

Application of complex radiation risk models and exposure uncertainty
propagation to large-scale radiation cohort survival analysis with Colossus

by

Eric Giunta

B.S., Kansas State University, 2019

B.S., Kansas State University, 2019

M.S., Kansas State University, 2022

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Alan Levin Department of Mechanical and Nuclear Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Abstract

Studies on the effects of low-dose-rate chronic radiation exposure in populations require large amounts of data to find statistically significant associations between low radiation doses and different carcinogenic and cognitive health outcomes of interest. These studies are important not only for furthering the understanding of safe limits on Earth but also for improving the safety of missions to space. Existing statistical software packages were not designed to run complex risk models on millions of rows of data, and generally cannot consider exposure uncertainty or competing events in the data.

This need was met with the development and testing of the R package Colossus, which can apply risk models that were not typically available in R packages and take advantage of the available hardware to run regressions on big datasets quickly. One important aspect of Colossus is that it is open-source and uses a modular design to allow future capabilities to be added as needed. A series of tests were performed to verify that Colossus performed as intended and could perform analyses similar to other existing software.

The first novel application of Colossus was the analysis of large-scale plasmode simulations. A hypothetical study was proposed to characterize the competing risk effects of lung cancer mortality on the risk model parameter estimates of smoking status on Parkinson's disease mortality. Standard methods of plasmode simulations could not increase the observation time of individuals, which was hypothesized to lead to competing risk effects in the model parameter estimates. A random forest algorithm was used to extend the observation of individuals, and event rate models were simulated in the extended and unextended datasets. Analysis of the unextended data with only Parkinson's disease mortality produced results similar to the extended data with multiple competing events modeled, demonstrating that the unextended plasmode simulations induced effects similar to competing risks. Thus, simulation studies investigating the effects of competing risks in existing data require a method

of extending the observation of individuals to simulate outcomes without the effects of events in the original data.

The availability of exposure uncertainty propagation methods is one of the most significant capabilities of Colossus. Functions were implemented to perform Frequentist Model Averaging and Monte Carlo Maximum Likelihood algorithms using multiple realizations of radiation exposure. These automated functions could analyze thousands of realizations several times faster than manual methods because of the large amount of time required to process the data for each realization. These capabilities allow researchers to propagate exposure uncertainties into their analyses.

Colossus can currently perform radiation cohort analysis, which was previously not available with public software; however, further work is required. Additional capabilities are planned to be added, particularly different logistic model regression options. A multiple-realization analysis with real data is planned, which will provide more meaningful regression results and a better understanding of the true uncertainty in the risk-model parameters. Finally, further work is required to optimize Colossus. Further research into available libraries and alternative code bases may open up optimizations that would allow these analyses to be performed with even less time and memory required.

Application of complex radiation risk models and exposure uncertainty
propagation to large-scale radiation cohort survival analysis with Colossus

by

Eric Giunta

B.S., Kansas State University, 2019

B.S., Kansas State University, 2019

M.S., Kansas State University, 2022

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Alan Levin Department of Mechanical and Nuclear Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Approved by:

Major Professor
Amir A. Bahadori

Copyright

© Eric Giunta 2025.

Abstract

Studies on the effects of low-dose-rate chronic radiation exposure in populations require large amounts of data to find statistically significant associations between low radiation doses and different carcinogenic and cognitive health outcomes of interest. These studies are important not only for furthering the understanding of safe limits on Earth but also for improving the safety of missions to space. Existing statistical software packages were not designed to run complex risk models on millions of rows of data, and generally cannot consider exposure uncertainty or competing events in the data.

This need was met with the development and testing of the R package Colossus, which can apply risk models that were not typically available in R packages and take advantage of the available hardware to run regressions on big datasets quickly. One important aspect of Colossus is that it is open-source and uses a modular design to allow future capabilities to be added as needed. A series of tests were performed to verify that Colossus performed as intended and could perform analyses similar to other existing software.

The first novel application of Colossus was the analysis of large-scale plasmode simulations. A hypothetical study was proposed to characterize the competing risk effects of lung cancer mortality on the risk model parameter estimates of smoking status on Parkinson's disease mortality. Standard methods of plasmode simulations could not increase the observation time of individuals, which was hypothesized to lead to competing risk effects in the model parameter estimates. A random forest algorithm was used to extend the observation of individuals, and event rate models were simulated in the extended and unextended datasets. Analysis of the unextended data with only Parkinson's disease mortality produced results similar to the extended data with multiple competing events modeled, demonstrating that the unextended plasmode simulations induced effects similar to competing risks. Thus, simulation studies investigating the effects of competing risks in existing data require a method

of extending the observation of individuals to simulate outcomes without the effects of events in the original data.

The availability of exposure uncertainty propagation methods is one of the most significant capabilities of Colossus. Functions were implemented to perform Frequentist Model Averaging and Monte Carlo Maximum Likelihood algorithms using multiple realizations of radiation exposure. These automated functions could analyze thousands of realizations several times faster than manual methods because of the large amount of time required to process the data for each realization. These capabilities allow researchers to propagate exposure uncertainties into their analyses.

Colossus can currently perform radiation cohort analysis, which was previously not available with public software; however, further work is required. Additional capabilities are planned to be added, particularly different logistic model regression options. A multiple-realization analysis with real data is planned, which will provide more meaningful regression results and a better understanding of the true uncertainty in the risk-model parameters. Finally, further work is required to optimize Colossus. Further research into available libraries and alternative code bases may open up optimizations that would allow these analyses to be performed with even less time and memory required.

Table of Contents

List of Figures	xii
List of Tables	xx
Acknowledgements	xxiii
Dedication	xxiv
1 Introduction and Background	1
1.1 Recent History of Radiation Epidemiology	1
1.2 Basis for Commonly Used Radiation Risk Models	7
1.3 Risk Model Debate at Low Doses	9
1.4 Limitations of Existing Software	11
1.4.1 Model complexity and big data applications	11
1.4.2 Risk analysis accounting for exposure uncertainty	12
1.4.3 Epidemiological analysis accounting for multiple causes of loss	12
1.5 Goal and Objectives of this work	12
2 Survival Analysis Scoring and Optimization	14
2.1 Basic Model and Person Attributes	14
2.2 Poisson Model Specifics	18
2.3 Cox Proportional Hazards Model Specifics	19
2.4 Fine-Gray Competing Risk Model Specifics	21
2.5 Cox Proportional Hazards Plotting Functions	23
2.6 Methods of Model Comparison	28

2.6.1	Score Comparison	28
2.6.2	Residual Calculation	28
2.6.3	Use of Multi-model Inference	29
3	Software Design, Data Flow, and Hardware Use	31
3.1	Data Flow and Modular Testing	32
3.1.1	Derivative Testing	33
3.1.2	Comparison to Existing Software	35
3.1.3	Large-scale Complex Model Testing	37
3.1.4	Large-scale Multi-state Modeling	40
3.2	Big Data Optimization	41
3.3	Hardware Usage and Software Performance	45
4	An Example of Standard Use	53
4.1	Cox Proportional Hazards Regression	55
4.2	Poisson Rate Regression	58
4.3	Effect Modification	62
5	Validation Testing: Use With Million Person Study Data	65
5.1	Los Alamos National Laboratory Worker Results	67
5.2	U.S. Nuclear Power Plant Worker Results	70
5.3	Validation Test Conclusions	75
6	Competing Risks Effects in Plasmode Simulations	76
6.1	Theory	78
6.1.1	Effects of competing events in time series analysis	78
6.1.2	Applying known risk models to sampled intervals	79
6.1.3	Methods to sample from existing data	81
6.2	Methods	83

6.2.1	Annual exposure modeling with Random Forest algorithms	83
6.2.2	Plasmode sampling	87
6.2.3	Risk model derivation	88
6.3	Results and Discussion	91
6.3.1	Unextended plasmode analysis	93
6.3.2	Extended plasmode analysis	96
6.3.3	Plasmode method comparison	99
6.4	Conclusions and Future Work	103
7	Theory and Applications of Covariate Uncertainty	105
7.1	Theory	106
7.1.1	Justification of Implemented Methods	107
7.1.2	Monte Carlo Maximum Likelihood theory	108
7.1.3	Frequentist Model Averaging theory	109
7.1.4	Application to big data	110
7.2	Methods	111
7.2.1	Implementation of analysis methods into Colossus	111
7.2.2	Simulated radiation cohort data	114
7.3	Results and Discussion	117
7.3.1	Benchmarking	117
7.3.2	Regression results	120
7.4	Summary of Results	124
8	Future Work and Potential Improvements	126
8.1	Future Code Updates and Testing	127
8.1.1	Incorporating new capabilities and analysis methods	127
8.1.2	Further benchmarking and optimization of existing capabilities	128
8.1.3	Changes to libraries and code base	128

8.2 Future Epidemiological Analyses	129
Bibliography	131

List of Figures

1.1	Size of an Exposed Group Required to Detect an Increase in Total Cancer Mortality with a Lifetime Follow-up, According to Dose. Reproduced from table 7.2 in “Radiation Dose Reconstruction for Epidemiologic Uses” ¹	4
1.2	Generic synopsis for the presentation of annual organ dose and its uncertainty, reproduced from previously published work ²	5
1.3	Example plotting the distinction between EAR and ERR. An EAR measures the change in absolute rate and an ERR measures a change in relative rate. .	8
1.4	Radiation risk models can be composed of combinations of different linear and piecewise effects, due to the linearity and threshold potential in different steps of the outcome pathway.	10
2.1	Plots of the hazard (Eq. 2.35), cumulative hazard (Eq. 2.36), and survival function (Eq. 2.37) for the simulated data. The predicted background hazard changed past age 10 because the risk model did not account for the change in relative risk.	25
2.2	Schoenfeld residuals for the example. Around age 10, the residuals switched from generally negative to positive. This matched the true shift from overestimating to underestimating risk.	26
2.3	Martingale residuals for the example. Very few event rows had negative martingale residuals. This matched the underestimate of risk above age 10. . . .	27
2.4	Plots of the true and predicted hazard ratios for $x = 1$	27

3.1	The user calls R functions, the input is converted to matrices which are processed in C++, the C++ results are converted into lists, and the R function returns the results to the user. Reproduced from previously published work ³ .	32
3.2	Different regression functions all follow the same sequence of C++ functions, but exit at different points (denoted by color). Reproduced from previously published work ³ .	33
3.3	Stage 1 of testing. Colossus was run at points on a grid, and intermediate calculation values were printed to the console. Three values were compared: the sum of the terms, the sum of the relative risks, and the log-likelihood. The calculated derivatives were compared to the difference in value between grid points.	34
3.4	At every point on the grid, the values and derivatives were output to the console. Then the calculated values (solid line) were compared to the true derivatives (points).	35
3.5	Stage 2 of testing, comparison with existing software. Examples provided by 32-bit Epicure ⁴ were run in Colossus to compare the regression and plotting functions. Other residual plotting options were compared with the R survival library ⁵ .	36
3.6	Stage 3 of testing. Rate models were simulated to create datasets and analyzed in Colossus to compare the estimated and true model parameter values. The hazard ratio distributions were compared to evaluate whether discrepancies were significant.	39
3.7	Comparison between the true and predicted excess relative risk curves, and the distribution in error in calculated hazard ratio.	40
3.8	Standard function call with one realization analyzed per call. By running each regression separately, several identical steps were repeated thousands of times.	43

3.9	Function call with multiple realizations analyzed per call. Switching to one function call per analysis allowed fewer steps to be repeated.	44
3.10	Gradient-based optimization algorithms do not distinguish between local optima. A regression starting below 2.6 converges to the wrong point. A common solution is to perform multiple regressions at different starting points.	45
3.11	Basic parallelization method applied for independent column calculations. The equation applied to each column was dependent on a shared constant matrix, so values were assigned to each column simultaneously.	46
3.12	Basic parallelization method applied for accumulated sums. Each thread accumulated a sum over independent batches, and then the intermediate results were combined to calculate the sum over the entire dataset.	46
3.13	Regression duration using one core across different regressions methods and data size. In general the duration increased with number of covariates, risk groups, and intervals.	48
3.14	Regression duration using one covariate across different regressions methods and data size. The Poisson regressions did not use parallelized calculations, so they did not benefit from additional cores. Cox regressions parallelize over risk groups, so they saw benefit based on the number of risk groups.	49
3.15	Regression duration using eight covariates across different regressions methods and data size. Due to the increased number of parallel calculations, every regression method benefited from additional cores to some degree.	50
3.16	Peak observed memory use for regressions on a single core by number of covariates. Due to the low number of parallel calculations, the baseline memory required could be measured. Memory required generally increased with the size of the dataset and number of covariates.	51

3.17	Observed range of memory use for regressions on eight covariates by numbers of cores. Due to the number of parallel calculations, the range of observed memory use increased drastically with more cores.	52
6.1	The original data was used to create risk models and samples of individuals. These were combined to create randomly assigned outcome datasets for analysis. These steps are expanded on in figures 6.9, 6.10, and 6.11	78
6.2	Survival curves for the individuals in each birth cohort who stop being observed due to causes other than lung cancer or PD mortality. The NPP cohort was an updating dataset, so censoring rates were significantly different between the year-of-birth cohorts.	80
6.3	A plasmode simulation starts by sampling n subjects from the cohort. A new event time was generated for each subject. This is repeated M times to analyze the distribution in hazard ratio estimates.	81
6.4	The data goes from missing higher age observations to having a shared final observation age (dashed line), by predicting missing exposures (red)	83
6.5	Histogram of intervals in each annual dose bin. A majority of intervals contained zero annual dose, and the most common non-zero dose bins were on the scale of 1 – 10 mGy.	84
6.6	Survival curves showing (left) the proportion of individuals who continue to be observed overall and (right) the proportion who survive PD mortality. Due to other events in the data, most individuals left the NPP dataset due to causes other than PD mortality.	85
6.7	Two-dimensional histograms counting the proportion of intervals within each age and annual exposure bin. a-b) Non-zero annual lung exposures and annual brain exposures drop off after age 70. c-d) The same data normalized to the population in each age bin showed a nearly constant distribution up past age 75.	86

6.8	Two-dimensional histograms counting the proportion of intervals within each age and annual exposure bin with extended survival. Exposures were weighted to force decay in the expected dose with age past retirement, starting around age 80.	87
6.9	The NPP dataset was sampled to define random populations. In unextended sampling, the newly sampled population was immediately used as the exposed population. In the extended sampling, new intervals were added and then used as the exposed population.	88
6.10	The NPP dataset was used to approximate the background and relative event rates for multiple outcomes. These models are combined independently to approximate different combinations of competing risk effects.	89
6.11	Risk models were applied to exposed populations to randomly assign outcomes, and datasets were generated for different combinations of events. . . .	91
6.12	The populations in each SES cohort have similar biological sex distributions, but lower SES status was also associated with earlier years of birth.	92
6.13	The distribution of final cumulative doses were very similar between SES categories. Final observation ages peaked in the range of 50-70, however the peak was less pronounced in lower SES categories.	92
6.14	In the unextended plasmode simulations, the survival curve decreased further below the original curve as more event types were added. Variation in survival between simulations (depicted as 95% confidence bands) was due to the variation in the original final observation times between sampled populations.	93
6.15	HR parameter estimates for SES status on PD mortality in unextended samples without competing risks applied. (a-b) The average parameter values converge over the simulations, and (c) the true parameter values (solid) were below the observed average (dashed).	94

6.16	Violin plots of the low SES HR distributions for each grouping of events in unextended data. The median HR (middle solid line) was below the mean HR (dashed line). All three had very similar median values near the true value, and the mean value for the full set of outcomes was furthest off.	95
6.17	Average survival curve for overall observation for each competing risk simulation in extended data. As more events were added, the survival dropped (shown by arrows) until the simulation with all events closely matched the original data. The variation in survival curves among simulations was too low to visualize.	98
6.18	Violin plots of the low SES HR distributions for each grouping of events in extended data. The distributions for only PD and PD & lung were much narrower, while the distribution for the full set of outcomes was wider but was still significant at a 95% confidence interval.	98
6.19	HR parameter estimates for SES status on PD mortality in the extended data without competing risks modeled. (a-b) The average parameter estimates converged within 500 simulations and (c) there was a much narrower cluster of points around the true values.	99
6.20	(a) As additional events were added, the extended simulations approached the original survival curve, and the unextended simulations dropped below the original survival and (b) the closest simulations were within 5% of the original survival curve on average. 95% confidence bands were produced for the survival differences, the unextended survival was not significantly different while the extended survival was significantly different.	100
6.21	Histograms and violin plots for low SES HR central estimate distributions between the unextended simulations with PD and the extended simulations with a full set of outcomes for both Cox and Fine-Gray regressions. In both representations, the distributions were very similar.	101

6.22	Distribution of PD mortality proportion (events per 10,000) by SES status comparing plasmode methods. The closest methods both underestimated the true proportion, and the extended simulations tended to underestimate the high SES event rate.	102
6.23	Comparison of low SES HR central estimates for each group of events and extension. In every comparison for the same events modeled, the unextended simulation distributions had higher variances.	102
7.1	When a user calls a regression function, there are multiple steps required to process the data and reserve memory that can require significant amounts of time for large datasets. In the case of multiple realizations, these steps were identical between realizations.	111
7.2	MCML analysis in Colossus ran similarly to standard regressions: scores were calculated for each realization and the accumulated weighted sum was used to estimate the optimal parameter set. After the analysis was complete, the results were returned.	112
7.3	FMA analysis in Colossus was based on typical user steps. Each realization was optimized and then used as an initial guess for the next realization. After the analysis was complete, the lists of estimates, uncertainties, and scores were returned for further processing.	112
7.4	Activity level was shared between subjects working in the same decade, as log-normal distributions capped at 1×10^3 units.	114
7.5	The measured exposure duration for each subject was shared by age and distributed normally. Exposures were rounded so that every subject had a non-negative exposure duration.	115
7.6	Each working group was assigned a shielding factor relative to working group one. The second working group generally received less radiation and the third working group generally received more radiation than baseline.	115

7.7	Dose probability distributions for the first five realizations generated. The level of shared uncertainties produce significant differences in the curves, with some realizations having no intervals above 100 mGy annual dose and others having over 90% of intervals with annual dose above 100 mGy.	116
7.8	Benchmarking for the exponential risk models. Using the multi-realization functions in Colossus reduced the average time per realization from 8 seconds down to under 1 second for the FMA and MCML analyses. 95% confidence intervals for duration showed very little variation.	118
7.9	Benchmarking for the linear risk models. Using the multi-realization functions in Colossus reduced the time per realization from 8 seconds down to around 1-2 seconds for the FMA and MCML analyses respectively. 95% confidence intervals for duration showed very little variation.	120
7.10	For the exponential risk model data and both combinations of shared and unshared uncertainty in dose, the predicted model parameter distributions for the known variables matched in both FMA and MCML analyses.	121
7.11	For the exponential risk model data, the FMA and MCML results differed between uncertainty types. The FMA and MCML analyses with only unshared uncertainty matched closely.	122
7.12	For the linear risk model data and both combinations of shared and unshared uncertainty in dose, the predicted model parameter distributions for the known variables matched in both FMA and MCML analysis.	122
7.13	For the linear risk model data, the FMA and MCML results differed between uncertainty types. The FMA and MCML analyses with only unshared uncertainty matched closely but the shared uncertainty data did not match. . . .	123
7.14	The FMA analyses with shared uncertainty produced skewed distributions. In both cases, the true value was included in the 95% confidence interval. . .	124

List of Tables

1.1	Excess Relative Risk (ERR) at 100 mGy brain dose for Parkinson’s disease within 6 cohorts within the Million Person Study. Reproduced from publication ⁶	7
2.1	Table representation of Equation 2.13.	17
3.1	Example 6.12, comparison between software	37
3.2	Multi-state model simulation, Cox proportional hazards model comparison between software.	41
3.3	Example of multiple realization data.	42
4.1	Lung cancer data, exponential hazard ratio model results.	57
4.2	Lung cancer data, linear hazard ratio model results.	57
4.3	Lung cancer data, hazard ratio model comparison (95% confidence interval).	58
4.4	Lung cancer data, exponential Poisson model results.	60
4.5	Lung cancer data, linear excess relative risk Poisson model results.	60
4.6	Lung cancer data, Poisson rate model comparison (95% confidence interval).	61
4.7	Lung cancer data, Cox proportional hazards and Poisson exponential relative risk model comparison (95% confidence interval).	62
4.8	Lung cancer data, linear hazard ratio effect modification model results.	64
4.9	Lung cancer data, treatment effect modification comparison (95% confidence interval).	64
5.1	Hazard Ratio at 100 mGy (95% Confidence Interval) for four causes of death among LANL workers comparing Colossus results with published results based on 32-bit Epicure software.	67

5.2	Hazard Ratio for each dose category displayed for dose response in the published article on LANL workers (95% Confidence Interval) for four causes of death comparing Colossus results with published results based on 32-bit Epicure Software.	67
5.3	Log-Linear Excess Relative Risk at 100 mGy (95% Confidence Interval) for four causes of death among LANL workers comparing Colossus results with the published results based on 32-bit Epicure software.	69
5.4	Linear Excess Relative Risk at 100 mGy (95% Confidence Interval) for four causes of death among LANL workers comparing Colossus based on 32-bit Epicure software.	70
5.5	Hazard Ratio at 100 mGy (95% Confidence Interval*) for five causes of death among nuclear power plant workers comparing Colossus results with published results based on 32-bit Epicure software.	71
5.6	Hazard Ratio for each dose category displayed for dose response in the published article on nuclear power plant workers (95% Confidence Interval*) for five causes of death comparing Colossus results with published results based on 32-bit Epicure software.	71
5.7	Log-Linear Excess Relative Risk at 100 mGy (95% Confidence Interval*) for five causes of death among nuclear power plant workers comparing Colossus results with the published results based on 32-bit Epicure.	74
5.8	Linear Excess Relative Risk at 100 mGy (95% Confidence Interval) for five causes of death among nuclear power plant workers comparing Colossus results with the published results based on 32-bit Epicure software.	74
6.1	PD Mortality by SES status in original data.	91
6.2	HR estimate comparison to true value in unextended data considering only PD mortality.	94

6.3	PD Mortality Proportion (events per 10,000 and 95% confidence interval) by SES status in unextended data.	96
6.4	PD Mortality Proportion (events per 10,000 and 95% confidence interval) by SES status in extended data.	97
7.1	Example of weight overflow and underflow by reference AIC. The difference between AIC was large enough that some subset of the exponential values had to be forced to near 0 or ∞	110
7.2	Average single realization regression times and 95% confidence interval (sec- onds) for the FMA and MCML alternative options and exponential risk model.	118
7.3	Average single realization regression times and 95% confidence interval (sec- onds) for the FMA and MCML alternative options and linear risk model. . .	119

Acknowledgments

I would like to acknowledge the contributions of the various members of the Million Person Study who helped me shape my work into something that future epidemiologists can use, the educational professionals at Kansas State University whose courses provided the educational framework for my PhD, and the graduate students who graciously tested my software and helped it be in the usable state it is now.

This work was sponsored by NASA Grants 80NSSC19M0161 and 80NSSC23M0129 through subcontracts from the National Council on Radiation Protection and Measurements (NCRP). This work is additionally supported by U.S. Nuclear Regulatory Commission (NRC) Grant No. 31310021M0030 and Kansas State University's Johnson Cancer Research Center.

Some parts of the computing for this project were performed on the Beocat Research Cluster at Kansas State University, which is funded in part by NSF grants CNS-1006860, EPS-1006860, EPS-0919443, ACI-1440548, CHE-1726332, and NIH P20GM113109. Beocat Director Daniel Andresen and other Beocat staff provided valuable technical expertise.

Dedication

I would like to dedicate this to my soon-to-be wife and my cat. Without them to encourage me, I do not know where I would be. I would also like to thank my friends and family for always encouraging me.

Chapter 1

Introduction and Background

1.1 Recent History of Radiation Epidemiology

Merriam-Webster defines epidemiology as “the sum of the factors controlling the presence or absence of a disease or pathogen”⁷. Radiation epidemiology studies how radiation exposure in a population changes the rate of diseases and disorders. Radiation epidemiology has historically focused on the impact of specific exposures in limited populations. Radiation exposure measurements require an existing system to detect and report individual radiation exposures, or a highly detailed understanding of the radiation source. This limitation led to much of the understanding of the health effects of radiation originating from the analysis of specific groups, such as the Life Span Study (LSS) of atomic bomb survivors⁸, uranium miners⁹, and those impacted by the Chornobyl reactor accident¹⁰. Epidemiologists analyze these cohorts to understand the effects of radiation and to help protect future populations from radiation overexposure. One common trend is that these studies often investigate significant exposures to cohorts with thousands to tens of thousands of subjects. These studies differ in the duration of exposure; some cover acute exposures on the scale of seconds, while others cover chronic exposures on the scale of years.

Early studies of atomic bomb survivors defined the steps followed by radiation epidemiologists. After the atomic bombs were dropped, Japanese and American groups began working

in the surrounding areas to conduct surveys on the injuries and immediate health impacts of radiation exposure^{8;11}. In the following years, a census was conducted to collect information on where the survivors were at the time of the bombings. The cohorts collected during the survey formed the basis of multiple longitudinal studies conducted by the Atomic Bomb Casualty Commission and continued by the Radiation Effects Research Foundation to the present day. These studies were the basis for the development of dosimetry systems that used a combination of information on the bombs, surrounding buildings, and the location of survivors to predict radiation exposure to the organs of individual survivors^{8;11}. These dosimetric systems have been used with health and mortality records to quantify the effects of radiation exposure on cancer mortality, cancer incidence, and leukemia, among other health outcomes. Across different studies, models have been fit for the risk to survivors, in-utero exposed, and the children of exposed survivors^{8;11}.

Studies of atomic bomb survivors have provided significant insights into the effects of acute radiation exposure. Studies on the effect of increasing radiation doses on solid cancer incidence have found differences in the excess relative risk between male and female survivors. The response in men up to 1 Gy was lower than that in women but featured an upward curve in response. Comparisons by age of exposure found that most excess cases of solid cancer came from survivors exposed before the age of 20 years. These studies also found a significant difference in the estimated risk between models solving for a ratio of risk instead of a difference in event rate¹². Similar analyses have been performed for the incidence of leukemia, which also found evidence of non-linear trends in excess relative risk and significant effects of the age of exposure. However, in leukemia analyses, there was no evidence of sex-dependent differences¹³. Similar effects were found for the effects of radiation in children and in-utero exposure, while other studies failed to find evidence of changes in cancer mortality and incidence in the children of survivors relative to the parental radiation dose¹¹. These groundbreaking studies of atomic bomb survivors have provided vital insights into how radiation, age, and personal characteristics may impact the rates of health outcomes in a population. However, these results were difficult to apply broadly. These results were specific to the population of Japan at the time, which had a lack of 20-30 year old

civilian men¹¹. The nature of the exposure also made it difficult to apply these risk models to chronic exposures⁸.

Overall, studies agree on the damage caused by acute exposure on the scale of hundreds of rads. Brief exposures above 6 rad per hour and above 50 rads to the GI system can cause vomiting, while brief exposures above 800 rad can cause death. However, the threshold for death attributable to GI system exposure due to protracted exposure, less than 6 rad per hour, is 1800 rads¹⁴. The effects of exposure of hundreds to thousands of rads at a rate of several rad per hour are well-documented; however, there are ongoing disputes regarding the chronic effects of even lower rates of exposure. This is largely due to the lack of data on low-dose-rate exposures. The number of samples required to detect an effect is inversely related to the size of the effect; therefore, detecting a small effect at low exposures requires a large number of samples. This effect is described in detail in the 1995 book, “Radiation Dose Reconstruction for Epidemiologic Uses”¹. In the text, the authors proposed a study in which a population was exposed to a specific radiation dose and followed over their full lifetime. The risk model recommended by the BEIR V report¹⁵ was used to project lifetime risk of radiation-induced fatal cancers at a rate of 7.5 per rem and 10,000 people. The population was assumed to have a background rate of 1,800 cancer deaths per 10,000 people. For any dose, a minimum sample size was required to detect a significant difference in excess cases at 80% power and 95% confidence¹. At 20 rem and 150 excess cancers per 10,000 people, 5,200 samples were required. At 0.25 rem and 1.9 excess cancers per 10,000, 32,000,000 samples were required¹ (Fig. 1.1).

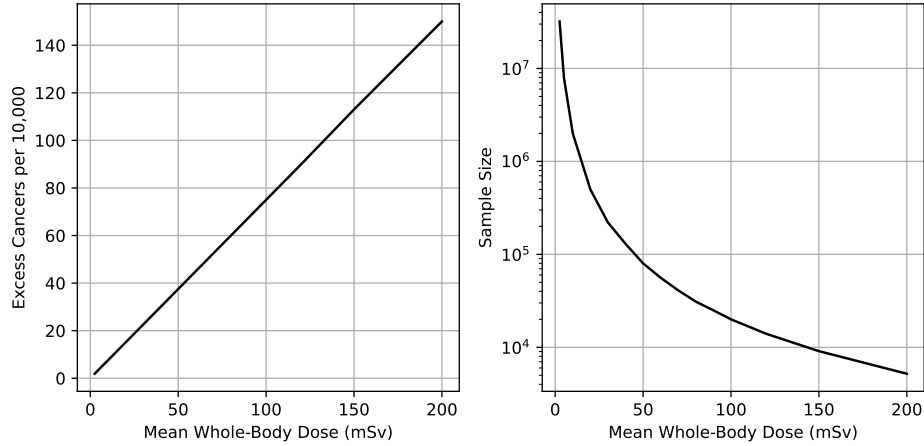


Figure 1.1: Size of an Exposed Group Required to Detect an Increase in Total Cancer Mortality with a Lifetime Follow-up, According to Dose. Reproduced from table 7.2 in “Radiation Dose Reconstruction for Epidemiologic Uses”¹

Multiple occupational exposure studies have attempted to solve this issue by collecting data on large populations exposed to low amounts of radiation over the past century. This includes the INWORKS study¹⁶ and the Million Person Study (MPS)¹⁷, the subject of this PhD work. The MPS was started as an effort to collect radiation exposure data on a population representative of the United States and provide a scientific understanding of the health effects of prolonged radiation exposure¹⁸. They aimed to accomplish this by collecting exposure and outcome data on cohorts of Department of Energy workers, nuclear weapon test participants, nuclear power plant workers, industrial radiographers, medical radiation workers, nuclear submariners, and other U.S. Navy personnel and radium dial painters, and applied evolving models for organ exposures and risk of health effects¹⁹.

Similar to atomic bomb survivor studies, publications from the Million Person Study were instrumental in demonstrating and standardizing methods for data collection and validation. To even have the data needed for risk analysis, significant work has been done to test methods of assigning vital statuses and causes of death²⁰ as well as methods of determining socioeconomic status differences from census data²¹. Statistical approaches from previous studies were compared to establish a standard for comparing studies using different types of risk models²². The MPS has been closely involved in efforts to improve dosimetric models and standardize the measurement and application of exposure uncertainty^{2;23}.

Dosimetry in the MPS followed a general process. For each cohort, the people and organs of interest were first identified. Then each unique external and internal source were identified. For every source of radiation, the available information was collected. This included information like the duration and exposure conditions. Then the information about each unique source of exposure was combined with ICRP published methodology to produce best estimates of annual organ doses and uncertainties. All of these sources would be combined to produce total annual organ doses from high and low linear-energy-transfer (LET) radiation sources (Fig. 1.2). A significant source of dosimetric information was the different radiation monitoring systems in place, like the Radiation Exposure Information and Reporting System (REIRS), DOE Radiation Exposure Monitoring System (REMS), and Landauer databases. The monitoring data was used to approximate exposure to different organs. This was complicated by possible exposure inhomogeneity over whole-body exposures, differences in protective gear between subjects, potential lack of compliance in wearing monitors, and changes in monitoring systems over time.

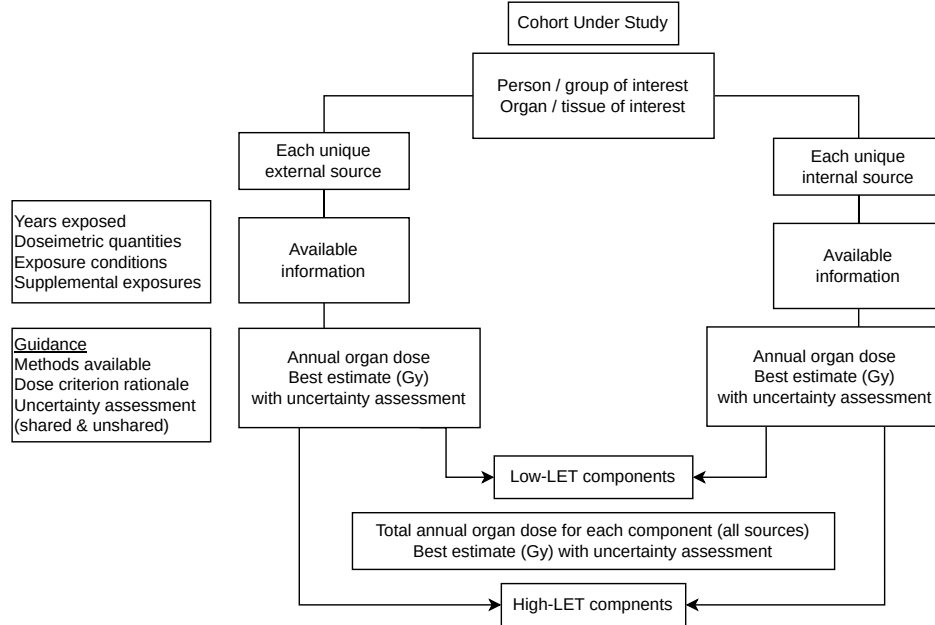


Figure 1.2: Generic synopsis for the presentation of annual organ dose and its uncertainty, reproduced from previously published work².

In addition to evaluating and publishing associations between radiation exposure and

causes of death among hundreds of thousands of workers, the MPS has made progress in studies of disease incidence²⁴. One goal of the study of chronic exposures is the increased risk of neurocognitive conditions such as Parkinson’s disease. Previous studies have found an association between chronic radiation exposure and Parkinson’s disease mortality; however, these longitudinal studies have generally reported a low number of neurocognitive mortality events²⁴. The incidence of disease is generally higher than mortality. Therefore, studies on the incidence of neurocognitive conditions could provide a clearer picture of the relationship between radiation exposure and neurocognitive conditions. This required finding a new source of incidence data, which has led to the incorporation of Medicare and Medicaid records²⁴. This provided a direct source of prescription data that could indicate when a patient was first diagnosed. Medicare and Medicaid records also provided information on cognitive testing in nursing homes and home health agencies. The Minimum Data Set (MDS) and Outcome and Assessment Information Set (OASIS) were systems that routinely assessed the physical and cognitive health of residents of facilities associated with Medicare and Medicaid, as well as associated home health agencies²⁴. Through these different sources, the MPS team developed procedures to infer information on incidence that was not possible with previous sources of vital status information.

To date, the Million Person Study has reported associations between radiation exposure and over 40 causes of death among over 600,000 workers²⁴. The goal of these publications was to report the current measurement of effects in individual cohorts, as well as to combine cohort results in meta-analyses to provide a measurement of the effects of radiation exposure across the collection of cohorts. The MPS has released publications reporting associations between organ doses and causes of death for individual cohorts, such as the Los Alamos National Laboratory²⁵ and Nuclear Power Plant²⁶ worker cohorts. The MPS study has also released publications focused on specific associations across multiple cohorts, such as the association between brain dose and Parkinson’s disease mortality⁶. A single summary estimate was produced through a meta-analysis of six different cohorts, each with different estimates of association. Each individual cohort prediction was not statistically significant, but the meta-analysis produced a statistically significant estimate (Table 1.1).

Table 1.1

Excess Relative Risk (ERR) at 100 mGy brain dose for Parkinson’s disease within 6 cohorts within the Million Person Study. Reproduced from publication⁶

MPS cohort	ERR at 100 mGy (95% CI)
Nuclear Power Plant Workers	0.24 (-0.02; 0.50)
Industrial Radiography Workers	0.24 (-0.13; 0.61)
Medical Radiation Workers	0.17 (-0.20; 0.54)
Los Alamos National Laboratory Workers	0.16 (-0.07; 0.40)
Rocky Flats Workers	0.13 (-0.11; 0.37)
Atomic Veterans	-0.22 (-0.90; 0.46)
Total	0.17 (0.05; 0.29)

1.2 Basis for Commonly Used Radiation Risk Models

The results of countless epidemiological studies were used to form the basis of the recommendations presented in the 2006 BEIR VII phase 2 report, “Health Risks from Exposure to Low Levels of Ionizing Radiation: Health Risks from Exposure to Low Levels of Ionizing Radiation”²⁷. This report presented a summary of the risk models used in atomic bomb studies, the biological basis for the model structure, and suggested a general series of formulae that could be used for future studies. In general, risk models describe the event rate in a cohort as a function of age (a, e), acute exposure (d), and study-specific risk factors (p). One of the most important model details presented is the different relationships between the background event rate and the cohort-specific event rate. In many models, the cohort-specific event rate is written as a function of the background rate (λ) and excess risk. The background rate is treated as a function of non-dose risk factors, the attained age (a) as well as other study-specific risk factors (p). The excess risk is treated as a function of dose (d) and any covariates the change the effect, generally the age of exposure (e), attained age (a), and other risk factors (p). An Excess Relative Risk (ERR) changes the ratio of cohort-specific

risk to background risk (Eq. 1.1) and an Excess Absolute Risk (EAR) changes the additive difference between cohort-specific risk and background risk (Eq. 1.2). The distinction is that an ERR model assumes a multiplicative effect, whereas an EAR model assumes an additive effect (Fig. 1.3).

$$Rate_{ERR} = \lambda(a, p)[1 + ERR(e, a, p, d)] \quad (1.1)$$

$$Rate_{EAR} = \lambda(a, p) + EAR(e, a, p, d) \quad (1.2)$$

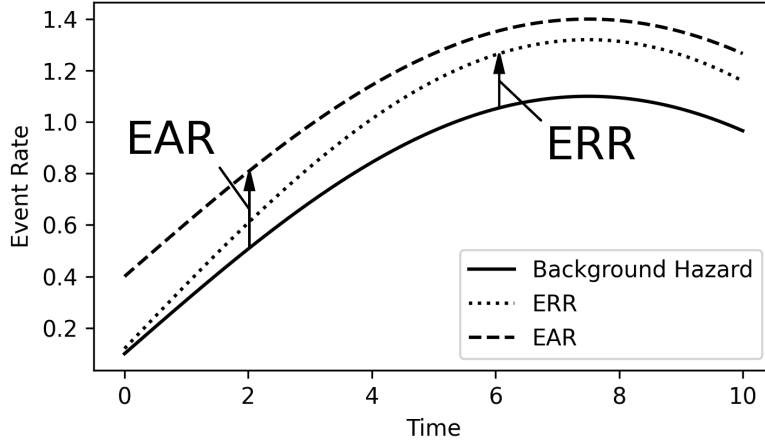


Figure 1.3: Example plotting the distinction between EAR and ERR. An EAR measures the change in absolute rate and an ERR measures a change in relative rate.

Another important result from the BEIR VII report was the suggested ERR/EAR dependence on the radiation dose²⁷. First, a previous analysis found that linear-quadratic risk models could describe the effects of higher acute exposures, particularly the relationship between higher exposures and cell death, reducing tumor formation rates. At lower doses, the quadratic terms approach zero, and the models become approximately linear. For lower doses, studies have investigated reparameterizing linear models as exponential models to reduce the bias in the predicted model parameter distributions²⁷. Based on the different risk models tested, the report presents a model applicable to a broad range of doses (Eq. 1.3).

$$f(d) = (\alpha_1 d + \alpha_2 d^2) \exp(-\alpha_3 d - \alpha_4 d^2) \quad (1.3)$$

1.3 Risk Model Debate at Low Doses

There is much debate regarding the functional form of risk models at low radiation doses. The goal of this PhD work was not to prove which model should be used, but it was important to test different models. Previous MPS studies have compared exponential and linear models²², and studies of atomic bomb survivors have compared linear and quadratic models⁸. There has been significant debate over the existence of a threshold for radiation-induced effects, such as cancer²⁸. Studies of atomic bomb survivors have demonstrated differences between models that assume exposure increases the relative probability of an event, as opposed to an additive effect⁸. To design software capable of applying various radiation dose risk models, it is important to understand the biological basis of different model formulae.

One method for predicting the effects of low doses of radiation is to extrapolate linear risk models from high to low radiation exposure. This is referred to as the linear no-threshold (LNT) model, which assumes that the excess risk increases proportionally to the dose²⁹. The LNT model was originally adopted based on atomic bomb survivor studies and formed the basis of public and occupational safety limits. The LNT model makes two key assumptions: the probability of an event increases linearly with the amount of radiation exposure, and any amount of radiation exposure increases the risk of an event. The chronic effects of radiation were believed to be due to incorrect repair of cellular damage. Increasing the amount of radiation exposure would increase the number of damaged cells, and therefore increase the probability of at least one damaged cell not being repaired correctly. Under these assumptions, any additional radiation exposure above background levels would increase the relative risk of chronic radiation effects. This is particularly important for radiation cleanup limits; the acceptable level of radiation following cleanup would be determined by the acceptable level of excess risk of radiation effects in the public.

Several issues with these assumptions have led to the proposal of alternative radiation risk models. There are three areas in which these models disagree with the LNT model: the assumption of no threshold, proportional damage to energy deposition, and risk strictly increasing with radiation exposure. Part of the disagreement arises from assumptions about cell repair mechanisms. The LNT model assumes that any damage can trigger DNA repair with near-constant accuracy. Some studies have found evidence of cells not initiating repair following low doses^{28;30} and only initiating repair after additional single-strand breaks were induced by a hydrogen-peroxide solution³¹. This could explain the potential thresholds in the risk model and suggests that damage from low doses may not be repaired by cells. If sufficiently low doses cause damage that is not repaired, then additional doses may be sufficient to initiate repair and reduce the risk of negative health effects. This could violate the assumption that the risk increases with additional doses. This also has a significant impact on radiation cleanup limits. This would suggest that past a point, further cleanup could be harmful. Further research into cell damage and the adoption of an Adverse Outcome Pathway (AOP) view of cancer formation³² could lead to drastically more complicated risk models than have been assumed in the past. This complexity should be reflected in the software designed to apply these models (Fig. 1.4).

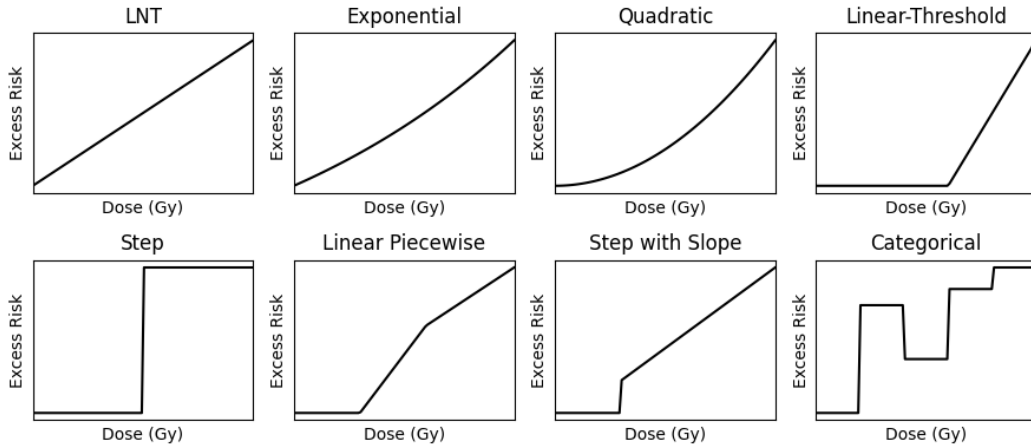


Figure 1.4: Radiation risk models can be composed of combinations of different linear and piecewise effects, due to the linearity and threshold potential in different steps of the outcome pathway.

1.4 Limitations of Existing Software

The basis of this PhD study was that the scale of data and complexity of risk models required by radiation epidemiology went beyond the limitations of existing software. Radiation epidemiological studies commonly use software like 32-bit Epicure⁴, SAS³³, and publicly available R packages. 32-bit Epicure has been a staple of radiation epidemiology since it was developed for atomic bomb survivor studies. It was designed to meet the model complexity required to describe radiation health effects but was limited in the size and structure of data that could be loaded. R packages, like the survival package⁵, can take advantage of other R packages to load and process very large datasets but are not designed for the level of model complexity required. A middle ground that could use both large datasets and complex risk models was lacking. The following sections briefly cover how existing packages fail to meet the needs of MPS analyses, which are expanded in later chapters.

1.4.1 Model complexity and big data applications

As previously discussed, there are a vast number of models that researchers want to use and compare, which have previously been available only through standalone software⁴ or functions developed by individual researchers and not distributed. Thus, the models that a researcher can use may be limited by either the software their system can run or by existing specialized code. This limitation could be avoided if there were a software package designed to run on multiple systems, apply a variety of risk models to different regression methods, and be modified by users to meet specific needs.

Another issue is the challenge of big data analysis. Analysis packages that were previously used to analyze cohorts with thousands of subjects may encounter bottlenecks when analyzing cohorts with millions of subjects. Access to software designed for big data will aid investigations of occupational studies in rapidly processing and analyzing data, by taking advantage of available hardware and existing software. The development of massive machine learning models has drastically improved the available methods for loading and processing large amounts of data. These principles and libraries should be leveraged to aid the

investigation of large longitudinal studies.

1.4.2 Risk analysis accounting for exposure uncertainty

One issue with radiation exposure is that dosimetry is not known with absolute certainty. The radiation source is typically known, and biological models are often used to approximate the expected amount of damage to a given person. Ultimately, the radiation exposure used for an analysis is the result of various models, approximations, and simulations. Importantly, this means that a probability distribution of exposures can be produced, but only a single estimate is used in the analysis. Methods are available to apply uncertain exposures²³, but can be computationally complex and may not be available for every desired risk model.

1.4.3 Epidemiological analysis accounting for multiple causes of loss

The lack of rare events is a common limitation of longitudinal studies in radiation epidemiology. Unlike case-control studies, the cross-sectional study design used in many large-scale radiation cohort studies cannot ensure a random distribution of risk factors in subjects with events. Risk factors correlated with higher rates of common events affect the distribution of individuals who survive long enough to experience rare events. For example, if individuals who smoke are more likely to die, then fewer smoking subjects will be at risk at older ages. This can lead to studies lacking the sample size required to detect a significant effect and introduce biases if unmeasured risk factors are correlated with multiple outcomes.

1.5 Goal and Objectives of this work

The goal of this work was to further the capabilities of the radiation epidemiological community so that they can perform the chronic radiation cohort studies necessary to better understand the health impacts of prolonged radiation exposure. This goal was divided into three main objectives. The first objective was to meet the level of model complexity required

and demonstrate that the approach could use the scale of the expected data. The second objective was to incorporate methods to propagate uncertainties in radiation exposure, in accordance with NCRP recommendations²³. The third objective was to incorporate and test methods of analysis for subjects at risk of multiple outcomes. This work was performed in conjunction with the MPS to incorporate the experience and needs of an internationally recognized team of researchers and directly support their studies on health effects in American cohorts¹⁸ and work with NASA on predicted health effects in astronauts³⁴. In summary, the culmination of the work during this PhD was the release of the R package Colossus^{35;36}. Colossus is an open-source package written in a combination of R and C++ code that provides a source of previously unavailable combinations of risk models and likelihood equations and is designed for the types of analyses required by modern radiation epidemiological studies.

Chapter 2

Survival Analysis Scoring and Optimization

Epidemiological studies are often interested in measuring both the overall and relative event rates²². Colossus is designed to perform Cox proportional hazards modeling, Poisson regression, and Fine-Gray competing risk modeling, each of which is used to fit an equation relating personal attributes to the probability of an event occurring.

2.1 Basic Model and Person Attributes

Colossus and survival analysis software in general assume that the data consist of several key columns: a measure of time, a measure of events, and covariates that may explain differences in event rates between rows. Generally, these columns are used to relate some equations for the event rate to the number of measured events. In Colossus, the risk (R) is a function of n terms ($T_0, T_1, \dots, T_i, \dots, T_{n-1}$). Each term, T_i , is the product of the linear and nonlinear subterms (S_j). Each subterm is a function of covariates ($\vec{\beta}$). The risk model can describe the hazard ratio, expected event rate, or excess risk.

Typically, risk is the sum or product of terms. The model formulae were based on those available in 32-bit Epicure, a widely used software by radiation epidemiologists⁴. The sim-

plest combinations are the Additive model (Eq. 2.1), Product Additive (Eq. 2.2), Product Additive Excess (Eq. 2.3), and Multiplicative Excess (Eq. 2.4). Even more intricate is the Geometric Mixture Model (Eq. 2.5). This model allows every term, T_i^* , to be a relative risk, T_i , or an excess relative risk, $1 + T_i$, and uses a weighting factor, θ , to adjust the combination of the additive and multiplicative effects.

$$R_A = \sum_{i=0} T_i \quad (2.1)$$

$$R_{PA} = T_0 \left(\sum_{i=1} T_i \right) \quad (2.2)$$

$$R_{PAE} = T_0 \left(1 + \sum_{i=1} T_i \right) \quad (2.3)$$

$$R_{ME} = T_0 \prod_{i=1} (1 + T_i) \quad (2.4)$$

$$R_{GMIX} = T_0 \left(\prod_{i=1} T_i^* \right)^\theta \left[1 + \sum_{i=1} (T_i^* - 1) \right]^{1-\theta} \quad (2.5)$$

Every term, T_i , is the product of linear and nonlinear subterms. Each subterm is a function of covariates (x_j) and model parameter estimates (β). The simplest subterms are linear (Eq. 2.6), product linear (Eq. 2.7), or log-linear (Eq. 2.8).

$$S_{i,L} = \sum_{j=0} x_j \cdot \beta_i \quad (2.6)$$

$$S_{i,PL} = 1 + \sum_{j=0} x_j \cdot \beta_i \quad (2.7)$$

$$S_{i,LL} = \exp \left(\sum_{j=0} x_j \cdot \beta_i \right) \quad (2.8)$$

Colossus also features a general, nonlinear subterm designed to encompass additional model formulae desirable for radiation epidemiology. The subterm is the sum of exponential functions, quadratic curves (Eq. 2.9), linear piecewise functions (F_{LT} , Eq. 2.10), and step func-

tion piecewise functions (F_{STP} , Eq. 2.11). The most straightforward use of this subterm would be to construct a linear piecewise function that could account for the threshold effects discussed in Section 1.3. The threshold could also be optimized by manually applying a threshold and solving for the optimum score for each set threshold. This subterm optimizes both the threshold and other model parameters, saving the time required to run multiple regressions.

$$S_{i,NL} = \sum_j (\alpha_i \cdot \exp(x_j \cdot \beta_i)) + \sum_j (\delta_i \cdot x_j^2) + \sum_j F_{i,LT} + \sum_j F_{i,STP} \quad (2.9)$$

$$F_{i,LT} = \begin{cases} \alpha_i \cdot (x_j - \beta_i) & , \text{ for } x_j > \beta_i \\ 0 & , \text{ otherwise} \end{cases} \quad (2.10)$$

$$F_{i,STP} = \begin{cases} \alpha_i & , \text{ for } x_j > \beta_i \\ 0 & , \text{ otherwise} \end{cases} \quad (2.11)$$

Each term is the product of the subterms used (Eq. 2.12), so a risk model could be as simple as a single log-linear term or as complicated as the product of multiple linear terms. As previously discussed, radiation epidemiology uses both ERR and EAR models²⁷. It was important that Colossus could model both sums and products of terms because that allowed Colossus to solve for both the ERR and EAR models.

$$T_i = S_{i,LL} \cdot S_{i,L} \cdot S_{i,PL} \cdot S_{i,NL} \quad (2.12)$$

In the preceding equations, the concatenation of the subterm model parameters forms a single vector of model parameters $\vec{\beta}$. Similarly, the corresponding column numbers, term numbers, and subterm formulae are stored as vectors. These vectors are the input values passed to the R functions. For example, the model represented in Equation 2.13 can be described using the columns in Table 2.1.

$$R = (1 + \beta_b \cdot x_b + \beta_c \cdot x_c) \exp(\beta_a \cdot x_a) \quad (2.13)$$

Table 2.1
Table representation of Equation 2.13.

Term number	subterm type	covariate
0	log-linear	a
1	linear	b
1	linear	c

Both Cox proportional hazards and Poisson modeling use an estimate of the event rate equation and the events in each row to calculate a log-likelihood (LL), representing how well the risk model explains differences in event rates in the data. The model is optimized by maximizing the log-likelihood via the Newton-Raphson method³⁷ by calculating the gradient and Hessian of the log-likelihood to take controlled steps toward an estimated root of the first derivative of the log-likelihood.

$$\Delta \vec{\beta} \approx -\frac{\partial LL}{\partial \vec{\beta}} \times \left(\frac{\partial^2 LL}{\partial \vec{\beta}^2} \right)^{-1} \quad (2.14)$$

The maximum step size is limited to user-provided limits, and half-steps are applied as necessary. If a step is taken and the log-likelihood does not improve, Colossus tries a half-step. As Colossus approaches the solution, the maximum step size is reduced to limit overstepping. The iterations continue until the maximum step or the largest absolute first derivative is less than a user-provided threshold. If the first derivatives are near zero, then the model parameters are at an optimum and the model is considered converged. However, this optimization method does not differentiate between local and global optima.

2.2 Poisson Model Specifics

The Poisson model predicts the average number of events per person and per unit time. Poisson modeling assumes that each row of the data has a measure of interval duration in person-years (P_i), the number of events during the interval (c_i), and the covariates (x_i) and parameters ($\vec{\beta}$) that define the event rate (λ). The true parameters are unknown; therefore, an estimate of the parameters ($\hat{\beta}$) is used to calculate the expected number of events for each row (Eq. 2.16). This estimate of the event rate (E_i) is compared to the number of observed events to measure how well the model fits the data.

$$\lambda_i = \lambda(\hat{\beta}, x_i) \quad (2.15)$$

$$E_i = P_i \cdot \lambda_i \quad (2.16)$$

There are two scores used to evaluate the goodness of fit of the Poisson model to the data: the log-likelihood (Eq. 2.17-2.19) and the deviance (Eq. 2.20). Deviance (D) is calculated as the difference between the log-likelihood using the estimated number of events and the log-likelihood substituting the true number of events.

$$LL = \sum_i^n (c_i \cdot \ln(P_i \cdot \lambda_i) - P_i \cdot \lambda_i) \quad (2.17)$$

$$\frac{\partial LL}{\partial \hat{\beta}_a} = \sum_i^n \left(c_i \cdot \frac{\frac{\partial}{\partial \hat{\beta}_a} \lambda_i}{\lambda_i} - P_i \cdot \frac{\partial}{\partial \hat{\beta}_a} \lambda_i \right) \quad (2.18)$$

$$\frac{\partial^2 LL}{\partial \hat{\beta}_a \partial \hat{\beta}_b} = \sum_i^n \left(c_i \cdot \left[\frac{\frac{\partial^2}{\partial \hat{\beta}_a^2} \lambda_i}{\lambda_i} - \frac{\frac{\partial}{\partial \hat{\beta}_a} \lambda_i}{\lambda_i} \cdot \frac{\frac{\partial}{\partial \hat{\beta}_b} \lambda_i}{\lambda_i} \right] - P_i \cdot \frac{\partial^2}{\partial \hat{\beta}_a \partial \hat{\beta}_b} \lambda_i \right) \quad (2.19)$$

$$D = 2 \sum_i^n \left(c_i \cdot \ln \left(\frac{c_i}{P_i \cdot \lambda_i} \right) - (c_i - P_i \cdot \lambda_i) \right) \quad (2.20)$$

A negative number of events is impossible; therefore, neither score is defined if the event rate is negative. For rows without counts, the first term of the deviance equation (Eq. 2.20) simplifies to zero using L'Hôpital's rule. The final equation used by Colossus must predict a

non-negative rate for every row of data, which is automatically checked. If a negative rate is calculated during a step, a half-step is used. If the user starts a regression at a point that gives a negative rate, Colossus returns an error that explains the issue.

Both log-likelihood and deviance are calculated by Colossus during the regression. After maximizing the log-likelihood, the Akaike information criterion³⁸ (AIC, Eq. 2.21), and Bayesian information criteria³⁹ (BIC, Eq. 2.22) provide scores to compare models. The AIC and BIC combine the optimized log-likelihood ($\hat{LL}$) with the number of parameters (k) and number of rows (n) of data, often to penalize the addition of low-information parameters to a model and avoid overfitting.

$$AIC = 2k - 2\hat{LL} \quad (2.21)$$

$$BIC = k \cdot \ln(n) - 2\hat{LL} \quad (2.22)$$

2.3 Cox Proportional Hazards Model Specifics

A Poisson model is used to estimate the event rate, but a Cox proportional hazard model is used to estimate the relative event rate compared to the baseline event rate in the cohort⁴⁰. The model assumes that each row of the data (i) has a measure of the start (TS_i) and end (TE_i) of a time interval, a classifier indicating whether an event occurred during the interval (δ_i), and both covariates (x_i) and parameters ($\vec{\beta}$) that define the relative risk (r_i). Cox proportional hazards models function by comparing the relative risk of rows that experienced events to the relative risk of rows that did not experience an event at each event time. The risk of an event is assumed to be the product of a baseline hazard and an independent hazard ratio; the hazard ratio can be estimated by comparing rows with the same baseline hazard. The relative hazard is always a constant hazard ratio, not a constant difference. This is an important distinction between Cox and Poisson models. An excess relative risk (ERR) can be modeled using both methods, but a Cox model can not model an excess absolute risk model (EAR) (Fig. 1.3).

The Cox proportional hazards model assumes that the hazard can be split into a baseline hazard shared by each stratum and a hazard ratio that accounts for individual risk factors. The baseline hazard is assumed to be a function of time, and the hazard ratio is assumed to be independent of time. The intervals are sorted into risk groups (R_j) for each event time (t_j). Risk groups can be further divided using stratification variables. If a row interval contains an event time (t_j), then the row is in the risk group for that event time (R_j) (Eq. 2.23). Each row can be part of multiple risk groups or not part of any risk group.

$$i \in R_j : TS_i < t_j \leq TE_i \quad (2.23)$$

The score used to evaluate the accuracy of the Cox proportional hazards model is the partial likelihood, which relates the hazard ratio in event rows to the hazard ratio in rows at risk⁴⁰. The simplest formula assumes that every risk group has a single event, but the formula can be extended to multiple events using the Breslow estimator⁴¹ (Eq. 2.24).

$$L(\hat{\beta}) = \prod_j^N \left(\prod_{i \in R_j, \delta_i=1} \left[\frac{r_i}{\sum_{l \in R_j} r_l} \right] \right) \quad (2.24)$$

The issue with the Breslow estimator is that the events occurred in an order, so each tied event should have a different risk group. However, the Breslow estimator uses the same denominator for each tied event. This approximation can be improved by assuming that every way of breaking the ties is equally likely. The Efron formula⁴² uses an additional term in the denominator that applies the average weight (Eq. 2.25), assuming that there are d_j events for the risk group R_j .

$$L(\hat{\beta}) = \prod_j^N \left(\frac{\prod_{i \in R_j, \delta_i=1} r_i}{\prod_{k=1}^{d_j} \left[\sum_{l \in R_j} r_l - k/d_j \sum_{i \in R_j, \delta_i=1} r_i \right]} \right) \quad (2.25)$$

The denominator is written as $r_j^*(k)$ and is assumed to be the result of the Efron approximation (Eq. 2.26); however, the corresponding Breslow estimator value can be substituted

without loss of generality. Colossus optimizes the natural log of the likelihood or the log-likelihood (Eq. 2.27-2.29).

$$r_j^*(k) = \sum_{l \in R_j} r_l - k/d_j \left(\sum_{i \in R_j, \delta_i=1} r_i \right) \quad (2.26)$$

$$LL(\hat{\beta}) = \sum_j^N \left(\sum_{i \in R_j, \delta_i=1} \ln(r_i) - \sum_{k=1}^{d_j} \ln(r_j^*(k)) \right) \quad (2.27)$$

$$\frac{\partial LL(\hat{\beta})}{\partial \beta_a} = \sum_j^N \left(\sum_{i \in R_j, \delta_i=1} \left[\frac{\frac{\partial}{\partial \beta_a} r_i}{r_i} \right] - \sum_{k=1}^{d_j} \left[\frac{\frac{\partial}{\partial \beta_a} r_j^*(k)}{r_j^*(k)} \right] \right) \quad (2.28)$$

$$\begin{aligned} \frac{\partial^2 LL(\hat{\beta})}{\partial \beta_a \partial \beta_b} = & \sum_j^N \left(\sum_{i \in R_j, \delta_i=1} \left[\frac{\frac{\partial}{\partial \beta_a \partial \beta_b} r_i}{r_i} - \frac{\frac{\partial}{\partial \beta_a} r_i}{r_i} \frac{\frac{\partial}{\partial \beta_b} r_i}{r_i} \right] \right) \\ & - \sum_j^N \left(\sum_{k=1}^{d_j} \left[\frac{\frac{\partial}{\partial \beta_a \partial \beta_b} r_j^*(k)}{r_j^*(k)} - \frac{\frac{\partial}{\partial \beta_a} r_j^*(k)}{r_j^*(k)} \frac{\frac{\partial}{\partial \beta_b} r_j^*(k)}{r_j^*(k)} \right] \right) \end{aligned} \quad (2.29)$$

Due to the use of logarithms, the derivative equations for both the Poisson and Cox models are dependent on the ratio of derivatives of risk to the risk ($\frac{\partial r}{\partial \beta}/r$, Eq. 2.18-2.19, 2.28-2.29). The same ratio was calculated multiple times throughout the log-likelihood calculations. Not only is it faster to calculate these values once, but it is often more accurate to calculate these ratios during risk calculation with the additional context of the entire risk model.

2.4 Fine-Gray Competing Risk Model Specifics

The Fine-Gray competing risk model⁴³ was designed to resolve issues with the Cox proportional hazards model. One of the assumptions of the Cox proportional hazards model is that every subject either experiences the event of interest or is randomly censored from the study. The censoring rate is generally assumed to be the same for each subject. Common examples include tracking every subject up to the same duration or final age, or if subjects

randomly leave the study. An important assumption is that censoring and event times are statistically independent. In many cases, there is more than one event to which the subject can succumb, and these events may not affect the subjects equally. An example of an investigation considering these types of events is a 2014 study into kidney cancer in French and German uranium miners, which applied a competing risk method to account for miners having higher rates of death due to circulatory system disease or lung cancer⁹. The risk groups for kidney cancer at later ages were biased by the subjects who were more likely to survive. Events that are not modeled but affect risk groups are called competing events, which are not considered in the standard Cox proportional hazards model.

The Fine-Gray model defines cause-specific hazards, which are combined to calculate the total probability of a subject surviving to a given age. Instead of using a binary indicator for experiencing an event during an interval, δ_i , the Fine-Gray model assumes that every row has a value that indicates whether the interval was censored or experienced one of the multiple possible events. The primary difference between the Fine-Gray and Cox proportional hazard models is the treatment of competing events. When a subject experiences a competing event in the Cox proportional hazards model, they can be treated as immediately censored or as never experiencing the event of interest. The Fine-Gray model uses a middle ground by computing the probability that the subject would not be censored prior to the subsequent risk groups. The risk groups in the Fine-Gray model consist of the union of the Cox proportional hazards risk groups (Eq. 2.23) and rows that already experienced a competing event (Eq. 2.30).

$$i \in R_{jk} : (TS_i < t_j \leq TE_i) \cup (TE_i < t_j, c_i \neq k) \quad (2.30)$$

The log-likelihood equations for the Fine-Gray model are almost identical to those for the Cox proportional hazards model. The risks for rows with competing events are weighted based on the relative censoring rate at the competing event time and the future risk group time (Eq. 2.32). Assume that a row has a competing event at time t_k and is being evaluated in a risk group for time t_j . The risk should be weighed by the probability that the subject

would not have been censored before t_j , given that the subject survived to t_k ($P(t \geq t_j \mid t \geq t_k)$). This probability is used as a weight (w_{kj} , Eq. 2.31) to add competing event contributions to risk groups (Eq. 2.32). The log-likelihood equations are calculated as shown in Equations 2.27-2.29, substituting an updated sum of risks (r^* , Eq. 2.32) and using the updated definitions for the risk groups (Eq. 2.30).

$$w_{kj} = P(t > t_j \mid t > t_k) = \frac{P(t \mid t > t_j)}{P(t \mid t > t_k)} \quad (2.31)$$

$$r_j^*(k) = \sum_{l \in R_j} (w_{lj} \cdot r_l) - k/d_j \left(\sum_{i \in R_j, \delta_i=1} r_i \right) \quad (2.32)$$

There are two equivalent ways to format the input for these regression analyses. Colossus allows every row to have an assigned weight. By default, it is assumed that rows with competing risks are flagged, $c_i = 2$, and every row has a weight for the non-censored fraction. Competing risk rows are added to the relevant risk groups, and the weighting applied to each row in the risk group is the ratio of the row weights. In this method, the code assigns rows to risk groups and calculates the relative weights. Alternatively, a user can construct a new dataset without competing risk events flagged and add additional weighted, censored rows with the same effect. Suppose that instead of adding a single row to multiple future risk groups, a new row is defined in the dataset for each future risk group. If the weights of these new rows are the relative weights calculated in the first method and every original row is assigned a weight of unity, then the two methods perform the same analysis. This allows both simple ratios of weights and more complicated functions of weights to be used.

2.5 Cox Proportional Hazards Plotting Functions

The following subsection covers the plotting capability in Colossus to estimate the baseline survival rate or test assumptions about the Cox model. To provide examples, data were simulated using a known hazard formula (CH_{ij} , Eq. 2.33), and were analyzed using an approximate relative risk that is not dependent on time (r_i , Eq. 2.34). This model violates

the assumption of a constant hazard ratio with time, as demonstrated in the results of the assumption testing.

$$CH_{ij} = \begin{cases} 0.03 \cdot t_j \cdot (1 + 0.1 \cdot x_i) & , \text{ for } t_j < 10 \\ 0.03 \cdot t_j \cdot (1 + 3.5 \cdot x_i) - 1.02 \cdot x_i & , \text{ for } t_j \geq 10 \end{cases} \quad (2.33)$$

$$r_i = 1 + x_i \quad (2.34)$$

A Cox proportional hazards model does not directly solve for the baseline hazard; however, an estimated baseline hazard function can be computed. If the covariates had no effect, then the baseline at any event time would be equal to the proportion of rows at risk that experienced events. If the proportional hazard equation indicates that one of the rows at risk is more or less likely to have an event, then one would expect that the number of events would be greater or lesser than the baseline events. The estimated hazard is approximately the number of observed events (d_j) weighted by the sum of the hazard ratios at the event time (Eq. 2.35).

$$\hat{\lambda}_0(t_j) = \frac{d_j}{\sum_{k \in R_j} r_{kj}} \quad (2.35)$$

The baseline hazard can be accumulated to solve for the cumulative hazard (Eq. 2.36) and the exponential of the negative cumulative baseline hazard provides an estimate of the baseline survival function ($\hat{S}$, Eq. 2.37). The true underlying relationship between age and survival rate is also known as the survival function. If the hazard ratios were all set to unity, then this survival function is the Kaplan-Meier estimate. This survival function can also be divided by the values of a factor to observe the effect of stratification.

$$\hat{\Lambda}_0(t) = \sum_{t_j \leq t} \hat{\lambda}_0(t_j) \quad (2.36)$$

$$\hat{S}_0(t) = \exp(-\hat{\Lambda}_0(t)) \quad (2.37)$$

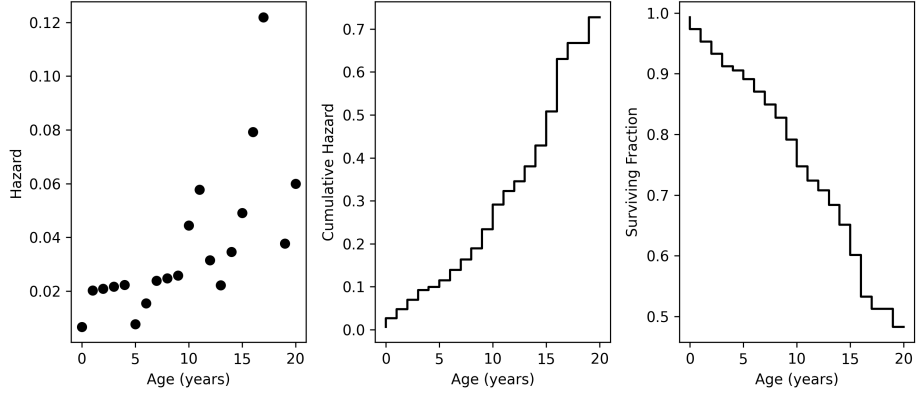


Figure 2.1: Plots of the hazard (Eq. 2.35), cumulative hazard (Eq. 2.36), and survival function (Eq. 2.37) for the simulated data. The predicted background hazard changed past age 10 because the risk model did not account for the change in relative risk.

Colossus also enables the user to calculate Schoenfeld residuals for a Cox model to test the proportional hazards assumption⁴⁴. Schoenfeld residuals show the difference between the average covariate value of rows with events and the risk-weighted average of the covariate over every row at risk (Eq. 2.38). Consistently large or small residuals indicate that the hazard is greater or less than the predicted values. Residuals correlated with event time may indicate that the hazard is time-dependent. The residual can be inversely weighted by the variance to find a scaled Schoenfeld residual, which represents the variation in hazard with time in time-dependent covariates⁴⁵.

$$s(t_j) = \frac{\sum_{i \in R_j, \delta_i=1} x_i}{d_j} - \frac{\sum_{i \in R_j} (x_i \cdot r_i)}{\sum_{i \in R_j} (r_i)} \quad (2.38)$$

For the example data, there was a clear trend in the Schoenfeld residuals (Fig. 2.2). As age increased, the average covariate value of the rows with events became larger than the risk-weighted average covariate value. The predicted relative risk model underestimated the risk below age 10 and overestimated the risk above 10 years of age. This violation of the model assumptions led to a non-random trend in the residuals with age.

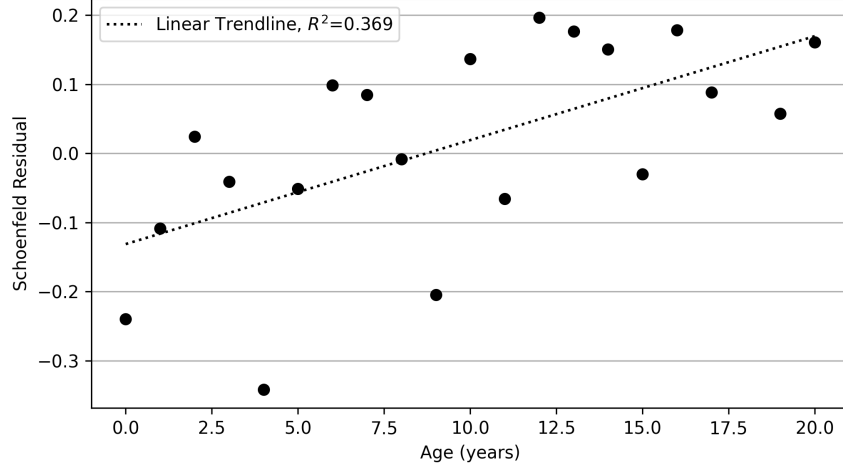


Figure 2.2: Schoenfeld residuals for the example. Around age 10, the residuals switched from generally negative to positive. This matched the true shift from overestimating to underestimating risk.

Colossus calculates and plots the Martingale residuals by survival time and covariate value⁴⁶. The Martingale residual is calculated as the difference between the event status and the cumulative hazard at the final event time for each subject (Eq. 2.39). The residual is bound in the open interval $(-\infty, 1)$. Negative values corresponded to rows that survived longer or at greater risk, and values near unity indicated an event at a combination of small risk or short survival time.

$$m_i = \delta_i - \int_{TS_i}^{TE_i} \left(\hat{\lambda}_0(t) r_i(t) \right) dt \quad (2.39)$$

For the example data, approximately 30% of the rows experienced no events and had a Martingale residual between -1.4 to -0.7 (Fig. 2.3). Very few rows experienced an event and had Martingale residuals below 0.0. This indicates that rows that experienced events did so despite a low cumulative hazard, suggesting that the rows had a greater risk than that predicted by the model.

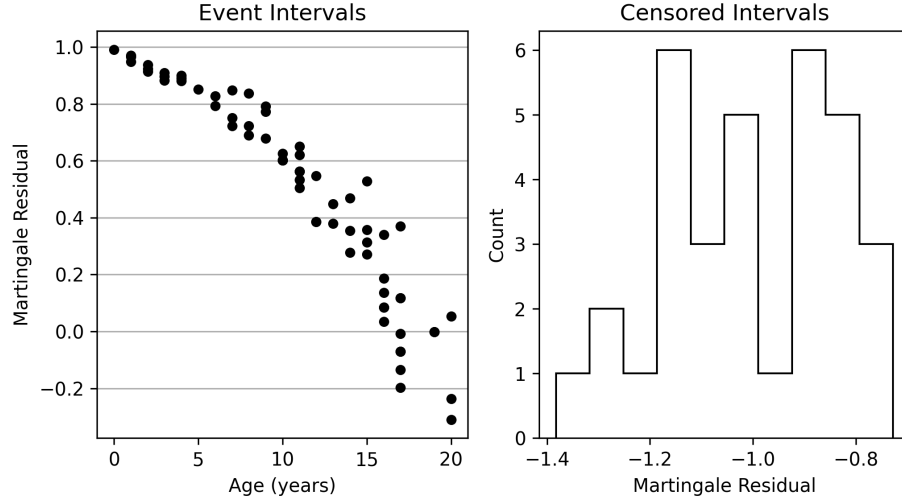


Figure 2.3: Martingale residuals for the example. Very few event rows had negative martingale residuals. This matched the underestimate of risk above age 10.

The fidelity of the Cox proportional hazards model can be tested using a combination of these methods. For this example, the residuals indicated that there may have been an unmodeled time-dependent effect that increased the risk as time increased. This unmodeled effect matched what one would expect from not including time dependence in the risk equation. One would expect to see an underestimate in the hazard ratio with increasing age and a change in residuals around age 10 (Fig. 2.4).

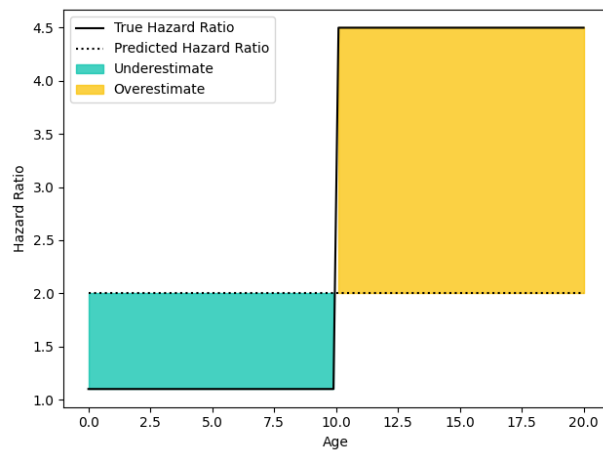


Figure 2.4: Plots of the true and predicted hazard ratios for $x = 1$

2.6 Methods of Model Comparison

In many cases, a researcher will fit multiple models for risk and must determine the best estimate. In some cases, a researcher might add new risk factors to a model and want to know if the model improved. In other cases, a researcher may have multiple different formulas to explain the effect of a risk factor and may want to know which formula best describes the true relationship. There are three general ways that models have been compared and best estimates have been produced.

2.6.1 Score Comparison

One common method of comparing models in radiation epidemiology is to compare the overall score metric, like the log-likelihood (Equations 2.17,2.27), AIC and BIC (Equations 2.21,2.22), or deviance (Eq. 2.20). When Colossus was first developed, the MPS team provided a list of necessary capabilities. Two of the items on the list were the ability to calculate the AIC and the ability to perform likelihood ratio tests. Many studies of radiation cohorts have focused on fitting multiple ERR and EAR models, with differences in linear, quadratic, and threshold effects, to determine the most appropriate risk model. Some atomic bomb survivor studies tested different risk models using AIC to compare similar models but kept ERR and EAR separate¹³. Others have presented differences in Deviance in Poisson models to indicate relative model performance¹². Another method for model comparison is to use parameter uncertainties in nested models. One method for comparing threshold models to non-threshold models is to measure the threshold parameter confidence interval¹². If a linear model with a threshold was not significantly different from a linear model without a threshold, a non-threshold model was determined to be more appropriate.

2.6.2 Residual Calculation

Score comparison is important for comparing existing models, but it does not provide information on errors in the model. Cox and Poisson models both have ways of calculating

residuals, which can provide information on how performance varies across the dataset. For a Cox model, Schoenfeld (Eq. 2.38)⁴⁴ and Martingale (Eq. 2.39)⁴⁶ residuals can be calculated to measure how well the predicted hazard ratios match the data. By plotting the residuals for two different models by radiation exposure, the relative performance at different radiation exposures can be compared. This could provide information on where different models could overestimate or underestimate the risk.

Poisson models can compare the predicted event rate to the observed event rate, which provide more easily interpreted residuals. Two options are deviance (Eq. 2.41) and Pearson residuals (Eq. 2.42)⁴. Similar to Cox residuals, these residuals can be plotted against risk factors to detect non-random trends. Once again, residuals could be calculated for two models to compare the difference in performance based on risk factors.

$$D_i = 2 \left(c_i \cdot \ln \left(\frac{c_i}{P_i \cdot \lambda_i} \right) - (c_i - P_i \cdot \lambda_i) \right) \quad (2.40)$$

$$r_i^D = \text{sgn}(c_i - P_i \cdot \lambda_i) \sqrt{D_i} \quad (2.41)$$

$$r_i^P = \frac{c_i - P_i \cdot \lambda_i}{\sqrt{P_i \cdot \lambda_i}} \quad (2.42)$$

2.6.3 Use of Multi-model Inference

Finally, some studies have demonstrated methods for combining the results of multiple models. In many cases, a researcher may look to the suggestions of research groups and committees that each endorse suggested radiation risk models, and may be confronted with multiple risk models that yield different estimates and uncertainties⁴⁷. Instead of picking a single model to use, Multi-Model Inference (MMI) could be used to produce a weighted estimate. The MMI ERR estimate is the weighted sum of the ERR from each model (Eq. 2.44), with weights calculated from the relative model performances (Eq. 2.43)⁴⁷. Monte Carlo based uncertainties estimates can be calculated by applying the model parameter uncertainties from each model to randomly sample MMI estimates. These methods have been tested with both AIC and BIC based weights^{47;48}, and studies have shown how the MMI approach

could be extended to estimates of the Radiation Attributed Decrease of Survival (RADS) in astronauts⁴⁸.

$$w_i = \frac{\exp(-0.5(AIC_i - \min(AIC)))}{\sum_{j=1}^m \exp(-0.5(AIC_j - \min(AIC)))} \quad (2.43)$$

$$ERR_{MMI} = \sum_{i=1}^m w_i \cdot ERR_i \quad (2.44)$$

Chapter 3

Software Design, Data Flow, and Hardware Use

Colossus was coded in both R and C++ so that calculations could be quickly run in compiled C++ code, but the user only needed to be familiar with R to use the package. The first section of this chapter discusses the general flow of data through common functions and the steps taken to perform verification testing*. The second section discusses the steps taken to optimize Colossus for big data and various types of analyses. The third section discusses the hardware usage and provides details on the performance scaling.

Modularity is an integral feature of Colossus. Overall, the user has access to a wide variety of different types of risk models that can be optimized and compared. These capabilities were designed such that alternative versions can be easily added. Every step of the analysis in Colossus was written as a separate function with standard inputs and outputs. Each iteration of a regression in Colossus requires a risk model to be applied to the data, a likelihood to be calculated, and the likelihood to be used to calculate a step to be taken toward the optimized parameter set. Risk model functions convert the parameters to risk, likelihood functions convert the risk into a score, and step calculation functions convert the score to updated parameters. New versions of each type of function can be easily added,

*Validation testing is discussed in [Chapter 5](#)

such that any combination of new and existing functions is available immediately.

3.1 Data Flow and Modular Testing

Colossus has two types of functions: calculation and analysis. Data cleaning and transformation are steps that the user can execute or Colossus can automatically run during regression functions. In most cases, the user does not need to call the calculation functions directly; instead, they are run as part of other functions. Thus, the functions available in Colossus were separated into calculation functions that could be quickly performed in C++ and analysis functions that the user needs to call and understand (Fig. 3.1).

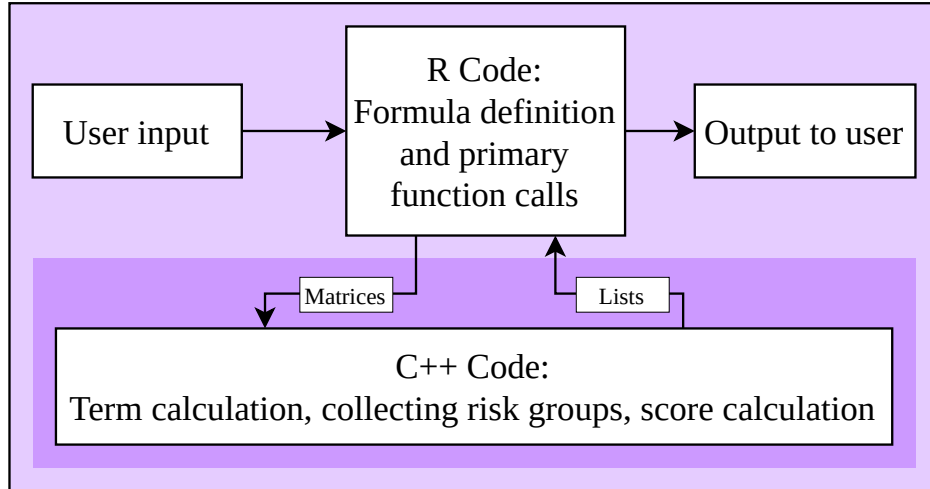


Figure 3.1: The user calls R functions, the input is converted to matrices which are processed in C++, the C++ results are converted into lists, and the R function returns the results to the user. Reproduced from previously published work³.

Every functionality in Colossus starts with an R function that passes a formula, table of covariate data, and list of control parameters to a C++ function. The general order of the main C++ functions is as follows:

1. Calculating the terms and risks for every row;
2. Calculating the log-likelihood; and
3. Taking a step to optimize the model.

The same order of calculations is repeated for each iteration. Cox proportional hazards functions have an additional step to determine the rows in each risk group. The risk group collection step is performed only once per regression. Different functions exit this cycle at different points (Fig. 3.2). The reference risk calculation exits the cycle after calculating the risk, because it does not depend on the risk groups. The hazard plotting function exits the cycle after collecting the risk groups because it does not depend on likelihood. The score calculation function exits after a single iteration, reaching an iteration limit, or converging. The modular structure of the code made it straightforward to test; each output could be used to determine the stage at which the errors originated.

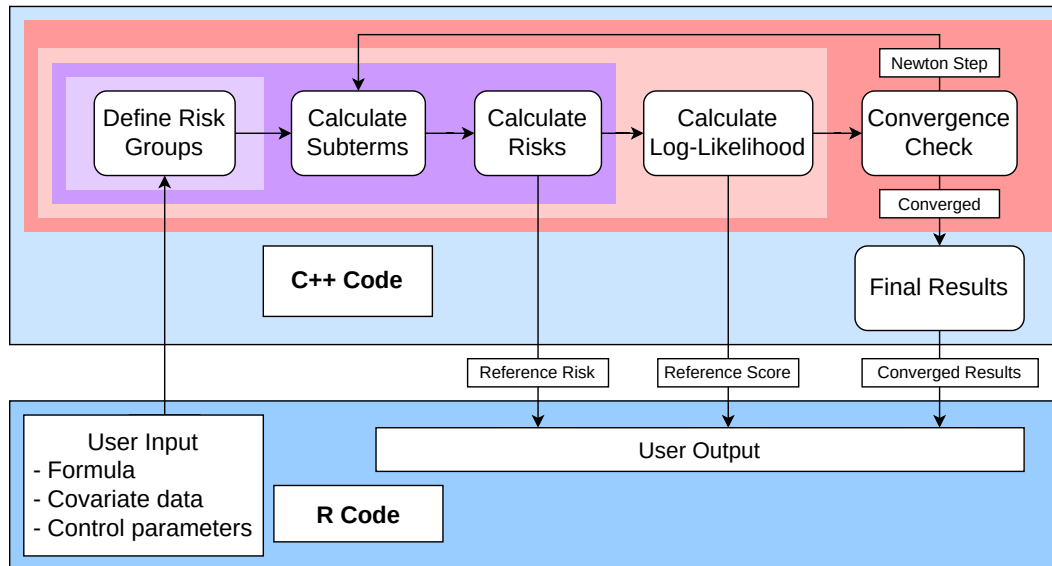


Figure 3.2: Different regression functions all follow the same sequence of C++ functions, but exit at different points (denoted by color). Reproduced from previously published work³.

3.1.1 Derivative Testing

The first stage of testing was performed to verify that the derivative calculations had been coded correctly. The calculated derivatives were compared with the empirical results by calculating different values and derivatives at points on a grid and comparing the calculated derivatives with the true change in value. At multiple points during a regression, equations are applied to calculate the term, risk, and log-likelihood values, along with the first and second derivatives. Any miscalculation would be expected to affect any subsequent calculations

(Fig. 3.3). Combinations of different subterm types and numbers of terms were applied to test the accuracy of the subterm calculations. Combinations of different numbers of terms and model formulae were applied to test whether the risk calculations were accurate, knowing that the subterm calculations were accurate. Finally, knowing that the subterm and risk calculations were correct, the log-likelihood calculations were compared with different stratification options, tied event time methods, and different numbers of tied events.

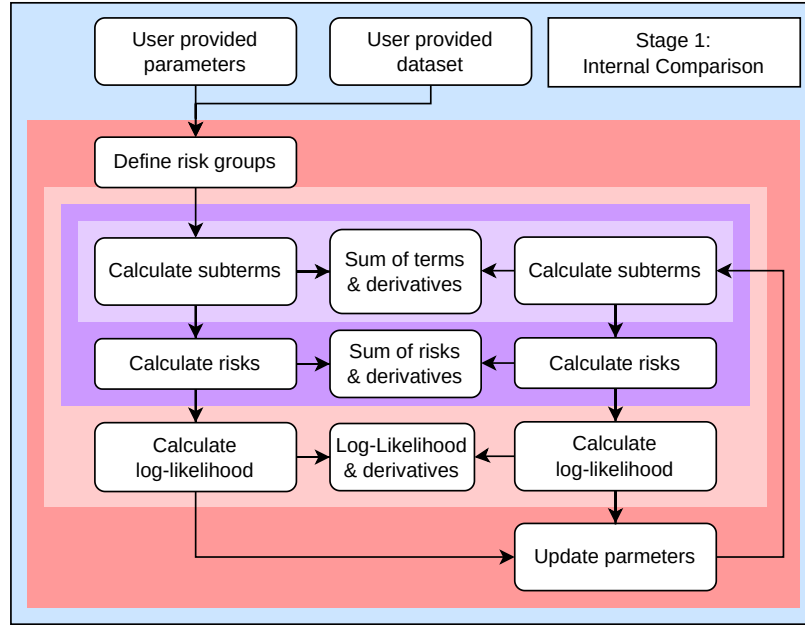


Figure 3.3: Stage 1 of testing. Colossus was run at points on a grid, and intermediate calculation values were printed to the console. Three values were compared: the sum of the terms, the sum of the relative risks, and the log-likelihood. The calculated derivatives were compared to the difference in value between grid points.

As an example, testing was repeated using data from the Veterans' Administration Lung Cancer study⁴⁹. Subterm, risk, and log-likelihood was output for a linear excess risk model. The relative risk was a function of the factored cell type ($C_1 - C_3$), Karnofsky score centered at 50 (K_{50}), and treatment group (Trt) (Eq. 3.1). The treatment parameter was varied from -0.26 to -0.17 at ten grid points to visualize the accuracy. The approximated derivatives (scatter points) were plotted against the predicted values (solid line), and in all cases, the derivative calculations in Colossus matched the true values within a narrow margin (Fig. 3.4). The same process was performed for different combinations of sub-terms, risk models,

likelihood models, and data sources.

$$HR = (1 + \beta_{trt} \cdot Trt) \exp(\beta_{c1} \cdot C_1 + \beta_{c2} \cdot C_2 + \beta_{c3} \cdot C_3 + \beta_{k50} \cdot K_{50}) \quad (3.1)$$

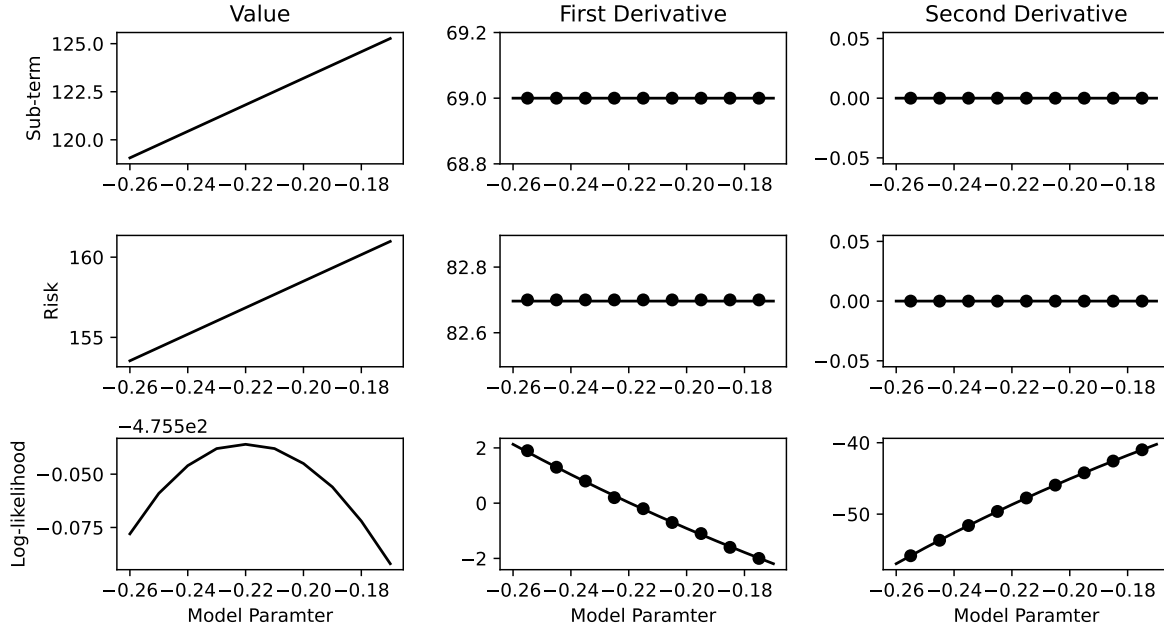


Figure 3.4: At every point on the grid, the values and derivatives were output to the console. Then the calculated values (solid line) were compared to the true derivatives (points).

3.1.2 Comparison to Existing Software

The second stage of testing was comparisons with other software. Due to the complex model formulae, the most appropriate software for comparison was 32-bit Epicure⁴. The Epicure manual contained 18 Cox proportional hazards examples and 18 Poisson examples. For each example, the optimal parameters and scores were calculated in both Epicure and Colossus and compared. A comparison of survival functions and relative risks was used to verify the risk calculation and risk group steps. Log-likelihood calculations and the converged result for each example were used to verify the score calculation and Newton step calculations (Fig. 3.5). The various residual plots not supported by Epicure were tested by comparison

with R packages, and the Fine-Gray capability was compared with weighted Cox regressions in the Survival package in R⁵. Only single-term log-linear models were tested because the Survival package in R did not support the same model options as Colossus.

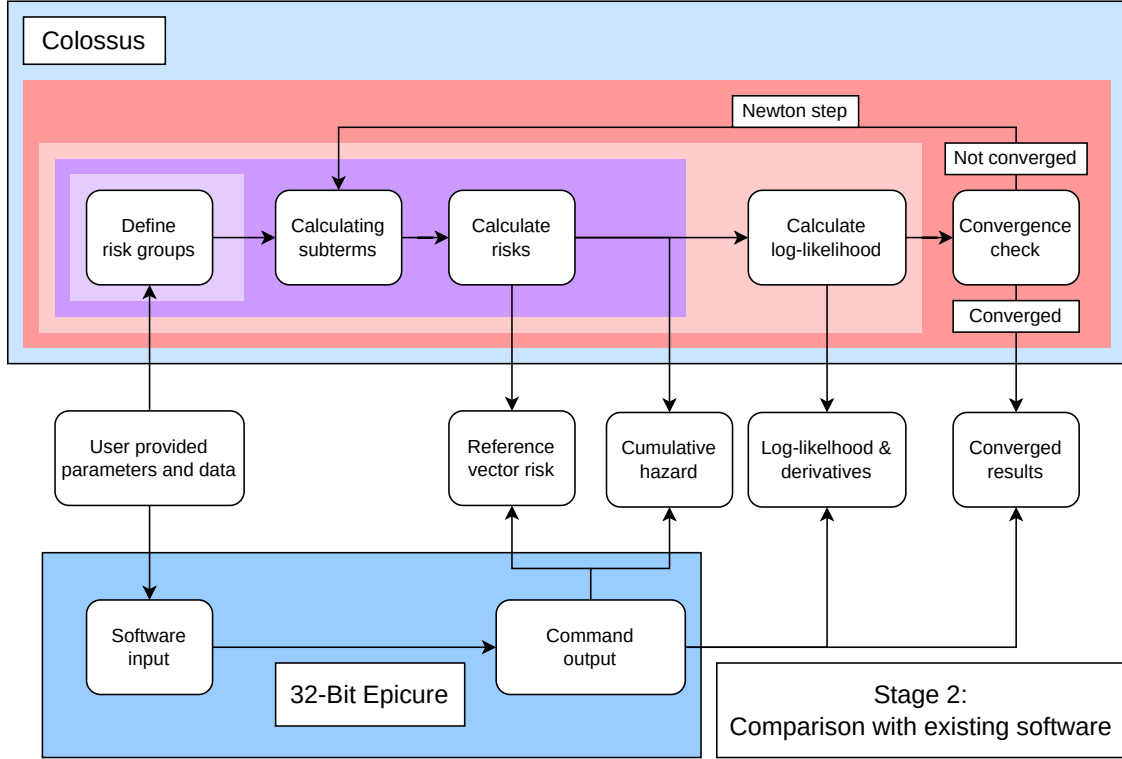


Figure 3.5: Stage 2 of testing, comparison with existing software. Examples provided by 32-bit Epicure⁴ were run in Colossus to compare the regression and plotting functions. Other residual plotting options were compared with the R survival library⁵.

As an example, the analysis performed in Example 6.12 of the 32-bit Epicure manual⁴ is presented below. The regression for output 6.19 was performed using 32-bit Epicure and Colossus. The same data was used in the example of stage 1 testing, except that in this model, the treatment effect was exponential (Eq. 3.2). The results matched with slight differences in the convergence criteria. A similar analysis was performed for the other datasets and risk models outlined in the Epicure manual, in addition to comparisons between the likelihood curves.

$$HR = \exp(\beta_{c1} \cdot C_1 + \beta_{c2} \cdot C_2 + \beta_{c3} \cdot C_3 + \beta_{k50} \cdot K_{50} + \beta_{trt} \cdot Trt) \quad (3.2)$$

Table 3.1
Example 6.12, comparison between software

	Epicure		Colossus	
Covariate	Central Estimate	Standard Error	Central Estimate	Standard Error
<i>cell_1</i>	-0.3926	0.2822	-0.39256	0.28220
<i>cell_2</i>	0.4225	0.2631	0.42251	0.26306
<i>cell_3</i>	0.7562	0.2984	0.75614	0.29844
<i>karno50</i>	-0.03119	0.0005156	-0.031193	0.00051562
<i>trt</i>	-0.2539	0.2005	-0.25392	0.20052
AIC:	961.012		961.0116	

3.1.3 Large-scale Complex Model Testing

The next stage of testing focused on verifying that Colossus could detect different strengths of effects in large datasets. This led to the most complex stage of testing, which was the analysis of simulated results. Simulated data were generated by constructing a table of subject intervals, calculating the probability of an event during each interval, and sampling a uniform distribution to assign event times. No noise or missing data were assigned, and every individual experienced the event of interest or survived the interval. To begin, the baseline hazard (BH) and relative risk (RR) equations were written as functions of survival time (t), biological sex (s_0, s_1), urban status (u_0, u_1), five work groups ($w_0, \dots, w_4$), smoking status (c_0, c_1), and cumulative dose (d). The baseline was dependent on survival time (Equation 3.3), and the relative risk was the product of the exponential sum of the

non-radiation covariates and a function of radiation exposure (Equation 3.4).

$$BH = a \cdot t^2 \cdot \exp(-b \cdot t) \quad (3.3)$$

$$RR = \exp(\beta \cdot s_1 + \sum_{i=1}^4 (\alpha \cdot w_i) + \nu \cdot u_1 + \mu \cdot c_1) \cdot f(d(t)) \quad (3.4)$$

Each individual was followed from the age of 18 years to a maximum age of 82 years. For each year of survival, individuals were exposed to a uniform random dose between 0.0 and 1.0 Gy up to 65 years of age. For each interval, a random surviving fraction was sampled, and if the predicted survival age was within the interval, the individual experienced the event and was not followed up further. The datasets were designed to contain tens of millions of rows of data.

The purpose of these tests was to establish whether Colossus could predict the true values of the parameters with different combinations of weak and strong, positive and negative correlations, and complicated non-linear subterms (Eq. 2.9). A strong negative and weak negative correlation reduced the relative risk to below 50% and within 90% of the unaltered value, respectively. Weak and strong positive correlations increased the relative risk to within 110% and above 200% of the unaltered values, respectively.

This step was divided into two series of calculations. The first treated the function of dose as an exponential and varied each covariate individually over combinations of correlation sign and strength. The second treated the function of dose as a constant plus a non-linear subterm (Equation 2.9). Because these models were more complex, only the dose was varied over combinations of correlation strengths and signs. For each risk model, a dataset of event times was generated, regressions were performed up to convergence, and the converged model parameters were compared to the true model parameters. Overall, Colossus converged within reasonable bounds of the true parameter values (Fig. 3.6). The error in the relative risk was determined to be within 0.1% and was attributable to statistical noise.

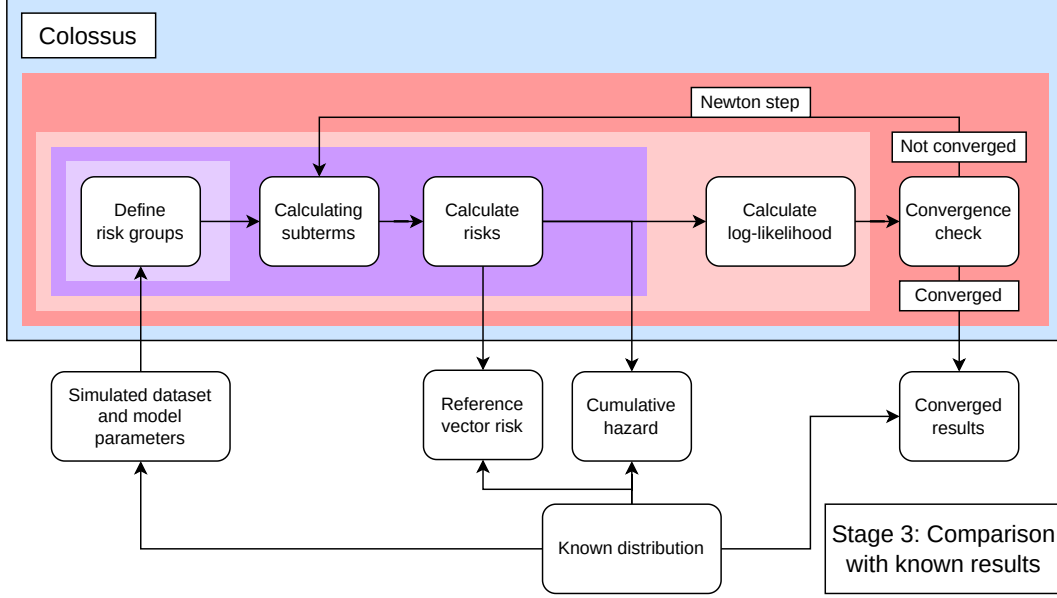


Figure 3.6: Stage 3 of testing. Rate models were simulated to create datasets and analyzed in Colossus to compare the estimated and true model parameter values. The hazard ratio distributions were compared to evaluate whether discrepancies were significant.

As an example, the results of one of the linear-threshold model tests is presented below. The true ERR increased by 0.067 per unit dose above a threshold of 10 dose, and the predicted ERR increased by 0.0652 per unit dose above a threshold of 9.91 dose (Eq. 3.5). The total relative risk was calculated for all 66 million intervals and for both the true and predicted model parameters (Fig. 3.7). Approximately 56% of the predicted hazard ratios were within 0.1% of the true value, and nearly 100% were within 1%. Similar to this example, comparisons between the model parameters and risk distributions were made for various combinations of risk models and associations. Across every combination, the error was found to be within 1% for over 99% of the data.

$$ERR = 1 + \begin{cases} \alpha \cdot (d - \beta) & , \text{ for } d > \beta \\ 0 & , \text{ otherwise} \end{cases} \quad (3.5)$$

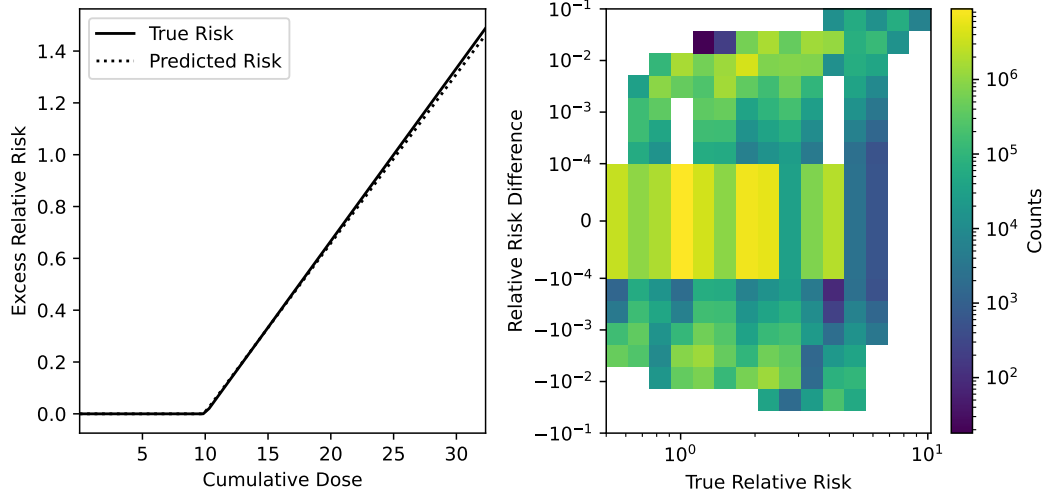


Figure 3.7: Comparison between the true and predicted excess relative risk curves, and the distribution in error in calculated hazard ratio.

3.1.4 Large-scale Multi-state Modeling

Finally, the third stage of testing was repeated with individuals at risk of multiple competing events. Every subject could leave the study due to a common or rare event. Common events occurred at an earlier age than rare events and at a much higher rate. The baseline hazard (BH) and relative risk (RR) equations for each event were written as functions of survival time (t), biological sex (s_0, s_1), economic status ($e_1, \dots, e_7$), smoking status (c_1, c_2, c_3), and cumulative dose (d). The baseline was dependent on survival time (Equation 3.3), and the relative risk was the exponential sum of the non-radiation covariates and radiation exposure (Equation 3.6).

$$RR = \exp \left(\beta \cdot s_1 + \sum_{i=2}^7 (\alpha \cdot e_i) + \sum_{i=2}^3 (\nu \cdot c_i) + \mu \cdot d \right) \quad (3.6)$$

The event rates were selected such that the presence of the common event would affect the predicted relative risk models of the rare event. The optimal model parameter values were expected to no longer match their true values; the goal of this test was to match the Survival package in R⁵. The results for one example are presented below. For this example, the Cox proportional hazard results matched within 0.01 (Table 3.2).

Table 3.2*Multi-state model simulation, Cox proportional hazards model comparison between software.*

Covariate	Survival		Colossus Difference	
	Central Estimate	Standard Error	Central Estimate	Standard Error
<i>Dose</i>	0.00304	0.00246	0.00003	0
<i>e_2</i>	0.07989	0.00946	0.00123	0
<i>e_3</i>	0.14609	0.00983	0.00239	0
<i>e_4</i>	0.1828	0.0102	0.00289	0
<i>e_5</i>	0.2422	0.01046	0.00382	0
<i>e_6</i>	0.27949	0.01073	0.00443	0
<i>e_7</i>	0.29621	0.01104	0.00473	0
<i>s_1</i>	0.37118	0.00632	0.00518	0
<i>c_2</i>	0.13551	0.00801	0.00201	0
<i>c_3</i>	0.49171	0.00651	0.00763	0

3.2 Big Data Optimization

In addition to being able to combine different types of risk models and likelihood functions, Colossus was designed to handle large amounts of data. Throughout the design process, benchmarking and testing were focused on the performance with big data. This process is exemplified by the implementation of regression considering covariate uncertainty in Colossus, which is described in greater detail in Chapter 7.

In short, the Million Person Study aims to perform risk analysis taking into account the uncertainty in annual dose estimates. Instead of using a single expected annual dose, uncertainties throughout the detailed dosimetric analysis should be considered. The details presented in NCRP Commentary 34²³ indicated that one next step in considering covariate uncertainty is often to generate multiple realizations of each annual organ dose, analyze each realization, and then combine the results. For any outcome and organ dose analyzed in this

way, obtaining a single parameter estimate would require running thousands of regressions. Even if each regression requires on the order of seconds, the final result may take hours. Optimizing Colossus for multiple realization analyses was beyond optimizing each realization individually and provided an example of how highly specific analyses were implemented and optimized in Colossus.

First, an outline of how multiple realizations are analyzed using standard software is provided. A hypothetical dataset would be composed of rows with time at risk, risk factors, and event indicators (Table 3.3). Multiple realizations of at least one risk factor exposure would be generated before running the analysis. Importantly, the time each row was at risk and the event columns would be constant between regressions, but the risk factors would change between realizations.

Table 3.3
Example of multiple realization data.

ID	Entry Age	Exit Age	Event	Dose 1	Dose 2	...
1	18	19	0	0.65	0.66	...
1	19	20	0	0.36	0.77	...
1	20	21	1	0.57	0.33	...
2	30	31	0	0.62	0.42	...
2	31	32	0	0.35	0.59	...
...	...	...	...	...	...	...

For a standard function call, the dataset is loaded, the analysis is run, and the results are recorded. A diagram of the functional flow within Colossus is provided below (Fig. 3.8); however, other comparable software follow a similar series of steps.

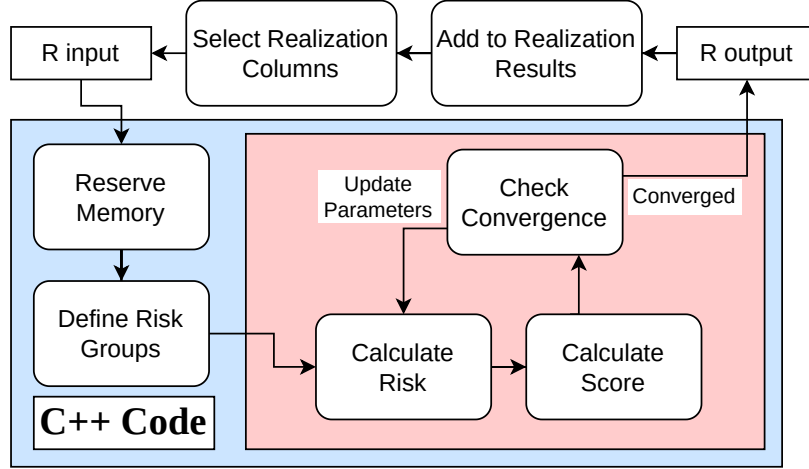


Figure 3.8: Standard function call with one realization analyzed per call. By running each regression separately, several identical steps were repeated thousands of times.

Colossus uses the Rcpp package⁵⁰ to run analyses with a combination of R and C++ functions and performs calculations in parallel using OpenMP⁵¹. If the analysis had been performed in Colossus, it would have started by processing the data in R, then the R variables would have been converted to C++ variables, memory would have been allocated for the matrices used for risk and score calculation, and iterations would have been run as previously described until the regression converged and the results were exported to R.

For this analysis, Cox regression was used, which would add an additional step to pre-determine rows at risk at each event time. Reserving memory and predetermining rows at risk are steps taken in Colossus to save time over large regressions. The sizes of the required matrices would be determined by the number of rows in the dataset and the number of covariates in the model. The rows at risk would be based on the event times and intervals for each row. None of these values would have changed between iterations; memory could be allocated, and the risk groups solved once per regression instead of once per iteration. Another important detail was that the analysis switches from R to C++ and back once, not multiple times per regression. The process of converting variables between R and C++ was time consuming. By designing the functions with these principles in mind, each independent regression function was optimized.

Optimizing each regression required identifying the steps that were constant over iter-

ations. Similarly, optimizing multiple regressions required identifying the steps that were constant over each realization. In the example given, the intervals at risk at each event time and the structure of the model would not change between realizations. The same memory allocations and risk group calculations would have been repeated for every realization. A truly optimized analysis would have performed these steps once per analysis, not once per realization. This was achieved by moving the realization column update into the C++ code block (Fig. 3.9). In doing so, the analysis followed the same optimization rules as those for single realization functions. The steps were repeated only as often as necessary, and the code did not switch excessively between R and C++.

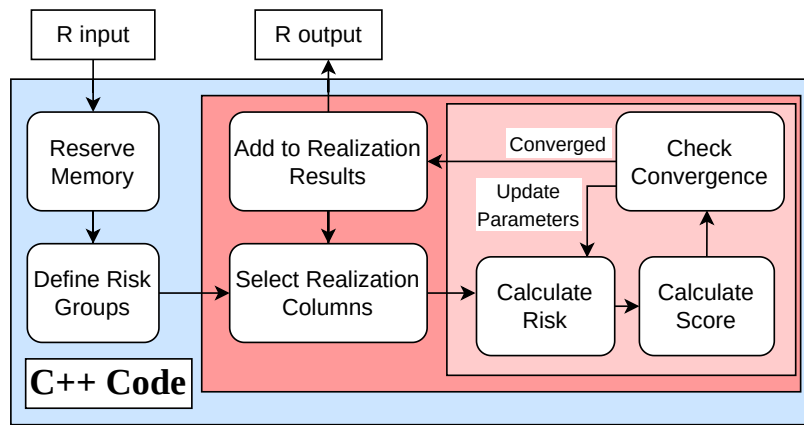


Figure 3.9: Function call with multiple realizations analyzed per call. Switching to one function call per analysis allowed fewer steps to be repeated.

The same steps were performed for various expected types of analyses. Another example was for a user trying multiple initial parameter sets. In many cases, a user may want to run several short regressions at different starting points and then run more iterations at the best point. So, Colossus was designed to accept multiple initial parameter lists. Within the same C++ function, Colossus runs a regression at each starting point, selects the best parameter set, and runs more iterations. This shortcut is important for risk models that may have several local optima. Without testing multiple different starting points, it may not be possible to differentiate local optima from global optima (Fig. 3.10). As the Million Person Study applies more complicated models to large cohorts, these simple approaches can save significant time in the long run.

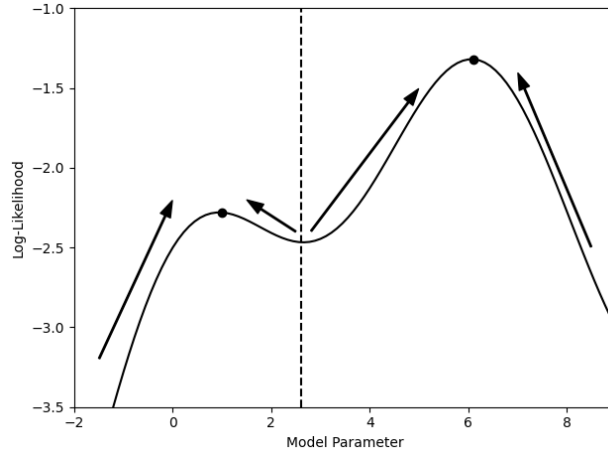


Figure 3.10: Gradient-based optimization algorithms do not distinguish between local optima. A regression starting below 2.6 converges to the wrong point. A common solution is to perform multiple regressions at different starting points.

3.3 Hardware Usage and Software Performance

Colossus runs calculations in parallel when possible, using data stored in Eigen matrices⁵⁰ or standard C++ vectors. Many of the calculations either assign values to matrix columns independently or assign a sum of values to a vector element. Both tasks were written as “for loops” in parallel. Each thread was assigned a calculation to a different column or vector element, so multiple calculations could be performed simultaneously. In general, the derivative calculations were performed as independent column calculations (Fig. 3.11) and the Cox proportional hazard log-likelihoods were calculated as batch sums over each risk group (Fig. 3.12). This type of analysis with OpenMP used shared-memory parallelism, which meant that these operations could be performed simultaneously without making copies of the data.

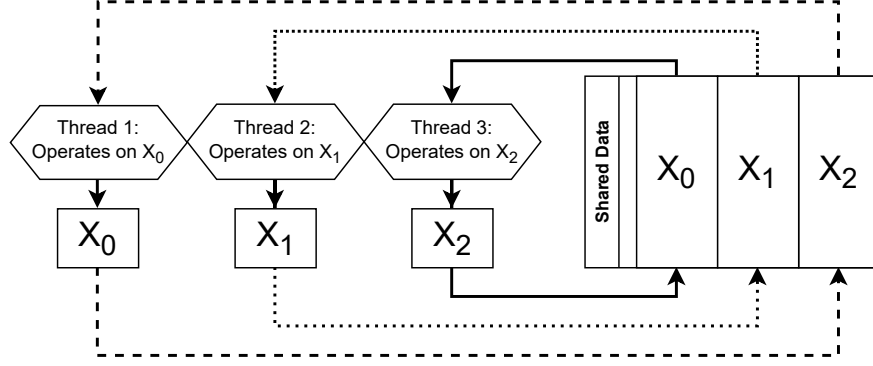


Figure 3.11: Basic parallelization method applied for independent column calculations. The equation applied to each column was dependent on a shared constant matrix, so values were assigned to each column simultaneously.

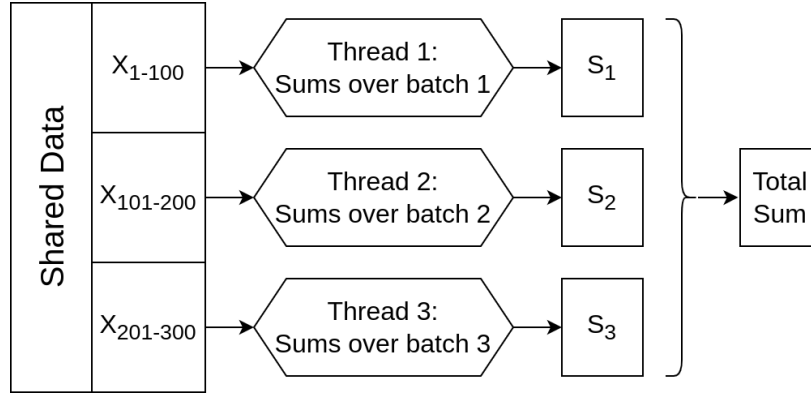


Figure 3.12: Basic parallelization method applied for accumulated sums. Each thread accumulated a sum over independent batches, and then the intermediate results were combined to calculate the sum over the entire dataset.

The sizes of the created matrices and the number of parallel loops are functions of the values known to users. These values are the number of rows (n_r), number of risk groups (n_g), number of covariates (n_c), and number of parallel cores (n_p). The term and risk calculations define matrices on the order of $n_r \times n_c$ and $n_r \times n_c^2$. The risk group collection defines vectors on the scale of $1 \times n_g$. Log-likelihood calculations for Cox proportional hazards models define matrices on the order of $n_g \times n_c$ and $n_g \times n_c^2$. The memory required for different sections was expected to increase with the number of rows and risk groups and potentially increase quadratically with the number of covariates.

The term and risk calculations were parallelized over the matrix columns, so the time

duration for these steps was expected to be proportional to the ratio of $n_r \cdot n_c^2/n_p$. The risk group collection step filters a vector with length n_r and was parallelized over risk groups, so the time duration was expected to be proportional to the ratio of $n_g \cdot n_r/n_p$. The log-likelihood calculation for Cox proportional hazards models was parallelized over columns and risk groups, so the step duration was expected to be proportional to the ratio of $n_g \cdot n_c^2/n_p$.

To visualize these effects, benchmarking was performed to compare the time requirements among regressions with differences in data scale and model complexity. The intention was to run regressions with differences in the number of intervals, risk groups, model covariates, and cores used. First, a simulated dataset was randomly generated with approximately 42 million intervals. Subsets of the dataset were generated by randomly sampling different percentages of intervals. Each subset was further divided into two versions with different numbers of risk groups by rounding the time columns to one or two decimal places. Each version was analyzed using Cox regression, and the original subset was analyzed using Poisson regression. Each regression was repeated for different numbers of covariates and cores. Changes in time by covariates were compared at 1, 16, and 128 cores. Changes in time by core were compared at 1, 4, and 8 covariates. In total, six different numbers of intervals were tested, ranging from 4 thousand to 40 million, and 216 regressions were performed for each dataset size.

Benchmarking was performed using the Kansas State University Beocat computing system with 128 core Warlock nodes. The regressions were split over six jobs, one for each interval sampling, so that regressions with the same number of intervals used the same hardware. Memory use was measured by the peak memory reported per job and tables of memory use at minute intervals. Regressions that took less than 1 minute could not be accurately described, so detailed memory use analysis was not performed for analyses using less than 1 million rows. Additionally, memory usage was found to vary significantly during parallel calculations. This meant that the observed memory use was not likely to fully capture the range of memory use during each regression.

The first expected trend was that the regression duration would increase with the number of covariates and the size of the dataset. The plots of regression duration using one core show the influence of dataset size, risk groups, and number of covariates (Fig. 3.13). At

4,226 intervals, the regression duration was too varied to form a meaningful trend. Above 4,200 intervals, the duration increased with the number of covariates, with a slight curvature. One would expect the Cox regressions to take longer than the Poisson regressions, and the Cox regressions with more risk groups to take longest. Above 42,000 rows this trend was consistent. As the difference in risk groups increased, the Cox regressions with more risk groups took significantly longer than those with fewer risk groups. These differences in duration were due to the number of calculations performed. Cox likelihood calculations are split by risk groups, while Poisson likelihoods are not. This meant that Cox regressions required more calculations than Poisson regressions, and the difference increased with the number of risk groups. The slight curvature was believed to be due to the calculation of second derivatives, which were expected to increase proportionally to the number of covariates squared.

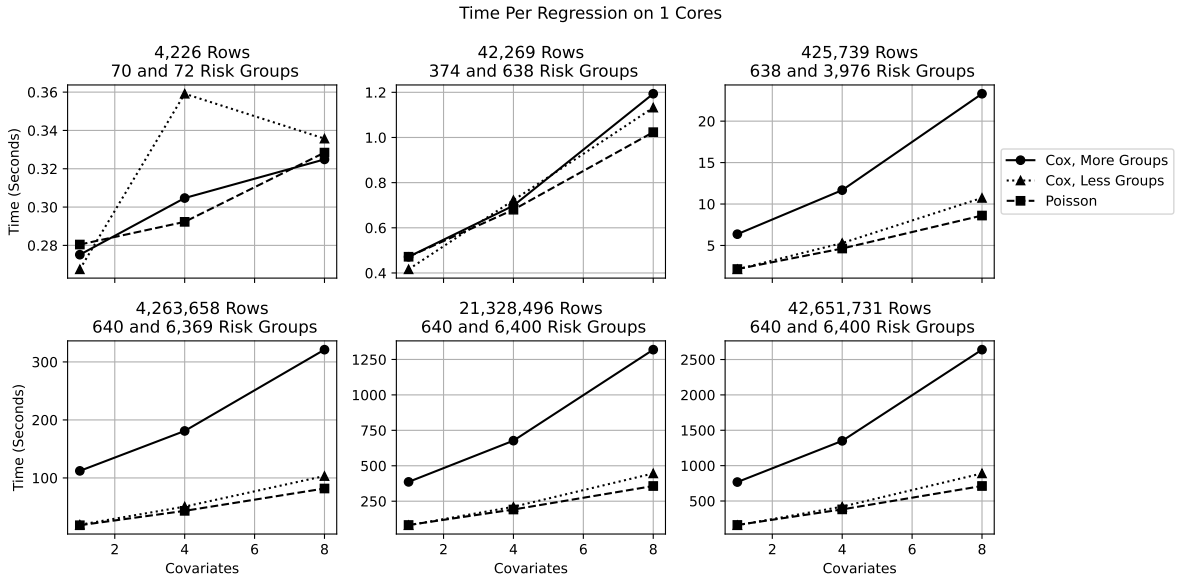


Figure 3.13: Regression duration using one core across different regressions methods and data size. In general the duration increased with number of covariates, risk groups, and intervals.

The second expected trend was the change in regression duration with the number of cores and regression model at one covariate. Increasing the number of cores was expected to decrease the time of parallelized calculations and increase the preparation time of the

calculations. The first benchmark was at one covariate (Fig. 3.14). Poisson regressions with a single covariate have no parallelized calculations, so duration was expected to only increase. Cox regressions with one covariate parallelize likelihood calculations over risk groups, so the duration was expected to increase based on the number of risk groups. This trend was consistent for all datasets. In general the Poisson durations increased with number of cores, the Cox regressions with fewer risk groups had slight improvements, the Cox regressions with more risk groups had more improvement, and duration generally increased at 128 cores. For each dataset size and regression method, there was an optimal number of cores. Poisson regressions were generally optimized at one core, and the Cox regressions were generally optimized between 8 to 32 cores.

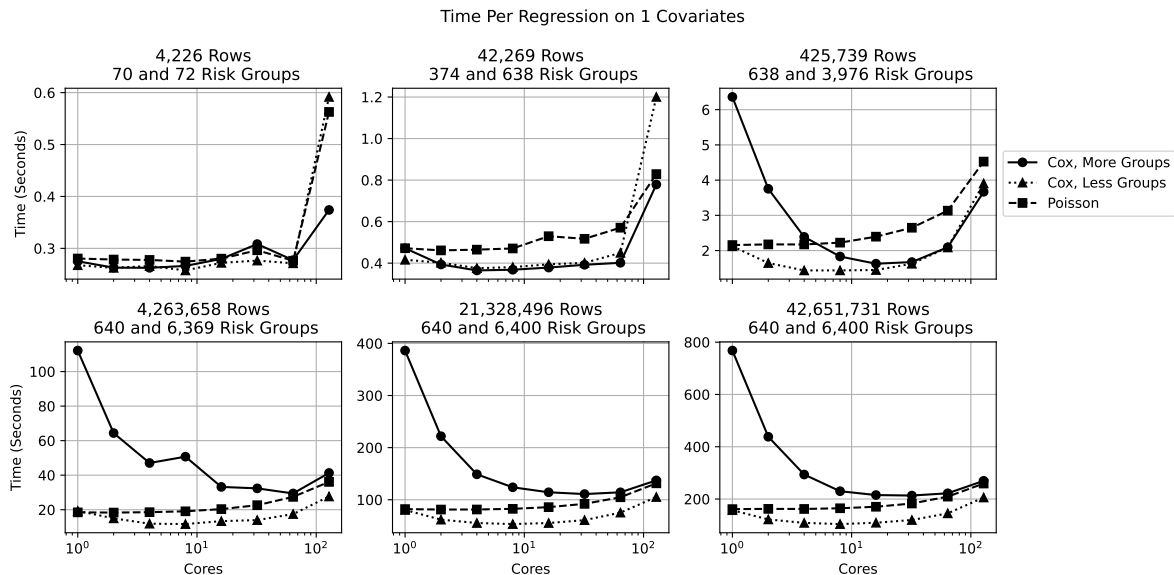


Figure 3.14: Regression duration using one covariate across different regressions methods and data size. The Poisson regressions did not use parallelized calculations, so they did not benefit from additional cores. Cox regressions parallelize over risk groups, so they saw benefit based on the number of risk groups.

The final expected trend was the change in regression duration with the number of cores and the regression model with eight covariates. As the number of covariates increases, the number of parallelized calculations increases for each regression model. This meant that every regression method was expected to improve with an increasing number of cores based on the number of calculations. Contrary to the benchmarking at one covariate, the Poisson

regression duration was expected to be optimized above one core. The Cox regressions were expected to be optimized at higher core numbers due to the increase in parallelized calculations. This trend was consistent with the data (Fig. 3.15). Even with as few as 4,200 rows, the Poisson regressions improved with increasing covariates. For regressions with more than one million intervals, the improvement continued up to 128 cores. In general, the reduction in regression duration with increasing number of cores improved as the number of covariates increased.

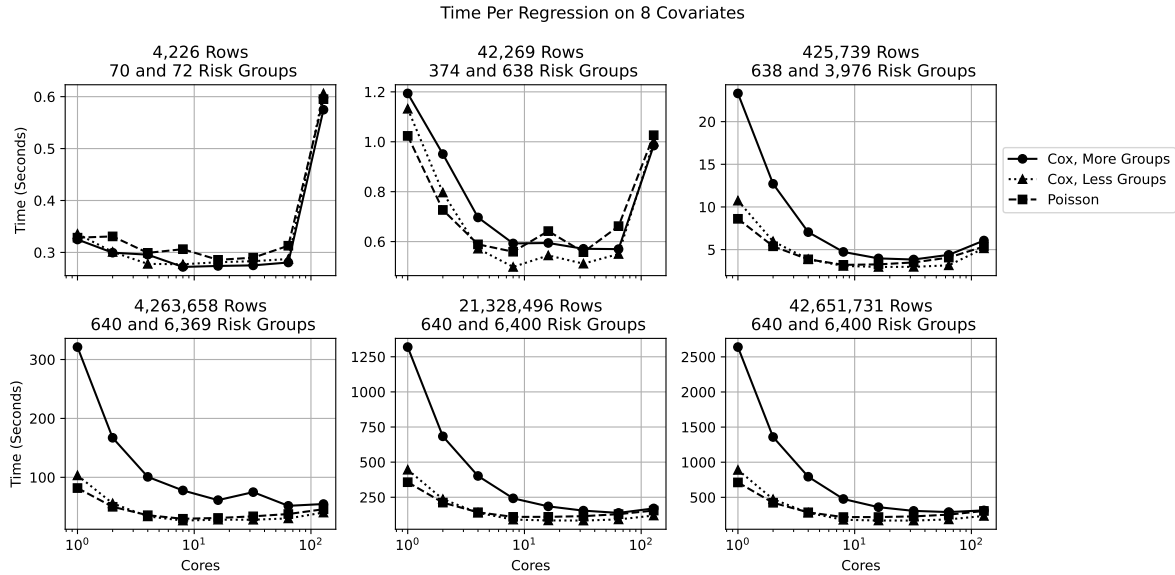


Figure 3.15: Regression duration using eight covariates across different regressions methods and data size. Due to the increased number of parallel calculations, every regression method benefited from additional cores to some degree.

The memory required was expected to increase with the number of intervals, risk groups, and covariates in the model. These values control the size of the matrices used to store the calculations throughout the regression. Another important factor that increased memory usage was the use of temporary variables to store intermediate values. The number of temporary values defined at once depended on the number of parallel calculations performed simultaneously. This meant that the memory use was dependent on a baseline level dependent on the size of the matrices storing the final results and a variable level that increased with the number of parallel calculations. The memory benchmarking only observed memory use at specific times, which meant that peak memory use for every regression was likely not

observed. Baseline memory could be approximated by running regressions on a single core, so very few temporary variables would be defined at once (Fig. 3.16). For a single core, the memory use ranged from 9 to 16 GB for the four million interval data and from 50 to 90 GB for the 42 million interval data. In general, the observed memory use increased with the number of covariates and intervals.

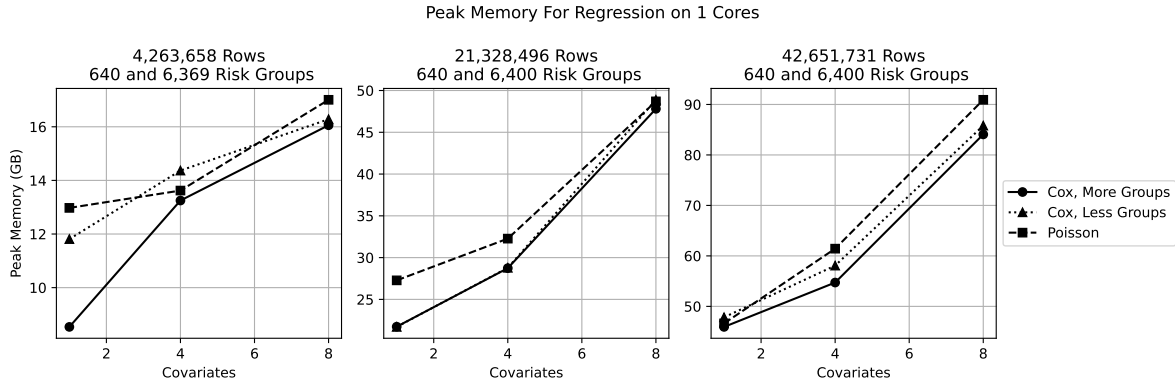


Figure 3.16: Peak observed memory use for regressions on a single core by number of covariates. Due to the low number of parallel calculations, the baseline memory required could be measured. Memory required generally increased with the size of the dataset and number of covariates.

The effect of temporary variables can be approximated by running regressions with multiple covariates with increasing numbers of cores. For each regression, the minimum and maximum observed memory use were recorded and plotted (Fig. 3.17). For 42 million interval data, the average minimum memory use was consistently on the scale of 50 to 90 GB, while the average maximum memory use above 32 cores increased up to 140 to 170 GB.

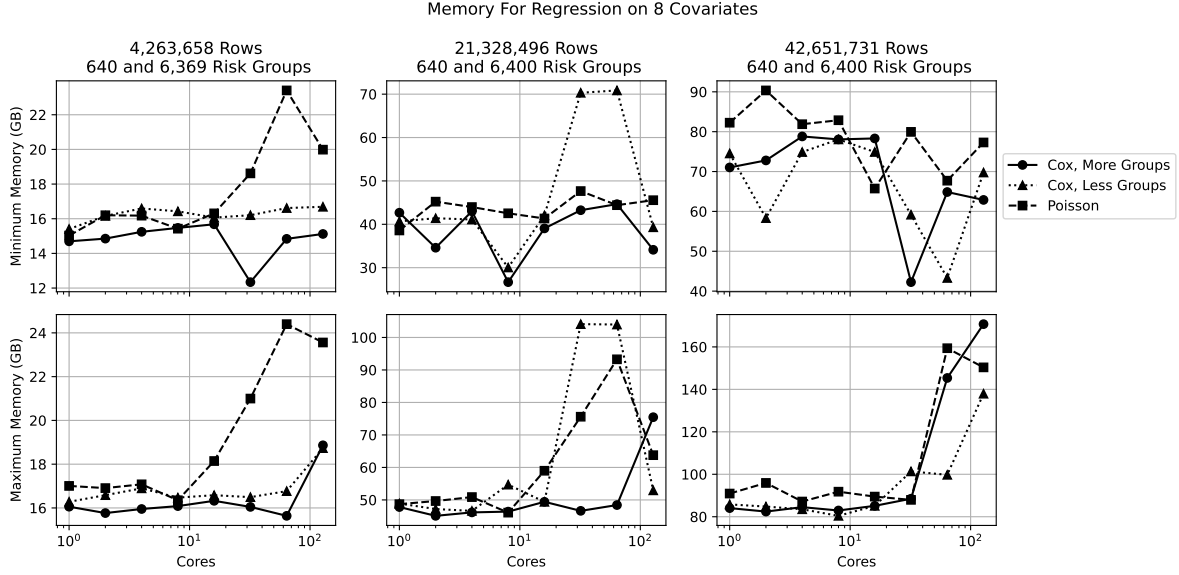


Figure 3.17: Observed range of memory use for regressions on eight covariates by numbers of cores. Due to the number of parallel calculations, the range of observed memory use increased drastically with more cores.

The exact regression durations varied significantly among the different systems and regression problems. Differences in the number of iterations and size of the data influenced which regression steps took the longest duration. In the benchmarking performed in Chapter 7, the regressions took approximately 8 seconds, most of which was spent processing the data. Other regressions performed during testing spent significantly more time in the iteration stage than in the processing stage. The general rule of thumb still applies: increasing the complexity of the model, the size of the data, or the fidelity of the event times generally increases the time and memory required. Conversely, reducing the model complexity, rounding event times, and analyzing subsets of the data would generally decrease the time and memory required.

Chapter 4

An Example of Standard Use

To demonstrate the use of Colossus[†], examples of typical Cox proportional hazards and Poisson model fitting are presented here. For further discussion of Colossus capabilities and example code, the reader is directed to the variety of vignettes included with the package at the CRAN³⁵ and GitHub repositories³⁶.

This analysis used data from the Veterans' Administration Lung Cancer study⁴⁹. This dataset was selected for its relevance to cancer research, availability to any R user, and use in the 32-bit Epicure manual. This dataset was available in R through the Survival package⁵. The Cox proportional hazards analysis mirrored Example 6.12 in the 32-bit Epicure manual⁴, so transformations were applied to convert the version in R to the version included with 32-bit Epicure (Listing 1). The following transformations were made:

1. Changing the 93rd Karnofsky performance score entry from 30 to 20;
2. Converting the treatment indicator to 0/1; and
3. Recoding the cell types as integers.

[†]This chapter is reproduced from previously published work³

```

library(Colossus)
library(survival)
library(dplyr)
data(cancer, package = "survival")
df <- veteran
# Make the same adjustments as Epicure example 6.5
karno <- df$karno
karno[93] <- 20
df$karno <- karno
df$trt <- df$trt - 1
df$trt <- as.integer(df$trt == 0)
cell_string <- df$celltype
cell <- case_when(
  cell_string == "squamous" ~ 1,
  cell_string == "smallcell" ~ 2,
  cell_string == "adeno" ~ 3,
  cell_string == "large" ~ 0
)
df$cell <- cell

```

Listing 1: Preparing Lung Cancer Data

Next, the covariates used for modeling were prepared (Listing 2). The model measured the effects of cell type, Karnofsky performance score, and treatment on different measures of risk. These risk models were functions of the factored cell type ($C_1 - C_3$), the Karnofsky score centered at 50 (K_{50}), and the treatment group (Trt). The factoring of cell type in this analysis was performed using the Colossus function *factorize*, which divided the columns in a list into factored columns and returned the final updated table.

```

df$karno50 <- df$karno - 50
# Convert the cell column into multiple factor columns
fcols <- c("cell")
val <- Colossus::factorize(df, fcols) # Colossus function
df <- val$df

```

Listing 2: Preparing Lung Cancer Covariates

4.1 Cox Proportional Hazards Regression

Two Cox proportional hazards models were used in these examples. The first applied treatment as an exponential effect (Eq. 4.1), and the second considered treatment to be a linear effect (Eq. 4.2). Note that treatment was a binary covariate, so both models were expected to produce the same hazard ratio central estimates.

$$HR_1 = \exp(\beta_{c1} \cdot C_1 + \beta_{c2} \cdot C_2 + \beta_{c3} \cdot C_3 + \beta_{k50} \cdot K_{50} + \beta_{trt} \cdot Trt) \quad (4.1)$$

$$HR_2 = (1 + \beta_{trt} \cdot Trt) \exp(\beta_{c1} \cdot C_1 + \beta_{c2} \cdot C_2 + \beta_{c3} \cdot C_3 + \beta_{k50} \cdot K_{50}) \quad (4.2)$$

First, the survival interval was specified (Listing 3). The default survival definition in Colossus was a counting interval, where every row had a starting time, ending time, and status indicator. Left-truncated and right-truncated survival problems are special cases in which every row has the same starting or ending point. This analysis only had a measure of survival duration, so the start time was considered shared. The interval end time was set to the survival time column, and the start time was set to `%trunc%`, which Colossus used to denote a shared end-point. The event indicator was the censoring status.

```
t0 <- "%trunc%"  
t1 <- "time"  
event <- "status"
```

Listing 3: Left Truncated Survival Interval

The next step was to define the model equation (Section 2.1). Every element in the equation had a column name, subterm type, term number, and initial parameter estimate. The default was a single term, multiplicative model, and exponential subterms (Eq. 4.1). This equation matches the typical Cox proportional hazards model. These values were entered into the Colossus functions via vectors (Listing 4). Only a vector of column names and a vector of subterm types were used in this case.

```
names <- c(
  "cell_1", "cell_2", "cell_3",
  "karno50", "trt"
)
tform_1 <- c(
  "loglin", "loglin", "loglin",
  "loglin", "loglin"
)
tform_2 <- c(
  "loglin", "loglin", "loglin",
  "loglin", "plin"
)

```

Listing 4: Cox Regression Model Definition

Cox regression functions in Colossus generally require a table of data, interval columns, an event column, and model definition vectors. Other details controlling the regression could have been provided through different control lists. In this analysis, the remaining arguments were set to default values (Listing 5). The regression functions returned a list of details regarding the optimized model and convergence status. The results were summarized and printed in a concise table using the *Interpret_Output* function (Tables 4.1-4.2).

```
e1 <- Colossus::RunCoxRegression(
  df, t0, t1, event,
  names = names, tform = tform_1
)
e2 <- Colossus::RunCoxRegression(
  df, t0, t1, event,
  names = names, tform = tform_2
)
Colossus::Interpret_Output(e1)
Colossus::Interpret_Output(e2)

```

Listing 5: Cox Regression Running and Printing

Table 4.1
Lung cancer data, exponential hazard ratio model results.

Covariate	Subterm	Central Estimate	Standard Error
<i>cell_1</i>	log-linear	-0.39259	0.28220
<i>cell_2</i>	log-linear	0.42251	0.26306
<i>cell_3</i>	log-linear	0.75616	0.29844
<i>karno50</i>	log-linear	-0.03119	0.00516
<i>trt</i>	log-linear	-0.25394	0.20052
-2*Log-Likelihood: 951.01			
AIC: 961.01			

Table 4.2
Lung cancer data, linear hazard ratio model results.

Covariate	Subterm	Central Estimate	Standard Error
<i>cell_1</i>	log-linear	-0.39259	0.28220
<i>cell_2</i>	log-linear	0.42251	0.26306
<i>cell_3</i>	log-linear	0.75616	0.29844
<i>karno50</i>	log-linear	-0.03119	0.00516
<i>trt</i>	product-linear	-0.22426	0.15555
-2*Log-Likelihood: 951.01			
AIC: 961.01			

In both analyses, solutions were found with the same AIC, 961.01, and the same central estimates for the hazard ratio (Table 4.3). The differences in the Wald confidence intervals originated from the assumptions violated by using a linear model. Wald confidence intervals assume that the distribution of the maximum likelihood estimators is approximately normal, which has been found to be violated by risk models accounting for additive effects⁵². True likelihood-based confidence intervals could have been calculated to correct for this approxi-

mation.

Table 4.3

Lung cancer data, hazard ratio model comparison (95% confidence interval).

Covariate	First Model HR	Second Model HR	HR Difference
<i>cell_1</i>	0.675 (0.388 , 1.174)	0.675 (0.388 , 1.174)	0.000 (0.000 , 0.000)
<i>cell_2</i>	1.526 (0.911 , 2.555)	1.526 (0.911 , 2.555)	0.000 (0.000 , 0.000)
<i>cell_3</i>	2.130 (1.187 , 3.823)	2.130 (1.187 , 3.823)	0.000 (0.000 , 0.000)
<i>karno50</i>	0.969 (0.960 , 0.979)	0.969 (0.960 , 0.979)	0.000 (0.000 , 0.000)
<i>trt</i>	0.776 (0.524 , 1.149)	0.776 (0.471 , 1.081)	0.000 (0.053 , 0.069)

4.2 Poisson Rate Regression

An important difference between the Cox proportional hazards model regression and Poisson model regression was the representation of time. Cox proportional hazards models used time interval endpoints, and Poisson models used the total interval duration or person-time. In this analysis, the survival time column represented the person-time at risk in each interval. Another difference was that Poisson models often include an intercept term to account for the average event rate. In Cox regression, a constant multiple is canceled out during the likelihood calculation. For Colossus functions, an intercept column can be added using the column name *CONST* to add a column of ones.

$$\lambda_1 = \exp(\beta_{c1} \cdot C_1 + \beta_{c2} \cdot C_2 + \beta_{c3} \cdot C_3 + \beta_{k50} \cdot K_{50} + \beta_{trt} \cdot Trt + \beta_{Const}) \quad (4.3)$$

$$\lambda_2 = (1 + \beta_{trt} \cdot Trt) \exp(\beta_{c1} \cdot C_1 + \beta_{c2} \cdot C_2 + \beta_{c3} \cdot C_3 + \beta_{k50} \cdot K_{50} + \beta_{Const}) \quad (4.4)$$

Another optional input for the functions in Colossus was the initial parameter estimate. Suppose that the Cox proportional hazards model regressions have already been run, and one wanted to use the optimized parameters as initial guesses for the parameters in the Poisson models. The intercept parameter, β_{Const} , was estimated using the average event rate. The initial parameter guess was a vector in the same order as the column name and subterm type vectors (Listing 6).

```
names <- c(
  "CONST", "cell_1", "cell_2",
  "cell_3", "karno50", "trt"
)
tform_1 <- c(
  "loglin", "loglin", "loglin",
  "loglin", "loglin", "loglin"
)
tform_2 <- c(
  "loglin", "loglin", "loglin",
  "loglin", "loglin", "plin"
)
avg_event <- sum(df$status) / sum(df$time)
print(log(avg_event))
# -4.868916
a_n1 <- c(-4.87, -0.39, 0.42, 0.76, -0.031, -0.25)
a_n2 <- c(-4.87, -0.39, 0.42, 0.76, -0.031, -0.22)
```

Listing 6: Poisson Regression Model Definition

The Poisson regression functions required nearly identical inputs, except that a single person-year column was used instead of the two interval endpoint columns (Listing 7). In this example, the optional parameter starting values were also provided. The Poisson regression functions returned lists of results similar to the Cox proportional hazards functions, and the

same function was used to summarize the results (Tables 4.4-4.5).

```
e1 <- Colossus::RunPoissonRegression(
  df, t1, event,
  names = names, tform = tform_1, a_n = a_n1
)
e2 <- Colossus::RunPoissonRegression(
  df, t1, event,
  names = names, tform = tform_2, a_n = a_n2
)
Colossus::Interpret_Output(e1)
Colossus::Interpret_Output(e2)
```

Listing 7: Poisson Regression Running and Printing

Table 4.4

Lung cancer data, exponential Poisson model results.

Covariate	Subterm	Central Estimate	Standard Error
<i>CONST</i>	log-linear	-4.47422	0.23938
<i>cell_1</i>	log-linear	-0.36859	0.27191
<i>cell_2</i>	log-linear	0.41982	0.25823
<i>cell_3</i>	log-linear	0.71318	0.29037
<i>karno50</i>	log-linear	-0.02969	0.00482
<i>trt</i>	log-linear	-0.19739	0.19304
-2*Log-Likelihood: 393.72			
AIC: 149.72, BIC: 423.24			

Table 4.5

Lung cancer data, linear excess relative risk Poisson model results.

Covariate	Subterm	Central Estimate	Standard Error
<i>CONST</i>	log-linear	-4.47422	0.23938
<i>cell_1</i>	log-linear	-0.36859	0.27191

<i>cell_2</i>	log-linear	0.41982	0.25823
<i>cell_3</i>	log-linear	0.71318	0.29037
<i>karno50</i>	log-linear	-0.02969	0.00482
<i>trt</i>	product-linear	-0.17913	0.15846
-2*Log-Likelihood: 393.72			
AIC: 149.72, BIC: 423.24			

Similar to the Cox proportional hazards model results, the hazard ratios due to the effect of each covariate were isolated. Both Poisson analyses found solutions with the same AIC, 149.72, and the same central estimates for the hazard ratio (Table 4.6). Once again, the treatment effects did not have matching confidence intervals due to Wald confidence interval assumptions.

Table 4.6
Lung cancer data, Poisson rate model comparison (95% confidence interval).

Covariate	First Model HR	Second Model HR	HR Difference
<i>CONST</i>	0.011 (0.007 , 0.018)	0.011 (0.007 , 0.018)	0.000 (0.000 , 0.000)
<i>cell_1</i>	0.692 (0.406 , 1.179)	0.692 (0.406 , 1.179)	0.000 (0.000 , 0.000)
<i>cell_2</i>	1.522 (0.917 , 2.524)	1.522 (0.917 , 2.524)	0.000 (0.000 , 0.000)
<i>cell_3</i>	2.040 (1.155 , 3.605)	2.040 (1.155 , 3.605)	0.000 (0.000 , 0.000)
<i>karno50</i>	0.971 (0.962 , 0.980)	0.971 (0.962 , 0.980)	0.000 (0.000 , 0.000)
<i>trt</i>	0.821 (0.562 , 1.198)	0.821 (0.510 , 1.131)	0.000 (0.052 , 0.067)

Finally, the hazard ratio estimates between the Cox proportional hazards and Poisson models were compared. Although the scores could not be compared, the hazard ratio estimates were similar (Table 4.7). This illustrated how the same measurement of risk could be solved using multiple models.

Table 4.7

Lung cancer data, Cox proportional hazards and Poisson exponential relative risk model comparison (95% confidence interval).

Covariate	Poisson Model HR	Cox Model HR	HR Difference
<i>cell_1</i>	0.692 (0.406 , 1.179)	0.675 (0.388 , 1.174)	0.016 (0.018 , 0.005)
<i>cell_2</i>	1.522 (0.917 , 2.524)	1.526 (0.911 , 2.555)	0.004 (0.006 , 0.031)
<i>cell_3</i>	2.040 (1.155 , 3.605)	2.130 (1.187 , 3.823)	0.090 (0.032 , 0.218)
<i>karno50</i>	0.971 (0.962 , 0.980)	0.969 (0.960 , 0.979)	0.001 (0.002 , 0.001)
<i>trt</i>	0.821 (0.562 , 1.198)	0.776 (0.524 , 1.149)	0.045 (0.039 , 0.049)

4.3 Effect Modification

The final example demonstrates how to apply and interpret effect modification. For this example, the Cox proportional hazards analysis was expanded by splitting the effect of treatment by cell type. The goal of this type of analysis was to compare the effects among strata to determine whether the effect was independent of the strata. So, this analysis determined whether the treatment effect differed among cell types.

$$HR_3 = \left(1 + \sum_{i=0}^3 (\beta_{trt \times C_i}) \cdot Trt \cdot C_i \right) \exp(\beta_{c1} \cdot C_1 + \beta_{c2} \cdot C_2 + \beta_{c3} \cdot C_3 + \beta_{k50} \cdot K_{50}) \quad (4.5)$$

The analysis started by defining the interaction columns and preparing the new equation vectors (Listing 8). The results of the previous Cox proportional hazards analysis provided an estimate of the optimal parameters. The null hypothesis was that every cell type was equally affected by the treatment, and so the optimal parameter for treatment across different levels of cell type would be the same as the non-stratified estimate found earlier. The results were summarized and compared with the previous model results (Table 4.8).

```
df$trt_cell_0 <- df$trt * df$cell_0
df$trt_cell_1 <- df$trt * df$cell_1
df$trt_cell_2 <- df$trt * df$cell_2
df$trt_cell_3 <- df$trt * df$cell_3
names_3 <- c(
  "cell_1", "cell_2", "cell_3",
  "karno50", "trt_cell_0",
  "trt_cell_1", "trt_cell_2",
  "trt_cell_3"
)
tform_3 <- c(
  "loglin", "loglin", "loglin",
  "loglin", "plin",
  "plin", "plin",
  "plin"
)
a_n3 <- c(-0.393, 0.423, 0.756,
          -0.031, rep(-0.22,4)
          )
e3 <- Colossus::RunCoxRegression(
  df, t0, t1, event,
  names = names_3, tform = tform_3,
  a_n = a_n3
)
Colossus::Interpret_Output(e3)
```

Listing 8: Effect Modification Regression

Table 4.8*Lung cancer data, linear hazard ratio effect modification model results.*

Covariate	Subterm	Central Estimate	Standard Error
<i>cell_1</i>	log-linear	-0.39259	0.28220
<i>cell_2</i>	log-linear	0.42251	0.26306
<i>cell_3</i>	log-linear	0.75616	0.29844
<i>karno50</i>	log-linear	-0.03119	0.00516
<i>cell_0 * trt</i>	product-linear	-0.40309	0.23771
<i>cell_1 * trt</i>	product-linear	0.39830	0.55214
<i>cell_2 * trt</i>	product-linear	-0.54858	0.14778
<i>cell_3 * trt</i>	product-linear	0.23492	0.52445
-2*Log-Likelihood: 944.63			
AIC: 960.63			

The new central estimates by cell type were not equal to the original grouped estimate, but the confidence intervals overlapped (Table 4.9). The new model had a smaller AIC than the previous model, suggesting that the model improvement justified the use of additional parameters. It cannot be concluded that the differences in the treatment effects among the cell types were statistically significant.

Table 4.9*Lung cancer data, treatment effect modification comparison (95% confidence interval).*

Covariate	HR	95% Confidence Interval
<i>trt</i>	-0.22426	(-0.52914 , 0.08062)
<i>cell_0 * trt</i>	-0.40309	(-0.86900 , 0.06282)
<i>cell_1 * trt</i>	0.39830	(-0.68389 , 1.48049)
<i>cell_2 * trt</i>	-0.54858	(-0.83823 , -0.25893)
<i>cell_3 * trt</i>	0.23492	(-0.79300 , 1.26284)

Chapter 5

Validation Testing: Use With Million Person Study Data

This chapter covers validation using data from two sources, workers at the Los Alamos National Laboratory (LANL)²⁵ and U.S. nuclear power plants (NPP)²⁶, based on analyses published in 2022 by the Million Person Study. For both cohorts, the original Epicure and SAS scripts were used to verify that the new results were in line with the published values[†].

For both datasets and each organ tested, calculations of Excess Relative Risk (ERR) at 100 mGy cumulative organ dose, Hazard Ratio (HR) at 100 mGy cumulative organ dose, and HR for multiple cumulative organ dose categories were repeated. The previous publications reported the log-linear ERR (ERR_{LL}) as an approximation. In addition to the log-linear ERR, the linear ERR (ERR_L) solved with Colossus was compared with the results obtained using 32-bit Epicure. Comparisons were made using the same Dose Weighting Factor (DWF) as in the original publications. The DWF was related to the intake of radionuclides at LANL, primarily plutonium. Cox proportional hazards models for non-CLL leukemia (DWF=1), lung cancer (DWF=20, combined sexes), ischemic heart disease (DWF=1), and Parkinson's disease (DWF=1) were tested using LANL data. The same models were tested for nuclear power plant worker data, in addition to the model for all solid cancers excluding leukemia.

[†]This chapter is reproduced from previously published work³

The ERR and HR were assumed to be functions of biological sex, level of education and economic categories, year of birth categories, and the annual cumulative dose. The organ doses for non-CLL leukemia, lung cancer, ischemic heart disease, all solid cancers, and Parkinson's disease were the active bone marrow, lung, heart, colon, and brain doses, respectively. These outcomes were selected to capture a variety of event rates commonly calculated by the MPS, at the request of the MPS team.

$$RR_{LL} = \exp \left(\beta_{LL} \cdot dose + \beta_1 \cdot sex + \sum_j \alpha_j \cdot DOB_j + \sum_k \gamma_k \cdot educ_k \right) \quad (5.1)$$

$$RR_L = (1 + \beta_L \cdot dose) \cdot \exp \left(\beta_1 \cdot sex + \sum_j \alpha_j \cdot DOB_j + \sum_k \gamma_k \cdot educ_k \right) \quad (5.2)$$

$$HR(dose, z) = \exp(\beta_{LL} \cdot dose \pm z \cdot \sigma_{\beta_{LL}} \cdot dose) \quad (5.3)$$

$$ERR(dose, z)_{LL} = \beta_{LL} \cdot dose \pm z \cdot \sigma_{\beta_{LL}} \cdot dose \quad (5.4)$$

$$ERR(dose, z)_L = \beta_L \cdot dose \pm z \cdot \sigma_{\beta_L} \cdot dose \quad (5.5)$$

The original publications calculated the ERR with the PEANUTS module in 32-bit Epicure⁴ and calculated hazard ratios with SAS software³³. Every regression was repeated in 32-bit Epicure and Colossus to verify that any differences from the published results were not due to a failure to converge. The published results were provided up to ± 0.01 , so any difference from the published results less than 0.01 was treated as within acceptable limits. The new Epicure analyses, the linear ERR estimates, were compared up to ± 0.001 , so any difference from the new run results less than 0.001 was treated as within acceptable limits. Any differences within these levels of precision might have been due to numerical precision differences or slight differences in the convergence criteria. Wald's method for confidence interval calculation was used for the log-linear ERR and hazard ratios, and likelihood-based bounds were used for the linear ERR to better account for linearity⁵².

5.1 Los Alamos National Laboratory Worker Results

Starting with the overall HR dose models (Table 5.1) and categorical dose HR models (Table 5.2), the published results matched the Colossus results within 0.008. Colossus and 32-bit Epicure regressions with the same convergence criteria found optimal parameters within 0.001, so the minor difference from the published results was concluded to be due to differences in the convergence criteria and rounding error. Overall, the log-linear hazard ratios estimated by Colossus matched both the published results and the 32-bit Epicure results.

Table 5.1

Hazard Ratio at 100 mGy (95% Confidence Interval) for four causes of death among LANL workers comparing Colossus results with published results based on 32-bit Epicure software.

Outcome	Published	Colossus	Difference
Non-CLL Leukemia	0.65 (0.33 , 1.28)	0.649 (0.331 , 1.275)	0.001 (0.001 , 0.005)
Lung Cancer	1.01 (0.98 , 1.04)	1.006 (0.980 , 1.033)	0.004 (0.0 , 0.007)
Ischemic Heart Disease	0.94 (0.85 , 1.04)	0.939 (0.848 , 1.040)	0.001 (0.002 , 0.000)
Parkinson's Disease	1.18 (0.93 , 1.49)	1.175 (0.927 , 1.489)	0.005 (0.003 , 0.001)

Table 5.2

Hazard Ratio for each dose category displayed for dose response in the published article on LANL workers (95% Confidence Interval) for four causes of death comparing Colossus results with published results based on 32-bit Epicure Software.

Non-CLL Leukemia			
Dose Category (mGy)	Published	Colossus	Difference
5 ≤ Dose	1.09 (0.75 , 1.59)	1.094 (0.750 , 1.594)	0.004 (0.000 , 0.004)

Lung Cancer			
Dose Category (mGy)	Published	Colossus	Difference
$5 \leq \text{Dose} < 50$	0.88	0.879	0.001
	(0.71 , 1.09)	(0.712 , 1.086)	(0.002 , 0.004)
$50 \leq \text{Dose} < 100$	0.71	0.707	0.003
	(0.4 , 1.26)	(0.399 , 1.254)	(0.001 , 0.006)
$100 \leq \text{Dose}$	1.02	1.022	0.002
	(0.7 , 1.49)	(0.702 , 1.487)	(0.002 , 0.003)
Ischemic Heart Disease			
Dose Category (mGy)	Published	Colossus	Difference
$5 \leq \text{Dose} < 50$	0.96	0.961	0.001
	(0.87 , 1.07)	(0.865 , 1.068)	(0.005 , 0.002)
$50 \leq \text{Dose} < 100$	1.11	1.104	0.006
	(0.87 , 1.41)	(0.869 , 1.403)	(0.001 , 0.007)
$100 \leq \text{Dose}$	0.91	0.905	0.005
	(0.7 , 1.17)	(0.699 , 1.172)	(0.001 , 0.002)
Parkinson's Disease			
Dose Category (mGy)	Published	Colossus	Difference
$5 \leq \text{Dose} < 50$	0.74	0.735	0.005
	(0.51 , 1.05)	(0.514 , 1.051)	(0.004 , 0.001)
$50 \leq \text{Dose} < 100$	1.83	1.827	0.003
	(1.06 , 3.17)	(1.056 , 3.162)	(0.004 , 0.008)
$100 \leq \text{Dose}$	1.05	1.049	0.001
	(0.53 , 2.06)	(0.534 , 2.062)	(0.004 , 0.002)

Finally, the results of the ERR models for the LANL workers are shown in Tables 5.3 and 5.4. Two ERR models were compared: log-linear and linear. Overall, there was virtually no

difference between the log-linear ERR published and the results from Colossus.

Table 5.3

Log-Linear Excess Relative Risk at 100 mGy (95% Confidence Interval) for four causes of death among LANL workers comparing Colossus results with the published results based on 32-bit Epicure software.

Outcome	Published	Colossus	Difference
Non-CLL Leukemia	-0.43 (-1.11 , 0.24)	-0.432 (-1.107 , 0.243)	0.002 (0.003 , 0.003)
Lung Cancer	0.01 (-0.02 , 0.03)	0.006 (-0.020 , 0.033)	0.004 (0.000 , 0.003)
Ischemic Heart Disease	-0.06 (-0.16 , 0.04)	-0.063 (-0.165 , 0.039)	0.003 (0.005 , 0.001)
Parkinson's Disease	0.16 (-0.07 , 0.4)	0.161 (-0.076 , 0.398)	0.001 (0.006 , 0.002)

When considering the different ERR results, there was an important difference between the Wald confidence intervals and likelihood-based intervals for the linear models. A likelihood-based interval can only be defined where the parameter was valid. Lower confidence interval bounds were not found for either software and almost every model tested because each dose parameter had a lower threshold ($-1/dose_{max}$). If the dose parameter was smaller than the lower threshold, the predicted hazard ratio for at least one row would have been negative. If the optimum score at the lower threshold was within the confidence interval, then the confidence interval did not have a defined lower bound. Wald confidence intervals assume that the model was defined for the full parameter range, so both bounds would always exist.

The only significant difference in the linear ERR comparison was observed in the regression for non-CLL leukemia. The optimum ERR found, -0.11979 , was near the threshold for a negative relative risk. Near this point, 32-bit Epicure was unable to solve for any likelihood-based bounds, while Colossus was able to solve for an upper bound. This was

believed to be a circumstantial issue that arose from the location of the optimal point and was not believed to be reflective of any failure to converge or calculate risks.

Table 5.4

Linear Excess Relative Risk at 100 mGy (95% Confidence Interval) for four causes of death among LANL workers comparing Colossus based on 32-bit Epicure software.

Outcome	Epicure	Colossus	Difference
Non-CLL Leukemia	-0.120 (None ¹ , None)	-0.120 (None ¹ , 0.186)	0.000 (NA , 0.186)
Lung Cancer	0.011 (None ² , 0.072)	0.011 (None ² , 0.072)	0.000 (NA , 0.000)
Ischemic Heart Disease	-0.068 (None ³ , 0.031)	-0.068 (None ³ , 0.031)	0.000 (NA , 0.000)
Parkinson's Disease	0.216 (-0.092 , 0.693)	0.215 (-0.092 , 0.693)	0.000 (0.000 , 0.000)

¹ non-CLL Leukemia has a minimum ERR of -0.1198 at 100 mGy

² Lung Cancer has a minimum ERR of -0.0069 at 100 mGy

³ Ischemic Heart Disease has a minimum ERR of -0.1119 at 100 mGy

5.2 U.S. Nuclear Power Plant Worker Results

The analysis of the nuclear power plant workers started with the overall HR dose models (Table 5.5) and categorical dose HR models (Table 5.6). Overall, the published results and the Colossus results match within 0.009. Once again, the same regressions were run in Colossus and 32-bit Epicure, and the optimal parameters were within 0.001. The differences between the published results and Colossus were believed to be due to differences in the convergence criteria, not a failure to converge.

Table 5.5

Hazard Ratio at 100 mGy (95% Confidence Interval) for five causes of death among nuclear power plant workers comparing Colossus results with published results based on 32-bit Epicure software.*

Outcome	Published	Colossus	Difference
Non-CLL Leukemia	1.16 (1.00 , 1.36)	1.164 (0.999 , 1.357)	0.004 (0.001 , 0.003)
Lung Cancer	0.96 (0.90 , 1.02)	0.957 (0.898 , 1.020)	0.003 (0.002 , 0.000)
Ischemic Heart Disease	0.99 (0.94 , 1.04)	0.991 (0.944 , 1.040)	0.001 (0.004 , 0.000)
All Solid Cancers	1.01 (0.97 , 1.05)	1.008 (0.970 , 1.047)	0.002 (0.000 , 0.003)
Parkinson's Disease	1.27 (0.98 , 1.65)	1.272 (0.979 , 1.652)	0.002 (0.001 , 0.002)

* 90% confidence interval used for non-CLL Leukemia per published value

Table 5.6

Hazard Ratio for each dose category displayed for dose response in the published article on nuclear power plant workers (95% Confidence Interval) for five causes of death comparing Colossus results with published results based on 32-bit Epicure software.*

Non-CLL Leukemia			
Dose Category (mGy)	Published	Colossus	Difference
$5 \leq \text{Dose} < 50$	1.20 (0.93 , 1.53)	1.196 (0.934 , 1.531)	0.004 (0.004 , 0.001)
$50 \leq \text{Dose} < 75$	1.51 (1.03 , 2.21)	1.507 (1.029 , 2.205)	0.003 (0.001 , 0.005)
$75 \leq \text{Dose} < 100$	1.43 (0.87 , 2.34)	1.429 (0.873 , 2.339)	0.001 (0.003 , 0.001)

100 $\leq$ Dose < 200	1.01 (0.62 , 1.66)	1.013 (0.619 , 1.658)	0.003 (0.001 , 0.002)
200 $\leq$ Dose < 300	2.05 (1.06 , 3.97)	2.050 (1.059 , 3.966)	0.000 (0.001 , 0.004)
300 $\leq$ Dose	2.52 (1.08 , 5.91)	2.520 (1.075 , 5.907)	0.000 (0.005 , 0.003)
Lung Cancer			
Dose Category (mGy)	Published	Colossus	Difference
25 $\leq$ Dose < 50	1.18 (1.08 , 1.29)	1.184 (1.082 , 1.295)	0.004 (0.002 , 0.005)
50 $\leq$ Dose < 100	0.99 (0.88 , 1.11)	0.991 (0.884 , 1.111)	0.001 (0.004 , 0.001)
100 $\leq$ Dose < 250	1.01 (0.88 , 1.17)	1.014 (0.880 , 1.168)	0.004 (0.000 , 0.002)
250 $\leq$ Dose	0.73 (0.52 , 1.03)	0.731 (0.521 , 1.027)	0.001 (0.001 , 0.003)
Ischemic Heart Disease			
Dose Category (mGy)	Published	Colossus	Difference
25 $\leq$ Dose < 50	1.02 (0.95 , 1.10)	1.020 (0.947 , 1.099)	0.000 (0.003 , 0.001)
50 $\leq$ Dose < 100	1.01 (0.93 , 1.11)	1.012 (0.925 , 1.107)	0.002 (0.005 , 0.003)
100 $\leq$ Dose < 250	1.00 (0.89 , 1.12)	0.996 (0.889 , 1.115)	0.004 (0.001 , 0.005)
250 $\leq$ Dose	0.95 (0.75 , 1.20)	0.948 (0.749 , 1.200)	0.002 (0.001 , 0.000)

All Solid Cancer			
Dose Category (mGy)	Published	Colossus	Difference
$25 \leq \text{Dose} < 50$	1.16	1.153	0.007
	(1.09 , 1.23)	(1.088 , 1.223)	(0.002 , 0.007)
$50 \leq \text{Dose} < 100$	1.00	1.000	0.000
	(0.93 , 1.08)	(0.930 , 1.076)	(0.000 , 0.004)
$100 \leq \text{Dose} < 250$	1.06	1.051	0.009
	(0.97 , 1.15)	(0.961 , 1.149)	(0.009 , 0.001)
$250 \leq \text{Dose}$	0.90	0.897	0.003
	(0.74 , 1.09)	(0.737 , 1.092)	(0.003 , 0.002)
Parkinson's Disease			
Dose Category (mGy)	Published	Colossus	Difference
$25 \leq \text{Dose} < 50$	1.03	1.028	0.002
	(0.64 , 1.65)	(0.640 , 1.651)	(0.000 , 0.001)
$50 \leq \text{Dose} < 100$	0.71	0.708	0.002
	(0.34 , 1.46)	(0.343 , 1.461)	(0.003 , 0.001)
$100 \leq \text{Dose} < 250$	1.74	1.735	0.005
	(0.93 , 3.25)	(0.927 , 3.246)	(0.003 , 0.004)
$250 \leq \text{Dose}$	2.35	2.353	0.003
	(0.75 , 7.44)	(0.745 , 7.435)	(0.005 , 0.005)

* 90% confidence interval used for non-CLL Leukemia per published value

The regression results for both ERR models are presented in Tables 5.7 and 5.8. Similar to the LANL evaluations, the published log-linear ERR matched the Colossus results, and the linear ERR matched between the Colossus and 32-bit Epicure results. These likelihood boundary solutions demonstrated both upper and lower-bound convergence.

Table 5.7

Log-Linear Excess Relative Risk at 100 mGy (95% Confidence Interval) for five causes of death among nuclear power plant workers comparing Colossus results with the published results based on 32-bit Epicure.*

Outcome	Published	Colossus	Difference
Non-CLL Leukemia	0.15 (0.000 , 0.31)	0.152 (-0.001 , 0.305)	0.002 (0.001 , 0.005)
Lung Cancer	-0.04 (-0.11 , 0.02)	-0.044 (-0.108 , 0.012)	0.004 (0.002 , 0.008)
Ischemic Heart Disease	-0.01 (-0.06 , 0.04)	-0.009 (-0.058 , 0.040)	0.001 (0.002 , 0.000)
All Solid Cancers	0.01 (-0.03 , 0.05)	0.008 (-0.031 , 0.046)	0.002 (0.001 , 0.004)
Parkinson's Disease	0.24 (-0.02 , 0.50)	0.241 (-0.021 , 0.502)	0.001 (0.001 , 0.002)

* 90% confidence interval used for non-CLL Leukemia per published value

Table 5.8

Linear Excess Relative Risk at 100 mGy (95% Confidence Interval) for five causes of death among nuclear power plant workers comparing Colossus results with the published results based on 32-bit Epicure software.

Outcome	Epicure	Colossus	Difference
Non-CLL Leukemia	0.206 (-0.038 , 0.562)	0.206 (-0.038 , 0.562)	0.000 (0.000 , 0.000)
Lung Cancer	-0.049 (None ¹ , 0.016)	-0.049 (None ¹ , 0.016)	0.000 (NA , 0.000)
Ischemic Heart Disease	-0.010 (-0.056 , 0.042)	-0.010 (-0.056 , 0.042)	0.000 (0.000 , 0.000)

All Solid Cancers	0.009	0.009	0.000
	(-0.030 , 0.051)	(-0.030 , 0.051)	(0.000 , 0.000)
Parkinson's Disease	0.318	0.318	0.000
	(-0.056 , 0.943)	(-0.056 , 0.943)	(0.000 , 0.000)

¹ Lung Cancer has a minimum ERR of -0.0922 at 100 mGy

5.3 Validation Test Conclusions

These results demonstrated that analyses previously performed using SAS and 32-bit Epicure could also be performed in R using Colossus. However, these results also demonstrated the importance of using multiple software for the validation of results and the importance of running the regression at multiple initial parameter estimates. The results of a single software and initial parameter estimate can be misleading. Combined with the verification testing in Chapter 3, Colossus has been shown to have both the modeling capabilities of 32-bit Epicure and the programming flexibility of R functions. Introducing these capabilities into the R language will allow complex risk models to be used in tandem with a wide variety of libraries for loading and processing big data.

Chapter 6

Competing Risks Effects in Plasmode Simulations

Many epidemiological studies use real event data to approximate event rates in a population. It can also be useful to apply a hypothetical event rate to a population and measure its potential effects. This can help answer questions about excess cases for different event rates or how many samples are needed to detect the effect of a risk factor. Methods exist for applying risk models to sampled data, but they should be used with an understanding of the biases from the original study design.

In an ideal study, every individual could be followed from the start of observation until they experience the event of interest. In many studies, individuals might be censored and leave the trial before experiencing any event, or they may experience a competing event, which is similar to censoring. However, there is an important distinction in that censoring is assumed to be noninformative, meaning that the censoring and event times for individuals in the study are statistically independent. Random censoring or censoring applied uniformly to a population reduces the sample size but does not introduce any bias. Competing event rates generally do not meet these criteria and have the potential to introduce bias into the analysis. This difference is the reason for the development of statistical methods to describe multi-state models⁵³. Multi-state models describe situations in which a subject can

transition from being alive to multiple different event states. This includes simple survival, multiple sequential events, competing events, and an illness-death model in which a subject can transition between being healthy or sick and die in either state. The focus of this work was on competing risk models.

The purpose of this chapter was to apply different methods of sampling data and observe how the effects of competing risks in the original data could be transferred to the sampled data. The analysis was based on a hypothetical study of the competing risks of lung cancer and Parkinson’s disease (PD) mortality in the Million Person Study (MPS)¹⁷ cohort of Nuclear Power Plant (NPP) workers²⁶. Analyses of large population studies have found that cigarette smoking reduces the risk of Parkinson’s disease^{6;54}. A study of male British doctors found that current smokers had a 30% lower risk of PD than those who never smoked⁵⁴. If this effect was not biologically significant or was not as strong as indicated, then another possible explanation was that heavy smokers tended to die from other causes before experiencing PD mortality. This bias could be tested by applying both events to a population with known event rates, performing a regression on the effects of the SES category on the risk of PD mortality, and then comparing the predicted hazard ratio to the “true” value.

In this chapter, the CEDR version of the NPP dataset⁵⁵ was used to generate estimates of the underlying baseline and relative risk models for multiple outcomes of interest and was randomly sampled to generate hypothetical study populations. Then, a method was tested to extend the observed survival of individuals sampled with plasmode simulations and remove any lingering effects of event times from the original data. The known event rates were applied to the sampled populations to generate random realizations of the event times for each sampled individual. Finally, the sampled datasets with applied events were analyzed using Colossus³⁵ to generate distributions of the predicted hazard ratios for low socioeconomic status, considering different combinations of competing events and sampling methods (Fig. 6.1).

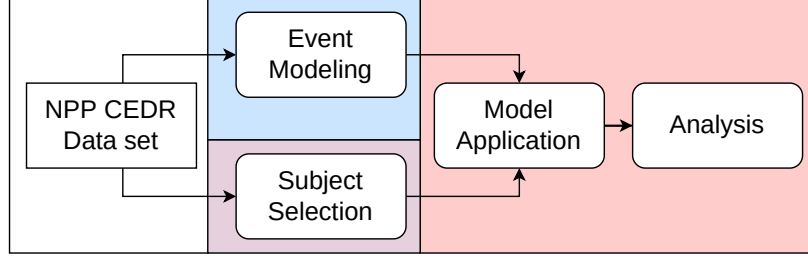


Figure 6.1: The original data was used to create risk models and samples of individuals. These were combined to create randomly assigned outcome datasets for analysis. These steps are expanded on in figures 6.9, 6.10, and 6.11

This study was designed to test the hypothesis that standard methods of plasmode simulation of the NPP cohort would inherit competing risk effects and increase the predicted protective effect, independent of the risk models applied to the data. The hypothesis was tested by generating simulated datasets with different combinations of sampling, with and without survival extension, and different combinations of applied competing events. If the unextended plasmode simulations without competing risks modeled produced estimates similar to the extended plasmode simulations with competing risks modeled, then there would be strong evidence that the default plasmode simulation introduced competing risk effects.

6.1 Theory

6.1.1 Effects of competing events in time series analysis

Competing risk effects arise from subsets of the cohort with different rates of competing events. Competing events have a similar effect to sampling bias; a well-designed study with an equal distribution of risk factors at age 18 will not necessarily have an equal distribution of risk factors at age 80. In a broad sense, this effect is conceptualized by treating every individual as if they have a known event time for every outcome (T_i) and censoring time (C). Each individual is tracked up to a final observation time (T), which is the minimum of the event and censoring times ($\min(T_0, T_1, \dots, T_n, C)$)⁵³.

There are two methods for correcting for the effects of competing risks. The first is to

treat rows with competing events as never experiencing the main event and contributing to later risk groups, weighted by the probability of not having been censored. This can be done using the Fine-Gray regression model⁴³, which is discussed in greater detail in Section 2.4. The second method is to include a counterfactual and substitute what would have happened if an individual had not experienced a competing event. This is rarely possible in observational studies but can be straightforward in a simulated study. In a simulation study, every event time can be generated, so the counterfactual is known for every individual.

One detail of this application that made the simulation study more complicated was the presence of cumulative radiation exposure, which varied over time. In survival analysis, individuals are typically represented by intervals with constant risk factors. If there are no time-dependent covariates, each individual can be represented by a single interval. In the nuclear power plant worker cohort, radiation exposure was a chronic occupational exposure, and individuals were represented by annual intervals. This meant that the event rates affected both the outcome column and the distribution of intervals observable at each age. In both cases, a competing event causes future events to be unobserved. In an analysis with time-dependent covariates, a competing event would also cause future exposures to be unobserved.

6.1.2 Applying known risk models to sampled intervals

During a simulation study, it is often necessary to take a known equation of the event rate and then randomly assign events to intervals. This can be done by taking a collection of intervals, calculating the cumulative hazard over each interval, and then randomly selecting a surviving proportion for each interval. The surviving proportion corresponds to an event time, and every interval containing their event time experiences the event.

The goal of the competing risk analysis was to apply event rates at levels comparable to those in the NPP cohort. Three events were selected for modeling: PD mortality, lung cancer mortality, and a collection of all other causes of mortality and censoring. PD and lung cancer mortalities were selected to directly isolate the effect of lung cancer mortality. The collection of all other causes of mortality and censoring was selected to compare the total

effects of every cause of mortality and censoring in the cohort with the effects of plasmode simulations without observation extension.

The cumulative hazard of each mortality event was solved similarly to the examples shown in Section 2.5, by solving for the hazard ratio and approximating the baseline hazard (Eqs. 2.35-2.36). Colossus³⁵ was used to solve the cumulative hazard of mortality from PD and lung cancer as functions of attained age, cumulative organ doses, year of birth cohort, biological sex, and SES. The cumulative hazard of mortality from all other causes was approximated as a function of age and year of birth cohort by taking the logarithm of the survival functions (Eq. 2.37). The stratification of all other causes by year of birth was vital because individuals born later were likely to still be alive and censored without events at their last follow-up (Fig. 6.2).

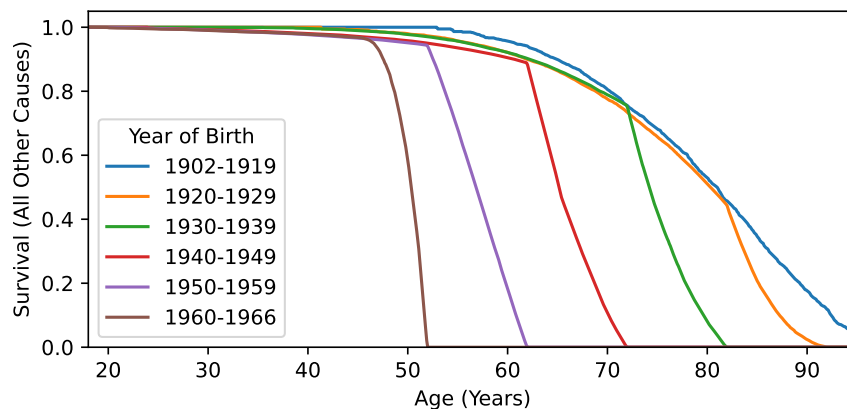


Figure 6.2: Survival curves for the individuals in each birth cohort who stop being observed due to causes other than lung cancer or PD mortality. The NPP cohort was an updating dataset, so censoring rates were significantly different between the year-of-birth cohorts.

A similar study could have been performed using a non-parametric or semi-parametric model for the mortality rate, which may have been a better model for the event rate in the original dataset. In this study, parametric risk models were chosen so that each survival model regression could be compared with known solutions. The risk models were parametrically defined using exponential models available in Colossus so that a risk model that matched the true simulated risk model formula could be fitted (Eq. 6.4-6.8). Any difference between the observed and true model parameters could be used to quantify the differences

in the accuracy of the models between the simulated datasets.

6.1.3 Methods to sample from existing data

For this study, the simulated population needed to be comparable to the study population, particularly in terms of the associations between demographics and radiation exposure. One method of generating a random population from a study population was to use a plasmode simulation⁵⁶, by randomly sampling individuals from an existing study population and then applying a model for expected events. Plasmode simulations were designed to preserve the correlations among covariates in the population.

In a typical plasmode simulation, the researcher starts with a population of N individuals, each having a series of constant covariates $\vec{x}$, a measure of survival time t , and an event indicator e . The researcher then samples n individuals from the population, uses the covariates to estimate the new survival time of each sampled individual, replaces the original survival time with the new survival time, and then analyses the new population. M independent samples are generated and analyzed (Fig. 6.3). The number of samples and trials is usually determined by weighing the preservation of the correlation with the computational resources required to perform the analysis⁵⁷.

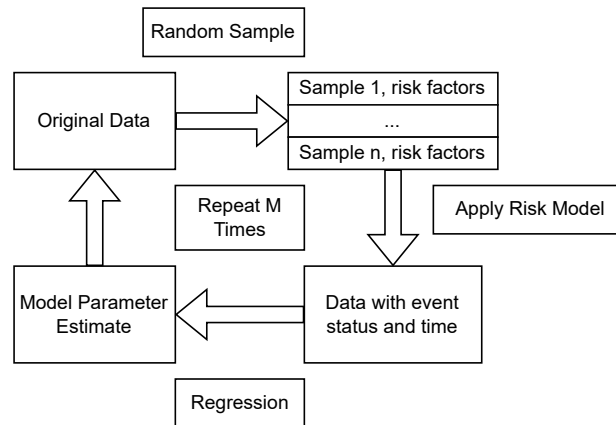


Figure 6.3: A plasmode simulation starts by sampling n subjects from the cohort. A new event time was generated for each subject. This is repeated M times to analyze the distribution in hazard ratio estimates.

Plasmode simulations remove the effects of the original event times by reassigning event times and assuming that future risk factors are known. If every individual were represented by a single interval with constant risk factors, the survival time could be freely adjusted to the final survival time. If one event is modeled, then every individual is observed from the start of the study until the event time. If multiple events are modeled, every individual is observed until the first event time.

This study included time-dependent covariates. The typical method of sampling time-series data for a plasmode simulation is to sample individuals and then include all intervals for each individual sampled⁵⁶. After applying the risk models, the first event for each individual is identified, and any subsequent intervals are removed. Because the sampled intervals are limited by the length of time that the original study population members survived, the interval distribution is inherently biased. Even if only one event was modeled, individuals cannot be observed beyond their original survival time. One way to remove this bias was to add new interval data to extend the observation of individuals beyond their original final observation time. Doing so incorrectly could ruin the preserved correlation between the constant and time-dependent covariates.

One option for performing a plasmode simulation without introducing competing risk effects was to fit a semi-parametric model for time-dependent exposure and use it to extend the observation of individuals. In theory, a semi-parametric model can be used to predict the annual dose without imposing parametric assumptions or restrictions. This method was limited by the available data because the model could only be trained using the recorded covariates. Notably, this meant that the model accuracy would suffer with age if there were not enough high-age intervals in the data.

The general simulation procedure began by training a machine learning model to predict the annual exposure using the available data. The sampled data started with only the observed intervals (Fig. 6.4, left). Every subject needed to be observed until a new final shared age, so new intervals were predicted for every subject that was removed from the dataset prior to the new final age (dashed line). Every new interval (red) was predicted using information from the previous interval (Fig. 6.4, middle). The process was complete

when each subject was observed until the new final shared observation time (Fig. 6.4, right).

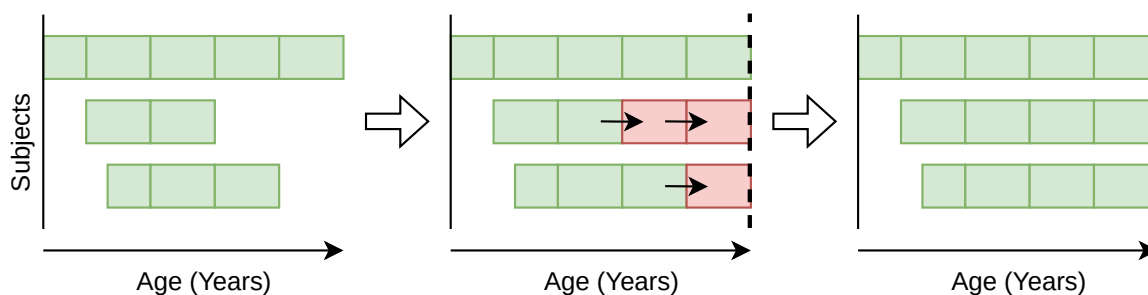


Figure 6.4: The data goes from missing higher age observations to having a shared final observation age (dashed line), by predicting missing exposures (red)

Because every individual would have intervals propagated to the same age, every individual could be observed up to any combination of events, independent of their original survival time. This method was dependent on a semi-parametric model being able to accurately predict future exposures within reason, which is discussed further in Section 6.2.1.

6.2 Methods

6.2.1 Annual exposure modeling with Random Forest algorithms

To maintain the correlation between radiation exposure and personal characteristics, a semi-parametric Random Forest model was selected to predict the annual radiation exposure to multiple organs⁵⁸. However, there was insufficient information to assign an exact annual exposure, so exposure was predicted up to a distribution of values. The annual exposures were categorized into bins and fitted with a categorical model to return the probability that the annual dose would fall within each bin. For each predicted interval, a bin was randomly assigned, and a dose within the bin was randomly selected with a uniform probability. Exposure categories were divided into one bin for zero annual exposure and 50 bins spaced logarithmically from 10^{-5} mGy to 220 mGy (Fig. 6.5). The bins were selected to fully cover the distribution of annual exposures, be thin enough that the dose within each bin would not vary drastically, and avoid bins with too few counts.

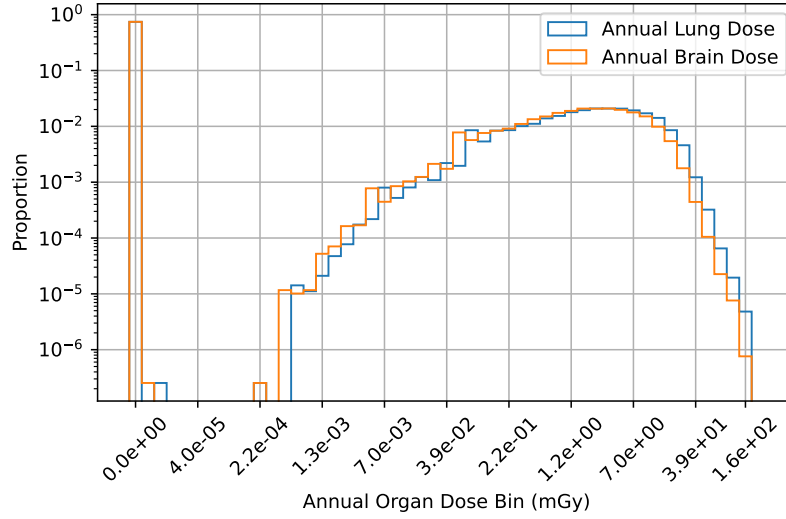


Figure 6.5: Histogram of intervals in each annual dose bin. A majority of intervals contained zero annual dose, and the most common non-zero dose bins were on the scale of 1 – 10 mGy.

To design the Random Forest model, relevant covariates were first identified. Individuals who worked at the same time in similar areas were likely to have similar exposures, which were estimated using a combination of age during each interval, year of the interval, hiring year, and level of education of the individual. Any whole-body external exposure that increases the dose to one organ is likely to increase the dose to other organs; therefore, annual exposures to different organs within the same year were associated. Finally, the annual dose for one year was used to provide information on the annual dose for the next year.

The model was split into two sub-models. The first sub-model used data from each interval to predict annual lung exposure in the subsequent year. The second sub-model used the same information, in addition to lung exposure in the following year, to predict the brain exposure in the following year. The sub-model performance was tuned by testing different numbers of trees and allowable tree depths. The model for lung exposure performed with 81% accuracy, and the model for brain exposure, assuming correct lung exposure, performed with 99% accuracy. A feature importance analysis was performed to determine which features contributed the most to improving the accuracy of the trees. The analysis found that the annual dose was the most important feature, followed by the year of hire, current year, and year of birth.

Finally, extrapolation to older ages was investigated. Almost 50% of the individuals were not observed past age 62 (Fig. 6.6). Within the data, PD occurred at greater ages; therefore, the high age intervals needed to be realistic. A correction factor was applied, based on observations in the data and expectations given the nature of occupational exposure in the cohort. Radiation exposure was a combination of external and internal sources. External radiation exposure should have been greatest while the individual worked at the plant and then have immediately gone to zero when they retired. Internal exposure should have decreased slowly as the isotopes decayed and were biologically eliminated. The correction factor also accounted for the 10-year lag in the cumulative dose. One would expect that each individual retired, their external exposures stopped 10 years later, and then any internal exposures continued to decrease.

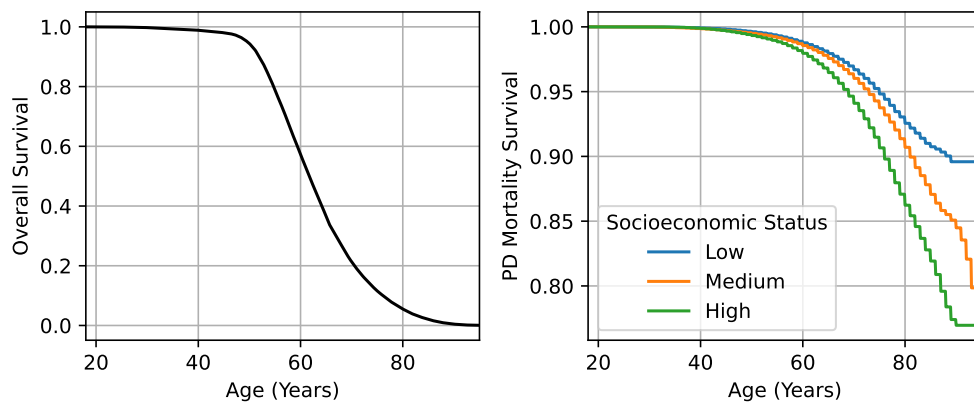


Figure 6.6: Survival curves showing (left) the proportion of individuals who continue to be observed overall and (right) the proportion who survive PD mortality. Due to other events in the data, most individuals left the NPP dataset due to causes other than PD mortality.

Based on histograms of the annual dose and age in the NPP data, most intervals had zero annual dose, and the probability of non-zero doses decreased as age increased past 70 years (Fig. 6.7.a-b). This is consistent with primarily external exposures; once individuals stopped working, they stopped being exposed and were much more likely to have no annual radiation dose. However, histogram normalization did not account for changes in population with age. When bin counts were normalized by the total number of intervals at each age, the decrease in non-zero bin counts at higher ages was less pronounced (Fig. 6.7.c-d).

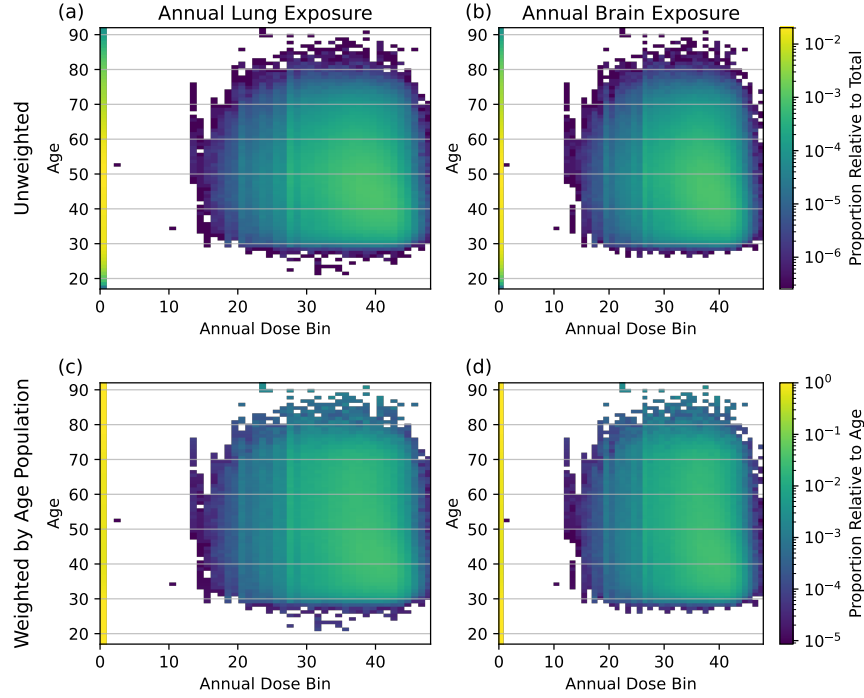


Figure 6.7: Two-dimensional histograms counting the proportion of intervals within each age and annual exposure bin. a-b) Non-zero annual lung exposures and annual brain exposures drop off after age 70. c-d) The same data normalized to the population in each age bin showed a nearly constant distribution up past age 75.

The lack of high-age data caused an issue when the Random Forest model was fitted to the data. According to the data, the dose distribution did not change significantly between ages 30-80, meaning that a Random Forest model that did not change after age 30 still performed well for the vast majority of the data. The trends in the data did not match the expected changes in exposure after retirement age. Instead of a sudden drop, a gradual decrease was observed. For this analysis, a decay was imposed (Eq. 6.1) on the annual dose such that individuals past age 90 effectively had no annual dose (Eq. 6.2, Fig. 6.8). This step may not be necessary as the follow-up for this cohort is extended. It was important to note that these corrections did not significantly change the cumulative dose for most individuals, who experienced the majority of their cumulative dose at younger ages.

$$C(t) = \exp((t - 90) * \ln(0.1)/(70 - 90)) \quad (6.1)$$

$$D^*(t) = \begin{cases} D(t) & C(t) \geq 1 \\ D(t) * C(t) & 1 > C(t) > 0 \\ 0.0 & C(t) \leq 0 \end{cases} \quad (6.2)$$

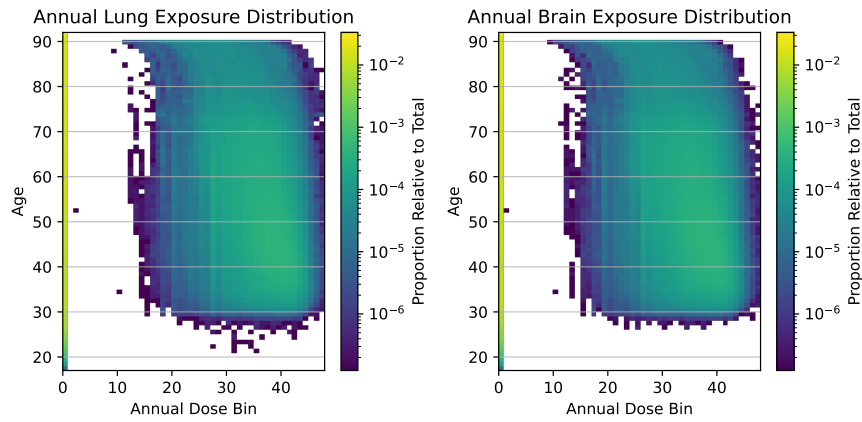


Figure 6.8: Two-dimensional histograms counting the proportion of intervals within each age and annual exposure bin with extended survival. Exposures were weighted to force decay in the expected dose with age past retirement, starting around age 80.

6.2.2 Plasmode sampling

In both cases of plasmode simulation, the analysis started by selecting a sample size (n) and the number of trials (M). The procedure outlined by Bickel et al. ⁵⁷ was followed to select the appropriate values. In short, a series of potential sample sizes was generated using equation 6.3 and $q = 0.97$, rounding to the nearest integer. For each sample size, the L_2 -norm of the covariance matrix was calculated for 500 trials. Finally, the Wasserstein metric was used to measure the change between the values of n . The goal of this process was to find the values of n that minimized the difference between the consecutive values of n . This process found that sampling 100,000 of the 130,000 individuals sufficiently reduced the change in the

covariance.

$$n_j = [q^j N] \quad (6.3)$$

For both methods, subject IDs were randomly sampled with replacement, and every interval for each sampled ID was included. Subjects were assigned new study IDs so that repeated subject IDs were treated as separate individuals. In the second method, new annual intervals were generated to extend the survival of every individual to age 95 (Fig. 6.9). New intervals were generated independently for each sample so that each sampling was an independent realization of future exposure. This preserved individual demographics from the original NPP data in both methods and maintained the distribution over time in the second method.

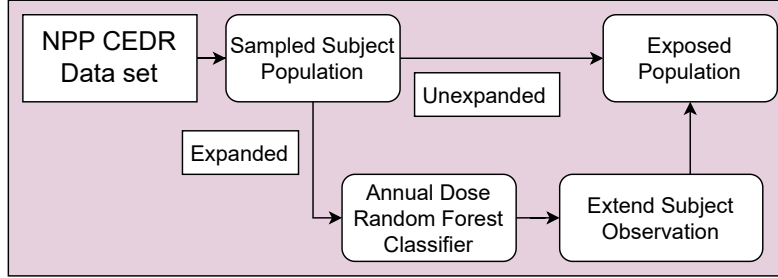


Figure 6.9: The NPP dataset was sampled to define random populations. In unextended sampling, the newly sampled population was immediately used as the exposed population. In the extended sampling, new intervals were added and then used as the exposed population.

6.2.3 Risk model derivation

As previously stated, the risk models were derived from the analysis of the NPP dataset (Fig. 6.10). First, exponential hazard ratio models were fitted for lung cancer and PD mortality (Eq. 6.4). The hazard ratio estimates were then used to estimate the cumulative hazard. The hazard at any point in time (t_i) was approximately equal to the number of events (d) divided by the sum of the hazard ratios for each row at risk at that time ($j \in R(t_i)$) (Eq. 6.5).

The cumulative hazard for all other causes was calculated by taking the natural logarithm of the survival curve for all other causes (Eq. 6.6).

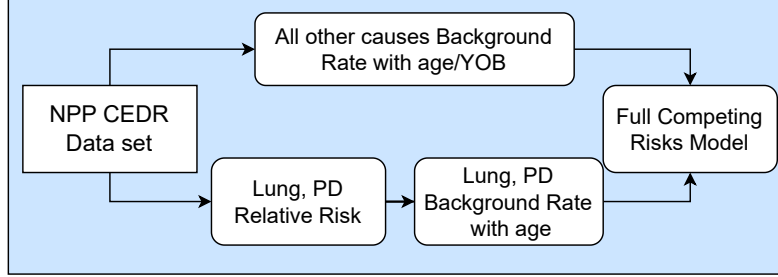


Figure 6.10: The NPP dataset was used to approximate the background and relative event rates for multiple outcomes. These models are combined independently to approximate different combinations of competing risk effects.

$$HR_{organ,k} = \exp \left(\sum_{i \in SES} \beta_i \cdot x_{i,k} + \sum_{j \in YOB} \beta_j \cdot x_{j,k} + \beta_{sex} \cdot x_{sex,k} + \beta_{organ,k} \cdot x_{organ,k} \right) \quad (6.4)$$

$$CH(t) \approx \sum_{t_i \leq t} \frac{d(t_i)}{\sum_{j \in R(t_i)} HR_{organ,j}} \quad (6.5)$$

$$CH(t) \approx -\ln(S(t)) \quad (6.6)$$

The cumulative hazard curves were approximated using exponential functions. The cumulative hazard curves for lung cancer and PD mortality were well approximated by the exponential of a quadratic function (Eq. 6.7). The cumulative hazard for all other causes was less straightforward. The cumulative hazard was informed by a combination of individuals who died from other causes and individuals who were censored because they were still alive at the last follow-up (Fig. 6.2). This was modeled with a spline effect so that the predicted survival in each birth cohort dropped significantly past a threshold (Eq. 6.8).

$$CH_{q,organ}(t) = \exp(\beta_{0,organ} \cdot t^2 + \beta_{1,organ} \cdot t + \beta_{2,organ}) \quad (6.7)$$

$$CH_{ls,B}(t) = \begin{cases} \exp(\beta_{0,B} \cdot t + \beta_{1,B}) + \beta_{3,B} & t < \alpha_B \\ \exp(\beta_{0,B} \cdot t + \beta_{1,B} + \beta_{2,B} \cdot (t - \alpha_B)) + \beta_{3,B} & t \geq \alpha_B \end{cases} \quad (6.8)$$

Finally, the models were combined to solve for the cumulative hazard in each interval (Eqs. 6.9-6.11). The interval widths were limited to one year so that the cumulative hazard was approximately linear within each interval (Eqs. 6.12-6.13). Events were randomly assigned by generating a random survival proportion (U) for each interval and then solving the survival equation for the expected event age (Eqs. 6.14-6.15). Every interval that contained the expected event age was assigned the event and updated interval length.

$$CH_{lung}(t) = CH_{q,lung} \cdot HR_{lung} \quad (6.9)$$

$$CH_{PD}(t) = CH_{q,PD} \cdot HR_{brain} \quad (6.10)$$

$$CH_{Else}(t) = CH_{ls,YOB} \quad (6.11)$$

$$S(t_{event}|t_{entry}) = \exp(-(CH(t_{event}) - CH(t_{entry}))) \quad (6.12)$$

$$S(t_{event}|t_{entry}) \approx \exp(-\Delta CH \cdot (t_{event} - t_{entry})) \quad (6.13)$$

$$U = S(t_{event}|t_{entry}) \quad (6.14)$$

$$t_{event} = t_{entry} - \log(U)/\Delta CH \quad (6.15)$$

Each event had columns for the event indicator and event time. Events were applied by determining the first observed events occurring in each interval and then filtering the data to remove any intervals after the first observed event for each subject (Fig. 6.11). The exit ages for each interval were rounded to the nearest 0.01 year to limit the number of unique event times.

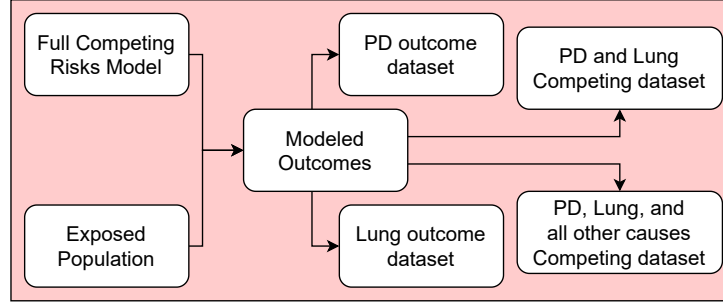


Figure 6.11: Risk models were applied to exposed populations to randomly assign outcomes, and datasets were generated for different combinations of events.

6.3 Results and Discussion

The goal of this work was to measure the effects of interval extension and competing events in each simulation and identify trends in bias and variance. The three metrics used were comparisons of the overall survival curves (Fig. 6.6), comparisons between the proportion of individuals with events in each SES category (Table 6.1), and estimated effects of SES category on hazard ratio using a Cox regression model. It is important to note that the values were not expected to match perfectly between the original data and plasmode simulations. These differences were attributable to the variance between the simulations and inaccuracies in the event rate models.

Table 6.1
PD Mortality by SES status in original data.

	PD Mortality	Other	Total
Low SES	28	33,610	33,638
Medium SES	66	67,894	67,960
High SES	46	33,689	33,735
Total	140	135,193	135,333

Before performing the regressions, it is important to understand any differences between the SES cohorts. If the distributions of risk factors among SES cohorts were not similar,

then any differences in event rates could have been due to factors other than SES. The distributions of biological sex and final cumulative doses were very similar among the SES subsets (Fig. 6.12-6.13). One trend was the difference in year of birth cohorts. Every SES population was primarily composed of subjects born between 1940 and 1959, but the low SES population contained more subjects born earlier. Generally, as SES increased, more subjects were born at later ages (Fig. 6.12). A related trend occurred in the final observation time, with a larger proportion of low SES subjects observed at later final ages (Fig. 6.13). The lowest SES population contained a higher proportion of subjects born earlier; therefore, there was a lower proportion of subjects still being tracked.

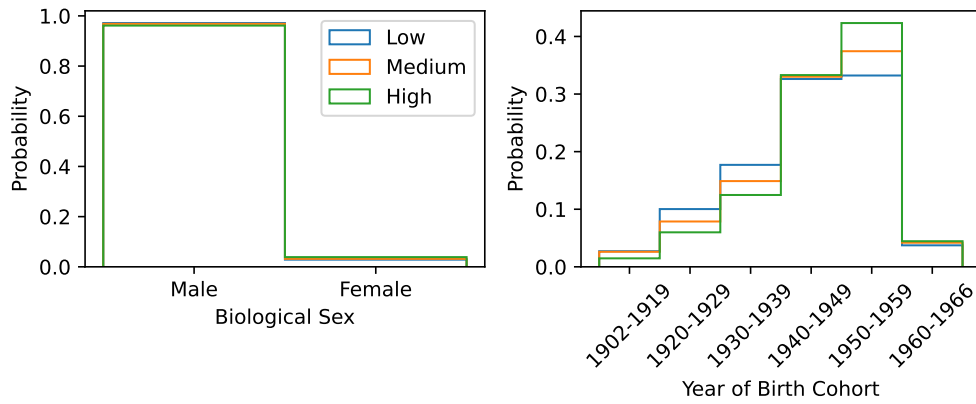


Figure 6.12: The populations in each SES cohort have similar biological sex distributions, but lower SES status was also associated with earlier years of birth.

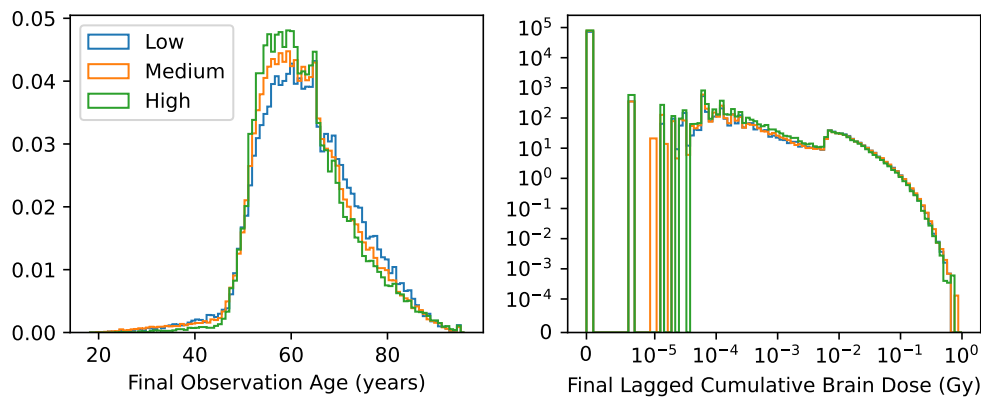


Figure 6.13: The distribution of final cumulative doses were very similar between SES categories. Final observation ages peaked in the range of 50-70, however the peak was less pronounced in lower SES categories.

6.3.1 Unextended plasmode analysis

A plasmode simulation without extension limited the observable time for individuals. Thus, one would expect that the data would have already included the loss of follow-up to all causes of outcome and censoring before any events were applied. Individual survival would decrease on average, so the overall survival curve was expected to be strictly below the survival curve of the original data (Fig. 6.14). The average survival in the unextended plasmode simulation considering only PD mortality varied within a narrow range around the original survival. This variation was likely due to the variations in the survival curves for each sampled population. Modeling both PD and lung cancer mortality reduced the survival rate, and including all other causes of censoring and mortality reduced it further. As more events were added, the survival central estimates and variance decreased. As expected, applying additional event models to the data without increasing the observable time only exacerbated the loss of follow-up.

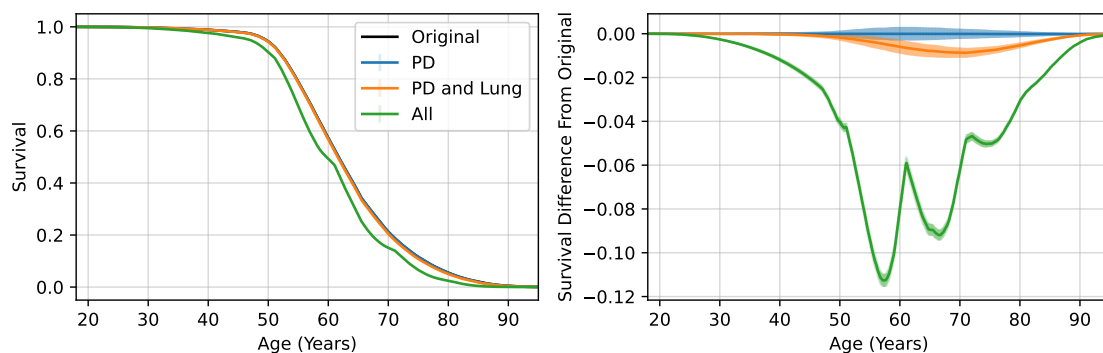


Figure 6.14: In the unextended plasmode simulations, the survival curve decreased further below the original curve as more event types were added. Variation in survival between simulations (depicted as 95% confidence bands) was due to the variation in the original final observation times between sampled populations.

Overall, the study found that in the unextended plasmode simulations that only modeled PD mortality:

1. The average estimates of HR for low and medium SES category converged over the 500 realizations (Fig. 6.15.a-b);

2. Overestimation of one HR tended to be associated with overestimation of the other (Fig. 6.15.c, Table 6.2); and
3. The average estimates of HR were close to the known model parameters, but there was considerable variation in the HR estimate. (Fig. 6.15.c)

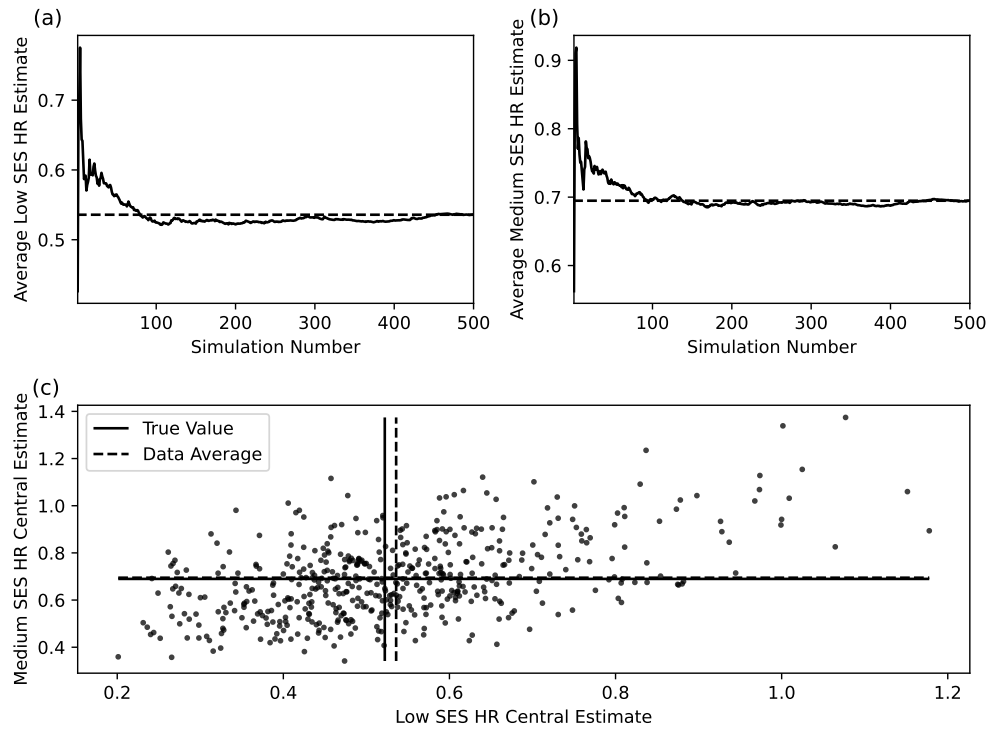


Figure 6.15: HR parameter estimates for SES status on PD mortality in unextended samples without competing risks applied. (a-b) The average parameter values converge over the simulations, and (c) the true parameter values (solid) were below the observed average (dashed).

Table 6.2

HR estimate comparison to true value in unextended data considering only PD mortality.

	Low SES Above	Low SES Below	Total
Medium SES Above	166	98	264
Medium SES Below	94	142	236
Total	260	240	500

$$\chi^2: 24.868$$

$$\text{p-value: } 6 \times 10^{-7}$$

Next, the regression results among different competing risk situations were compared. Three combinations of outcomes were simulated: individuals subjected only to PD mortality, those subjected to both PD and lung cancer mortality, and those subjected to a collection of different censoring and mortality causes. In all cases, individuals were observed until their first assigned event or their original final observation time. Survival rates decreased as more event types were included (Fig. 6.14), so the dataset with the full set of outcomes was the least appropriate representation of the original survival curve. The distributions of the predicted HR for just PD mortality and PD & lung cancer mortality were very similar, and all three distributions were skewed toward the null. For the distribution of the full set of outcomes, 8.4% of simulations produced a HR estimate above 1 (Fig. 6.16).

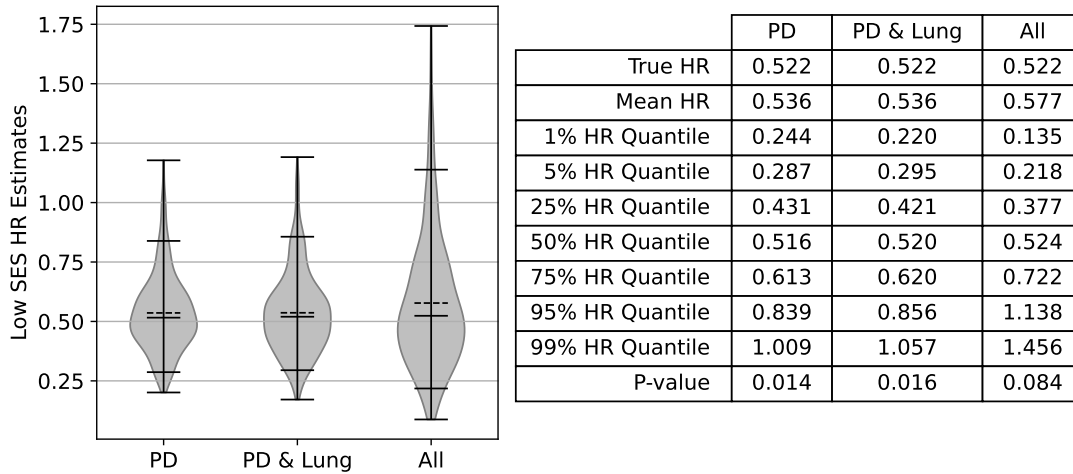


Figure 6.16: Violin plots of the low SES HR distributions for each grouping of events in unextended data. The median HR (middle solid line) was below the mean HR (dashed line). All three had very similar median values near the true value, and the mean value for the full set of outcomes was furthest off.

For the proportion of individuals with PD events, the most accurate representation was the dataset with only PD mortality simulated. On average, 9.2 out of every 10,000 individuals in the unextended plasmode simulations subjected only to PD mortality experienced a PD event, compared with 10.3 in the original data (Table 6.3). As lung cancer and additional

outcomes were added, over 50% of the individuals who would have experienced a PD event left the dataset due to a competing event.

Table 6.3

PD Mortality Proportion (events per 10,000 and 95% confidence interval) by SES status in unextended data.

	Original Data	PD Only	PD and lung	All Causes
Low SES	8.32	7.54 (4.11, 10.97)	6.75 (3.46, 10.04)	3.53 (1.16, 5.90)
Medium SES	9.71	8.56 (6.17, 10.95)	7.82 (5.55, 10.09)	3.90 (2.23, 5.57)
High SES	13.64	12.32 (7.85, 16.79)	11.69 (7.34, 16.04)	5.66 (2.66, 8.66)
Total	10.35	9.24 (7.42, 11.06)	8.52 (6.77, 10.27)	4.25 (2.99, 5.51)

6.3.2 Extended plasmode analysis

Interval extension was performed to mitigate the effects of competing events. One would expect that the datasets that only considered PD mortality would have overestimated the number of PD events, and the final dataset with additional outcomes would have matched the original data more closely. This was supported by the average proportion of individuals with events in each SES category (Table 6.4). With only PD mortality included, 664.8 out of every 10,000 individuals in the simulated data experienced PD events. As additional outcomes were added, this number decreased, and the simulations of all causes with extended intervals found that an average of 8.5 individuals out of every 10,000 experienced a PD event (Table 6.4). As previously noted, most individuals in the original dataset left because of competing events. In the extended plasmode simulations, 99% of individuals who would have experienced a PD mortality event first experienced a competing event.

Table 6.4

PD Mortality Proportion (events per 10,000 and 95% confidence interval) by SES status in extended data.

	Original Data	PD Only	PD and lung	All Causes
Low SES	8.32	464.86 (440.58, 489.14)	415.08 (391.95, 438.21)	7.45 (3.88, 11.02)
Medium SES	9.71	631.45 (609.81, 653.09)	577.30 (557.27, 597.33)	8.20 (5.69, 10.71)
High SES	13.64	915.15 (880.89, 949.41)	870.90 (837.72, 904.08)	10.21 (6.21, 14.21)
Total	10.35	664.84 (650.33, 679.34)	614.31 (600.70, 627.93)	8.54 (6.77, 10.31)

The survival curves for each competing event combination showed similar trends. One would expect that individuals only at risk of PD mortality would have mostly survived, while individuals at risk of the full collection of outcomes would have survived at rates closely matching the original data. The average survival curves for the 500 simulations showed a clear decrease in survival with additional events up to a close match to the original dataset (Fig. 6.17). There were slight differences between the original survival curve and the simulations with the full set of outcomes, which were believed to be the result of inaccuracies in the cumulative hazard models for the full set of outcomes and variance in the simulation survival curves.

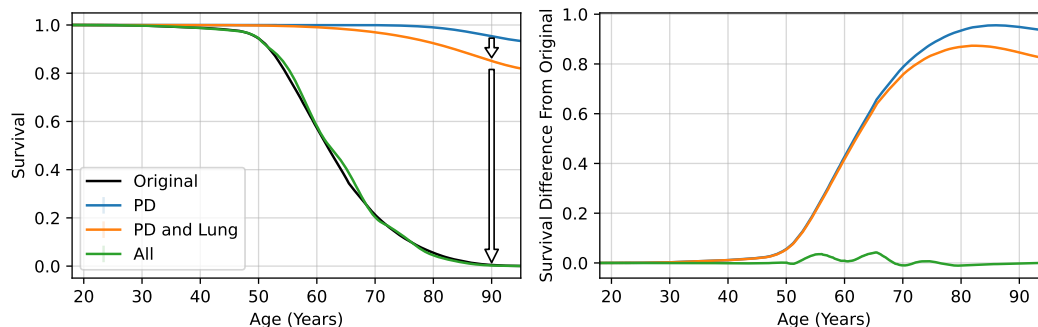


Figure 6.17: Average survival curve for overall observation for each competing risk simulation in extended data. As more events were added, the survival dropped (shown by arrows) until the simulation with all events closely matched the original data. The variation in survival curves among simulations was too low to visualize.

Finally, the regression results are presented. Once again, the average hazard ratio estimates converged within 500 simulations (Fig. 6.19). In general, there was a much narrower range of values than in the unextended analysis results (Fig. 6.15). This trend continued for the different event combinations. The simulations with the full set of outcomes tended to be the furthest from the true value, similar to the case with the unextended simulations. However, the distribution was narrower than the unextended sampling counterpart and significant at a 95% confidence level (Fig. 6.18).

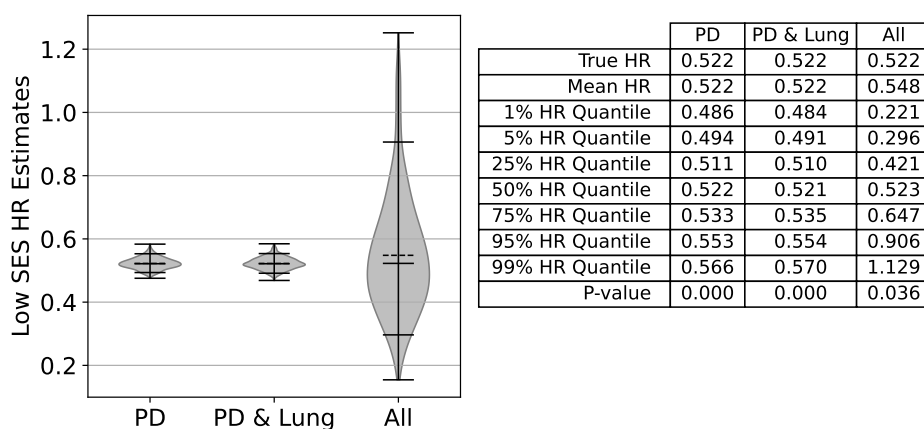


Figure 6.18: Violin plots of the low SES HR distributions for each grouping of events in extended data. The distributions for only PD and PD & lung were much narrower, while the distribution for the full set of outcomes was wider but was still significant at a 95% confidence interval.

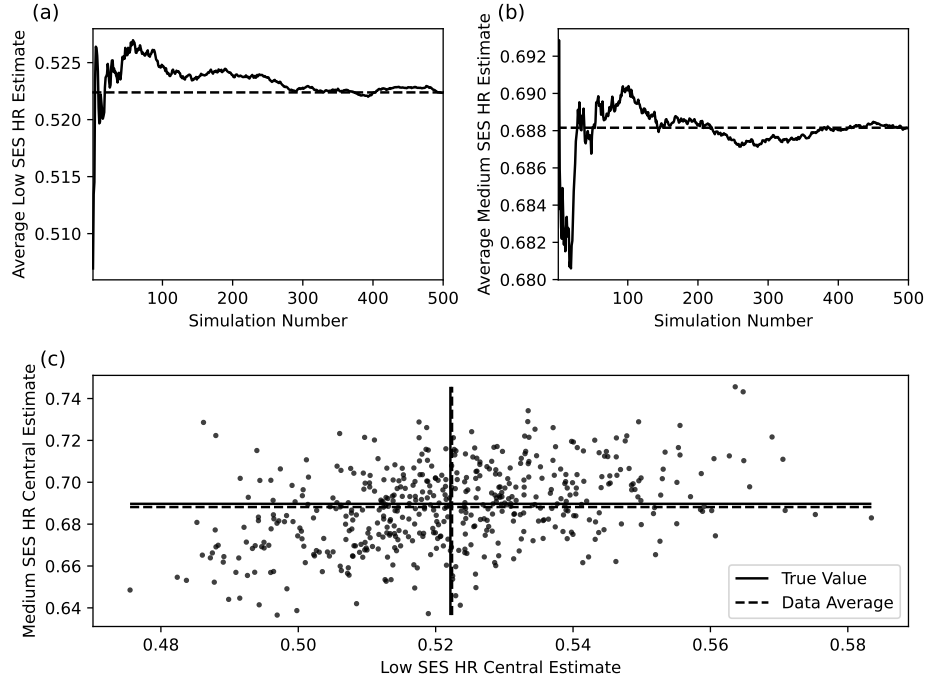


Figure 6.19: HR parameter estimates for SES status on PD mortality in the extended data without competing risks modeled. (a-b) The average parameter estimates converged within 500 simulations and (c) there was a much narrower cluster of points around the true values.

6.3.3 Plasmode method comparison

Extended plasmode simulations were used to simulate both single events and combinations of events, and unextended plasmode simulations were used to simulate events in addition to those in the original data. As more events were added to the simulations, the extended plasmode simulation survival curves approached the original data (Fig. 6.20.a.). The distributions in final observation times between the original data, extended competing risk simulations, and unextended PD simulations were all within a 5% difference (Fig. 6.20.b.).

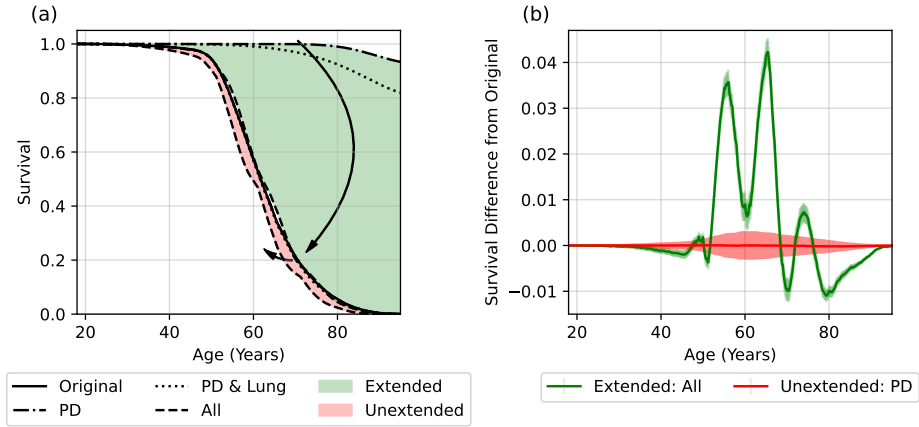


Figure 6.20: (a) As additional events were added, the extended simulations approached the original survival curve, and the unextended simulations dropped below the original survival and (b) the closest simulations were within 5% of the original survival curve on average. 95% confidence bands were produced for the survival differences, the unextended survival was not significantly different while the extended survival was significantly different.

Overall, a similar trend was observed in the HR distributions as was found with the survival curves. The unextended simulations with only PD modeled produced HR estimates similar to those of the Cox regressions of the extended simulations with the full set of outcomes modeled. Simulations with multiple modeled events were analyzed using both Cox and Fine-Gray regressions to provide additional insight into any competing risk effects. The Fine-Gray and Cox regressions for the extended simulations with the full set of outcomes produced very similar model parameter distributions. This also indicated that lung cancer mortality did not significantly bias the Cox regression analysis results.

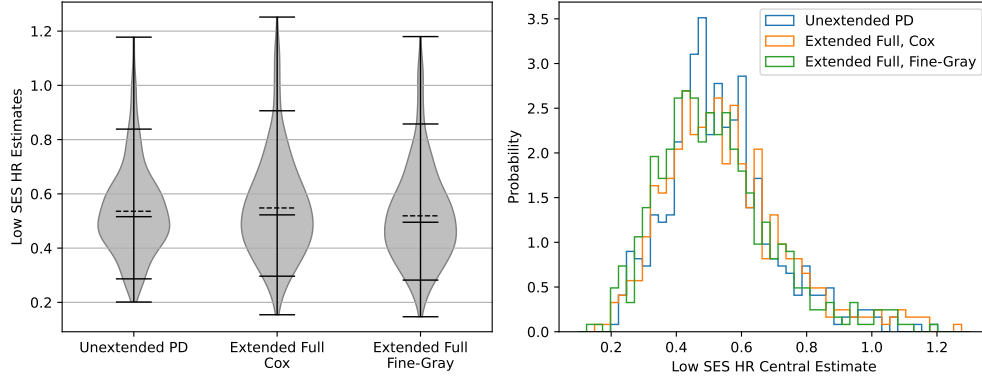


Figure 6.21: Histograms and violin plots for low SES HR central estimate distributions between the unextended simulations with PD and the extended simulations with a full set of outcomes for both Cox and Fine-Gray regressions. In both representations, the distributions were very similar.

Ultimately, the extended plasmode simulations with the full set of outcomes produced results similar to those of the original data. So, the models for hazard and background rates produced similar survival curves and relative proportions of events as the effects of competing risks in the original data. Similar results were found between the unextended plasmode simulations considering only PD mortality and the extended plasmode simulations considering all causes. Although not identical, the proportion of individuals with events (Fig. 6.22) and regression result distributions (Fig. 6.21) matched closely. This supported the hypothesis that single-event unextended plasmode simulations were equivalent to simulations with multiple competing events. The most significant difference between event proportions was in the subjects with high SES, which the extended simulations more noticeably under-predicted. This suggested that the loss of follow-up in the high SES subjects due to competing events did not match the true effect in the data, which could be a result of error in the hazard equations.

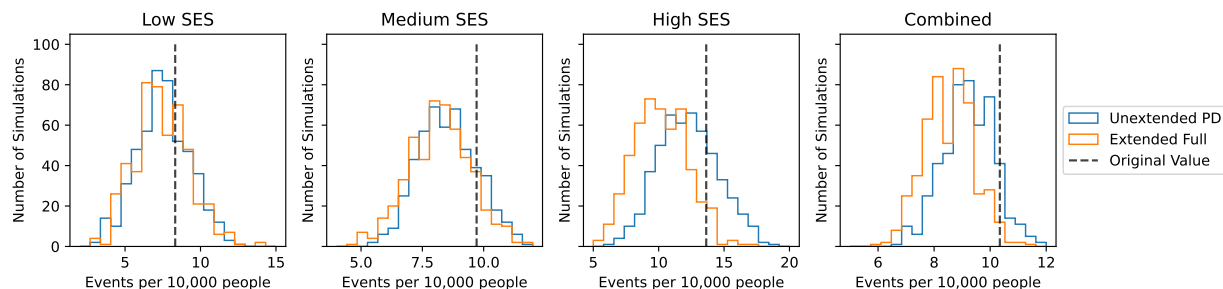


Figure 6.22: Distribution of PD mortality proportion (events per 10,000) by SES status comparing plasmode methods. The closest methods both underestimated the true proportion, and the extended simulations tended to underestimate the high SES event rate.

Finally, it is important to discuss how the similar effects observed in the unextended and extended simulation results were related to the original analysis. The proposed analysis was designed to test whether the predicted effects of smoking on PD mortality could have been biased by subjects at risk of lung cancer mortality. If this were true, one would expect the HR estimate to have been lower, further from 1, than the true value. However, the unextended simulations slightly overestimated the HR (Fig. 6.21). There were no significant differences in the median central estimates; however, large changes in the variance of the distributions were observed (Fig. 6.23). This was likely due to changes in the number of individuals at risk and the number of observed events. In this study, the presence of competing events produced significant variance in the hazard ratio. However, no significant bias was observed.

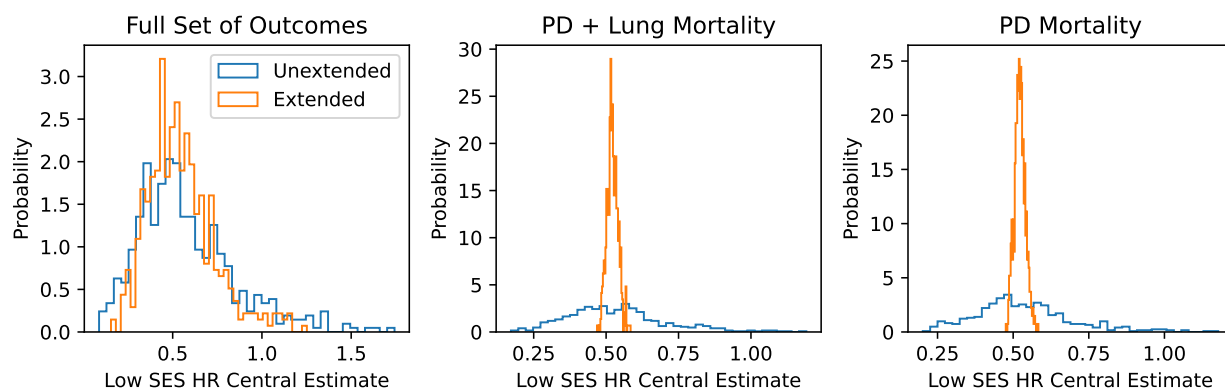


Figure 6.23: Comparison of low SES HR central estimates for each group of events and extension. In every comparison for the same events modeled, the unextended simulation distributions had higher variances.

6.4 Conclusions and Future Work

Ultimately, this analysis failed to find evidence of significant bias forming from competing events. However, through comparisons between the estimated hazard ratios and event rates, the lingering effects of the original survival times were identified and compared with the effects of the modeled events. The unextended plasmode simulations were not appropriate for the counterfactual analysis necessary to analyze competing event effects between PD and lung cancer mortality. However, this did not mean that the unextended plasmode simulations were not useful for competing risk analysis. The data would have been appropriate for simulations in which the simulated event could only have occurred before the original survival times. One example would be modeling the additional risk of an outcome and measuring excess cases. Suppose a researcher compared mortality rates between two similar cohorts and found that the second cohort was twice as likely to die from a workplace accident. They may be interested in knowing how the difference in workplace accidents may have changed the observed event distribution in the second cohort. They could simulate the difference in workplace accident probability on the population of the first cohort to estimate a distribution of additional workplace accident mortality events. The difference between the original and newly observed event rates in the first cohort could provide insight into the effect on event rates in the second cohort.

Multiple steps could be taken to improve the risk models and exposure prediction for a hypothetical future competing risk study using NPP cohort data. The first would be the use of the original, non-lagged organ doses. In this analysis, the only available dataset was the analytic working file, which contained only lagged organ doses. The non-lagged doses contained information on the next 10 years of exposures, which would have reduced the number of interval exposures that would need to be predicted and increased the amount of information on exposures at older ages. Another step that could be taken is to improve the accuracy of the risk model for other causes of death. In this analysis, all other causes of mortality and censoring were combined into a single event, depending on age and year of birth. The collection of mortality and censoring causes could have been broken down into

separate models for each mortality and censoring cause. If the hazard ratio and background rate for each competing event were modeled separately, the effect of risk factors on the combined risk could have been more accurately modeled. This would allow the study to more accurately model the effects of organ doses and personal characteristics on the combined risk of all other causes of mortality and censoring. Finally, other methods for predicting future exposure should be tested and compared. A breakdown of accuracy by risk factors would provide a clear visual of where these models performed well and what steps should be taken to more accurately fill in missing data.

Chapter 7

Theory and Applications of Covariate Uncertainty

Research in radiation epidemiology often takes the form of retrospective observational studies, which require the characterization of radiation exposure through measurements, calculations, or a combination of the two. The absorbed dose to organs for each person and time interval is estimated using dosimetric models. In the Million Person Study¹⁷, dosimetrists produce point estimates for the annual organ dose, which are then used to fit risk models for health outcomes⁵⁹. Parametric models of the risk of an outcome for a cumulative organ dose can be produced, including the characterization of model parameter uncertainties. However, these model uncertainties typically do not account for uncertainties in exposure variables.

Although methods to account for exposure uncertainty have been discussed in the literature, they are often developed for proprietary models and may be difficult to apply to large datasets^{60;61}. Methods that consider dosimetry uncertainty could be made more widely available by integrating them into a publicly available analysis package designed to accommodate complex risk models and large amounts of data. For this work described in this chapter, the capabilities of Colossus³⁵ were expanded by incorporating methods to propagate dosimetric uncertainty into the model parameter uncertainties by optimizing multiple realizations of exposure, based on recommendations made in NCRP Commentary 34, “Recommendations on

Statistical Approaches to Account for Dose Uncertainties in Radiation Epidemiologic Risk Models”²³. These methods typically optimize each realization separately before merging the results, or optimize the integrated log-likelihood.

The goal of this chapter was to apply the multiple-realization analysis functions available in Colossus to a simulated radiation exposure cohort with both shared and unshared exposure uncertainties. The results from the different analysis methods were compared to illustrate the differences. The main focus of this study was to compare the efficiency of performing these analyses with specially designed functions compared to manually performing each regression. These functions allow complex risk models to be applied to big data with complex uncertainties. Making these methods more widely available in open-source software will allow ongoing and future research studies to make full use of uncertain measurements in survival analyses.

7.1 Theory

All the specifics of how dose realizations are generated and where uncertainties originate have been described in great detail in the literature²³. The Two-dimensional Monte Carlo method⁵⁹ is briefly outlined in Section 7.2.2 to demonstrate one such method.

To begin, every row (i) of the dataset had an outcome, a measure of time, a measure of observed dose (d_i), and measures of other risk factors used in the model ($\vec{\theta}_i$). Cox proportional hazards models were used to solve for the hazard ratio of the outcome (r_i) as either a linear (Eq. 7.1) or exponential (Eq. 7.2) function of the cumulative dose.

$$r_{i,l} = (1 + \alpha_l \cdot d_i) \exp(\vec{\beta} \cdot \vec{\theta}_i) \quad (7.1)$$

$$r_{i,e} = \exp(\alpha_e \cdot d_i + \vec{\beta} \cdot \vec{\theta}_i) \quad (7.2)$$

There are two methods of accounting for exposure uncertainty that were focused on: Monte Carlo Maximum Likelihood (MCML) and Frequentist Model Averaging (FMA). Both

methods required optimizing the log-likelihood of a model with respect to the realizations, but the two differed in how the results were combined²³.

7.1.1 Justification of Implemented Methods

The most significant determining factor in which methods to incorporate was the most appropriate methods for the types of uncertainties in MPS cohort analyses. Random uncertainties can be divided into combinations of shared/unshared and classical/Berkson uncertainties. Shared errors arise from factors that have a shared value or distribution among subjects in a group, and unshared errors arise from factors that are statistically independent among subjects²³. A group of people in a house near a nuclear weapon test has the shared factor of house position relative to the bomb but has unshared factors for their relative positions in the house. For classical errors, the estimate varies independently from the true value by a measurement uncertainty. For Berkson error, the true value varies independently from the estimate by individual deviation. The distinction is which measurement is independent of uncertainty²³. Berkson error commonly occurs when every member of a group is assigned a single averaged value; thus, the true values are distributed around the assigned estimate. Classical errors commonly occur when a measurement instrument has some imprecision; therefore, the estimates are distributed around the true value.

Ideally, methods implemented in Colossus could be used for a wide variety of cohorts, regardless of the types of errors present. Several methods presented in NCRP commentary 34 are designed only for specific types of errors. Simulation extrapolation (SIMEX) only works for purely classical errors, and Corrected information matrix (CIM) analysis is only relevant for data with primarily Berkson errors. Some methods cannot account for shared errors, like first-order regression calibration²³. The report concluded by listing the most generally applicable methods as regression calibration, Monte Carlo Maximum Likelihood (MCML), Bayesian Markov Chain Monte Carlo (MCMC), and Frequentist Model Averaging (FMA)²³. From these, MCML and FMA were selected as two generally appropriate methods that could handle the level of uncertainty and complexity required for future MPS cohort

analyses. MCML and FMA were selected as two common methods for combining the results of multiple models. MCML requires a code to average the log-likelihood over multiple models, and FMA requires a code to optimize multiple models separately. These methods may be useful outside exposure uncertainty analysis, like combining estimates from multiple similar models.

7.1.2 Monte Carlo Maximum Likelihood theory

The first method for accounting for multiple realizations was to optimize the expected log-likelihood. At the highest level, this required the integration of the log-likelihood over the probability distribution of the doses. Assuming that a sufficient number of independent random realizations were generated, the integral could be approximated using a sum (Eq. 7.3).

$$E \left[LL(\alpha, \vec{\beta}, \vec{d}_j, \boldsymbol{\theta}) \right] \approx \frac{1}{N} \sum_{j=1}^N \left(LL(\alpha, \vec{\beta}, \vec{d}_j, \boldsymbol{\theta}) \right) \quad (7.3)$$

In practice, this could have been done in two ways using existing software:

1. Using an existing regression function to evaluate every realization on a multi-dimensional grid of parameter values, and then averaging the log-likelihoods to find the optimum grid point.
2. Manually using the Newton-Raphson algorithm by iteratively calculating the derivative matrices for each realization at a given parameter set and then updating the parameter values based on the average derivatives.

The first method was the easiest to implement and could have been performed using existing analysis software. However, this could require a significant number of grid points for complicated risk models. The second method could have been performed in Colossus, which returns derivative matrices and would require fewer calculations than evaluation on a

grid. In both cases, a significant number of processing steps would have to be repeated for every realization calculation. The motivating factor for implementing the MCML analysis function in Colossus was to make MCML more generally available.

7.1.3 Frequentist Model Averaging theory

The second method was to optimize each realization and then use the relative Akaike information criterion (AIC) scores to produce weighted means and confidence intervals for the model parameters for each realization. The AIC is a metric that accounts for both the log-likelihood of the model and the number of covariates used (k)³⁸. In many circumstances, the AIC is used to weight the improvement in a model by the cost of overfitting associated with adding low-information variables. The analysis starts by optimizing the log-likelihood for each realization and then calculating the AIC (Eq. (7.4)). The set of AIC values would then be used to generate a set of weights for each realization (Eq. (7.5)). Each realization result would have a weight and a defined model parameter uncertainty distribution. A large number of total samples (M) would be selected, and $M * FW_j$ values would be drawn from each realization model parameter distribution. These combined samples produce a new distribution for the dose model parameter, accounting for both exposure and model parameter uncertainties.

$$AIC_j = 2(k - LL_j) \tag{7.4}$$

$$FW_j = \frac{\exp(-AIC_j/2)}{\sum_{l=1}^N \exp(-AIC_l/2)} \tag{7.5}$$

Compared to MCML, this was much easier to implement with existing functions. However, depending on the number of realizations and the scale of the data, a substantial amount of time may be spent on repeated steps. The degree of repeated steps is analyzed in greater detail in Section 7.3.1.

7.1.4 Application to big data

It was critical to consider how these methods needed to be adapted to be applied to thousands of realizations, each with potentially millions of rows of data. While the original formulae were mathematically correct, they were impractical to implement without simplification. The most immediate issue was the rounding of the FMA weights (Eq. 7.5). It would not have been unusual for the AIC to be on the scale of tens of thousands, which meant that the exponential terms were likely to be very close to zero. Rounding issues were avoided by adjusting the AIC using a reference value (Eq. 7.6-7.7).

$$FW_j = \frac{\exp(AIC_{ref}/2)}{\exp(AIC_{ref}/2)} \frac{\exp(-AIC_j/2)}{\sum_{l=1}^N \exp(-AIC_l/2)} \quad (7.6)$$

$$FW_j = \frac{\exp(-\Delta AIC_j/2)}{\sum_{l=1}^N \exp(-\Delta AIC_l/2)} \quad (7.7)$$

Three reasonable guesses for the appropriate reference AIC were the minimum, average, and maximum AIC. If the range of the AIC values was on a scale of 10 to 100, then all three would be reasonable. However, if the differences were on the scale of thousands, there would have been an overflow (Table 7.1). In all cases, setting the reference AIC to the minimum AIC kept the exponential terms below 1. This also limited the sum of the exponential terms to N and kept the sum from experiencing overflow.

Table 7.1

Example of weight overflow and underflow by reference AIC. The difference between AIC was large enough that some subset of the exponential values had to be forced to near 0 or ∞ .

AIC	$\exp(-AIC/2)$	$\exp(-\Delta_{min}/2)$	$\exp(-\Delta_{avg}/2)$	$\exp(-\Delta_{max}/2)$
10,000	0	1	7.23e86	1.40e217
10,100	0	1.93e-22	1.39e65	2.71e195
10,500	0	2.67e-109	1.93e-22	3.75e108
11,000	0	7.12e-218	5.15e-131	1

7.2 Methods

7.2.1 Implementation of analysis methods into Colossus

When the user calls a regression function, Colossus runs several functions to clean and prepare the data before running iterations (Fig. 7.1). These include steps such as removing rows and columns that are not used during the analysis, checking for errors in the function inputs, and predetermining indices for rows in each risk group. In a multiple realization analysis, the time and event columns would not change, and the same row filtering and risk group calculations would be repeated. Repeating these steps could be avoided by designing a new regression function that performs these processing steps on the entire dataset of realizations and then switches between realizations in the code.

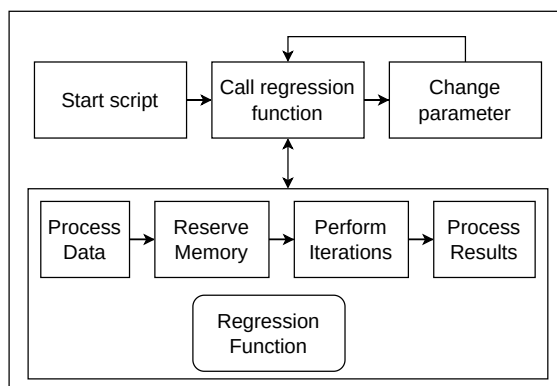


Figure 7.1: When a user calls a regression function, there are multiple steps required to process the data and reserve memory that can require significant amounts of time for large datasets. In the case of multiple realizations, these steps were identical between realizations.

For the MCML analysis using the Newton-Raphson method³⁷, the function started by calculating the log-likelihood and derivatives and accumulated the weighted sum for each realization. The final accumulated values were then used to estimate the next step using a derivative-based algorithm (Fig. 7.2). This method returned a single model parameter central estimate and uncertainty. For the FMA analysis, the function iteratively ran regres-

sions on each realization and then stored the results to combine at the end (Fig. 7.3). In many cases, the model parameters for known exposures may be relatively constant. So, the function used the results from one regression as the initial parameter estimate for the next. In the best-case scenario, doing so started each regression close to the optimal parameter set and reduced the regression time.

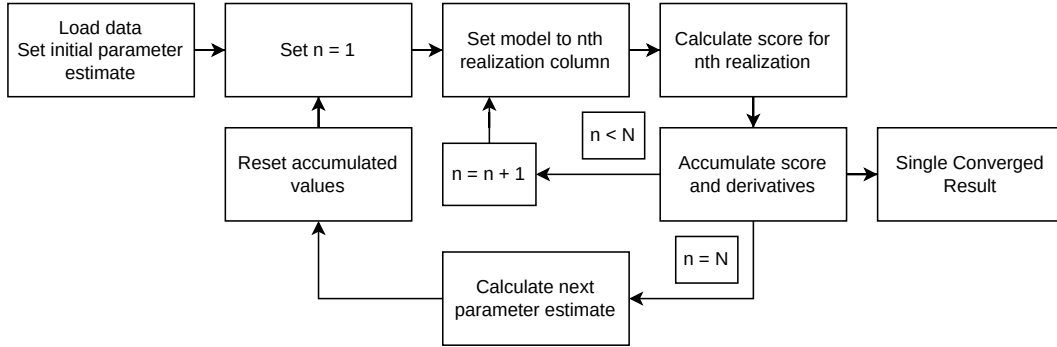


Figure 7.2: MCML analysis in Colossus ran similarly to standard regressions: scores were calculated for each realization and the accumulated weighted sum was used to estimate the optimal parameter set. After the analysis was complete, the results were returned.

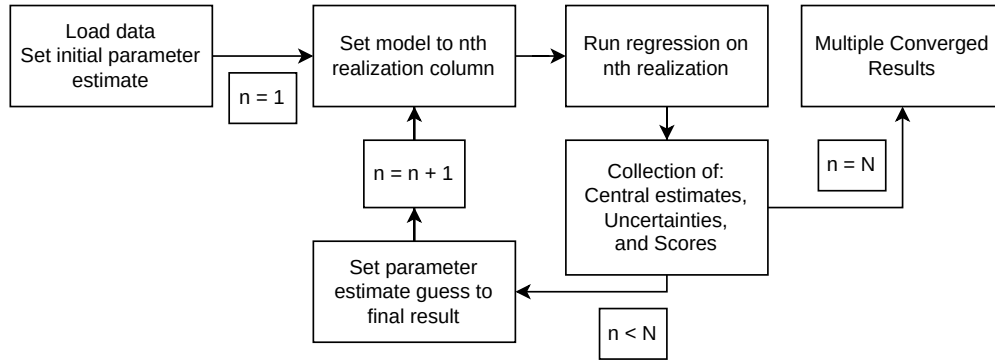


Figure 7.3: FMA analysis in Colossus was based on typical user steps. Each realization was optimized and then used as an initial guess for the next realization. After the analysis was complete, the lists of estimates, uncertainties, and scores were returned for further processing.

As discussed earlier, one of the goals of this study was to benchmark the time taken by each method using Colossus. Functions were written in Colossus for FMA and MCML analyses, so reasonable alternative analysis methods were prepared to measure the speedup. For the FMA analysis, the alternative was clear; each realization could be optimized individually, so the alternative time required per realization for FMA was the average time required

to optimize a single realization.

The alternative to MCML was less straightforward. Colossus, by default, returns the vector and matrix of the first and second derivatives of the log-likelihood with respect to each of the model parameters. The score and derivatives can be calculated for every realization and parameter set and then manually combined outside Colossus. This would likely require fewer calculations than the evaluation of points on a multidimensional grid. The number of iterations used for the analysis in the next section was based on the average number of iterations reported during the benchmarking. The total number of calculations performed during a regression is a combination of the number of iterations and any half-step calculations; however, Colossus does not report the number of half-steps taken. This meant that the estimate of five iterations per analysis was likely an underestimate of the true number of calculations necessary.

Another important difference to note was how to interpret the change in time per realization between different numbers of realizations. For both cases, adding more realizations was expected to increase the time required to process and store the underlying data. Suppose benchmarking was performed at N and $N + 1$ realizations. Optimizations are performed independently for the FMA analysis, so the first N realizations would have been optimized with the same number of iterations in both regressions. Optimizations are not performed independently for the MCML analysis, and adding more realizations may have changed the number of iterations required between regressions. The benchmarking time for MCML was defined as the time required to optimize the model, which was not necessarily the time required to run a set number of iterations. If the average time per iteration decreased and the average number of iterations increased, then the overall time required per realization could have still increased. This distinction was important for understanding how the benchmarking results could be applied to future work.

7.2.2 Simulated radiation cohort data

To begin, a study population of 20,000 people was randomly generated. Each person entered the study between the ages of 18 and 35 and between the years of 1930 and 1950, with a constant smoking status, employment category, and sex. Annual interval data were generated for each person from their entrance age to 95 years. The annual radiation exposures were generated following the examples in a previously published work on two-dimensional Monte Carlo methods⁵⁹.

The annual dose for each person was calculated using a parametric exposure model as a function of activity level, exposure duration, shielding, and conversion factor from radiation exposure to damage. The activity level was shared by exposures in each decade and was distributed as a log-normal distribution, peaking in the 1960 to 1969 bin. Values above 1×10^3 were rounded down to keep the cumulative exposure under one Gray (Fig. 7.4).

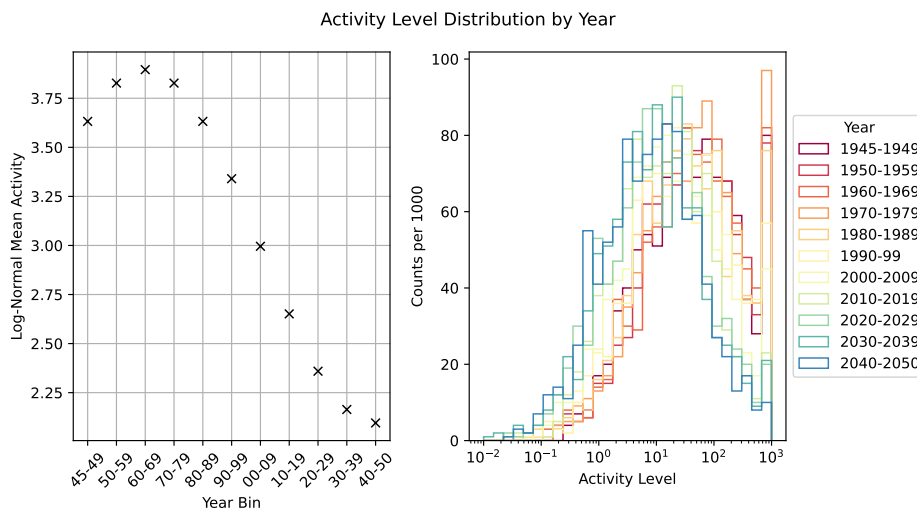


Figure 7.4: Activity level was shared between subjects working in the same decade, as log-normal distributions capped at 1×10^3 units.

The exposure duration was shared by age in 10-year bins and was normally distributed. The exposure peaked in the subjects aged 30 to 40 and was close to zero above age 80. Random durations below zero were rounded up so that every subject had a non-negative exposure duration (Fig. 7.5).

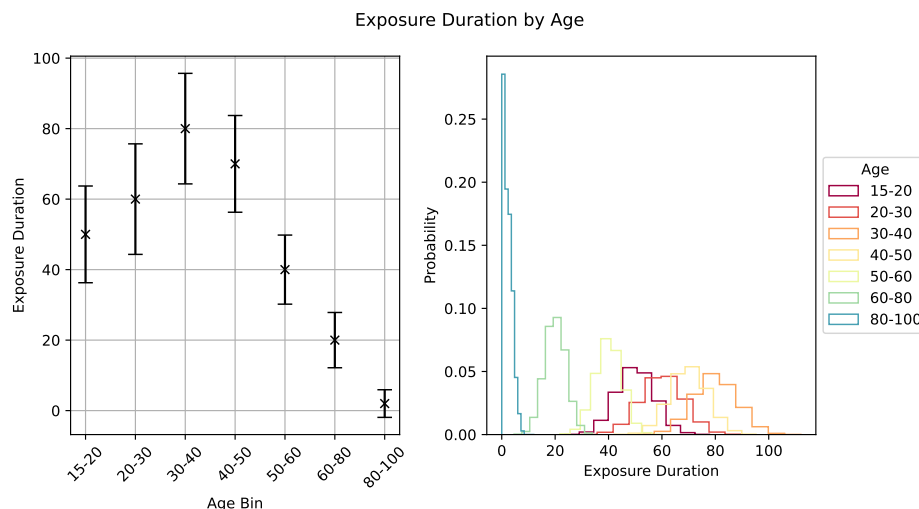


Figure 7.5: The measured exposure duration for each subject was shared by age and distributed normally. Exposures were rounded so that every subject had a non-negative exposure duration.

The shielding was shared among the three working cohorts and was distributed uniformly. The first working group was assigned a baseline protective factor of one. The second working group received between 20% and 99% of the baseline exposure, and the third working group received between 90% of the baseline and 50% higher than the baseline. On average, the three working groups were assigned relative shielding factors of 1.0, 0.595, and 1.2 (Fig. 7.6).

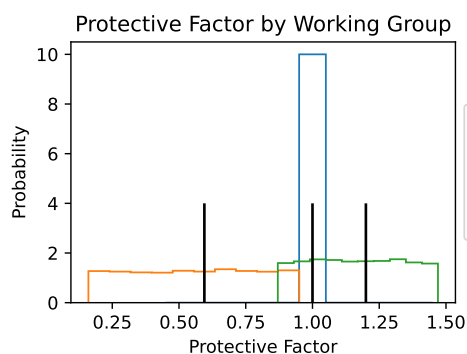


Figure 7.6: Each working group was assigned a shielding factor relative to working group one. The second working group generally received less radiation and the third working group generally received more radiation than baseline.

Finally, a conversion factor from 1.5×10^3 to 4.5×10^3 was uniformly assigned to each interval. These values were selected arbitrarily to create significantly different dose histories

between subjects born in different years and assigned to different working cohorts. The mean values and standard deviations were adjusted to avoid outlier subjects receiving cumulative doses on the scale of hundreds of Gray. Overall, the subjects tended to receive annual doses between 1×10^{-3} and 1×10^3 mGy, and the shared uncertainties produced large differences between the dose distributions in each realization (Fig. 7.7).

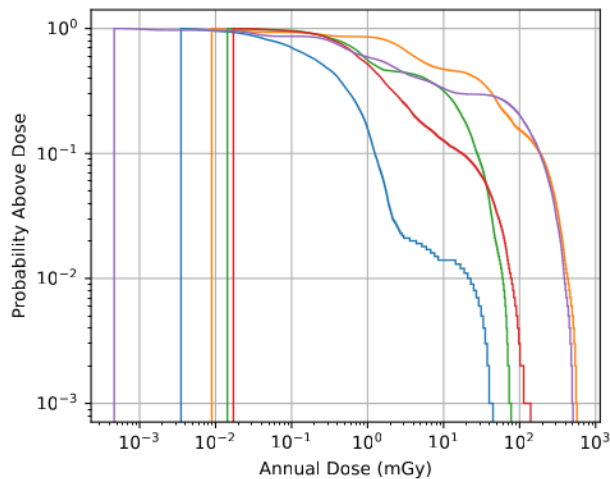


Figure 7.7: Dose probability distributions for the first five realizations generated. The level of shared uncertainties produce significant differences in the curves, with some realizations having no intervals above 100 mGy annual dose and others having over 90% of intervals with annual dose above 100 mGy.

Using the two-dimensional Monte Carlo generation method, each realization was created by generating the shared parameters once per realization and generating the unshared parameters uniquely for each interval. The first realization was then used to generate event times for each person, and the person data were filtered into intervals before and including the first event time for each person. Finally, the constant covariate data was joined with the dose realizations to create a dataset for analysis.

Multiple uncertainty combinations and event rate models were applied and analyzed independently in this study. Two versions of the realizations were created. The first used the full exposure model with both shared and unshared uncertainties. The second method used the average value of the shared factors and only applied unshared uncertainties (Figures 7.4-7.6). Two hazard ratio equations were applied: one linear with respect to the cumulative

dose and the other exponential with respect to the cumulative dose (Eqs. 7.1-7.2). All four combinations were analyzed independently to compare the regression results and differences in benchmarking. Although the main intent of this study was not to compare the results between FMA and MCML, the regression results were compared to visualize the effects of shared error and linear risk models on the method performance. The most significant difference in benchmarking was the change in the hazard ratio model. In Colossus, purely exponential models are analyzed using simplifications in the likelihood equations, which decrease the average time spent per iteration. Running benchmarking with both linear and exponential models provided a more complete view of the potential speedups and variations in the time required for different datasets.

7.3 Results and Discussion

The regression results and run time were used for comparison. For every combination of uncertainties (primarily shared or only unshared) and risk models (exponential or linear), 5000 realizations were analyzed using the FMA and MCML methods to compare the parameter uncertainties. Each combination of uncertainties, risk models, and analysis methods was also benchmarked with different numbers of realizations to compare the average runtimes. The benchmarking was performed using 16 cores on a system with two Intel[®] Xeon[®] Silver 4210R, 2.40 GHz processors, and 500 GiB of DDR4 RAM.

7.3.1 Benchmarking

First, the exponential risk model analysis is presented. Each realization was optimized in approximately 8 seconds, and 0.5 seconds were spent running iterations (Table 7.2). The majority of the time spent in the regression was due to repeated steps, such as processing the input table, assigning memory, and locating the risk groups. Thus, one would expect that the time spent per realization would decrease significantly as the number of realizations increases.

Table 7.2

Average single realization regression times and 95% confidence interval (seconds) for the FMA and MCML alternative options and exponential risk model.

	Full Optimization	Single Iteration
Shared and Unshared	8.42 (8.12, 8.72)	8.02 (7.74, 8.30)
Only Unshared	7.59 (7.23, 7.96)	7.10 (6.81, 7.39)

Benchmarking was performed by running the multi-realization regression function using only the first N realizations. N was varied logarithmically from one to 5000, and each regression was repeated five to 15 times. As expected, the average time per realization decreased with an increase in the number of realizations (Fig. 7.8). The serial analysis (FMA) took approximately 8 seconds per realization at one realization and less than 1 second per realization for more than 20 realizations. The integrated analysis (MCML) took approximately 8 seconds per realization for one realization and less than 1 second per realization for more than 50 realizations. For both analysis methods, the automated functions performed much faster than the alternatives. The estimated alternative run-time duration for the MCML analysis was approximately 40 seconds per realization.

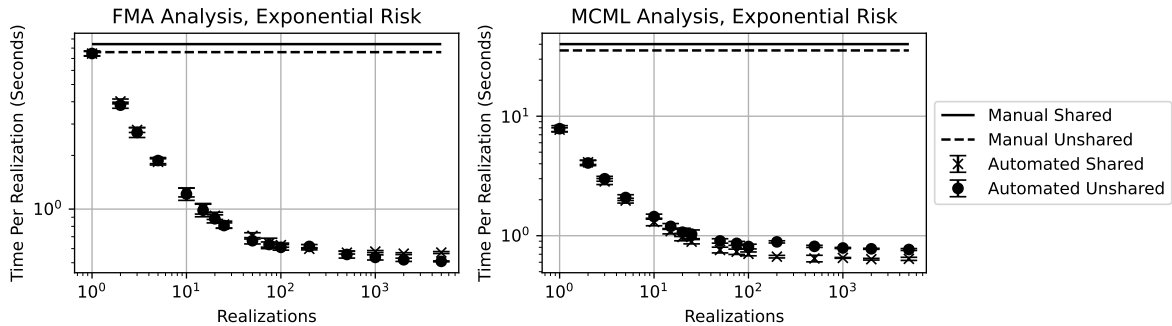


Figure 7.8: Benchmarking for the exponential risk models. Using the multi-realization functions in Colossus reduced the average time per realization from 8 seconds down to under 1 second for the FMA and MCML analyses. 95% confidence intervals for duration showed very little variation.

Next, the benchmarking of the linear risk models is presented. As previously mentioned, Colossus applies several simplifications to improve the speed of the purely exponential model

regressions. This meant that linear risk models were expected to run iterations more slowly than exponential risk models. In general, each realization optimization was approximately 8 seconds and 0.8 seconds were spent running iterations (Table 7.3). A majority of the time spent during each realization was spent on processing steps, so one would expect to have seen improvements in the multiple-realization functions. The average time spent iterations was higher, so one would expect to see less of an improvement than what was found in the exponential model benchmarking.

Table 7.3

Average single realization regression times and 95% confidence interval (seconds) for the FMA and MCML alternative options and linear risk model.

	Full Optimization	Single Iteration
Shared and Unshared	7.74 (7.41, 8.07)	6.96 (6.63, 7.28)
Only Unshared	8.19 (7.92, 8.46)	7.36 (7.06, 7.66)

Similar to the exponential model benchmarking, the serial analysis with one realization and the alternative of manually optimizing realizations both were around 8 seconds per realization, and the improvement continued up to approximately 100 realizations. For the linear model, the asymptote was approximately 1 to 2 seconds per realization on average. The increase in the time spent running iterations translated to a higher average time per realization. There was still considerable speedup, particularly in the MCML analysis, which was estimated to have taken around 40 seconds per realization on average.

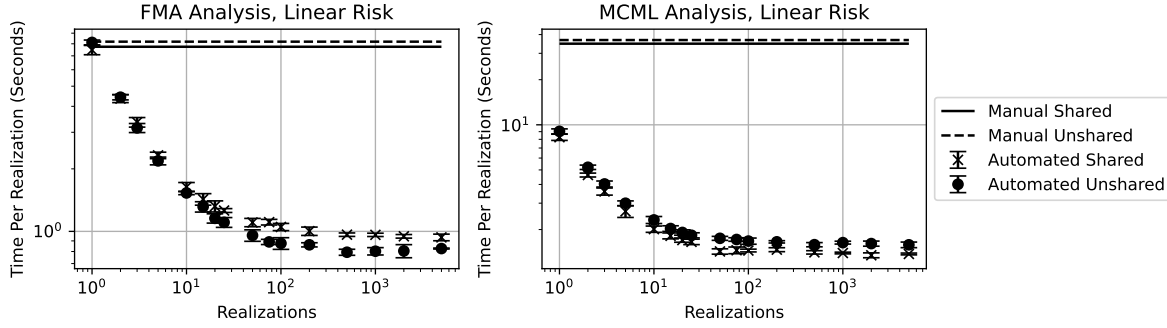


Figure 7.9: Benchmarking for the linear risk models. Using the multi-realization functions in Colossus reduced the time per realization from 8 seconds down to around 1-2 seconds for the FMA and MCML analyses respectively. 95% confidence intervals for duration showed very little variation.

The automated functions in Colossus were several times faster than the alternative calculation methods, particularly the MCML analyses. These results were specific to this dataset; however, in general, the multi-realization functions will reduce the time required. These functions avoided the large number of repeated processing steps encountered in alternative methods and made MCML analysis feasible for complex risk models.

7.3.2 Regression results

Next, the regression results are discussed. For every combination of uncertainty and risk models, all 5000 realizations were analyzed using both the FMA and MCML methods. The FMA realization optimizations were combined using the weights from Equation 7.6, and one million samples were used. The weighted FMA distributions were compared with the normal distributions produced from the MCML analyses. These were repeated for known and uncertain variables.

The model parameter distributions produced by the FMA and MCML analyses for the known variables closely matched for the exponential risk model data (Fig. 7.10). In both cases, the predicted parameter distributions were normally distributed. This is an important distinction because the MCML parameter distributions used Wald confidence intervals. Thus, the MCML results could only produce a normal distribution.

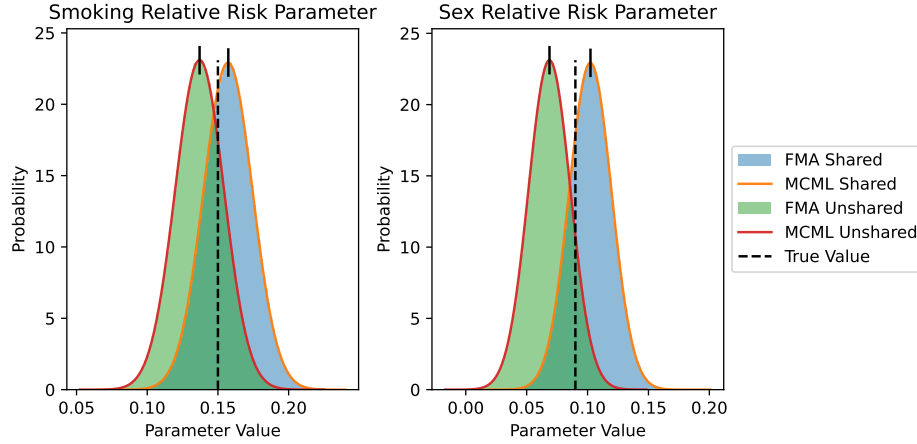


Figure 7.10: For the exponential risk model data and both combinations of shared and unshared uncertainty in dose, the predicted model parameter distributions for the known variables matched in both FMA and MCML analyses.

There were considerable differences in the distributions of the cumulative dose parameters. For the exponential risk model and purely unshared uncertainty, the FMA and MCML estimates closely matched. However, for shared uncertainty, the FMA analysis produced a non-normal distribution with a different central estimate from the MCML estimate (Fig. 7.11). In all cases, the central estimates were biased toward the null, but the unshared data produced a closer central estimate than the shared data. There were two likely causes for the differences in the distributions produced from the primarily shared uncertainty data. The difference in the central estimate was believed to be due to differences in the biases between the two methods. The difference in the distribution shape was believed to be due, in part, to the use of Wald confidence intervals. NCRP commentary 34 mentions that FMA could be used with both Wald confidence intervals as well as likelihood-profile based standard error estimates²³. Likelihood-based confidence intervals could be calculated for the MCML analysis to test whether the true parameter distribution matched a normal distribution.

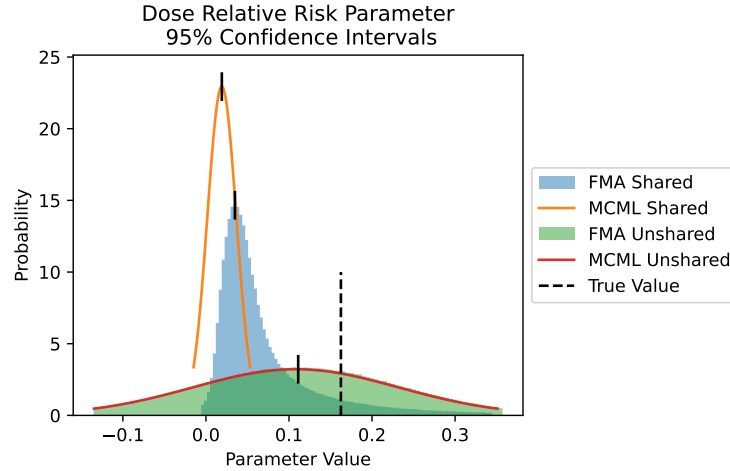


Figure 7.11: For the exponential risk model data, the FMA and MCML results differed between uncertainty types. The FMA and MCML analyses with only unshared uncertainty matched closely.

Next, the linear model regression results are presented. The model parameter distributions for smoking status and biological sex between the linear and exponential models were similar. In all cases, the FMA and MCML analyses predicted matching normal model parameter distributions (Fig. 7.12).

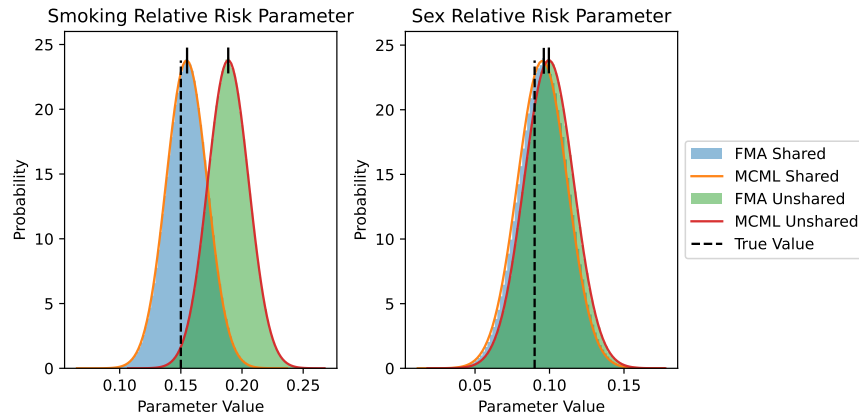


Figure 7.12: For the linear risk model data and both combinations of shared and unshared uncertainty in dose, the predicted model parameter distributions for the known variables matched in both FMA and MCML analysis.

Similar to the exponential risk model, the dose model parameter distributions for the linear model and shared uncertainty did not match. The analysis of the primarily shared uncertainty realizations was biased toward the null, with differently shaped distributions.

However, the unshared distribution analyses matched. Once again, the difference in the central estimate and distribution shape was believed to be due to differences in bias between methods and possible inaccuracies in the Wald confidence interval assumptions. Both the MCML and analysis and the individual realization analyses, the model parameter distributions likely violated the Wald confidence interval assumptions.

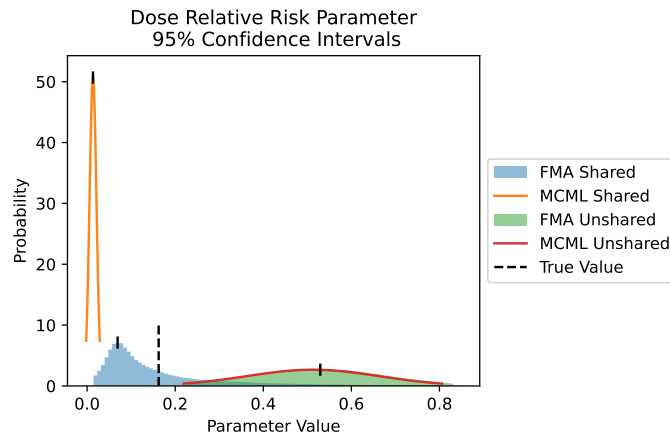


Figure 7.13: For the linear risk model data, the FMA and MCML results differed between uncertainty types. The FMA and MCML analyses with only unshared uncertainty matched closely but the shared uncertainty data did not match.

The FMA analyses with shared uncertainty for both models included the true value in a 95% confidence interval (Fig. 7.14). The two distributions were slightly different; the exponential distribution was narrower but included more negative values. However, it is important to note that both distributions were skewed right and underestimated the true effect of the dose. The analyses with unshared uncertainties for both models matched the MCML counterparts, and the linear model distribution did not include the true value in a 95% confidence interval. This suggested that the influence of shared uncertainties changed the level of bias and model parameter distribution shapes between methods.

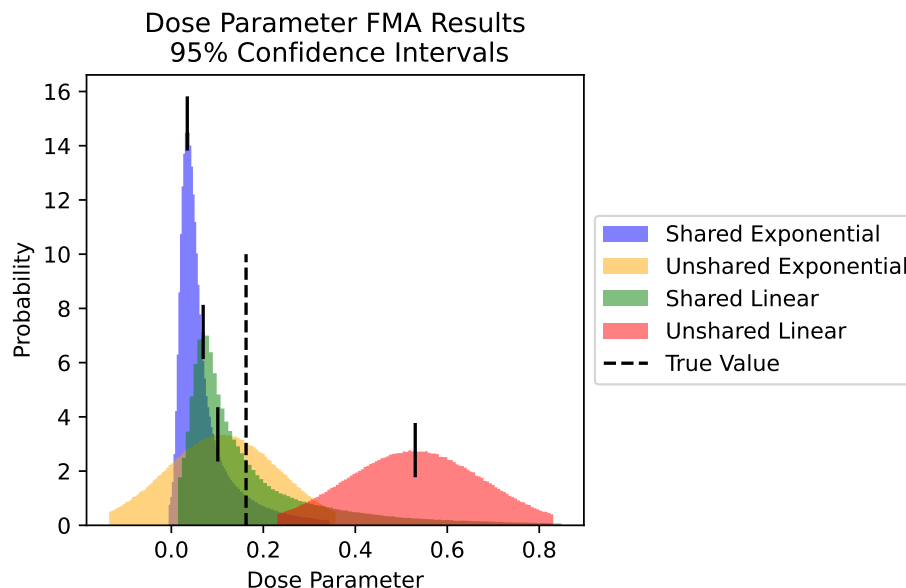


Figure 7.14: The FMA analyses with shared uncertainty produced skewed distributions. In both cases, the true value was included in the 95% confidence interval.

7.4 Summary of Results

For various combinations of uncertainty and risk models, Colossus was able to replicate methods that were not available in other open-source software. This not only highlighted the capabilities of Colossus for researchers wanting to apply complex models to uncertain data, but also demonstrated the importance of having software that was designed for these purposes. Widely available software packages could have been used to run FMA analysis, but doing so would have been time-consuming. In Colossus, using multi-realization functions reduced the time required by a factor of 4-8. The potential speedup for MCML analyses was even more significant, even more so in comparison to calculating scores on a grid.

This was the first step in incorporating multiple-realization analysis into Colossus. At this stage, the functions were tested for accuracy and were demonstrated to perform faster than the alternative. The next steps are to apply these methods to real data and incorporate other methods of applying exposure uncertainty. These methods were developed in preparation for a planned analysis of the Tennessee Eastman Corporation (TEC) MPS cohort⁶². An

important step in this future analysis will be to determine which methods of incorporating exposure uncertainty are the most appropriate, given the types of uncertainty in the data. Additional exposure uncertainty analysis methods are planned to be added to Colossus at the request of the MPS team. Importantly, additional discussion and testing will focus on scaling the performance to 50 million person-year datasets.

Chapter 8

Future Work and Potential Improvements

The goal of this PhD work was to investigate the issues encountered in current studies of prolonged radiation exposure and to test software and methods capable of meeting emerging needs. The release of Colossus made the application of complex risk models to large amounts of data possible in conjunction with the data loading and processing functions available in R. The further development of Colossus to include FMA and MCML functions made the combination of uncertain exposures and complex risk models both widely available and feasible to perform over shorter time scales. Finally, the plasmode simulations indicated the potential for biases in time-series data and demonstrated how a machine learning model could produce plasmode simulations without bias. Colossus is currently being used by the MPS team to perform their analyses entirely in R and will continue to be updated with future analyses. In particular, new exposure uncertainty methods will be added based on MPS cohort uncertainty, and software development will continue to further reduce memory requirements.

8.1 Future Code Updates and Testing

The Colossus package will continue to be developed and used beyond this PhD work, so the following sections briefly cover different improvements and studies that are ongoing or planned.

8.1.1 Incorporating new capabilities and analysis methods

The most attainable goal for future Colossus development is the incorporation of additional capabilities. This includes both risk and likelihood model expansion, as well as the inclusion of additional data and result processing methods. One goal presented in NCRP Commentary 34 was that “A software repository should be created for as many methods of analysis as possible [...] Such methods and computational programs should be accessible to the public.”²³.

The framework of Colossus, a publicly available R package, places it in a position to help realize this ambition. Currently, only Cox proportional hazards models can be used for FMA and MCML analyses in Colossus. Future versions of Colossus are planned to include Poisson models and any new likelihood models. Another planned improvement is the addition of new likelihood and risk models. Currently, matched case-control regression is being added and tested, and logistic regression can also be added similarly.

In its current state, Colossus was able to meet the model-fitting requirements of the MPS. Further work is planned to improve the processing of data and results. One significant area of improvement is the methods used to compare models. Section 2.6 discussed existing methods of comparing model performance, and the methods commonly used in radiation epidemiology. Other methods are used in other disciplines, but have not been commonly used in radiation epidemiology. One example is concordance index and AUC calculations, which are difficult to apply to data with high rates of censoring^{63;64}. Alternatives that could be more appropriately applied to radiation cohort data have been proposed. These alternative metrics could be applied to radiation cohorts and implemented in Colossus. Providing more methods to compare models would allow the radiation epidemiology community to evaluate

risk models in depth and produce higher quality radiation risk models.

8.1.2 Further benchmarking and optimization of existing capabilities

Throughout the development of Colossus, the driving motivation behind code optimization has been testing with new data and analysis methods. Optimizations in risk group ordering and storage were performed because a user noticed that their regressions took longer when the intervals were grouped. The functions used to convert the input data into data-tables were changed because the multiple realization datasets were larger than those in previous tests, and the data-table conversion became the longest step. One goal of future development is to continue testing different types of data and models to identify bottlenecks. One type of stress test planned is benchmarking scripts with multiple input files and regressions to test how well memory is allocated and freed over a long script. A part of the appeal of Colossus was its ability to prepare a complicated script that could be run on a computing cluster. This scale of analysis can realistically include hundreds of regressions with different data files. Testing was previously performed using scripts that loaded a single data file and then ran a series of regressions. These tests may not have been sufficiently intensive to detect any issues.

8.1.3 Changes to libraries and code base

The most ambitious planned future development would be to change the underlying packages or switch the programming language entirely. As datasets become larger, it can be more time and memory-efficient to use libraries that operate on data without loading all the data into RAM. In particular, this could benefit multiple-realization functions, which store and access tables with thousands of columns. One popular option is Apache Arrow⁶⁵, instead of the Eigen matrices currently used. Eigen was designed to perform linear algebra and Arrow was designed to perform table operations. Most analyses in Colossus involve transforming and filtering tables, not matrix operations. There is a potential speedup in switching to

a library intended for the types of processing performed in Colossus. An early version of Colossus was written using Arrow tables to assign risk group indices and calculate risk values. At the time, this development version could run large regressions twice as fast as the default version of Colossus, which could be improved further by taking advantage of the changes to Colossus since then. The development version using Arrow was put on hiatus because of the difficulty in setting up the compilation instructions for the package. A Colossus version using Arrow would likely be faster but would require a more complicated installation. This issue largely arose from the fact that the necessary C++ library dependencies were managed separately from the R-package dependencies. Another option would be to switch programming languages to something like Python, which would directly handle the necessary dependencies. Ultimately, this future work would be dependent on balancing usability with performance, but taking advantage of new libraries will be important for improving Colossus.

8.2 Future Epidemiological Analyses

First, a multiple-realization analysis of the Tennessee Eastman Corporation MPS cohort⁶² is planned. This will mark the first large-scale, multiple-realization analysis for the Million Person Study using Colossus and will provide insight into how radiation exposure uncertainty may have affected previously published results. The analysis will include an in-depth discussion of the uncertainty structure in the data, the appropriateness of different uncertainty analysis methods, and a comparison with previously published results.

Next, an analysis of subjects exposed to multiple stressors is planned. One of the ongoing projects for the Million Person Study is investigating the combined effects of radiation exposure, isolation, and confined spaces, among other stressors, in a cohort of nuclear submariners⁶⁶ to predict the effects of similar stressors during space flight. The combination of stressors to which submariners are exposed is similar to the stressors experienced in space travel, and the submariner dataset provides a source of well-recorded combinations of stressors. This analysis will involve investigating the scope of effect modification and testing how well risk model structures describe the interaction.

Finally, Colossus was ultimately designed for future analyses of harmonized, merged MPS datasets. Other researchers are hard at work investigating how to combine different MPS cohorts, including how to handle individuals in multiple cohorts and how socioeconomic categories in different cohorts should be combined into a single harmonized scale. This will provide not only a quantification of how well different risk models fit, but also direct comparisons between the harmonized datasets and meta-analysis of the individual cohorts. This will also involve a detailed analysis of how the effects of socioeconomic status and dose differ between cohorts. This will be done by stratifying the treatment effects by cohort or modeling the effect modification by cohort. Stratification will allow the cohorts to be analyzed separately and produce a model that optimizes the average log-likelihood weighted by cohort size. Effect modification allows for specific treatment effects to be cohort-specific and other effects to be shared across all cohorts. This will provide further insight into the chronic radiation effects with sample sizes not previously available.

Bibliography

- [1] Committee on an Assessment of CDC Radiation Studies. *Radiation Dose Reconstruction for Epidemiologic Uses*. Online access: National Academy of Sciences National Academies Press. National Academies Press, 1995. ISBN 9780309051873. URL <https://books.google.com/books?id=x2grAAAAYAAJ>.
- [2] Lawrence T. Dauer, André Bouville, Richard E. Toohey, John D. Boice Jr., Harold L. Beck, Keith F. Eckerman, Derek Hagemeyer, Richard W. Leggett, Michael T. Mumma, Bruce Napier, Kathy H. Pryor, Marvin Rosenstein, David A. Schauer, Sami Sherbini, Daniel O. Stram, James L. Thompson, John E. Till, R. Craig Yoder, and Cary Zeitlin. Dosimetry and uncertainty approaches for the million person study of low-dose radiation health effects: overview of the recommendations in ncrp report no. 178. *International Journal of Radiation Biology*, 98(4):600–609, 2022. doi: 10.1080/09553002.2018.1536299. URL <https://doi.org/10.1080/09553002.2018.1536299>. PMID: 30452303.
- [3] Eric Giunta, Dawson Stutzman, Sarah S. Cohen, Benjamin French, Linda Walsh, Lawrence T. Dauer, John D. Boice Jr., Steve Blattnig, Dan Andresen, and Amir A. Bahadori. Colossus: software for radiation epidemiological studies with big data. *Journal of Radiological Protection*, 45(2):021504, apr 2025. doi: 10.1088/1361-6498/adcd80. URL <https://dx.doi.org/10.1088/1361-6498/adcd80>.
- [4] Dale L. Preston, Jay H. Lubin, Donald A. Pierce, Mary E. McConney, and Natalia S. Shilnikova. Epicure risk regression and person-year computation software: Command summary and user guide, 2015. URL <http://epicurehelp.risksciences.com/index.html#!Documents/userguide.htm>.

- [5] Terry M Therneau. *A Package for Survival Analysis in R*, 2023. URL <https://CRAN.R-project.org/package=survival>. R package version 3.5-7.
- [6] Lawrence T. Dauer, Linda Walsh, Michael T. Mumma, Sarah S. Cohen, Ashley P. Golden, Sara C. Howard, Grace E. Roemer, and John D. Boice. Moon, mars and minds: Evaluating parkinson’s disease mortality among u.s. radiation workers and veterans in the million person study of low-dose effects. *Zeitschrift für Medizinische Physik*, 34 (1):100–110, 2024. ISSN 0939-3889. doi: <https://doi.org/10.1016/j.zemedi.2023.07.002>. URL <https://www.sciencedirect.com/science/article/pii/S0939388923000843>. Special Issue: Space Radiation Research.
- [7] Merriam-Webster. epidemiology, 2025. URL <https://www.merriam-webster.com/dictionary/epidemiology>.
- [8] Dale Preston. Atomic bomb survivor studies: Overview and recent findings, 2019.
- [9] Damien Drubay, Sophie Ancelet, Alain Acker, Michaela Kreuzer, Dominique Laurier, and Estelle Rage. Kidney cancer mortality and ionizing radiation among french and german uranium miners. *Radiation and environmental biophysics*, 53, 05 2014. doi: 10.1007/s00411-014-0547-4.
- [10] UNSCEAR. Assessments of the radiation effects from the chornobyl nuclear reactor accident, 2021.
- [11] Kotaro Ozasa, Eric J Grant, and Kazunori Kodama. Japanese legacy cohorts: The life span study atomic bomb survivor cohort and survivors’ offspring. *Journal of Epidemiology*, 28(4):162–169, 2018. doi: 10.2188/jea.JE20170321.
- [12] Eric J. Grant, Alina Brenner, Hiromi Sugiyama, Ritsu Sakata, Atsuko Sadakane, Mai Utada, Elizabeth K. Cahoon, Caitlin M. Milder, Midori Soda, Harry M. Cullings, Dale L. Preston, Kiyohiko Mabuchi, and Kotaro Ozasa. Solid Cancer Incidence among the Life Span Study of Atomic Bomb Survivors: 1958–2009. *Radiation Research*, 187(5):

- 513 – 537, 2017. doi: 10.1667/RR14492.1. URL <https://doi.org/10.1667/RR14492.1>.
- [13] Wan-Ling Hsu, Dale L. Preston, Midori Soda, Hiromi Sugiyama, Sachiyo Funamoto, Kazunori Kodama, Akiro Kimura, Nanao Kamada, Hiroo Dohy, Masao Tomonaga, Masako Iwanaga, Yasushi Miyazaki, Harry M. Cullings, Akihiko Suyama, Kotaro Ozasa, Roy E. Shore, and Kiyohiko Mabuchi. The Incidence of Leukemia, Lymphoma and Multiple Myeloma among Atomic Bomb Survivors: 1950–2001. *Radiation Research*, 179(3):361 – 382, 2013. doi: 10.1667/RR2892.1. URL <https://doi.org/10.1667/RR2892.1>.
- [14] Richard E. Faw and J. Kenneth Shultis. *Radiological Assessment: Sources and Doses*. American Nuclear Society, 1999. ISBN 9780894484551.
- [15] National Research Council. *Health Effects of Exposure to Low Levels of Ionizing Radiation: BEIR V*. The National Academies Press, Washington, DC, 1990. ISBN 978-0-309-03995-6. doi: 10.17226/1224. URL <https://nap.nationalacademies.org/catalog/1224/health-effects-of-exposure-to-low-levels-of-ionizing-radiation>.
- [16] Ghassan B Hamra, David B Richardson, Elisabeth Cardis, Robert D Daniels, Michael Gillies, Jacqueline A O’Hagan, Richard Haylock, Dominique Laurier, Klervi Leuraud, Monika Moissonnier, Mary Schubauer-Berigan, Isabelle Thierry-Chef, and Ausrele Kesminiene. Cohort profile: The international nuclear workers study (inworks). *International Journal of Epidemiology*, 45(3):693–699, 07 2015. ISSN 0300-5771. doi: 10.1093/ije/dyv122. URL <https://doi.org/10.1093/ije/dyv122>.
- [17] John D. Boice Jr., André Bouville, Lawrence T. Dauer, Ashley P. Golden, and Richard Wakeford. Introduction to the special issue on the us million person study of health effects from low-level exposure to radiation. *International Journal of Radiation Biology*, 98(4):529–532, 2022. doi: 10.1080/09553002.2021.1989906. URL <https://doi.org/10.1080/09553002.2021.1989906>. PMID: 34612764.

- [18] John D. Boice Jr., Sarah S. Cohen, Michael T. Mumma, and Elizabeth D. Ellis. The million person study, whence it came and why. *International Journal of Radiation Biology*, 98(4):537–550, 2022. doi: 10.1080/09553002.2019.1589015. URL <https://doi.org/10.1080/09553002.2019.1589015>. PMID: 30831042.
- [19] John D. Boice Jr., Brian Quinn, Isaf Al-Nabulsi, Armin Ansari, Paul K. Blake, Steve R. Blattnig, Emily A. Caffrey, Sarah S. Cohen, Ashley P. Golden, Kathryn D. Held, Derek W. Jokisch, Richard W. Leggett, Michael T. Mumma, Caleigh Samuels, John E. Till, Sergei Y. Tolmachev, R. Craig Yoder, Joey Y. Zhou, and Lawrence T. Dauer. A million persons, a million dreams: a vision for a national center of radiation epidemiology and biology. *International Journal of Radiation Biology*, 98(4):795–821, 2022. doi: 10.1080/09553002.2021.1988183. URL <https://doi.org/10.1080/09553002.2021.1988183>. PMID: 34669549.
- [20] Michael T. Mumma, Sarah S. Cohen, Jennifer L. Sirko, Elizabeth D. Ellis, and John D. Boice Jr. Obtaining vital status and cause of death on a million persons. *International Journal of Radiation Biology*, 98(4):580–586, 2022. doi: 10.1080/09553002.2018.1539884. URL <https://doi.org/10.1080/09553002.2018.1539884>. PMID: 30412007.
- [21] Sarah S. Cohen, Michael T. Mumma, Elizabeth D. Ellis, and John D. Boice Jr. Validating the use of census data on education as a measure of socioeconomic status in an occupational cohort. *International Journal of Radiation Biology*, 98(4):587–592, 2022. doi: 10.1080/09553002.2018.1549758. URL <https://doi.org/10.1080/09553002.2018.1549758>. PMID: 30451561.
- [22] Ashley P. Golden, Sarah S. Cohen, Heidi Chen, Elizabeth D. Ellis, and John D. Boice. Evaluation of statistical modeling approaches for epidemiologic studies of low-dose radiation health effects. *International Journal of Radiation Biology*, 98(4):572–579, 2022. doi: 10.1080/09553002.2018.1554924. URL <https://doi.org/10.1080/09553002.2018.1554924>. PMID: 30499762.

- [23] Jonine L. Bernstein, Harry M. Cullings, Michael B. Bellamy, Benjamin C. French, Mark P. Little, Carmen D. Tekwe, and Helen A. Grogan. Recommendations on statistical approaches to account for dose uncertainties in radiation epidemiologic risk models (2024). *NCRP Commentary*, 34, 2024.
- [24] Lawrence T. Dauer, Michael T. Mumma, Julie C. Lima, Sarah S. Cohen, Daniel Andresen, Amir A. Bahadori, Michael Bellamy, David A. Bierman, Steve Blattnig, Benjamin French, Eric Giunta, Kathryn Held, Nolan Hertel, Laura Keohane, Richard Leggett, Loren Lipworth, Kathleen B. Miller, Ryan B. Norman, Caleigh Samuels, Kali S. Thomas, Sergei Y. Tolmachev, Linda Walsh, and John D. Boice Jr. A Million Person Study Innovation: Evaluating Cognitive Impairment and other Morbidity Outcomes from Chronic Radiation Exposure Through Linkages with the Centers for Medicaid and Medicare Services Assessment and Claims Data. *Radiation Research*, 202(6):847 – 861, 2024. doi: 10.1667/RADE-23-00186.1. URL <https://doi.org/10.1667/RADE-23-00186.1>.
- [25] John D. Boice Jr., Sarah S. Cohen, Michael T. Mumma, Ashley P. Golden, Sara C. Howard, David J. Girardi, Elizabeth Dupree Ellis, Michael B. Bellamy, Lawrence T. Dauer, Caleigh Samuels, Keith F. Eckerman, and Richard W. Leggett. Mortality among workers at the los alamos national laboratory, 1943–2017. *International Journal of Radiation Biology*, 98(4):722–749, 2022. doi: 10.1080/09553002.2021.1917784. URL <https://doi.org/10.1080/09553002.2021.1917784>. PMID: 34047625.
- [26] John D. Boice Jr, Sarah S. Cohen, Michael T. Mumma, Derek A. Hagemeyer, Heidi Chen, Ashley P. Golden, R. Craig Yoder, and Lawrence T. Dauer. Mortality from leukemia, cancer and heart disease among u.s. nuclear power plant workers, 1957–2011. *International Journal of Radiation Biology*, 98(4):657–678, 2022. doi: 10.1080/09553002.2021.1967507. URL <https://doi.org/10.1080/09553002.2021.1967507>. PMID: 34669562.
- [27] "National Research Council". "Health Risks from Exposure to Low Lev-

- els of Ionizing Radiation: BEIR VII Phase 2*". "The National Academies Press", "Washington, DC", 2006. ISBN "978-0-309-09156-5". doi: "10.17226/11340". URL "<https://nap.nationalacademies.org/catalog/11340/health-risks-from-exposure-to-low-levels-of-ionizing-radiation>".
- [28] Richard William Harbron. Cancer risks from low dose exposure to ionising radiation – is the linear no-threshold model still relevant? *Radiography*, 18(1):28–33, 2012. ISSN 1078-8174. doi: <https://doi.org/10.1016/j.radi.2011.07.003>. URL <https://www.sciencedirect.com/science/article/pii/S1078817411000691>. Cancer: The Contribution of Imaging and Radiotherapy to Improving Outcomes.
- [29] NRC. Federal register volume 86, issue 156 (august 17, 2021), 2021.
- [30] B. Marples, B. G. Wouters, S. J. Collis, A. J. Chalmers, and M. C. Joiner. Low-Dose Hyper-radiosensitivity: A Consequence of Ineffective Cell Cycle Arrest of Radiation-Damaged G2-Phase Cells. *Radiation Research*, 161(3):247 – 255, 2004. doi: 10.1667/RR3130. URL <https://doi.org/10.1667/RR3130>.
- [31] Saskia Grudzenski, Antonia Rath, Sandro Conrad, Claudia E. Rube, and Markus Löbrich. Inducible response required for repair of low-dose radiation damage in human fibroblasts. *Proceedings of the National Academy of Sciences*, 107(32):14205–14210, 2010. doi: 10.1073/pnas.1002213107. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1002213107>.
- [32] National Academies of Sciences Engineering and Medicine, editor. *Leveraging Advances in Modern Science to Revitalize Low-Dose Radiation Research in the United States*. National Academies Press (US), 2022. ISBN 0-309-27577-6.
- [33] SAS. *SAS (r) Proprietary Software 9.4 (TS1M5)*. SAS Institute Inc., Cary, NC, USA, 2016.
- [34] John D. Boice. The million person study relevance to space exploration and mars. *International Journal of Radiation Biology*, 98(4):551–559, 2022. doi: 10.1080/09553002.

- 2019.1589020. URL <https://doi.org/10.1080/09553002.2019.1589020>. PMID: 30831048.
- [35] Eric Giunta. *Colossus*, 2024. URL <https://cran.r-project.org/web/packages/Colossus/index.html>.
- [36] Eric Giunta. Colossus. <https://github.com/ericgiunta/Colossus>, 2024.
- [37] Tjalling J. Ypma. Historical development of the newton–raphson method. *SIAM Review*, 37(4):531–551, 1995. doi: 10.1137/1037125. URL <https://doi.org/10.1137/1037125>.
- [38] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.
- [39] Gideon Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2): 461–464, July 1978.
- [40] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 12 2018. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1972.tb00899.x. URL <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>.
- [41] NE Breslow. Discussion on professor cox’s paper. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):202–220, 1 1972. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1972.tb00900.x. URL <https://doi.org/10.1111/j.2517-6161.1972.tb00900.x>.
- [42] Bradley Efron. The efficiency of cox’s likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565, 1977. doi: 10.1080/01621459.1977.10480613. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1977.10480613>.
- [43] Jason P. Fine and Robert J. Gray and. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–

- 509, 1999. doi: 10.1080/01621459.1999.10474144. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10474144>.
- [44] David Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, 04 1982. ISSN 0006-3444. doi: 10.1093/biomet/69.1.239. URL <https://doi.org/10.1093/biomet/69.1.239>.
- [45] Patricia M. Grambsch and Terry M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994. ISSN 00063444. URL <http://www.jstor.org/stable/2337123>.
- [46] TERRY M. THERNEAU, PATRICIA M. GRAMBSCH, and THOMAS R. FLEMING. Martingale-Based Residuals for Survival Models. *Biometrika*, 77(1):147–160, 03 1990. ISSN 0006-3444. doi: 10.1093/biomet/77.1.147. URL <https://doi.org/10.1093/biomet/77.1.147>.
- [47] Alberto Stabilini, Luana Hafner, and Linda Walsh. Comparison and multi-model inference of excess risks models for radiation-related solid cancer. *Radiation and Environmental Biophysics*, 62(1):17–34, January 2023. ISSN 1432-2099. doi: 10.1007/s00411-022-01013-0. URL <http://dx.doi.org/10.1007/s00411-022-01013-0>.
- [48] Luana Hafner and Linda Walsh. Application of multi-method-multi-model inference to radiation related solid cancer excess risks models for astronaut risk assessment. *Zeitschrift für Medizinische Physik*, 34(1):83–91, February 2024. ISSN 0939-3889. doi: 10.1016/j.zemedi.2023.06.003. URL <http://dx.doi.org/10.1016/j.zemedi.2023.06.003>.
- [49] John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, 1980.
- [50] Douglas Bates and Dirk Eddelbuettel. Fast and elegant numerical linear algebra using the rcppeigen package. *Journal of Statistical Software*, 52(5):1–24, 2013. doi:

- 10.18637/jss.v052.i05. URL <https://www.jstatsoft.org/index.php/jss/article/view/v052i05>.
- [51] OpenMP Architecture Review Board. Openmp application program interface version 5.2, November 2021. URL <https://www.openmp.org/wp-content/uploads/OpenMP-API-Specification-5-2.pdf>.
- [52] Suresh H. Moolgavkar and David J. Venzon. General relative risk regression models for epidemiologic studies. *American Journal of Epidemiology*, 126(5):949–961, 11 1987. ISSN 0002-9262. doi: 10.1093/oxfordjournals.aje.a114733. URL <https://doi.org/10.1093/oxfordjournals.aje.a114733>.
- [53] Terry M Therneau, Cynthia Crowson, and Elizabeth Atkinson. *Multi-state models and competing risks*, 2024. URL <https://cran.r-project.org/web/packages/survival/vignettes/compete.pdf>.
- [54] Benjamin Mappin-Kasirer, Hongchao Pan, Sarah Lewington, Jennifer Kizza, Richard Gray, Robert Clarke, and Richard Peto. Tobacco smoking and the risk of parkinson disease. *Neurology*, 94(20):e2132–e2138, 2020. doi: 10.1212/WNL.0000000000009437. URL <https://www.neurology.org/doi/abs/10.1212/WNL.0000000000009437>.
- [55] CEDR. Mortality from leukemia, cancer and heart disease among u.s. nuclear power plant workers, 1957–2011, 2023. URL https://oriseapps.ornl.gov/CEDR/search_results.aspx?DataSet=NPP23A01.
- [56] Youssra Souli, Xavier Trudel, Awa Diop, and Denis Talbot. Longitudinal plasmode algorithms to evaluate statistical methods in realistic scenarios: an illustration applied to occupational epidemiology. *BMC Medical Research Methodology*, 23, 10 2023. doi: 10.1186/s12874-023-02062-9.
- [57] Peter Bickel and Anat Sakov. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, 18, 07 2008.

- [58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [59] Steven L. Simon, F. Owen Hoffman, and Eduard Hofer. The Two-Dimensional Monte Carlo: A New Methodologic Paradigm for Dose Reconstruction for Epidemiological Studies. *Radiation Research*, 183(1):27 – 41, 2014. doi: 10.1667/RR13729.1. URL <https://doi.org/10.1667/RR13729.1>.
- [60] Deukwoo Kwon, Steven L. Simon, F. Owen Hoffman, and Ruth M. Pfeiffer. Frequentist model averaging for analysis of dose–response in epidemiologic studies with complex exposure uncertainty. *PLOS ONE*, 18(12):1–20, 12 2023. doi: 10.1371/journal.pone.0290498. URL <https://doi.org/10.1371/journal.pone.0290498>.
- [61] You Wu, F. Owen Hoffman, A. Iulian Apostoaei, Deukwoo Kwon, Brian A. Thomas, Racquel Glass, and Lydia B. Zablotska. Methods to account for uncertainties in exposure assessment in studies of environmental exposures. *Environ Health*, 18(31), 2019. doi: 10.1186/s12940-019-0468-4. URL <https://doi.org/10.1186/s12940-019-0468-4>.
- [62] Michael Bellamy, Keith Eckerman, and Lawrence Dauer. Reconstructed lung doses for the million person study cohort of 26,650 tennessee eastman corporation workers employed between 1942 and 1947. *Journal of Radiological Protection*, 43(1):013503, feb 2023. doi: 10.1088/1361-6498/acb1be. URL <https://dx.doi.org/10.1088/1361-6498/acb1be>.
- [63] Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. URL <http://jmlr.org/papers/v21/20-729.html>.
- [64] Yilong Zhang, Xiaoxia Han, and Yongzhao Shao. The roc of cox proportional hazards cure models with application in cancer studies. *Lifetime Data Analysis*, 27(2):195–215,

January 2021. ISSN 1572-9249. doi: 10.1007/s10985-021-09516-6. URL <http://dx.doi.org/10.1007/s10985-021-09516-6>.

- [65] Neal Richardson, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoş Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. *arrow: Integration to 'Apache' 'Arrow'*, 2025. URL <https://github.com/apache/arrow/>. R package version 19.0.1.
- [66] Loren P. Lipworth. Subs & space: Million person study multiple stressor study, 2025. URL <https://ncrponline.org/shop/videos/subs-space-million-person-study-multiple-stressor-study/>.