis, the group of observations with valid data on ln(yearly cumulative external gamma exposure) is compared with the group of observations with missing ln(yearly cumulative external gamma exposure) in terms of the ln(yearly median lymphocyte counts). The comparison is made using an ANOVA test which assesses whether ln(yearly median lymphocyte counts) depends of the ln(yearly cumulative external gamma exposure)'s missingness. Based on ANOVA test results the test statistic d^2 proposed by Little is calculated as follows:

$$d^2=(n-1) *F/(n-2+F)=(400-1)*0.24/(400-2+0.24)=0.24$$

Notation explanation: n=number of observations, F=the value of the statistical test

Under the null hypothesis d^2 is distributed as $\chi^2_{(\sum p_j - p)}$ where p_j is the number of variables observed within each analyzed pattern and p is the number of continuous variables of interest. In this case, $\sum p_j$ is 2+1=3 since there are two patterns included in the analysis: one pattern with both variables present ($p_1=2$) and one pattern with one variable present and one variable missing

($p_2=1$). The continuous variables of interest are the $\ln(\text{yearly median lymphocyte counts})$ and the $\ln(\text{yearly cumulative external gamma exposure})$ and therefore $P=2$. Thus $(\sum p_j - p) = 3 - 2 = 1$. $\chi^2_{(1)} = 3.84$ (from the χ^2 table) $0.24 < 3.84$ and therefore we can not reject the null hypothesis that states that the data are missing completely at random.

In the second part of the analysis, the group of observations with valid data on $\ln(\text{yearly median lymphocyte counts})$ is compared with the group of observations with missing $\ln(\text{yearly median lymphocyte counts})$ in terms of the $\ln(\text{yearly cumulative external gamma exposure})$. The comparison is made using an ANOVA test which assesses whether $\ln(\text{yearly cumulative external gamma exposure})$ depends on the $\ln(\text{yearly median lymphocyte counts})$'s missingness. Based on ANOVA test results the test statistic d^2 proposed by Little is calculated as follows:

$$d^2 = (n-1) * F / (n-2+F) = (301-1) * 0.42 / (301-2+0.42) = 0.42$$

Notation explanation: n=number of observations, F=the value of statistical test

Under the null hypothesis d^2 is distributed as $\chi^2_{(\sum p_j - p)}$ where p_j is the number of variables observed within each analyzed pattern and p is the number of continuous variables of interest. In this case, $\sum p_j$ is $2+1=3$ since there are two patterns included in the analysis: one pattern with both variables present ($p_1=2$) and one pattern with one variable present and one variable missing ($p_2=1$). The continuous variables of interest are the $\ln(\text{yearly median lymphocyte counts})$ and the $\ln(\text{yearly cumulative external gamma exposure})$ and therefore $P=2$. Thus, $(\sum p_j - p) = 3 - 2 = 1$. $\chi^2_{(1)} = 3.84$ (from the χ^2 table) $0.42 < 3.84$ and therefore we can not reject the null hypothesis that states that the data are missing completely at random.

In order to assess the mixed patterns missing data mechanisms two statistical tests have been applied. This approach introduces a multiple comparison issue. However, the multiple

comparison issue can be ignored because the results are not significant and there are only two comparisons.

Mixed patterns missing data assessment in males

In the first part of the analysis, the group of observations with valid data on ln(yearly cumulative external gamma exposure) is compared with the group of observations with missing ln(yearly cumulative external gamma exposure) in terms of the ln(yearly median lymphocyte counts). The comparison is made using an ANOVA test which assesses whether ln(yearly median lymphocyte counts) depends on the ln(yearly cumulative external gamma exposure)'s missingness. Based on ANOVA test results the test statistic d^2 proposed by Little is calculated as follows:

$$d^2 = (n-1) * F / (n-2+F) = (253-1) * 0.12 / (253-2+0.12) = 0.12$$

Notation explanation: n=number of observations, F=the value of statistical test

Under the null hypothesis d^2 is distributed as $\chi^2_{(\sum p_j - p)}$ where p_j is the number of variables observed within each analyzed pattern and p is the number of continuous variables of interest. In this case, $\sum p_j$ is 2+1=3 since there are two patterns included in the analysis: one pattern with both variables present ($p_1=2$) and one pattern with one variable present and one variable missing ($p_2=1$). The continuous variables of interest are the ln(yearly median lymphocyte counts) and the ln(yearly cumulative external gamma exposure) and therefore P=2. Thus $(\sum p_j - p) = 3 - 2 = 1$. $\chi^2_{(1)} = 3.84$ (from the χ^2 table) $0.12 < 3.84$ and therefore we can not reject the null hypothesis that states that the data are missing completely at random.

The second part of the analysis is no longer possible since in males, there are not data with the ln(yearly median lymphocyte counts) missing and ln(yearly cumulative external gamma exposure) present.

Mixed patterns missing data assessment in females

In the first part of the analysis, the group of observations with valid data on ln(yearly cumulative external gamma exposure) is compared with the group of observations with missing ln(yearly cumulative external gamma exposure) in terms of the ln(yearly median lymphocyte counts). The comparison is made using an ANOVA test which assesses whether ln(yearly median lymphocyte counts) depends on the ln(yearly cumulative external gamma exposure)'s missingness. Based on ANOVA tests results the test statistic d^2 proposed by Little is calculated as follows:

$$d^2 = (n-1) * F / (n-2+F) = (147-1) * 0.65 / (147-2+0.65) = 0.65$$

Notation explanation: n=number of observations, F=the value of statistical test

Under the null hypothesis d^2 is distributed as $\chi^2_{(\sum p_j - p)}$ where p_j is the number of variables observed within each analyzed pattern and p is the number of continuous variables of interest. In this case, $\sum p_j$ is 2+1=3 since there are two patterns included in the analysis: one pattern with both variables present ($p_1=2$) and one pattern with one variable present and one variable missing ($p_2=1$). The continuous variables of interest are the ln(yearly median lymphocyte counts) and the ln(yearly cumulative external gamma exposure) and therefore P=2. Thus $(\sum p_j - p) = 3 - 2 = 1$. $\chi^2_{(1)} = 3.84$ (from the χ^2 table) $0.65 < 3.84$ and therefore we can not reject the null hypothesis that states that the data are missing completely at random.

In the second part of the analysis, the group of observations with valid data on ln(yearly median lymphocyte counts) is compared with the group of observations with missing ln(yearly median lymphocyte counts) in terms of the ln(yearly cumulative external gamma exposure). The comparison is made using an ANOVA test which assesses whether ln(yearly cumulative external gamma exposure) depends on the ln(yearly median lymphocyte counts)'s missingness. Based on ANOVA tests results the test statistic d^2 proposed by Little is calculated as follows:

$$d^2 = (n-1) * F / (n-2+F) = (101-1) * 0.21 / (101-2+0.21) = 0.21$$

Notation explanation: n=number of observations, F=the value of the statistical test

Under the null hypothesis d^2 is distributed as $\chi^2_{(\sum p_j - p)}$ where p_j is the number of variables observed within each analyzed pattern and p is the number of continuous variables of interest. In this case, $\sum p_j$ is $2+1=3$ since there are two patterns included in the analysis: one pattern with both variables present ($p_1=2$) and one pattern with one variable present and one variable missing ($p_2=1$). The continuous variables of interest are the ln(yearly median lymphocyte counts) and the ln(yearly cumulative external gamma exposure) and therefore $P=2$. Thus, $(\sum p_j - p) = 3 - 2 = 1$. $\chi^2_{(1)} = 3.84$ (from the χ^2 table) $0.21 < 3.84$ and therefore we can not reject the null hypothesis that states that the data are missing completely at random.

In order to assess the mixed patterns missing data mechanisms two statistical tests have been applied. This approach introduces a multiple comparison issue. However, the multiple comparison issue can be ignored because the results are not significant and there are only two comparisons.

3.2.3.5 Missing data analysis conclusion

It can be concluded that all tests performed for log-transformed yearly median lymphocyte counts missing data analysis have shown the following: the occurrence of drop-outs and mixed

missing data patterns can be considered completely at random in the overall study cohort and after stratification by sex.

This important result allows the use of the GEE (Generalized Estimating Equation) statistical method without anticipating biases for the parameter estimates.

3.3 STATISTICAL MODELING RESULTS

The statistical models formulated in this dissertation address the following questions presented in the methodology section:

- 1) Are lymphocyte counts affected by long term occupational exposure to gamma radiation in Mayak PA workers?
- 2) Are there differences between males and females in terms of inhibition of lymphocyte counts by occupational exposure to gamma radiation in Mayak PA workers?
- 3) Is there a differential effect on lymphocyte counts at lower gamma doses (below 5 rads) compared to higher gamma doses? If yes, is the effect size similar in males and females?
- 4) Does the effect on the lymphocyte counts vary over time in workers exposed to occupational external gamma radiation? If yes is the time effect similar in males and females?

As stated in the methodology section, the main objective of this study is to assess the effects of external gamma exposure on lymphocyte counts, adjusting for sex, baseline lymphocyte counts, work location related to Plutonium exposure, and lifestyle variables.

In order to address the main objective and to answer the research questions, six statistical models have been constructed and compared. The model that fits the data best, and satisfies the assumptions required by the GEE method will be chosen as the optimal model.

The variables included in the six models have been selected based on theoretical and practical considerations. The theoretical considerations are derived from information already existing in the literature and from current regulations on occupational exposure to radiation. The practical considerations have been established based on the available data on the study cohort.

The six statistical models that will be evaluated in this dissertation have the following characteristics:

- The outcome variable is the log-transformed yearly median lymphocyte count $\ln(Y_{it})$. The log-transformation satisfies the normality assumption of the GEE methods.
- The predicted value is defined as $E[\ln(Y_{it})]$. The link function used for these models is the identity function.
- The explanatory covariate is the log-transformed yearly cumulative external gamma dose. The log-transformation of the yearly cumulative external gamma dose satisfies the linearity assumption of the GEE method. Linear splines are implemented for the log-transformed yearly cumulative external gamma dose in some of the models. In these models the splines become the main covariates.
- Sex is always included in the models and stratification by sex is always performed since the descriptive analysis suggests differential effects of external gamma radiation exposure in males and females.

- Work location related to Plutonium exposure is always included in the model as a dummy variable. This variable allows adjustment for Plutonium exposure.
- Lifestyle variables smoking history and alcohol consumption at start of employment are also included in the model as adjustment variables.

Each of the six models is fit separately for males and females, omitting the covariate for sex in the models.

The models analyzed in this dissertation can be written mathematically using the following notation introduced in the methodology section.

Model#1:

$$E[\ln(Y_{it})] = \beta_0 + \beta_1 \ln(X_{1_{it}}) + \beta_2 \ln(X_{2_i}) + \beta_3 X_{3_i} + \beta_4 X_{4_i} + \beta_5 X_{5_{it}} + \beta_6 X_{6_{it}} + \beta_7 X_{7_i}$$

Model #1 includes as the explanatory covariate the log-transformed yearly cumulative external gamma dose and as adjustment variables sex, baseline lymphocyte counts, work location related to Plutonium exposure, smoking history, and alcohol consumption at start of employment.

Model#2:

$$E[\ln(Y_{it})] = \beta_0 + \beta_1 \ln(X_{1_{it}}) + \beta_2 \ln(X_{2_i}) + \beta_3 X_{3_i} + \beta_4 X_{4_i} + \beta_5 X_{5_{it}} + \beta_6 X_{6_{it}} + \beta_7 X_{7_i} + \beta_9 X_{9_{it}}$$

Model#2 includes, in addition to the variables described for Model #1, an interaction term between the log-transformed yearly cumulative external gamma dose and sex.

Model#3:

$$E[\ln(Y_{it})] = \beta_0 + \beta_1 \ln(X_{1_{it}}) + \beta_2 \ln(X_{2_i}) + \beta_3 X_{3_i} + \beta_4 X_{4_i} + \beta_5 X_{5_{it}} + \beta_6 X_{6_{it}} + \beta_7 X_{7_i} + \beta_8 X_{8_i}$$

Model#3 includes, in addition to the variables described for Model #1, an interaction term between the log-transformed yearly cumulative external gamma dose and work location related to plutonium exposure.

Model#4:

$$E[\ln(Q_{it})] = \beta_0 + \beta_1 X_{1_{it}} + \beta_2 X_{2_{it}} + \beta_3 X_{3_{it}} + \beta_2 \ln(X_{2_i}) + \beta_3 X_{3_i} + \beta_4 X_{4_i} + \beta_5 X_{5_{it}} + \beta_6 X_{6_{it}} + \beta_7 X_{7_i}$$

Model#4 is characterized by the implementation of an optimization known as linear splines. Linear splines are applied to the log-transformed yearly cumulative external gamma dose in order to address the non-linear relationship between the yearly median lymphocyte counts and yearly cumulative external gamma dose. The non-linear relationship between the yearly median lymphocyte counts and yearly cumulative external gamma dose was presented in the descriptive part of the results. An important aspect of linear splines implementation is the knots locations. The knots of the linear splines are set in Model#4 according to the distribution of the residuals, and according to a specific gamma radiation dose value (5 rads), considered as a cut-off in the occupational regulations for radiation workers.

The value of the ln(yearly cumulative external gamma dose) at which the residuals' distribution changes or begins to show a pattern is set as the first knot. The location of the first knot is established after the examination of the residuals plots. Thus, the first knot is located where ln(yearly cumulative external gamma dose)=0 which corresponds to a yearly cumulative external gamma dose of 1 rad.

The value of the ln(yearly cumulative external gamma dose) which corresponds to 5 rads gamma dose is set as the second knot since 5 rads represents the cumulative yearly occupational

radiation exposure highest limit accepted for radiation workers in the US. Thus, the second knot is located at $\ln(5)=1.609437$.

Therefore, the first knot is data derived from the residuals' distribution, and the second knot corresponds to a meaningful value of the covariate based on the current exposure standard.

Model#5:

$$E[\ln(Y_{it})] = \beta_0 + \beta_1 X_{1_{it}} + \beta_2 X_{2_{it}} + \beta_3 \ln(X_{2_i}) + \beta_4 X_{3_i} + \beta_5 X_{4_i} + \beta_6 X_{5_{it}} + \beta_7 X_{6_{it}} + \beta_8 X_{7_i}$$

Model #5 also includes the implementation of linear splines. In Model#4, there is only one knot located where $\ln(\text{yearly cumulative external gamma exposure})=\ln(5)=1.609437$.

This knot location corresponds to a 5 rads external gamma dose, which represents the cumulative yearly occupational radiation exposure cut-off accepted for radiation workers in the United States. This unique knot location leads to the implementation of two linear splines. The linear splines are used to evaluate whether there is any differential in the effect of gamma radiation at doses below 5 rads versus doses higher than 5 rads.

Model#6:

$$E[\ln(Y_{it})] = \beta_0 + \beta_1 \ln(X_{1_{it}}) + \beta_2 \ln(X_{2_i}) + \beta_3 X_{3_i} + \beta_4 X_{4_i} + \beta_5 X_{5_{it}} + \beta_6 X_{6_{it}} + \beta_7 X_{7_i} + \beta_8 X_{10_i}$$

Model #6 includes, in addition to the variables already described in Model #1, the number of years from first external gamma exposure. Model #5 explores the time effect on the yearly median lymphocyte counts.

The number of years since the first external gamma exposure is the temporal variable used for all the GEE models analyzed in this dissertation.

3.3.1 Assessment of the coefficients in the models

As described in the methodology section, the estimated β coefficients and their standard errors are computed using a quasi-likelihood based method known as GEE. The p-values are calculated by testing the null hypothesis that the coefficients equal zero ($H_0: \beta=0$) against the two-sided alternative hypothesis that the coefficients are different from zero ($H_A: \beta \neq 0$). The work location related to Plutonium exposure is fitted as dummy variable, thereby calculating two β coefficients, β_5 and β_6 with two corresponding p-values.

Since the goal of the project is to analyze the global effect of work location related to Plutonium exposure, an additional analysis is necessary. The additional analysis consists of simultaneous testing of β_5 and β_6 , using the χ^2 test. ($H_0: \beta_5=0$ and $\beta_6=0$, $H_A: \beta_5 \neq 0$ and $\beta_6 \neq 0$)

The β coefficients used in this analysis are defined as follows:

β_1 =the coefficient of the log-transformed yearly cumulative external gamma dose

$\beta_{1_1}, \beta_{1_2}, \beta_{1_3}$ =the coefficients of the linear splines applied to the log-transformed yearly cumulative external gamma dose

β_2 =the coefficient of log-transformed baseline lymphocyte counts

β_3 =the coefficient of smoking history variable

β_4 =the coefficient of alcohol consumption at start of employment variable

β_5, β_6 =the coefficients of work location related to Plutonium exposure variable fitted as dummy variable

β_7 =the coefficient of the sex variable

β_8 =the coefficient of the interaction term between ln(yearly cumulative external gamma dose) and work location related to Plutonium exposure

β_9 =the coefficient of the interaction term between ln(yearly cumulative external gamma dose) and sex

β_{10} =the coefficient for the number of years from the first external gamma exposure

The coefficients, standard errors and p-values for the overall study cohort and after stratification by sex are presented in Tables 53-55.

Table 53: Total; β -Coefficients, Standard Errors, and p-values calculated by using GEE Method for the Six Models

Model	β Coefficients* (Semi-robust standard error) (p-value)												
	β_1	β_{1_1}	β_{1_2}	β_{1_3}	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
1	-0.03 (0.00) (<i><0.0005</i>)	NA	NA	NA	0.29 (0.03) (<i><0.0005</i>)	0.01 (0.03) (0.69)	0.01 (0.02) (0.60)	0.00 (0.02) (0.15)**	-0.06 (0.03) (0.15)**	0.01 (0.03) (0.69)	NA	NA	NA
2	-0.04 (0.01) (<i><0.0005</i>)	NA	NA	NA	0.29 (0.03) (<i><0.0005</i>)	0.01 (0.03) (0.68)	0.01 (0.02) (0.60)	0.002 (0.02) (0.17)**	-0.06 (0.03) (0.17)**	-0.02 (0.04) (0.65)	NA	0.01 (0.01) (0.08)	NA
3	-0.02 (0.01) (<i><0.0005</i>)	NA	NA	NA	0.29 (0.03) (<i><0.0005</i>)	0.01 (0.03) (0.71)	0.01 (0.02) (0.60)	0.03 (0.02) (0.16)**	-0.02 (0.04) (0.16)**	0.01 (0.03) (0.66)	-0.01 (0.01) (0.08)	NA	NA
4	NA	0.02 (0.02) (0.30)	-0.004 (0.01) (0.79)	-0.05 (0.01) (<i><0.0005</i>)	0.29 (0.03) (<i><0.0005</i>)	0.01 (0.02) (0.61)	0.01 (0.02) (0.64)	0.01 (0.02) (0.11)**	-0.06 (0.03) (0.11)**	0.02 (0.03) (0.62)	NA	NA	NA
5	NA	0.01 (0.01) (0.54)	NA	-0.05 (0.01) (<i><0.0005</i>)	0.29 (0.03) (<i><0.0005</i>)	0.01 (0.02) (0.61)	0.01 (0.02) (0.64)	0.01 (0.02) (0.11)**	-0.06 (0.03) (0.11)**	0.02 (0.03) (0.62)	NA	NA	NA
6	-0.03 (0.00) (<i><0.0005</i>)	NA	NA	NA	0.29 (0.03) (<i><0.0005</i>)	0.01 (0.03) (0.71)	0.01 (0.02) (0.60)	0.002 (0.02) (0.23)**	-0.05 (0.03) (0.23)**	0.01 (0.03) (0.65)	NA	NA	0.003 (0.00) (0.10)

NA=non-applicable since the term is not in the model

* β coefficients as defined in text

**p-value is calculated by applying a global χ^2 test to the coefficients of the work location dummy variables

As shown in Table 53, in the overall study cohort the main explanatory covariate which is the log-transformed yearly cumulative external gamma dose is a highly statistically significant negative predictor for the ln(lymphocyte count) in all models in which it is included (Models 1, 2, 5 and 6). When splines are implemented, the splines which correspond to doses higher than 5 rads are highly statistically significant negative predictors for ln(lymphocyte count) (Models 3 and 4). Moreover, Model 4 shows that gamma doses below 5 rads have a statistically non-significant increase in the log-transformed lymphocyte counts ($p=0.54$). Additionally, gamma doses above 5 rads predict a statistically significant inhibition of the lymphocyte counts.

There is no statistically significant interaction effect between the log-transformed dose and work location related to Plutonium exposure (Model#3).

There is also no statistically significant time effect on the log-transformed lymphocyte counts where time is defined as the number of years from the first external gamma exposure.

In regards to the adjustment variables, the baseline lymphocyte count is the only statistically significant term. Work location related to Plutonium exposure, smoking history and alcohol consumption at start of employment are non-significant terms.

Table 54: Males: β -Coefficients, Semi-Robust Standard Errors, and p-values Calculated for the Six Models

Model	Coefficients (Semi-robust standard error) (p-value)											
	β_1	β_{1_1}	β_{1_2}	β_{1_3}	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
1	-0.04 (0.01) (<i><0.0005</i>)	NA	NA	NA	0.32 (0.04) (<i><0.0005</i>)	0.01 (0.03) (0.70)	0.04 (0.03) (0.12)	0.01 (0.02) (0.39)**	-0.04 (0.03) (0.39)**	NA	NA	NA
2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	-0.03 (0.01) (<i><0.0005</i>)	NA	NA	NA	0.32 (0.04) (<i><0.0005</i>)	0.01 (0.03) (0.77)	0.04 (0.03) (0.15)	0.04 (0.03) (0.25)**	0.02 (0.05) (0.25)**	NA	-0.01 (0.01) (0.08)	NA
4	NA	0.04 (0.04) (0.28)	0.01 (0.02) (0.54)	-0.06 (0.01) (<i><0.0005</i>)	0.32 (0.04) (<i><0.0005</i>)	0.01 (0.03) (0.62)	0.03 (0.02) (0.19)	0.01 (0.02) (0.23)**	-0.05 (0.03) (0.23)**	NA	NA	NA
5	NA	0.02 (0.01) (0.09)	NA	-0.06 (0.01) (<i><0.0005</i>)	0.32 (0.04) (<i><0.0005</i>)	0.01 (0.03) (0.61)	0.03 (0.02) (0.19)	0.01 (0.02) (0.23)**	-0.05 (0.03) (0.23)**	NA	NA	NA
6	-0.03 (0.01) (<i><0.0005</i>)	NA	NA	NA	0.32 (0.04) (<i><0.0005</i>)	0.01 (0.03) (0.72)	0.04 (0.02) (0.16)	0.01 (0.02) (0.56)**	-0.03 (0.04) (0.56)**	NA	NA	0.004 (0.00) (0.11)

NA=non-applicable since the term is not in the model

* β coefficients as defined in text

**p-value is calculated by applying a global χ^2 test to the coefficients of the work location dummy variables

As shown in Table 54, in males the explanatory covariate, the log-transformed yearly cumulative external gamma dose is a highly statistically significant negative predictor for ln(lymphocyte counts) in all models in which it is included (Models 1, 3 and 6). When splines are implemented, the splines which correspond to doses higher than 5 rads are highly statistically significant negative predictors for the ln(lymphocyte counts) (Models 4 and 5). Moreover, Model 5 indicates that gamma doses below 5 rads lead to a borderline non-significant stimulation of the log-transformed lymphocyte counts ($p=0.09$). Additionally, gamma doses above 5 rads determine a statistically significant inhibition of the lymphocyte counts. There is no statistically significant interaction effect between the log-transformed dose and work location related to Plutonium exposure (Model 3). There is also no statistically significant time effect on the log-transformed lymphocyte counts, where time is defined as the number of years from the first external gamma exposure (Model 6).

In regards to the adjustment variables, the baseline lymphocyte count is the only statistically significant term. Work location related to Plutonium exposure, smoking history and alcohol consumption at start of employment are non-significant terms.

Table 55: Females: Coefficients, Semi-Robust Standard Errors, and p-values Calculated for the Six Models

<i>Model</i>	B Coefficients* (Semi-robust standard error) (p-value)											
	β_1	β_{1_1}	β_{1_2}	β_{1_3}	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
1	-0.02 (0.01) (<i><0.0005</i>)	NA	NA	NA	0.24 (0.05) (<i><0.0005</i>)	-0.07 (0.07) (0.35)	-0.15 (0.05) (0.01)	0.003 (0.04) (0.33)**	-0.08 (0.06) (0.33)**	NA	NA	NA
2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	-0.01 (0.01) (0.11)	NA	NA	NA	0.24 (0.05) (<i><0.0005</i>)	-0.06 (0.07) (0.41)	-0.15 (0.05) (0.01)	0.02 (0.03) (0.37)**	-0.05 (0.06) (0.37)**	NA	-0.01 (0.01) (0.33)	NA
4	NA	0.01 (0.02) (0.65)	-0.03 (0.02) (0.09)	-0.03 (0.01) (0.01)	0.23 (0.05) (<i><0.0005</i>)	-0.06 (0.07) (0.36)	-0.15 (0.05) (0.01)	0.01 (0.02) (0.34)**	-0.08 (0.06) (0.34)**	NA	NA	NA
5	NA	-0.01 (0.01) (0.16)	NA	-0.03 (0.01) (0.01)	0.23 (0.05) (<i><0.0005</i>)	-0.07 (0.07) (0.34)	-0.15 (0.05) (0.01)	0.01 (0.03) (0.32)**	-0.08 (0.06) (0.32)**	NA	NA	NA
6	-0.02 (0.01) (<i><0.0005</i>)	NA	NA	NA	0.23 (0.05) (<i><0.0005</i>)	-0.07 (0.07) (0.33)	-0.15 (0.05) (0.01)	0.003 (0.03) (0.37)**	-0.08 (0.06) (0.37)**	NA	NA	0.002 (0.00) (0.55)

NA=non-applicable since the term is not in the model

* β coefficients as defined in text.

**p-value is calculated by applying a global χ^2 test to the coefficients of the work location dummy variables

As shown in Table 56, in females the explanatory covariate, the log-transformed yearly cumulative external gamma dose, is a highly statistically significant negative predictor for the outcome in all models in which it is included. This result is illustrated in Models 1, 2 and 5. When splines are implemented, the splines which correspond to doses higher than 5 rads are highly significant negative predictors for the outcome (Models 4 and 5). In females, dissimilar from males, Model 5 shows that gamma doses below 5 rads leads to a statistically non-significant inhibition of the log-transformed lymphocyte counts ($p=0.16$). Additionally, gamma doses above 5 rads are associated with a statistically significant inhibition of the lymphocyte counts ($p=0.01$). There is no statistically significant interaction effect between the log-transformed dose and work location related to Plutonium exposure (Model 3). There is also no statistically significant time effect on the log-transformed lymphocyte counts where the time is defined as the number of years from the first external gamma exposure (Model 6).

In regards to the adjustment variables, the baseline lymphocyte count and alcohol consumption at start of employment are statistically significant covariates. Work location related to Plutonium exposure and smoking history are non-significant terms.

3.3.2 Assessment of goodness of fit of the models

This section presents the results of GEE model fitting. As discussed in the methodology section, the main goal of the model building is to find appropriate GEE models to analyze the effect of external gamma radiation (the explanatory covariate) on the lymphocyte counts (the outcome variable) while adjusting for sex, lifestyle variables and work location related to Plutonium exposure.

The model building requires the assessment of the goodness of fit using graphical as well as computational procedures.

The graphical assessment of the goodness of fit is performed for each model. It consists of three sets of graphs drawn for the overall study cohort and separately for males and females. The three sets of graphs are:

- 1) histograms of the residuals' distribution by years since the first external gamma exposure occurred
- 2) scatter plots of the residuals against the main covariate which consists of log-transformed yearly cumulative external gamma exposure; the scatter plots are drawn by years since the first external gamma exposure occurred
- 3) scatter plots of the residuals against the fitted values; the scatter plots are drawn by years since the first external gamma exposure occurred

As presented in the methodology section, residuals (res) are calculated as:

$$res_{it} = Y_{it} - \hat{Y}_{it},$$

where:

y_{it} = the observed variable measured for i-th individual at time t

\hat{y}_{it} = the fitted or predicted value for i-th individual at time t

The fitted or predicted value is a calculated value. It is derived for each fitted GEE model. Each model leads to different predicted values, thereby resulting in different residuals.

The goodness of fit procedures are applied to the previously described five GEE models. These five models are contrasted in terms of their goodness of fit.

The graphical goodness of fit comparison is performed through the three sets of plots described above as follows:

- 1) Histograms of the residuals' distribution by years since the first external gamma exposure occurred. The comparison is made for the overall study cohort and after stratification by sex.

1a. Overall study cohort

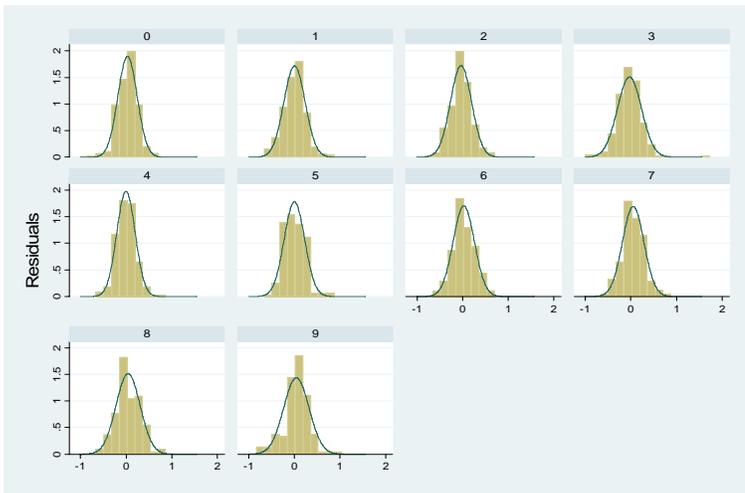


Figure 17: Total, Model#1; Histogram of Residuals' Distribution by Years since the First External Gamma Exposure

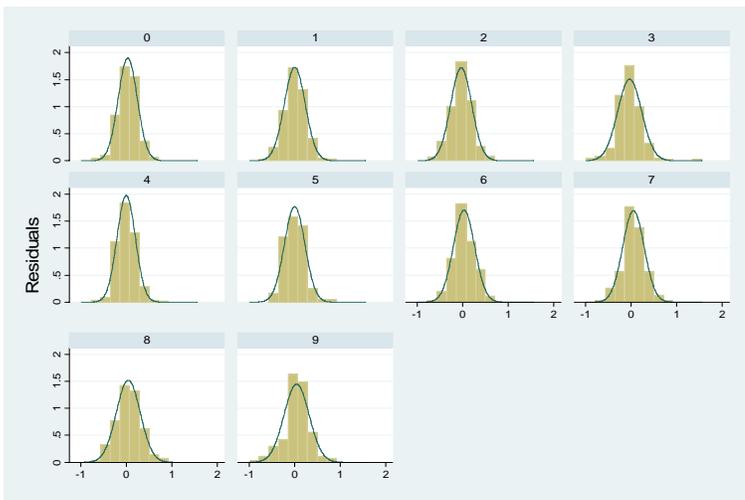


Figure 18: Total, Model#2 Histogram of Residuals' Distribution by Years since the First External Gamma Exposure

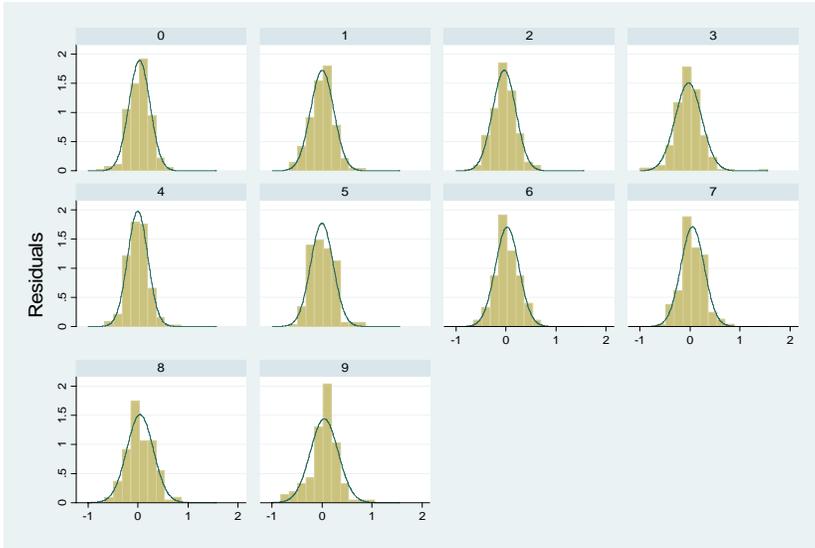


Figure 19: total, Model#3; Histogram of Residuals' Distribution by Years since the First External Gamma Exposure

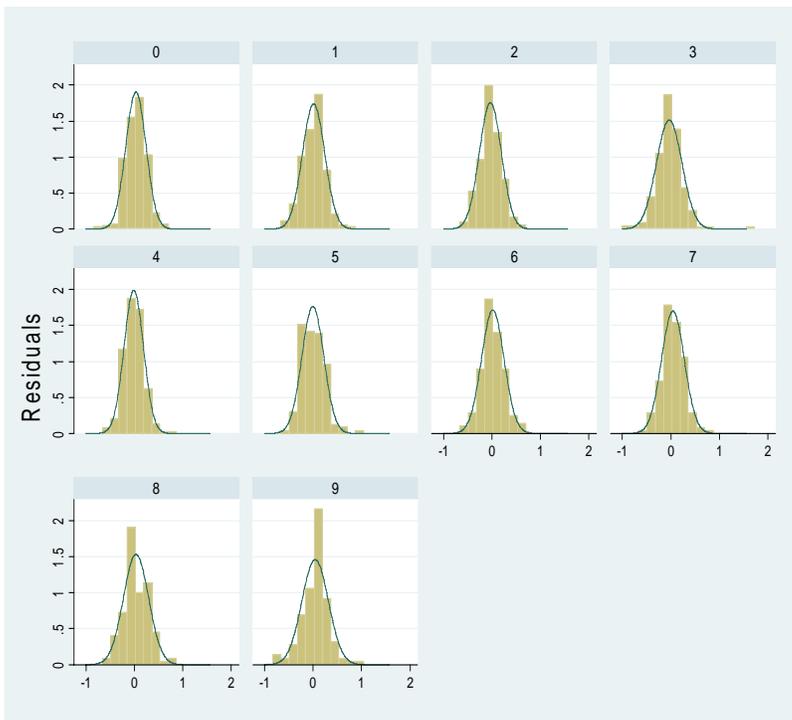


Figure 20: Total, Model #4; Histogram of Residuals' Distribution by Years since the First External Gamma Exposure

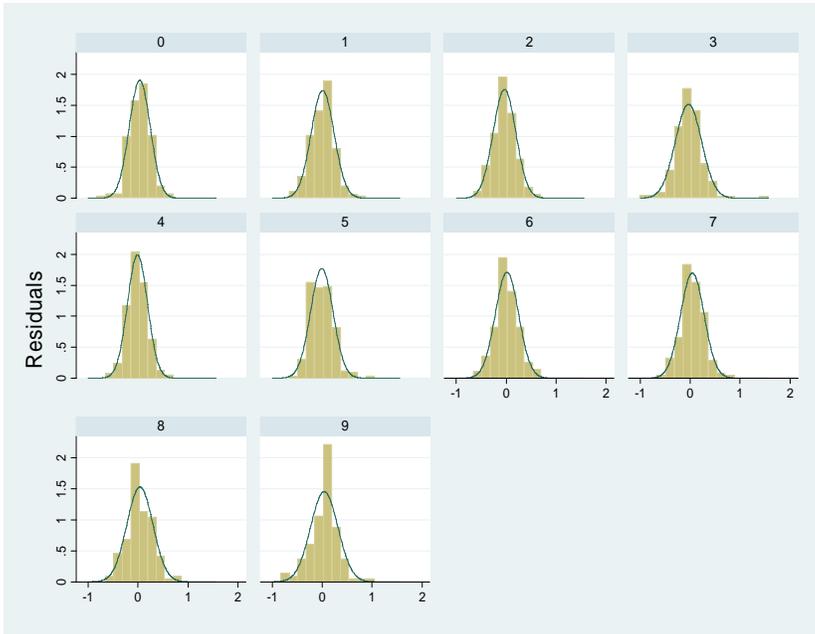


Figure 21: Total, Model#5; Histogram of Residuals' Distribution by Years since the First External Gamma Exposure

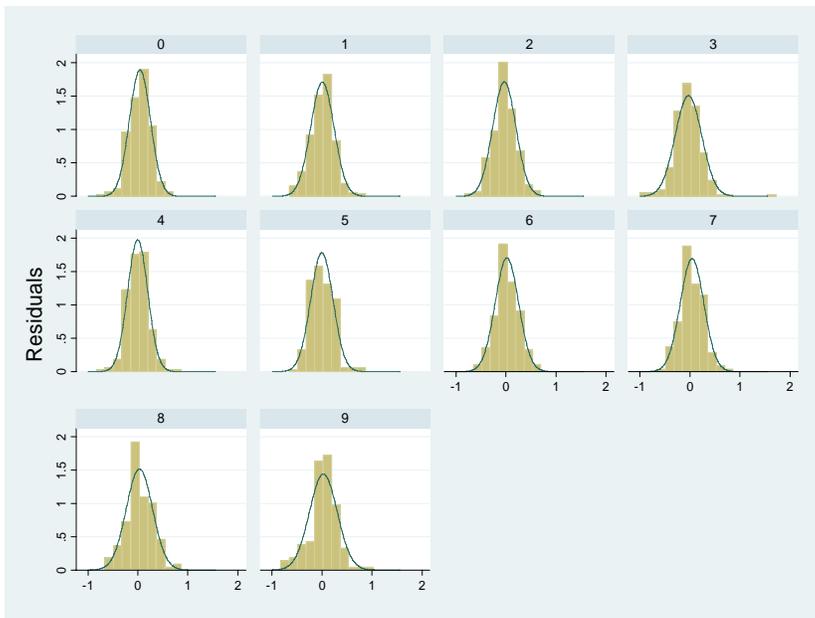


Figure 22: Total; Model#6; Histogram of Residuals' Distribution by Years since the First External Gamma Exposure

In the overall study cohort, the distribution of the residuals looks very similar in all six models. The residuals distribution can be considered normal or close to normal in all models and for most of the time points.

1b. Study cohort - males

The graphical distribution of the residuals is analyzed in males by years from the first external gamma exposure.

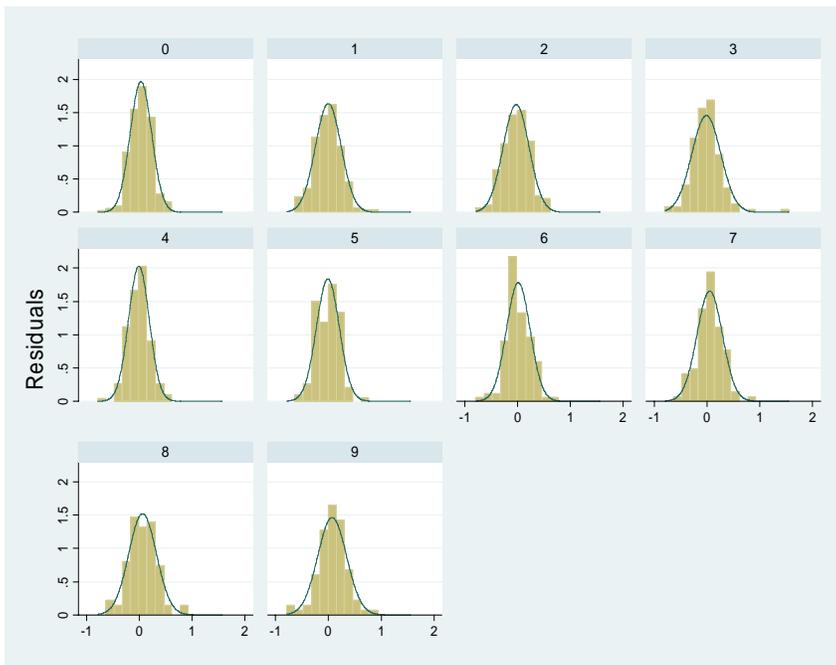


Figure 23: Males; Model#1; Histogram of Residuals' Distribution by Years since the First External Gamma Exposure

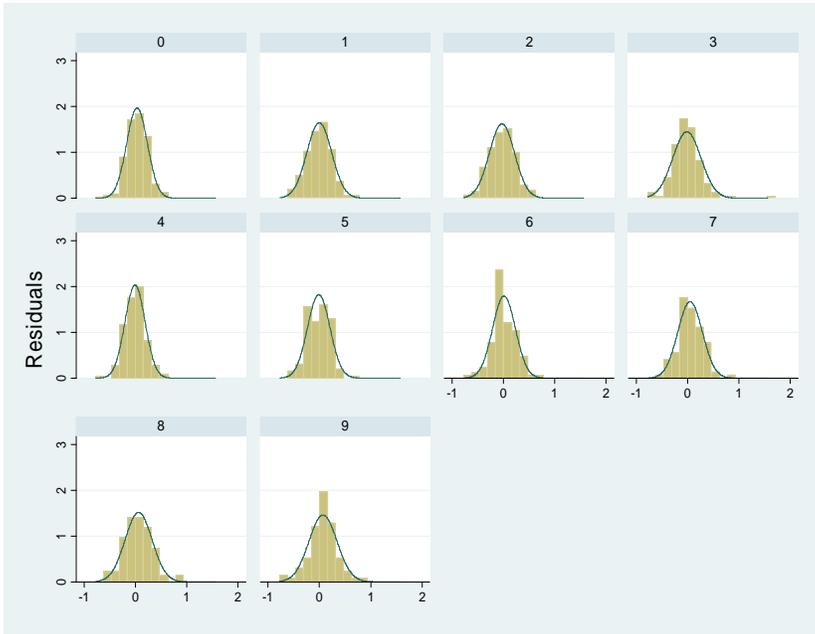


Figure 24: Males; Model#3; Histogram of Residuals' Distribution by Years since the First External

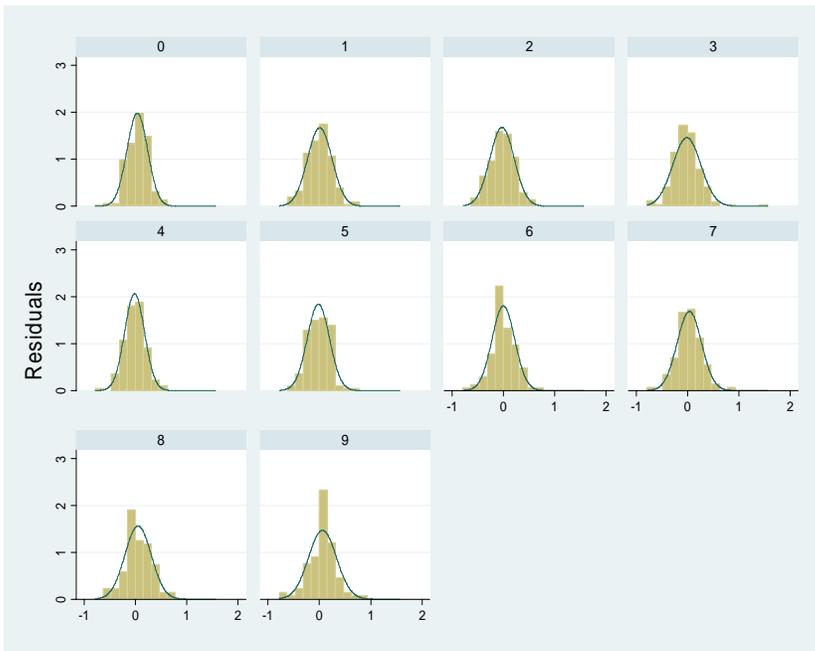


Figure 25: Males; Model#4; Histogram of Residuals' Distribution by Years since the First External Gamma Exposure

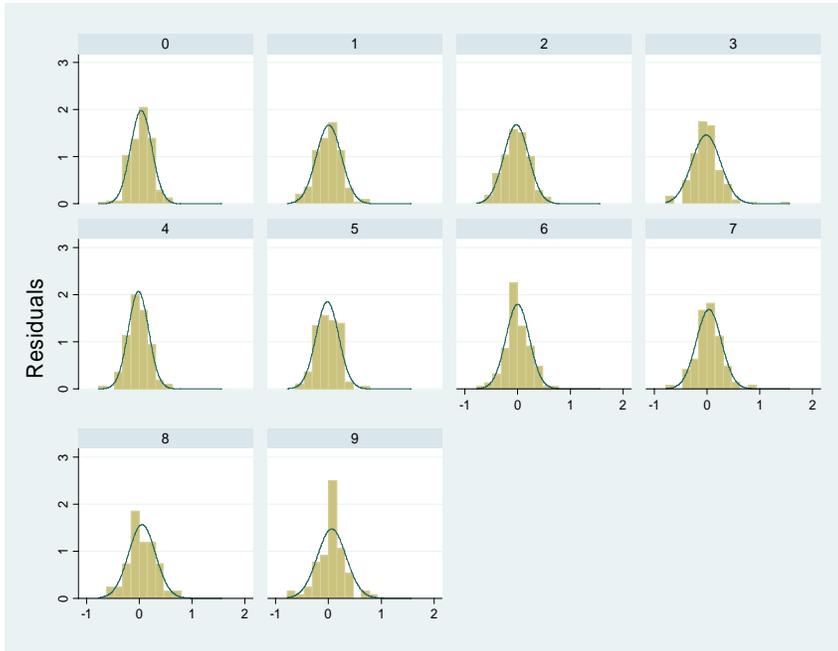


Figure 26: Males; Model#5; Histogram of Residuals' Distribution by Years since the First External Gamma Exposure

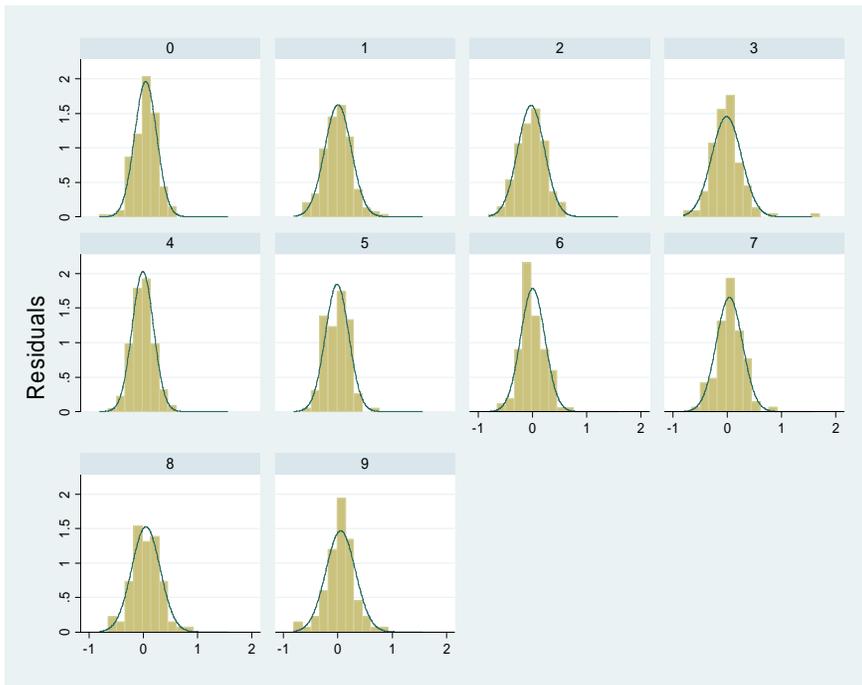


Figure 27: Males; Model #6; Histogram of Residuals' Distribution by Years since the First External Gamma Exposure

In males, the distribution of the residuals looks very similar in all five models. The residuals distribution can be considered normal or close to normal in all models and for most of the time points.

1c. Study cohort - females

The graphical distribution of the residuals is also analyzed in females, by years from the first external gamma exposure.

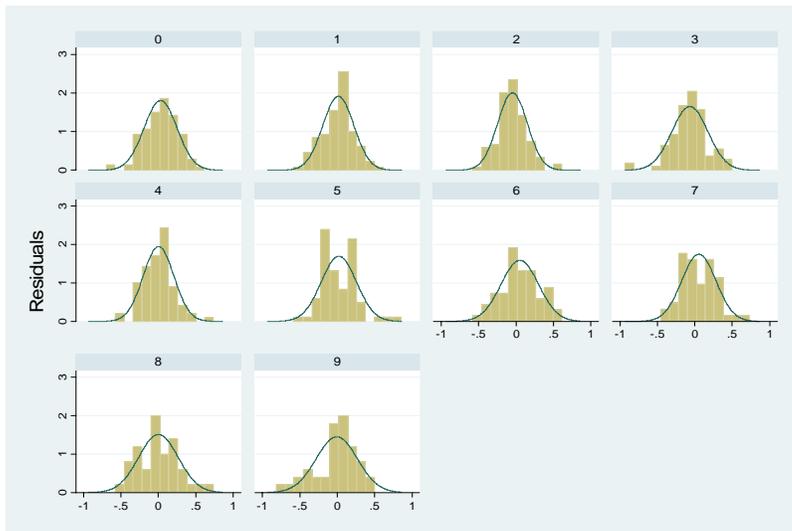


Figure 28: Males; Model #6; Histogram of Residuals' Distribution by Years since the First External Gamma Exposure

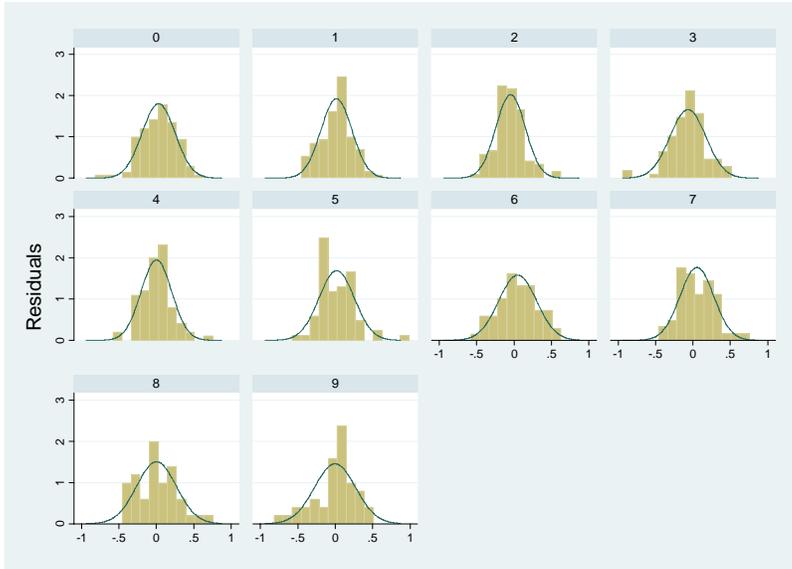


Figure 29 Females; Model#3; Histogram of Residuals' Distribution by Years since the First External Exposure

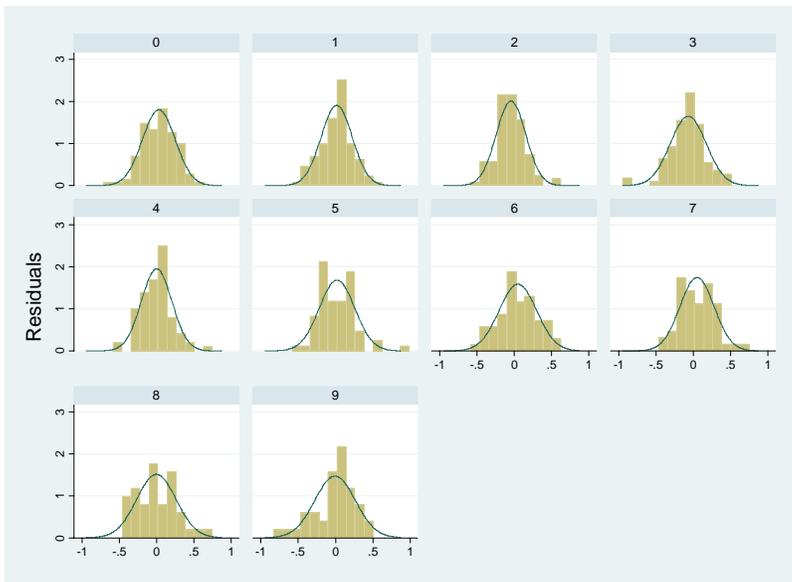


Figure 30: Females; Model#4; Histogram of Residuals' Distribution by Years since the First External Gamma Exposure

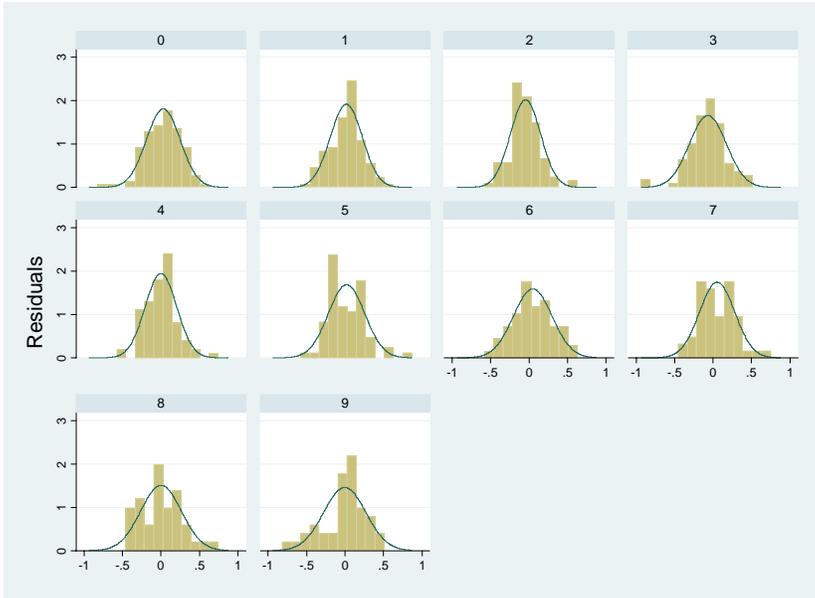


Figure 31: Females; Model#5; Histogram of Residuals' Distribution by Years since the First External Gamma Exposure

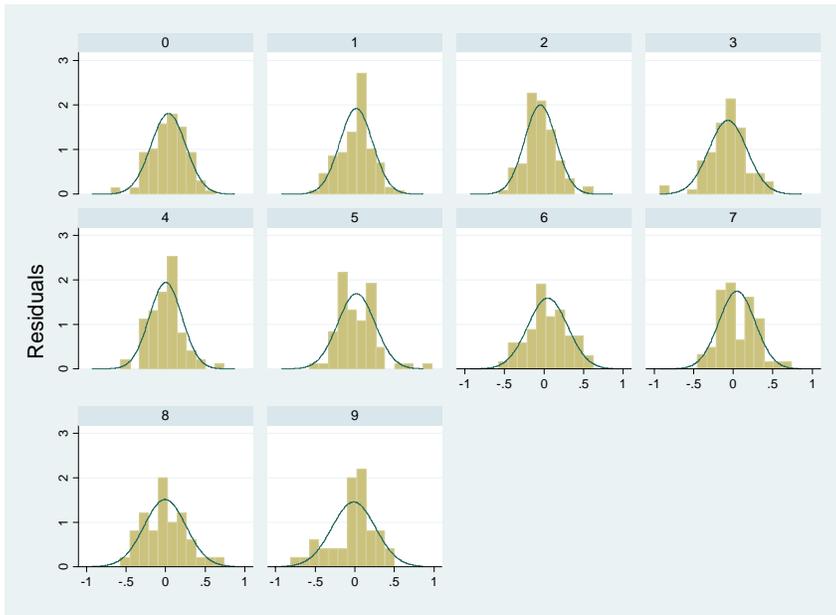


Figure 32: Females; Model#6; Histogram of Residuals' Distribution by Years since the First External Gamma Exposure

In females the residuals distribution looks similar in all five analyzed models. In terms of normal distribution, for most of the time points the residuals' distribution looks close to normal.

However, in females in all models the residuals seem to have a higher departure from normality than in males.

- 2) The goodness of fit of the six models is also compared using the scatter plot of the residuals against the explanatory covariate (log-transformed yearly cumulative external gamma exposure). Scatter plots are drawn by years since the first external gamma exposure occurred. If the models fit well, the scatter plots should show no pattern.

2a. Overall study cohort

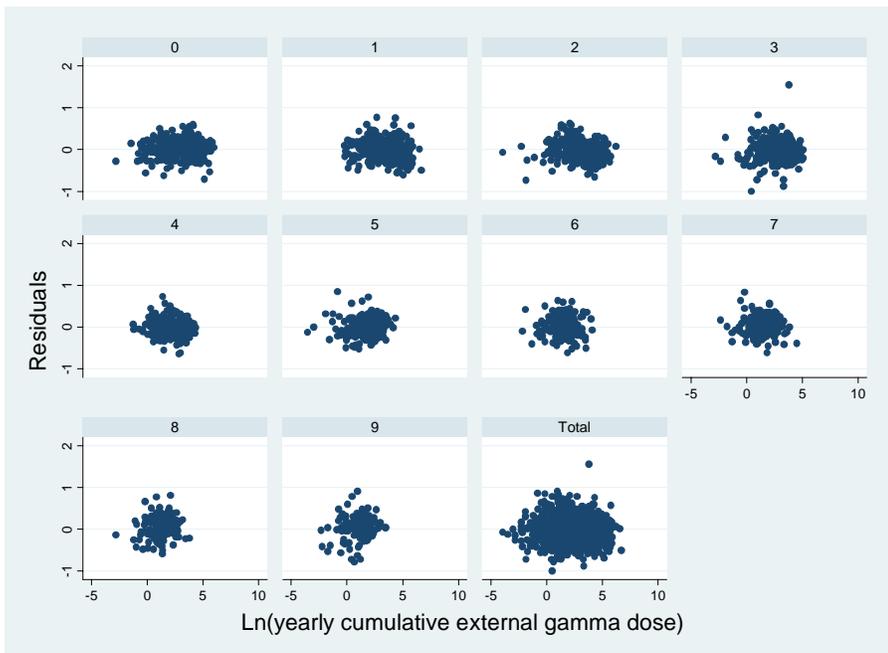


Figure 33: Total; Model#1; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years since the First External Gamma Exposure

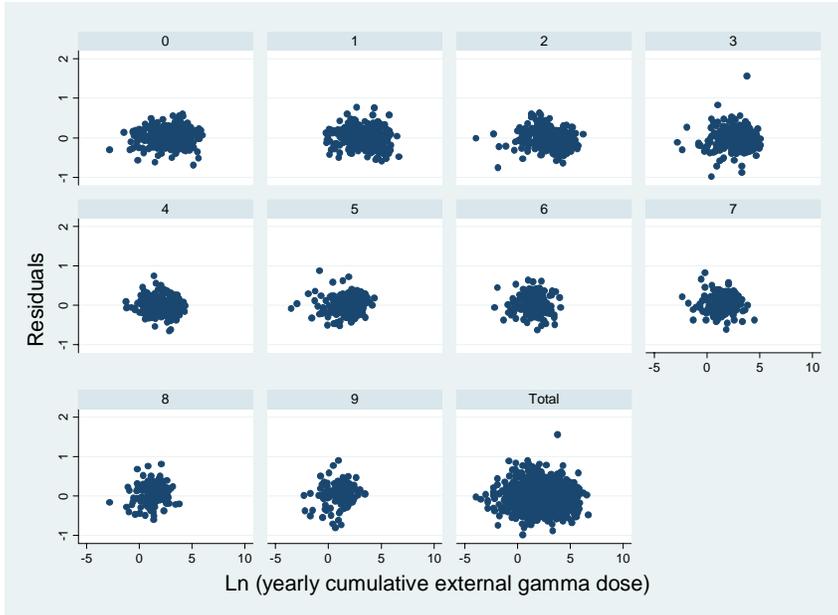


Figure 34: Total; Model#2; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years since the First External Gamma Exposure

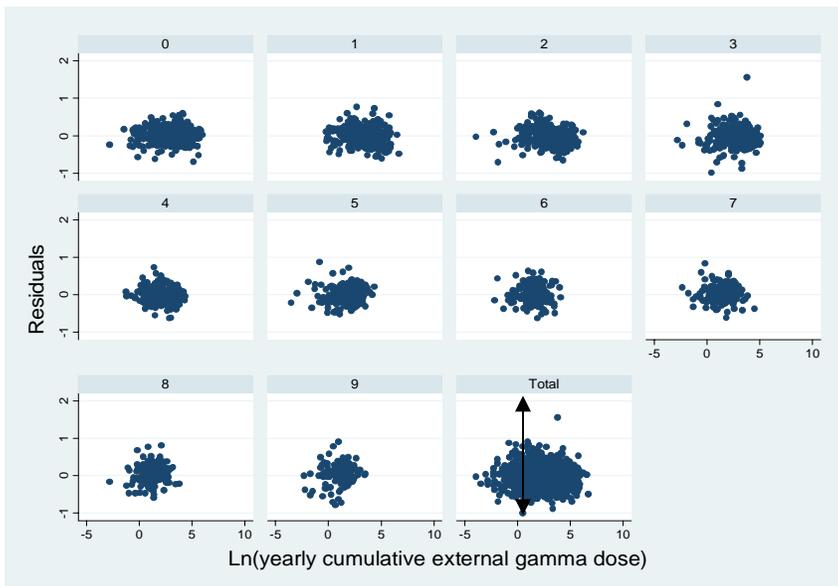


Figure 35: Total; Model #3; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure

↕ indicates the 1nd knot location used for splines implementation in Model 3 according to the change in the distribution of the residuals; it will be used for explaining the location of the 1st knot

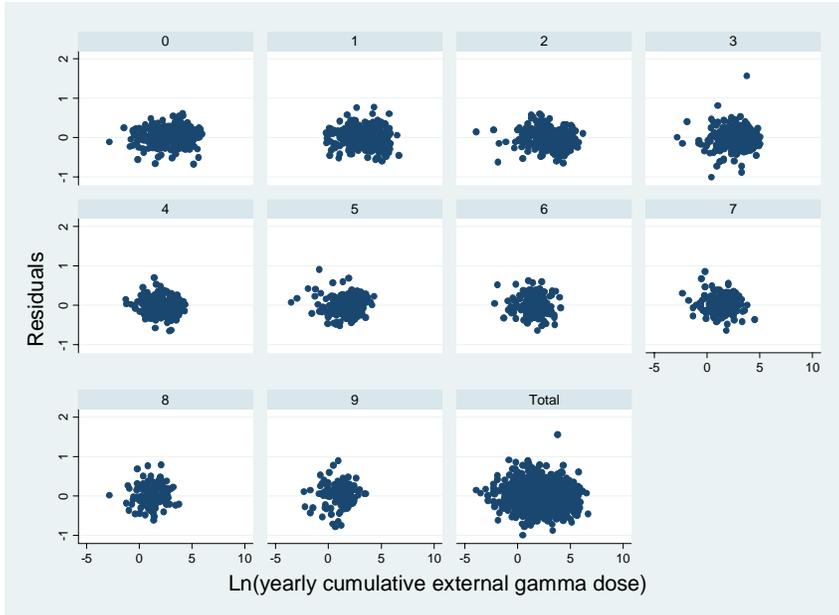


Figure 36: Total; Model#4; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure

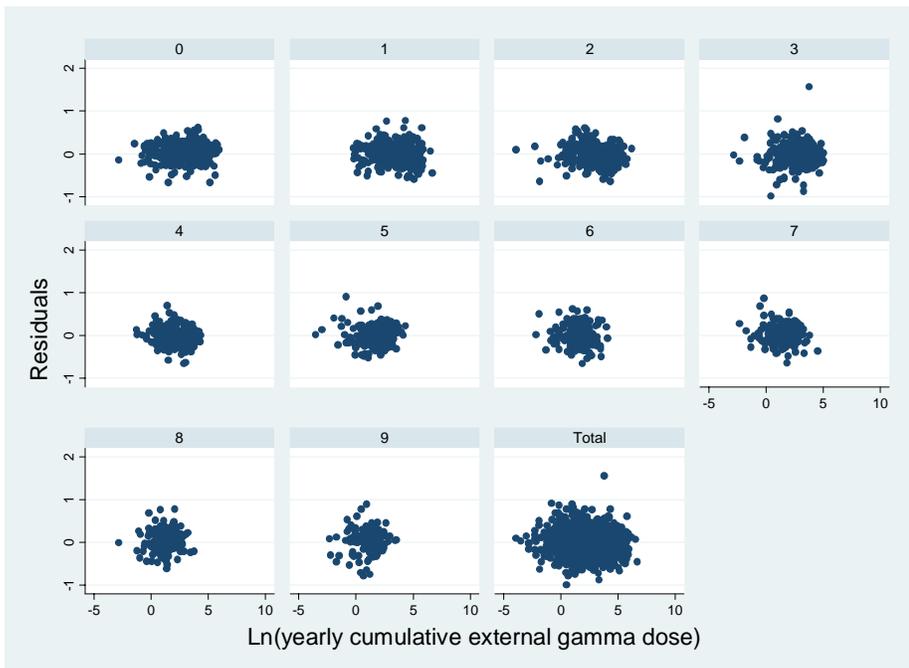


Figure 37: Total; Model#5; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure

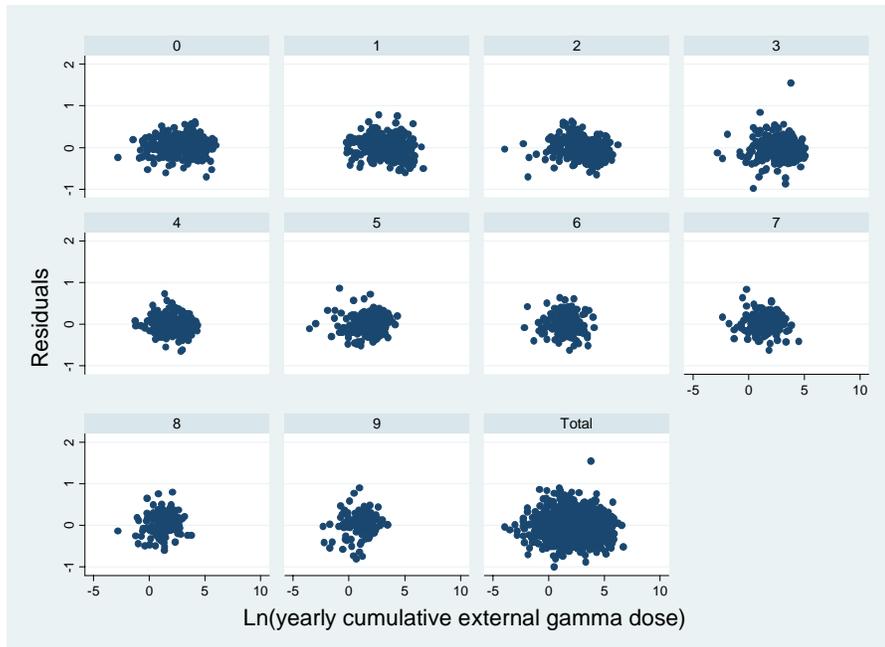


Figure 38: Total; Model#6; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure

In the overall study cohort, in all the six models the residuals appear to be randomly distributed against the explanatory covariate which is the log-transformed yearly cumulative external gamma dose. This is illustrated by the “Total” scatter plot. Furthermore, analyzing every year from the first external gamma exposure, the residuals do not show any pattern when plotted against the explanatory covariate which is the log-transformed yearly cumulative external gamma dose.

2b. Study cohort - males

The goodness of fit of the five models is compared in males separately using the scatter plot of the residuals against the main covariate (log-transformed yearly cumulative external gamma exposure.) Scatter plots are drawn in males by years since the first external gamma exposure occurred. If the models fit well, the scatter plots should show no pattern.

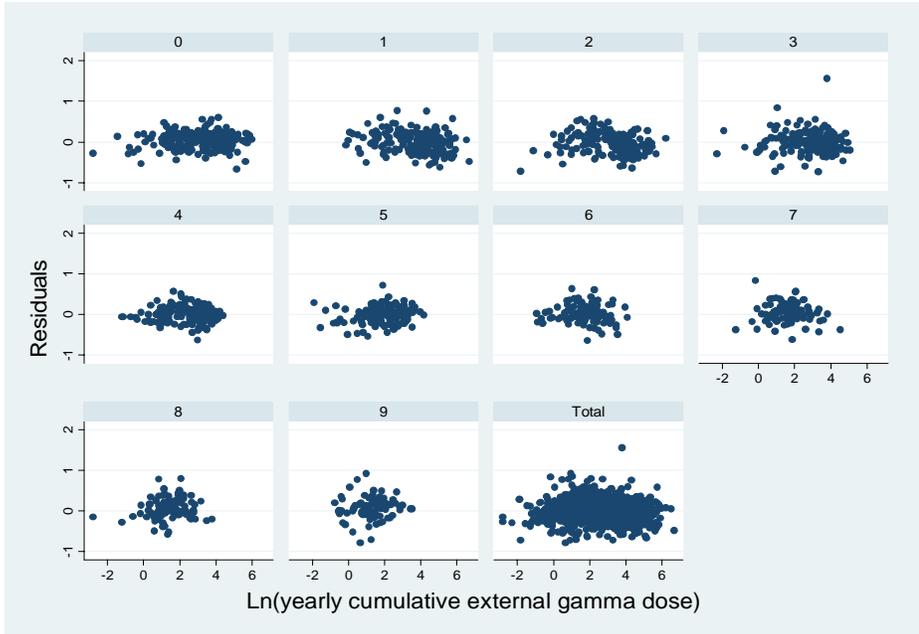


Figure 39: Males; Model#1; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure

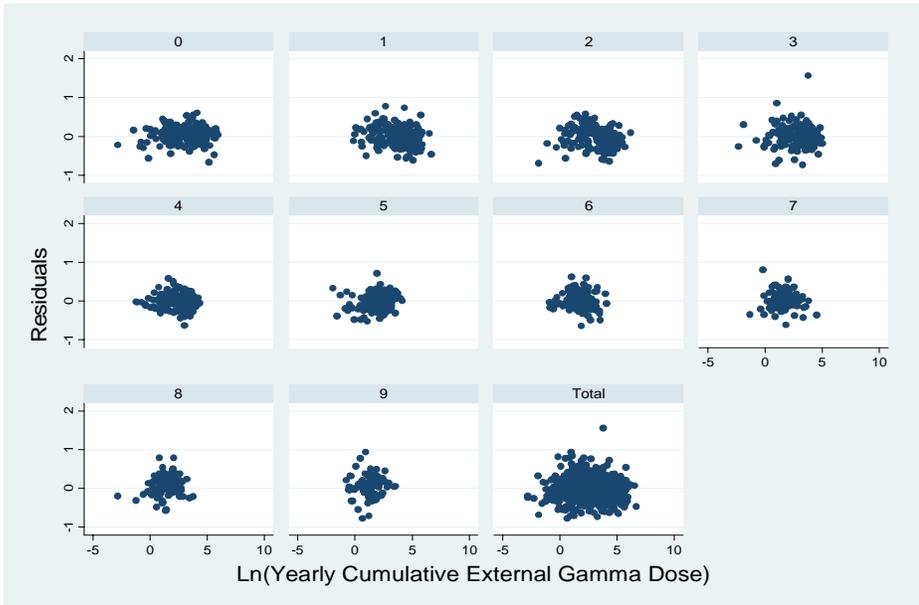


Figure 40: Males; Model#3; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure

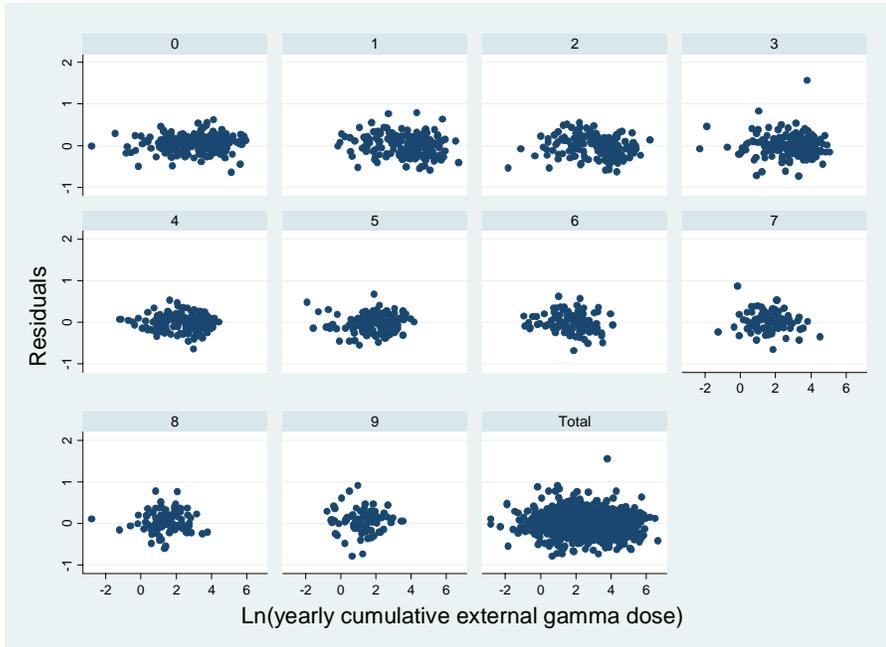


Figure 41: Males: Model#4; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure

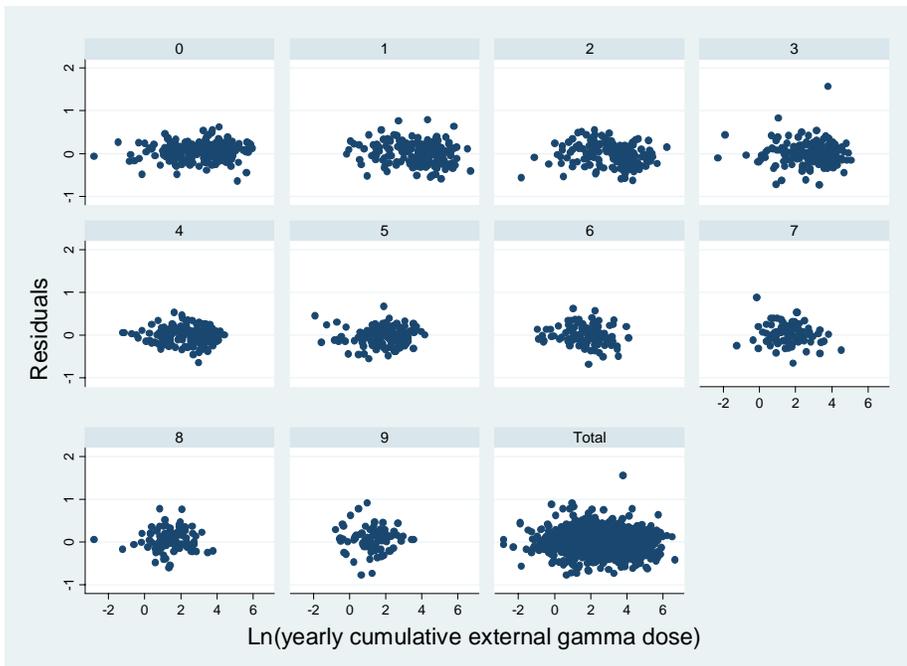


Figure 42: Males; Model#5; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure

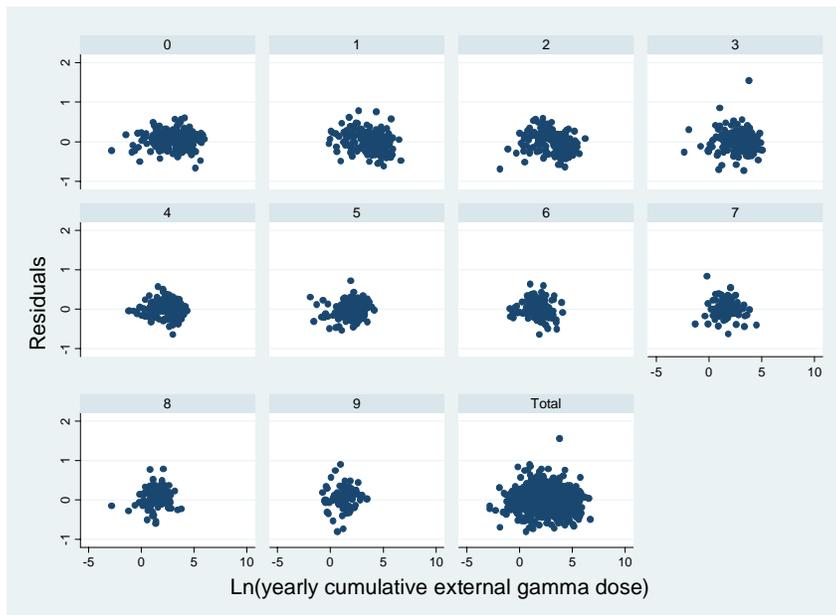


Figure 43: Males; Model#6; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure

In males, in all six models the residuals appear randomly distributed against the explanatory covariate which is the log-transformed yearly cumulative external gamma dose. This is illustrated by the “Total” scatter plot. Furthermore, looking at every year from the first external gamma exposure, the residuals do not show any pattern for most of the points in time when plotted against the explanatory covariate which is the log-transformed yearly cumulative external gamma dose.

2c. Study cohort – females

The goodness of fit of the six models is also compared in females separately using the scatter plot of the residuals against the explanatory covariate (log-transformed yearly cumulative external gamma exposure.) Scatter plots are drawn in females by years since the first external gamma exposure occurred. If the models fit well, the scatter plots should show no pattern.

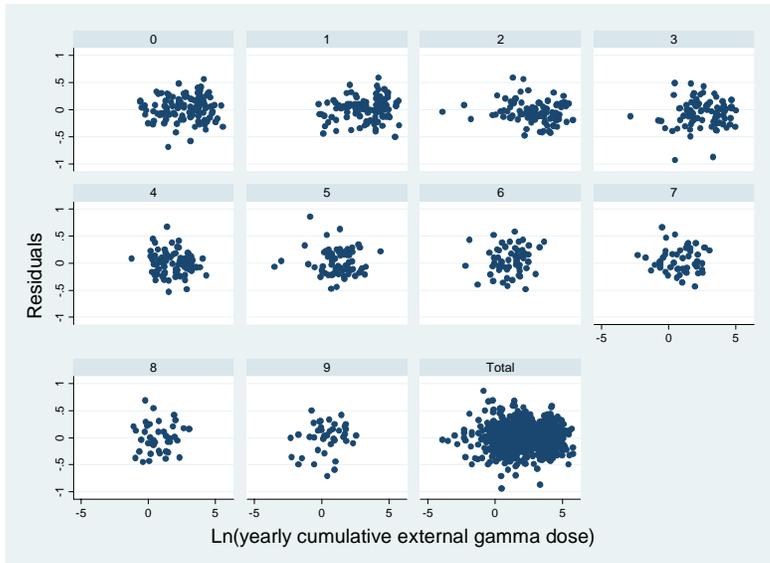


Figure 44: Females; Model#1; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure

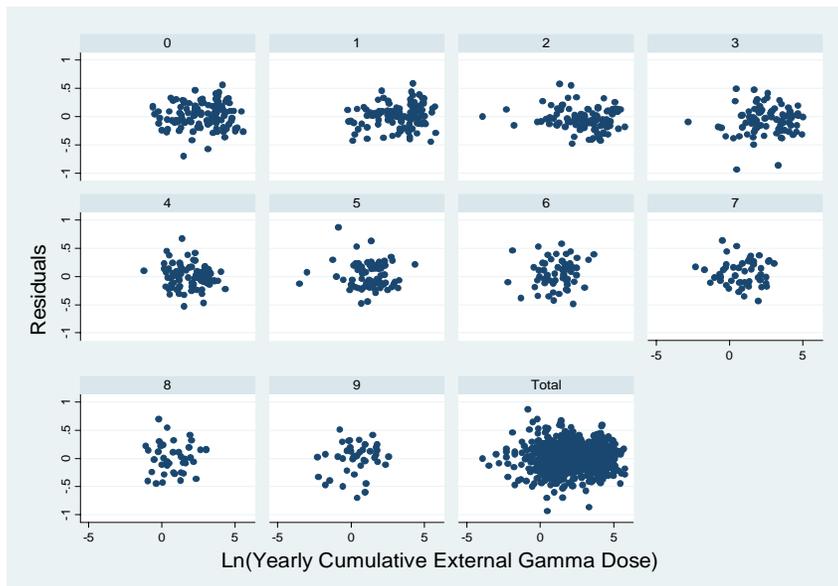


Figure 45: Females; Model#3; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure

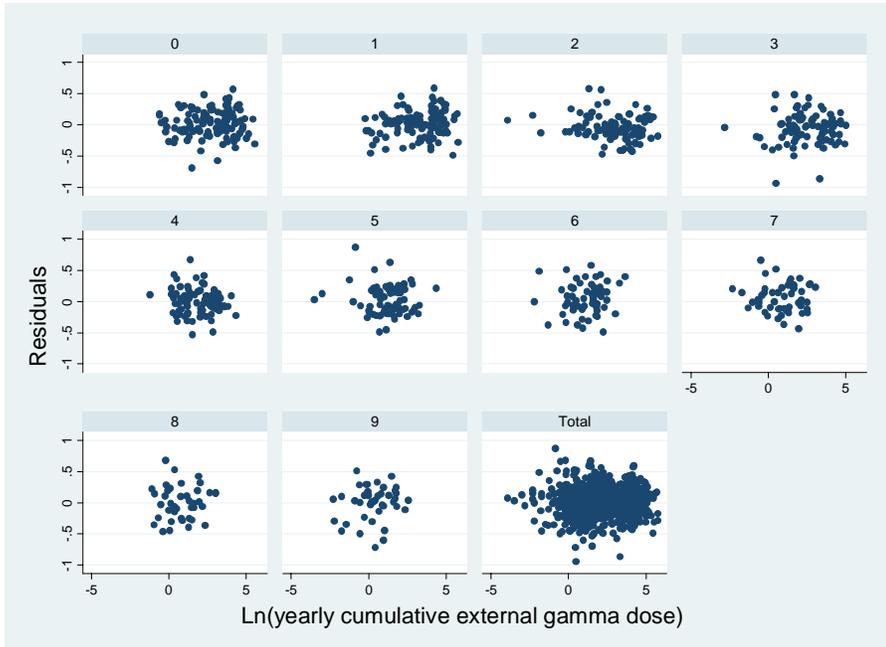


Figure 46: Females; Model#4; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure

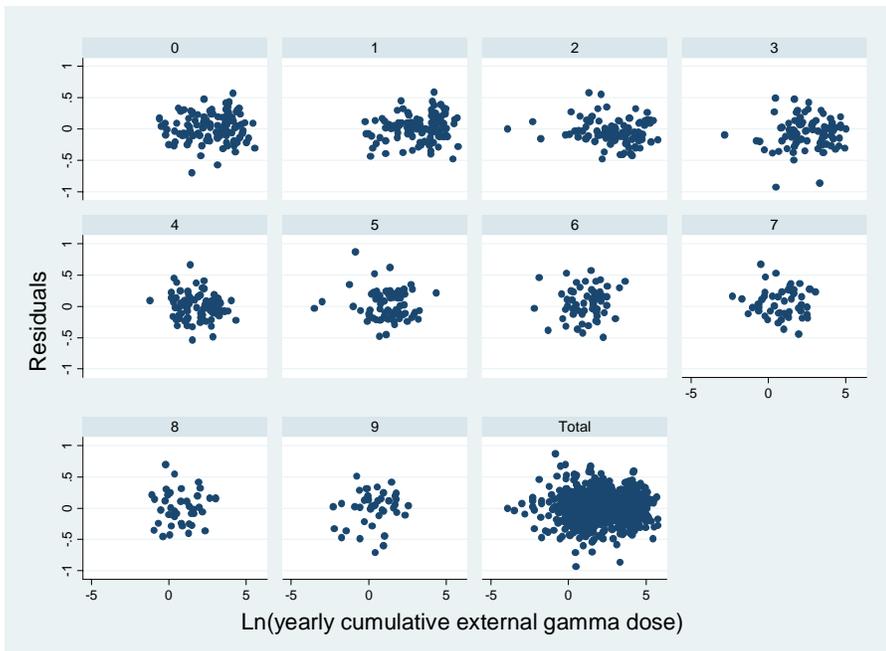


Figure 47: Females; Model #5; Scatter Plot of Residuals versus the Log-Transformed Yearly Cumulative External Gamma Exposure by Years Since the First External Gamma Exposure

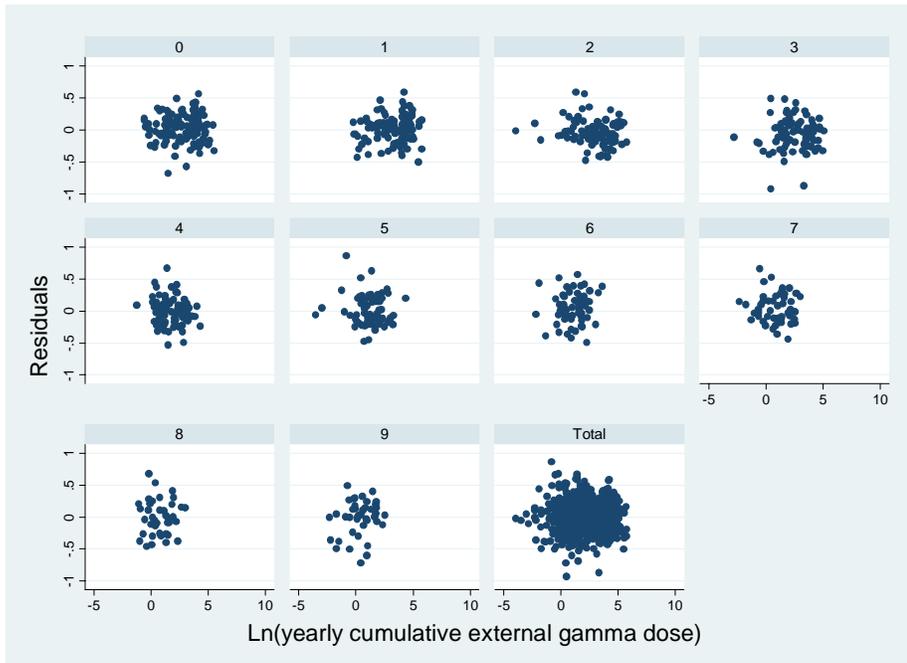


Figure 48: Females; Model#6; Scatter Plot of Residuals versus the Log-Transformed YearlyCumulative External Gamma Exposure by Years Since the First External Gamma Exposure

In females, in all five models the residuals appear randomly distributed against the explanatory covariate which is the log-transformed yearly cumulative external gamma dose. This is illustrated by the “Total” scatter plot. Furthermore, looking at every year from the first external gamma exposure, the residuals do not show any pattern for most of the points in time when plotted against the explanatory covariate which is the log-transformed yearly cumulative external gamma dose.

3) Scatter plots of the residuals against the fitted values. Scatter plots are drawn by years since the first external gamma exposure occurred.

3a. Overall study cohort

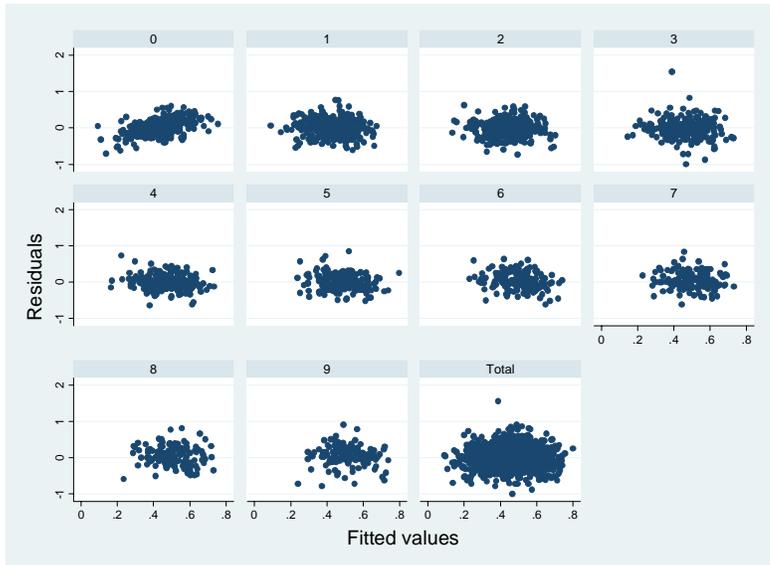


Figure 49: Total; Model#1; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

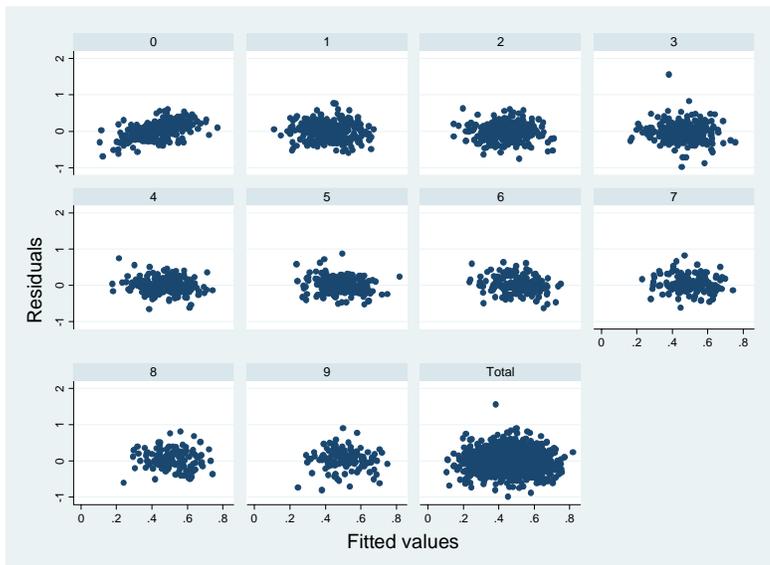


Figure 50: Total; Model#2; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

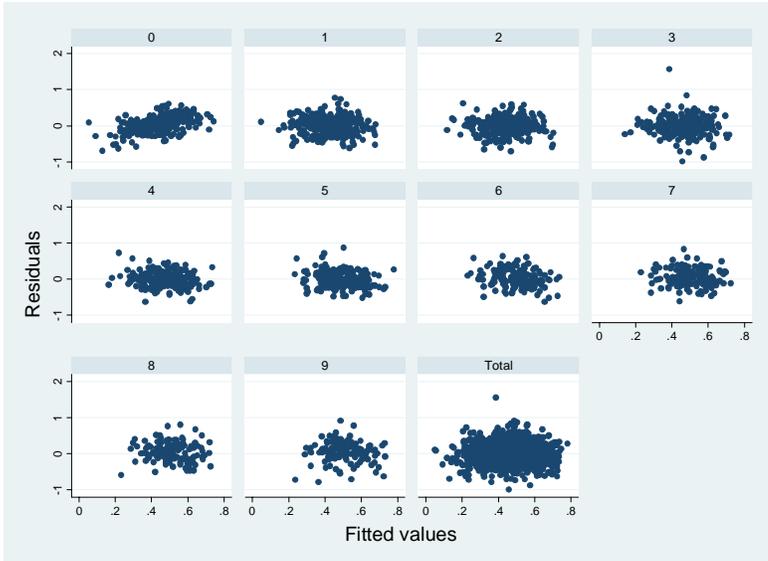


Figure 51: Total; Model#3; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

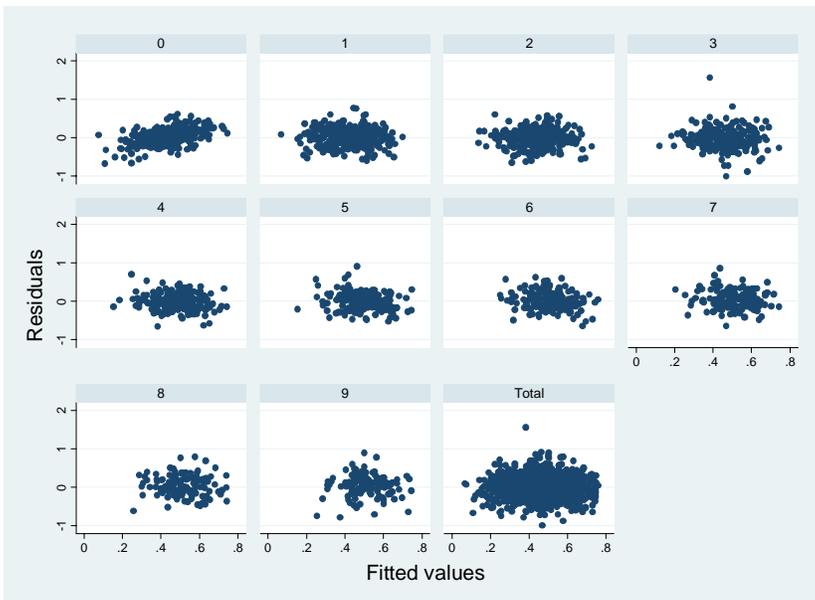


Figure 52: Total; Model#4; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

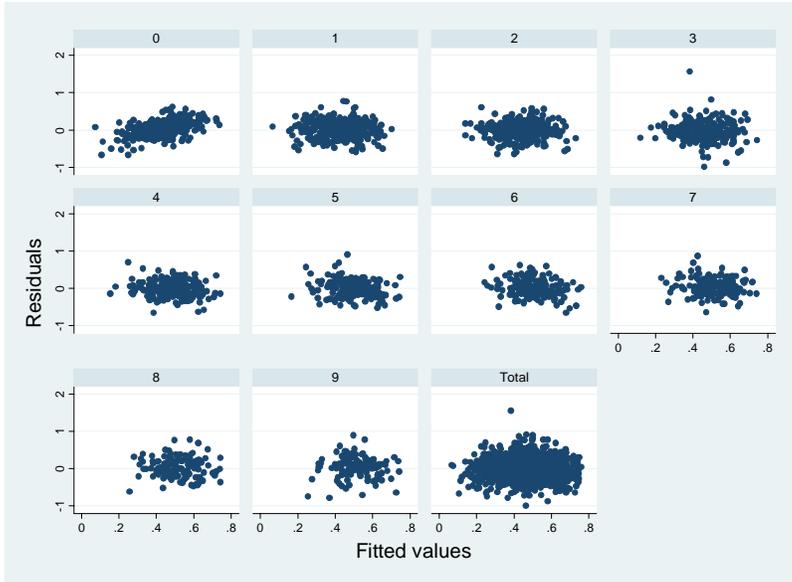


Figure 53: Total; Model#5; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

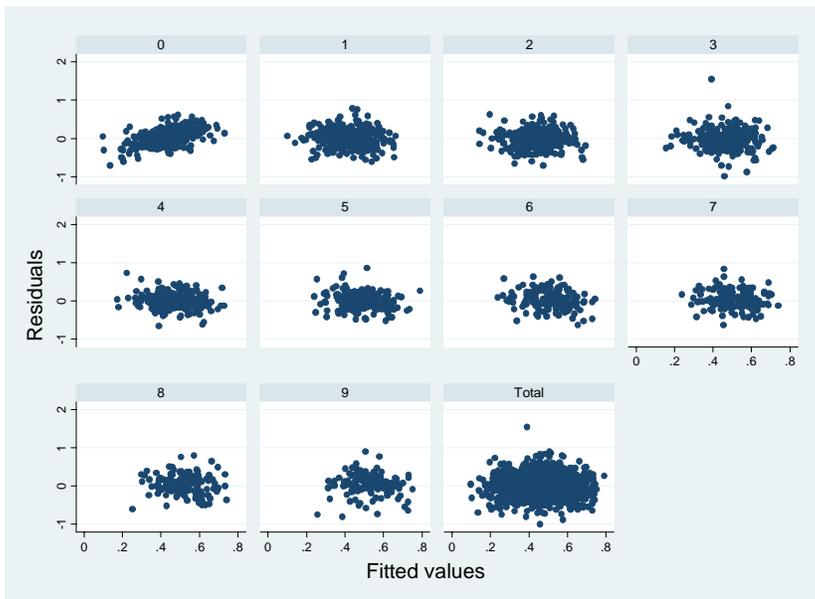


Figure 54: Total; Model#6; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

As illustrated in the “Total” scatter plot, in all the five models the plots of the residuals against the fitted values do not show any pattern. Furthermore, looking at every year from the first external gamma exposure, the residuals do not show any pattern for most of the points in time when plotted against the fitted values.

3b. Study cohort - males

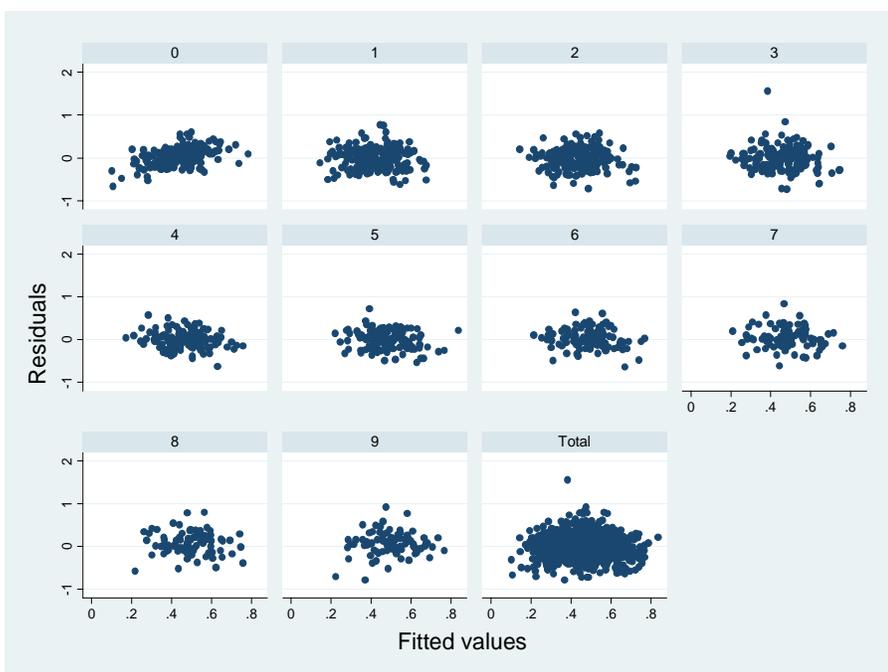


Figure 55: Males; Model#1; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

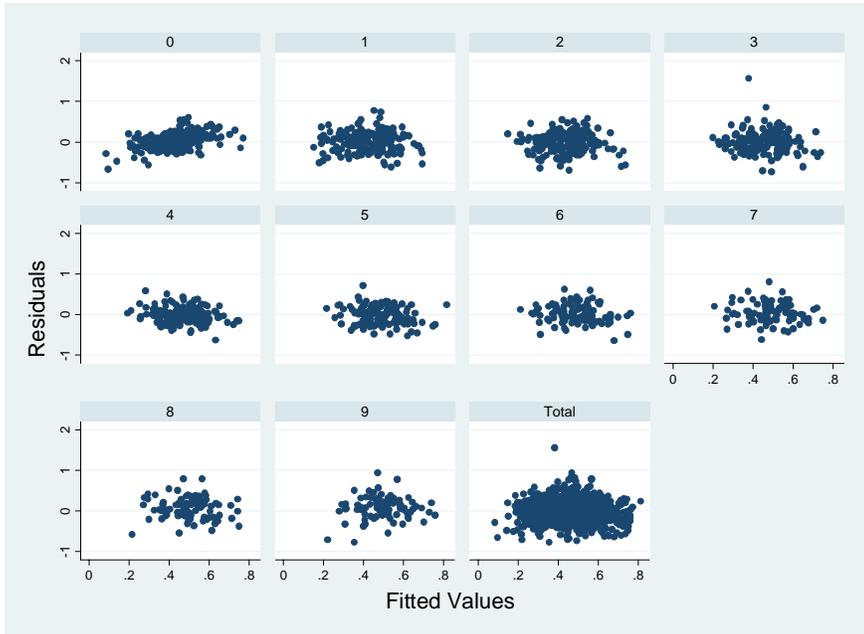


Figure 56: Males; Model#3; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

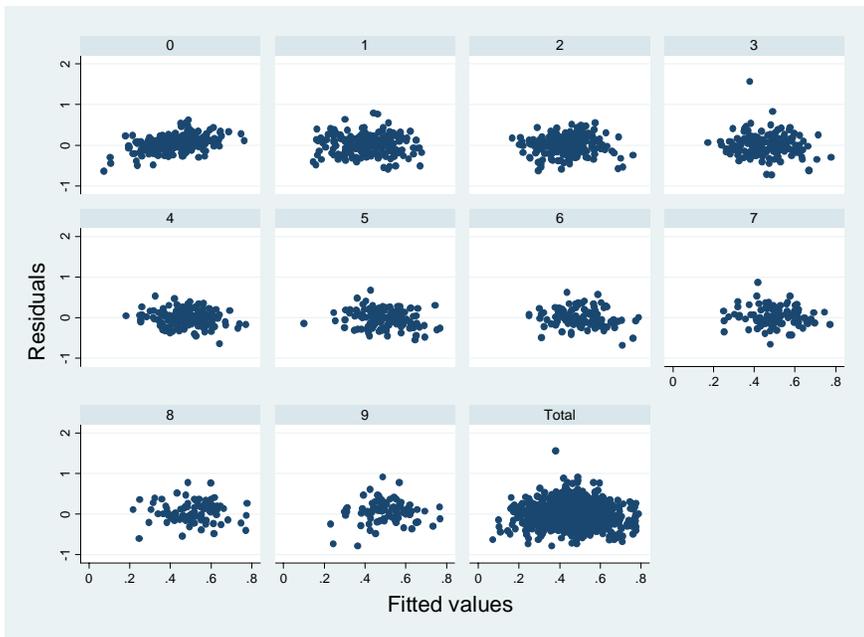


Figure 57: Males; Model#4; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

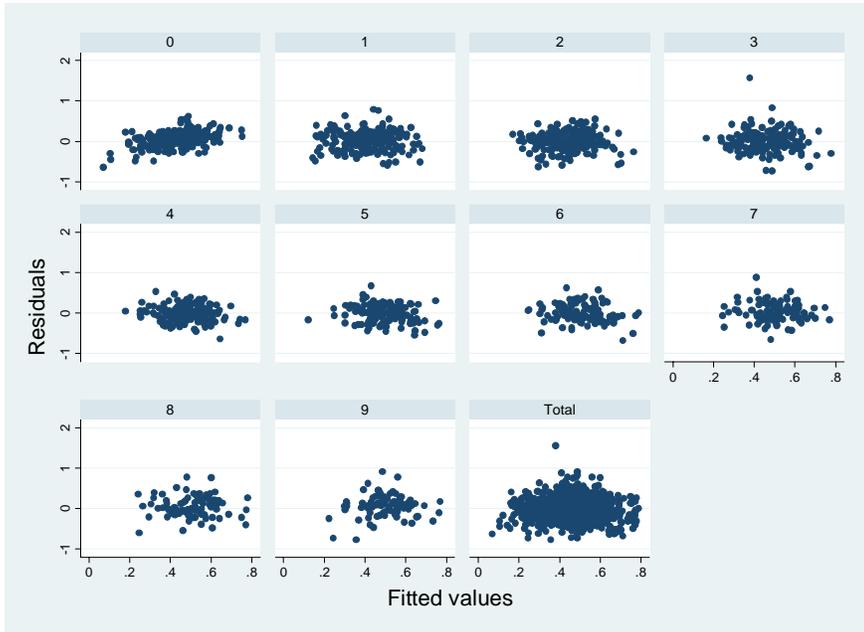


Figure 58: Males; Model#5; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

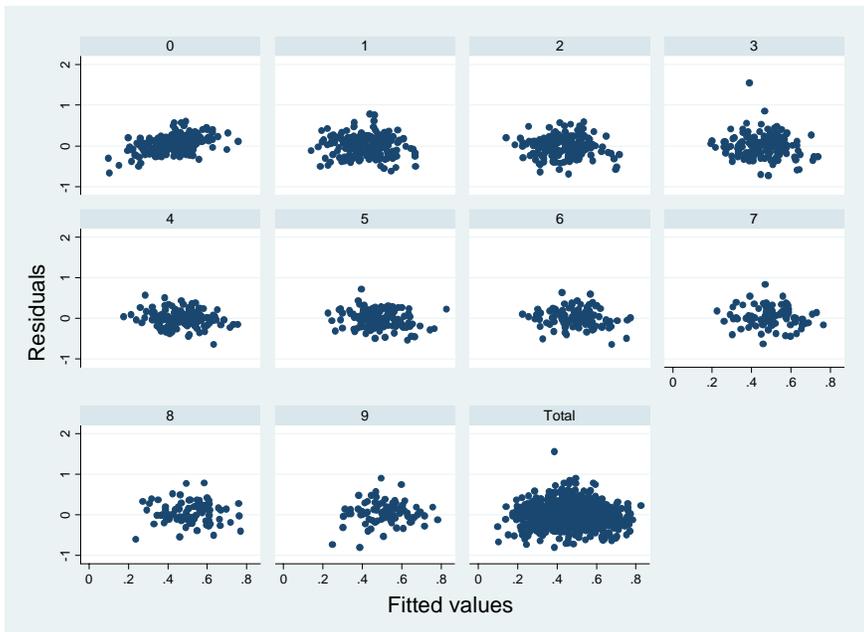


Figure 59: Males; Model#6; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

As illustrated in the “Total” scatter plots, in males in all five models, the plots of the residuals against the fitted values do not show any pattern. Furthermore, looking at every year from the first external gamma exposure, the residuals plotted against the fitted values do not show any pattern for most of the points in time.

3c. Study cohort - females

The goodness of fit of the five models is also compared in females separately using the scatter plot of the residuals against the fitted values; scatter plots are drawn in females by years since the first external gamma exposure. If the models fit well, the scatter plots should have no pattern.

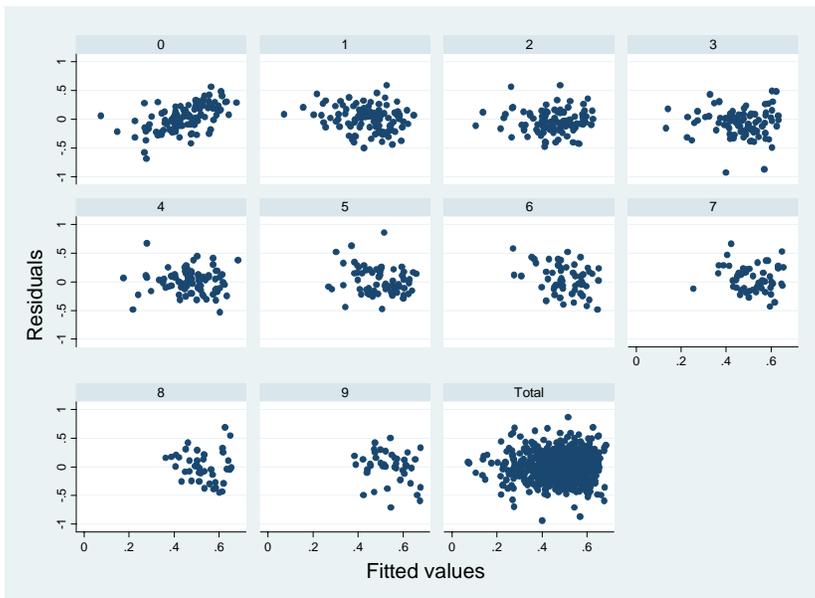


Figure 60: Females; Model#1; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

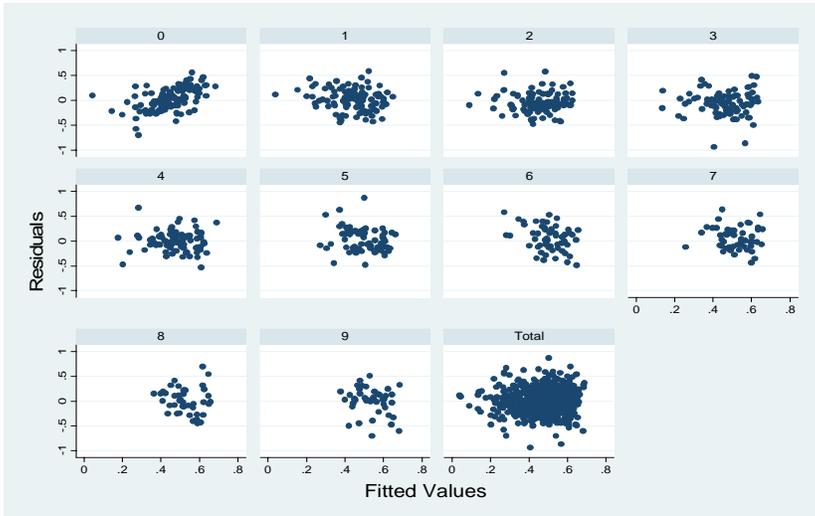


Figure 61: Females: Model#3; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

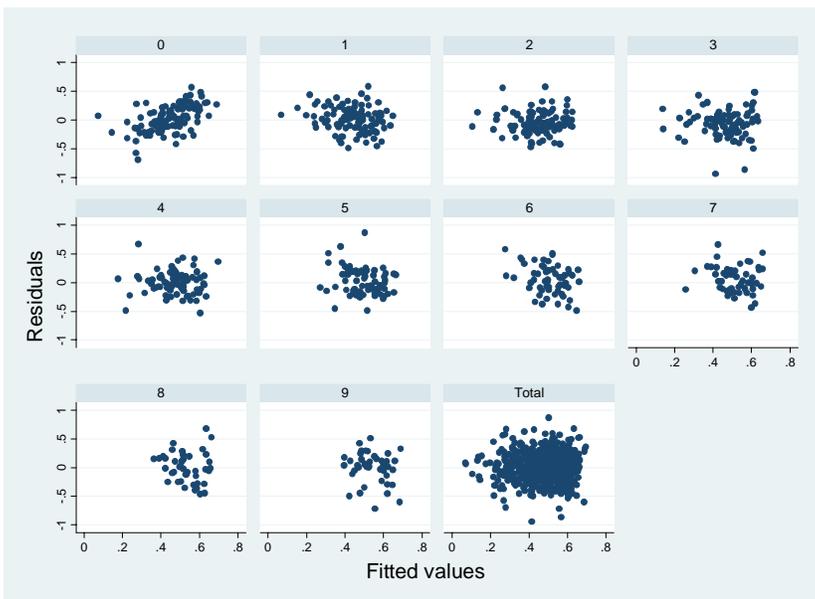


Figure 62: Females; Model#4; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

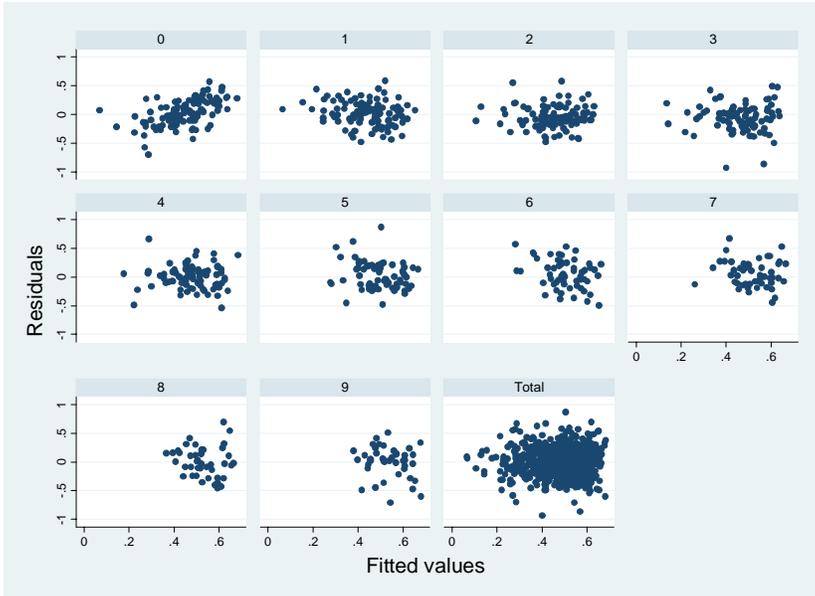


Figure 63: Females; Model#5; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

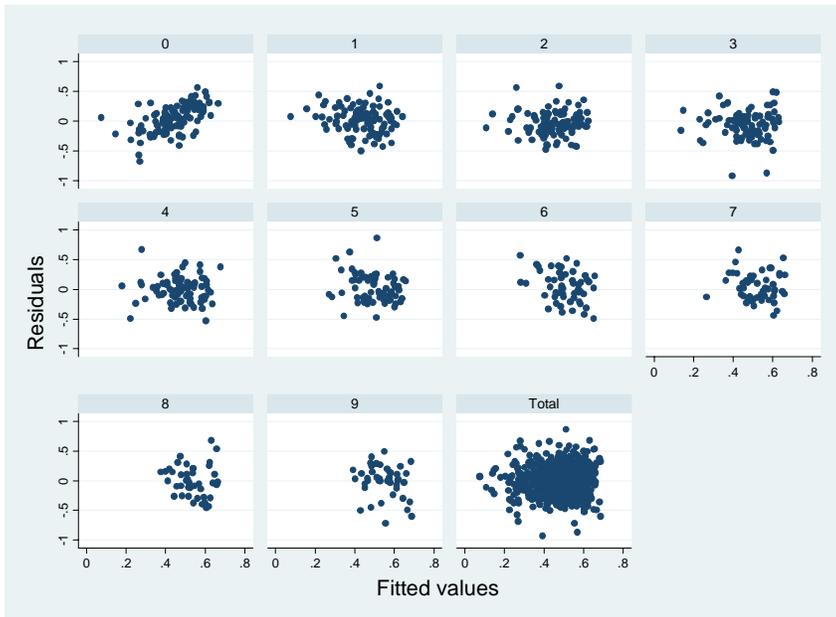


Figure 64: Females: Model#6; Scatter Plot of Residuals versus the Fitted Values by Years Since the First External Gamma Exposure

As illustrated in the “Total” scatter plots, in females in all five models the plots of the residuals against the fitted values do not show any pattern. Furthermore, looking at every year

from the first external gamma exposure, the residuals plotted against the fitted values do not show any pattern for most of the points in time.

The graphical assessment of the goodness of fit indicates that the five models fit well in males, females, and the overall study cohort.

3.3.3 Numerical assessment of goodness of fit of the models

The indicators used for goodness of fit assessment are described in detail in the methodology section. They are summarized as follows:

1) the Wald-Wolfowitz statistic, used to test if the residuals have a random distribution in a repeated measures setting⁶³

$$W_z = \frac{T - E(T)}{\sqrt{V(T)}} \quad \text{where:} \quad E(T) = \frac{2n_p n_n}{n_n + n_p} + 1$$

$$V(T) = \frac{2n_p n_n (2n_p n_n - n_n - n_p)}{(n_n + n_p)^2 (n_n + n_p - 1)}$$

n_n = number of negative residuals

n_p = number of positive residuals

T = number of runs = how many times the sign of residual changes

Under the null hypothesis, W_z is distributed as normal (0,1). Therefore, values of W_z less than -1.96 or greater than 1.96 are associated with non-random distribution of the residuals.

2) GEE-R² computation according to the formula:

$$R^2_m = 1 - \frac{\sum_{i=1}^{T_i} \sum_{i=1}^n (Y_{it} - \hat{Y}_{it})^2}{\sum_{i=1}^T \sum_{i=1}^n (Y_{it} - \bar{Y})^2}$$

where:

y_{it} =the observed variable measured for i-th individual at time t

\hat{y}_{it} =the predicted variable for i-th individual at time t

$$\bar{Y} = \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n Y_{it} = \text{the overall mean}$$

3) QICu adapted for longitudinal data is calculated⁵⁵ according to the formula:

$$QICu = -2Q(g^{-1}(y)) + 2p$$

where $Q = -\frac{1}{2} \sum (y - \hat{y})^2$ is the value of quasi-likelihood function calculated for the independence model but with the regression coefficients fitted for the hypothesized correlation structure.

y =the observed value

\hat{y} =the fitted values

The numerical assessment of goodness of fit is performed in each model. The Wald-Wolfowitz statistic, GEE-R² and the QICu are calculated in each model for the overall study cohort and separately for males and females. The results are presented in Tables 57-59.

Table 56: Total; Goodness of Fit Assessment (Residuals distribution, GEE-R², QICu) in Each of the Five Models

	Residuals distribution around 0	GEE-R ²	QICu
Model #1	random	0.61	12.11
Model#2	random	0.61	14.10
Model #3	random	0.61	14.08
Model #4	random	0.60	16.09
Model #5	random	0.60	14.08
Model #6	random	0.61	14.11

Table 57: Males; Goodness of Fit Assessment (Residuals distribution, GEE-R², QICu) in Each of the Five Models

	Residuals distribution around 0	GEE-R ²	QICu
Model #1	random	0.62	10.00
Model#2	NA	NA	NA
Model #3	random	0.62	12.00
Model #4	random	0.60	14.00
Model #5	random	0.60	12.00
Model #6	random	0.62	12.00

Table 58: Females; Goodness of Fit Assessment (Residuals distribution, GEE-R², QICu) in Each of the Five Models

	Residuals distribution around 0	GEE-R ²	QICu
Model #1	random	0.58	10.10
Model #2	NA	NA	NA
Model #3	random	0.58	12.07
Model #4	random	0.58	14.10
Model #5	random	0.58	12.10
Model #6	random	0.58	12.11

The results presented in Tables 57-59 show that in all models the residuals can be considered randomly distributed around zero. About 60% of the variation of the outcome variable is explained by the model. This is considered a good percent of variation explained for a repeated measures design⁴⁸.

QICu has only slight variation among all analyzed models. In the same models, after stratification by sex, the sex variable is omitted and QICu decreases as the number of covariates decreases by one.

Thus, all five models have a similar goodness of fit. After stratification by sex, the goodness of fit level stays similar across models. In terms of the differential between males and females, Tables 58 and 59 show that males and females have a similar goodness of fit level for the five analyzed models. However, in females compared with males there is a slight decrease in

the GEE- R^2 and a slight increase in QICu. This result shows that the five statistical models fit slightly better in terms of GEE- R^2 and QICu in males than in females.

3.3.4 Predicting the lymphocyte counts using the five models

The models predict the lymphocyte counts as a function of external gamma dose in males and females separately, while adjusting for baseline lymphocyte counts, work location related for Plutonium exposure, smoking history and alcohol consumption at start of employment.

Thus, the five models are applied to the study cohort. Lymphocyte counts are predicted for yearly cumulative external gamma doses of 1 rad and 5 rads which represent the knots location for the linear splines. In terms of the adjustment variables, the baseline lymphocyte count is set to equal the median baseline lymphocyte count. The categorical variables of smoking history, alcohol consumption and work location related to Plutonium exposure are set at baseline.

The first prediction of the lymphocyte counts is performed for a yearly cumulative external gamma dose of 1 rad. All five models are run in males and females separately. Lymphocyte counts are predicted for the second year of follow-up in non-smokers, non-alcohol consumers, who worked at the reactor or at other locations and who had a baseline lymphocyte count of $1.6 \times 1000/\text{mm}^3$. The following calculations have been done:

For external gamma dose=1 rads

Dose=1 rad $\ln(1)=0$

The term including the $\ln(\text{dose})$ equals zero for dose=1 rad. Moreover, in Model 4, the three splines calculated at dose=1 rads with knots implemented at dose=1 rads and dose =5 rads

are zero. Finally, in Model 5, the two splines were calculated at dose=1 rads with one knot implemented at dose=5 rads. Both of the implemented splines are zero.

The baseline lymphocyte count=1.6 and the $\ln(\text{baseline lymphocyte counts})=0.47$

Model#1 males

$$E[\ln(\text{lymphocyte counts})]=0.35+0.32*0.47=0.5004$$

$$\text{lymphocyte counts}=e^{0.5004}=1.65$$

Model#1 females

$$E[\ln(\text{lymphocyte counts})]=0.41+0.24*0.47=0.5228$$

$$\text{lymphocyte counts}=e^{0.5228}=1.69$$

Model#3 males

$$E[\ln(\text{lymphocyte counts})]=0.34+0.32*0.47=0.49$$

$$\text{lymphocyte counts}=e^{0.49}=1.63$$

Model#3 females

$$E[\ln(\text{lymphocyte counts})]=0.40+0.24*0.47=0.5128$$

$$\text{lymphocyte counts}=e^{0.5128}=1.67$$

Model#4 males

$$E[\ln(\text{lymphocyte counts})]=0.32+0.32*0.47=0.4704$$

$$\text{lymphocyte counts}=e^{0.4704}=1.6$$

Model#4 females

$$E[\ln(\text{lymphocyte counts})]=0.43+0.23*0.47=0.5381$$

$$\text{lymphocyte counts}=e^{0.5381}=1.71$$

Model#5 males

$$E[\ln(\text{lymphocyte counts})]=0.31+0.32*0.47=0.4604$$

$$\text{lymphocyte counts}=e^{0.4604}=1.58$$

Model#5 females

$$E[\ln(\text{lymphocyte counts})]=0.41+0.23*0.47=0.5181$$

$$\text{lymphocyte counts}=e^{0.5181}=1.68$$

Model#6 males

$$E[\ln(\text{lymphocyte counts})]=0.33+0.004+0.32*0.47=0.4844$$

$$\text{lymphocyte counts}=e^{0.4844}=1.62$$

Model#6 females

$$E[\ln(\text{lymphocyte counts})]=0.40+0.002+0.23*0.47=0.5101$$

$$\text{lymphocyte counts}=e^{0.5101}=1.67$$

Dose=5 rads

$$\text{Dose}=5 \text{ rads} \quad \ln(5)=1.61$$

In Model 4, the three splines were calculated at dose=5 rads with knots implemented at dose=1 rads and dose=5 rads. The three splines are: spline1=0, spline2=1.61, spline3= 9.43×10^{-7} .

In Model 5, the two splines were calculated at dose=5 rads with one knot implemented at dose=5 rads. Thus, spline1=1.61, spline2= 9.43×10^{-7} .

$$\text{The baseline lymphocyte count}=1.6 \quad \text{and the } \ln(\text{baseline lymphocyte counts})=0.47$$

Model#1 males

$$E[\ln(\text{lymphocyte counts})]=0.35+0.32*0.47-0.04*1.61=0.436$$

$$\text{lymphocyte counts}=e^{0.436}=1.55$$

Model#1 females

$$E[\ln(\text{lymphocyte counts})]=0.41+0.24*0.47-0.02*1.61=0.4906$$

$$\text{lymphocyte counts}=e^{0.4906}=1.63$$

Model#3 males

$$E[\ln(\text{lymphocyte counts})]=0.34+0.32*0.47-0.03*1.61=0.4421$$

$$\text{lymphocyte counts}=e^{0.4421}=1.56$$

Model#3 females

$$E[\ln(\text{lymphocyte counts})]=0.40+0.24*0.47-0.01*1.61=0.4967$$

$$\text{lymphocyte counts}=e^{0.4967}=1.64$$

Model#4 males

$$E[\ln(\text{lymphocyte counts})]=0.32+0.32*0.47+1.61*0.01-0.06*9.43*10^{-7}=0.49$$

$$\text{lymphocyte counts}=e^{0.49}=1.63$$

Model#4 females

$$E[\ln(\text{lymphocyte counts})]=0.43+0.23*0.47-1.61*0.03-9.43*10^{-7}*0.03=0.49$$

$$\text{lymphocyte counts}=e^{0.49}=1.63$$

Model#5 males

$$E[\ln(\text{lymphocyte counts})]=0.31+0.32*0.47+0.02*1.61-0.06*9.43*10^{-7}=0.49$$

$$\text{lymphocyte counts}=e^{0.49}=1.63$$

Model#5 females

$$E[\ln(\text{lymphocyte counts})]=0.41+0.23*0.47-0.01*1.61-0.03*9.43*10^{-7}=0.50$$

$$\text{lymphocyte counts}=e^{0.50}=1.65$$

Model#6 males

$$E[\ln(\text{lymphocyte counts})]=0.33+0.005+0.32*0.47-0.03*1.61=0.44$$

$$\text{lymphocyte counts}=e^{0.44}=1.55$$

Model#6 females

$$E[\ln(\text{lymphocyte counts})]=0.4+0.002+0.23*0.47-0.02*1.61=0.4779$$

$$\text{lymphocyte counts}=e^{0.4779}=1.61$$

Table 59: Expected Values of the Lymphocyte Counts in Males and Females, Non-Smokers, Non-Alcohol Consumers, Working at the Reactor or Other Locations with a Baseline Lymphocyte Count of $1.6*1000/\text{mm}^3$ and Being Exposed at 1 rads and 5 rads Yearly Cumulative External Gamma Radiation. Models 1,2,3,4, and 5 are fit in males and females separately

	Dose=1 rads		Dose=5 rads	
	Males	Females	Males	Females
Model#1	1.65	1.69	1.55	1.63
Model#2	NA	NA	NA	NA
Model#3	1.63	1.67	1.56	1.64
Model#4	1.60	1.71	1.63	1.63
Model#5	1.58	1.68	1.63	1.65
Model#6*	1.62	1.67	1.55	1.61

*Model 6 includes the number of years from the first external gamma exposure; the second year was considered in this calculation

The above table shows a decline in lymphocyte counts at 5 rads versus 1 rad yearly cumulative gamma exposure. The decrease of the lymphocyte counts occurs in both males and females. Females have slightly higher expected values of lymphocyte counts than males.

To summarize, the comparative evaluation of the five models suggests that all models have a similar and satisfactory goodness of fit. When applied to the study cohort data the models lead to results that are numerically close.

Important theoretical and regulatory considerations³ in conjunction with the evaluation of the goodness of fit results recommend Model#1 as the most adequate choice for the study analysis.

Within Model #4, the linear splines implemented with a knot located at 5 rads reveal differences between males and females at exposures below 5 rads. These differences pertain

mainly to the direction of the coefficients. The linear spline generated at doses below 5 rads is a borderline statistically non-significant positive predictor in males ($p=0.09$), while it is a statistically non-significant negative predictor in females ($p=0.16$). Interestingly, when predicting the expected values of the $\ln(\text{lymphocyte counts})$ separately in males and females, the numerical results are very close. This is an aspect that requires further discussion.

4.0 DISCUSSION

Prior to more detailed discussion of the findings, it is worthwhile to summarize the key results of the statistical analysis:

- a) There is a statistically significant relationship between the log-transformed lymphocyte counts and the log-transformed external gamma dose. As the log-transformed external gamma dose increases, the log-transformed lymphocyte counts decrease.
- b) The linear radiation dose-response model is considered appropriate for the data of the study cohort.
- c) There are not statistically significant differences between males and females regarding the effect of occupational radiation exposure on the lymphocyte counts.

The results of the descriptive analysis along with the results of the statistical models raise many important concerns. The discussion focuses on the following issues: the special characteristics of the study cohort, the radiation dose-response modeling issues, the sophisticated statistical methodology used in this project, and on the new statistical tools developed and implemented in this dissertation.

A valuable feature of this dissertation research relates to the structure of the study cohort. The steps followed to generate the study cohort were described in detail in the methodology section. The study cohort consists of 223 males and 130 females hired by Mayak PA nuclear facility between 1948 and 1960 who were exposed for many years to radiation. For the purposes

of this dissertation analysis, the study cohort is followed-up for ten years from the first exposure to external gamma radiation.

Although the number of workers decreases during the ten years of follow-up due to the drop out phenomenon, the proportions of males and females remain stable. At the end of the follow-up period the study cohort still includes 141 males and 82 females. This study cohort allowed a detailed sex-stratified statistical analysis.

The age distribution of the workers included in the cohort shows a young study population. The workers included in the study cohort are on the average 25 years old at the start of employment. There is a difference of 2 years between the average age of males and females, females being slightly younger than males. Despite this difference, the males and females who are included in the study cohort can be considered to have similar age distributions.

Other characteristics of the study cohort derive from the political, educational, and health status criteria used by Mayak PA in selecting radiation workers. The workers hired at Mayak PA were the first Russian professionals specialized in radioactive Plutonium processing. These workers were young, well educated and they had to be healthy in order to be able to work long hours in a dangerous environment⁶⁹. Due to this rigorous selection, the Mayak PA employees exhibit over time an overall health status better than that of the general population. As a consequence, the effects of dangerous occupational exposures may be less obvious in this pre-selected healthy subgroup of the population than they would be in the general population. It can be assumed that the younger the workers, the healthier and more resilient to occupational exposures they are likely to be.

The interpretation of the results of this study has to take into consideration the excellent health status that characterizes Mayak PA workers at start of their employment in order to avoid

underestimating the damaging effects of occupational radiation exposure on workers. Illustrative of this situation within the framework of this study is the trend of yearly median lymphocyte counts by year since the first external gamma exposure. The descriptive analysis shows a decrease of the yearly median lymphocyte counts as an effect of external gamma radiation exposure which is consistent with the lymphocytes' high sensitivity to radiation described in the literature^{4,5,7,28}. However, most of the yearly median lymphocyte counts do not drop below the normal range presented in Appendix D and cannot be considered an abnormal value of the blood test or an expression of disease. This result does not signify that the long term exposure to radiation does not affect the health of the workers included in the study cohort, but that the good health status and the youth of Mayak PA employees mask to some extent the occupational radiation effects, at least within the follow up interval. Although the decrease of the lymphocyte counts does not reach abnormal levels and cannot be considered a disease, it may lead to immunity disorders, cancer susceptibility or other chronic conditions after long term exposure³⁻⁵. Given all these considerations, the lymphocyte counts drop must be interpreted as an important effect of occupational radiation exposure.

The lymphocyte trends identified in the study are a result of the long term exposure to external gamma radiation. Therefore, the external gamma dose exposure values are critical in the analysis of the study cohort. When the Mayak PA nuclear facility started to operate in 1948, there was almost no attention paid to the radiation protection of workers. According to the results of the descriptive analysis performed in this dissertation, during the second year of follow-up, the workers were exposed to a median external gamma dose of 45 rads/year. This is nine times more than the 5 rads/year which is the highest limit currently accepted for occupational radiation exposure in the US, and this situation makes the current analysis

exceptionally unique as there is no previous research on the effects of such high long-term occupational radiation exposure on the lymphocyte.

The results of the descriptive analysis in this dissertation show that females are exposed to lower occupational radiation doses than males. However, radiation doses received by the female workers are still high. For example, during the second year of follow-up females received a yearly median external gamma dose of 28 rads/year which is still a high gamma radiation dose compared to the limits currently accepted in the US. Interestingly, the yearly external gamma dose decreases in time. The highest exposures of the study cohort are recorded during the first 4 years of follow-up. This exposure encompasses the period of time before 1960, when the peak of radiation exposures at Mayak, PA⁷⁰ were recorded. After 1960 the international regulations regarding occupational radiation exposure limits were implemented and the radiation exposure substantially decreased⁷¹.

The analysis in this dissertation focuses on the effect of the external gamma radiation exposure on the lymphocyte counts. Besides these two variables considered crucial for this dissertation, the statistical models include a number of control variables for which the statistical analysis is adjusted. The control variables are considered potentially to affect the outcome variable and therefore are included in the statistical models although they do not represent the main concern of the study.

The most important control variable in the analysis is the baseline lymphocyte count which is defined as the lymphocyte count recorded at the beginning of the follow-up. This lymphocyte count represents the starting point and is considered important for determining the trend of the lymphocyte counts during the follow-up. The selection of the baseline lymphocyte creates a number of concerns that require further discussion.

As explained in the methodology chapter, the workers are aligned according to the year of the first external gamma exposure. Ideally, the baseline lymphocyte count would be the count that precedes the first radiation exposure. This baseline would be ideal since these values are presumably not affected by any radiation exposure. Thus, the ideal baseline lymphocyte count should be chosen as the last count recorded in the year preceding the first external gamma exposure occurrence. However, as described in the Methodology section, 221 workers do not have any lymphocyte count in the year preceding the first external gamma exposure. Setting the baseline as the first lymphocyte count recorded in the year of the first external gamma exposure would lead to missing baseline counts in 221 out of 353 workers. To avoid this situation the baseline lymphocyte count has been defined as the first baseline count during the year of the first external gamma exposure, as explained in the methodology section.

This study examined the possibility that the first count performed in the year when the first exposure occurred was already affected by the radiation exposure as described in the methodology section. The specific approach used in analyzing this aspect was to compare the last lymphocyte count performed in the year preceding the first exposure with the first lymphocyte count performed during the year when the first exposure occurred. The comparison has been performed in 119 workers who have lymphocyte counts recorded during the year preceding the first exposure and also during the year when first exposure occurred.

The results of the analysis show that in the same group of workers the last lymphocyte count performed during the year preceding the first external gamma exposure can be considered distributed similarly to the first lymphocyte count performed in the year of the first exposure. The lymphocyte counts performed before the first external gamma exposure occurred are not affected by radiation. The lymphocyte counts performed during the year when the first external

gamma exposure occurred might be influenced by radiation if the blood test has been performed after the radiation exposure occurred. The Mayak records specify the month, the day, and the year when each blood count was performed. However, in contrast to the situation with the blood counts, the radiation exposure records do not have any occurrence date and the yearly cumulative external gamma dose is specified for an entire given year. Therefore, during the year of the first radiation exposure it is not possible to establish when the first blood count was performed relative to the first radiation exposure occurrence. The statistical analysis indicated that the first blood count performed during the year of the first external gamma exposure is a reasonable baseline. Although this first blood count may underestimate the actual starting point of the lymphocyte counts due to the beginning of radiation exposure, this underestimation applies to the whole study cohort. Thus the adjustment for the baseline lymphocyte count is consistent in all workers and it appears to be justified for the ensuing statistical analysis. It is worthwhile to mention that the distribution of the baseline lymphocyte counts is similar in males and females and therefore it can be considered that both sexes have a common starting point in terms of the lymphocyte counts.

Another control variable that has a major potential impact on this analysis is the exposure to Plutonium. Since Plutonium exposure may affect the lymphocyte counts^{31,70-73}, it is absolutely necessary to adjust for it. The adjustment for Plutonium exposure is a challenging issue since a large amount of Plutonium dose information is missing. Therefore, this adjustment has been performed indirectly by generating a new variable named “work location related to Plutonium exposure.” This variable is created by recoding the work location variable through steps described in detail in the methodology section. Briefly, the goal of the recoding procedure is to create groups of work locations that fall within distinct categories of Plutonium exposure. The

results of the statistical analysis show that work location is not a statistically significant predictor for the lymphocyte counts. However, it is still considered an important variable for which the analysis should be adjusted.

It is also worthwhile to recall that the work location variable consists of the plant at which the workers were employed for the longest time during a current year. Interestingly, the descriptive analysis of the dynamics of the work location shows that most of the workers do not change work sites or change locations at most one time during the follow up. Therefore, the work location variable is considered a time dependent variable with a slow dynamic.

In this statistical analysis it is absolutely required to adjust for smoking history since there are published studies that illustrate that the lymphocyte counts increase in smokers⁷⁴. The adjustment for alcohol consumption is also highly recommended since there are studies that show the suppressive effect of alcohol on the lymphocytes⁷⁵. In the study cohort, most of the smokers and alcohol consumers are among males. Most of the females do not have a smoking history and are not recorded as alcohol consumers at start of employment.

A variable of special interest is the temporal one which consists of the number of years from the first external gamma exposure occurrence. The temporal variable represents a main starting point in setting up a longitudinal data analysis. Two important considerations relate to this temporal variable. First, it is part of the longitudinal data model specification since the repeated measurements must refer to a time variable. Second, when the temporal variable is included in the GEE models as a predictor, it might introduce a time effect on the outcome variable, and this time effect has to be tested. This test has been performed using Model#6, in which it can be noted that the number of years from the first external gamma exposure is not a statistically significant predictor for the lymphocyte counts and therefore leads to the conclusion

that there is not a significant difference between the lymphocyte counts recorded at the beginning of follow-up compared to the counts recorded at the end of follow-up. This result can be explained by the negative relationship between the log-transformed yearly median lymphocyte counts and log-transformed yearly cumulative external gamma dose. According to the results presented in the descriptive analysis, the yearly cumulative external gamma doses show a short initial increase followed by a subsequent decrease. This suggests that during the follow up period the increase in the number of years since the first external gamma exposure does not equate with a constant increase of external gamma dose. On the contrary, as the number of years from the first exposure increases, the radiation dose decreases and allows the recovery of the lymphocyte counts. This observation regarding the yearly external gamma dose is associated with an inverse trend of the lymphocyte counts. The yearly median lymphocyte counts show an initial decrease followed by a subsequent increase and a tendency to rebound to the initial values. Therefore, at the end of follow-up the lymphocyte counts are at about the same level as they were at the beginning of the follow-up.

In the existing literature there is an important debate regarding the most appropriate theoretical model that illustrates the radiation dose-response relationship^{6,15,17,18,20,24,26,76-85}. The most relevant radiation dose-response models are presented in the introduction of this dissertation. The main issue regarding the radiation dose-response relationship is the debate between the linear non-threshold model and the non-linear threshold model. Although the linear non-threshold model has been adopted as the base for occupational regulation regarding permissible radiation doses, the non-linear threshold model has been discussed and applied to different situations^{7,8,11,13,15,17,20,24,26,77-79,83,84,86-89}.

This dissertation study contributes significantly and uniquely to this debate, as it performs the comparison of both linear and non-linear approaches when applied to the study cohort. The statistical procedures in this dissertation consist of the analysis of six radiation dose-response models. Four models are constructed and analyzed assuming a linear radiation dose-response relationship while two models are constructed and analyzed assuming a non-linear radiation dose-response relationship. The inverse relationship between log-transformed yearly median lymphocyte counts and log-transformed yearly cumulative external gamma dose is consistent in all analyzed models. The coefficients which correspond to the same variable are similar across all models, thereby showing similarities among models. The predicted values calculated for the lymphocyte counts in males and females are similar in all models. Since the similarity of predicted values and the similarity of coefficients are observed in models assuming linear as well as non-linear radiation dose-response relationships, it can be concluded that the linear radiation dose-response relationship provides a good fit for the study cohort data. The linear radiation dose-response model is currently applied in the implementation of regulations regarding acceptable occupational radiation exposure levels. This dissertation analysis is unique since it tests the applicability of the theoretical radiation dose-response model on the study cohort data. It is the first study that contrasts radiation dose-response models of lymphocyte counts in an occupational setting. The models are tested on a large number of workers including a relatively large number of female workers who were exposed to radiation doses five times higher than currently accepted.

The study cohort consists of a large number of males and females and separate models were fitted in males and females thereby being the first study that assesses and compares occupational radiation effects by sex. The analysis performed separately in males and females

shows the inverse dose-response relationship between the radiation exposure and the lymphocyte counts in both sexes.

Interestingly, some differences between males and females are suggested by models which assume a non-linear radiation dose-response model. It is important to recall that the GEE models used in this statistical analysis assume equality between the expected value of the outcome variable and the linear combination of the parameters. In other words, a linear relationship between the expected value of the log-transformed lymphocyte counts and the log-transformed yearly cumulative external gamma dose is assumed, while adjusting for lifestyle variables and work locations.

Since the non-linear radiation dose-response model has been discussed in the literature, it was necessary to find a statistical tool to deal with the nonlinearity issue within the framework of GEE techniques which assume linearity. This study is believed to be the first to implement a theoretical concept named linear splines in an occupational study of radiation exposure. The linear splines are aimed at addressing the potential non-linearity between log-transformed lymphocyte counts and yearly cumulative external gamma dose within GEE models applied to the study cohort data. An important issue of linear spline implementation consists of the location of the so-called knots. The dissertation analysis implements the simultaneous application of both theoretical and data derived criteria. The data derived criteria refer to the residuals distribution. The knot is located when the residuals distribution changes. The theoretical criteria implemented by this dissertation refer to the occupational radiation exposure upper limit currently accepted in the US which equals 5 rads. The reason for implementing the 5 rads knot location is to investigate if the radiation dose-response is different for radiation doses below 5 rads compared

with doses higher than 5 rads. In other words, this knot location helps to investigate if 5 rads can be considered a threshold for the occupational radiation dose-response relationship.

The analysis shows that in males and in females there is definitely a significant inverse relationship between the log-transformed lymphocyte counts and the log-transformed yearly cumulative external gamma dose. Interestingly, the statistical analysis suggests some differences between males and females at doses lower than 5 rads. In males, there is a slightly stimulation of the lymphocyte counts at doses below 5 rads. However, this stimulative effect is statistically non-significant (Model #5, p-value=0.09). In females, according to the same model the stimulation effect of the lymphocyte counts at radiation doses below 5 rads is replaced by an inhibition. In females the inhibitive effect of occupational radiation exposure on the lymphocyte counts is statistically non-significant (Model#5, p-value=0.16). Thus, these models applied to the study cohort suggest the occurrence in males of the adaptive response described by Russian authors as an increase of the lymphocyte counts at low doses²⁶. However, due to the lack of statistical significance, there is no strong statistical support in the present dissertation for concluding that there is an adaptive response.

The scientists in Western countries define the adaptive response as an adaptation phenomenon consisting of a decreased sensitivity to acute radiation exposure that follows a low-dose radiation exposure. According to the models fitted and analyzed in this dissertation one can not make any statement about the adaptive response as defined by scientists in the West since there is no acute high radiation exposure following a low-dose radiation exposure. The increase in the lymphocyte counts noticed in males according to the models fitted in this dissertation at doses lower than 5 rads corresponds to the definition of hormesis. However, the lack of statistical

significance does not allow any strong statement about a hormetic radiation dose-response model in males.

The statistical issues involved by the use of GEE models require a special attention since this is the first analysis on a large occupational database using this sophisticated analytical methodology. A main concern when fitting marginal models using the GEE technique is that the goodness of fit assessment is not standardized, and the computational tools are not provided by most software packages. This problem arises due to the multiple measurements recorded in each individual, thereby inducing correlation among measurements that belong to the same individual. In this dissertation a set of models' goodness of fit assessment tools are implemented according to guidelines specified in critical reference sources^{42,47-51,54-56,64,90-96}. Thus, the model comparison using these statistical tools results in similarities among the six models in terms of the goodness of fit. The similarities between the models assuming linear dose-response relationship and the models assuming non-linear dose-response relationship are very important since they show that assuming non-linearity and implementing the linear splines do not affect the goodness of fit. These results also hold for the estimated parameter coefficients which are similar in models assuming a linear radiation dose-response model and in models assuming non-linearity by using the linear splines. Thus, the most parsimonious models are selected as optimal models and they are the models fitted under the assumption of a linear radiation dose-response relationship.

The four models assuming a linear radiation dose-response response show a statistically significant inverse relationship between the log-transformed yearly median lymphocyte counts and the log-transformed cumulative external gamma dose while adjusting for sex, baseline counts, smoking and alcohol consumption, work location and work location related to Plutonium exposure. These models fitted for the overall study cohort and including sex as a covariate show

that sex is not a statistically significant covariate ($p > 0.6$). However, although sex is not a statistically significant covariate, there are differences between males and females indicated by the descriptive analysis that suggest the possibility that sex is an effect modifier.

In this dissertation two approaches are used in order to analyze the possibility of differences between males and females with regard to their sensitivity to occupational radiation exposure.

The first approach is applied to the model assuming a linear radiation dose-response relationship. It consists of fitting an interaction term between sex and log-transformed yearly cumulative external gamma dose (Model 2). The interaction term is fitted in addition to the main effects variables which consist of the log-transformed external gamma dose, sex and the variables for which the analysis controls: the baseline lymphocyte count, work location related to Plutonium exposure, smoking history and alcohol consumption at start of employment. The analysis of Model 2 shows that the interaction term between log-transformed external gamma dose and sex is borderline statistically non-significant ($p\text{-value} = 0.08$). Therefore, due to the borderline p -value that corresponds to the interaction term and due to the differences by sex suggested by the descriptive analysis, it is considered appropriate to fit the models in males and females separately.

The second approach was applied to all the models analyzed in the study. It consists of the calculation of the predicted value of the lymphocyte count for each model in males and females separately using the same values for the covariates in both sexes. These analyses lead to the finding that the predicted values of the lymphocyte counts are similar in males and females in each model.

The application of the two approaches of testing the differences between males and females regarding their sensitivity to occupational radiation exposure identifies a common result which is that there is a similar response to occupational radiation exposure observed in males and females.

However, the statistical analysis of the models employing the linear splines indicates that there may be differences between males and females regarding occupational radiation sensitivity at external gamma doses below 5 rads. The differences suggested by the coefficients of the linear splines are not statistically significant. In spite of the differences suggested by the models including linear splines, the predicted values of the lymphocyte counts are similar in males and females in each of the six models, irrespective of the linearity assumptions about radiation dose-response relationship. Thus, the differences between males and females observed earlier are not statistically confirmed.

Another important issue is the choice of the statistical model that fits best the analyzed data. Since there is not enough evidence to reject the linearity assumption, the optimal model for this data has to be chosen from the models fitted according to a linear radiation dose-response relationship in an occupational setting. The main interest in this study is to analyze the effect of occupational external gamma radiation exposure on the lymphocyte counts while controlling for some important covariates which consist of sex, baseline count, work location related to Plutonium exposure, smoking history, and alcohol consumption at start of employment. In correspondence with this interest, the optimal model is considered to be Model#1 which includes all these variables. Model#1 is preferred to the models including the linear splines, since it is more parsimonious, the goodness of fit is similar to that of the other models, and the findings are consistent with those derived from the other, more complex models.

4.1 STRENGTHS OF THE ANALYSIS

One of the strengths of this project refers to the use of a subset of the unique Mayak PA database which includes a large number of males and females exposed to long-term occupational radiation exposure. The selection of the subset of data named the study cohort used in this analysis was possible due to the large quantity and good quality of data available on workers of both sexes who were employed at Mayak PA at the opening of this nuclear facility. The analysis encompasses the first ten years of operation at Mayak PA when the male and female workers were exposed to the highest occupational radiation doses. Due to the existence of records containing information on the blood tests in these workers who were employed during the early years of Mayak PA operation, the statistical analysis is feasible and informative.

Another strong point of this analysis consists of testing the theoretical radiation dose-response relationship on a subset of Mayak PA workers data. This challenging task is accomplished by building sophisticated statistical models which correspond to the most widely accepted radiation dose-response models. Thus, this dissertation approaches the complicated issue of building GEE statistical models for the analysis of longitudinal data with unbalanced number of observations recorded at each point in time. Furthermore, the assessment of goodness of fit is performed in these complex models using a set of tools that are developed, implemented and tested as part of this dissertation research. It is worthwhile to mention that the assessment of goodness of fit in GEE model is not provided by any statistical software as a standard procedure.

Another important strength of this dissertation project is the approach to the potential non-linearity suggested by the descriptive analysis of the radiation dose-response relationship in Mayak PA workers. The linear splines, applied to the GEE models as an optimization technique allows formal statistical tests to determine whether there is a statistically significant threshold at

two specific low doses of radiation which define the knot locations. Usually the knot locations are values of the covariate of interest that are specified according to the residuals values, and therefore the knot locations are data derived. In this study, there are two knot locations. One of them corresponds to the change of pattern in the residuals and it is data derived. However, the second knot corresponds to the value of radiation exposure established as the upper occupational limit currently accepted in the US which is 5 rads and therefore is based on a theoretical value and not on the data analyzed. Thus, the analysis implements linear splines defined both according to data derived and theoretical criteria.

The approach to missing data mechanisms assessment represents another strong point of this dissertation. Although statistical tests for longitudinal missing data mechanisms have been theoretically developed and published^{42,44,61,66}, the computational tools have not been implemented in the most popular software packages. Since the assumptions of the estimation technique used in this analysis require that data are missing completely at random (MCAR,) it is considered an important contribution to the programming and implementation of tools for testing missing data; these tools are derived in accordance with theoretical statistical tests^{12,44,60,66-68}. The results of statistical testing for the mechanisms of missing data indicate that missing data can be considered completely at random, and therefore that the GEE techniques utilized are appropriate.

4.2 LIMITATION OF THE ANALYSIS

Some of the limitations of this study derive from the constraints of using an existing database that was not designed specifically for this study. In addition to this general limitation, the data

collected on Mayak PA workers have a long history and it is likely that the many changes that occurred during the timeframe of interest in the broader scientific, social and political environment are also reflected in the data collection processes.

One important limitation of this study is represented by the lymphocyte baseline counts. As previously discussed, for purposes of optimal use of the existing data, this study implemented a special procedure for establishing the baseline lymphocyte counts. This procedure has been discussed in detail in this dissertation. However, it is important to note that this procedure, although fully justified from a statistical point of view, cannot preclude the possibility that some of the lymphocyte baseline counts are affected by earlier exposure to radiation.

Another limitation of this dissertation relates to the data regarding external gamma dosimetry. In all the occupational studies reviewed for this project, the radiation dosimetry issue raises critical challenges. Radiation doses are usually reconstructed from the film badges worn by workers. The technologies used for the manufacture of film badges, and the radiation dose reconstruction methodologies have evolved over time. This study is focused on workers hired between 1948 and 1960, and the procedures used for the reconstruction of the doses bears the historical limitations of the respective period. Despite the problems related to dosimetry, the Mayak PA workers database is a unique database which provides information about radiation doses in male and female workers on a consistent basis and encompasses many years of relatively high doses of occupational radiation exposure.

The Plutonium exposure data is recorded in selected workers only, thereby making it impossible to use the Plutonium dose as a covariate itself. The methodological difficulties regarding the adjustment for Plutonium exposure represent an important limitation of this study, since the Plutonium exposure is considered to affect the lymphocyte counts^{31,70-73}. Therefore, in

order to minimize this limitation, the adjustment for Plutonium exposure has been performed through a variable named work location related to Plutonium exposure which approximates the actual exposure data.

Another limitation consists of the distribution of smoking and alcohol consumption variables by sex (Tables 26-27). There is a reduced number of females smokers and alcohol consumers at baseline thereby, involving a reduced number of subjects in these categories after stratification by sex. This situation could lead to unstable models. However, smoking and alcohol consumption were used only as control variables. Moreover, when statistical models were fitted without smoking and alcohol consumption, the coefficients for the explanatory variables were similar.

Finally, an important limitation of this analysis relates to the recovery of the lymphocyte counts following the decrease in radiation exposure during the latter years of follow-up. Since the recovery of the lymphocyte counts is a complex phenomenon, special modeling techniques are required⁹⁷. The issue of modeling the lymphocyte counts recovery is not specifically addressed in this dissertation. In order to address this issue, one could employ pharmacodynamic models⁹⁷, but such modeling is beyond the scope of the present dissertation. However, this study provides an important scientific basis for future work on the dynamics of the recovery of lymphocyte counts following different patterns of exposure.

5.0 CONCLUSIONS

The results and the discussion presented in this dissertation lead to the following conclusions:

1) Occupational external gamma radiation exposure is associated with a statistically significant decrease in lymphocyte counts after adjusting for baseline lymphocyte counts, sex, work location related to Plutonium exposure, smoking history and alcohol consumption at start of employment. This conclusion applies to the overall study cohort as well as to the males and females separately.

2) A linear dose-response relationship between the lymphocyte counts and external gamma dose best describes the overall study cohort data as well as males and females separately.

3) Although some differential sensitivity is observed in occupational radiation response between males and females, the differences observed do not achieve statistical significance.

4) The issue of differential sensitivity between males and females merits further investigation. It is worthwhile to consider whether the recovery process of the lymphocyte counts differs in males and females following the decrease in occupational radiation exposure. The study of lymphocyte counts recovery requires further work using special modeling techniques.

5) Most of the yearly median lymphocyte counts do not drop below the normal range and cannot be considered an abnormal value of the blood test or an expression of disease. This result does not signify that long term exposure to radiation does not affect the health of the workers

included in the study cohort, since the drop of the lymphocyte counts may contribute to long-term health effects.

5.1 PUBLIC HEALTH SIGNIFICANCE

This research is important for the public health community as it has implications regarding the current regulations for occupational radiation exposure. Specifically, the findings imply that: 1) Based on the analysis of a subset of Mayak PA workers, it is reasonable to use the linear radiation dose-response model for regulatory purposes, and 2) Since differential sensitivity to occupational radiation exposure between males and females is not confirmed by these analyses, there is no strong rationale for different regulatory standards for males and females.

APPENDIX A - Radiation physics

The stability of an atom depends on the forces of its nuclear components. An unstable atom tends to become stable by releasing energy in different ways, often by emission of ionizing radiation. The ionizing radiation represents a small part of the electromagnetic spectrum which includes radio waves, radar, microwaves and ultraviolet radiation. Ionization consists of the ejection of an electron from the atom due to the transfer of energy by radiation which overcomes the binding energy of the electron. The ionization may be direct or indirect according to the method by which the radiation interacts with the medium.

Alpha and beta particles are considered directly ionizing radiation. They react with target molecules as oxygen and water striking the tissue or medium directly.

Electromagnetic radiations such as x-rays and gamma are indirectly ionizing: they release energy as a result of various interactions. The energy is used to produce a fast-moving charged particle such as an electron. The electron may secondarily react with a target molecule^{98,99}. Indirectly ionizing radiation are considered more penetrating than directly ionizing radiation^{98,99}. Penetration refers to the amount of radiation which reaches a certain depths in the tissue.

Gamma radiation, which is the exposure of interest in this study, is defined as ionizing low LET electromagnetic radiation. *LET* refers to the linear energy transfer, or the amount of energy deposited in a unit of track length. LET (linear energy transfer) is expressed in KeV/ μ (kilo electron volt/micron). LET (linear energy transfer) is an important concept since it is positively correlated with the biologic effectiveness of radiation. However due to technical problems related to different radiation qualities used in experiments, LET (linear energy transfer)

can not be used to assess the biological effectiveness of radiation. Therefore, a more general term RBE (relative biological effectiveness) of a given radiation dose is used in order to assess the biological effectiveness of radiation.

RBE (relative biological effectiveness) is related to the LET (linear energy transfer) (Table1). RBE (relative biological effectiveness) is calculated by comparing the biologic effectiveness of a given type of experimental radiation against to 250 kilovolt X-rays. The biologic effectiveness of ionizing radiation is due to the localized deposition of energy which may affect important structures such as the genetic material (e.g. DNA). Most gamma and x-rays have relative biological effectiveness, RBE~1. Table1 illustrates ranges of RBE values. It is important to mention that ranges are presented since LET (linear energy transfer) and RBE (relative biological effectiveness) may vary from tissue to tissue. Moreover LET (linear energy transfer) and RBE (relative biological effectiveness) may be different if early and late radiation effects are compared.

Table 60. Approximate LET* and RBE** for Different Types of Ionizing Radiations

Type of Radiation	LET (keV/ μ)	RBE
Gamma and X-rays	0.3-10	1
Beta radiation	0.5-15	1-2
Neutrons	20-50	2-5
Alpha radiation	80-250	5-10

Source: Mettler FA, Moseley RD. Medical effects of ionizing radiation. Orlando, FL: Grune & Stratton, 1985.

*linear energy transfer **relative biological effectiveness

According to Table53, the gamma rays are characterized by lower LET (linear energy transfer) and RBE (relative biological effectiveness) compared to other types of radiation.

It is important to mention that the total amount of energy delivered in a lethal dose of radiation is extremely small but effectively utilized. For instance, a total body external gamma dose of 7 Gray=700rad corresponds to an absorption of only 1 cal in a 70 kilogram man which corresponds to a temperature increase of less than $0.002^{\circ}\text{C}^{98}$.

The unit of measurement for radiation exposure is the Gray (Gy), $1 \text{ Gy} = 1 \text{ J/kg}$ and *Rad* measures the absorbed energy. *Rem* (roentgen equivalent man) is used to assess the biological response and to compare radiation effects. For specific radiation categories the relationship between rem and rad is the following: **$1 \text{ rem} = \text{QF} \times 1 \text{ rad}$** . QF=1 for x-rays and gamma rays, meaning that the relative biological effectiveness of X and gamma radiation is the same.

Thus, LET (linear energy transfer) analysis via QF (quality factor) shows that for gamma and x-ray exposures, the rad and rem as measurement units are interchangeable. Other important relationships between radiation units are the following: $1 \text{ Sv} = 100 \text{ rem}$ and $1 \text{ Gy} = 100 \text{ rads}$ ^{98,99}.

The permissible radiation levels have changed over time from high levels which were not safe for the workers to lower levels considered likely to be safe. For example in Russia, in 1946, permissible exposures were high, 0.2 rem/day or 60 rem/year. Two years later, they were lowered to 0.1 rem/day or 30 rem/year. Since 1962 permissible exposure levels have been close for United States and Russia¹⁰⁰. For many years, in the United States the acceptable occupational exposures have been set below 5rem/year. An important issue is the high rate radiation exposures during radiation accidents. Although some permissible levels for the radiation accidents have been mentioned in a Russian report about radiation sickness, as high as 25 rem in 15 minutes or 100 rem/year¹⁰⁰, it is difficult to do such an assessment because accidental exposure doses are beyond control and an accident itself is unpredictable and to be avoided. A report about nuclear criticality accidents involving uncontrolled nuclear fission describes seven such accidents at Mayak PA, four of them involving plutonium processing and three of them uranium processing. All of the accidents resulted in significant exposures, but four involved fatalities. These accident records are very valuable pieces of information because they illustrate a case of a real *in vivo*

challenging dose applied to humans. Some of them can be retrieved in the Mayak PA workers database^{69,101,102}.

APPENDIX B - MWECE database format

Table 61: MWECE Database Format - Lifestyle Variables Definition and Coding

Variable Name	Variable definition	Variable Code
Tobacco_Use	Did the subject ever use any kind of tobacco?	-1 = unknown 0 = no 1 = yes, but quit 2 = yes, never quit
Yr-Start_Tob	Year started using tobacco.	-1 = unknown -2 = not applicable (never smoked) 19xx = date
Yr_End_Tob	Year quit using tobacco.	-1 = unknown -2 = not applicable (never smoked or never quit) 19xx = date
Base_CPD	Cigarettes per day smoked at start of work.	-1 = unknown -2 = not applicable (never smoked) -3 = used tobacco, but did not smoke cigarettes xx = number cigarettes per day
Base_Alc	Overall pattern of alcohol use at start of work.	-1 = unknown 0 = does not drink alcohol 1 = moderate use 2= domestic alcohol abuse 3 = alcoholism
48 - 54_Alc	Overall pattern of alcohol use from 1948 -1954.	-1 = unknown 0 = does not drink alcohol 1 = moderate use 2= domestic alcohol abuse 3 = alcoholism
55-end_Alc	Overall pattern of alcohol use from 1955 until present or end of available data.	-1 = unknown 0 = does not drink alcohol 1 = moderate use 2= domestic alcohol abuse 3 = alcoholism

Table 62: MWECE Database Format - Hematological Variables Definition and Coding

Variable Name	Variable Code
Identification number (Clinic_Id)	
Radiation sickness diagnosis (Primary_Dx)	000 = uninjured worker 001 = PPn only 011 = PPn and ARS 101 = PPn and CRS 010 = ARS only 100 = CRS only 110 = ARS and CRS 111 = PPn, ARS, and CRS
Radiation related diagnosis (Secondary_Dx)	00 = no secondary diagnoses 10 = radiation cataract only 01 = other radiation related condition 11 = cataract and other condition
Laboratory test day	-1 = Missing -2 = Not Applicable 1-31 = day
Laboratory test month	-1 = Missing -2 = not Applicable 1-12 = month
Laboratory test year	-1 = Missing -2 = Not Applicable 00-xx = year
Erythrocytes (10 ¹² /l)	-1 = Missing Erythrocytes = xx.xx
Hemoglobin (g/l)	-1 = Missing Hemoglobin = xxxx.x
Reticulocytes (%)	-1 = Missing Reticulocytes = xx.x
Thrombocytes (10 ⁹ /l)	-1 = Missing Thrombocytes = xxxx.xx
Leukocytes (10 ⁹ /l)	-1 = Missing Leukocytes = xxx.xxx
Basophils (%)	-1 = Missing Basophils = xx.xx
Bands (%)	-1 = Missing Bands = xx.xx
Polymorphonuclear Leukocytes (%)	-1 = Missing Poly Leuk = xxxx.xx
Lymphocytes (%)	-1 = Missing Lymphocytes = xxx.xx
Monocytes (%)	-1 = Missing Monocytes = xxx.xx
Plasma Cells (%)	-1 = Missing Plasma Cells = xx.x

Table 63: MWECE Database Format – Work Location Variables Definition and Coding

Variable Name	Variable definition	Variable Code
Day_Start_Work	Day for date started work	-1 = Missing 1-31 = day
Month_Start_Work	Month for date started work	-1 = Missing 1-12 = month
Year_Start_Work	Year for date started work	-1 = Missing 00-xx = year
Day_End_Work	Day for date stopped work	-1 = Missing 1-31 = day
Month_End_Work	Month for date stopped work	-1 = Missing 1-12 = month
Year_End_Work	Year for date stopped work	-1 = Missing 00-xx = year
Emplcode	Employee Code	1 = Plant A 2 = Plant B 3 = Plant C 4 = Plant D 8 = not employed in a radiation exposed area in plant A,B,C or D or at any other site 9 = Unknown

Number of obs = 400 R-squared = 0.0006
 Root MSE = .226771 Adj R-squared = -0.0019

Source	Partial SS	df	MS	F	Prob > F
Model	.012200143	1	.012200143	0.24	0.6265
pat_ly_dose	.012200143	1	.012200143	0.24	0.6265
Residual	20.4672075	398	.051425144		
Total	20.4794076	399	.051326836		

Mixed patterns missing data assessment

Testing if ln(yearly cumulative gamma dose) depends on ln(yearly median lymphocyte counts)missingness

. anova lndose pat_ly_dose if pat_ly_dose!=1

Number of obs = 301 R-squared = 0.0014
 Root MSE = 1.74681 Adj R-squared = -0.0019

Source	Partial SS	df	MS	F	Prob > F
Model	1.28634082	1	1.28634082	0.42	0.5167
pat_ly_dose	1.28634082	1	1.28634082	0.42	0.5167
Residual	912.347606	299	3.05132978		
Total	913.633946	300	3.04544649		

Mixed patterns missing data assessment -Males

Testing if ln(yearly cumulative gamma dose) depends on ln(yearly median lymphocyte counts) missingness

. anova lnmedly pat_ly_dose if pat_ly_dose!=2

Number of obs = 253 R-squared = 0.0005
 Root MSE = .234342 Adj R-squared = -0.0035

Source	Partial SS	df	MS	F	Prob > F
Model	.006855233	1	.006855233	0.12	0.7241
pat_ly_dose	.006855233	1	.006855233	0.12	0.7241
Residual	13.7839249	251	.054916035		
Total	13.7907801	252	.054725318		

. anova lndose pat_ly_dose if pat_ly_dose!=1

Number of obs = 200 R-squared = 0.0000
 Root MSE = 1.73259 Adj R-squared = 0.0000

Source	Partial SS	df	MS	F	Prob > F
--------	------------	----	----	---	----------

Model		0	0	
pat_ly_dose		0	0	
Residual		597.374417	199	3.00188149

Total		597.374417	199	3.00188149

Mixed patterns missing data assessment - Females

Testing if ln(yearly cumulative gamma dose) depends on ln(yearly median lymphocyte counts)missingness

. anova lnmedly pat_ly_dose if pat_ly_dose!=2

Number of obs = 147 R-squared = 0.0044
 Root MSE = .210604 Adj R-squared = -0.0024

Source	Partial SS	df	MS	F	Prob > F
Model	.028709964	1	.028709964	0.65	0.4224
pat_ly_dose	.028709964	1	.028709964	0.65	0.4224
Residual	6.43134792	145	.044354124		

Total	6.46005789	146	.044246972		

. anova lndose pat_ly_dose if pat_ly_dose!=1

Number of obs = 101 R-squared = 0.0021
 Root MSE = 1.76163 Adj R-squared = -0.0080

Source	Partial SS	df	MS	F	Prob > F
Model	.650224708	1	.650224708	0.21	0.6481
pat_ly_dose	.650224708	1	.650224708	0.21	0.6481
Residual	307.22928	99	3.10332606		

Total	307.879505	100	3.07879505		

APPENDIX D - Lymphocyte counts range in humans

Categories	Lymphocyte counts range*
Severe Lymphocytopenia	<500/mm ³
Mild Lymphocytopenia	501-1000/mm ³
Moderate Lymphocytopenia	1001-1500/mm ³
Normal Lymphocyte Counts	1501-4000/mm ³
Lymphocytosis	>4000/mm ³

*according to Wintrobe-Wald criteria

BIBLIOGRAPHY

1. Gilbert ES. Invited Commentary: Studies of Workers Exposed to Low Doses of Radiation. *Am. J. Epidemiol.* 2001;153(4):319-322.
2. Yoshinaga S, Mabuchi K, Sigurdson AJ, Doody MM, Ron E. Cancer Risks among Radiologists and Radiologic Technologists: Review of Epidemiologic Studies. *Radiology* 2004;233(2):313-321.
3. BEIR VII Phase 2. 2006.
4. Alberts B. *Molecular biology of the cell.* 3rd ed. New York: Garland Pub., 1994.
5. Lee GR, Wintrobe MM. *Wintrobe's clinical hematology.* 9th ed. Philadelphia: Lea & Febiger, 1993.
6. Seong J, Suh CO, Kim GE. Adaptive response to ionizing radiation induced by low doses of gamma rays in human cell lines. *Int J Radiat Oncol Biol Phys* 1995;33(4):869-74.
7. Wolff S, Afzal V, Wiencke JK, Olivieri G, Michaeli A. Human lymphocytes exposed to low doses of ionizing radiations become refractory to high doses of radiation as well as to chemical mutagens that induce double-strand breaks in DNA. *Int J Radiat Biol Relat Stud Phys Chem Med* 1988;53(1):39-47.
8. Sorensen KJ, Attix CM, Christian AT, Wyrobek AJ, Tucker JD. Adaptive response induction and variation in human lymphoblastoid cell lines. *Mutation Research* 2002;519(1-2):15-24.
9. Thierens H, Vral A, Barbe M, Meijlaers M, Baeyens A, Ridder LD. Chromosomal radiosensitivity study of temporary nuclear workers and the support of the adaptive response induced by occupational exposure. *International Journal of Radiation Biology* 2002;78(12):1117-26.
10. Wolff S, Jostes R, Cross FT, Hui TE, Afzal V, Wiencke JK. Adaptive response of human lymphocytes for the repair of radon-induced chromosomal damage. *Mutation Research* 1991;250(1-2):299-306.
11. Mortazavi SMJ. *An Introduction to Radiation Hormesis.*

12. Sinclair WK. The linear no-threshold response: why not linearity?[erratum appears in Med Phys 1998 May;25(5):794]. Medical Physics 1998;25(3):285-90; discussion 300.
13. Linear Vs Non-linear dose response. Vol. 2006. Waterloo, Ontario: University of Waterloo Radiation Safety Training.
14. Rigaud O. The adaptive response to ionizing radiation: low dose effects unpredictable from high dose experiments. Human & Experimental Toxicology 18, no 7: 443 1999.
15. Bonner WM. Low-dose radiation: Thresholds, bystander effects, and adaptive responses. PNAS 2003;100(9):4973-4975.
16. Ballarini F, Ottolenghi A. Low-dose radiation action: possible implications of bystander effects and adaptive response. Journal of Radiological Protection 2002;22(3A):A39-42.
17. Wolff S. The adaptive response in radiobiology: evolving insights and implications. Environ Health Perspect 1998;106 Suppl 1:277-83.
18. Ikushima T. Radioadaptive response: responses to the five questions. Human & Experimental Toxicology 1999;18(7):433-435.
19. Cai L. Research of the adaptive response induced by low-dose radiation: where have we been and where should we go? Human & Experimental Toxicology 18, no 7: 419 1999.
20. Pettersen EO. Low-dose hypersensitivity and adaptive responses to radiation. University of Oslo, 2002.
21. Adelstein SJ. Biologic responses to low doses of ionizing radiation: adaptive response versus bystander effect. J Nucl Med 2003;44(1):125; author reply 125-6.
22. Kadhim MA, Moore SR, Goodwin EH. Interrelationships amongst radiation-induced genomic instability, bystander effects, and the adaptive response. Mutat Res 2004;568(1):21-32.
23. Mosse I, Kostrova L, Subbot S, Maksimenya I, Molophei V. Melanin decreases clastogenic effects of ionizing radiation in human and mouse somatic cells and modifies the radioadaptive response. Radiation & Environmental Biophysics 2000;39(1):47-52.
24. Olivieri G. Adaptive response and its relationship to hormesis and low dose cancer risk estimation. Human & Experimental Toxicology 18, no 7 (1999): 440.
25. Gourabi H, Mozdarani H. A cytokinesis-blocked micronucleus study of the radioadaptive response of lymphocytes of individuals occupationally exposed to chronic doses of radiation. Mutagenesis 1998;13(5):475-480.

26. Akleev AV, Aleshchenko AV, Gotlib V, Kudriashova OV, Semenova LP, Serebrianyi AM, Khudiakova OI, Pelevina, II. [Adaptive response of blood lymphocytes of the inhabitants of the South Ural chronically exposed to radiation]. *Radiats Biol Radioecol* 2004;44(4):426-31.
27. Okladnikova ND, Pesternikova VS, Azizova TV, Sumina MV, Kabasheva N, Belyaeva ZD, Fevralev AM. Sostoianie zdorov'ia personala zavoda po pererabotke otrabotavshogo iadernogo topliva. *Medicsina Truda i Promyshlennaia Ekologiya* 2000(6):10-4.
28. Manual on Radiation Haematology. Vienna: International Atomic Energy Agency, 1971.
29. Gajendiran N, Tanaka K, Kumaravel TS, Kamada N. Neutron-induced adaptive response studied in go human lymphocytes using the comet assay. *Journal of Radiation Research* 2001;42(1):91-101.
30. Guskova A K, Boysogolov G D. Radiation sickness in man. 1971.
31. Claycamp HG, Okladnikova ND, Azizova TV, Belyaeva ZD, Boecker BB, Pesternikova VS, Scott BR, Shekhter-Levin S, Sumina MV, Sussman NB, Teplyakov, II, Wald N. Deterministic effects from occupational radiation exposures in a cohort of Mayak PA workers: data base description. *Health Phys* 2000;79(1):48-54.
32. Berrington A, Darby SC, Weiss HA, Doll R. 100 years of observation on British radiologists: mortality from cancer and other causes 1897-1997. *Br J Radiol* 2001;74(882):507-519.
33. Sponsler R, Cameron JR. Nuclear Shipyard Worker Study (1980-1988): A Large Cohort Exposed to Low Dose Rate Gamma Radiation.
34. Evans HJ, Buckton KE, Hamilton GE, Carothers A. Radiation-induced chromosome aberrations in nuclear-dockyard workers. *Nature* 1979;277(5697):531-4.
35. Sont WN, Zielinski JM, Ashmore JP, Jiang H, Krewski D, Fair ME, Band PR, Letourneau EG. First Analysis of Cancer Incidence and Occupational Radiation Exposure Based on the National Dose Registry of Canada. *Am. J. Epidemiol.* 2001;153(4):309-318.
36. Mohan AK, Hauptmann M, Linet MS, Ron E, Lubin JH, Freedman DM, Alexander BH, Boice JD, Jr., Doody MM, Matanoski GM. Breast Cancer Mortality Among Female Radiologic Technologists in the United States. *J Natl Cancer Inst* 2002;94(12):943-948.
37. Matanoski GM, Seltser R, Sartwell PE, Diamond EL, Elliott EA. The current mortality rates of radiologists and other physician specialists: deaths from all causes and from cancer. *Am. J. Epidemiol.* 1975;101(3):188-198.
38. Andersson M, Engholm G, Ennow K, Jessen KA, Storm HH. Cancer risk among staff at two radiotherapy departments in Denmark. *Br J Radiol* 1991;64(761):455-460.

39. Gilbert ES, Petersen GR, Buchanan JA. Mortality of Workers at the Hanford Site: 1945-1981.11-25.
40. Smith PG, Douglas AJ. Mortality of workers at the Sellafield plant of British Nuclear Fuels. *British Medical Journal Clinical Research Ed.* 1986;293(6551):845-54.
41. Sperati A, Abeni DD, Tagesson C, Forastiere F, Miceli M, Axelson O. Exposure to indoor background radiation and urinary concentrations of 8-hydroxydeoxyguanosine, a marker of oxidative DNA damage. *Environ Health Perspect* 1999;107(3):213-5.
42. Fitzmaurice GM, Laird NM, Ware JH. *Applied longitudinal analysis.* Wiley series in probability and statistics. Hoboken, N.J.: Wiley-Interscience, 2004.
43. Diggle P, Liang K-Y, Zeger SL. *Analysis of longitudinal data.* Oxford statistical science series; 13. Oxford New York: Clarendon Press; Oxford University Press, 1994.
44. Diggle PJ. Testing for Random Dropouts in Repeated Measurement Data. *Biometrics* 1989;45(4):1255-1258.
45. Hardin WJ, Hilbe MJ. *Generalized Estimating Equations.* Boca Raton: Chapman & Hall/CRC, 2003.
46. Rochon J. Analyzing Bivariate Repeated Measures for Discrete and Continuous Outcome Variables. *Biometrics* 1996;52(2):740-750.
47. Lloyd JE. Modern statistical techniques for the analysis of longitudinal data in biomedical research. *Pediatric Pulmonology* 2000;30(4):330-344.
48. Ballinger GA. Using Generalized Estimating Equations for Longitudinal Data Analysis. *Organizational Research Methods* 2004;7(2):127-150.
49. Pregibon D. Goodness of Link Tests for Generalized Linear Models. *Applied Statistics* 1980;29(1):15-24.
50. Horton NJ, Lipsitz SR. Review of Software to Fit Generalized Estimating Equation Regression Models. *American Statistician* 1999;53(2):160-169.
51. Paul Burton LGPS. Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in Medicine* 1998;17(11):1261-1291.
52. Liang K-Y, Zeger SL. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* 1986;73(1):13-22.

53. Twisk JWR. Longitudinal Data Analysis. A Comparison Between Generalized Estimating Equations and Random Coefficient Analysis. *European Journal of Epidemiology* 2004;19(8):769.
54. Pan W. Model selection in estimating equations. *Biometrics* 2001;57(2):529-34.
55. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2001;57(1):120-5.
56. Hedeker D, Gibbons RD. *Applied longitudinal data analysis*. Hoboken, N.J.: Wiley; Chichester: John Wiley [distributor], 2005.
57. Wegman EJ, Wright IW. Splines in Statistics. *Journal of the American Statistical Association* 1983;78(382):351-365.
58. Wold S. Spline Functions in Data Analysis. *Technometrics* 1974;16(1):1-11.
59. Wikipedia-Contributors. *Spline Interpolations*. 2006.
60. Little RJA, Rubin DB. *Statistical analysis with missing data*. Wiley series in probability and statistics. 2nd ed. New York: Wiley, 2002.
61. Rubin DB. Inference and Missing Data. *Biometrika* 1976;63(3):581-592.
62. Wang JX, Zhang LA, Li BX, Zhao YC, Wang ZQ, Zhang JY, Aoyama T. CANCER INCIDENCE AND RISK ESTIMATION AMONG MEDICAL X-RAY WORKERS IN CHINA, 1950-1995.455-466.
63. Yue-Cune C. Residuals analysis of the generalized linear models for longitudinal data. *Statistics in Medicine* 2000;19(10):1277-1293.
64. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986;42(1):121-30.
65. Smith PL. Splines as a Useful and Convenient Statistical Tool. *The American Statistician* 1979;33(2):57-63.
66. Ridout MS, Diggle PJ. Testing for Random Dropouts in Repeated Measurement Data. *Biometrics* 1991;47(4):1617-1621.
67. Little RJA. Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Association* 1995;90(431):1112-1121.
68. Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association* 1988;83(404):1198-1202.

69. Larin V. Mayak's walking wounded. *The Bulletin of the Atomic Scientists* 1999.
70. Koshurnikova NA, Shilnikova NS, Okatenko PV, Kreslov VV, Bolotnikova MG, Sokolnikov ME, Khokhriakov VF, Suslova KG, Vassilenko EK, Romanov SA. Characteristics of the cohort of workers at the Mayak nuclear complex. *Radiat Res* 1999;152(4):352-63.
71. Okladnikova ND, Pesternikova VS, Sumina MV, Doshchenko VN. Occupational diseases from radiation exposure at the first nuclear plant in the USSR. *Sci Total Environ* 1994;142(1-2):9-17.
72. Hande MP, Azizova TV, Burak LE, Khokhryakov VF, Geard CR, Brenner DJ. Complex chromosome aberrations persist in individuals many years after occupational exposure to densely ionizing radiation: An mFISH study. *Genes Chromosomes Cancer* 2005.
73. Voelz GL, Stevenson AP, Stewart CC. Does Plutonium Intake in Workers Affect Lymphocyte Function? *Radiat Prot Dosimetry* 1989;26(1-4):223-226.
74. Nakata A, Tanigawa T, Araki S, Sakurai S, Iso H. Lymphocyte Subpopulations Among Passive Smokers. *JAMA* 2004;291(14):1699-a-1700.
75. Glassman AB, Bennett CE, Randall CL. Effects of ethyl alcohol on human peripheral lymphocytes. *Arch Pathol Lab Med* 1985;109(6):540-2.
76. Ziegler A, Kastner C, Brunner D, Blettner M. Familial associations of lipid profiles: a generalized estimating equations approach. *Statistics in Medicine* 2000;19(24):3345-57.
77. Bosi A, Olivieri G. Variability of the adaptive response to ionizing radiations in humans. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 1989;211(1):13.
78. Calabrese EJ. RADIATION HORMESIS: ITS HISTORICAL FOUNDATIONS AS A BIOLOGICAL HYPOTHESIS. *BELLE Newsletter* 1999;8(2).
79. Feinendegen LE. The role of adaptive responses following exposure to ionizing radiation. *Human & Experimental Toxicology* 18, no 7: 426 1999.
80. Hain J, Jaussi R, Burkart W. Lack of adaptive response to low doses of ionizing radiation in human lymphocytes from five different donors. *Mutation Research* 1992;283(2):137-44.
81. Joksic G, Petrovic S. Lack of adaptive response of human lymphocytes exposed in vivo to low doses of ionizing radiation. *Journal of Environmental Pathology, Toxicology & Oncology* 2004;23(3):195-206.

82. Khandogina EK, Mutovin GR, Zvereva SV, Antipov AV, Zverev DO, Akifyev AP. Adaptive response in irradiated human lymphocytes: radiobiological and genetical aspects. *Mutation Research* 1991;251(2):181-6.
83. Luckey TD. Radiation hormesis. Boca Raton, Fla.: CRC Press, 1991.
84. Sankaranarayanan K, Duyn Av, Loos MJ, Natarajan AT. Adaptive response of human lymphocytes to low-level radiation from radioisotopes or X-rays. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 1989;211(1):7.
85. Shadley JD, Wiencke JK. Induction of the adaptive response by X-rays is dependent on radiation intensity. *International Journal of Radiation Biology* 1989;56(1):107-18.
86. Ito A. CELLULAR RESPONSES TO LOW DOSES OF RADIATION.
87. Maffei F, Angelini S, Forti GC, Lodi V, Violante FS, Mattioli S, Hrelia P. Micronuclei frequencies in hospital workers occupationally exposed to low levels of ionizing radiation: influence of smoking status and other factors. *Mutagenesis* 2002;17(5):405-409.
88. Mossman K. RADIATION EXPOSURE AND ADAPTIVE PROCESSES. *BELLE NEWSLETTER* 1999;7(3).
89. Muller WU, Dietl S, Wuttke K, Reiners C, Biko J, Demidchik E, Streffer C. Micronucleus formation in lymphocytes of children from the vicinity of Chernobyl after (131)I therapy. *Radiation & Environmental Biophysics* 2004;43(1):7-13.
90. Andreas Ziegler UG. The Generalised Estimating Equations: A Comparison of Procedures Available in Commercial Statistical Software Packages. *Biometrical Journal* 1998;40(3):245-260.
91. Beiyao Z. Summarizing the goodness of fit of generalized linear models for longitudinal data. *Statistics in Medicine* 2000;19(10):1265-1275.
92. John AN. A large class of models derived from generalized linear models. *Statistics in Medicine* 1998;17(23):2747-2753.
93. Lipsitz SR, Fitzmaurice GM, Orav EJ, Laird NM. Performance of Generalized Estimating Equations in Practical Situations. *Biometrics* 1994;50(1):270-278.
94. McCullagh P, Nelder JA. Generalized linear models. Monographs on statistics and applied probability; 37. 2nd ed. London; New York: Chapman and Hall, 1989.
95. Prentice RL, Zhao LP. Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses. *Biometrics* 1991;47(3):825-839.

96. Zeger SL, Liang K-Y, Albert PS. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics* 1988;44(4):1049-1060.
97. Zamboni WC, D'Argenio DZ, Stewart CF, MacVittie T, Delauter BJ, Farese AM, Potter DM, Kubat NM, Tubergen D, Egorin MJ. Pharmacodynamic Model of Topotecan-induced Time Course of Neutropenia. *Clin Cancer Res* 2001;7(8):2301-2308.
98. Mettler FA, Moseley RD. Medical effects of ionizing radiation. Orlando, FL: Grune & Stratton, 1985.
99. Mettler FA, Upton AC. Medical effects of ionizing radiation. 2nd ed. Philadelphia: W.B. Saunders, 1995.
100. Kossenko M, Akleyev A, Degteva M, Kozheurov V, Degtayaryova R. Analysis of Chronic Radiation Sickness Cases in the Population of the Southern Urals. In: *Medicine URCfR*, ed, 1994.
101. McLaughlin T, Monahan SP, Pruvost NL, Frolov VV, Ryazanov BG, Sviridov VI. A Review of Criticality Accidents. Los Alamos National Laboratories, 2000.
102. Vargo GJ. A brief history of nuclear criticality accidents in Russia--1953-1997. *Health Phys* 1999;77(5):505-11.