



PHD

Inclusion of prevalent cohorts to study the causal impact of Systemic Sclerosis on cancer

Barry, Eleanor

Award date:
2023

Awarding institution:
University of Bath

[Link to publication](#)

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

Copyright of this thesis rests with the author. Access is subject to the above licence, if given. If no licence is specified above, original content in this thesis is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC-ND 4.0) Licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>). Any third-party copyright material present remains the property of its respective owner(s) and is licensed under its existing terms.

Take down policy

If you consider content within Bath's Research Portal to be in breach of UK law, please contact: openaccess@bath.ac.uk with the details. Your claim will be investigated and, where appropriate, the item will be removed from public view as soon as possible.

Inclusion of prevalent cohorts to study the causal impact of Systemic Sclerosis on cancer

Eleanor Barry

Thesis submitted for doctoral degree



CDT Statistical Applied Mathematics at Bath, SAMBa

Bath University

Bath, UK

January 25, 2023

Copyright notice

Attention is drawn to the fact that copyright of this thesis/ portfolio rests with the author and copyright of any previously published materials included may rest with third parties. A copy of this thesis/ portfolio has been supplied on condition that anyone who consults it understands that they must not copy it or use material from it except as licenced, permitted by law or with the consent of the author or other copyright owners, as applicable.

Declaration of any previous submission of the work


The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree.

Author's signature: 

Eleanor VH Barry

Declaration of authorship

I am the author of this thesis, and the work described therein was carried out by myself personally.

Author's signature: 

Eleanor VH Barry

Contents

Abstract	xxix
Acknowledgements	xxxi
Abbreviations and definitions	xxxiii
1 Introduction: background, overview of thesis aims and introduction to concepts	1
1.1 Background	1
1.2 Thesis aims and objectives	3
1.3 Thesis structure and originality	6
2 Previous research into Systemic Sclerosis	11
2.1 Introduction	11
2.2 Systemic Sclerosis epidemiology	12
2.3 Cancer in SSc patients	15
2.4 SSc and mortality	19
2.5 Discussion	21
3 Dataset characteristics	23
3.1 Introduction	23

3.2	The Clinical Practice Research Datalink	23
3.3	Notable adjustments to the dataset and missing data	26
3.4	Onset, start of follow up, and incident or prevalent status	30
3.5	Possible outcomes	33
3.6	Correlation between age at diagnosis and truncation time	40
3.7	Covariates	43
4	Systemic Sclerosis and mortality	51
4.1	Introduction	51
4.2	Cohort characteristics	53
4.3	Standardized mortality ratio	57
4.4	Survival, and the Kaplan-Meier curve	62
4.5	Proportional hazards model	66
4.6	Summary	72
5	Combining incident and prevalent cohorts to study SSc mortality	73
5.1	Introduction	73
5.2	Prevalent cohort characteristics	75
5.3	Notation and left truncation definition	76
5.4	Nonparametric estimation of the survivor function	79
5.5	Proportional hazards model	89
5.6	Possible pitfalls of prevalent inclusion	93
5.7	Summary	95

6	Cancer analysis with competing risks	99
6.1	Introduction	99
6.2	Cancer standardized incidence ratios	100
6.3	Competing risk theory	104
6.4	Cumulative incidence risks	110
6.5	Cause-specific and subdistribution proportional hazards model . .	120
6.6	Summary	134
7	G-formula: Theory and application to the incident cohort	137
7.1	Introduction	137
7.2	Causal inference overview	139
7.3	G-formula methodology	143
7.4	Application of g-formula to incident cohort	154
7.5	Alternative to the g-formula: IPCW	168
7.6	Summary	174
8	G-formula: Application to a prevalent cohort	177
8.1	Introduction	177
8.2	Estimating regression coefficients in a prevalent cohort	180
8.3	Using the prevalent cohort for regression coefficients, applied to the incident pseudo-population	183
8.4	Consequence of unadjusted prevalent inclusion	192
8.5	Methodology for weighting the g-formula	196
8.6	Simulation of proportion estimation and weighted g-formula estimator	199
8.7	Estimating the baseline covariate distribution	204

8.8	Weighted g-formula estimation using the prevalent cohort	220
8.9	Weighted g-formula risk estimation using the combined dataset	223
8.10	Alternative method: IPCW with left truncated data	228
8.11	Summary	236
9	Accounting for temporal trends	239
9.1	Introduction	239
9.2	Dependent truncation and the quasi-stationarity assumption	242
9.3	Testing quasi-stationarity in the SSc dataset	249
9.4	Accounting for conditionally independent truncation using weightings	256
9.5	Application of temporal trend allowances to SSc dataset	263
9.6	Temporal trends in the g-formula	273
9.7	Conclusions	282
10	Discussion	285
10.1	Summary of work undertaken	285
10.2	Specific conclusions	286
10.3	Summary risk ratios of cancer in SSc patients depending on model used	290
10.4	Strengths and limitations	294
10.5	Original contribution	299
10.6	Possibilities for future work	300
10.7	Overall conclusion	301
11	References	303

A Appendix	315
A.1 Relating to Chapter 5: Simulation of truncation times	315
A.2 Relating to Chapter 9: Simulation of conditionally independent truncation	318
A.3 ISAC Application form	328

List of Tables

4.1	Cohort characteristics of the incident cohort. SD is standard deviation. Death here includes those who may have had cancer prior to death. Age at death is calculated from the subset of patients who died, and study entry to death (years) is the time between SSc diagnosis to death in those who died. Smoking is based on 803 SSc patients and 4690 non-SSc patients. BMI is based on 615 SSc patients and 3286 non-SSc patients.	54
4.2	Mortality rates in SSc patients stratified by age and sex, compared with the background population.	59
4.3	Mortality rates in non-SSc patients stratified by age and sex, compared with the background population.	60
4.4	SMR for SSc and non-SSc patients with 95% confidence intervals.	61
4.5	Summary table for survival after 5-year, 10-year, and 15-year for SSc and non-SSc patients. Survival is the percentage who have survived after t years from diagnosis. The risk ratio at these times is on the right of the table.	66

4.6	Cox model for the hazard ratios in the incident cohort. On the left are the results from the full model with all covariates, on the right are the results of the reduced model after covariates are removed due to lower AIC. Location has a baseline of London, BMI (kg/m ²) has baseline of a 'normal' health range, 18.5-24.9, the baseline for smoking is 'non-smoking' and the baseline for drinking is 'drinker'.	70
4.7	Chi-Squared Goodness-of-Fit test for proportional hazards assumption. The data was the reduced model which used only the incident cohort.	71
5.1	Cohort characteristics. For prevalent patients, BMI is based on 620 SSc patients and 3274 non-SSc patients. Also for prevalent patients, smoking is based on 771 SSc patients and 4506 non-SSc patients. For incident patients, smoking is based on 803 SSc patients and 4690 non-SSc patients, and for BMI is based on 615 SSc patients and 3286 non-SSc patients. Age at death is calculated from the subset of patients who died, and study entry to death (years) is the time between SSc diagnosis/match to death in those who died. Follow-up is time from study entry to death or censoring.	77
5.2	Cumulative incidence of mortality for incident, prevalent and combined datasets at time points 5-, 10- and 15- years, also the ratio between them (SSc over non-SSc). The 95% confidence intervals are provided in brackets.	83

5.3	Cox model for hazard ratios of mortality, using the combined incident and prevalent dataset. On the left are the results from the full model with all covariates, on the right are the results of the reduced model after covariates are removed due to lower AIC. BMI (kg/m ²) has a baseline of a 'normal' health range, 18.5-25, the baseline for smoking is 'non-smoking' and the baseline for drinking is 'drinker'. The grey shading shows significant values at the 95% confidence level.	91
5.4	Chi-Squared test for the proportional hazards assumption for mortality.	92
5.5	Cox regression model for reduced covariates, combined dataset. . .	92
5.6	Possible 'bias' from including prevalent cohorts for three simulations. Time is 10, 20 and 30 years after diagnosis. The grey cells are when the bias is significant at the 95% level (percentile confidence intervals).	96
6.1	Cancer SIRs in SSc patients stratified by age and sex, compared with the general population (CRUK).	102
6.2	Cancer SIRs in non-SSc patients stratified by age and sex, compared with the general population (CRUK).	103
6.3	SIR for SSc and non-SSc patients with 95% confidence intervals. Here, total is our whole dataset, which is 83% female.	104
6.4	Naive one minus Kaplan-Meier (marginal cumulative incidence) of cancer for incident and combined datasets at time points 5, 10, 15, 20 and 30 years after SSc diagnosis, also the ratio between them in bold (SSc over non-SSc). The 95% confidence intervals are given in brackets.	117

6.5	Cause-specific cumulative incidence of cancer for incident and combined datasets at time points 5, 10, 15, 20 and 30 years after SSc diagnosis, also the ratio between them in bold (SSc over non-SSc). The 95% confidence intervals are given in brackets. . . .	118
6.6	Simulated data to aid interpretation of the Fine and Gray model. The hazard ratio is shown for covariates L_1 and L_2 for outcome Y . The ratio is comparing those with the covariate to those without. .	122
6.7	Cause-specific proportional hazards model for cancer, from the incident cohort only. On the left are the results from the full model with all covariates, on the right are the results of the reduced model after covariates are removed due to minimisation of AIC. Location has a baseline of London, BMI (kg/m^2) has a baseline of a ‘normal’ health range (18.5-24.9), the baseline for smoking is ‘non-smoking’, and the baseline for drinking is ‘drinker’.	124
6.8	Scaled Schoenfeld residuals, cause-specific proportional hazards for cancer, incident dataset.	125
6.9	Cause-specific model for the hazard ratios for cancer from both the incident and prevalent data. On the left are the results from the full model with all covariates, on the right are the results of the reduced model after covariates are removed due to minimisation of AIC. Location has a baseline of London, BMI (kg/m^2) has a baseline of a ‘normal’ health range, 18.5-24.9, the baseline for smoking is ‘non-smoking’ and the baseline for drinking is ‘drinker’.	126
6.10	Scaled Schoenfeld residuals, cause-specific proportional hazards for cancer, combined (incident and prevalent) dataset.	127

6.11	Fine and Gray model for cancer in the incident cohort. On the left are the results from the full model with all covariates, on the right are the results of the reduced model after covariates are removed due to minimisation of AIC. Location has a baseline of London, BMI (kg/m ²) has a baseline of a ‘normal’ health range, 18.5-24.9, the baseline for smoking is ‘non-smoking’ and the baseline for drinking is ‘drinker’.	130
6.12	Scaled Schoenfeld residuals, subdistribution proportional hazards for cancer, incident dataset.	131
6.13	Fine and Gray model for cancer from both the incident and prevalent data. On the left are the results from the full model with all covariates, on the right are the results of the reduced model after covariates are removed due to minimisation of AIC. Location has a baseline of London, BMI (kg/m ²) has the baseline of a ‘normal’ health range, 18.5-24.9, the baseline for smoking is ‘non-smoking’ and the baseline for drinking is ‘drinker’.	132
6.14	Scaled Schoenfeld residuals, subdistribution proportional hazards for cancer, incident and prevalent dataset.	133
7.1	Example table of notation. In a realistic/clinical setting, A and Y are observable, however $Y^{a=1}$ and $Y^{a=0}$ are not.	140
7.2	Coefficient values for logistic regression on the incident cohort. The baseline comparator for sex is female, for A is non-SSc ($A=0$), and the baseline for smoking is ‘non-smoker’. The 95% confidence intervals are in brackets.	157
7.3	Cumulative incidence and risk ratio for the outcome of cancer for the marginal NPMLE (1-KM) and direct effect g-formula methods using solely the incident data. The 95% confidence intervals are given in brackets.	166

7.4	Cumulative incidence and risk ratio for the outcome of cancer for the cause-specific NPMLE and total effect g-formula methods using solely the incident data. The 95% confidence intervals are given in brackets.	167
7.5	Coefficient values for censoring logistic regression, estimated using the incident cohort only. The baseline for A is non-SSc, the baseline for sex is male and the baseline for smoking and ex-smoker is non-smoker.	171
7.6	Approximate significance of smooth terms, estimated using the incident cohort only.	171
8.1	Coefficient values for logistic regression, estimated using the prevalent cohort only. The baseline for A is non-SSc, the baseline for sex is female and the baseline for smoking and ex-smoker is non-smoker. Age at SSc diagnosis is increase in hazard per year (as opposed to decade).	185
8.2	Coefficient values for logistic regression, estimated using both the prevalent and incident cohorts. The baseline for A is non-SSc, the baseline for sex is female and the baseline for smoking and ex-smoker is non-smoker. Age at SSc diagnosis is increase in hazard per year (as opposed to decade).	186
8.3	Cumulative incidence risk ratio comparisons for the left truncation Kaplan-Meier and direct g-formula estimator methods using both the incident and prevalent data to model the GLM but then applying it to the incident pseudo-population. The NPMLE is estimated from both the incident and prevalent patients. The 95% confidence intervals are given in brackets.	190

8.4	Cumulative incidence risk ratio comparisons for the left truncation cause-specific and total g-formula effect estimator methods using both the incident and prevalent data to model the GLM but then applying it to the incident pseudo-population. The NPMLE is estimated from both the incident and prevalent patients. The 95% confidence intervals are given in brackets.	191
8.5	Example incident patients.	205
8.6	Percentage in each age group for the incident and prevalent cohorts, and then estimated proportions based on NPMLE and logistic regression methods. The prevalent cohort is defined as all SSc patients (i.e. no set window of diagnoses dates).	219
8.7	Percentage in each age group for the incident and prevalent cohorts, and then estimated proportions based on NPMLE and logistic regression methods. Reduced model with patients diagnosed after 1980, with 1502 SSc patients.	220
8.8	Cumulative incidence risk ratios for cancer, comparisons for the left truncated Kaplan-Meier and direct effect g-formula methods, using both the incident and prevalent data for both methods. The 95% confidence intervals are given in brackets.	227
8.9	Cumulative incidence risk ratios for cancer, comparisons for the Aalen-Johansen and total effect g-formula methods using both the incident and prevalent data for both methods. The 95% confidence intervals are given in brackets.	228
8.10	Coefficient values for censoring logistic regression, estimated using the incident cohort only. The baseline for A is non-SSc, the baseline for sex is male and the baseline for smoking and ex-smoker is non-smoker.	232
8.11	Approximate significance of smooth terms, estimated using the incident cohort only.	232

9.1	Cox regression model for calendar time of diagnosis per 10 years, with coefficient and hazard ratio (HR) and 95% confidence intervals, left, along with the global Schoenfeld residual test with p-values, right. Grey highlights the statistically significant hazard ratios.	251
9.2	Cox PH regression model, calendar time of diagnosis as categorical variable, with coefficients and hazard ratio (HR) with 95% confidence intervals, left, along with Schoenfeld residual test with p-values for each category, right. Grey highlights the statistically significant hazard ratios.	252
9.3	Cox proportional hazard regression model, calendar time of diagnosis as a continuous covariate, with hazard ratio and 95% confidence intervals on the left and Schoenfeld test p-values on the right, where the p-value assigned to ex-smoker is for the smoking category as a whole. The baseline for sex and smoking are female and non-smoker, respectively. Grey highlights the statistically significant hazard ratios.	254
9.4	Cox regression models with calendar time of diagnosis as covariate (left), along with Cox proportional hazards test (right).	263
9.5	Direct effect on cancer for incident and combined datasets at time points 10, 20 and 30 years, also the ratio between them (SSc over non-SSc). The 95% confidence intervals are provided in brackets. .	272
9.6	Total effect on cancer for incident and combined datasets at time points 10, 20 and 30 years, also the ratio between them (SSc over non-SSc). The 95% confidence intervals are provided in brackets. .	272

9.7	Coefficient values for logistic regression, estimated using both the prevalent and incident cohorts. The baseline for A is non-SSc, the baseline for sex is female and the baseline for smoking and ex-smoker is non-smoker. Age is the age at which SSc patients were diagnosed. The baseline for calendar time is W=0 in 18/09/2017, and each yearly increment is further back in calendar time.	276
9.8	Cumulative incidence of cancer for incident and combined datasets at time points 10-, 20- and 30-years, also the ratio between them (SSc over non-SSc). The 95% confidence intervals are provided in brackets.	281
9.9	Cumulative incidence of cancer for incident and combined datasets at time points 10-, 20- and 30-years, also the ratio between them (SSc over non-SSc). The 95% confidence intervals are provided in brackets.	281
A.1	Cox model for simulation, with calendar time of diagnosis compared to those diagnosed with SSc after 2000. The hazard ratio is the mean of the 500 samples.	319
A.2	Cox model for simulation of the hazard ratios for cancer, Y , with calendar year of SSc diagnosis compared to baseline for year 2000+ .	323
A.3	Cox model for simulation of the hazard ratios for death without cancer, D , with calendar year of SSc diagnosis compared to baseline for year 2000+	323

List of Figures

3.1	Distribution of diagnosis dates over calendar time.	33
3.2	Number in the risk set by type of patients, whether they are incident or prevalent and if they are SSc or non-SSc.	34
3.3	End events by year in the incident SSc set. The number corresponds to the number of events in total and the percentage is the proportion that had that event in the incident SSc group. . . .	35
3.4	End events by year in the incident non-SSc set. The number corresponds to the number of events in total and the percentage is the proportion that had that event in the incident non-SSc group. . . .	36
3.5	End events by year in the prevalent SSc set. The number corresponds to the number of events in total and the percentage is the proportion who had that event in the prevalent SSc group. . . .	36
3.6	End events by year in the prevalent non-SSc set. The number corresponds to the number of events in total and the percentage is the proportion who had that event in the prevalent non-SSc group. . . .	37
3.7	Lexis plot for incident (left) and prevalent (right) patients.	38
3.8	Lexis plot for incident SSc patients. The plot is the density of follow-up time subdivided by age and calendar year. The width of each square is 2 years.	39

3.9	Lexis plot for prevalent SSc patients, from diagnosis of SSc to end date. The plot is the density of follow-up time subdivided by age and calendar year. The width of each square is 2 years.	40
3.10	Truncation time versus age at SSc diagnosis. Left is non-SSc and right SSc. Colour scheme: grey is censored, red is cancer, blue is death.	41
3.11	Truncation time versus time to event. Left is non-SSc and right SSc. Colour scheme: grey is censored, red is cancer, blue is death.	41
3.12	Age at SSc diagnosis versus time to event. Left is non-SSc and right SSc. Colour scheme: grey is censored, red is cancer, blue is death.	42
3.13	BMI distribution. NA is when patient information is not recorded.	45
3.14	Smoking habit distribution. NA is when patient information is not recorded.	46
3.15	Alcohol usage distribution. NA is when patient information is not recorded.	47
3.16	Distribution of GP locations.	48
4.1	Directed Acyclic Graph, DAG, demonstrating the possible causal pathways of our study. The dashed line is due to unknown causal pathways between the unmatched covariates and our exposure.	56
4.2	Nelson-Aalen cumulative hazard (left) and Kaplan-Meier survival probability (right), from SSc diagnosis (or match) to death, in the incident set only. The shaded areas represent the 95% confidence intervals.	65
4.3	Scaled Schoenfeld residual plots for age at SSc diagnosis.	71
5.1	Example patients plotted over calendar time.	79

5.2	Example patients plotted from time since diagnosis to event, with the number in each risk set at time t.	80
5.3	Cumulative hazard rate (left) and survival probability (right) when the risk set is from SSc diagnosis or SSc match diagnosis to event (time at entry not included in the nonparametric estimate), prevalent cohort only alongside incident cohort only, with 95% confidence intervals.	81
5.4	Cumulative hazard rate (left) and survival probability (right) when the patient is included in the risk set is from study entry to event, prevalent cohort only alongside incident cohort only, with 95% confidence intervals.	82
5.5	Number at risk in each set over time, Left - Incident, Right- Prevalent.	85
5.6	Estimation of left truncation distribution.	86
6.1	Naive Kaplan-Meier estimator or marginal risk (1-KM) with 95% confidence intervals, Left: Cancer, Right: Death.	111
6.2	Cause-specific cumulative incidence function for different competing risk hazards (CIF) with 95% confidence intervals, Left: Cancer, Right: Death.	112
6.3	Naive Kaplan-Meier with 95% confidence intervals, incident and prevalent cohorts, Left: Cancer, Right: Death.	114
6.4	Cause-specific cumulative incidence function for incident and prevalent cohorts with 95% confidence intervals. Left: Cancer, Right: Death.	115
6.5	Estimation of left truncation distribution	119
6.6	Scaled Schoenfeld residual plots for age at SSc diagnosis, cause-specific hazard ratios, combined dataset.	128

6.7	Scaled Schoenfeld residual plots for age at SSc diagnosis, subdistribution hazard, incident dataset.	129
7.1	Causal diagram 1, A is exposure, Y is outcome.	142
7.2	Causal diagram 2, A is exposure, Y is outcome, Z is confounder.	142
7.3	Causal diagram 3, A is exposure, Y is outcome, Z is the collider.	143
7.4	Causal DAG, as depicted by Young et al. The dashed arrows must be absent when the direct effect is being considered for exchangeability to hold (see below), but can be present when the total effect is being considered.	146
7.5	G-formula risk estimation for risk of cancer (left) and death (right), direct effect (elimination of competing events), with the regression and prediction based on incident cohort only. The dashed lines are the incident NPMLE 1-Kaplan-Meier estimation.	163
7.6	G-formula risk estimation for risk of cancer (left) and death (right), total effect (without elimination of competing events), with the regression and prediction based on incident cohort only. The dashed lines are the incident NPMLE Aalen-Johansen estimation.	164
7.7	Plots of the smmothing functions used to model the heazard of censoring. The grey shading is the 95% confidence intervals.	172
7.8	IPCW for risk of cancer (left) and death (right), direct effect, with data from the incident cohort. The unadjusted 1-Kaplan-Meier estimations are shown as the solid lines, and the IPCW are the dashed lines.	173
7.9	IPCW for risk of cancer (left) and death (right), total effect, with data from the incident cohort. The unadjusted Aalen-Johansen estimations are shown as the solid lines, and the IPCW are the dashed lines.	174

8.1	G-formula for risk of cancer (left) and death (right), direct effect, with the logistic regression model being fitted on either solely the incident cohort (blue/orange dashed lines), the prevalent cohort (red/green dot dashed line) or both the incident and prevalent cohort (red/green solid line), but with the pseudo-population based solely on the incident SSc cohort.	187
8.2	G-formula for risk of cancer (left) and death (right), total effect, with the logistic regression model being fitted on either solely the incident cohort (blue/orange dashed lines), the prevalent cohort (red/green dot-dashed line) or both the incident and prevalent cohort (red/green solid line), but with the pseudo-population based solely on the incident SSc cohort.	188
8.3	G-formula for risk of cancer (left) and death (right), direct effect, with the hazards being predicted from the prevalent cohort, and either applied to the incident SSc pseudo-population (solid line) or the prevalent SSc pseudo-population (dashed line).	194
8.4	G-formula for risk of cancer (left) and death (right), total effect, with the hazards being predicted from the prevalent cohort, and either applied to the incident cohort (solid line) or the prevalent cohort (dashed line).	195
8.5	Simulation of biased prevalent covariate distribution and adjusted estimated baseline distribution for males. The points are the mean of 300 simulations and the bars are the 95% confidence intervals.	201
8.6	The cumulative risk curves produced by simulation. The black curve is the 1-KM estimated from an incident cohort, the red is the biased prevalent curve and the green is the g-formula bias adjusted curve.	204
8.7	Estimation of the baseline covariate proportion for females in the incident cohort depending on the hypothetical study start.	207

8.8	Estimation of the baseline covariate proportion for age at SSc diagnosis in the incident cohort depending on the hypothetical study start.	209
8.9	Estimation of the baseline covariate proportion for smoking types at SSc diagnosis in the incident cohort depending on the hypothetical study start.	210
8.10	Truncation time and probability of surviving until that truncation time for each patient, depending on sex of the patient.	213
8.11	Estimation of the baseline covariate proportion for females in the prevalent cohort depending on the hypothetical recruitment start (from when SSc patients are diagnosed), including logistic regression based survival.	214
8.12	Truncation time and probability of surviving until that truncation time for each patient, depending on age at diagnosis. The number shows the exact age of the patient at diagnosis, and the colour of the number also depicts age at SSc diagnosis with blue being younger and red being older.	215
8.13	Estimation of the baseline covariate proportion for age at SSc diagnosis in the prevalent cohort depending on the hypothetical diagnosis start, including logistic regression based survival.	216
8.14	Truncation time and probability of surviving until that truncation time for each patient, depending on smoking status.	217
8.15	Estimation of the baseline covariate proportion for smoking status in the prevalent cohort depending on the hypothetical diagnosis start, including logistic regression based survival.	218

8.16	G-formula for risk of cancer (left) and death (right), direct effect, with hazard fitted from the prevalent GLM only and applied to the a) incident SSc cohort (solid lines), b) prevalent SSc cohort (dashed red/green lines) and then a weighted prevalent (dashed orange/blue lines).	222
8.17	G-formula for risk of cancer (left) and death (right), total effect, with hazard fitted from the GLM and applied to the a) incident (solid lines), b) prevalent (dashed red/green lines) and then a weighted prevalent (dashed orange/blue lines), full dataset.	223
8.18	G-formula for risk of cancer (left) and death (right), direct effect, with the g-formula using both incident and prevalent cohorts weighted based on survival, and the nonparametric method which is also estimated from the combined dataset.	224
8.19	G-formula for risk of cancer (left) and death (right), total effect, with the g-formula effect estimator using both incident and prevalent cohorts weighted based on survival probability, and the nonparametric cause-specific cumulative curves which are also estimated from the combined dataset.	225
8.20	Plots of the smoothing functions used to model the hazard of censoring. The dark grey shading is the 95% confidence intervals.	233
8.21	IPCW for risk of cancer (left) and death (right), direct effect. The hazard is fitted from the GLM of the combined incident and prevalent cohorts and we then show the a) unweighted/unadjusted cause-specific cumulative incidence (solid red/green), b) IPCW including a weighting for loss to follow-up and competing event (dashed red/green lines) and c) IPCW where loss to follow-up is not weighted but the competing event is (dashed orange/blue).	234

8.22	IPCW for risk of cancer (left) and death (right), total effect, with hazard fitted from the GLM and applied to the a) unweighted, b) IPCW (dashed red/green lines).	235
9.1	Cumulative incidence using the independent Kaplan-Meier (solid lines) and the temporally adjusted recent cumulative incidence direct effect (dashed line). Cancer (left) and death (right).	265
9.2	Cumulative incidence using the independent Kaplan-Meier (solid lines) and the temporally adjusted historic cumulative incidence direct effect (dashed line). Cancer (left) and death (right).	266
9.3	Comparing the non-adjusted cause-specific cumulative incidence with the recent cumulative incidence curves. The non-adjusted cause-specific cumulative incidence are the solid lines, and the temporal adjusted recent cumulative incidence curves are the dashed lines. Cancer (left) and death without cancer (right). . . .	267
9.4	Comparing the non-adjusted cause-specific cumulative incidence with the historic cumulative incidence. The non-adjusted cause-specific cumulative incidence are the solid lines, and the temporal adjusted historic cumulative incidence curves are the dashed lines. Cancer (left) and death without cancer (right). . . .	269
9.5	G-formula recent direct effect on cancer (left) and death (right), with the logistic regression model being fitted on either the incident and prevalent cohort with temporal trends (solid line, red and green) or the Breslow estimated recent risk with both the incident and prevalent cohort (dot-dashed line, red and green).	278

9.6	G-formula recent total effect on cancer (left) and death (right), total effect, with the logistic regression model being fitted on either the incident and prevalent cohort with temporal trends (solid line, red and green) or the competing risk estimated recent risk with both the incident and prevalent cohort (dot-dashed line, red and green).	279
10.1	Risk ratios for the direct effect of SSc on cancer depending on methods used throughout the thesis. The times taken are at 10, 20, and 30 years after SSc diagnosis.	293
10.2	Risk ratios for the total effect of SSc on cancer depending on methods used throughout the thesis. The times taken are at 10, 20, and 30 years after SSc diagnosis.	294
A.1	Simulations of cumulative distribution functions for left truncation distribution. Left: S1, demonstrating a uniform distribution. Middle: S2, demonstrating changes in SSc incidence over calendar time. Right S3, demonstrating changes in survival probability over calendar time	317
A.2	Simulation of truncation distribution in the one outcome event setting.	320
A.3	Simulation of survival given dependent left truncation in the one event setting. The black is the simulated 1-KM curve of the incident data. The red is the same for the prevalent data using the left truncated Kaplan-Meier estimator. The green is the dependent-adjusted recent risk estimator. The blue is the dependent-adjusted historic risk. The grey is the theoretical cumulative incidence of the patients with the longest truncation.	322

A.4	Simulation of survival given dependent left truncation in the two event setting under the assumption of the elimination of the competing event (i.e naive Kaplan-Meier). Left: Cancer, Right: Death without cancer. The black is the simulated 1-KM curve of the incident data. The red is the same for the prevalent data using the left truncated Kaplan-Meier estimator. The green is the dependent-adjusted recent risk estimator. The blue is the dependent-adjusted historic risk. The grey is the theoretical cumulative incidence of the patients with the longer truncation. . .	325
A.5	Simulation of truncation distribution. The black line is the independent distribution assuming left truncation distribution (as shown in Chapter 5), and the green is the dependent-adjusted left truncation distribution.	326
A.6	Simulation of survival given dependent truncation	327

Abstract

Systemic Sclerosis (SSc) is a rare autoimmune disease. In this thesis we investigate the risk of cancer (the outcome of interest) in patients diagnosed with SSc (the exposure of interest) and compare this against the risk of cancer in people without SSc. A large UK primary care dataset provides the data for this study. This dataset contains 806 patients who were diagnosed with SSc over the chosen study period 1998 to 2018, forming the incident cohort group, termed ‘SSc’. An additional 780 patients diagnosed with SSc prior to their entry to the UK database (which may be later than 1998) and cancer free at this time, form the prevalent cohort, potentially a valuable additional resource for analysis as its addition greatly increases sample size and length of follow-up. A pre-requisite for inclusion of this prevalent cohort is that it is consistent with the incident cohort and does not distort the study population. The thesis examines how to account for prevalent patients in a competing risk framework, including in the presence of informative censoring, and the issues that could be encountered by left truncated data with a long follow-up.

Each patient diagnosed with SSc, in both incident and prevalent cohorts, is matched to 6 other patients who will be used as comparators. These matches form a group termed ‘non-SSc’ and this group enables the risk of cancer in patients without an SSc diagnosis to be assessed. For both SSc and non-SSc patients there is a competing risk, death, and the thesis considers methods of analysis in a competing risk setting. As common nonparametric methods of analysis are often limited by confounding, we investigate the parametric g-formula to permit

causal interpretation, as opposed to an association, between SSc and cancer. The expansion of the parametric g-formula to include prevalent patients, including the derivation of a weighted form, is formulated. Methods to adjust for differences observed due to date of diagnosis of SSc are developed, with particular relevance to the potential adjustment of patients in the prevalent cohort to provide a better estimation of current risk.

The thesis concludes that prevalent inclusion is often beneficial, following pre-analysis of differences between the incident and prevalent cohorts and use of the Cox model to test for temporal trends. Including prevalent patients in this analysis without temporal trends suggests that there is an increase in risk of cancer in those with SSc when compared to those without (approximate 1.25 times the risk), however once temporal trends are accounted for there is no longer a statistically significant difference in the risk of cancer between SSc patients and non-SSc patients. Due to the increased mortality of SSc patients, the cause-specific risk ratio between SSc and non-SSc patients is not significant at the 5% level and decreases over calendar time. The parametric g-formula method allowing for prevalent patients was used for this study but no change in outcome between this method and the prior nonparametric methods was observed, possibly due to the matching having already accounted for the majority of confounding. We recommend the parametric g-formula when there are differentiating covariate distributions between the exposed and unexposed.

The results of this study will support those who are interested in the epidemiology of SSc, those who are considering inclusion of prevalent patients and how best to do this, and lastly those who may be interested in using the g-formula with left truncated data.

Acknowledgements

Firstly I would like to thank my supervisors Anita McGrogan and Jonathan Bartlett. My skills have improved so much, thanks to your patient guidance and expertise. I have thoroughly enjoyed researching this area and finding these results. I also appreciate the input of Julia Snowball for the organisation of the dataset.

To the SAMBa CDT who made the PhD a varied and fulfilling experience in what can be a lonely and solitary job. To Susie, Jess, Helena, Lindsey, and the whole executive team who worked so hard to support all students during the difficult last few years. To Karsten Matthies with his efficiency with sorting many issues in the last year. To Matt Nunes for patiently double checking my work and making much of my garbling coherent. Also, to Theresa Smith and Darren Ashcroft for taking the time to read my thesis and making the viva process interesting the thought provoking. Thank you, all.

To all the authors whose names are mentioned in this thesis for their excellent work and inspiration. Without this earlier work my thesis could not have existed.

To the wonderful friends and colleagues who have enriched the last few years. The superb comradery of my fellow maths PhD students, which I doubt can be rivalled at any other university. Thank you for putting up with my cake emails (and me) for years. To the laughter provided by my game friends: my Wednesday D&D group and the Bookem lot. To Chaps and Charlotte for the wine and walks. And with special thanks to George S, James A and Will G, who supported me and guided me far more than any of them will ever take credit for.

And finally to my wonderful parents, who have led by example, and whose love, compassion and support have made this even remotely possible.

Abbreviations and definitions

Association A statistical relationship between two variables, where an outcome is more/less likely in those with an exposure. This does not necessarily mean the exposure is the reason for the outcome, as there may be confounding factors present.

Causation A statistical relationship between two variables, where an exposure produces an effect.

Censoring When the event of interest is not measured and therefore a censored patient provides incomplete data. We have right censoring in this study, which can either occur at the end of the study period (i.e., administrative censoring) or when a person fails to return for a study visit (loss to follow-up).

Confounding When a factor creates a distorted association between an exposure and an outcome. See Section 7.2 for more detail.

Left truncation A patient may be included if they have experienced the initiating event but have not experienced the outcome of interest, therefore patients enter the study at different times of disease progression. However someone who has the outcome of interest prior to study entry is not included. The data arising here is *left truncated*, or patients with *delayed entry*.

Notation	Interpretation
A	Indicator of event ‘assignment’, (A=1 for SSc, A=0 for non-SSc)
ATE	Average treatment effect
ATT	Average treatment effect in the treated
C	Censoring time
CI	Confidence interval
CIF	Cumulative incidence function
CIR	Cumulative incidence ratio
D	Event occurrence of a competing risk, used in causal models
Δ	Indicator of censoring status, 1 is an event was observed, 0 if censoring occurred
E	Competing risk set
F	Distribution of survival times
G	Distribution of left truncation
H	Distribution of censoring
HR	Hazard ratio
i	Index for patient number
IPW	Inverse probability weighting
IPCW	Inverse probability censored weighting
j	Index for competing risks
K	Time unit when considering competing risks in time steps or set of competing risks
KM	Kaplan-Meier
L	Left truncation time
NPMLE	Nonparametric maximum likelihood estimator (often used for 1-KM or Aalen-Johansen)
S	Probability of survival
SSc	Systemic Sclerosis
T	Time of event
W	Calendar time of diagnosis
X	Time of study exit ($X = \min(T, C)$)
Y	Event occurrence of the event of interest in competing risks
Z	Covariates

Chapter 1

Introduction: background, overview of thesis aims and introduction to concepts

1.1 Background

Epidemiology is the study of the distribution and determinants of disease frequency in human populations and the use of this information to aid public health. For example:

- Qualifying and quantifying associations between risk factors and outcomes, with the aim of minimizing those factors found to have a significant negative effects on health, to improve public health
- Identifying emerging health problems
- Establishing public health priorities for a population
- Evaluating the effectiveness of intervention programs

Analysis of an entire population is implausible, but useful observations can be made on a study sample, which is selected in some way from the target

population. The confidence that can be placed in conclusions drawn from such samples depends in part on sample size. Small samples can be unrepresentative by chance, and the scope for such chance errors can be quantified statistically. More problematic are biases incurred from the way the data has been collected, such as bias due to censoring or differing covariates (either between two comparison groups or the dataset being a poor representation of the true population).

This thesis focuses on an example of the first aim of epidemiology in the list above, namely the identification of an association between an exposure and an outcome, to aid healthcare advice by reducing¹ the relevant exposure, or mitigating its effects. We utilise the approach of comparing exposed and unexposed groups with the aim of identifying an attributable risk. The focus of our study is the risk of cancer (the outcome of interest) in patients diagnosed with Systemic Sclerosis (the exposure of interest), in comparison with the risk of cancer in those without an SSc diagnosis.

Systemic Sclerosis, subsequently denoted by SSc, is classified as a subset of Scleroderma. Scleroderma is an autoimmune disease which most notably affects the skin, but which can also affect internal organs of the body: SSc has the capacity to affect the heart, oesophagus, blood vessels, kidneys, lungs, and digestive system. Risk factors associated with SSc are reported in literature although there are differences in these risk factors between studies. However, it is agreed with consistency that the female sex, genetics, and exposure to silica are all risk factors for the development of SSc. Mortality is greatly increased in those diagnosed with SSc by a factor in the range of approximately 2-6, depending on study size, time period of study, and location. The development of SSc in an individual is believed to lead to a predisposition to other, potentially life-threatening, conditions. Cancer is hypothesised to be one of these. Small studies into the potential association between SSc and cancer have been undertaken (see Section 2.2), but have suffered from various deficiencies, for example small sample sizes, insufficient follow-up period, or an insufficiently

¹Conversely, if the outcome is positive, the aim may be to increase the exposure.

defined temporal relationship (as in, for example, whether the cancer was defined prior to or after SSc diagnosis). In addition, these studies were rarely done in the UK. This study analyses a large dataset of UK patients, and addresses the deficiencies present in earlier studies.

1.2 Thesis aims and objectives

The aims and objectives of this thesis can be split into the epidemiological and the statistical. The epidemiological aim is to answer the question ‘Does SSc increase the risk of cancer?’. The statistical research aim addresses the question of ‘Which methodologies can be used and how well do these methodologies answer the epidemiological question?’. Therefore, this thesis describes the work undertaken to address these aims and how more complex ideas were developed as further considerations arose.

1.2.1 Epidemiological aim: ‘Does SSc increase the risk of cancer?’

This aim is addressed through the analysis of data from a large, matched², dataset available from the Clinical Practice Research Datalink (CPRD). CPRD is a large, UK-wide research service which routinely collects data. As of the publication of this thesis, there are two possible databases that our dataset could come from: CPRD GOLD or CPRD Aurum. We specifically use CPRD GOLD (see Section 3.2 for more detail). From this selected dataset, 780 patients were diagnosed with SSc between 1998 to 2018, forming the incident cohort. An additional 806 patients were diagnosed with SSc prior to when the patients entered the CPRD, forming the prevalent cohort, potentially a valuable additional resource for analysis as it greatly increases sample size and length of follow-up. Each patient diagnosed with SSc, in both incident and prevalent cohorts, is matched to 6 other patients in the database. These matches form a group termed the non-SSc patients and enable the risk of cancer in patients without an SSc diagnosis to be assessed. However,

²See Section 3.3.1 for details on matching.

these prevalent patients have an additional requirement of being alive and cancer free at study entry.

The purpose of this aim - ‘Does SSc increase the risk of cancer?’ - directly relates to improving public health and the best allocation of resources. If it can be demonstrated that SSc has a causal impact on cancer, increased awareness of this among medical practitioners permits the provision of increased preventative measures such as screening SSc patients earlier to mitigate the risk and increase the survival of those with SSc, who already have an increased mortality. Alternatively, if an increased risk is not found, then no increased allocation of resources to aid those with SSc is required.

1.2.2 Statistical aim: ‘Which methodologies can be used and how well do these methodologies answer the epidemiological question?’

The study began with a review of available methodologies and their applicability to our analysis. From here, three key areas were identified that were necessary for this study.

Objective 1: Prevalent cohorts To assess the impact and possible value of including a prevalent cohort in an epidemiological study. Prevalent patients are those who were diagnosed with SSc prior to entry to the CPRD dataset, but are event free³. The inclusion of our prevalent cohort of patients more than doubles the number of patients which can be used for the study and will also provide longer durations of follow-up. This is important because a disease with a long duration of time between diagnosis and cancer (or death) such as SSc means that a study with a short follow-up time will not fully observe patients in later years. However, there is a bias incurred in the inclusion of prevalent patients if standard methods are used without correction, as patients with smaller survival times are less likely to be

³See Section 5.3 for an example of prevalent patients and what we mean by event-free.

included, leading to an oversampling of those with longer survival times. Therefore, the thesis discusses methods that are most often used to account for this bias. We then discuss other statistical methods concerning the inclusion of prevalent patients, including adjusting a causal model. Lastly, the inclusion of prevalent cohorts is justifiable if their disease progression is comparable to current disease progression. However, this may not be a valid assumption, for example an improvement in healthcare over time may result in patients from earlier times receiving less effective treatment than more recent patients. Inclusion of these patients would not represent the current status of risk as well as the solely incident cohort, leading to bias. We term these temporal trends, and we review methods to account for these changes over time.

Objective 2: Competing risks To form estimates of risk within a competing framework to avoid treating the competing event, death, as non-informative censoring, potentially biasing results. SSc is a life-limiting disease, and patients with SSc have a higher mortality than those without. Therefore death is a ‘competing event’, i.e., it is an event which may hinder or prevent the event of interest from being observed. The results from our data must be interpreted with this in mind. The established framework for competing risks is discussed. Consideration of the issue of competing risks is important throughout the interpretation of results from this study.

Objective 3: Causal models To assess commonly used models and their possible limitations, and place current work in a causal framework. A frequent purpose of epidemiological investigations is to identify an association between an exposure and some effect or outcome, in the hopes that this may aid healthcare advice by reducing (or increasing, if the outcome is positive) the relevant exposure, or mitigating its effects. Studies therefore often compare exposed and unexposed groups in attempting to identify an association. However, the ideal aim is to find a causal relationship, not just an association, as associations may be due to spurious

relationships between covariates. If an exposure has an associative link with an outcome, but not a causal one, then healthcare intervention may be directed incorrectly, leading to an ineffective and potentially costly misallocation of resources. Therefore, adjusting for confounding is vital. Also, performing studies or gathering data is not simple and insufficient timescales or poor information may lead to bias. Therefore, we investigate an additional method, the g-formula, to potentially provide a better interpretation of causal effect. We research the use of causal models, which are growing in popularity, as a possible way to improve current prevalent cohort methods, and as part of this we have developed a model for prevalent inclusion. Consideration is given to whether such models provide an improvement to established models and what the drawbacks associated with them are.

This thesis describes the work undertaken to address these aims and objectives, how more complex ideas were developed as further considerations arose, and the results obtained from the study.

1.3 Thesis structure and originality

This thesis is structured as follows:

- In Chapter 2 we review some of the key findings from relevant epidemiological papers about SSc, to compare against our own findings and highlight the importance of our large UK based study.
- In Chapter 3 we introduce the CPRD dataset and its key characteristics.
- In Chapter 4 we review commonly known methodologies found in survival analysis (Kaplan-Meier curve, Cox proportional hazards model) and apply them to our dataset. This is to lay the foundations for future chapters and act as a comparison for later analysis. Note that these applications are

performed only on mortality in SSc, in order to temporarily avoid the need to introduce competing risk methodology at this stage.

- In Chapter 5 we cover prevalent cohort methodology (Lynden-Bell left truncation curve, left truncation adjusted Cox model) with application to our dataset. These are known methodologies found in the literature. Again, note that this is done from the perspective of mortality in those with SSc.
- Chapter 6 covers the framework for a competing risk setting (cause-specific cumulative incidence, Fine-Gray model), along with prevalent cohort inclusion. Competing risks are a known, if underused, area. From this chapter onward, we shall be investigating the impact of SSc on cancer.
- Chapter 7 introduces the theory of causal inference, and considers the use of the parametric g-formula and IPCW (Inverse Probability Censored Weighting). The use of these is undertaken under a competing risk framework. This chapter then applies this theory to the incident cohort.
- Chapter 8 continues the theory of Chapter 7, but discusses possible modification to allow for prevalent inclusion. This includes weightings to estimate the covariates we would expect in the incident cohort. We also discuss differences in the incident and prevalent cohorts due to covariate distributions.
- Chapter 9 considers the problem of possible temporal trends when a dataset with such a long follow-up time is used. Using methodology from the limited literature available, we discuss how to adjust nonparametric curves for possible changes over calendar time. We discuss how changes could be made to the g-formula to allow for temporal trends.
- Chapter 10 provides a summary of results and key findings. The impacts of this thesis are discussed, and suggestions for future work building on these results are made. The originality of the thesis is summarised.

1.3.1 Original contributions in this thesis

The original contributions can be split into the epidemiological and the statistical:

1. Epidemiological novelty - As we shall see in Chapter 2, there have only been a few publications on the health impact of SSc in the UK in general, and only a couple of these mention cancer in passing. Even on an international level, cancer is infrequently studied as death is the most studied outcome of SSc. Research into the linkage between SSc and cancer is often performed on small secondary care datasets, where comparisons between SSc patients and non-SSc patients must be estimated using national risk data. We have the advantage of a robust, large dataset, with corresponding matched non-SSc patients to act as comparators, which allows for a better comparison than a national dataset with grouped age categories and the survival rates only (usually) taken over one calendar year. From Chapter 4 onward, we are performing new calculations on a dataset not previously utilised for this purpose. Chapter 4 and Chapter 5 cover known methods, but are new in relation to this dataset, and the application of a large prevalent dataset with the comparison to an incident dataset is a good example of the possibility and benefits of including prevalent patients. In Chapter 6, we perform analysis in the context of competing risks, which (to our knowledge) has not been undertaken previously with an SSc dataset. From Chapter 7 onward we continue to use methods which have not been applied to SSc before. We investigate temporal trends, which gives insight into covariate changes and risk changes (of cancer and death) over calendar time, which is of interest when studying SSc in general.
2. Statistical methodology novelty - Chapter 4 is introducing and applying very common methodology. Chapters 5 and 6 use fewer common methods, but nevertheless these methods are available in research papers, hence these chapters are laying foundation knowledge for future chapters while in themselves providing limited new material. Chapter 7 is based on

methodology previously published in a single article, but has not been applied outside this one work which was clinical rather than epidemiological. Chapter 8 then adjusts this methodology to account for prevalent cohorts. This addition of prevalent cohorts will allow for more patients and hence a larger sample size, improving accuracy. Including prevalent cohorts in causality models is an original approach not previously applied (to our best knowledge), and hence fills a gap in SSc literature. Chapter 9 considers temporal trends based on rarely used methods from a select few papers, and these papers have used a method to adjust for 'left truncation' and not for calendar time as undertaken in this study. This study of calendar time will hopefully improve confidence in the use of prevalent cohorts.

Chapter 2

Previous research into Systemic Sclerosis

2.1 Introduction

In this chapter, we give an overview of a) SSc as a disease, b) previous research into the relationship between SSc and cancer and c) the already well researched increase in mortality for SSc patients. In providing this overview we:

- Summarise key characteristics of SSc patients which may be of importance to the study of cancer, such as average age of diagnosis or environmental exposures.
- Review the literature on cancer research to demonstrate the lack of a clear study of the link between SSc and cancer thus far, but to also highlight possible areas to focus on.
- Review which statistical models has been used so far in the study of cancer, to demonstrate possible areas of improvement.
- Review data on cancer risk to establish a basis against which the results of this study can be compared.
- Highlight the increased risk of death in SSc patients to demonstrate the necessity of a competing risk framework in future chapters.

2.2 Systemic Sclerosis epidemiology

Systemic Sclerosis (SSc) is classified as a subset of scleroderma, which is an autoimmune disease of the skin, but in SSc the internal organs are affected as well as the skin. SSc has the capacity to affect the heart, oesophagus, blood vessels, kidneys, lungs and digestive system. Risk factors of SSc are reported in the literature but there are discrepancies in these risk factors depending on the study. However, it is agreed with consistency that the female sex, previous family history and exposure to silica are all risk factors for SSc (A. Gabrielli et al. (2009), Abbot et al. (2018)). The risk of SSc peaks between the ages of 45 and 64 years, but incidence in all adult age groups have been reported (Abbot et al., 2018).

Systemic sclerosis (SSc) incidence and prevalence rates vary considerably between countries and are also variable over time. Reasons for this variability include the following:

- Different standards of healthcare between countries, resulting in reduced accuracy and speed in diagnosis. This may in part explain why there are such different prevalence rates between countries.
- Changing standards of diagnosis over time, internationally. The criteria for diagnosing systemic sclerosis has changed over time from country to country depending on legislation and improved guidelines, resulting in inaccurate or variable classification of the disease. Prior to 1980, there was no set diagnosis criteria for SSc between countries, and the disease was less well understood, and so was often under-diagnosed. In 1980, the criteria for the classification of systemic sclerosis were developed by the American College of Rheumatology (ACR) (Masi et al., 1980). However these 1980 criteria were based on study of patients with longstanding SSc and were therefore believed to miss both early onset SSc and a small subset of patients presenting with unusual symptoms. There was a reclassification of criteria for the diagnosis of SSc in 2013, resulting from a joint committee of the ACR and the European League Against Rheumatism (EULAR) (Van Den Hoogen et al.,

2013), and as a result a new set of patients were included in those diagnosed with SSc who would not have been diagnosed under the 1980 classification. This would have included, for example, patients with less severe symptoms and a consequence of that would be differing survival times. This highlights the difficulties associated with the diagnosis of SSc and variation over time.

- Improved treatment of other diseases. With longer survival times and improved survival rates in other fatal diseases, we are more likely to see SSc present itself in patients who would in previous times not have survived long enough to display SSc.
- Improved healthcare and lifestyle. It is possible that healthier lifestyles over time lead to a decrease in the occurrence of SSc, such as decreasing exposure to environmental risk factors. For example, it is known that exposure to silica is a risk factor for SSc and exposure limits for silica now exist.

Considering SSc prevalence internationally, reported prevalence of SSc has varied in the US over the recent decades, rising from 138/million (diagnosed between 1950-79) to 276/million (diagnosed between 1989-1991) (Mayes et al., 2003). A more recent study performed in the Tuscany region of Italy looked at SSc in great detail (Coi et al., 2021). Based on 924 cases of SSc diagnosed during the 2003–2017 period collated in the Rare Disease Registry of Tuscany, prevalence is estimated at 222/million (95% CI: 207–238/million).

Prior to a few recent studies, there was believed to be a lower prevalence of SSc in the UK compared to other countries. This was largely based on a study of prevalence in the northeast of the UK that used hospital admission data, local membership records from major UK scleroderma charities, data from local primary care centres and the responses to an approach to all relevant clinicians (Allcock et al., 2004). This study reported all cases of SSc seen over a 12-month period, and yielding an estimated prevalence of SSc of 88/million (95% CI: 68–108/million) in northeast of England, much lower than the US studies and the above Italian study. After adjustment for the whole of the UK, this reduced

to 82/million (95% CI: 62–98/million).

Two recent studies have been performed in the UK. Both use the CPRD GOLD dataset, therefore we note a crossover of patients between these studies and ours. The first reported a prevalence of 307/million (95% CI: 290-323/million) based on a dataset of 1,327 patients recorded between 1994-2013 (Royle et al., 2018). An important note for this study, however, is that many of these cases were not validated, hence the large prevalence. The other study reported a prevalence of 235.5/million (95% CI: 207.2-245.7/million) based on 1,757 patients between 1999-2017 (Pauling et al., 2021). They estimated that the incidence between 1999 and 2017 was slightly decreasing but the prevalence was increasing, hypothesising this is due to the decreased mortality in those with SSc as well as the general population. In particular, this study was performed on the same cohort of patients as this thesis, with a few extra patients. Both these studies have a much higher prevalence than Allcock et al’s study.

Lastly, there has been a recent UK cohort study (Rodríguez et al., 2019) using a different primary care database, The Health Improvement Network (THIN). In this 2019 article there is a particular emphasis on how the data was validated, with unconfirmed cases being discarded. Using a study period between 2000-2012, 1,321 cases of SSc (689 incident and 632 prevalent) had a mean follow-up of 7.6 years. Mean age at diagnosis was 59.1 years. Notably, the prevalence of SSc increased from 171.3/million (95% CI, 149.7-195.1/million) to 253.8/million (95% CI, 236.8-271.6/million) over the study period but incidence did not increase. As prevalence is due to both incidence and length of duration from exposure to outcome, the authors therefore believe that the length of duration from diagnosis to outcome has increased. The authors say that there was no clear mortality trend established to explain this. The authors mention that the reclassification of SSc diagnosis may be a reason, leading to SSc patients being diagnosed earlier and so increasing length of time under study, however as incidence has not changed they rule this out. They conclude that it may be due to “dynamic nature of population cohorts”, such as changes in covariates or the time for which SSc patients are under

observation.

Women are at considerably higher risk of developing SSc than men are, with a female/male ratio having been reported to be in the range of 3:1 to 14:1 depending on the study (A. Gabrielli et al., 2009). Royle et al. (2018) reported an adjusted incidence rate ratio of 4.7 (95% CI: 4.1-5.4). The Tuscany study reported a ratio of female-to-male of 6.5 (Coi et al., 2021), noted that males have a significantly worse survival, and also found an average onset age of 59.4 ± 14.6 .

It is suggested that genetics play an additional role, with evidence of linkage between SSc and familial clustering and also with differences in phenotypes among racial and ethnic groups (Englert et al., 1999). Also, environmental challenges (e.g., viruses, drugs, vinyl chloride, and silica) may be further factors in the development of SSc (Nietert & Silver, 2000). However, there does not seem to be a link between smoking and SSc, or alcohol usage and SSc (Abbot et al., 2018).

A cross-sectional study undertaken at the University of Bath in 2013 (Strickland et al., 2013) involved 204 SSc patients and reported some key statistics. Among this group, the mean age at diagnosis of SSc was 51.6 years with a mean duration of follow-up of 12.5 years. 53 patients (of the 204) died during the study with a mean age of 72.0 years at death. The mean disease duration at death was 14.2 ± 8.5 years. The most common cause of SSc-related mortality was pulmonary complications. From the previously mentioned THIN study (Rodríguez et al., 2019), the incidence of SSc was highest among individuals aged 60-79 for both males (1.36 cases per 100,000 person-years, 95% CI: 1.01-1.81) and females (7.06 cases per 100,000 person-years, 95% CI: 6.25-7.95).

2.3 Cancer in SSc patients

The development of SSc in an individual is believed to lead to a predisposition to other, potentially life-threatening, conditions. One of these is hypothesized to be cancer. Small studies into the potential association between SSc and cancer have been undertaken, but have suffered from various difficulties, for example too

small sample sizes, short follow-up times, or an insufficiently defined temporal relationship (as in, for example, whether the cancer was defined prior to or after SSc diagnosis). In addition, these studies were seldom done in the UK.

A comprehensive article (Weeding et al., 2020) has discussed the possible link between SSc and cancer. Weeding et al. review recent articles and consider potential mechanisms which may lead to a relationship between SSc and cancer¹. On the basis of their review of 14 sources, Weeding et al. suggest that the risk of cancer in those with SSc ranges from 1.5 to 4 times higher than the risk of the general population. They cite a particularly large meta-analysis of 16 observational studies which suggests a relative risk ratio of 1.75 (95% CI: 1.41, 2.18), with notable increased relative risk ratios for lung cancer (RR 4.35; 95% CI: 2.08, 9.09) and haematological neoplasms (RR 2.24; 95% CI: 1.53, 3.29)(Bonifazi et al., 2013). Weeding et al. proffer several possible explanations for the increased risk:

- Firstly, a link between cancer and SSc may arise either because SSc predisposes the development of cancer or because, in the other direction, cancer may lead to an increased risk of SSc. Cancer treatments such as chemotherapy, radiation therapy, and immunotherapy may lead to a biological response that increases SSc risk. Weeding et al. hypothesise that SSc is specifically a paraneoplastic disease², and that SSc may be an abnormal immune system response to a cancerous tumour. This direction of risk is the opposite of the effect we are studying (i.e. whether an SSc diagnosis predisposes a patient to an increased risk of cancer), and our data deliberately does not contain patients who are diagnosed SSc after having been diagnosed with cancer. It is possible there may be some patients in our dataset who had undetected cancer before they are diagnosed with SSc and then are either diagnosed with cancer or possibly not even diagnosed

¹We note that the relationship may be that SSc patients have an increased risk of cancer than the general population, or it may be that cancer patients are more likely to get SSc than the general population

²An effect arising from cancer in the body other than effects directly due to the physical presence of cancerous tissue in the area affected.

with cancer at all. This may give the wrong impression of the cause and effect relationship between SSc and cancer. The impact would hopefully be minimal, as there would either be a negligible number of patients or only a small difference in time between SSc diagnosis and cancer diagnosis.

- Another possible explanation may be that biological damage from SSc disease may lead to a malignant transformation within the same target tissue. This is a possible explanation for the increased risk of lung cancer.
- Weeding et al. speculate that SSc treatment, and in particular some cytotoxic therapies, may contribute to the development of cancer but they consider that there is not enough information about immunosuppressive drugs in general or cytotoxic therapies in particular to clarify this risk.
- Patients with SSc may have received exposure to ionizing radiation from medical diagnostic procedures.

Finally, although not hypothesized in Weeding et al., we do wonder if more, possibly slow-acting or slow-developing, cancers are being detected in SSc patients due to more effective and available health screening due to their disease status. This would be more likely to be seen in recent years.

The type, quality and size of studies investigating cancer incidence in those diagnosed with SSc varies greatly, and hence not surprisingly a wide range of results have been reported. Most studies, international as well as UK, have demonstrated an increased risk of cancer compared to the general population. A few have not demonstrated a significant difference, for example Chatterjee et al. found a standardized incidence ratio, SIR, of 0.91 (95% CI: 0.66-1.22) based on 538 SSc patients in the US between 1973-2004 (Chatterjee et al., 2005). Also, Thomas et al. estimated an SIR of 1.15 (95% CI: 0.81-1.60) based on 652 SSc patients in Scotland between 1981-1996 (E. Thomas et al., 2000).

However the majority of studies have demonstrated a statistically significant increased risk. One US study of 769 SSc patients estimated the SIR for all

cancers diagnosed after a diagnosis of SSc at 1.55 (95% CI: 1.16-1.93), in particular oesophageal cancer at 15.9 (95% CI: 4.2-27.6) and oropharyngeal cancer at 9.6 (95% CI: 3.0-16.3) (Derk et al., 2006).

A Danish study of 2,040 SSc patients between 1977-2006 similarly found an overall SIR for cancer of 1.5 (95% CI: 1.3-1.7), with an increased risk for men at 2.2 (95% CI: 1.7-2.8) compared to 1.3 for women (95% CI: 1.1-1.6) (Olesen et al., 2010). There was an increase in the SIR for lung cancer within the first 12 months of diagnosis of SSc with an SIR of 3.3 (95% CI: 1.9-5.3), however this reduced to 1.6 (95% CI: 1.2-2.0) when only considering cancer after the first 12 months³. Considering only patients with a cancer diagnosis at least 12 months later than their SSc diagnosis, haematological cancer had an SIR of 2.5 (95% CI: 1.5-4.0) and immune related cancers an SIR of 1.4 (95% CI: 1.0-1.9).

A south Australian cohort study of cancer diagnosis after SSc diagnosis, with 441 SSc diagnoses and 70 cancer diagnoses between 1993-2000, reports an overall SIR of 2.0 (95% CI: 1.5-2.7), with a noticeable increase for lung cancer with SIR of 5.9 (95% CI: 3.1-10.3) (Hill et al., 2003). A more recent Australian study of 1,727 patients between 2008-2015 estimated an overall SIR of 2.1 (95% CI: 1.8-2.5), where the most common cancers were breast, melanoma, hematologic, and lung (Morrisroe et al., 2020).

A Swedish study of cancer diagnosis after systemic sclerosis diagnosis, with 917 patients diagnosed between 1965-1983, shows an overall SIR of 1.5 (95% CI: 1.2-1.9), in particular there were high SIRs for lung cancer (SIR=4.9; 95% CI: 2.8-8.1), non-melanoma skin cancers (SIR=4.2; 95% CI: 1.4-9.8), and primary liver cancer (SIR=3.3; 95% CI: 1.1-7.6) (Rosenthal et al., 1995).

Due to the established relationship between SSc and the lungs, a study was performed looking solely at the link between SSc and lung cancer (Peng et al., 2020). Based on a meta-analysis involving 12,218 from 10 studies (many of which are included above), an odds ratio (OR) of 2.80 was estimated (95% CI: 1.55-5.03),

³Cases were also looked at only after 12 months from diagnosis in order to not accidentally include patients who were diagnosed with SSc after having been diagnosed with cancer.

showing a significantly increased risk of lung cancer among patients with SSc. This was observed particularly in males (OR 4.11, 95% CI: 1.92-8.79) compared with female patients (OR 2.73, 95% CI: 1.41-5.27).

2.4 SSc and mortality

SSc carries a high mortality risk compared to the average population; there are many studies indicating this from a range of countries, over a wide range of time periods. These include studies from primary and secondary care datasets.

A recent 2020 University of Bath study of 1,757 patients found a standardised mortality ratio (SMR) of 3.51 (95% CI: 3.19-3.84) (Pauling et al., 2021). An earlier paper with a smaller number of patients looked at a secondary dataset of studies with 204 SSc patients, with a mean duration of follow-up of 12.5 years (Strickland et al., 2013). This study compared the survival of these patients with that from general statistics in England and Wales (from the Office of National Statistics) and found an SMR for the whole cohort of 1.34 (95 % CI: 1.00-1.75), where the SMR was higher in males, 1.54 (95% CI: 0.67-3.04) compared to females, 1.30 (95% CI: 0.95-1.74). In the 53 who died, the leading cause was infection (n=13), followed by respiratory (n=11), malignancy (n=10), and cardiovascular reasons (n=8). It should be noted that this is a small study with only 204 patients considered without direct comparators. The THIN study (Rodríguez et al., 2019) placed the mortality rate ratio (when adjusted for age and sex) as 2.82 (95% CI, 2.55-3.13) among those diagnosed with SSc compared to the rest of the population.

A few meta-analyses of small studies have been done. One meta-analysis utilised 9 studies spanning a 40 year time period and reported a pooled SMR of 3.53 (95% CI: 3.03-4.11) (Elhai et al., 2012). However, interestingly, the adjusted meta-regression analysis did not show significant change in SMR over time ($p = 0.523$). Another meta-analysis looked at 7 studies from the United States, Europe, and Japan with a total of 1,645 incident cases, and found a range of SMRs between 1.5 and 7.2 (Ioannidis et al., 2005). This large range of values

resulted in the authors being reluctant to conclude a value for a pooled SMR. A meta-analysis of 17 worldwide studies with a total of 9,239 patients spanning the years 1964 to 2005 estimated an SMR of 2.72 (95% CI: 1.93-3.83), noting a small improvement in the SMR in recent years (Rubio-Rivas et al., 2014). This study split death between those diagnosed with SSc before and after 1990, noting that causes of death differ significantly between the two, due to an increase in lung involvement and a decrease in renal involvement. Deaths involving the lung increased from 37.5% pre-1990 to 56.3% after 1990, deaths involving the heart increased from 24.0% pre-1990 to 26.9% after 1990, and deaths involving the kidney decreased from 26% pre-1990 to 12.3% after 1990. Note that the authors define these ‘involvements’ as SSc-related deaths. While they do not go into detail, they mention cancer deaths as a category of ‘non-SSc’ related deaths, and the authors attribute these to 12.2% of non-SSc mortality causes.

A study in Denmark from 1995-2010 looked at mortality and incidence over time (Butt et al., 2018). The authors note that some studies have shown a small increase in survival over time, implying improvement, and this agrees with their own findings, although it is only a very small increase. They note an SMR of 5.7 (95% CI: 4.7–6.4), higher than other studies, and they suggest that this finding is due to their study being larger (2,778 incident SSc) and more encompassing with a greater variety of ages. While older age at diagnosis was linked to higher mortality rates, the SMR was larger for younger ages at diagnosis due to the low mortality in the general population. Comparing causes of death, one fifth of all deaths was attributable to cardiovascular causes, a quarter to pulmonary diseases, and 15% were due to cancer.

A recent French study (Pokeerbux et al., 2019) of 625 patients diagnosed between 2000-2013 estimated an SMR of 5.73 (95% CI: 4.68–6.94), with 1-, 5- and 10-year survival at approximately 98.0%, 85.9%, and 71.7% respectively. They discuss the unusually high SMR in their study and suggest it could be due to a high presence of males and a high presence of those with a particular antibody.

The Tuscany study reported an average age of death in those with SSc as 74.0, with

1-, 5- and 10-year survival at approximately 98.4%, 91.6%, and 79.4% respectively (Coi et al., 2021). No SMR was estimated.

2.5 Discussion

Systemic sclerosis is a life-shortening disease, with a direct cause not yet fully established (although there are some hypothesised risk-factors). The reclassification in 2013 is notable, as this reclassification may affect our dataset intake, and we may see a different type of SSc patient entering the study. Based on the reclassification, there is likely to be a higher number of patients with less serious SSc, therefore it is to be anticipated that there will be longer survival times. While mortality in those diagnosed with SSc varies between decade and country, it is certainly much higher than the mortality observed in the underlying base populations. While lung involvement, such as pulmonary hypertension, is considered the leading cause of death in those with SSc, deaths due to cancers were also observed in a number of studies. While cancer SIRs vary between studies, there is the appearance that there is an increase, particularly with cancers that have lung involvement.

No notable study investigating cancer in those with SSc has been performed in the whole of the UK, and none with as large a dataset as ours, hence the motivation for our study. As seen above, many studies compare SSc to a national database as opposed to a matched comparator study. Notably, our study has:

- A robust, large number of SSc patients (780 incident and 806 prevalent)
- A UK based primary database, specifically constructed for the study of cancer
- A matched cohort-comparator study, which will be more accurate than an SIR/SMR based on national datasets which many of the studies above use.

However a noticeable disadvantage is that we do not have codes for

limited/diffusive SSc sub-types, which many of the above studies do identify and analyze separately, due to the different severity of the sub-types. We will also not investigate sub-types of cancer.

Chapter 3

Dataset characteristics

3.1 Introduction

This chapter provides a background to the dataset used in the study, highlighting its strengths and weaknesses. The contents are:

1. CPRD description
2. Notable adjustments to the dataset and missing data
3. Onset, start of follow-up, and incident or prevalent classification
4. Possible outcomes
5. Covariates

3.2 The Clinical Practice Research Datalink

CPRD collects fully-coded patient electronic health records from GP practices. Since 2017, this has been divided into two distinct databases depending on which software is used, using the Vision[®] or EMIS[®] software systems. CPRD collects data from practices using VisionVR software that contribute to the CPRD GOLD database, which has been used in epidemiological research for 30 years. There is

the more recent option of CPRD Aurum, which contains routinely-collected data from practices using EMIS[®] Web electronic patient record system software. While Aurum is new and larger, it was not available when the data for this study was first collated, therefore we use CPRD GOLD data. A recent paper by Wolf et al. (2019) details CPRD Aurum in more depth, but for the rest of this thesis when we refer to CPRD we mean CPRD GOLD.

The Clinical Practice Research Datalink (CPRD) GOLD is a primary care database of de-identified medical records derived from the records of UK general practitioners, with coverage of over 11.3 million patients from 674 practices in the UK. As this is a primary care database, it includes patients with a range of different ages, locations and health/illness statuses. It is important that data of the kind collated in CPRD GOLD is both clean and consistently derived and interpreted. To achieve this, only 4.4 million patient records – i.e., about 39% of the patients in the database - have been classified as usable. Because of this reduced number, the CPRD GOLD dataset available to this study only accounts for 6.9% of the UK population. Nevertheless, the available data has a good range of sex, age and ethnicity and therefore is believed to be an accurate portrayal of human health in the UK (Herrett et al., 2015).

CPRD GOLD has some notable strengths compared to other datasets. The breadth of data is wide, including factors such as morbidity, lifestyle, prescription details and information on methods of secondary care (although the latter is dependent on who the patients were referred to by their general practitioners). CPRD GOLD is a large primary care database, offering to studies such as the one in this thesis more precise estimates than those obtained from less extensive datasets. It has the potential for long term follow-up, as most patients included in the database will continue to have a registered GP and hence will continue to be included in extensions of the database over time. The prescriptions linked with patients give a good indication of the progress of medical treatment and also provide some indication of the patient’s medical history, including other illnesses or medication history.

However, there are weaknesses in CPRD GOLD which should be borne in mind in the application of this dataset. There is a chance of data being missed and excluded from the dataset, due to a variety of reasons but including erroneous reporting and misinformation. Examples are patients withholding information, doctors preferentially sampling (for example reporting BMI if they believe the patient is overweight) or differing interpretations of symptoms. There is also a loss of information when a patient is passed to secondary care centres, or their GP practice changes, and while it is preferred that the secondary care centres send information back to the primary care centre this may not always happen, and this can lead to a loss of follow-up. This may affect SSc patients who often go to secondary care centres for specialised treatment. However overall CPRD GOLD is a robust dataset, and its size is useful for a rare disease study such as that of SSc as considered in this study.

3.2.1 Government permissions

This study is based in part on data from the Clinical Practice Research Datalink obtained under licence from the UK Medicines and Healthcare products Regulatory Agency. The data is provided by patients and collected by the NHS as part of their care and support strategy. The interpretation and conclusions contained in this study are those of the authors alone.

The ISAC (Independent Scientific Advisory Committee) number for this study is 17_109. Appendix A3 contains the relevant form for the departmental study of SSc.

3.2.2 Validation of SSc cases

The CPRD GOLD SSc dataset that we are using was devised for multiple studies, not just cancer incidence. A paper by Pauling et al. (2021) goes into great detail about the validation that the SSc patients went through to be included in our dataset. The checks were extensive, and detailed thoroughly in the

section ‘Development and validation of case ascertainment strategy’. We briefly summarise. The key diagnosis was the read codes, devised to provide uniformity within primary care electronic health records (Chisholm, 1990). Firstly, relevant read codes that highly signify SSc was found, but were insufficient, so primary supporting evidence was also required. Even after this, the positive predictive value (PPV) was not sufficiently high, so secondary evidence was also reviewed. There was also criteria which gave penalties to patients, such as a diagnosis for SSc appears early in the patient record. After these requirements and many cases being individually reviewed, a PPV of >0.75 was achieved.

Similarly, for cancer outcomes, all relevant medical, referral and test codes to identify records of cancer diagnosis were used and then verified that these are accurate.

3.2.3 Linkage to hospital data

There was the possibility of linking data to Hospital Episode Statistics (HES), which offers a comprehensive resource for inpatient admissions, outpatient appointments and Accident & Emergency attendance records in England. This may provide us with more information on the diagnosis of SSc patients, cancer incidence or covariates which may be missing. However, it was decided prior to data analysis that the cost for this data was not justified as for CPRD GOLD there is only a linkage for 60% of patients.

3.3 Notable adjustments to the dataset and missing data

We have made several modifications to the dataset first received for this study, and these are noted here for transparency.

The first is that the initial dataset had some patients under the age of 18 (22 SSc patients). We have taken the decision to remove young cases of SSc and their

matches, reducing the total dataset by 154 patients. This decision was taken for several reasons; these include the rarity of SSc in young patients, that it appears to act in a different way to older cases of SSc, and that misdiagnosis is thought to be more likely.

Secondly, there were four patients in the initial database who were randomly assigned to be a comparator to SSc patients, but who after their match date then themselves received a diagnosis of SSc, and become cohort patients with their own set of comparators. There were two options for how to handle these patients:

1. Treat the patients as being censored when they are diagnosed with SSc.
2. Remove the patients from being a comparator in the dataset.

The second option was chosen. An advantage of this is that they may have had SSc prior to their diagnosis and therefore would not have been true comparators.

Their removal means we do not have a completely matched dataset. As there are only four such patients, this will not significantly affect the study. A further advantage is that it prevents these four patients from being used as both SSc patients and controls, hence preventing them being used twice in the study.

Lastly, there is one SSc patient whose time between SSc diagnosis and entry to the study is larger than the time from diagnosis to an observable event (death or cancer) in any other patient. As this adds complications later in this thesis, they and their matches have also been removed. This is also the most historic case, having been diagnosed with SSc in 01/01/1956.

After these four patients, the juvenile cases and the most historic case are removed along with their comparators, we have 11,098 patients remaining, both SSc and non-SSc (5,642 incident, 5,456 prevalent). We note that there are a few SSc patients whose date at SSc diagnosis is very close to or the same as their end event. In fact, there were 5 SSc patients who died on the same day as diagnosis. We suspect this is due to the patient being diagnosed with SSc soon after death. Again, there is a choice of excluding these patients due to unreliable times between

SSc diagnosis and event, or to include them and suggest diagnosis occurred extremely close to exit time. The decision was made to opt for a very short time between diagnosis and exit time and hence to retain these patients in the study. It would be interesting to examine the effect of this decision by redoing the analysis with these patients excluded, but as they are only 5 SSc patients out of 806 SSc incident patients and 780 SSc prevalent patients, we suspect there will be very limited impact on the study.

There was a small number of patients who have a smoking status of ‘giving-up’ as opposed to yes, no, or ex-smoker. We have treated those ‘giving-up’ as ‘yes’ (smokers) due to the small number and as they were, until recently, smokers.

Lastly, there is the question of missing data, a common issue in epidemiological studies. If missing data is present for a particular patient, the Cox model (see Chapters 4, 5, 6 and 9) will remove all such patients from the analysis completely, reducing the size of the dataset. We shall also be using covariates in Chapters 7, 8 and 9 where patients with missing covariates cannot be used. If patients with missing covariates are removed, then this may introduce bias when there is a relationship between a patient having a missing variable and the health of the patient. For example, BMI in the healthy range may be recorded less often due to the emphasis of some GPs on the recording of high BMIs, or a patient in poor health is more likely to have their data recorded. Therefore, by removing patients with no BMI information, we may be reducing the number of healthier patients, leading to the impression of worse health.

Methods to overcome this selection bias exist (Kang, 2013), but are not applied here. Instead, when using the Cox model we give two alternatives, one with covariates which contain missing values where patients with missing values are removed, and one without the covariates which contain missing values. In later chapters, when covariates can be used to model survival, the covariates of BMI and alcohol consumption (the two covariates with the most missing data) were found not to be efficient estimators, and are therefore not used, so removal of patients with these missing characteristics was not necessary. For a handful of

patients, smoking data was missing, but smoking status was found to be a good predictor of cancer and mortality when using the Cox model. Therefore, starting at Chapter 7 patients with missing smoking data will be removed so that smoking is retained as a covariate.

3.3.1 Matching

As discussed above, we have a matched dataset. A matched dataset is when those exposed are allocated corresponding matches, with the matching based on patients having similar covariate distributions. This is a method that aims to reduce confounding by allowing for *exchangeability* between the exposed and unexposed cohorts (see Section 7.1 of this thesis). For example, consider a clinical trial, when a drug is given to both the exposed and unexposed patients where the exposed are entirely older patients and the unexposed entirely younger patients and monitored for an outcome. If an outcome of death is more common in the exposed group, then it may be assumed that it is the drug that is causing this effect. However, the same effect could be observed, even if the drug had no effect, as increased age is associated with higher mortality. This is an extreme example to demonstrate the benefit of matching.

We have matched on three covariates:

1. Sex
2. Year of birth
3. GP practice

Sex is often matched due to it being a common confounder. Having a match based on year of birth means that the matched patients will have the same age at diagnosis (or ‘hypothetical’ diagnosis if non-SSc, see next section), and will live through the same time period. Matching based on GP practice adjusts marginally for location, which can be a proxy for other things associated to location e.g., a rural versus urban environments; socioeconomic factors; patient management.

However, this effect will not apply to all, as patients may have moved into locations, and may therefore have spent time in other locations during parts of their lives relevant to this study. Further, location is not always matched to exact GP practice due to a lack of possible patients, in which case other adjacent practices in the same area were found and used instead.

3.4 Onset, start of follow up, and incident or prevalent status

We have a retrospective cohort study, and patients are selected based on having an SSc diagnosis. We match each of these cases with comparators, patients who did not have SSc at the time of the diagnosis of their SSc match and are alive and cancer free when the SSc patient enters the study. The outcome of key interest to this study is a cancer diagnosis.

Death is also an outcome which can be recorded for all patients, even if they have previously had cancer, if death occurs before they leave the study. This is useful for both Chapters 4 and 5 where we consider mortality.

The official study start date is 01/01/1998, as this date is believed to be the start of reliable patient data (partially due to the start of electronic recordings). But the recruitment date of any one participant may be later than the study start date depending on when the practice first started contributing to CPRD GOLD. Therefore, note that for terminology, there is a difference between the *overall* date of study start (01/01/1998) and a *patient's* start date, as some prevalent patients will not join the study until after 01/01/1998. The last date of follow-up is either when the patient's practice last contributed data to CPRD GOLD before the end of the study, or the date of the patient's death. The latest date a prevalent patient enters the study is 24/10/2016.

In most cohort studies, patients who have experienced the initiating event will be identified after the study start date and monitored onward for an outcome.

We define patients who are diagnosed with SSc while they are in our dataset as our ‘incident cases’. However, we have the opportunity to additionally include patients who were diagnosed with SSc prior to entry, but who are still cancer and death free. These patients constitute our prevalent cases. Inclusion of these patients increases sample size, allowing for greater statistical power and greater information. This is often a positive, especially with rare diseases with a long time period between exposure and outcome, such as would occur with SSc and cancer. However, there are disadvantages. The most significant is that in order for a prevalent patient to be included they need to have survived and be cancer free up until their recruitment into the study. This often means we have a length biased sample. We discuss this and how it is typically accounted for in the next chapter.

Prevalent SSc patients are recruited if they have been diagnosed with SSc prior to entry to recruitment to CPRD GOLD. Each prevalent SSc patient is then assigned 6 matches (comparators), the same as for the incident SSc patients, with each match being required to a) not have SSc prior to the prevalent SSc patient’s entry time, b) be alive and cancer free at the time of entry to the study and c) be in CPRD GOLD during the prevalent SSc patient’s date of entry. Matches will have dates of entry into our prevalent cohort identical to their matched SSc patient.

For the majority of this work we shall be comparing the risk starting from SSc diagnosis, if they are an SSc patient, or the diagnosis date of the SSc patient that a comparator (termed as non-SSc patients) is matched to. That is to say, time zero is from SSc diagnosis date. Regardless of whether they are SSc or non-SSc, we shall call their age at this date the ‘age at SSc diagnosis’. For SSc patients this approach is logical, however it could be argued to be less so for the non-SSc patients as we are supposing that the time their match is made is the time at which they could have been ‘assigned’ SSc, comparable to when a treatment is assigned in a clinical trial. This could be argued to be incorrect, as SSc is not an ‘allocatable’ disease. However, we believe this is the best measure of causal

risk and comparison between the two groups. This also allows this work to be applicable to drug trials or other future studies where the exposure is assigned.

The diagnosis dates of SSc patients in the study are of interest, as shown in Figure 3.1. The incident patients are diagnosed from 1998 to the study end, and prevalent patients range from 1957 to the end of the study. Patients enter the study after 1998 only (by the self-imposed limits of our study start date) however several practices did not join CPRD GOLD until after 1998, which is why prevalent patients continue to be diagnosed after this time. If a practice had an SSc patient diagnosed in 1999, but the practice did not join CPRD GOLD until 2001, then we define this as still being a prevalent patient, for example. We discuss this more in Section 5.3. This partially explains why we do not have as large a recruitment of incident SSc patients in 1998-2000, as there are fewer practices contributing to CPRD GOLD. There are more prevalent patients joining in the later calendar times of 1980s and 1990s than the earlier calendar times, which is consistent with the definition of prevalent cohorts, as patients are much more likely to enter the study having survived 1 year rather than 20 years, for example.

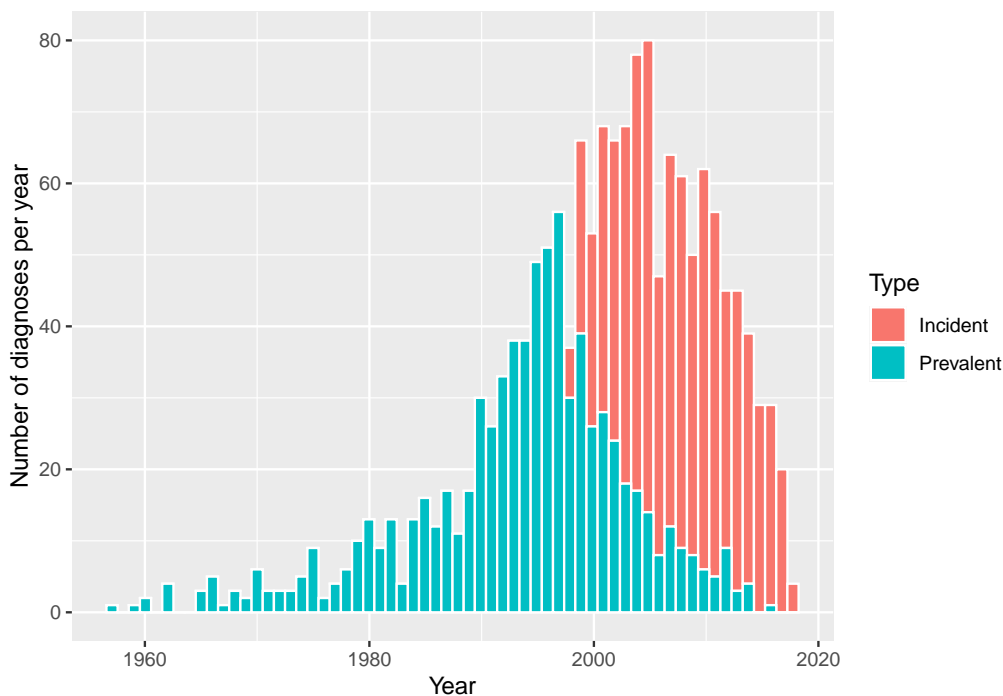


Figure 3.1: Distribution of diagnosis dates over calendar time.

3.5 Possible outcomes

Patients have three possible end events: cancer, death and censoring.

Censoring occurs if the study ends and the patient has not been diagnosed with cancer or has died before this time. Figure 3.2 shows the number in 4 subgroups of the data (incident or prevalent, SSc or non-SSc) over calendar time. The incident cohorts by definition do not enter the study until after the study start, so we see the number in the study increase from 0 at the start of the study. After a certain calendar time the number of people leaving the study eventually becomes greater than those entering the study, hence the decline. In the prevalent cohort, we do start with a number of patients entering the study at 01/01/1998 as they were diagnosed with SSc prior to this date. This number continues to increase as some enter the study later (due to CPRD GOLD still registering practices), however again after a certain time the number having an event or being censored is greater

than those entering, hence the decline.

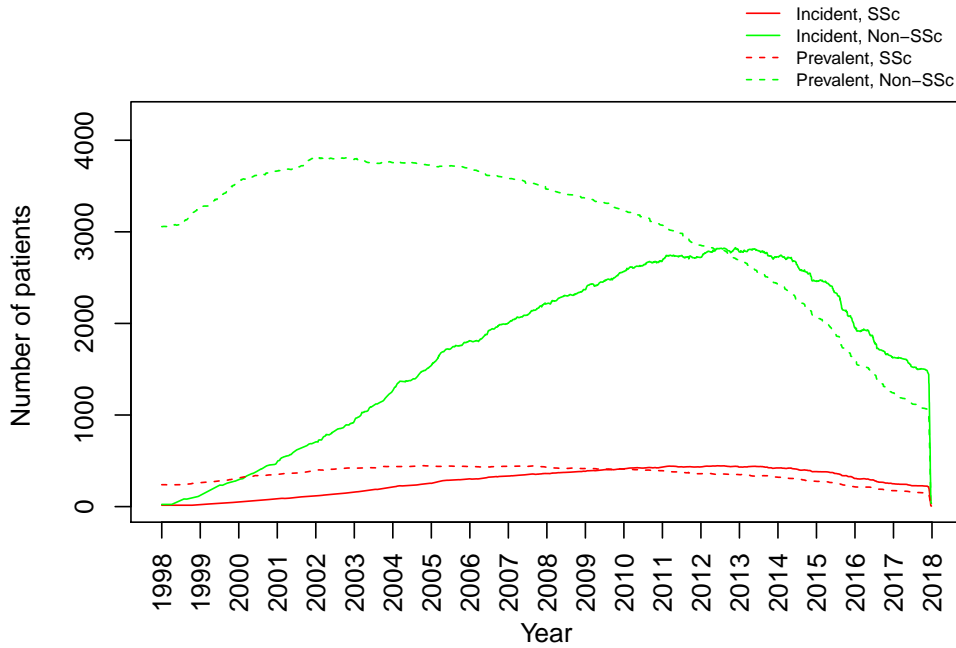
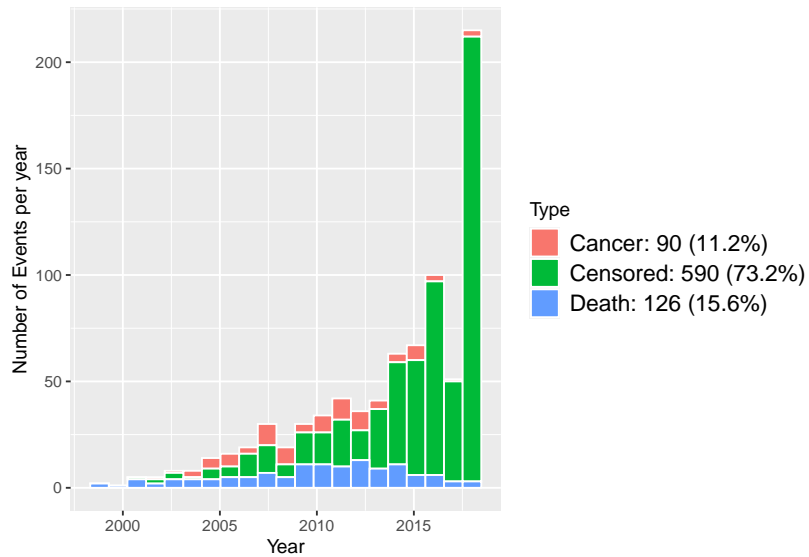


Figure 3.2: Number in the risk set by type of patients, whether they are incident or prevalent and if they are SSc or non-SSc.

Figures 3.3, 3.4, 3.5 and 3.6 show the distribution of events over calendar year. The proportion who have each type of event in each of the subgroups is shown in the key of the figures. In all plots we can see the sudden jump in the number of censoring cases in the final study year, however censoring events do occur before this. Reasons for this censoring include losing patients from the dataset, for example due to emigration. It could also reflect the last time that the practice updated their records to CPRD GOLD, leading to the large increase in cases being censored in the last few years. Unfortunately there is no way to distinguish which type of censoring has occurred for each patient. The number of deaths without cancer is greater in both the incident and prevalent SSc groups, which can be attributed to the higher mortality of SSc. There are more events (less censoring) in the prevalent cohort. This is likely due to both the age difference between the incident and prevalent cohorts (the incident cohort has a lower average age

during the study, see Section 5.1) and that the incident cases have less time with SSc (which is a maximum of 20 years due to study length). Both of these factors result in fewer events being observed in the incident cohort, and lend support to the use of prevalent cohorts as more events will be seen.



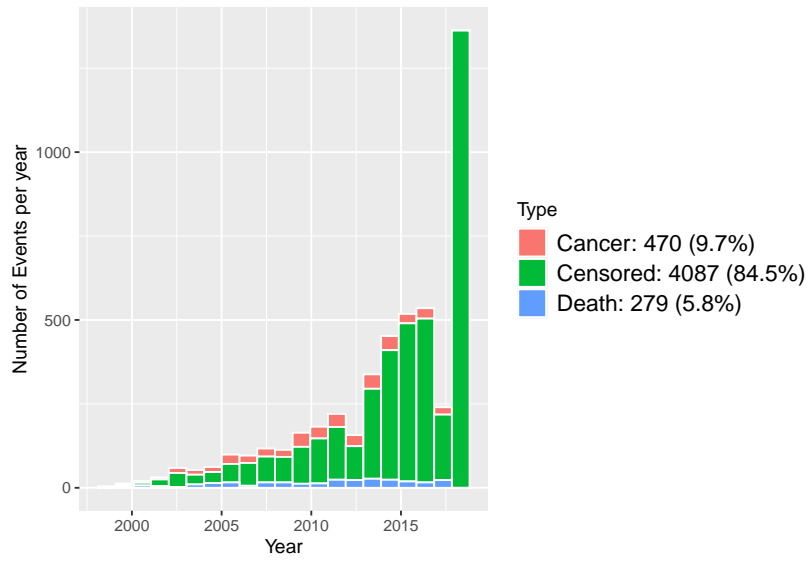


Figure 3.4: End events by year in the incident non-SSc set. The number corresponds to the number of events in total and the percentage is the proportion that had that event in the incident non-SSc group.

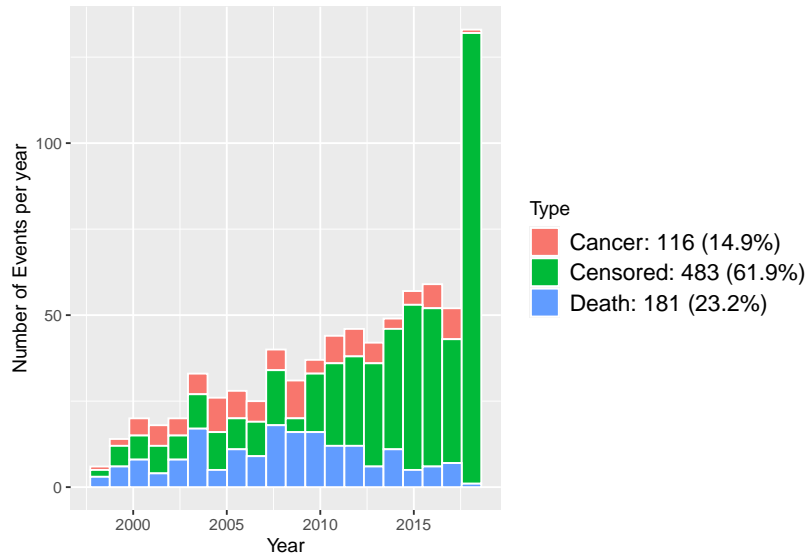


Figure 3.5: End events by year in the prevalent SSc set. The number corresponds to the number of events in total and the percentage is the proportion who had that event in the prevalent SSc group.

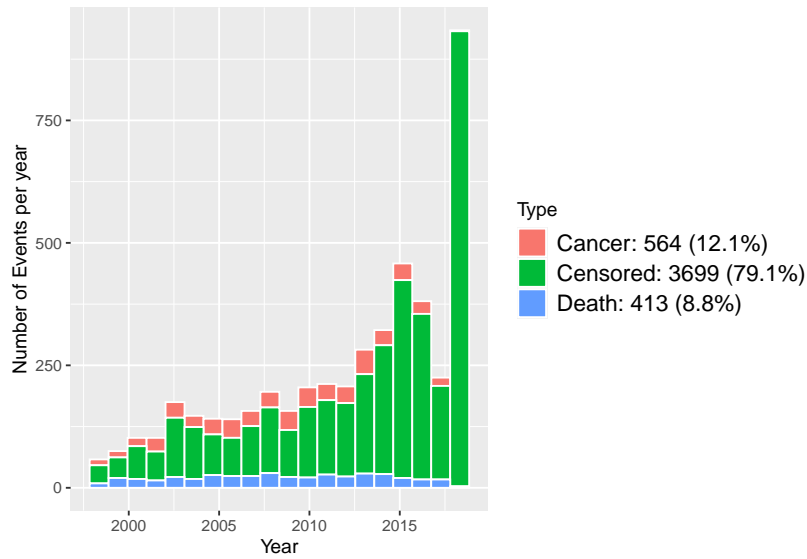


Figure 3.6: End events by year in the prevalent non-SSc set. The number corresponds to the number of events in total and the percentage is the proportion who had that event in the prevalent non-SSc group.

3.5.1 Lexis diagrams

We include Lexis plots of the incident (left) and prevalent (right) times from diagnosis of SSc to study exit in Figure 3.7. These plots specifically show SSc patients. Each line represents one or more patients, depending on if there are overlapping lines. In the incident plot, the green line is the time from when a patient was diagnosed to their study exit. Prevalent patients have two connecting lines, where the grey is the time from diagnosis to their study entry, and the green is from study entry to study exit. If an event occurred at study exit, then it is marked with a cross, where cancer is red and death is blue.

The plot on the right demonstrates the amount of time for which prevalent patients had SSc prior to their study entry, which in many cases was decades. This indicates the bias towards the younger age of diagnosis for prevalent patients, but also indicates the older ages at their study entry time.

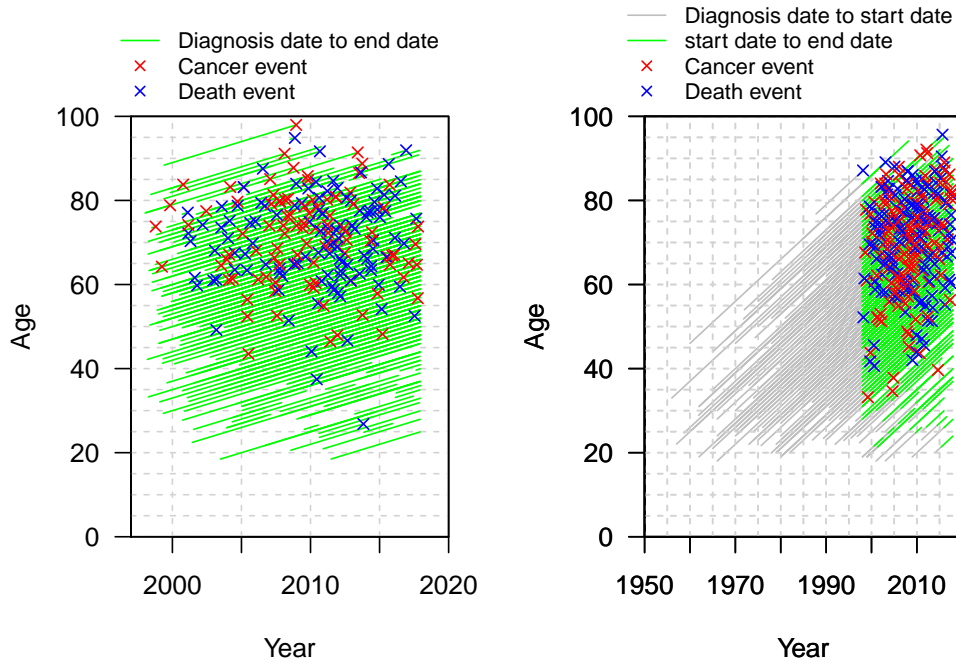


Figure 3.7: Lexis plot for incident (left) and prevalent (right) patients.

It is difficult to discern information from the above plots, due to the number of patients in the datasets, and the multiple overlapping lines shown. Therefore, we also provide heatmaps of the Lexis plots, where the value of each block is the total time covered by patients in the study (by calendar year and age in that year), including patients' time from their diagnosis of SSc to the time they exit the study. Figure 3.8 shows this for the incident SSc cohort, and Figure 3.9 for the prevalent SSc cohort. The dark red areas are where the calendar time and age under study has the greatest follow-up, and dark blue is follow-up time contributed by only a few patients. For example, for the prevalent cohort there are approximately 40 years of follow-up time provided by patients who were 42-44 years old in the year 1994-1996.

Again, we can see that the prevalent cohort covers a much wider spread of calendar time between time of diagnosis and exit from the study (or study end date). In the incident plot, we can see that the highest density of period under study occurs in 2009-2011 with patient age 60-69. In the prevalent set, the highest density of

follow-up times occurs in 1999-2005, with patient age approximately 48-65. These figures demonstrate two important differences between our two cohorts: firstly, that the prevalent set includes a much earlier calendar time period, and there may therefore be differences in survival over time which contribute to the results obtained for the incident and prevalent patients, and secondly that the patient ages under study for the prevalent cases tend to be significantly younger than those for the incident cases.

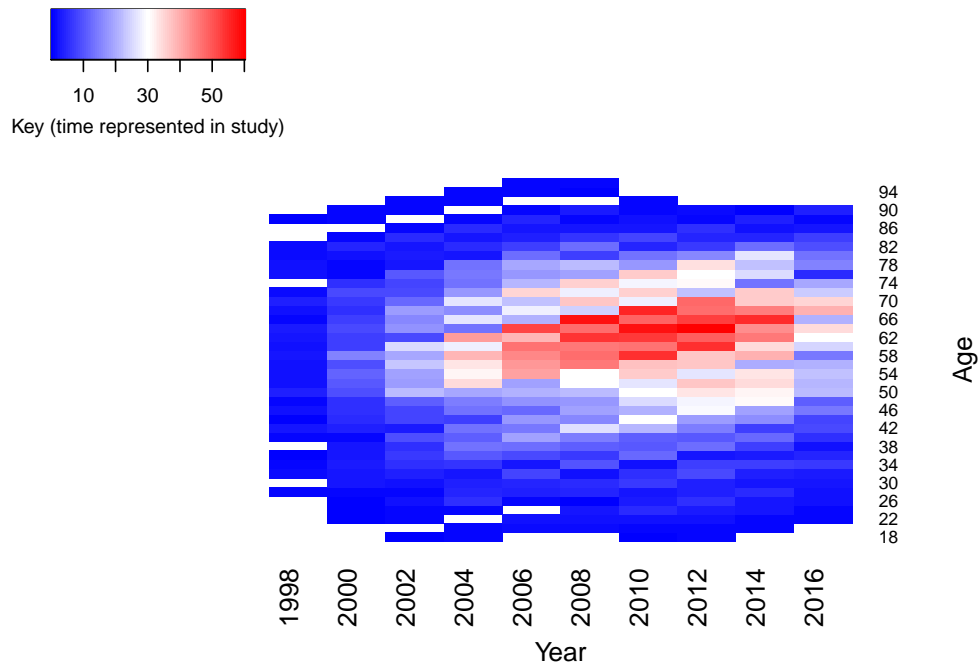


Figure 3.8: Lexis plot for incident SSc patients. The plot is the density of follow-up time subdivided by age and calendar year. The width of each square is 2 years.

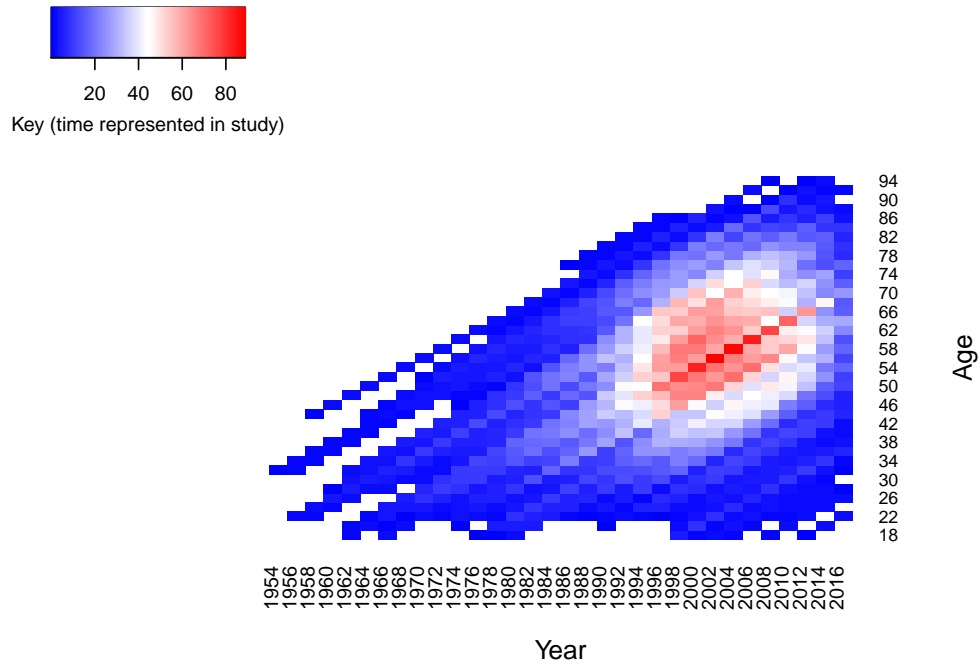


Figure 3.9: Lexis plot for prevalent SSc patients, from diagnosis of SSc to end date. The plot is the density of follow-up time subdivided by age and calendar year. The width of each square is 2 years.

3.6 Correlation between age at diagnosis and truncation time

In future chapters (Chapters 7, 8, and 9) we will be concerned about a correlation between a patient's time between diagnosis of SSc and entry into the study, and their age at SSc diagnosis. To investigate this, below are plots (Figures 3.10, 3.11 and 3.12) showing the relationship between a) the truncation time and age at SSc diagnosis, b) truncation time and time to event, and c) age at SSc diagnosis and time to event. Each cross represents a patient, and the colour corresponds to an event (red, blue and grey for cancer, death and censoring, respectively). Truncation time is the time from diagnosis of SSc to entry into the study, and time to event is from SSc diagnosis to event/censoring. It should be noted that there could be overlapping patients, in particular in the non-SSc patients for (a),

as we know by study design that 6 patients will have the same truncation time and age at entry combination due to matching on date of birth. We expect there to be differences for time to event, as this will vary in every patient, but again there may be overlap.

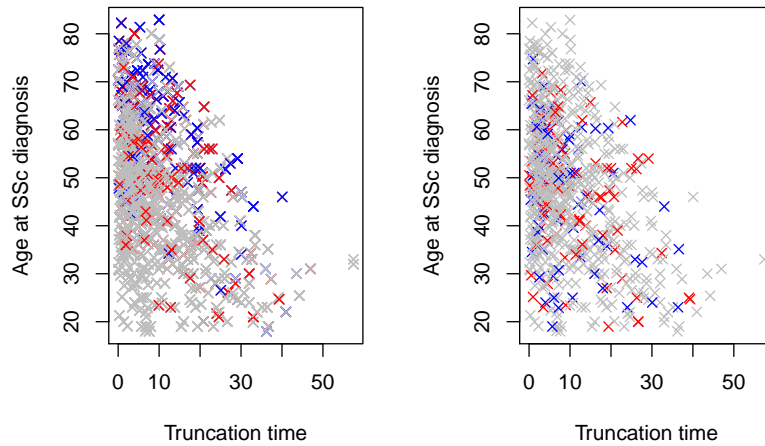


Figure 3.10: Truncation time versus age at SSc diagnosis. Left is non-SSc and right SSc. Colour scheme: grey is censored, red is cancer, blue is death.

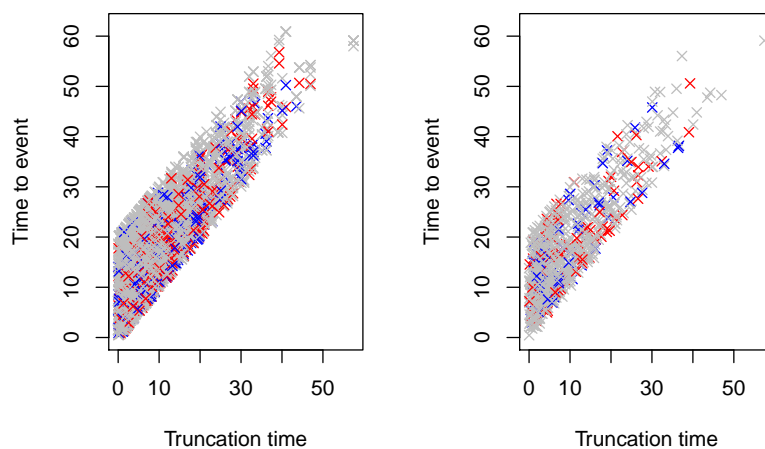


Figure 3.11: Truncation time versus time to event. Left is non-SSc and right SSc. Colour scheme: grey is censored, red is cancer, blue is death.

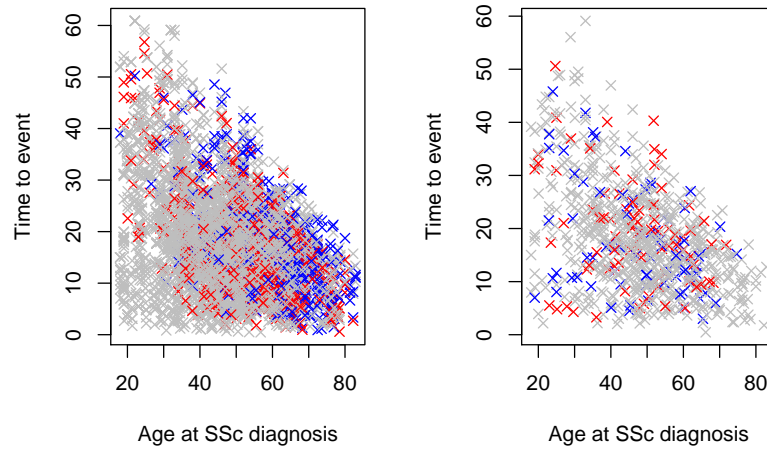


Figure 3.12: Age at SSc diagnosis versus time to event. Left is non-SSc and right SSc. Colour scheme: grey is censored, red is cancer, blue is death.

In Figure 3.10 there is a strong relationship between truncation time and age at SSc diagnosis, as we require patients to be event free at the start of 1998, therefore greater truncation times correspond to lower age at diagnosis. In the plots, the patients with the highest ages at entry to the study will be along the diagonal (where age at entry plus truncation time is highest). Therefore, we would expect deaths, and to a lesser extent cancer, to occur more frequently along the diagonal. We indeed do see more events at higher ages at SSc diagnosis time, as these patients entered the study older and were observed for 20 years and had a relatively high likelihood of an event. This is particularly noticeable for the non-SSc group, but in the SSc group we observe more deaths over a wider range of time points, not just those with higher age at events, with fewer censoring events. The points on the non-SSc and the SSc graphs are the same relative to each other, due to the matched nature of the dataset, as they share common ‘age at SSc diagnosis’ and ‘truncation time’ information. This also means there are overlapping points in the non-SSc plot as six patients will have the same age at SSc diagnosis and truncation time.

Figure 3.11 shows the relationship between truncation time and event time. The

constraints on the plot are due to the truncation time and event time being a maximum of 20 years apart. There are no strong observations to be made from the plots, however we might expect more cases at higher truncation times due to the longer disease duration.

In Figure 3.12, for the non-SSc group we observe that a higher age at diagnosis results in more death and cancer events. Again, there is the expectation of events along the diagonal, as a long time from SSc diagnosis to event implies a higher age during the study, however there is a noticeable lack of events at higher times to event and at lower age at SSc diagnosis. The SSc group has more sporadic events.

Some prior prevalent cohort research has been focused on the idea that the study starts at one time-point, which is centralised on the study having a set calendar time of recruitment, where patients are only included if they have the exposure of interest at this time (for examples, see M.-C. Wang (1991) and Keiding (1991)). However we cannot claim to have observed this, as we have some prevalent patients who were diagnosed after 1998 but are still considered prevalent as the practice did not start to contribute data until after this date. This means that there is no longer a one-to-one correspondence between truncation time and calendar time, and our study is not the same as a cross-sectional study. Therefore, while truncation time and calendar time may be highly correlated, and while the literature discussed often uses truncation time, we shall often be using calendar time.

3.7 Covariates

We have a few covariates in our model that we have the option of utilizing. Firstly there are the three that cohort and comparator patients are matched on: year of birth, sex, and practice ID. When there was not an appropriate match at the same practice another was found at a practice in the same broad geographical location. Therefore, we shall only discuss the 11 levels of area in England, and also Scotland and Wales. Note that 85.3% of the dataset is women (82.8% in

incident set, 87.8% in prevalent set), due to the high number of women with SSc. We also have data on the body mass index (BMI), alcohol intake and smoking status.

BMI (kg/m^2) is divided into underweight (<18.5), normal (18.5-25.4), overweight (25.5-29.9), obese (30-34.9), severely obese (35-39.9) and morbidly obese (40+). BMI is recorded prior to study entry. Proportions of BMI within the SSc and non-SSc sets are shown in Figure 3.13. The largest category is 18.5-25.4, which is the recommended healthy weight, with the second highest category being slightly overweight. This is in keeping with the national population. There appears to be a lower average BMI for those with SSc, with more than double the proportion being underweight and the two highest BMI categories lower by a factor of about 2. Whether this is due to doctor's recommendation to keep a low weight, a cause or even an effect of SSc is unclear. However, we also note the larger number of missing observations, especially in the SSc set, and it may be that through biased reporting we see a difference between the two groups as non-SSc patients may only have their weight recorded if their practitioner was concerned based on interview, biasing the recorded weights to extremes. The number of missing patients is greater for SSc patients, which after inspection is due to a much later number missing for prevalent patients (33.7%). Why there is this bias for prevalent patients is unknown.

If we assumed that missing BMI was proportionally distributed between the BMI levels (missing at random), then the BMI for the non-SSc set would be 1.8%, 40.4%, 32.7%, 19.8% and 2.7% for underweight, normal, overweight, obese and morbidly obese respectively. We can attempt to compare this to the general population using the 2019 Health survey of England (NHS, 2020). If we took our data to be those aged 45-54 years (the average age for those diagnosed with SSc in our dataset) in 2017, the last year under study, for women the BMI percentages were 0.9%, 32.0%, 30.1%, 37.0% and 7.2% for underweight, normal, overweight, obese and morbidly obese respectively. From this comparison, our CPRD GOLD

dataset appears to have a lower BMI in general¹. If we were to take an earlier year then we might expect a (slightly) lower BMI in the HSE, or if we chose a lower age category we might also expect a lower BMI. However, we also have approximately 15% men, who on average have a higher BMI.

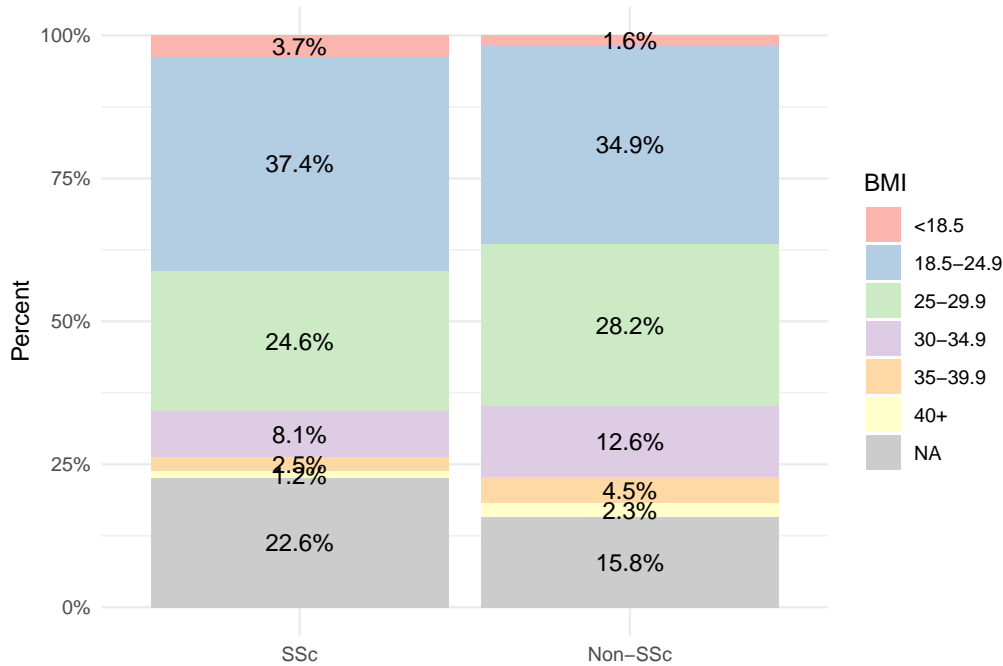


Figure 3.13: BMI distribution. NA is when patient information is not recorded.

Smoking can be categorized as “Smoker”, “Non-smoker” or “Ex-smoker”. The covariate distribution in the dataset is shown in Figure 3.14. Most are non-smokers, however, together smokers and ex-smokers make up just under half of the patients. There is a very small number of patients with missing data (2.95% of overall patients), with more of the non-SSc patients having missing data. There does not appear to be a large difference between the SSc and the non-SSc patients.

If we were to assume that the missing values proportionally distributed between the three smoking categories, then the split would be would have 52.2% for

¹HSE provides a BMI for all women of all ages as 1.8%, 36.7%, 31.5%, 30.0% and 4.6%, which is more in-keeping with our study but will have a lower age on average than our study.

non-smokers, 26.2% ex-smokers and 20.8% for smokers for SSc patients, and 54.6% for non-smokers, 23.3% for ex-smokers and 21.2% for smokers for non-SSc patients. The HSE puts the proportion at 57.8% non-smokers, 21.8% ex-smokers and 20.4% for smokers, for women in the year 2008 for those in the age range 45-54 (NHS, 2020). Our proportions are therefore similar.²

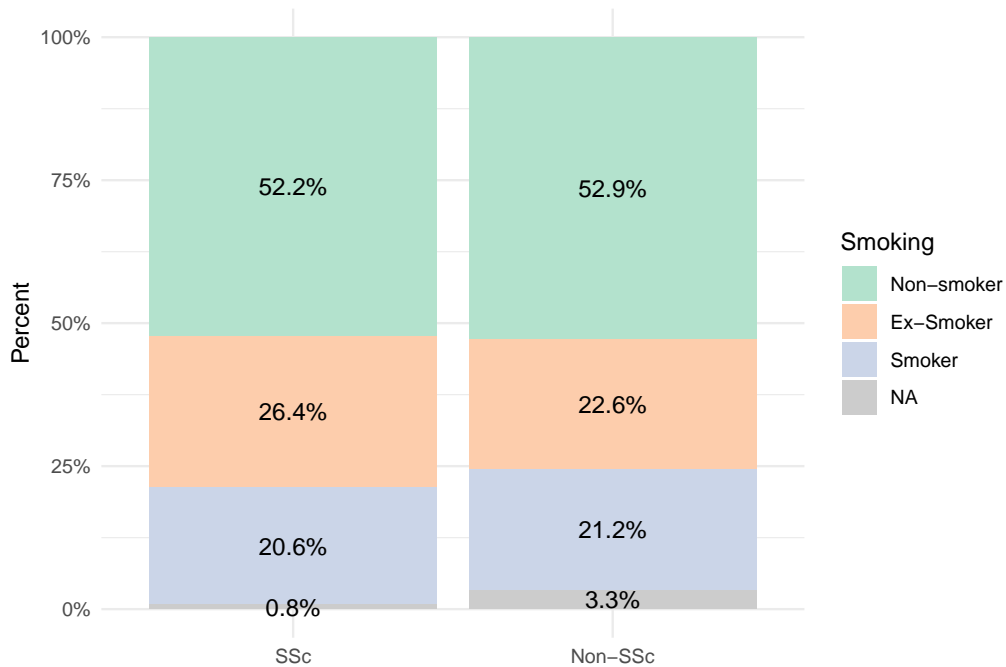


Figure 3.14: Smoking habit distribution. NA is when patient information is not recorded.

We have alcohol status as “Drinker”, “Non-drinker”, “Heavy-drinker” and “Ex-drinker”, with the distribution shown in Figure 3.15. Being a drinker is the most common type, followed by non-drinker. Approximately 9% of all patients are missing alcohol information. Notably, slightly more SSc patients are heavy drinkers and non-drinkers, however, this may be a consequence of less missing

²We have chosen 2008 at the year due to it being midway between our study start and study end. Females are a good comparison as the majority of the study is female, however males from the same age category are slightly higher smokers, with 50.4%, 27.4% and 22.2% for non-smoker, ex-smoker and smoker respectively. Women in the age group 45-54 are the most common in our study due to SSc being diagnosed later in life, however for women across all ages the number of smokers is reduced at 58.0%, 22.1% and 19.9% for non-smoker, ex-smoker and smoker respectively.

observations in this category as well.

If we assume missing alcohol usage is split in the same proportions as the non-missing then the proportions for SSc patients would be 70.1%, 6.5%, 4.4% and 18.6% for drinker, ex-drinker, heavy drinker and non-drinker. For non-SSc patients this is 75.3%, 5.69%, 2.5% and 15.7% for drinker, ex-drinker, heavy drinker and non-drinker, respectively. Comparison with a national database is difficult due to the different metrics of drinking between databases and surveys. A study investigating alcohol read codes in the UK puts the distribution at 20.3% non-drinkers, 3.7% ex-drinkers and 76.1% current drinkers based on 862 642 from the CPRD GOLD dataset (Mansfield et al., 2019). If we assumed heavy drinkers are part of current drinkers, these proportions are very similar to non-SSc patients. The HSE places the proportion in 2011 for women between 45 and 54 at 14.6%, 78.5% and 6.9% for non-drinker, drinker (0.1 to 35 units per week) and heavy drinker (>35 units per week) (NHS, 2020). Note that this does not include ex-drinkers.

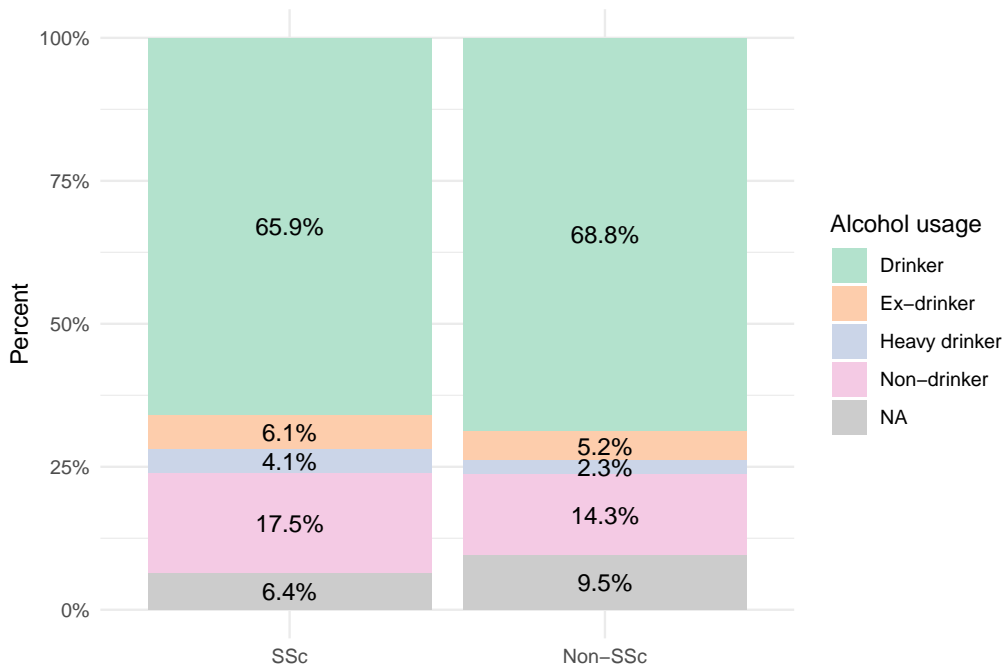


Figure 3.15: Alcohol usage distribution. NA is when patient information is not recorded.

Each patient has a practice ID, which will be one of 11 areas in England, or Scotland or Wales. This is shown in Figure 3.16. All patients have a GP practice, therefore they all have a location. There appears to be a reasonable number of patients in all geographical locations. There are large similarities between GP practice locations, which is expected due to matching, however there are some small differences in each location.

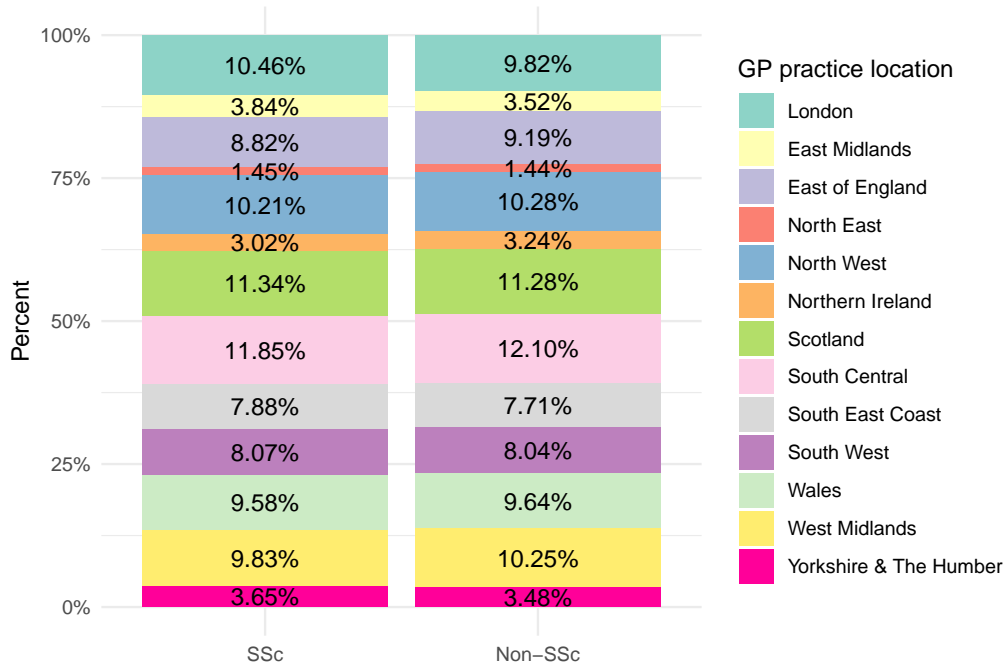


Figure 3.16: Distribution of GP locations.

The covariates BMI, alcohol intake and smoking habits were taken before entry into the study. In later chapters, our models will assume that the data was taken prior to exposure to SSc, which is not the case for prevalent patients, whose diagnosis of SSc (or date of match) might be closer to entry than diagnosis. This is a concern, as these covariates will change over time and could even be impacted by SSc. Both smoking and alcohol intake may reduce after a diagnosis of SSc, for example, due to greater health concerns. BMI may rise due to prevalent patients being older when they enter the study and BMI increase is often correlated with increasing age. As we justify below, we shall not use BMI, which we suspect will

have the largest change, or alcohol. We will include smoking, as we believe that smoking history will be more informative than no smoking history. Depending on the study, or the time between diagnosis and study entry, using covariates taken after diagnosis may not be advisable, especially if the diagnosis directly impacts the covariates, such as a change in medication.

There is also the concern of missing data, where missing data not at random will impact the study. We discuss this in more depth in later chapters, but due to the high number of missing data for alcohol and BMI, we do not use these in the study from chapter 7 onward. We do include smoking, as even if there is a hypothesised decrease in smoking use in SSc prevalent patients from their diagnosis to their study entry, we believe this will be a small number and the significance of smoking as a predictor outweighs this bias.

Chapter 4

Systemic Sclerosis and mortality

4.1 Introduction

We previously noted the need to consider competing risks, which arises when an event not under study precedes the event of interest, preventing us from observing the event of interest. This need to consider competing risks derives from our prior knowledge of SSc, where SSc patients are at an increased risk of mortality compared to those without an SSc diagnosis. We have already seen it has been widely established that SSc leads to a much higher mortality than that observed in the general population (Section 2.4), although estimates do vary. This was confirmed in our initial investigation into the data described in the previous chapter, where we have demonstrated that approximately 20% of SSc patients died during the study, compared to just 8% in the non-SSc group. As we wish to start with the more commonly used methodology and from there go on to include prevalent cohorts and then account for competing risks, this chapter (and the next) will focus on mortality in SSc patients only.

The aims of this chapter are:

- As we have a large robust UK dataset with mortality information included, it is of interest to quantify the increased risk of death in those with SSc. Therefore, in this chapter, we investigate mortality in our SSc dataset. We shall quantify the survival over time and also in comparison to our comparator dataset. We will also quantify the hazard ratio with and without covariates, and compare mortality with the general population.
- This chapter will also serve as an overview of the Kaplan-Meier curve and the Cox proportional hazards model. The overall goal of this thesis, looking at cancer, is complicated by competing risks. To avoid the complexities of competing risks, it will be beneficial to start with an analysis that does not suffer from this issue. It will be discussed in future chapters why the methods in this chapter are of themselves not sufficient when there are competing risks present. We will investigate Standardized Mortality Ratios (SMRs) in order to compare our CPRD GOLD data with the general population. We also investigate covariates, by introducing the Cox model and assessing covariate risks.

We have selected only the incident dataset for this chapter's analysis - the inclusion of the prevalent cases will be considered in the next chapter. Note that in this analysis we have used a dataset which is not entirely representative of true mortality rates. As our study was designed for the analysis of the risk of cancer, patients were included on the basis that they were cancer-free at the time of SSc diagnosis (or, if they are a comparator, cancer free at the time of their match to an SSc patient). Patients who had a cancer diagnosis either prior to SSc or prior to a possible match were not included, therefore we are missing a subset of patients who may be less healthy than the general population, due to us having fewer cancer patients than would be expected in the average population.

4.2 Cohort characteristics

The characteristics of the incident patient population are listed in Table 4.1. The onset of disease is defined as the time when the patient was diagnosed with SSc, or their match date if they are non-SSc patients. The age at entry is therefore the age at which SSc patients were diagnosed with SSc, or the age at which non-SSc patient’s match was diagnosed with SSc.

The majority of patients in the dataset are women. This agrees with previous research which has indicated that women are more at risk of SSc than men. Just over a tenth of SSc patients had a cancer diagnosis during the study, and 20.1% of patients died (with or without being diagnosed with cancer first), which is more than double that of the percentage of cancer deaths in the non-SSc group. The average age at which our SSc patients were diagnosed with SSc is higher than we have seen in some literature (58.1 years) : for example a US study from 2010 estimated a mean age of 45.5 years in 2,300 patients (Manno et al., 2011), and a Spanish study estimated 45 years in 1,037 patients (Alba et al., 2014). However, the THIN study from Chapter 2 also reports a very similar average age of 59 years (Rodríguez et al., 2019). In our study, in those who have died, the mean time between entry to study and death (and average time of follow-up) is short, but so is our average follow-up. This shorter timescale is a consideration, especially with a long-term, chronic disease such as SSc. Average BMI is notably higher in non-SSc patients.

4.2.1 Matching

Our study is matched, which means that every SSc patient has 6 matched non-SSc patients who have the same DOB, sex, and approximate geographical location (GP practice). This is done to ensure an equal distribution of the variables which are believed to be confounders among exposed and unexposed patients. However, depending on the matching and the dataset, further adjustment may be needed. It was decided early in the study that for the analysis of the study we would treat

	SSc (n=806)		Non-SSc (n=4836)	
	Number	(% of SSc)	Number	(% of non-SSc)
Female	667	82.8	4002	82.8
Male	139	17.2	834	17.2
Cancer diagnosed prior to death	90	11.2	470	9.7
Death during study	162	20.1	433	9.0
Smokers	161	20.0	1035	22.1
Ex-smokers	212	26.4	1184	25.2
	Mean	SD	Mean	SD
Age at entry (years)	58.1	13.6	58.1	13.6
Age at death (years)	73.1	10.2	76.0	10.7
Time between study entry and death (years)	5.3	4.0	6.0	4.1
Follow-up (years)	6.8	4.6	7.1	4.7
BMI (kg/m ²)	25.8	5.2	27.0	5.6

Table 4.1: Cohort characteristics of the incident cohort. SD is standard deviation. Death here includes those who may have had cancer prior to death. Age at death is calculated from the subset of patients who died, and study entry to death (years) is the time between SSc diagnosis to death in those who died. Smoking is based on 803 SSc patients and 4690 non-SSc patients. BMI is based on 615 SSc patients and 3286 non-SSc patients.

the study as unmatched, and still include matched covariates in the modelling if they appeared significant (as age and sex often does). Matching is usually done to adjust for confounding (see Section 7.2). In case-control studies, matching often needs to be adjusted for in order to not incur bias. Matching in cohort studies is less well researched. However, we shall now consider Sjölander & Greenland (2013), who discuss when further adjustment may be required for cohort studies.

In a matched study, accounting for the matched covariates when there are no other unmatched covariates is unnecessary. However, the researcher may wish to account for the matching covariates, depending on the model used. For example, in the Cox model, not including the matched covariates provides a consistent estimator of the hazard ratio in a population where the matching variables are distributed as in the matched cohort. Due to the non-collapsibility of the Cox model, if matching was included it would change the measure under study from a marginal one to one conditional on the covariates used.

An issue arises when there are unmatched confounders, with Sjölander & Greenland (2013) saying it is usually not valid to ignore the matching variables when adjusting for additional confounders (stated with a counterexample). However, there are conditions where adjusting for the matched covariates is not necessary (if at least one of the following holds):

1. The unmatched covariates are conditionally independent of the exposure, given the matching variables.
2. The unmatched covariates are conditionally independent of the matching variables, given the exposure.
3. The outcome is conditionally independent of the matching variables, given the exposure and the unmatched covariates.

We discuss these criteria for our study. The following DAG (see Section 7.2), Figure 4.1, shows the relationship between our exposure, outcome and matched/unmatched covariates.

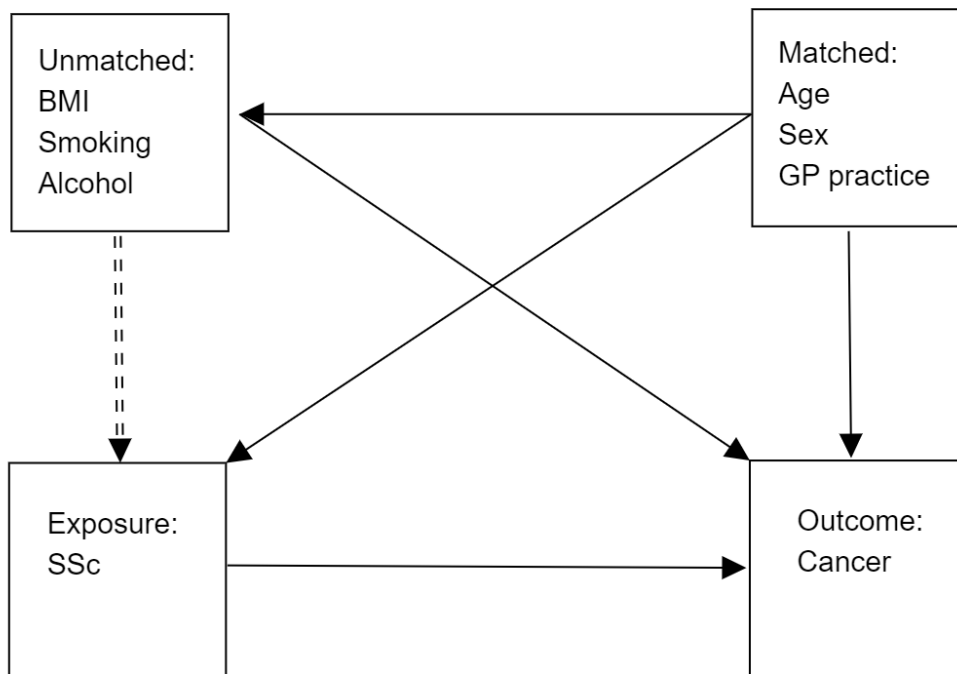


Figure 4.1: Directed Acyclic Graph, DAG, demonstrating the possible causal pathways of our study. The dashed line is due to unknown causal pathways between the unmatched covariates and our exposure.

From above, (2) does not hold as the matched covariates affect the unmatched covariates (both age and sex affect lifestyle habits, and GP practice is correlated to location, which will also impact lifestyle habits). Again, (3) does not hold as the matching variables are confounders for the exposure and outcome. However, whether (1) holds is debatable (hence the dashed line). There either is, or conversely is not, an impact of the unmatched covariates on the exposure. If there is not, then the unmatched covariates are conditionally independent of the exposure given the matching variables, and hence (1) holds. Conversely, if there is an impact of the unmatched covariates on the exposure then (1) does not hold and adjusting for matching is necessary. While there has been no evidence that BMI, smoking or alcohol intake affect SSc onset (Abbot et al., 2018), this may be due to too small a sample size of existing studies or insufficient research in general. We discuss these two possible scenarios:

- If there is not believed to be an impact of the unmatched covariates on the exposure, then adjusting for matching is unnecessary. However, we have continued to do so for age and sex in this body of work. For the Cox model, this changes the measure from a marginal one to conditional based on age and sex. For the logistic regression (Chapters 7-9) this was necessary for the weightings that were used.
- If the arrow is present, then (1) does not hold and therefore accounting for matching is necessary. As stated above, we have done so for age and sex, however we have not done so for GP practice. With hindsight it may have been better to include GP practice as a random effect term¹ in the model, but this was not done within the analysis. Matching could have been covered in more depth, with GP practice added as a random effect, therefore this is mentioned in 10.4.2 Limitations.

Although not covered here, articles that may be of interest for those doing a case-control study are Mansournia et al. (2013) and Mansournia et al. (2018), which approaches the bias from a causal framework with directed acyclic graphs (DAGs, see Section 7.2).

4.3 Standardized mortality ratio

Standardized Mortality Ratios (SMRs, or Standardized Incidence Ratios, SIRs, when mortality is not the event of interest) are used to compare the occurrence of an event in a cohort with that observed in a given population. SMRs have been used in other SSc studies to compare the mortality of those with SSc compared with the general population. In particular, a meta-analysis (Rubio-Rivas et al., 2014) compared 17 SSc studies, where the SMRs ranged between 1.05 and 5.40, with an overall SMR of 2.42 in studies whose mid-cohort year was set after 1990.

Mathematically, SMR is the ratio of observed events and expected events:

¹A random effect term due to most patients having a unique GP practice ID, and would therefore be inefficient as a separate covariate.

$$SMR = \frac{\sum_j d_j}{\sum_j n_j r_j} = \frac{O}{E}$$

where O is the observed events in the cohort population and E is the expected number of events. For each stratum, j , we define d_j as the total observed events, n_j as the person-years contributed by our patients, and r_j as the expected rate in the reference population. Our comparison (our expected rate) comes from the Office of National Statistics (ONS) (ONS, 2020). In particular, we take the mortality rates from 2017, which is the closest year to our study end date. The standard error is approximately $SE = SMR/\sqrt{O}$ and therefore confidence intervals are given by:

$$CI = \frac{O}{E} \pm z \frac{\sqrt{O}}{E}$$

where $z = \Phi^{-1}(1 - \alpha/2)$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution. Table 4.2 demonstrates the derivation of the SMR for SSc patients, Table 4.3 demonstrates the derivation of the SMR for non-SSc patients, and Table 4.4 provides a summary of results.

Age group	Female						Male						Total	
	Person years (n)	Number in group	Number of events (d)	ONS mortality rate (r)	Expected events (n * r)		Person years (n)	Number in group	Number of events (d)	ONS mortality rate (r)	Expected events (n * r)		Total Observed	Total Expected
15-19	1.5	1	0	2e-04	0		1.5	1	0	3e-04	0	0	0	
20-24	20.1	8	0	2e-04	0		6.4	2	0	5e-04	0	0	0	
25-29	50.0	18	0	3e-04	0.01		12.6	7	1	6e-04	0.01	1	0.02	
30-34	81.0	29	0	4e-04	0.03		22.8	8	0	8e-04	0.02	0	0.05	
35-39	143.0	57	0	7e-04	0.10		25.1	9	0	0.0011	0.03	0	0.13	
40-44	244.7	84	0	0.0011	0.27		33.4	11	0	0.0017	0.06	0	0.33	
45-49	322.0	111	2	0.0016	0.52		68.7	24	0	0.0026	0.18	2	0.70	
50-54	460.2	156	2	0.0025	1.15		135.6	42	1	0.0037	0.50	3	1.65	
55-59	609.0	194	5	0.0037	2.25		133.9	48	2	0.0057	0.76	7	3.01	
60-64	733.8	240	13	0.0060	4.40		140.4	52	8	0.0092	1.29	21	5.69	
65-69	698.1	234	21	0.0094	6.56		123.5	43	5	0.0145	1.79	26	8.35	
70-74	561.0	199	27	0.0151	8.47		78.2	30	3	0.0227	1.77	30	10.24	
75-79	385.7	130	26	0.0277	10.69		62.6	19	7	0.0398	2.49	33	13.18	
80-84	206.9	81	19	0.0506	10.47		20.7	11	7	0.0694	1.44	26	11.91	
85-89	69.5	27	6	0.0975	6.77		2.0	2	1	0.1241	0.25	7	7.02	
90+	23.0	10	6	0.2141	4.92		0	0	0	0.2370	0	6	4.92	
				$O_F=127$	$E_F=56.6$				$O_M=35$	$E_M=10.6$		$O_T=162$	$E_T=67.2$	

Table 4.2: Mortality rates in SSc patients stratified by age and sex, compared with the background population.

Age group	Female					Male					Total	
	Person years (n)	Number in group	Number of events (d)	ONS mortality rate (r)	Expected events (n * r)	Person years (n)	Number in group	Number of events (d)	ONS mortality rate (r)	Expected events (n * r)	Total Observed	Total Expected
15-19	8.7	6	0	2e-04	0	9.1	6	0	3e-04	0	0	0
20-24	98.8	47	0	2e-04	0.02	29.9	12	0	5e-04	0.01	0	0.03
25-29	257.6	104	0	3e-04	0.08	86.8	40	1	6e-04	0.05	1	0.13
30-34	448.1	169	1	4e-04	0.18	157.7	54	0	8e-04	0.13	1	0.31
35-39	797.1	332	1	7e-04	0.56	134.7	54	0	0.0011	0.15	1	0.71
40-44	1354.0	485	1	0.0011	1.49	194.9	63	0	0.0017	0.33	1	1.82
45-49	1841.0	640	3	0.0016	2.95	381.8	145	2	0.0026	0.99	5	3.94
50-54	2680.5	916	4	0.0025	6.70	773.5	243	1	0.0037	2.86	5	9.56
55-59	3573.7	1159	15	0.0037	13.22	797.2	285	1	0.0057	4.54	16	17.76
60-64	4178.5	1428	28	0.006	25.07	976.7	324	12	0.0092	8.99	40	34.06
65-69	4344.6	1370	29	0.0094	40.84	827.6	285	11	0.0145	12.00	40	52.84
70-74	3627.4	1233	65	0.0151	54.77	570	202	9	0.0227	12.94	74	67.71
75-79	2708.8	850	59	0.0277	75.03	397.1	127	19	0.0398	15.8	78	90.83
80-84	1747.9	601	79	0.0506	88.44	238.4	74	7	0.0694	16.54	86	104.98
85-89	670.2	242	47	0.0975	65.34	101.3	43	12	0.1241	12.57	59	77.91
90+	163.0	65	21	0.2141	34.91	19.1	9	5	0.2370	4.52	26	39.43
					$O_F=353$					$E_F=409.6$	$O_T=433$	$E_T=502.0$

Table 4.3: Mortality rates in non-SSc patients stratified by age and sex, compared with the background population.

	SSc		Non-SSc	
	SMR	95% CI	SMR	95% CI
Female	2.24	[1.85, 2.63]	0.86	[0.77, 0.95]
Male	3.30	[2.21, 4.40]	0.87	[0.68, 1.06]
Total	2.41	[2.04, 2.78]	0.86	[0.78, 0.94]

Table 4.4: SMR for SSc and non-SSc patients with 95% confidence intervals.

According to our estimates, there are 2-3 times the number of deaths in our SSc population than we would expect in the same age groups of the UK population, indicating a higher mortality, as expected. The mortality appears to be higher for males, but it should be noted that there is a smaller number of males in the dataset, and the SMR's have wide confidence intervals. For non-SSc patients, our data appears to have a slightly lower mortality than we would expect. A reason for this may be the preferential sampling of comparators who did not have a diagnosis of cancer prior to their match date. This might imply the SSc SIR is an underestimation as well.

SMRs are often used due to a lack of comparators, however we have the benefit of matched non-SSc patients. This enables a better analysis of risk with the methods we are about to use. We have matched on location and date of birth such that those matched were born in the same year, so using the methods described below allows us to work on a finer scale than the grouped age categories of the SMR method above. Also, the below methods allow SSc patients to be compared to patients living through similar times, as opposed to a set comparison rate of risk set in the year 2017. Lastly, the SMR is one result of one number, implying no variation over time from SSc diagnosis, where we might be interested in how risk changes over time, an effect which the Kaplan-Meier will allow for.

4.4 Survival, and the Kaplan-Meier curve

We use the Kaplan-Meier to estimate survival. In our study, survival will be measured from a diagnosis of SSc (or in the case of their matches, the hypothetical diagnosis date). We define the survival function as the probability of survival beyond time t ,

$$S(t) = \Pr(T > t)$$

where T is a random variable denoting the time that the event occurs. We also define the cumulative distribution function (CDF) as

$$F(t) = \Pr(T \leq t) = \int_0^t f(u)du = 1 - S(t)$$

where $f(t)$ is the probability density function (PDF). This can be interpreted as the probability of having an event prior to time t . In the absence of competing risks, the Kaplan-Meier estimate of the survival function is frequently used for estimating the survival function, and has the relationship $CIF(t) = F(t) = 1 - S(t)$, where CIF is the cumulative incidence function. This definition will differ when competing risks are introduced.

The hazard rate (or hazard function) is a useful and commonly used method to assess time to event date. It is the instantaneous rate of failure (of an individual experiencing the event) at time t given that an individual is alive at time t , formulaically written

$$\begin{aligned} \lambda(t) &= \lim_{h \rightarrow 0} \frac{P(t \leq T < t + h | T \geq t)}{h} \\ &= \lim_{h \rightarrow 0} \frac{1}{h} \frac{P([t \leq T < t + h] \cap [T \geq t])}{P(T \geq t)} \\ &= \lim_{h \rightarrow 0} \frac{1}{h} \frac{P(t \leq T < t + h)}{P(T \geq t)} \end{aligned}$$

Therefore $\lambda(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d \log(S(t))}{dt}$

This leads to the cumulative hazard function defined as $\Lambda(t) = \int_0^t \lambda(u) du = -\log(S(t))$ which leads to the survival function as

$$S(t) = \exp\{-\Lambda(t)\}$$

Therefore we could use any distribution defined for $t \in [0, \infty)$ to model a survival distribution.

We now estimate these functions using a common nonparametric method. We define the Kaplan-Meier estimate of the survival function (Kaplan & Meier, 1958)

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d(t_j)}{r(t_j)}\right)$$

where we define the observation time of events at $t_1 < t_2 < \dots < t_n$ of n patients, and $r(t_j)$ is the number of individuals at risk (which is defined as those who have neither been censored nor experienced the outcome of interest by this time) prior to time t_j , and $d(t_j)$ is the number of individuals who experience the outcome of interest at time t_j .

The confidence interval is traditionally calculated using Greenwood's calculation of the variance (Greenwood, 1926):

$$\widehat{Var}\{\hat{S}(t)\} = \hat{S}^2(t) \sum_{t_j \leq t} \frac{d_j}{r_j(r_j - d_j)}$$

such that we have confidence intervals:

$$\left[\hat{S}(t) e^{-z_{1-\alpha/2} \sqrt{\widehat{Var}\{\hat{S}(t)\}}}, \hat{S}(t) e^{z_{1-\alpha/2} \sqrt{\widehat{Var}\{\hat{S}(t)\}}} \right]$$

4.4.1 Cumulative hazard function

The Nelson-Aalen estimator is a nonparametric estimator which may be used to estimate the cumulative hazard rate function from censored survival data. The Nelson-Aalen produces an increasing right-continuous step function with increments $d(t_j)/r(t_j)$ at observed failure times. It is a discrete form of the cumulative hazard as defined above, $\Lambda(t) = \int_0^t \lambda(u)du$, and it takes the form

$$\hat{\Lambda}(t) = \sum_{t_j \leq t} \frac{d(t_j)}{r(t_j)}$$

From this we can also estimate the survivor function $\hat{S}(t) = \exp(-\hat{\Lambda}(t))$.

The curvature of the Nelson-Aalen estimator gives an idea of the hazard rate curve. If it is straight, it implies a constant hazard.

In order to find the Kaplan-Meier survival function and the cumulative hazard function, we use the R Package `survival` with the command `survfit`. This gives the number at risk and number of events at each observed time, and provides point wise confidence intervals using Greenwood's formula.

Figure 4.2 presents the Nelson-Aalen cumulative hazard curves and the Kaplan-Meier curves for our SSc and non-SSc incident dataset. The SSc curves indicate a slight increase in slope from 7- to 10-years after diagnosis, indicating a slight increase in hazard over those years.

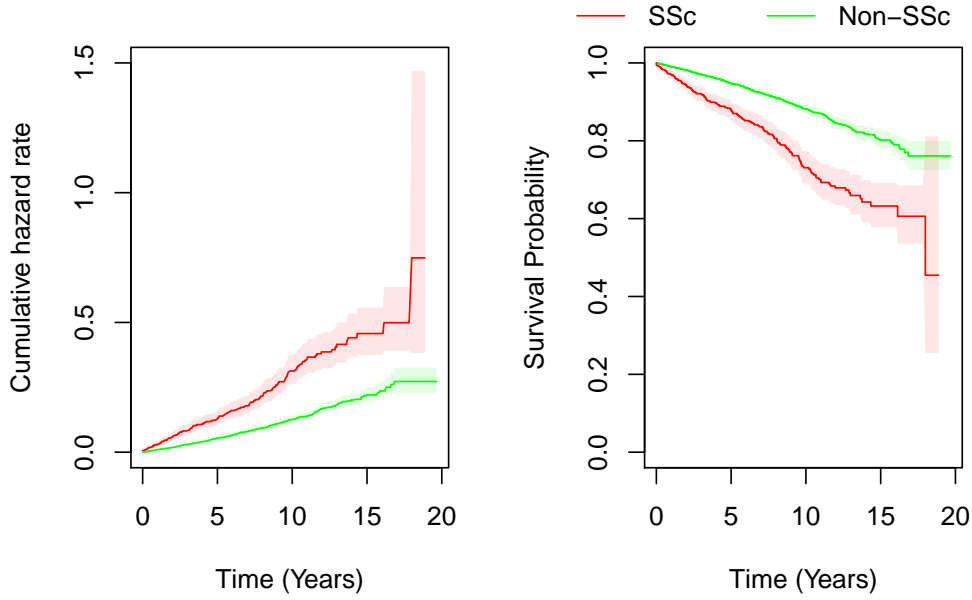


Figure 4.2: Nelson-Aalen cumulative hazard (left) and Kaplan-Meier survival probability (right), from SSc diagnosis (or match) to death, in the incident set only. The shaded areas represent the 95% confidence intervals.

The plot on the right of Figure 4.2 shows the survival curves by SSc status. Those with an SSc diagnosis have a lower survival than our comparators, and using a log-rank test this is significant with $\chi^2 = 91.4$ with 1 degree of freedom (p-value < 0.0001). SSc patients clearly have a worse mortality.

Table 4.5 summarises the survival at 5-, 10- and 15-years and the relative risk ratios at these times, $RR(t)$. The risk ratio is $RR(t) = F_{SSc}(t)/F_{nonSSc}(t)$. The confidence intervals are calculated as (Stegherr et al., 2021)

$$CI(t) (95\%) = RR(t) \exp \left(\pm 1.96 \sqrt{\left(\frac{se(F_{SSc}(t))}{F_{SSc}(t)} \right)^2 + \left(\frac{se(F_{nonSSc}(t))}{F_{nonSSc}(t)} \right)^2} \right)$$

The SSc group have a total of 5476.67 patient-years at risk and survival of 88% after 5 years and 73% after 10 years. The comparators have better survival than

the SSc patients, with the SSc 5-year survival rate being similar to that of the non-SSc 10 year survival rate. The risk ratios imply double the mortality in SSc patients.

Time from SSc diagnosis	Survival				Risk ratio	
	SSc (%)	[95% CI]	Non-SSc (%)	[95% CI]	RR(t)	[95% CI]
5 years	88.0	[85.5,90.4]	94.8	[94.1,95.5]	2.30	[1.81, 2.94]
10 years	73.1	[69.0,77.4]	88.1	[86.9,89.4]	2.27	[1.88, 2.73]
15 years	63.2	[57.8,69.2]	80.2	[78.0,82.4]	1.85	[1.53, 2.25]

Table 4.5: Summary table for survival after 5-year, 10-year, and 15-year for SSc and non-SSc patients. Survival is the percentage who have survived after t years from diagnosis. The risk ratio at these times is on the right of the table.

4.5 Proportional hazards model

We shall also use the proportional hazards model (also known as the Cox proportional hazards model, or ‘the Cox model’), a regression model commonly used for investigating the association between the survival time of patients and one or more predictor variables. The proportional hazards model allows us to simultaneously evaluate the effect of several factors on survival, and the influence these factors have on our event of interest.

The proportional hazards model defines the hazard given covariates as

$$\lambda(t) = \lambda_0(t) \times \exp(b_1 z_1 + b_2 z_2 + \dots + b_p z_p)$$

where t is the survival time, the coefficients (b_1, b_2, \dots, b_p) measure the impact of covariates (z_1, z_2, \dots, z_p) , and $\lambda_0(t)$ is the baseline hazard. The quantities $\exp(b_i)$ are the hazard ratios (HR), where a hazard ratio above 1 indicates a covariate that is positively associated with the event hazard, and is thus negatively associated with the length of survival. Conversely, a hazard ratio below 1 indicates a covariate that is negatively associated with the event probability, and is thus positively

associated with the length of survival. There are, however, three assumptions for appropriate use of the Cox proportional hazards regression model. These are:

- independence of survival times between distinct individuals in the sample,
- a multiplicative relationship between the predictors and the hazard, and
- a constant hazard ratio over time.

We assume the first to hold in our study. We shall use the Nelson-Aalen curves and the Schoenfeld residuals to verify the proportional hazards assumption: that when the predictor variables do not vary over time, the hazard ratio comparing any two observations is constant with respect to time.

We can use the function `coxph` in R's survival analysis package, `survival`, to calculate the hazard ratios.

For our study, the Cox model is used to fit the model of SSc status and covariates, giving hazard ratios and 95% confidence intervals. From previous knowledge, we would expect the risk of death to increase with age.

The covariates are:

- SSc status, whether they enter the study as an SSc patient or a match.
- The age at which the patient was diagnosed with SSc. From previous knowledge, we would expect the risk of death to increase with age. We report the hazard ratio as 'per 10 years' so indicating the increased/decreased hazard had someone been diagnosed with SSc 10 years later.
- Sex, with females as baseline.
- Geographical location, which refers to one of 13 larger administrative areas of England, Scotland and Wales, with the baseline being the area 'London'.
- BMI, compared against a baseline of the healthcare recommended range 18.5-24.9 (kg/m²).

- Alcohol, which has four categories: non-drinker, drinker, heavy drinker and ex-drinker, with drinker the most common, and therefore the baseline covariate.
- Smoking, which has three categories: non-smoker, smoker, and ex-smoker, with non-smoker as the baseline comparison.

Without the inclusion of covariates, other than SSc status, we have a hazard ratio (HR) of 2.35 [95% CI: 1.96, 2.82]. We then include the covariates and observe what effect this has on the hazard ratio of SSc status.

The breakdown of the proportional hazards model for the incident cohort is in Table 4.6. Note that if a patient has a covariate missing then that patient is not used in the model. Therefore, the left of the table (the full model) is based on 4,711 patients (see Missing Data in Section 3.3). The hazard ratio (which is for every 10 year increase in age of diagnosis) rises steeply from 2.35 to 3.06 when all covariates are included. Age at entry to the study has a large hazard ratio, indicating that a higher age at entry leads to an increased mortality. The hazard almost triples with each additional decade older a patient is when they enter the study. Males have a significantly increased hazard of death compared to females, with a hazard ratio of 1.58, and a confidence interval that does not include 1 [95% CI: 1.24, 2.01]. Having a BMI of higher than 35 is indicative of worse survival, as expected. Similarly, being a smoker or someone with smoking history increases risk. Being a heavy drinker increases the risk, as does being a non-drinker or ex-drinker. The higher risk in the non-drinker or ex-drinker may be due to advice from medical professionals if a patient has been asked to not drink for health reasons. However, there is no way to confirm this. None of the locations reveal a significant increased mortality risk compared to London.

However, this may not be the best fitting model. The Cox model deals with missing data by removing the patients with at least one missing data (listwise deletion). Therefore, we would not want to include a variable if it is unrelated to the outcome and its inclusion reduces the size of the dataset. We therefore

perform variable selection, to reduce the number of variables in the models in order to increase interpretability and model effectiveness. For this, we use a backwards stepwise model to minimise the Akaike information criterion (AIC) of nested models within the full model. It is recognised that there are many ways to test for model selection, such as the Wald test or LASSO. AIC was chosen for uniformity over this thesis, as AIC is used in other chapters. However, we note that as we are comparing nested models, a Likelihood Ratio Test would also have been appropriate. Backward selection was used (as opposed to forward or stepwise) as it is the simple method and we suspect more variables will stay in the model than be removed, hence a more efficient method.

In order to achieve a minimal AIC, location is removed, but the model supports the inclusion of the other factors. Therefore the reduced model is also estimated from 4,711 patients. The reduced model is shown on the right of Table 4.6. With the model reduced, the hazard ratio for SSc hardly changes, and there are no significant difference in any of the covariates.

As previously stated, the above table required the removal of patients. Therefore, we lastly note that the hazard ratio for SSc when age and sex are included (hence all 5,642 incident patients are included, while still allowing for 2 covariates) is 2.73 [95 % CI: 2.28, 3.27].

4.5.1 Test for proportional hazard assumption

As mentioned above, the Cox proportional hazard model requires that the assumption of proportionality is met (i.e., a constant hazard ratio over time), and that the explanatory variable only changes the chance of failure - not the timing of periods of high hazard.

We use the reduced model on the right of Figure 4.6 and test that the proportional hazards assumption holds. To do this, we test the proportional hazards assumption for each covariate, by checking the correlation of the corresponding set of scaled Schoenfeld residuals with time to check for independence for each

	Full model			Reduced model	
	HR	95% CI		HR	95% CI
SSc status (Non-SSc=0, SSc=1)	3.06	[2.49, 3.75]		3.07	[2.51, 3.76]
Age at SSc diagnosis (10 years)	2.79	[2.52, 3.08]		2.80	[2.54, 3.10]
Sex (Female=0, Male=1)	1.58	[1.24, 2.01]		1.65	[1.30, 2.08]
Location: East Midlands	1.00	[0.52, 1.92]			
Location: East of England	1.07	[0.63, 1.84]			
Location: North East	2.06	[0.96, 4.40]			
Location: North West	1.17	[0.73, 1.85]			
Location: Northern Ireland	0.80	[0.36, 1.77]			
Location: Scotland	1.33	[0.85, 2.08]			
Location: South Central	1.03	[0.64, 1.66]			
Location: South East Coast	1.08	[0.66, 1.77]			
Location: South West	1.00	[0.60, 1.67]	→		
Location: Wales	1.33	[0.84, 2.09]	After reduction		
Location: West Midlands	1.30	[0.81, 2.10]			
Location: Yorkshire & The Humber	1.06	[0.54, 2.05]			
BMI: <18.5	1.57	[0.97, 2.55]		1.51	[0.94, 2.43]
BMI: 25-29.9	0.92	[0.74, 1.14]		0.93	[0.74, 1.15]
BMI: 30-34.9	1.20	[0.89, 1.60]		1.20	[0.90, 1.60]
BMI: 35-39.9	1.55	[1.05, 2.30]		1.55	[1.05, 2.29]
BMI: 40+	2.84	[1.77, 4.56]		2.90	[1.81, 4.63]
Ex-smoker	1.46	[1.17, 1.81]		1.47	[1.18, 1.83]
Smoker	2.17	[1.71, 2.75]		2.25	[1.78, 2.85]
Ex-drinker	1.37	[1.01, 1.86]		1.37	[1.01, 1.86]
Heavy drinker	2.46	[1.53, 3.96]		2.58	[1.61, 4.13]
Non-drinker	1.51	[1.19, 1.91]		1.49	[1.18, 1.88]

Table 4.6: Cox model for the hazard ratios in the incident cohort. On the left are the results from the full model with all covariates, on the right are the results of the reduced model after covariates are removed due to lower AIC. Location has a baseline of London, BMI (kg/m^2) has baseline of a 'normal' health range, 18.5-24.9, the baseline for smoking is 'non-smoking' and the baseline for drinking is 'drinker'.

covariate and as a global test (Schoenfeld, 1982). Table 4.7 shows the results in our reduced model.

	chisq	df	p-value
SSc status	0.914	1	0.339
Age at SSc diagnosis (10 years)	4.433	1	0.035
Sex	0.121	1	0.727
BMI	2.078	5	0.838
Smoking	0.804	2	0.669
Alcohol	4.308	3	0.230
Global	13.349	13	0.421

Table 4.7: Chi-Squared Goodness-of-Fit test for proportional hazards assumption. The data was the reduced model which used only the incident cohort.

Age at entry to study is statistically significant at the 5% level, implying that the null hypothesis of the proportionality assumption may not be valid here. A graphical diagnostic using the function `ggcoxzph()` (in the `survminer` package) produces graphs of the scaled Schoenfeld residuals against the transformed time. We do this for age at SSc diagnosis (or age at study entry for the incident set), Figure 4.3. For the proportionality assumption to hold we would expect to see a flat line across time. However, we observe deviations as these results suggest that the residual increases with time.

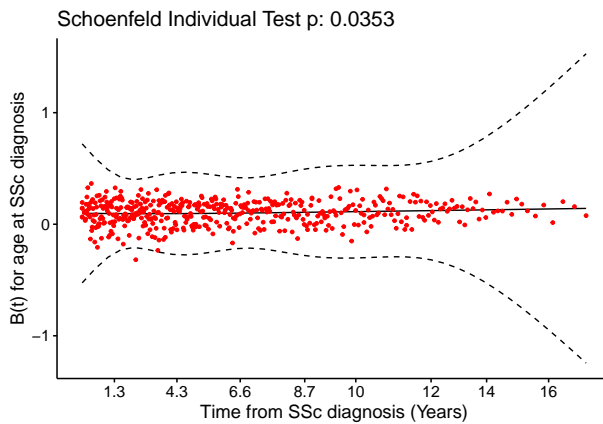


Figure 4.3: Scaled Schoenfeld residual plots for age at SSc diagnosis.

Methods to solve this issue include stratification of the variable or partitioning on the x-axis (subset the survival times, e.g. 0-5 years, 5-10 years, etc). However, we note that the 0.035 is close to the 5% level, and the residuals seem close to being constant over time. In addition, in the next chapter when the prevalent cohort is included we do not observe this issue with the residuals anymore. Therefore we will not take this any further here.

4.6 Summary

We use this chapter to demonstrate two of the most commonly used methods in survival analysis: the Kaplan-Meier and the Cox model. We also used the SMR due to its common use in SSc mortality studies.

Systemic sclerosis is known to have a higher risk of mortality, with the probability of survival beyond 20 years from diagnosis being under 50% in other literature (Rubio-Rivas et al., 2014). Our dataset agrees with this earlier research, that mortality is increased when compared to the general population. The SMRs imply more than double the mortality rate. If death is an endpoint which is often observed, and/or certain patients are at a higher risk of both the event of interest and the competing event, then it would be incorrect to treat death as a form of censoring, leading to the need for consideration of competing risks. However, we note that we are missing a subset of patients, namely those who had cancer prior to entry, therefore we would not propose that the results of this chapter are used (without this consideration) in the quantification of the mortality risk of SSc patients.

We note the short disease duration and follow-up times of our SSc set. Due to SSc being a long-lasting disease, it would be preferable to have a longer follow-up time, and the inclusion of prevalent cohorts will allow for a longer follow-up. The next chapter will focus on the inclusion of the prevalent cohort.

Chapter 5

Combining incident and prevalent cohorts to study SSc mortality

5.1 Introduction

As we have discussed in Chapter 1, there are two main methodology considerations in our study: the inclusion of prevalent patients and the analysis of competing risks. This chapter discusses the inclusion of prevalent cohorts.

Prevalent cohorts have sampling bias due to left truncation. As our study was originally designed to study the impact of SSc on cancer, to be included in our study our prevalent patients are required to be alive and cancer-free when they are first available to be recruited into the dataset. Hence, their starting event, or exposure, occurs prior to the study start time, and patients who have the outcome of interest prior to the study are not under observation. Note that data can also be right truncated¹, and studies that have both left and right truncation are called ‘doubly truncated’ studies. However, our data is an example of the

¹Right truncated: When patients are not included in the study if they do not have the event of interest prior to study end (or other random time, the ‘right truncation’ time).

most commonly observed in left truncation studies, which is that of being left truncated and right censored² (LTRC).

When including prevalent cohorts, there is known to be an oversampling of healthier patients, as these are patients who have survived long enough to be in the study. Individuals with longer survival times measured from the onset of the disease are more likely to be included in the cohort, whereas those with shorter survival times are more likely to be excluded. This issue is discussed in this chapter.

The inclusion of prevalent cohorts in this analysis is possible, and perhaps preferred, due to study design. A minimum study start date of 01/01/1998 and maximum study end date of 31/12/2017 only provides us with a maximum of 20 years for incident patients under study. If prevalent patients are included, subject to correcting for the characteristics of a prevalent cohort, we will have a larger sample size. This gives us more information and potentially longer duration of survival.

In this chapter, we demonstrate the inclusion of prevalent cohort patients under the key assumption of independence between a patient's left truncation time and event time. We shall apply nonparametric likelihood estimate (NPMLE) methods, and the Cox proportional hazards model for left truncated data. We shall use NPMLE methods to estimate the distribution of truncation times. The key assumption of independence made in this section will be addressed in later chapters.

As with in the previous chapter, we estimate mortality in SSc patients compared to non-SSc patients to introduce the theory of prevalent cohort inclusion without competing risk theory. We begin looking at cancer in the next chapter. Again, we are missing some patients due to the dataset being formed for the study of cancer rather than mortality. Therefore, we are missing patients who were diagnosed

²Right censored: When the event of interest is not observed over follow-up, and the end time recorded is the last known time of the patient being in the study. Not observing the event could be due to a patient withdrawing from or leaving the study, or the study end being prior to the event of interest.

with cancer prior to study entry. We therefore expect these mortality results to be an underestimation of risk.

This section provides the basis for discussion in subsequent chapters concerning competing risks. In each of these chapters we shall discuss how the methods are adapted when prevalent datasets are included. Following this, we shall revisit the assumption of independence between entry time and event time, and discuss adapting this chapter's methods in Chapter 9.

5.2 Prevalent cohort characteristics

The characteristics of the prevalent cohort are listed in Table 5.1, and are set against the characteristics of the incident cohort for comparison. As with the incident cohort, the majority of the patients in the prevalent cohort are women, with a higher percentage being female than in the incident set. As males have a higher mortality, this could be due in part to more males dying prior to entry, a notable feature of left truncated data. In both the SSc and non-SSc cohorts of the prevalent cohort dataset, a greater proportion of patients have a diagnosis of cancer prior to death, and a greater number also die during the study than in the incident cohort dataset. Both of these are most likely due to the advanced stage of SSc at which patients enter the study, due to the lag between SSc diagnosis and entry to study. Age at SSc diagnosis is either the age at which SSc patients were diagnosed, or the age at which non-SSc patients were matched. The average age at which incident SSc patients are diagnosed (58.1 years) is older than the average age of diagnosis in the prevalent cohort (48.1 years). This is logical, as in order to enter the study our prevalent patients have to be alive, therefore truncation time is correlated with younger ages of diagnosis (see Figure 3.10). There is no noticeable difference between incident and prevalent patients in the ages at which the patients die, however. The longer follow-up times here are an advantage of including prevalent patients. There is no large difference in the smoking habits between the incident and prevalent cohorts. There does seem to be a difference

in BMI for SSc patients, with incident patients having slightly higher BMI. As the BMI is recorded close to study entry, this could be because prevalent patients are further along in SSc progression, which may lead to weight-loss. There is no evidence of a difference in BMI for non-SSc patients between the incident and prevalent cohorts. Note that smoking and BMI percentages are only calculated using patients who have a data entry for that characteristic.

5.3 Notation and left truncation definition

Let patients be denoted by $i = 1, \dots, n$. Each patient has an SSc diagnosis date (or match date if they are non-SSc), an ‘entry to study’ date, and a death or censoring date. The time from diagnosis to study entry is denoted by L , and the time from SSc diagnosis to either death time, T , or censoring time, C , is $X = \min(T, C)$. We denote our censoring indicator by $\delta \in (0, 1)$ where $\delta_i = \mathbb{I}(T_i < C_i)$. Therefore, patient information provides the component triple $\{(L_i, X_i, \delta_i), i = 1, \dots, n\}$. We note some independence assumptions. Firstly we assume that for all patients the left truncation time, event time and censoring time are independent and identically distributed, respectively, to allow for the use of NPMLE methods (Breslow & Crowley, 1974). This is done for patients that are observed, i.e., we assume that $T \perp (L, C) | L \leq T$ and $L \perp C | L \leq T$, which are commonly made assumptions when using left truncated data (Tsai et al. (1987), M.-C. Wang (1991)). We relax these assumptions in later chapters.

Figure 5.1 shows hypothetical patients who are representative of our datasets when considering mortality.

- Patient 1 is an example of a patient who will not be included in our study of mortality. This example patient had cancer prior to 1998, and therefore did not meet the criteria for entry to the original study (which was looking at cancer). This is a potential flaw in this study of mortality, as while cancer could be thought of as censoring, if there is dependence between death time and cancer time we may be introducing informative censoring.

	Prevalent				Incident			
	SSc (n=780)		Non-SSc (n=4676)		SSc (n=806)		Non-SSc (n=4836)	
	Number	(% of SSc)	Number	(% of non-SSc)	Number	(% of SSc)	Number	(% of non-SSc)
Female	685	87.8	4106	87.8	667	82.8	4002	82.8
Male	95	12.2	570	12.2	139	17.2	834	17.2
Cancer diagnosed prior to death/censoring	116	14.9	564	12.1	90	11.2	470	9.7
Death during study	238	30.5	593	12.7	162	20.1	433	9.0
Smokers	166	21.5	981	21.8	161	20.0	1035	22.1
Ex-smokers	206	26.7	966	21.4	212	26.4	1184	25.2
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age at SSc diagnosis (years)	48.1	14.2	48.1	14.2	58.1	13.6	58.1	13.6
Age at death (years)	71.8	11.5	76.9	10.3	73.1	10.2	76.0	10.7
SSc diagnosis to death (years)	17.4	9.7	18.4	10.1	5.3	4.0	6.0	4.1
Follow-up (years)	18.5	9.9	19.0	10.2	6.8	4.6	7.1	4.7
BMI (kg/m ²)	24.5	4.7	26.6	5.6	25.8	5.2	27.0	5.6

Table 5.1: Cohort characteristics. For prevalent patients, BMI is based on 620 SSc patients and 3274 non-SSc patients. Also for prevalent patients, smoking is based on 771 SSc patients and 4506 non-SSc patients. For incident patients, smoking is based on 803 SSc patients and 4690 non-SSc patients, and for BMI is based on 615 SSc patients and 3286 non-SSc patients. Age at death is calculated from the subset of patients who died, and study entry to death (years) is the time between SSc diagnosis/match to death in those who died. Follow-up is time from study entry to death or censoring.

- Patient 2 is an example of an incident patient, whose time in the study is from the onset of SSc to their death.
- Patient 3 is another example of a patient who died before the start of the study, and is therefore not included in our study.
- Patient 4 is an example of a prevalent patient, who was included after the start date of the study. This patient developed cancer during the study, and this would have been the event of interest if we were studying cancer, but as we are only looking at mortality here this event is not important, only their death event is. This patient highlights that once a patient is in the study we do observe their death event, even if cancer diagnosis occurred prior to death.
- Patient 5 is an example of a prevalent patient who was diagnosed with SSc after 1998, representing patients who were diagnosed with SSc after the study start but were not recruited into CPRD GOLD until after 1998. Of all prevalent patients, 30% entered exactly on the study start date, 01/01/1998, and the others had entry dates after this. This category of patients results in us not having what is termed a cross-sectional sampling study. This is important as a number of studies make this assumption about their prevalent patients.
- Patient 6 is an example of a prevalent patient who entered the study in 1998 and did not have an observable event during the study and is therefore right-censored.

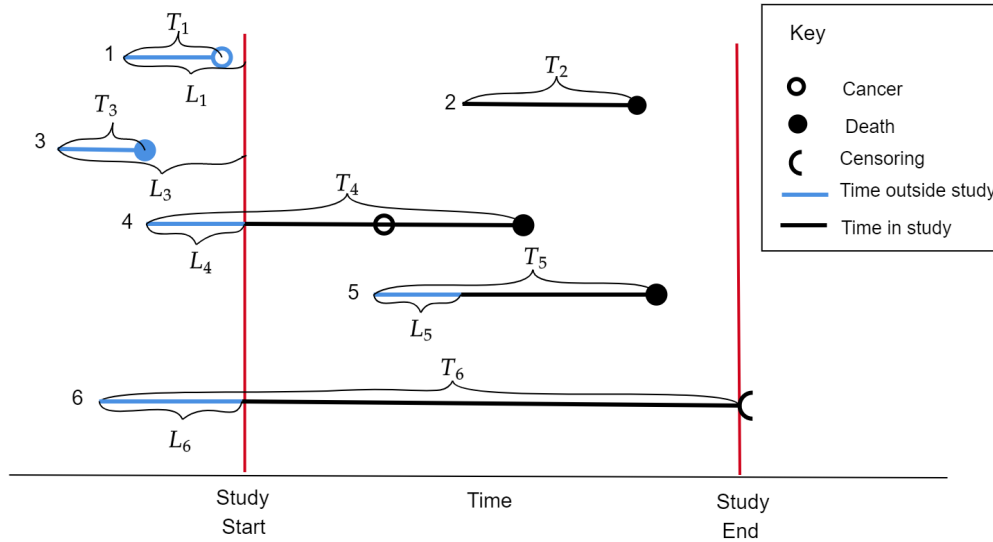


Figure 5.1: Example patients plotted over calendar time.

5.4 Nonparametric estimation of the survivor function

The most common method for estimating the survival function found in the literature is the nonparametric maximum likelihood estimator (NPMLE). The type of NPMLE we will use allows for left truncated and right censored data. It was first proposed by Lynden-Bell (1971), with the asymptotic properties studied by Woodroffe (1985) and by Tsai et al. (1987). It is termed the left truncated Kaplan-Meier in that it takes a very similar form to the full Kaplan-Meier, but only allows prevalent patients into the risk set once they are under study.

Let $S_T(t) = P(T > t)$ be the survivor function of T . We have ordered, distinct failure times $t_1, < t_2 < \dots < t_n$. If there is the assumption of independence between T and L , the nonparametric maximum likelihood estimator (NPMLE) of $S_T(t)$ is the product limit estimator

$$\hat{S}_T(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d(t_i)}{r(t_i)}\right)$$

where $d(t)$ is the number of subjects observed to fail at the observed failure time, and $r(t)$ is the number in the group under study at that time, $r(t) = \sum_{i=1}^n I(L_i < t \leq X_i)$, where n is the number of patients under study. We term this the **Left truncated Kaplan-Meier (LTKM)**.

To show how this affects prevalent patient inclusion, Figure 5.2 plots the example patients from Figure 5.1 with the x-axis being the time from diagnosis. Patients 1 and 3 are not included because their event occurred prior to the study. Patient 2 is included in the risk set from time 0, however the others do not join the risk set until they enter the study. Patients leave the risk set after an event or censoring.

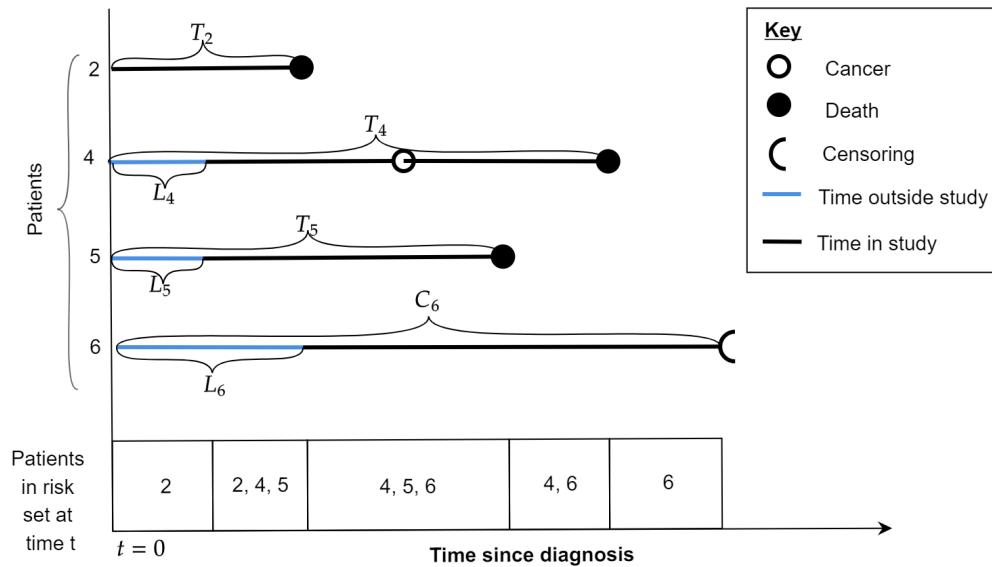


Figure 5.2: Example patients plotted from time since diagnosis to event, with the number in each risk set at time t .

We demonstrate the consequences of including prevalent patients from SSc onset, regardless of the time they entered the study. This is effectively ignoring left truncation, and demonstrates the bias of including patients who have longer

survival times. Figure 5.3 demonstrates this in our prevalent cohort (without the incident cohort), with the incident set from the previous chapter included as a comparison. We can see the cumulative hazard rate for the non-truncated adjusted Kaplan-Meier is substantially lower and their survival is an underestimation of risk. The x-axis is time from SSc diagnosis (which is time from hypothetical SSc date for non-SSc patients).

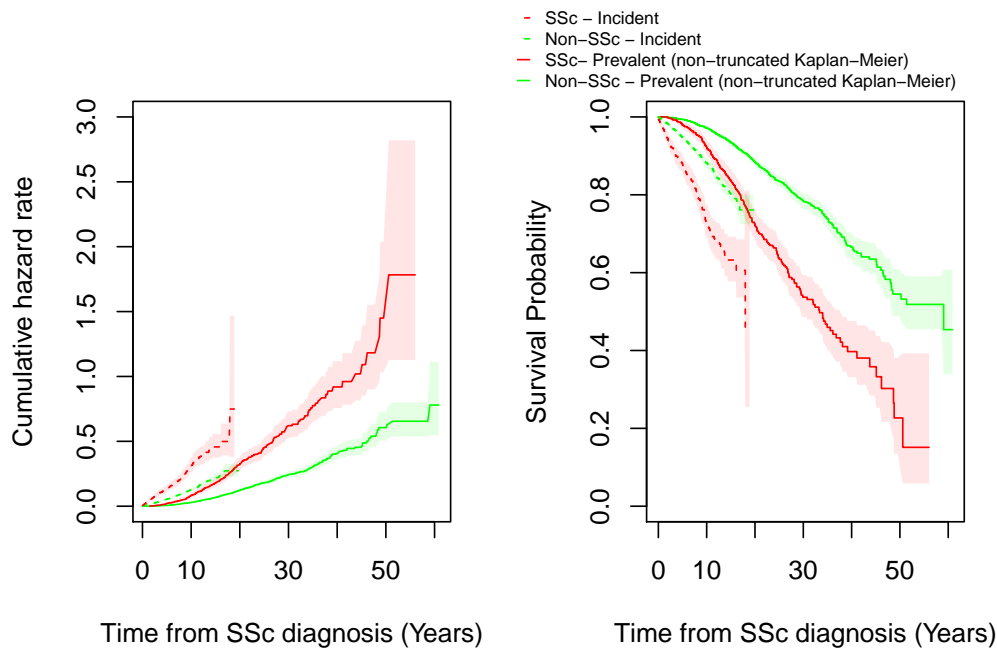


Figure 5.3: Cumulative hazard rate (left) and survival probability (right) when the risk set is from SSc diagnosis or SSc match diagnosis to event (time at entry not included in the nonparametric estimate), prevalent cohort only alongside incident cohort only, with 95% confidence intervals.

To account for this underestimation, we adjust the risk set as specified above, where the model will now only include patients in the risk set once they enter the study. Figure 5.4 demonstrates the corrected inclusion of prevalent patients using the truncated Kaplan-Meier.

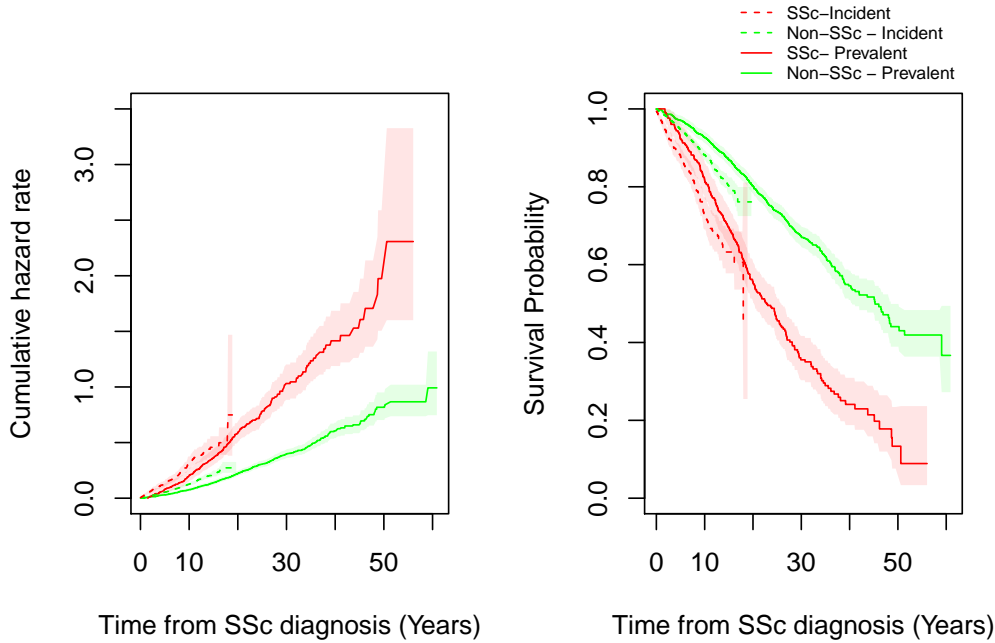


Figure 5.4: Cumulative hazard rate (left) and survival probability (right) when the patient is included in the risk set is from study entry to event, prevalent cohort only alongside incident cohort only, with 95% confidence intervals.

This modification prevents the underestimation of risk, as it now accounts for the delayed entry time. In Figure 5.4, there is a small difference between the incident and prevalent curves. Both the SSc and non-SSc prevalent patients appear to have a better survival than the incident sets at the times covered by the incident set. This appears counter-intuitive, as we may expect the prognosis of patients to get better over calendar time due to improved healthcare, as we expect someone with a more recent SSc diagnosis to have better care (we know prevalent patients tend to be alive through earlier calendar times, and there is a correlation between calendar time of diagnosis and left truncation time). This observation raises the question of whether the prevalent dataset is an accurate representation of the true risk in the recent population, and whether it should be used to guide future medical decisions. However, it could also be that prevalent patients who have survived long enough to make it into the study are a) biased to be healthier, or b) maybe have a weaker form of SSc. Also, the effect may arise from intrinsic

differences between the two datasets.

With methodology used in the previous chapter, and akin to the results shown in Table 4.5, in Table 5.2 we compare the risk ratios (or relative risk) for death of SSc compared with the non-SSc comparators at times 5, 10 and 15 years. We do this for the incident only cohort, the prevalent only cohort, and the combined incident and prevalent cohort (the prevalent only not plotted here).

Time from SSc diagnosis (Years)	Incident [95% CI]	Prevalent [95% CI]	Combined [95% CI]
5	SSc 0.120 [0.096, 0.145]	SSc 0.077 [0.038, 0.115]	SSc 0.119 [0.097, 0.141]
	Non-SSc 0.052 [0.045, 0.059]	Non-SSc 0.029 [0.017, 0.041]	Non-SSc 0.048 [0.042, 0.054]
	Ratio 2.304 [1.806, 2.939]	Ratio 2.662 [1.390, 5.096]	Ratio 2.502 [2.003, 3.126]
10	SSc 0.269 [0.226, 0.310]	SSc 0.185 [0.138, 0.229]	SSc 0.245 [0.215, 0.274]
	Non-SSc 0.119 [0.106, 0.131]	Non-SSc 0.073 [0.058, 0.088]	Non-SSc 0.103 [0.094, 0.111]
	Ratio 2.268 [1.882, 2.734]	Ratio 2.523 [1.835, 3.469]	Ratio 2.388 [2.058, 2.771]
15	SSc 0.368 [0.308, 0.422]	SSc 0.303 [0.252, 0.351]	SSc 0.354 [0.318, 0.388]
	Non-SSc 0.198 [0.176, 0.220]	Non-SSc 0.131 [0.114, 0.147]	Non-SSc 0.165 [0.154, 0.177]
	Ratio 1.854 [1.531, 2.246]	Ratio 2.320 [1.886, 2.854]	Ratio 2.139 [1.894, 2.415]

Table 5.2: Cumulative incidence of mortality for incident, prevalent and combined datasets at time points 5-, 10- and 15- years, also the ratio between them (SSc over non-SSc). The 95% confidence intervals are provided in brackets.

The risk ratio for those with SSc compared with the comparators appears to decrease over the 5 year time intervals, and this effect is observed in both the incident, prevalent and combined sets. The prevalent only risk ratios continue to imply that there is a large increase in mortality in those with SSc compared to the comparators, but have much wider confidence intervals, particularly at lower times, most likely a result of the small number in the risk set. However, the combined estimates have narrower confidence intervals compared to both the incident and the prevalent estimates due to more patients in the risk set, highlighting the benefit of including prevalent patients.

5.4.1 The Lai-Ying estimator

As first noted by Pan & Chappell (1999), one of the key concerns with using an NPMLE with only prevalent data is the small number of patients in the risk set close to the origin time. This is due to the definition of a prevalent cohort, as prevalent patients will all join after time 0. Therefore, the NPMLE can be very biased at early time-points due to the small number of people at risk, where there is a large jump in the estimated hazard at early time-points if an event is observed. This early bias will affect the rest of the survival curve.

Lai & Ying (1991) proposed a nonparametric estimator which adjusts the risk set in the presence of left truncation, by excluding hazards when there is only a small number of patients in the risk sets. The Lai-Ying estimator is defined as

$$\hat{S}_{LY}(t) = \prod_{i:t_i < t} \left(1 - \frac{d(t_i)}{r(t_i)} \mathbb{I}[r(t_i) \geq ck^\alpha] \right)$$

where k is the number of subjects in the cohort, and $c > 0$ and $0 < \alpha < 1$ are constants that can be chosen by the data analyst to prevent large hazards due to small numbers in the risk set.

We note here that this is not a concern for us when considering mortality, and nor will it be a concern in the next chapter when considering cancer. This is partially due to our robust prevalent dataset, as we have a large number in the risk set at the time of the first event. However, this would be a greater concern for smaller datasets in other studies. In consideration of our dataset, Figure 5.5 plots the number in the risk set at the time-points where there are events, e.g. if the first event is at time 0.5 years, then the plot will first show the number in the risk set at this time. The left of the figure is the incident dataset, and the right is the prevalent dataset. The left demonstrates the recruitment of patients at diagnosis (time 0) and then the decreasing number at risk due to patients leaving the risk set due to death or censoring. In the prevalent set, the time at the first death for the non-SSc patients is 0.91 years, when there are 319 at risk, and for the SSc

it is 1.78 years, when there are 106 at risk. These are large numbers, therefore this issue of a limited number in the risk set is not a concern for this particular study, but has been considered for completeness and may be relevant for other studies and different prevalent datasets. D. Wolfson et al. (2019) discusses the benefit of including incident patients alongside prevalent patients to prevent the large jumps in hazard.

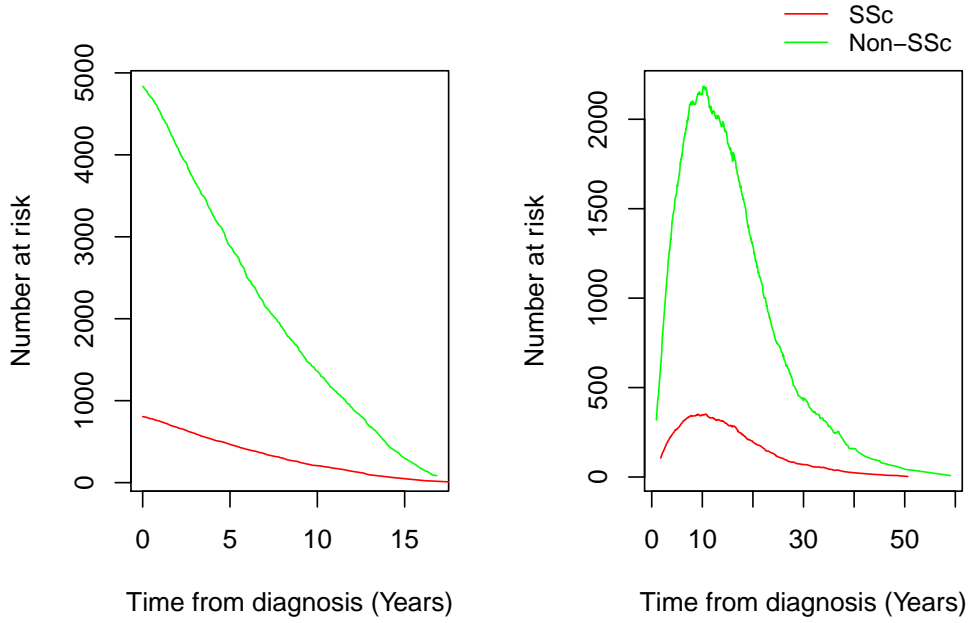


Figure 5.5: Number at risk in each set over time, Left - Incident, Right- Prevalent.

5.4.2 Left truncation distribution

Alongside the truncated Kaplan-Meier to estimate survival time, there is the Lynden-Bell estimator for the distribution of left truncation time, $G(\ell) = P(L \leq \ell)$. Let $\ell_{(1)} < \ell_{(2)} < \dots < \ell_{(n)}$ be the distinct truncation times of our prevalent patients. The distribution of truncation times, $G(\ell)$, is estimated by

$$\hat{G}_{LB}(\ell) = \prod_{k:\ell_{(k)} > \ell} \left(1 - \frac{\sum_{i=1}^n \mathbb{I}(L_i = \ell_{(k)})}{r(\ell_{(k)})} \right)$$

However, this is only accurate in the scenario of no right censoring. As censoring is common in our study (see Chapter 3), this is not an accurate method. We therefore refer to Wang’s work (M.-C. Wang, 1991), which, as part of a larger body of work, provides an NPMLE for the distribution in the presence of censoring,

$$\hat{G}(\ell) = \left(\sum_{i=1}^n \frac{1}{\hat{S}(L_i)} \right)^{-1} \sum_{i=1}^n \frac{\mathbb{I}(L_i \leq \ell)}{\hat{S}(L_i)} \quad (5.1)$$

where $S(L_i) = P(T_i > L_i)$, the probability of surviving until entry to the study. We calculate this on our dataset for the prevalent SSc and non-SSc cohorts, Figure 5.6. However, as stated by Qian & Betensky (2014), this holds on the assumption that $\Pr(L \leq C) = 1$, and not $\Pr(L \leq C) < 1$. In our study this may be slightly questionable, as patients may leave dataset by emigration from the UK or exiting the dataset. We hope to have a lack of patients who are lost to follow-up prior to the study, and would also anticipate that patients opting out of the dataset is non-informative, and therefore would not significantly impact the results.

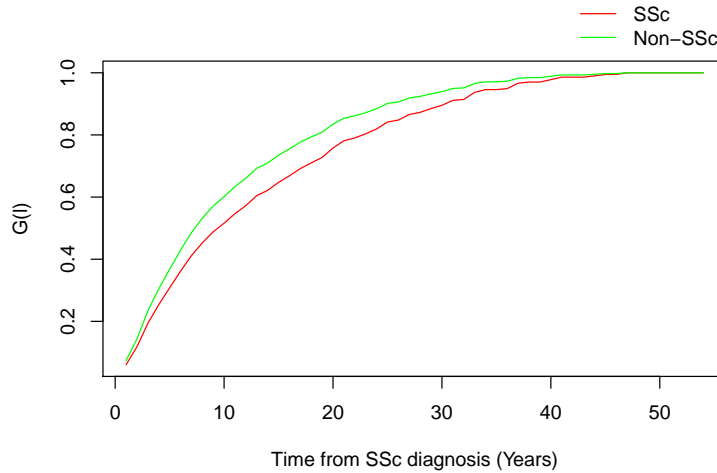


Figure 5.6: Estimation of left truncation distribution.

Note that with similar derivations to $\hat{G}(\ell)$ above, the probability distribution of event times, $F(t)$, can be estimated as an inverse probability weighted product likelihood estimator (PLE),

$$\hat{F}^{IPW}(t) = \left(\sum_{i=1}^n \frac{1}{\hat{G}(T_i)} \right)^{-1} \sum_{i=1}^n \frac{\mathbb{I}(T_i \leq t)}{\hat{G}(T_i)} \quad (5.2)$$

as mentioned by Shen (2003), where Equation (5.1) and Equation (5.2) are examples of Inverse Probability Weighting (IPW), with $\hat{F}^{IPW}(t)$ being an inverse probability of truncation weighted estimator. Often in studies, all observations are treated with equal importance. However, sometimes it is preferable to weight observations, for example because of sampling bias. If a subgroup of patients is less likely to be included in the study, these patients could have higher weightings to represent people missing in the study. Sometimes weighting is used when there is missing outcome data. For this, inverse probability weighting can be used with weights estimated from a logistic regression model for predicting non-response or censoring. Shen also demonstrated the equivalence of $1 - \hat{F}^{IPW}(t)$ to the left truncated Kaplan-Meier.

Returning to M.-C. Wang (1991), she notes that if the disease in question holds under the stationarity assumption then using the parametric form of $G(t)$ in the IPW can be more accurate than using the truncated Kaplan-Meier alone. The stationarity assumption is one where $G(t)$ coincides with a uniform distribution, or to phrase it another way, that disease onset (SSc) and its progression are independent of calendar time. If the stationarity assumption holds, then Vardi's estimate (Vardi, 1989) is a better estimate. If the stationarity assumption does not hold, but the underlying distribution of $G(t)$ is known, or can be established from the nonparametric estimation, then survival $S(t)$ can be modeled by a semiparametric model (M.-C. Wang, 1989). However, as can be seen from Figure 5.6, neither of these hold for our study.

Wang recommends checking the stationarity assumption visually by comparing $\hat{G}(t)$ to $U(a, b)$. In Wang's example, the study had a calendar recruitment start and end date of τ_1 and τ_2 respectively, and an enrolment date of τ , hence $a = \tau - \tau_2$ and $b = \tau - \tau_1$. The choices for a and b for us here are less clear than if we had a cross-sectional study with a fixed recruitment time, where we restrict our dataset

to those who were diagnosed with SSc between two time-points. As our study lacks these fixed recruitment time points, we may choose to make the approximation that $a = \min(L_i)$ and $b = \max(L_i)$. However, this would be very dependent on the largest truncation time, and may give this one data point too much importance and therefore bias the curve. If we wished to investigate this further, an extension of this study could be to subset the data to restrict recruitment so that SSc diagnoses fall within a calendar range, for example post-1980, for a better comparison. However, regardless of this, our curve does not have the straight CDF of a uniform distribution, and we therefore cannot assume a uniform distribution of SSc over calendar time and therefore will not be using Vardi's estimator. The curve implies that we disproportionately have more people with small truncation times, even when survival weights are taken into account. We hypothesise three categories of reasons why we may be seeing this curve (Figure 5.6):

- Incidence of SSc - Figure 5.6 might imply that the number of cases of SSc has been rising over the last 40 years. There could be a number of reasons for this.
 - Firstly, the population size of the UK has risen from 52.4 million in 1960 to 58.8 million in 1998 and to 66 million in 2017 (Randall, 2017), therefore an increase of SSc diagnoses over time in the UK would be expected, hence more with smaller left truncation times.
 - Secondly, due to improved healthcare, there could be more diagnoses due to better identification.
 - Thirdly, it could be that risk of developing SSc is rising, however this is unlikely as this has not been reported in literature (Rodríguez et al., 2019).
- Recruitment differences - These may arise due to the use of data in the original CPRD GOLD and in the extraction into this study. For example:
 - Early cases of SSc may not be recorded accurately or not at all. It could be that SSc was misdiagnosed, diagnosed too late, or entered into the

study on a delayed date from actual diagnosis. All these would result in fewer patients with long truncation times.

- Patients may also have been recorded under the diagnosis code for scleroderma, the old umbrella term for the disease, and hence not appear in our dataset. Recruitment differences such as these could give rise to the appearance of more cases over calendar time, but this cannot be verified from the data alone.
- Some patients were not included if there was doubt about their suitability for the study. This could bias the study away from earlier and possibly less reliable data. However, the number of these cases should be very small.
- Changes in survival over time - It could be that patients have a better survival depending on the period of calendar time they were alive for, for example due to medical and healthcare improvements we may expect more recent patients to have longer survival. Therefore, less patients with long truncation times will enter the study, and therefore there is the appearance of fewer patients making it into the study than expected from the weightings at higher truncation times (see simulation 3 below). We can adjust for this by weighting the survival dependent on their entry time, to account for dependent left truncation. We discuss this further in Chapter 9.

Appendix A.1 has simulations demonstrating how these three reasons may affect the truncation curve. Any of the above three reasons could lead to the curve we see in Figure 5.6. Another point could be that we deliberately did not recruit patients with cancer prior to entry, which may lead to bias.

5.5 Proportional hazards model

In a similar way to the NPMLE modification, we can alter the risk set used to calculate the partial likelihood so that patients are only included when they

are at risk. With the hazard function given covariates z specified by $\lambda(t|z) = \lambda_0(t) \exp(\beta z_i)$, then the partial likelihood, as specified by Breslow (1974), is

$$L(\beta) = \prod_{i=1}^N \left(\frac{\exp(\beta z_i)}{\sum_{j \in r(t_i)} \exp(\beta z_j)} \right)^{\delta_i}$$

where for prevalent cohort inclusion, the risk set is $r_i(t) = \sum_{j=1}^n \mathbb{I}(L_j < t \leq X_j)$. From this we can derive the Breslow estimator for the baseline cumulative hazard function, where the Breslow estimator is given by

$$\hat{\Lambda}_0(t) = \sum_{t_i \leq t} \hat{\lambda}_0(t_i) = \sum_{t_i} \frac{1}{\sum_{j \in r_i} \exp(\hat{\beta} x_j)}$$

We perform a similar analysis to that described in the last chapter, but now for the combined dataset, including both incident and prevalent patients. With just SSc type in the proportional hazards model, having SSc has a hazard ratio of 2.47 [95% CI: 2.20, 2.77], which is a small increase compared with the incident set only (HR 2.35 [95% CI: 1.96, 2.82]). We again use the full model with all covariates, as shown on the left of Table 5.3, then the reduced table without location (due to backwards step AIC minimisation, see Chapter 4) on the right. Due to missing covariates, when the model features BMI, alcohol and smoking status, the number of patients used in the model is reduced from 11,098 patients to 8,951 patients. Again, the AIC is used to check if covariates with missing data should be included or not, however the risk would be that we are removing variables which benefit the model but due to insufficient sample size are not statistically significant. However, only location removal made the AIC smaller, hence we used a reduced number of patients.

Again, on the right of Table 5.3, there are no significant changes to the hazard ratios when the additional prevalent patients are added, other than perhaps sex, where the hazard for males is reduced from 1.65 in the reduced incident model (Table 4.3) to 1.52 in the reduced combined model, and BMI, where the confidence interval for a BMI of 35-39.9 is no longer significant. When the prevalent cohort

	Full model			Reduced model	
	HR	95% CI		HR	95% CI
SSc status (Non-SSc=0, SSc=1)	3.13	[2.73, 3.59]		3.16	[2.75, 3.62]
Age at SSc diagnosis (10 years)	2.72	[2.55, 2.90]		2.71	[2.54, 2.88]
Sex (Female=0, Male=1)	1.53	[1.29, 1.81]		1.52	[1.28, 1.79]
Location: East Midlands	0.73	[0.46, 1.15]			
Location: East of England	1.02	[0.74, 1.41]			
Location: North East	1.27	[0.79, 2.06]			
Location: North West	0.95	[0.71, 1.27]			
Location: Northern Ireland	0.84	[0.56, 1.26]			
Location: Scotland	1.08	[0.81, 1.43]			
Location: South Central	0.88	[0.65, 1.18]			
Location: South East Coast	0.82	[0.60, 1.14]			
Location: South West	0.89	[0.65, 1.21]	————→		
Location: Wales	0.97	[0.72, 1.30]	After		
Location: West Midlands	0.89	[0.65, 1.20]	reduction		
Location: Yorkshire & The Humber	0.70	[0.45, 1.09]			
BMI: <18.5	1.45	[1.06, 1.99]		1.44	[1.06, 1.97]
BMI: 25-29.9	0.92	[0.80, 1.06]		0.92	[0.80, 1.07]
BMI: 30-34.9	1.16	[0.96, 1.40]		1.16	[0.97, 1.42]
BMI: 35-39.9	1.31	[0.97, 1.76]		1.32	[0.98, 1.78]
BMI: 40+	2.33	[1.67, 3.27]		2.39	[1.71, 3.35]
Ex-smoker	1.31	[1.13, 1.51]		1.31	[1.14, 1.52]
Smoker	2.30	[1.97, 2.68]		2.35	[2.01, 2.73]
Ex-drinker	1.50	[1.19, 1.87]		1.51	[1.21, 1.89]
Heavy drinker	2.61	[1.88, 3.64]		2.65	[1.91, 3.69]
Non-drinker	1.40	[1.21, 1.63]		1.40	[1.20, 1.62]

Table 5.3: Cox model for hazard ratios of mortality, using the combined incident and prevalent dataset. On the left are the results from the full model with all covariates, on the right are the results of the reduced model after covariates are removed due to lower AIC. BMI (kg/m^2) has a baseline of a 'normal' health range, 18.5-25, the baseline for smoking is 'non-smoking' and the baseline for drinking is 'drinker'. The grey shading shows significant values at the 95% confidence level.

is added, the hazard ratio for SSc type increases from 3.07 [95% CI: 2.51, 3.76] to 3.16 [95% CI: 2.75, 3.62]. The size of the confidence intervals have reduced due to the greater number of patients in the dataset, which is encouraging. Checking the Schoenfeld residuals, Table 5.4, age at SSc diagnosis no longer appears to fail the proportional hazard assumption.

	chisq	df	p-value
SSc status	0.162	1	0.69
Age at SSc diagnosis	0.305	1	0.58
Sex	9.26e-05	1	0.99
BMI	1.630	5	0.90
Smoking	2.720	2	0.93
Alcohol	2.719	3	0.26
Global	5.85	13	0.95

Table 5.4: Chi-Squared test for the proportional hazards assumption for mortality.

There is a small trade-off with the inclusion of some covariates due to missing data which may lead to selection bias. Therefore, in order to utilise all patients, we produce a Cox model solely looking at the covariates: SSc status, sex and age at diagnosis/match (Table 5.5). This reduces the hazard ratio for SSc status compared with when BMI, smoking and alcohol intake are included. The hazard ratio is then smaller.

	HR	95% CI
SSc status (Non-SSc=0, SSc=1)	2.88	[2.56, 3.23]
Age at SSc diagnosis (10 years)	2.61	[2.47, 2.74]
Sex (Female=0, Male=1)	1.41	[1.23, 1.62]

Table 5.5: Cox regression model for reduced covariates, combined dataset.

5.6 Possible pitfalls of prevalent inclusion

While including prevalent cohorts can give us information about survival and improve the power of our analysis, inclusion is only relevant if the prevalent cohort is representative of the current distribution of risk. There is a concern that past cases may not be representative of the risk today, and their inclusion in our analysis could bias results.

To demonstrate what we mean by this, we produce simulations. We demonstrate three pitfalls that may arise via simulations, and compare the difference in survival estimates between an incident cohort and a prevalent cohort. The simulations are performed on 10,000 patients for 1,000 iterations, although as we are working with left truncation only a small portion will be used in calculations. We shall fit a Kaplan-Meier curve and quantify the difference in survival for the incident and prevalent cohorts. Patients are born uniformly between 1800 and 2100. SSc diagnosis is simulated with distribution $N \sim (50, 10)$, to broadly mimic an average SSc onset age. Let T be time to event (death) from SSc diagnosis where time to event will change depending on the situation under study. The cohorts are taken from the simulated patients, with patients being in the incident cohort if their SSc is diagnosed after 2000, and in the prevalent cohort if they are diagnosed prior to 2000 but are still alive at 2000. We report the mean bias (the probability of surviving to time t in the SSc group minus the probability of surviving to time t in the non-SSc group), and the 95% percentile confidence interval.

- Simulation 1: Covariate selection bias
 - As we have seen from the cohort information, when we recruit patients we may be selecting patients on their likelihood of entering the study, as opposed to what we might expect in the general population. For example, recruiting prevalent patients may bias the data towards patients with lower age at entry. Therefore, we simulate patients such that older patients are at a higher risk of SSc. If a patient develops

SSc after age 50, their risk is $T_{SSc>50} \sim Exp(0.2)$, but prior to age 50 it is $T_{SSc<50} \sim Exp(0.1)$. As can be seen in Table 5.6, there is no bias reported in this situation for the NPML method. However, in a future chapter (Chapter 8), we shall consider a g-computation method where this oversampling of younger patients may need to be considered due to the need to create a pseudo-population.

- Simulation 2: Decreased/increased hazard over calendar time.
 - This would occur if the risk of the event was not constant over calendar time. There have, for example, been reports that SSc has had reduced mortality over the last 40 years due to improved understanding and treatment of SSc (Steen & Medsger, 2007), (Butt et al., 2018). Therefore the inclusion of patients who no longer represent our current population may lead to bias. We simulate patients such that the risk is higher prior to a particular calendar time. If a patient develops SSc prior to 1950, their risk is $T_{Diag<1995} \sim Exp(0.2)$, but after this it is $T_{Diag>1995} \sim Exp(0.1)$. There is a bias in this situation, with the prevalent set underestimating the survival of what we may consider the ‘general population’, the incident set. This is the disadvantage of including patients who may not accurately portray the current risk. This difference in risk is difficult to predict, however we shall discuss this more in Chapter 9.
- Simulation 3: Differing covariate risks which leads to differing hazards.
 - It may be that the risk of the disease itself is not changing, but due to changing demographics we may be recruiting prevalent patients who are uncharacteristic of the current population. An example could be smoking, in that smoking trends have decreased in the UK, and smoking is associated with increased mortality. By including prevalent cohorts, we could be including more patients who smoke. This is comparable to the above situation (hazard over calendar

time), as we are suggesting that a prevalent population will have a different overall risk compared to the incident, but it is the distinction between (a) a change in the risk of SSc type and (b) a change in covariate demographic. We simulate patients such that patients before a certain calendar time are more at risk. Patients have a 75% chance of being a smoker if they are born prior to 1950, and 25% if they are born after 1990. Smokers have double the risk, $T_{Smoker} \sim Exp(0.2)$, whereas non-smokers have risk according to the risk distribution $T_{Non-smoker} \sim Exp(0.1)$. The bias is greater than zero, with the prevalent set underestimating survival.

Simulation 1 is specific to left truncated data, however Simulation 2 and Simulation 3 are illustrating changes over calendar time, and while there is a correlation between prevalent and more historic cases, this is not specific to prevalent patients. Also, we note that these are only three simple simulations, and further simulations with different variables/constraints could produce differing results. However, these simple simulations suggest how prevalent inclusion may not lead to accurate predictors of current risk and therefore indicate that such situations will need to be adjusted for. Selection bias (simulation 1) is not a concern at present, but could become so when using the g-formula estimator (Chapter 8). Changes in calendar time will usually be difficult to verify and adjust for, as this information is often harder to observe due to the nature of prevalent cohorts, however we shall investigate how much accounting for temporal trends can benefit the analysis (Chapter 9).

5.7 Summary

We have observed a large difference in the risk ratios with the inclusion of the prevalent cohorts, however less of a change between the hazard ratios. As the hazard ratio is taking covariates into account, it may be that there is a difference due to covariate differences between the incident and prevalent cohorts. From our

	Time 10	Time 20	Time 30
	Bias [95% CI]	Bias [95% CI]	Bias [95% CI]
Sim 1: Selection Bias	1.44e-04 [-2.2e-02, 2.39e-02]	6.4e-05 [-4.68e-03, 4.45e-03]	2.18e-05 [-1.23e-03 ,1.16e-03]
Sim 2: Risk changes of SSc over calendar time	3.1e-02 [-1.31e-03, 6.68e-02]	4.32e-02 [2.92e-02, 5.69e-02]	2.33e-02 [1.71e-02, 2.92e-02]
Sim 3: Covariate change over calendar time	5.83e-02 [4.14e-02, 8.1e-02]	2.92e-02 [2.38e-02, 3.54e-02]	1.18e-02 [9.17e-03, 1.43e-02]

Table 5.6: Possible 'bias' from including prevalent cohorts for three simulations. Time is 10, 20 and 30 years after diagnosis. The grey cells are when the bias is significant at the 95% level (percentile confidence intervals).

analysis, we maintain the previous chapter's conclusion that SSc increases the risk of death approximately 3-fold.

The benefits of including prevalent patients are clear, as we have extended observation times and obtained narrower confidence intervals due to a larger sample size. However, these models do not allow for certain features of the dataset and have drawbacks. Firstly, the simple form of these models does not allow for competing risks. While we do not need to consider competing risks in this analysis of the dataset because death is recorded even after a cancer diagnosis, this is not the case when the event under observation is cancer, which has death as a competing risk. This issue is discussed further in the next chapter.

Secondly, the use of the left truncated Kaplan-Meier estimator assumes that there is independence between the left truncation time and the event/censoring time. We shall discuss how these methods can be adapted to allow for conditional independent truncation later, in Chapter 9. Also, we will consider methods which better accommodate covariates. Patients who were diagnosed much earlier than their incident counterparts will have different exposure covariates in life, as well as different healthcare treatments. While we hope that matching adjusts for much of the confounding, we will need to take this into account when considering studies that do not have matching, and it may still be worthwhile for our non-matched covariates.

In conclusion, while the left truncated Kaplan-Meier is a quick and viable method, we wish to understand these methods under a competing risk framework in order to study cancer, which we will investigate in the next chapter.

Chapter 6

Cancer analysis with competing risks

6.1 Introduction

We have previously noted the need for studies of the type presented in this chapter, to consider competing risks. Death is an event commonly observed in our dataset and which may occur prior to cancer onset or diagnosis, and, as shown in the last chapter, this may occur even more for SSc patients. We must define what we mean when we say ‘risk’ of cancer.

In one case, if cancer was the only event (death is treated as right censoring), we would see the probability of the event tend to one over time from diagnosis. But we know many SSc patients will never have cancer in their lifetime, as death precedes any cancer diagnosis. It is possible that the SSc patient’s mortality is so much greater than that of a typical member of the general population that their risk of cancer development relative to the general population is not significant in comparison, but treating death as censoring would not identify this. In the other case, if we were to treat death as its own event, this interpretation might disguise the causal relationship between SSc and cancer then this may lead to unfounded concern in the SSc group, a lack of concern when there should be, or

a poor allocation of resources if there is misrepresentation.

In this chapter we calculate the SIRs for cancer for our dataset, as this is the most common calculation found in literature when assessing cancer risk in those with SSc. We then provide as background a summary of methods of analysis of competing risks, and discuss how they can be used to find the cause-specific cumulative incidence function: the probability of observing a type of event over the time period under study. We shall apply these methods to our own dataset for the incident cohort, and from there we shall demonstrate the application of different methods to include left truncated data. We shall implement the Fine and Gray model for our dataset, and highlight the different interpretations between this and the Cox model.

There is an issue in the analysis of this chapter, which is informative censoring¹. If we were to treat death as a form of censoring when estimating survival, we may incorrectly establish a biased risk of cancer in those with SSc by treating mortality as a form of non-informative censoring. This leads to the theory used in the next chapter.

It should be noted that this chapter will estimate the risk of SSc on cancer, and this will also be the case in future chapters. From this point in the thesis any further analysis of death is in the terms of ‘competing risk’, where death is the alternative event to cancer.

6.2 Cancer standardized incidence ratios

In Chapter 4, the SSc cohort’s mortality was compared to mortality in the UK general population using SMRs. Many studies investigating the link between SSc and cancer take their SSc population from secondary healthcare databases. This may be, for example, a database which contains only SSc patients and is usually done for ease and accuracy of finding SSc patients. Such datasets inevitably lack

¹Non-informative censoring: when each subject has a censoring time that is statistically independent of their failure time, alternatively defined as the distribution of survival times (T), provides no information about the distribution of censorship times (C).

comparator non-SSc patients and hence it is necessary for such studies to use SIRs to compare their SSc cancer rates to those expected in the general population. Therefore, we shall do the same, and will compare the number of observed cancers in our dataset with those expected in the general UK population. The expected rates are from Cancer Research UK (CRUK) (CancerResearchUK, 2020), with comparison data risks estimated from events between the years 2015-2017. Table 6.1 shows the breakdown of the calculations for the SIRs for SSc patients, Table 6.2 shows the breakdown of the calculations for the SIRs for non-SSc patients, and Table 6.3 is the summary table including confidence intervals, estimated using the theory from Chapter 4.3. Please note that all three tables show results for the incident cohort only.

From Table 6.3, it can be seen that there appears to be a higher number of observed cancer diagnoses than expected from the UK population data for both the SSc and the non-SSc groups, and from the confidence intervals this is significant at the 95% level. SSc patients appear to have higher SIRs than the non-SSc patients. Why non-SSc patients have significantly higher SIRs in our dataset than the general population is not clear, but it may be due to the assumptions we are making for the calculations. For example, it is possible that cancer diagnosis and/or cancer rates changed over the time period our patients were observed. However, we are only using rates from 2015-2017 so this would appear to be unlikely to have had a significant effect. Male SSc patients in particular have a high SIR. This is consistent with literature, see Section 2.2. However, while the confidence intervals do not include 1, they are wide for the SSc group.

While these findings are interesting, this method for comparing cancer risk is not optimal compared to the other methods we shall focus on and is therefore not crucial to this study. SIRs involve rates of cancer incidence for the general population, usually taken from one year or a small range of years (for us, 2015-2017), as opposed to the 20 years under study, which is a poor approximation. Our dataset benefits from matched comparators, which means

Age group	Female					Male					Total	
	Person years (n)	Number in group	Number of events (d)	CRUK rate (r)	Expected events (n * r)	Person years (n)	Number in group	Number of events (d)	CRUK rate (r)	Expected events (n * r)	Total Observed	Total Expected
15-19	1.5	1	0	0.00021	0	1.5	1	0	0.00020	0	0	0
20-24	20.1	8	0	0.00035	0.01	6.4	2	0	0.00030	0	0	0.01
25-29	50.0	18	0	0.00074	0.04	12.6	7	0	0.00047	0.01	0	0.05
30-34	81.0	29	0	0.00114	0.09	22.8	8	0	0.00067	0.02	0	0.11
35-39	140.5	57	1	0.00169	0.24	25.1	9	0	0.00086	0.02	1	0.26
40-44	239.6	83	2	0.00260	0.62	33.4	11	0	0.00122	0.04	2	0.66
45-49	306.2	109	3	0.00410	1.26	68.7	24	0	0.00213	0.15	3	1.41
50-54	442.3	152	2	0.00576	2.55	132.4	42	2	0.00382	0.51	4	3.06
55-59	592.6	191	7	0.00732	4.34	123.4	47	3	0.0069	0.85	10	5.19
60-64	695.8	233	13	0.00980	6.82	127.0	50	4	0.01146	1.46	17	8.28
65-69	648.3	224	14	0.01290	8.36	107.6	41	3	0.01741	1.87	17	10.23
70-74	507.9	187	9	0.01510	7.67	62.7	26	4	0.02294	1.44	13	9.11
75-79	344.8	121	11	0.01836	6.33	50.9	16	1	0.02892	1.47	12	7.80
80-84	175.5	75	4	0.02085	3.66	17.5	10	2	0.03136	0.55	6	4.21
85-89	57.9	24	4	0.02233	1.29	2.0	2	0	0.03448	0.07	4	1.36
90+	19.0	9	1	0.02027	0.39	0	0	0	0.03297	0	1	0.39
			$O_F=71$		$E_F=43.7$		$O_M=19$		$E_M=8.4$	$O_T=90$	$E_T=52.103$	

Table 6.1: Cancer SIRs in SSc patients stratified by age and sex, compared with the general population (CRUK).

Age group	Female					Male					Total	
	Person years (n)	Number in group	Number of events (d)	CRUK rate (r)	Expected events (n * r)	Person years (n)	Number in group	Number of events (d)	CRUK rate (r)	Expected events (n * r)	Total Observed	Total Expected
15-19	8.7	6	0	0.00021	0	9.1	6	0	0.00020	0	0	0
20-24	98.8	47	0	0.00035	0.03	29.9	12	0	0.00030	0.01	0	0.04
25-29	256.9	104	1	0.00074	0.19	86.8	40	0	0.00047	0.04	1	0.23
30-34	438.8	168	1	0.00114	0.50	157.7	54	0	0.00067	0.11	1	0.61
35-39	779.0	330	2	0.00169	1.32	134.7	54	0	0.00086	0.12	2	1.44
40-44	1328.3	481	7	0.00260	3.45	194.9	63	0	0.00122	0.24	7	3.69
45-49	1807.0	637	11	0.00410	7.41	381.5	145	1	0.00213	0.81	12	8.22
50-54	2624.6	907	21	0.00576	15.13	765.5	242	4	0.00382	2.92	25	18.05
55-59	3458.8	1141	35	0.00732	25.33	772.4	282	7	0.00690	5.33	42	30.66
60-64	3994.5	1392	52	0.00980	39.13	950.3	316	10	0.01146	10.89	62	50.02
65-69	4104.6	1331	62	0.01290	52.94	787.7	281	18	0.01741	13.71	80	66.65
70-74	3382.5	1185	66	0.01510	51.08	499.3	190	20	0.02294	11.45	86	62.53
75-79	2469.6	805	70	0.01836	45.35	326.6	111	10	0.02892	9.45	80	54.80
80-84	1572.5	551	33	0.02085	32.78	208.9	68	8	0.03136	6.55	41	39.33
85-89	582.0	220	24	0.02233	13.00	81.9	36	5	0.03448	2.82	29	15.82
90+	135.7	55	2	0.02027	2.75	12.8	7	0	0.03297	0.42	2	3.17
			$O_F=387$		$E_F=290.4$		$O_M=83$		$E_M=64.9$	$O_T=470$	$E_T=355.272$	

Table 6.2: Cancer SIRs in non-SSc patients stratified by age and sex, compared with the general population (CRUK).

	SSc		Non-SSc	
	SIR	95% CI	SIR	95% CI
Female	1.63	[1.25, 2.00]	1.33	[1.20, 1.47]
Male	2.25	[1.24, 3.26]	1.28	[1.00, 1.55]
Total	1.73	[1.37, 2.08]	1.32	[1.20, 1.44]

Table 6.3: SIR for SSc and non-SSc patients with 95% confidence intervals. Here, total is our whole dataset, which is 83% female.

that the baseline covariates of our comparators should be the same as those for the SSc patients, allowing for better comparison due to the exchangeability between patients.

6.3 Competing risk theory

Epidemiological studies frequently investigate the impact of an exposure (or treatment) on an outcome of interest, but often this outcome is not the only event that can occur. Sometimes these events obscure or actually prevent the event of interest taking place. In our study, we are concerned about mortality, potentially obscuring the risk of cancer.

A naive approach would be to treat death without cancer as non-informative censoring, i.e. that censoring occurs independently of the risk of the outcome of interest (‘naive’ is the phrase coined by Andersen et al. (2012b)). However, ignoring the risk of death will lead to an upward bias in incidence by not changing the survival probability to reflect the loss of patients to other risks. We do not expect 100% of patients to have cancer. Therefore, there is a need to alter the cumulative incidence function (CIF) so that a competing event contributes to the survival probability.

There are three different hazards in competing risk to consider, namely marginal hazard, cause-specific hazard and subdistribution hazard. The marginal hazard is the hazard under the assumption that other events are considered to be a form

of censoring (i.e. they do not exist). This is dismissed in competing risk literature (‘naive’ (Andersen et al., 2012b) or ‘crude’ (Austin et al., 2016)), on the grounds that the interpretation of risk is not possible due to the nature of competing events. However, treating death as a censoring event does have more impact in a causal framework (interpreted as the direct effect of SSc on cancer). This is because incorporating competing events when calculating risk may obscure the relationship between an exposure and the event of interest. The Kaplan-Meier will treat other events as censoring, hence hypothetically removing the interaction of other events. However, this becomes an issue if there is dependent/informative censoring, which the standard Kaplan-Meier cannot account for. We shall cover this interpretation and give a more structured framework in the next chapter, but for now we must assume non-informative censoring.

As we will see below, the cause-specific hazard quantifies the rate of the event given that the competing event has not yet occurred. However, the cause-specific hazard means there is no longer a one-to-one correspondence between the effect of a covariate on the cause-specific hazard and the effect on the cumulative scale, therefore the subdistribution hazard was developed. The subdistribution hazard differs from the cause-specific hazard in that even if a subject experiences a competing event, they will remain in the risk set. This allows for the development of the Fine and Gray model.

6.3.1 Cumulative incidence function

We say that there are K types of competing events. Let (T, E) denote the event time and event type $E \in \{1, \dots, K\}$. Without competing risks, we define $F(t) = P(t \leq T) = 1 - S(t)$. When competing risks are present, we define the cumulative incidence function, CIF, for event type k as

$$F_k(t) = P(T \leq t, E = k)$$

Note that on the assumption E is inclusive of all potential events,

$$\lim_{t \rightarrow \infty} \sum_{k=1}^K F_k(t) = 1$$

The overall survival is the probability of not having an event up until t ,

$$S(t) = 1 - \sum_{k=1}^K F_k(t)$$

In order to approximate the cumulative incidence function, we require the hazard rate.

6.3.2 Cause-specific hazard rate

With the cause-specific cumulative incidence defined as $F_k(t) = P(T \leq t, E = k)$, we define the transition rate to cause k as

$$\lambda_k(t) = \lim_{dt \rightarrow 0} \frac{(t \leq T < t + dt, E = k | T \geq t)}{dt}$$

where the overall hazard of any event is $\lambda(t) = \sum_{k=1}^K \lambda_k(t)$.

On the condition of mutually exclusive events, the cause-specific cumulative hazard rate for all event types K ,

$$\Lambda_k(t) = \int_0^t \lambda_k(s) ds$$

Overall survival, the probability of being free from any event up to time t , is defined as

$$S(t) = \exp\left(-\sum_{k=1}^K \Lambda_k(t)\right) = \exp(-\Lambda(t))$$

The cause-specific cumulative incidence is determined by all cause-specific hazards. For continuous distributions we have

$$F_k(t) = P(T \leq t, E = k) = \int_0^t S(s) \lambda_k(s) ds$$

From here, we can use approximations. Let $t_i \in [t_1, \dots, t_n]$ be the distinct event times of patients. The cause-specific hazard can be estimated by

$$\widehat{\lambda}_k(t_i) = \frac{d_k(t_i)}{r(t_i)}$$

where d_k is the number of events of type k at time t_i , and r is the number who are at risk. We therefore define the estimator of the cause-specific cumulative incidence function for F_k ,

$$\widehat{F}_k^{AJ}(t) = \sum_{i:t_i \leq t} \widehat{S}(t_i-) \times \widehat{\lambda}_k(t_i)$$

where $S(t_i-)$ is the survival up until the previous timepoint. Note the notation, AJ , as this is a special case of the Aalen-Johansen. Also, the Kaplan-Meier estimator of overall survival of **all** event types is

$$\widehat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d(t_i)}{r(t_i)}\right)$$

where $d(t_i) = \sum_{k=1}^K d_k(t_i)$. The confidence intervals provided come from the R package `survfit`, and the standard errors for this function are “estimates computed using an infinitesimal jack-knife”. In the two case system (alive \rightarrow dead, for example), this is identical to the traditional Greenwood estimate for the variance of the survival curve, $S(t)$ (Therneau et al., 2020). While not covered by Therneau, we presume the confidence intervals estimated using `survfit` in the competing risk setting are comparable to the theoretical estimate (Allignol et al., 2010):

$$\begin{aligned} \widehat{Var}(\widehat{F}_k(\tau)) &= \sum_{t=1}^{\tau} [\widehat{F}_k(\tau) - \widehat{F}_k(t)]^2 \frac{d(t)}{\{r(t)[r(t) - d(t)]\}} \\ &\quad + [\widehat{S}(t-)]^2 \frac{d_k(t)}{r(t)^2} \left[1 - \frac{d_k(t)}{r(t)}\right] \\ &\quad - 2[\widehat{F}_k(\tau) - \widehat{F}_k(t)][\widehat{S}(t-)] \frac{d_k(t)}{n^2(t)} \end{aligned}$$

with confidence intervals

$$\exp \left\{ \ln [\hat{F}_k(t)] \pm z_{1-\alpha/2} \frac{\sqrt{\widehat{Var}(\hat{F}_k(t))}}{\hat{F}_k(t)} \right\}$$

where α is the probability a confidence interval will not include the population parameter. Adapting the Aalen-Johansen method to include left truncated data is undertaken by adjusting the risk set, as with the Kaplan-Meier, such that prevalent patients are included once they have entered the study, i.e. $r(t) = \sum_{i=1}^n \mathbb{I}(L_i < t \leq X_i)$ (Andersen et al. (2012a) proves this under a martingale framework, and Allignol et al. (2010) provide a good summary). This holds under the assumption of independent right censoring and independent left truncation.

6.3.3 Subdistribution approach

Gray (1988) notes that there is no longer a one-to-one correspondence between the effect of a covariate on the cause-specific hazard and the effect on the cumulative scale. This is problematic for proportional hazards model calculations as the effect of a covariate on the cause-specific hazard for a particular cause may be different from its corresponding effect on the probability of the event occurring. To overcome this, we now consider the subdistribution approach. We define the overall survival $P(T > t) = \exp(-\int_0^t h(s)ds)$ and cause-specific cumulative incidence of $F_k(t) = P(T \leq t, E = k)$ and define the subdistribution random variable as $T_k = T \times I(E = k) + \infty I(E \neq k)$.

The subdistribution hazard is defined as

$$\begin{aligned} h_k(s) &= \lim_{ds \rightarrow 0} \frac{1}{ds} P(s \leq T_k < s + ds | T_k \geq s) \\ &= \lim_{ds \rightarrow 0} \frac{\frac{1}{ds} P(s \leq T_k < s + ds, E = k)}{P(T \geq s \cup (T < s, E \neq k))} \end{aligned} \tag{6.1}$$

We note that this interpretation of the hazard is controversial, as it is not a rate in the epidemiological sense, as individuals that experience other competing events remain in the risk set. However, it is theory that is necessary for use in the Fine and Gray model, a proportional hazards model similar to the Cox model that

assesses hazard when competing risks are present. The subdistribution hazard at time t_i is estimated by

$$\widehat{h}_k(t_j) = \frac{d_k(t_j)}{r^*(t_j)}$$

where $d_k(t_j)$ is the number of events at t_j of type k and $r^*(t_j)$ is a modified number at risk, which includes weighted contributions from individuals who experienced the competing event prior to time t_j . As the subdistribution does not account for non-administrative censoring without weightings, we do not know when patients in the competing risk set may have been censored. For this, Geskus recommends that each individual in the risk set shall have weights (Geskus, 2011), so that $r^*(t_j) = \sum_{l=1}^N w_l(t_j)$ where

$$\omega_l(t_i) = \begin{cases} 1 & \text{if event-free and under observation until } t_i \\ \frac{1-\widehat{\Gamma}(t_i-)}{1-\widehat{\Gamma}(t_j-)} & \text{if } l \text{ has competing event observed at } t_j < t_i \\ 0 & \text{otherwise} \end{cases} \quad (6.2)$$

where $\widehat{\Gamma}(t)$ is the estimator of the censoring distribution. The $\Gamma(t)$ has a product-limit form that reverses the role of events and censoring:

$$\widehat{\Gamma}(t) = 1 - \prod_{j:c_j \leq t} \left(1 - \frac{m(c_j)}{r(c_j)}\right)$$

where $m(c_j)$ is the number of censorings at c_j . If there is left truncation, as with prevalent cohorts, then this too needs to be accounted for in the weightings for the risk set (Geskus, 2011). We find approximation $\Phi(t)$ as

$$\widehat{\Phi}(t) = \widehat{P}(L \leq t) = \widehat{P}(-L \geq -t) = \prod_{j:-\ell_j < -t} \left(1 - \frac{w(\ell_j)}{r(\ell_j)}\right) = \prod_{j:\ell_j > t} \left(1 - \frac{w(\ell_j)}{r(\ell_j)}\right)$$

where $w(\ell_j)$ is the number of entries at ℓ_j . Therefore, we have new weightings

$$\omega_l(t_i) = \begin{cases} 1 & \text{if event-free and under observation until } t_i \\ \frac{(1-\widehat{\Gamma}(t_i-))\widehat{\Phi}(t_i-)}{(1-\widehat{\Gamma}(t_j-))\widehat{\Phi}(t_j-)} & \text{if competing event observed at } t_j < t_i \\ 0 & \text{otherwise} \end{cases} \quad (6.3)$$

Geskus notes two forms of the cause-specific cumulative incidence function using the subdistribution hazard which are algebraically equivalent to the Aalen-Johansen form (Geskus, 2011). We include the formula for one below (the ‘Product-Limit estimate’), but we shall use the Aalen-Johansen in our calculations for the cumulative incidence function.

$$F_k^{PL}(t) = 1 - \prod_{j:t_j \leq t} (1 - \hat{h}_k(t_j))$$

Geskus proves that this and the Aalen-Johansen are equivalent on the condition that entry time and the censoring time are independent.

6.4 Cumulative incidence risks

We firstly estimate the marginal cumulative risk (1-KM) and the cause specific cumulative incidence function (CIF) for the incident case and then the prevalent case. This is to highlight possible difference when we have the additional prevalent patients. The Aalen-Johansen is estimated using the `survival` package in R.

6.4.1 Incident only

In this section we only use the incident cohort. Figure 6.1 is the one minus naive Kaplan-Meier (1-KM) under the assumption that the competing event is a form of censoring (death is treated as a form of censoring when we consider the risk of cancer, and cancer is a form of censoring when we consider the risk of death), which is termed the marginal cumulative risk. The cause-specific cumulative incidence is shown in Figure 6.2. For both, the risk of cancer is on the left and

the risk of death prior to cancer is on the right of the figures. A summary of the risk ratios for the 1-KM and the CIF are in Tables 6.4 and 6.5, respectively.

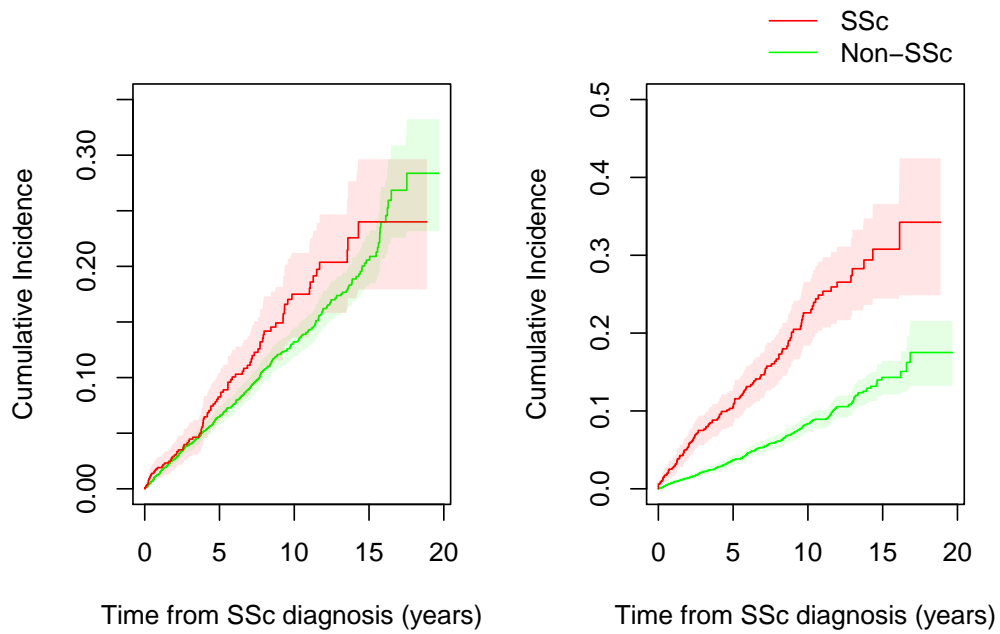


Figure 6.1: Naive Kaplan-Meier estimator or marginal risk (1-KM) with 95% confidence intervals, Left: Cancer, Right: Death.

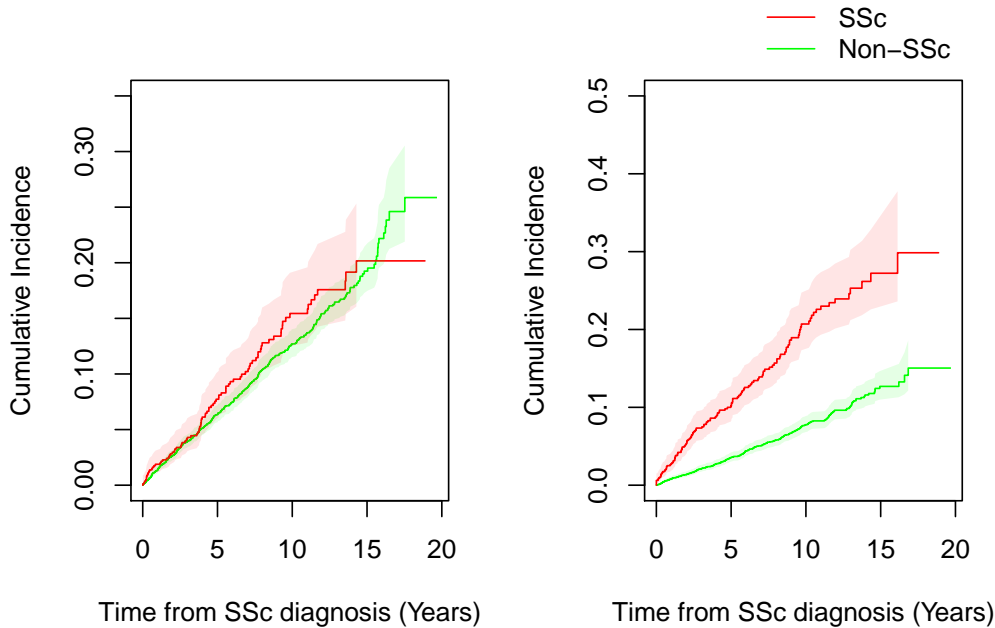


Figure 6.2: Cause-specific cumulative incidence function for different competing risk hazards (CIF) with 95% confidence intervals, Left: Cancer, Right: Death.

There are few differences between the naive estimator and the cause-specific CIF, except for a small reduction in incidence for the cause-specific risk, as expected.

In both KM and CIF for cancer, for the first four years there seems to be very little difference in the cumulative incidence between those who have SSc and those who do not. This changes as time increases. This could arise because there is a delay in the impact of SSc on the body, and hence any significant effect on the body is not observed over the short term, however this work will not consider the clinical aspects of SSc development. In later years of the 1-KM there is a greater risk of cancer in those with SSc than for the non-SSc patients, with a ratio between 1.1-1.4. This small increase is consistent with literature reviewed. We also note the relatively short timescale, with maximum 20 years follow-up, and that the SSc curve becomes more variable in later years due to a smaller number of patients in the risk set in later years. As most literature reviewed indicates an average onset of cancer from SSc diagnosis of between 10-16 years, this lack of observable time

is of some concern and it gives greater incentive for the use of prevalent cohorts. On the right of the Figures and 6.1 and 6.2 we have the risk of death without cancer, which shows approximately 2.5 times the incidence over time for those with SSc compared to non-SSc. Again, this increased risk is consistent with literature, although we remember that this is ‘death without cancer’ as opposed to deaths which may have cancer preceding death, therefore it is not a true comparison.

6.4.2 Including prevalent patients

We add prevalent patients to the incident cohort to make the combined cohort of patients. Here, patients are included in the risk set from the time they enter the study, as with the methodology of Chapter 5. The naive left truncated Kaplan-Meier is shown in Figure 6.3 and the left truncated cause-specific CIFs are shown in Figure 6.4. In this chapter we shall again assume we have random/independent right censoring and left truncation, however we relax these assumptions in the following chapters.

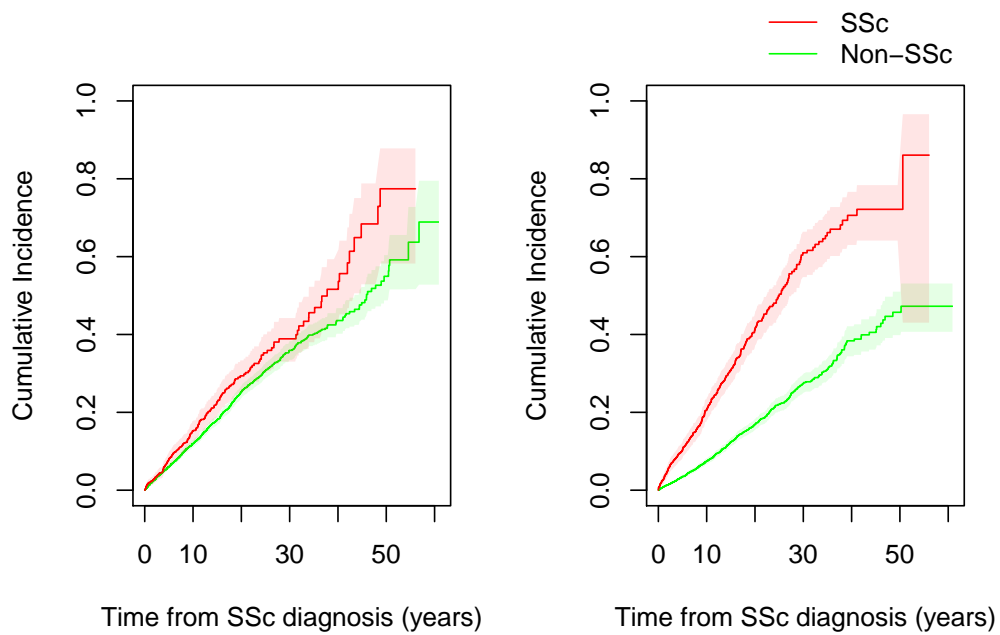


Figure 6.3: Naive Kaplan-Meier with 95% confidence intervals, incident and prevalent cohorts, Left: Cancer, Right: Death.

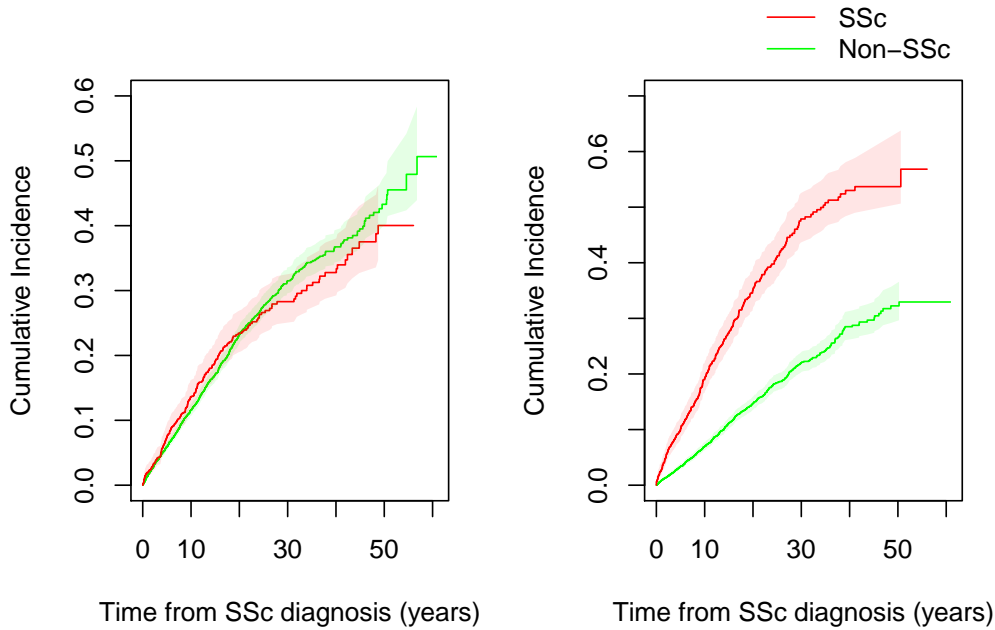


Figure 6.4: Cause-specific cumulative incidence function for incident and prevalent cohorts with 95% confidence intervals. Left: Cancer, Right: Death.

With regards to cancer onset in the naive Kaplan-Meier, Figure 6.3, there is a small difference between the SSc and the non-SSc groups, with the SSc group being more at risk. Table 6.4 shows that while the risk ratios of the incident cohort and the combined cohort are similar to each other, the confidence intervals are narrower, due to the increase in the number of patients and therefore higher accuracy. Also, we are able to observe longer times after SSc diagnosis, a large benefit of prevalent patient inclusion. As with the solely incident cohort's 1-KM estimations, the risk ratio suggests there is possibly a small effect of SSc on cancer risk. However, this result is on the assumption of informative censoring, which may not hold. For example, we suspect that patients who enter the study at higher ages are at both an increased risk of death and an increased risk of cancer. In the cause-specific CIF, while the SSc group have a higher risk of cancer than the comparators prior to 25 years after SSc diagnosis, this effect actually appears to reverse after this time. This reflects the increased mortality in SSc patients,

which precedes cancer events. However, the curves are close. This suggests that, in the presence of this competing risk, we would not observe a greater number of SSc patients being diagnosed with cancer than the general population. As with the incident cohort, we observe that the curves are coincident for approximately the first 4 years. The risk of death prior to cancer in those with SSc continues to be approximately 2.5 times that of the non-SSc patients within 30 years of SSc diagnosis.

With the addition of prevalent patients, there is a noticeable difference between the naive Kaplan-Meier (Figure 6.3) and the cause-specific cumulative incidence function (Figure 6.4). Table 6.4 illustrates the risk ratios for the naive Kaplan-Meier estimator when using the incident and combined cohorts, and Table 6.5 is the same for the cause-specific risk ratio. If death is treated as censoring, there is a larger risk ratio between SSc and non-SSc patients, where the combined cohort implies a significant risk in SSc patients, but then the risk ratio decreases further from time of diagnosis. However, with death as a competing risk, the cancer risk ratio is no longer significant. If we assumed independent censoring (that there is independence between a patient's hazard of cancer and a patient's hazard of death), these differences between the types of risk might imply a causal link between SSc and cancer, but no increased risk in the observable world as SSc patients have a much higher mortality.

6.4.3 Truncation distribution

In reference to Section 5.4.2, we produce the truncation distribution in the setting of two competing events. It is required that patients are both cancer free and alive when they enter the study, therefore the survival is based on these two conditions. Using the survival estimate from the Aalen-Johansen method:

$$\hat{G}(t) = \left(\sum_{i=1}^n \frac{1}{\hat{S}(L_i)} \right)^{-1} \sum_{i=1}^n \frac{\mathbb{I}(L_i < t)}{\hat{S}(L_i)}$$

Time from SSc diagnosis (years)	Incident [95% CI]	Combined [95% CI]
5	SSc 0.083 [0.060, 0.104] Non-SSc 0.064 [0.057, 0.072] Ratio 1.281 [0.957, 1.714]	SSc 0.080 [0.061, 0.098] Non-SSc 0.060 [0.053, 0.066] Ratio 1.333 [1.025, 1.732]
10	SSc 0.175 [0.136, 0.212] Non-SSc 0.132 [0.119, 0.145] Ratio 1.325 [1.043, 1.682]	SSc 0.152 [0.125, 0.178] Non-SSc 0.120 [0.110, 0.129] Ratio 1.268 [1.049, 1.533]
15	SSc 0.240 [0.179, 0.296] Non-SSc 0.206 [0.182, 0.228] Ratio 1.167 [0.892, 1.526]	SSc 0.226 [0.192, 0.258] Non-SSc 0.180 [0.168, 0.192] Ratio 1.252 [1.065, 1.472]
20		SSc 0.294 [0.252, 0.334] Non-SSc 0.252 [0.236, 0.268] Ratio 1.166 [1.001, 1.359]
30		SSc 0.389 [0.330, 0.442] Non-SSc 0.359 [0.335, 0.383] Ratio 1.082 [0.923, 1.268]

Table 6.4: Naive one minus Kaplan-Meier (marginal cumulative incidence) of cancer for incident and combined datasets at time points 5, 10, 15, 20 and 30 years after SSc diagnosis, also the ratio between them in bold (SSc over non-SSc). The 95% confidence intervals are given in brackets.

Time from SSc diagnosis (years)	Incident [95% CI]	Combined [95% CI]
5	SSc 0.077 [0.059, 0.101] Non-SSc 0.063 [0.056, 0.071] Ratio 1.220 [0.912, 1.631]	SSc 0.075 [0.059, 0.095] Non-SSc 0.059 [0.053, 0.066] Ratio 1.272 [0.978, 1.653]
10	SSc 0.154 [0.125, 0.191] Non-SSc 0.127 [0.115, 0.14] Ratio 1.215 [0.961, 1.536]	SSc 0.136 [0.114, 0.161] Non-SSc 0.116 [0.107, 0.125] Ratio 1.175 [0.972, 1.42]
15	SSc 0.202 [0.161, 0.253] Non-SSc 0.192 [0.173, 0.214] Ratio 1.048 [0.814, 1.348]	SSc 0.190 [0.165, 0.220] Non-SSc 0.170 [0.159, 0.182] Ratio 1.119 [0.953, 1.314]
20		SSc 0.234 [0.205, 0.268] Non-SSc 0.231 [0.217, 0.246] Ratio 1.013 [0.872, 1.176]
30		SSc 0.283 [0.248, 0.322] Non-SSc 0.315 [0.296, 0.336] Ratio 0.898 [0.777, 1.038]

Table 6.5: Cause-specific cumulative incidence of cancer for incident and combined datasets at time points 5, 10, 15, 20 and 30 years after SSc diagnosis, also the ratio between them in bold (SSc over non-SSc). The 95% confidence intervals are given in brackets.

The truncation distribution is shown in Figure 6.5, which uses prevalent patients only (prevalent patients are used to estimate survival as well).

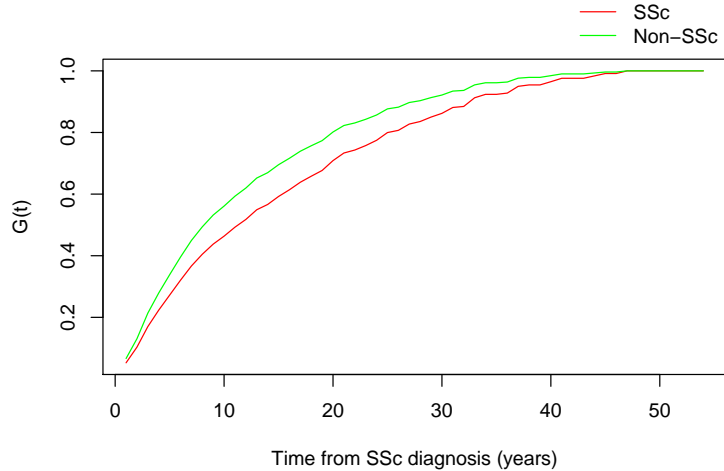


Figure 6.5: Estimation of left truncation distribution

There is a small difference between this truncation distribution and the truncation distribution for the mortality study in the previous chapter, with this new survival probability producing a more uniform curve. This is due to the differences in recruitment, as our dataset was recruited for the outcome cancer. In Chapter 5 we are missing a small subset of patients who had cancer prior to study entry but could still have been alive during recruitment. Therefore, this is not an ideal dataset for the study of mortality, and would estimate incorrect survival probability. Therefore, the curves in Figure 6.5 shows a better estimate of the truncation distribution, but we continue to have the appearance of more patients with shorter truncation times than large. Again, this could be for any of the reasons specified in the last chapter.

It is questionable to include non-SSc patients, and weight non-SSc patients by their survival, here. Prevalent patients were recruited into the study due to being matches for SSc patients. Therefore, their recruitment is not based on their survival but on the SSc patient's survival. We discuss this more in future chapters, as this recruitment criteria affects our methodology.

6.5 Cause-specific and subdistribution proportional hazards model

The cause-specific proportional hazards model estimates the effect of covariates on the rate at which events occur in subjects who are currently event free. Prentice & Kalbfleisch (1979) first proposed the consideration of a Cox-type model that estimates the effects of covariates on the cause-specific hazard rate

$$\lambda_k(t|x) = \lambda_{k,0} \exp(\beta_k^T x)$$

where $\lambda_{k,0}$ is the cause-specific baseline hazard for event type k .

The subdistribution hazard ratios obtained from the Fine and Gray model describe the relative effect of covariates on the subdistribution hazard function. Hence, the covariates in this model can also be interpreted as having an effect on the cumulative incidence function.

Similar to above, the proportional subdistribution hazards model is then:

$$h_k(t|x) = h_{k,0} \exp(\phi_k^T x)$$

where $h_{k,0}$ is the subdistribution baseline hazard for event type k . We shall not go deeply into the derivation of the hazard ratios, as it has extensive theory similar to that of the Cox proportional hazards model of Chapter 4 and 5 and it is not used beyond this chapter. For reference, we refer to the original Fine and Gray paper (Fine & Gray, 1999), and when considering left truncation we refer primarily to Geskus (Geskus, 2011). To implement this in R we could use the `finegray` function in the `survival` package but instead we use the `crprep` function in the `mstate` package, which also finds the weights for censoring. After this analysis, these weights can be used in combination with `coxph` in the `survival` package.

6.5.1 How to interpret the Fine and Gray model compared to the cause-specific proportional hazards model

The issue discussed in this section was first defined in Allison (2018), and it discusses the important differences between the two proportional hazards models and is therefore reiterated here. Simulations were also provided in Allison’s work, although the ones shown here are different.

We discuss how to interpret the Fine and Gray model with the emphasis on the comparison between this and the cause-specific proportional hazards model. Regression using a cause-specific hazard function provides hazard ratios that describe the effect of a covariate on the relative increase in the rate of the event of interest. This leads to the examination of the casual relationship between risk factors and an event, and therefore gives insight into the potential effect of an exposure in the case of the elimination of competing events. This differs from the subdistributional hazards model, which aims to establish hazard due to the presence of competing risks, and can lead to some counter-intuitive interpretations. Therefore, while relationships between an event of interest and its competing event(s) can be analysed by the use of the Fine and Gray model, it cannot be used to establish causal relationships. As we have stated in a previous chapter, the use of the subdistribution hazard is introduced as a way to model the impact of covariates on the cumulative incidence functions, however it is not a ‘natural’ interpretation of hazard. For example, if a variable increases the cause-specific hazard of event Y , it will lead to a decrease in the subdistribution hazard for competing event D .

To demonstrate this we perform a simple simulation. Each patient will have two binary covariates, L_1 and L_2 , with the first only affecting the hazard of Y and the latter only the hazard of D . The patients are randomly assigned L_1 and L_2 both with probability 0.5. Suppose patients have outcomes Y and D , both distributed with an exponential distribution. The rate for the distribution of times for outcome Y is double if they have exposure L_1 and the rate for the

distribution of times for outcome D is double if exposed to outcome L_2 . We simulate $Y_{L_1=1} \sim \text{Exp}(0.4)$, $Y_{L_1=0} \sim \text{Exp}(0.2)$, and $D_{L_2=1} \sim \text{Exp}(0.2)$, $D_{L_2=0} \sim \text{Exp}(0.1)$. Whichever of Y and D is observed first is the end event and the minimum of the two times is the end time, $X = \min(Y, D)$. The mean of the hazard ratios for outcome Y of 1000 simulations are shown in Table 6.6, along with the 95% confidence intervals (CI). Let b_i be the vector of point estimates at time τ , where $i = 1, \dots, B$, with mean $\bar{b} = \frac{1}{B} \sum_{i=1}^n b_i$ and confidence intervals (95%) for the sampling are included $(\bar{b} \pm z_{(1-\alpha)/2} \frac{s}{\sqrt{B}})$, where $s = \sqrt{\frac{\sum_{i=1}^n (b_i - \bar{b})^2}{B-1}}$.

	Cause-specific proportional hazards model		Subdistribution proportional hazards model	
	HR	95% CI	HR	95% CI
L_1	1.999	[1.998, 2.000]	1.818	[1.817, 1.819]
L_2	1.000	[1.000, 1.001]	0.696	[0.696, 0.696]

Table 6.6: Simulated data to aid interpretation of the Fine and Gray model. The hazard ratio is shown for covariates L_1 and L_2 for outcome Y . The ratio is comparing those with the covariate to those without.

From the cause-specific proportional hazards model we can see the causal relationship, with twice the hazard of event Y given the patient has covariate L_1 and with no increased or decreased hazard with covariate L_2 . It can be seen that the Fine and Gray model produces a hazard ratio below 2 for L_1 and below 1 for L_2 due to the inclusion of the possible event D , which leads to less Y events being observed. Therefore, we should not use the subdistribution PH model to estimate causal parameters. However, the subdistribution PH model is useful for understanding the hazard given the current hazard of competing events, and therefore is perhaps preferable for explaining the current hazard of cancer. It is worth remembering that this result would no longer hold if there was a change in mortality for SSc patients or the general population.

6.5.2 Cause-specific proportional hazards model for cancer

We use the cause-specific proportional hazards model to identify risk factors for cancer. The theory for this was covered in Chapter 4 in reference to mortality, however we now perform this for cancer.

We start with the incident dataset only. Without the addition of covariates, the hazard ratio for cancer in those with SSc compared to those without SSc is 1.22 [95% CI: 0.97-1.53], implying no significant difference between the two groups. We use a proportional hazards model to find the hazard ratio of our covariates of interest, Table 6.7. This includes all covariates that we could use, and then a reduction of the model based on minimizing the AIC. Removing location and BMI lowered the AIC. After adding back in patients who had missing BMI data, the number of patients went up from 4,711 to 5,139. After this, the other covariates were tested but none minimised the AIC further, hence the use of the reduced model, as shown.

When accounting for covariates this ratio rises to 1.38 then reduces to 1.30 with the removal of extraneous variables, which are alcohol intake and BMI, therefore including SSc status, sex, age at SSc diagnosis and smoking status, as all of these hazard ratios are significant.

It can be seen from Table 6.7 that having SSc implies an increased hazard of cancer, with the 95% confidence interval not including 1. Additionally, age at SSc diagnosis also implies an increased hazard of cancer. Males have a higher hazard too.

We verify the proportional hazards assumption for the reduced model, Table 6.8. As none of the p-values are below 0.05, the proportionality assumption is satisfied.

6.5.3 Inclusion of prevalent cohort

We now include the prevalent cohort and note the differences between the combined cohorts and the solely incident cohort. We work on the assumption

	Full model		Reduced model	
	HR	95% CI	HR	95% CI
SSc status (Non-SSc=0, SSc=1)	1.38	[1.09, 1.76]	1.30	[1.04, 1.63]
Age at SSc diagnosis (10 years)	1.56	[1.44, 1.69]	1.63	[1.52, 1.77]
Sex (Female=0, Male=1)	1.31	[1.02, 1.67]	1.28	[1.00, 1.54]
Location: East Midlands	0.94	[0.48, 1.85]		
Location: East of England	1.31	[0.78, 2.18]		
Location: North East	1.81	[0.79, 4.16]		
Location: North West	1.58	[1.01, 2.48]		
Location: Northern Ireland	1.33	[0.68, 2.63]		
Location: Scotland	1.61	[1.03, 2.49]		
Location: South Central	1.27	[0.81, 2.01]		
Location: South East Coast	1.37	[0.84, 2.24]		
Location: South West	1.40	[0.86, 2.31]		
Location: Wales	1.67	[1.07, 2.60]		
Location: West Midlands	1.57	[0.99, 2.51]		
Location: Yorkshire & The Humber	1.06	[0.54, 2.09]		
BMI: <18.5	0.89	[0.46, 1.75]		
BMI: 25-29.9	0.86	[0.70, 1.06]		
BMI: 30-34.9	0.68	[0.50, 0.92]		
BMI: 35-39.9	1.04	[0.71, 1.53]		
BMI: 40+	0.75	[0.37, 1.53]		
Ex-smoker	1.21	[0.98, 1.50]	1.20	[0.98, 1.47]
Smoker	1.17	[0.92, 1.49]	1.25	[1.00, 1.57]
Ex-drinker	1.33	[0.96, 1.83]	1.38	[1.01, 1.88]
Heavy drinker	1.47	[0.84, 2.58]	1.61	[0.99, 2.64]
Non-drinker	0.93	[0.71, 1.22]	0.96	[0.75, 1.23]

————→
After
reduction

Table 6.7: Cause-specific proportional hazards model for cancer, from the incident cohort only. On the left are the results from the full model with all covariates, on the right are the results of the reduced model after covariates are removed due to minimisation of AIC. Location has a baseline of London, BMI (kg/m²) has a baseline of a ‘normal’ health range (18.5-24.9), the baseline for smoking is ‘non-smoking’, and the baseline for drinking is ‘drinker’.

	Chi-squared	Degrees of freedom	p-value
SSc status	0.0349	1	0.852
Age at SSc diagnosis	3.4984	1	0.061
Sex	0.5257	1	0.468
Smoking	1.2170	2	0.544
Alcohol	3.1033	3	0.376
Global	8.5060	8	0.386

Table 6.8: Scaled Schoenfeld residuals, cause-specific proportional hazards for cancer, incident dataset.

of independent left truncation. This analysis may not hold for dependent truncation. Work has been done in this area, for example as in Keiding & Gill (1990) who recommend including left truncation time as a covariate, or Rennert & Xie (2021) who propose an expectation-maximization to relax the independence assumption to conditionally independent truncation. We discuss this more in later chapters.

Without additional covariates, the hazard ratio for SSc status is greater than that found for the incident cohort alone and continues to be significant, with HR=1.27 [95% CI: 1.09-1.49]. Table 6.9 shows the full model and the reduced model when AIC minimisation is used. Comparing the full model and the reduced model, there is little difference for the hazard ratios of SSc type. The reduced model recommended the removal of location and BMI, as with the incident cohort.

The hazard ratio for SSc is greater than 1, implying a causal relationship of SSc on cancer. An increased age at SSc diagnosis is associated with an increased hazard of cancer, which is expected. There is the indication that males have a higher hazard. Being a smoker is associated with higher cancer hazard, as expected from the relationship between smoking and cancer. Being a non-drinker has a higher hazard ratio than being a mild/moderate drinker, which may be due to people with poorer health being advised by medical professionals to reduce alcohol intake.

We check the proportional hazards assumption in Table 6.10. Age at SSc

	Full model		Reduced model	
	HR	95% CI	HR	95% CI
SSc status (Non-SSc=0, SSc=1)	1.36	[1.15, 1.62]	1.35	[1.16, 1.57]
Age at SSc diagnosis (10 years)	1.58	[1.49, 1.67]	1.61	[1.53, 1.70]
Sex (Female=0, Male=1)	1.35	[1.13, 1.60]	1.31	[1.12, 1.54]
Location: East Midlands	0.96	[0.59, 1.54]		
Location: East of England	1.33	[0.96, 1.83]		
Location: North East	1.03	[0.58, 1.85]		
Location: North West	1.22	[0.90, 1.66]		
Location: Northern Ireland	1.07	[0.71, 1.61]		
Location: Scotland	1.26	[0.93, 1.70]		
Location: South Central	1.13	[0.83, 1.53]		
Location: South East Coast	1.15	[0.83, 1.60]		
Location: South West	1.01	[0.72, 1.41]		
Location: Wales	1.37	[1.01, 1.85]	→	
Location: West Midlands	1.08	[0.78, 1.48]	After	
Location: Yorkshire & The Humber	1.21	[0.80, 1.84]	reduction	
BMI: <18.5	0.69	[0.41, 1.16]		
BMI: 25-29.9	0.97	[0.84, 1.11]		
BMI: 30-34.9	0.76	[0.62, 0.94]		
BMI: 35-39.9	0.89	[0.66, 1.21]		
BMI: 40+	0.80	[0.50, 1.28]		
Ex-smoker	1.07	[0.92, 1.24]	1.03	[0.89, 1.18]
Smoker	1.20	[1.02, 1.41]	1.20	[1.03, 1.39]
Ex-drinker	1.34	[1.04, 1.71]	1.27	[1.00, 1.61]
Heavy drinker	1.00	[0.61, 1.64]	1.08	[0.71, 1.64]
Non-drinker	0.85	[0.71, 1.02]	0.87	[0.75, 1.02]

Table 6.9: Cause-specific model for the hazard ratios for cancer from both the incident and prevalent data. On the left are the results from the full model with all covariates, on the right are the results of the reduced model after covariates are removed due to minimisation of AIC. Location has a baseline of London, BMI (kg/m²) has a baseline of a ‘normal’ health range, 18.5-24.9, the baseline for smoking is ‘non-smoking’ and the baseline for drinking is ‘drinker’.

diagnosis fails the proportional hazards assumption. Plotting the Schoenfeld residual against time, Figure 6.6, we note that the failure occurs at greater age at SSc diagnosis. Otherwise the line seems to be close to constant.

	chisq	df	p-value
SSc type	1.78	1	0.182
Age at SSc diagnosis	4.53	1	0.033
Sex	1.05e-04	1	0.992
Smoking	1.52	2	0.468
Alcohol	1.93	3	0.587
Global	9.57	8	0.297

Table 6.10: Scaled Schoenfeld residuals, cause-specific proportional hazards for cancer, combined (incident and prevalent) dataset.

We have three principal options to adjust for this failure:

- Stratification - The covariates with non-proportional effects may be incorporated into the model as stratification factors rather than regressors. We investigated several different stratifications of age at SSc diagnosis, and these did not appear to be a great improvement (not shown here). One that did raise the p-value of the Schoenfeld residual test from 0.033 to 0.46 was dividing SSc age into <40, 40-60 years and 60+ years. The hazard ratio for 40-60 years compared to <40 years is 2.13 [95% CI: 1.65, 2.75] and 60+ years compared to <40 years is 4.99 [95% 3.81, 6.54], and the hazard ratio for SSc type does not change greatly (1.33, 95% CI 1.14-1.55), hence complete results are not shown here.
- Partitioning - This is the partitioning of the time axis, in the hope that the proportional hazards assumption may hold at least approximately over short time periods. This was done for this dataset for differing partitions of time (not shown here) however the p-value did not improve.
- Interactions - Interactions between terms were tested, but no combinations

for age at SSc diagnosis were sufficient to satisfy the proportionality assumption.

After this, the recommendation is to use a different model, where an accelerated failure time or additive hazards model might be more appropriate for the data. From viewing the Schoenfeld residual plots in Figure 6.6 the fit appears close to constant over time, and we therefore do not explore this further, however a more flexible model, such as the one we investigate in the next chapter, may be an improvement.

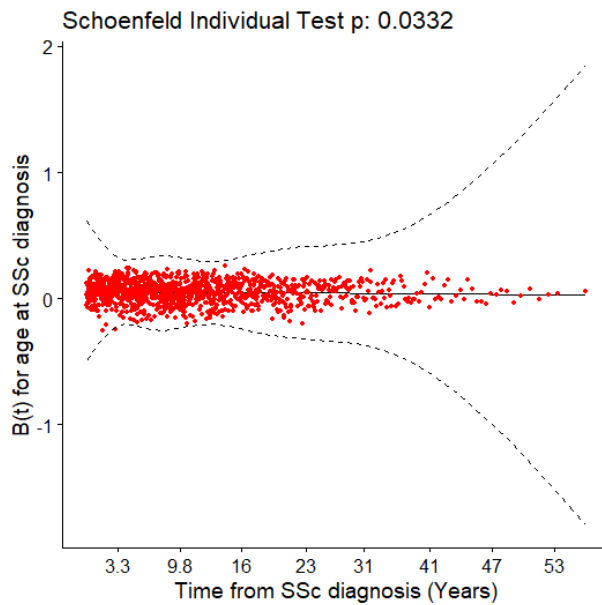


Figure 6.6: Scaled Schoenfeld residual plots for age at SSc diagnosis, cause-specific hazard ratios, combined dataset.

6.5.4 Proportional subdistribution hazards model for cancer in the incident set

We begin with only using the incident cohort. Without other covariates, the Fine and Gray model (the subdistribution proportional hazards model) estimates a hazard ratio for SSc type of 1.11 [95% CI: 0.89-1.39]. This is significantly lower than the cause-specific hazard ratio for SSc type, and implies that in the presence of competing risks those with SSc are not more at risk of cancer than the general

population due to SSc having higher mortality. The subdistribution proportional hazards models for the incident set with all covariates and the reduced set are given in Table 6.11. Using backward AIC, we find that the optimal model is to use age at SSc diagnosis, sex and smoking as the only variables to include in the model. For minimal AIC, SSc type is not required. However, due to the nature of our project investigating the impact of SSc on cancer risk we will retain SSc status in the model.

The hazard ratio for SSc status is not significant, with an HR of 1.10 [95% CI: 0.88-1.38]. Age at SSc diagnosis is significant, with hazard increasing by 0.43 for every decade older on entry to the study. Sex is still significant, as in the cause-specific hazard ratio, with males having a higher hazard. However, the Schoenfeld residuals for the subdistribution proportional hazards model, as shown in Table 6.12, indicate that the age at SSc diagnosis strongly violates the proportional hazards assumption. Figure 6.7 indicates a slow downward trajectory with increasing time. This could be due to death now being included in the model, as the hazard for death will certainly increase with age, and will most likely not have a linear effect.

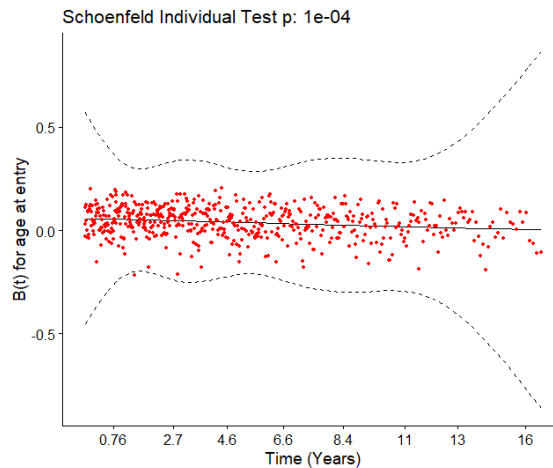


Figure 6.7: Scaled Schoenfeld residual plots for age at SSc diagnosis, subdistribution hazard, incident dataset.

	Full model		Reduced model	
	HR	95% CI	HR	95% CI
SSc status non-SSc=0, SSc=1)	1.17	[0.92, 1.47]	1.10	[0.88, 1.38]
Age at SSc diagnosis (10 years)	1.43	[1.33, 1.53]	1.49	[1.39, 1.59]
Sex (Female=0, Male=1)	1.24	[0.98, 1.57]	1.19	[0.96, 1.48]
Location: East Midlands	0.91	[0.47, 1.77]		
Location: East of England	1.28	[0.77, 2.13]		
Location: North East	1.65	[0.74, 3.69]		
Location: North West	1.57	[1.00, 2.47]		
Location: Northern Ireland	1.35	[0.69, 2.63]		
Location: Scotland	1.52	[0.98, 2.35]		
Location: South Central	1.25	[0.80, 1.97]		
Location: South East Coast	1.34	[0.82, 2.19]		
Location: South West	1.34	[0.82, 2.20]		
Location: Wales	1.64	[1.06, 2.54]		
Location: West Midlands	1.53	[0.97, 2.42]		
Location: Yorkshire & The Humber	1.11	[0.57, 2.15]		
BMI: <18.5	0.76	[0.40, 1.47]		
BMI: 25-29.9	0.89	[0.72, 1.08]		
BMI: 30-34.9	0.67	[0.49, 0.91]		
BMI: 35-39.9	1.04	[0.71, 1.52]		
BMI: 40+	0.67	[0.34, 1.36]		
Ex-smoker	1.17	[0.95, 1.44]	1.17	[0.96, 1.42]
Smoker	1.12	[0.88, 1.41]	1.25	[1.01, 1.54]
Ex-drinker	1.20	[0.88, 1.64]		
Heavy drinker	1.31	[0.75, 2.28]		
Non-drinker	0.88	[0.67, 1.15]		

Table 6.11: Fine and Gray model for cancer in the incident cohort. On the left are the results from the full model with all covariates, on the right are the results of the reduced model after covariates are removed due to minimisation of AIC. Location has a baseline of London, BMI (kg/m²) has a baseline of a ‘normal’ health range, 18.5-24.9, the baseline for smoking is ‘non-smoking’ and the baseline for drinking is ‘drinker’.

	chisq	df	p-value
SSc type	1.19	1	0.2751
Age at SSc diagnosis	16.20	1	5.8e-05
Sex	0.47	1	0.493
Smoking	1.095	2	0.5784
Global	18.33	5	0.0026

Table 6.12: Scaled Schoenfeld residuals, subdistribution proportional hazards for cancer, incident dataset.

6.5.5 Proportional subdistribution hazards with prevalent patients

Geskus (2011) proposed the inclusion of left truncation data via the use of weightings in the risk set, as specified in Section 6.3.3. With the addition of the prevalent cases, the inclusion of covariates BMI and alcohol intake minimises the AIC in addition to age at SSc diagnosis and sex, but not smoking. This is quite different to when only incident patients are included. With no additional covariates, the hazard ratio for SSc type is 1.05 [95% CI: 0.91-1.21]. The full model and reduced model are shown in Table 6.13.

The results suggest that having SSc does not lead to an increased hazard of cancer when death is considered as a competing event, as SSc patients are at a higher risk of death in general. However a causal relationship between SSc and cancer could still exist. It could be that as SSc patients have a higher hazard of death than non-SSc patients, it is more likely that they will die of other causes before a cancer event. This can be interpreted as ‘currently, SSc patients are not more likely to observe cancer compared to the general population’, but we recognise that this conclusion may no longer be valid if there are developments in the future treatment of mortality in SSc patients, narrowing the hazard of death between SSc patients and non-SSc patients.

However, there are clearly significant breaches of the proportional hazards assumption, as shown in Table 6.14.

	Full model			Reduced model	
	HR	95% CI		HR	95% CI
SSc status (Non-SSc=0, SSc=1)	1.02	[0.87, 1.21]		1.02	[0.87, 1.21]
Age at SSc diagnosis (10 years)	1.22	[1.17, 1.27]		1.22	[1.17, 1.27]
Sex (Female=0, Male=1)	1.25	[1.06, 1.48]		1.26	[1.07, 1.48]
Location: East Midlands	0.96	[0.60, 1.53]			
Location: East of England	1.34	[0.98, 1.84]			
Location: North East	1.04	[0.59, 1.83]			
Location: North West	1.32	[0.98, 1.79]			
Location: Northern Ireland	1.17	[0.79, 1.74]			
Location: Scotland	1.24	[0.92, 1.68]			
Location: South Central	1.22	[0.91, 1.65]			
Location: South East Coast	1.22	[0.89, 1.69]			
Location: South West	1.10	[0.79, 1.53]	→		
Location: Wales	1.42	[1.06, 1.91]	After		
Location: West Midlands	1.15	[0.84, 1.57]	reduction		
Location: Yorkshire & The Humber	1.33	[0.89, 1.99]			
BMI: <18.5	0.59	[0.35, 0.98]		0.59	[0.35, 0.97]
BMI: 25-29.9	1.02	[0.89, 1.16]		1.03	[0.90, 1.18]
BMI: 30-34.9	0.74	[0.60, 0.91]		0.75	[0.61, 0.92]
BMI: 35-39.9	0.86	[0.63, 1.16]		0.87	[0.64, 1.18]
BMI: 40+	0.67	[0.42, 1.06]		0.67	[0.42, 1.06]
Ex-smoker	1.05	[0.91, 1.21]			
Smoker	1.03	[0.88, 1.20]			
Ex-drinker	1.15	[0.91, 1.47]		1.15	[0.91, 1.46]
Heavy drinker	0.76	[0.46, 1.24]		0.76	[0.47, 1.24]
Non-drinker	0.83	[0.70, 0.98]		0.82	[0.69, 0.97]

Table 6.13: Fine and Gray model for cancer from both the incident and prevalent data. On the left are the results from the full model with all covariates, on the right are the results of the reduced model after covariates are removed due to minimisation of AIC. Location has a baseline of London, BMI (kg/m²) has the baseline of a ‘normal’ health range, 18.5-24.9, the baseline for smoking is ‘non-smoking’ and the baseline for drinking is ‘drinker’.

	chisq	df	p-value
SSc type	4.2137	1	0.040
Age at SSc diagnosis	111.4710	1	<2e-16
Sex	0.0523	1	0.819
BMI	8.2800	5	0.141
Alcohol	9.2283	3	0.026
GLOBAL	130.3823	11	<2e-16

Table 6.14: Scaled Schoenfeld residuals, subdistribution proportional hazards for cancer, incident and prevalent dataset.

We used the previously mentioned methods to account for this lack of proportionality:

- Stratification - We attempted to stratify age at SSc diagnosis. After several attempts of stratifying by age at SSc diagnosis, none of them appeared to eliminate the non-proportionality (not shown here). Also, stratifying is not possible for SSc type, as this is a binary covariate.
- Partitioning - While partitioning ‘time from SSc diagnosis’ into smaller intervals did work for very small partitions, it then lacked sufficient power because there were fewer events in each interval. Therefore, this was not successful.
- Interactions - Interactions between terms were tested, but no combinations with age at SSc diagnosis or SSc status were sufficient to satisfy the proportionality assumption.

In the light of these results, the recommendation is to use a different model, where an accelerated failure time or additive hazards model might be more appropriate for the data.

The failing of the proportionality assumption in the cause-specific proportional hazard model is concerning, however we are utilizing different methods in later

chapters, and are mainly considering the cumulative incidence function. For this chapter, both the incident and combined subdistribution PH models suggest that SSc patients are not at a significant risk of cancer when compared to the non-SSc when hazard of death is taken into account.

6.6 Summary

We have outlined the common methods used to account for competing risks using cumulative incidence curves. By considering death as a separate event and not as censoring, the risk ratio difference between SSc patients compared to non-SSc patients becomes narrower.

Prevalent patients were included. These patients narrowed the confidence intervals, and allowed for longer follow-up time. When considering both the naive Kaplan-Meier and cause-specific curves, the inclusion of prevalent cohorts decreased the risks compared to the solely incident curves. The risk ratios between SSc and non-SSc, however, remained similar to the solely incident risk ratios.

The cumulative incidence curves presented above, when covariates are not taken into account, suggest a ‘causal’ relationship when we consider the naive Kaplan-Meier, with SSc patients having between a 10% to 30% higher ‘causal’ risk of cancer. However, due to the current higher risk of death in the SSc group compared with the non-SSc group, the SSc group does not have a higher probability of observing cancer compared to the non-SSc. The proportional hazards models agree with these findings, where the cause-specific hazard ratios suggest an increased hazard of cancer in those with SSc when death is treated as censoring, but not when death is treated as a competing event using the Fine and Grey model.

In the next chapter, we shall discuss competing risks from a causal perspective under a causality framework for a more rigorous definition, and introduce the g-formula, which is a method that utilizes covariates and accounts for informative

censoring. We then aim to modify this method to allow for the inclusion of prevalent patients.

Chapter 7

G-formula: Theory and application to the incident cohort

7.1 Introduction

Until now we have focused on commonly used nonparametric methods (Kaplan-Meier survival curve, Aalen-Johansen cause-specific cumulative incidence) and the Cox model. However, the Kaplan-Meier has a couple of drawbacks. When the covariate distribution differs between exposure groups, differences in the Kaplan-Meier curve can inform us of an association between exposure and outcome, but does not always demonstrate a causal link. Matching is often thought to correct for this, but is a costly process and it is almost impossible to match on all covariates. Confounders can be stratified, but this reduces the risk set used for the NPMLE estimates to the size of the strata, reducing efficiency. In addition, the Kaplan-Meier curve has a strong assumption of independent censoring (sometimes termed non-informative censoring) where the probability of a patient being lost to follow-up is independent of their covariates. This may not hold in our study, particularly due to death being

treated as a form of censoring in the Kaplan-Meier.

This chapter introduces the g-formula. The g-formula is a discrete time method that estimates the average causal effect of an exposure on an outcome, taking into account the possibility of confounders. It was first described by Robins (1986) as part of a family of g-methods. The aim therefore is that the g-formula will provide knowledge of the causal effect of SSc on cancer, rather than information on association, which is provided by the Kaplan-Meier curve. The g-formula is often used when flexibility is required to include time-varying covariates (although these are absent in our study). Also, the g-formula does not assume that the hazard ratio is homogeneous over the levels of the confounders, but instead permits us to obtain risk ratios without stratification over covariates. Lastly, it differs in how it handles right censoring in that patients remain in the risk set even after censoring, allowing for adjustment of dependent (informative) censoring.

This chapter will predominantly make use of the competing risk framework provided by Young et al, 2020, which allows for both a parametric g-formula approach and an Inverse Probability Censoring Weighting (IPCW) approach. While this thesis focuses on the g-formula, we do also demonstrate results for the IPCW. The majority of this chapter is based on previously existing methodology, however the method provided in the Young et al. paper was only applied to a small clinical trial dataset, and we shall apply it to a large exposure study. Taubman was the first to use the g-formula in an epidemiological¹ study, and there has only been a handful of epidemiological studies using the g-formula since (Taubman et al., 2009). Also the Young et al. paper, and most other g-formula research, uses the average treatment effect (ATE), whereas we use the average treatment effect in the treated (ATT). Therefore, the discussion of assumptions in our dataset and the application of the g-formula in this setting is original.

Here, we will focus on the g-formula estimator and the chapter is structured as follows.

¹As opposed to clinical.

- Review of causality and causation
- Introduction of the g-formula
- Review of Young et al. terminology
- Application of the g-formula to the incident cohort
- Application of the IPCW to the incident cohort

In the next chapter we will adjust this method to allow for the inclusion of prevalent cohorts. In doing this, we will explore a method to estimate an unbiased covariate distribution of prevalent datasets, and we will also explore a weighted g-formula to allow for g-formula estimation of prevalent cohorts.

7.2 Causal inference overview

Causal inference is a very large topic area which is growing in popularity, and this section only covers what is needed for this chapter. Please see Hernan & Robins (2020) for further detail. Often, the use of the letter $A = a$ represents treatment in a clinical trial, however we shall be using the term ‘exposure’, in our case SSc exposure (taken to be SSc diagnosis due to recording in the dataset).

In survival analysis we are often interested in assessing and quantifying an effect of exposure or treatment on an outcome, and in particular if an event/intervention will cause an outcome/result. It could be that we observe a change in the exposure and then see a change in the outcome. If we see an association or correlation between the two, it may be natural to think the two are causally related. If we incorrectly assume that it is solely the change in the exposure that leads to the change in outcome we would be assuming a causal relationship. However, it could be that a covariate leads to both an increase of exposure and outcome, and it is this factor which leads to a change in outcome. This incorrect assumption of the interaction between exposure, outcome and covariates can lead to an incorrect conclusion of causal effect.

Let us define notation we will use. Let Y be the indicator of event of interest, such that $Y = 1$ if we observe the event and $Y = 0$ if not. Let $Y^{a=1}$ be the patient's outcome had they been assigned treatment and $Y^{a=0}$ be the outcome had they not been assigned treatment. We recall that there is a *causal effect for an individual* if $Y^{a=1} \neq Y^{a=0}$. Often the goal is to find the average causal effect of exposed versus unexposed:

$$\psi = E(Y^{a=1}) - E(Y^{a=0}) = E(Y^{a=1} - Y^{a=0})$$

However, in most cases we shall only observe one of the outcomes. If exposed, we do not observe $Y^{a=0}$ and if unexposed we do not observe $Y^{a=1}$. The event we do not observe is termed the *counterfactual outcome*. Often, all we are able to observe is one outcome, $P[Y = y|A = a]$. This will give us information about a possible association, but not causation. Table 7.1 is an example of patient data, where on the left is what might be observed in an epidemiological study, and on the right is the consequences given an exposure, and only one of the outcomes is observed.

Patient ID	Observed		Unknown hypothetical	
	A	Y	$Y^{a=1}$	$Y^{a=0}$
No 1	1	1	1	1
No 2	1	0	0	0
No 3	1	1	1	1
No 4	0	1	1	1
No 5	0	1	1	1
No 6	0	0	1	0

Table 7.1: Example table of notation. In a realistic/clinical setting, A and Y are observable, however $Y^{a=1}$ and $Y^{a=0}$ are not.

In Table 7.1, $P[Y = 1|A = 1] = \frac{2}{3}$, and $P[Y = 1|A = 0] = \frac{2}{3}$, therefore the *associational risk difference* is $P[Y = 1|A = 1] - P[Y = 1|A = 0] = 0$, implying no association. However, on the impossible assumption we know the counterfactual,

we have $P[Y^{a=1} = 1] = \frac{5}{6}$ and $P[Y^{a=0} = 1] = \frac{4}{6}$, and the average causal risk difference is $\frac{1}{6}$. Therefore, we may be mistaken and assume there is no causal relationship when there is one. We have introduced *confounding* for the effect of A on Y . This is often due to lack of consideration for covariates, henceforth denoted Z .

Patients will often enter the study with differing characteristics which can cause an imbalance between groups, for example we may see more males in the exposed group compared to the unexposed. If not accounted for, these characteristics may distort the relationship between our exposure and outcome. Randomization between exposure groups allows for the concept of *exchangeability*, written symbolically as $Y^a \perp\!\!\!\perp A$ for all a , where we would expect to see the same outcome in both an exposed group and an unexposed group had the exposure been swapped between the groups. This allows for independence, $(Y^{a=1}, Y^{a=0}) \perp\!\!\!\perp A$. In clinical trials, the treatment could be given randomly to allow for exchangeability between groups. However, due to the nature of observational studies, this is often not the case, as SSc cannot be assigned so we cannot choose which patients to put into a ‘treated’ or ‘untreated’ group, or randomize who gets SSc. We know certain characteristics are more likely to lead to SSc more than others. We do have matching in our study, another method to account for covariates, however we note other studies may not have benefited from this. We revisit exchangeability and two other assumptions needed for our application later in the chapter.

We can use Directed Acyclic Graphs (DAGs) to graphically model causal effects in order to understand the relationships between the variables and infer methods to estimate causal effects. We plot the most basic DAG demonstrating the causation of exposure A on outcome Y (Figure 7.1). The arrow goes from the exposure to the outcome, demonstrating a causal relationship.

We remind ourselves of the definition of a confounder, a variable which has an effect on both the exposure and outcome (Figure 7.2). A classic example is that of the effect of birth order on the occurrence of Downs syndrome (Rothman, 2012).



Figure 7.1: Causal diagram 1, A is exposure, Y is outcome.

There is a clear association between Down's syndrome and birth order, with the risk increasing as birth order increases. However, there is a confounder present, which is the mother's age at the time of birth. The risk of a woman giving birth to a child with Down's syndrome increases with maternal age due to errors in meiosis which happen at increased frequency in older mothers. There is clearly a positive correlation between birth order and maternal age, as the second and subsequent births will occur at increasing maternal age. Thus it is maternal age which is the causal factor, with an association between maternal age and birth order. Once the age of the mother at the time of birth and birth order are analysed simultaneously, birth order no longer has a significant effect. Another example is alcohol consumption on mortality, where there is a large array of factors that could affect both alcohol consumption and mortality, such as age, sex, lifestyle and income. We wish to control for confounders, or condition on the variable (i.e., stratification or regression) in order to reduce the bias introduced.

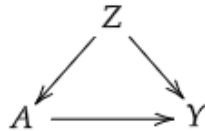


Figure 7.2: Causal diagram 2, A is exposure, Y is outcome, Z is confounder.

There is the possibility of colliders (Figure 7.3). These are variables which are affected by the exposure and the outcome. In such cases it is not appropriate to condition on variable Z . An example is where locomotor disease (exposure) and respiratory disease (outcome) may have a collider of hospitalization (Sackett, 1979). In the hospitalized sample there was an odds ratio of 4.06, indicating an association. However, in the general public the odds ratio was much lower at 1.06. Stratification based on hospitalization status led to the conclusion that there was

collider bias in the estimation of the effect of A on Y . We note that we most likely do not have colliders in our study, as our covariates occurred prior to (or close to) SSc diagnosis, however it is noted due to its importance.

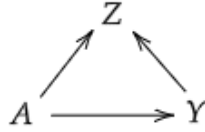


Figure 7.3: Causal diagram 3, A is exposure, Y is outcome, Z is the collider.

This theory on causality is a very brief introduction. Of interest may be the theory of d-separation and selection bias (Hernan & Robins, 2020), (Pearl, 2009).

7.3 G-formula methodology

The g-formula (or g-computation) is a type of g-method, also called Robin’s g-methods due to his derivations (Robins, 1986). The ‘g’ stands for generalized in that the methods are designed for application to generalized contrasts involving treatments. They provide consistent estimates of differences of exposure/treatment of average potential outcomes under a less restrictive set of identification conditions than standard regression methods. The average causal effect is a marginal effect because it averages over all individual-level effects in the population. It can be interpreted as a function of averages that would be observed if everybody in the population were exposed (or unexposed) at the same time.

The average treatment effect (ATE) is the risk difference which can be estimated using observational data using the following estimators:

$$ATE = \sum_z [E(Y|A = 1, Z = z) - E(Y|A = 0, Z = z)]P(Z = z) \quad (7.1)$$

The algorithm follows a series of steps:

- Step 1: Fit a flexible model for outcome Y on A and all relevant covariates, Z . For example, $E[Y^a|Z] = \beta_0 + \beta_1 A + \beta_2 Z + \beta_3 AZ$.
- Step 2: Create two (or more, depending on number of treatments) copies of the datasets, which we will term **pseudo-populations**, where each copy is be assigned a treatment, a . Then generate $\hat{E}(Y|A = a, Z = z)$ for their original covariates and assigned treatment.
- Step 3: On the basis of the results, we can estimate the counterfactual mean for n patients

$$\hat{E}(Y^a) = \frac{1}{n} \sum_{i=1}^n \hat{E}(Y|A = a, Z = z_i)$$

The estimate of the average treatment effect is the difference between the treated and untreated, $\hat{E}(Y^{a=1} = 1) - \hat{E}(Y^{a=0} = 1)$. Note that this does not include outcomes over time, and we shall expand on this to allow for time and competing risks.

From here, in order to estimate standard errors and construct confidence intervals, bootstrapping is recommended. To do this, steps 1 to 3 should be repeated with J bootstrapped samples with replacement of patients. Then the point estimates from step 3 can be used as parameter estimates.

When using g-methods, the three key assumptions of causal inference need to be considered: exchangeability, consistency and positivity. These are required in observational studies to treat the study as a conditionally randomized experiment, allowing for the use of counterfactual outcomes. We discuss them later in the chapter specifically for our framework.

7.3.1 Discrete time g-formula and Young et al's methodology

In this section we discuss the terminology used in Young et al's paper, which uses a causal framework to differentiate between hypothetical measures of risk in a competing risk setting. Let Y and D denote indicators of the event of interest and the competing event, respectively. The focus in our study is the two event system,

but this theory can be altered easily to include more events. We use discrete time-steps for each event occurring, $k = 0, \dots, K$ where K is selected to be the maximum follow-up time of interest. Each individual will have a counterfactual (or potential) outcome Y_{k+1}^a to indicate the event of interest occurring by time $k + 1$ if the individual is under treatment $A = a$. Let overbar notation denote the history of an event, $\bar{Y}_k = (Y_0, \dots, Y_k)$, and an underbar denote the future of the event, $\underline{Y}_{k+1} = (Y_{k+1}, \dots, Y_K)$.

We define the counterfactual risk of the event of interest by $k+1$ had all individuals in the population been assigned $A = a$ as

$$\Pr[Y_{k+1}^a = 1]$$

Therefore, the average causal effect at time $k + 1$ is

$$\Pr[Y_{k+1}^{a=1} = 1] - \Pr[Y_{k+1}^{a=0} = 1]$$

The discrete time hazard of the event of interest in interval $k + 1$ under a is

$$\Pr[Y_{k+1}^a = 1 | Y_k^a = 0]$$

When considering competing events, we have the risk of the event of interest defined as $\Pr[Y_{k+1}^a = 1]$, which is comparable to the cause-specific cumulative incidence function. We define the equivalent cause-specific hazard as $\Pr[Y_{k+1}^a = 1 | D_{k+1}^a = Y_k^a = 0]$ and the subdistribution hazard as $\Pr[Y_{k+1}^a = 1 | Y_k^a = 0]$. These are termed *total effects*, as we are quantifying the total impact of the exposure on outcome. Note we use the ordering of D_{k+1} as the event prior to Y_{k+1} as opposed to D_k as indicated in the causal DAG, Figure 7.4. For this calculation, we shall estimate the hazard at time-step k for Y and D for each patient using logistic regression with assigned treatment, A , time, k , and patient baseline covariates (sex, age at SSc diagnosis, and smoking status).

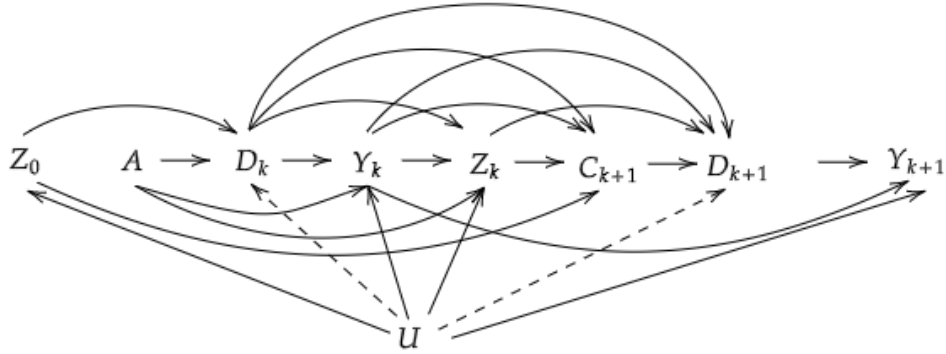


Figure 7.4: Causal DAG, as depicted by Young et al. The dashed arrows must be absent when the direct effect is being considered for exchangeability to hold (see below), but can be present when the total effect is being considered.

The marginal cumulative incidence from the previous chapter (the ‘naive Kaplan-Meier’) is comparable here to the cumulative risk under elimination of competing events, $\Pr[Y_{k+1}^{a, \bar{d}=\bar{0}} = 1]$, with corresponding ‘hazard under elimination of competing events’, $\Pr[Y_{k+1}^{a, \bar{d}=\bar{0}} = 1 | Y_k^{a, \bar{d}=\bar{0}} = 0]$, with $\bar{d} = \bar{0}$ suggesting that d does not occur at any previous time-step. This supposes that competing events had somehow been eliminated, which in practice is not feasible, for example death cannot be prevented. The differences in risk is referred to in Young et al’s work as *direct effects*, as they quantify the direct impact of the exposure on the outcome of interest, not mediated via the effect of the exposure on the competing event. We will estimate the hazard at time-step k for event Y for each patient based on logistic regression based on assigned treatment and covariates. For the direct effect, the hazard for the competing event shall be set to zero (see Equation (7.2)). This way of accounting for the competing event, setting the competing risk to 0 and computing the hazard of the event of interest dependent on covariates, is of particular interest in our study due to accounting for informative censoring better than the nonparametric Kaplan-Meier.

In Young et al., the authors make a specific definition of what they consider a

censoring event. In some literature, a competing risk event D is a form of censoring as it prevents the observation of Y . However, in this study it is the strict definition of administrative censoring (loss to follow-up), termed C_{k+1} where $C_0 \equiv 0$.

Under this notation, the direct effect is identified by the following function of the observed data (Young et al., 2020):

$$\begin{aligned} \Pr[Y_{K+1}^{a, \bar{c}=\bar{d}=0} = 1] &= \sum_{\bar{z}_k} \sum_{k=0}^K \Pr[Y_{k+1} = 1 | \bar{Z}_k = \bar{z}_k, \bar{Y}_k = \bar{C}_{k+1} = \bar{D}_{k+1} = 0, A = a] \\ &\times \prod_{j=0}^k (\Pr[Y_j = 0 | \bar{Z}_{j-1} = \bar{z}_{j-1}, \bar{C}_j = \bar{D}_j = \bar{Y}_{j-1} = 0, A = a]) \\ &\times f(z_j | \bar{z}_{j-1}, \bar{C}_j = \bar{D}_j = \bar{Y}_j = 0, a) \end{aligned}$$

and for the total effect,

$$\begin{aligned} \Pr[Y_{K+1}^{a, \bar{c}=\bar{0}} = 1] &= \sum_{\bar{z}_k} \sum_{k=0}^K \Pr[Y_{k+1} = 1 | \bar{Z}_k = \bar{z}_k, \bar{Y}_k = \bar{C}_{k+1} = \bar{D}_{k+1} = 0, A = a] \\ &\times \prod_{j=0}^k (\Pr[Y_j = 0 | \bar{Z}_{j-1} = \bar{z}_{j-1}, \bar{C}_j = \bar{D}_j = \bar{Y}_{j-1} = 0, A = a]) \\ &\times \Pr[D_{j+1} = 0 | \bar{Z}_j = z_j, \bar{C}_{j+1} = \bar{D}_j = \bar{Y}_j = 0, A = a] \\ &\times f(z_j | \bar{z}_{j-1}, \bar{C}_j = \bar{D}_j = \bar{Y}_j = 0, a) \end{aligned}$$

where $\Pr[Y_{k+1} = 1 | \bar{Z}_k = \bar{z}_k, \bar{C}_{k+1} = \bar{D}_{k+1} = \bar{Y}_k = 0, A = a]$ is the observed discrete-time hazard at $k + 1$ of the event of interest conditional on treatment and covariate history among those still free of the competing event and not lost to follow-up, which shall be found for each patient via logistic regression, and $f(z_j | \bar{z}_{j-1}, \bar{C}_j = \bar{D}_j = \bar{Y}_j = 0, a)$ is the conditional density of Z_j . For $j = 1$, $f(z_j | \bar{z}_{j-1}, \bar{C}_j = \bar{D}_j = \bar{Y}_j = 0, a) \equiv f(z_0)$.

As mentioned, when time-varying covariates are present, the above two equations can be estimated using Monte-Carlo simulation, however if it can be assumed

that exchangeability only needs to hold conditional on Z_0 , then this leads to much simpler risk estimators, Equation (7.2) and Equation (7.3)(Young et al., 2020). This holds for us, as we do not have time-varying covariates. Risk can then be estimated using the parametric g-formula estimator. For the direct effect this is

$$\frac{1}{n} \sum_{i=1}^n \sum_{k=0}^K p(a, z_{0i}, k; \hat{\theta}) \prod_{j=0}^{k-1} [1 - p(a, z_{0i}, j; \hat{\theta})], \quad (7.2)$$

And for the total effect

$$\frac{1}{n} \sum_{i=1}^n \sum_{k=0}^K p(a, z_{0i}, k; \hat{\theta}) [1 - q(a, z_{0i}, k; \hat{\eta})] \prod_{j=0}^{k-1} [1 - p(a, z_{0i}, j; \hat{\theta})] [1 - q(a, z_{0i}, j; \hat{\eta})], \quad (7.3)$$

Where z_{0i} are the baseline covariates for patient i , $p(a, z_0, k; \hat{\theta})$ and $q(a, z_0, k; \hat{\eta})$ are models for the observed event of interest and competing event hazards conditional at time k on $A = a$, $Z_0 = z_0$, respectively, achieved by regression models. Here, $\hat{\theta}$ and $\hat{\eta}$ are the consistent estimators of the index parameters θ and η , respectively.

7.3.2 Assumptions

We visit the three assumptions and their application for a) direct effect and b) total effects. For each $k = 0, \dots, K$, consider the following three identifying and untestable assumptions:

1. Exchangeability

Exchangeability occurs when the risk of the outcome in one group of patients would have been the same as the risk of the outcome in another group, had individuals in the first group received the treatment given to those in the second and vice versa. For the g-formula, we require conditional exchangeability, that is the counterfactual outcome is independent of the exposure, given the confounders Z of the data, $Y^a \perp\!\!\!\perp A | Z$. This (ideally) holds for randomized trials by

randomizing the covariates, and in observational studies this holds if, conditional on Z , there are no unmeasured confounders. This is a strong assumption for our study, which we cannot guarantee and are unable to test. We are only using three covariates, and it is possible we have unmeasured confounders. For example, we do not know the environmental exposures of each patient (e.g. we know SSc is believed to be correlated to silica exposure, see Section 2.2).

We specify this for our study. For direct effect this is:

$$\bar{Y}_{K+1}^{a, \bar{c}=\bar{d}=0} \perp\!\!\!\perp A | Z_0 \quad (7.4)$$

and

$$\bar{Y}_{k+1}^{a, \bar{c}=\bar{d}=0} \perp\!\!\!\perp (C_{k+1}, D_{k+1}) | \bar{Z}_k = \bar{z}_k, \bar{Y}_k = \bar{C}_k = \bar{D}_k = 0, A = a \quad (7.5)$$

The first consideration, (7.4), is that the counterfactual outcome and the actual treatment are independent given the measured baseline covariates of each patient, i.e., that the treated and the untreated groups would have experienced the same outcome of interest if they had received the same treatment level given differing baseline covariates. Our disease was not ‘given’, and while SSc has risk factors we are still not certain which covariates lead to SSc, therefore the relationship between who develops SSc and the outcome/censoring may be dependent on unmeasured confounders. Hence this exchangeability assumption, which is based solely on our measured confounders, may not be valid. For example, individual silica exposure is not recorded, which we know affects the development of SSc and both cancer and mortality (Calvert et al., 2003) (Steenland, 2005). The second condition is the independence between the counterfactual outcome and both forms of censoring in the case of direct effect (loss to follow-up and death) conditional on events at previous times steps, patient characteristics, and assigned treatment. Note that the above is applicable for time-varying covariates, Z_k , however we only have baseline covariates in our study, so only Z_0 needs to hold. The loss to follow-up

condition is less of a concern, as censoring due to anything other than the end of the study (or the last time the practice contributed data) is rare in the dataset. However, independence between the counterfactual outcome and the competing event based solely on our three covariates is unrealistic, and a strong assumption to make. In reference to the DAG, Figure 7.4, we note that exchangeability holds for this DAG in the absence of unblocked backdoor paths between:

- A and both Y_k and Y_{k+1} , conditional on Z_0 ,
- D_k and both Y_k and Y_{k+1} conditional on A and Z_0 ;
- C_{k+1} and Y_{k+1} conditional on Z_k, Y_k, D_k, A and Z_0 ;
- D_{k+1} and Y_{k+1} conditional on $C_{k+1}, Z_k, Y_k, D_k, A$ and Z_0 .

If these connections are unaccounted for in our dataset we lose exchangeability (if terminology of unblocked backdoor paths is not known, see d-separation theory, for example Fine Point 6.1, Hernan, 2020).

For the total effect, exchangeability holds if:

$$\bar{Y}_{K+1}^{a, \bar{c}=0} \perp\!\!\!\perp A | Z_0 \quad (7.6)$$

and

$$\bar{Y}_{k+1}^{a, \bar{c}=0} \perp\!\!\!\perp C_{k+1} | \bar{Z}_k = \bar{z}_k, \bar{D}_k = \bar{d}_k, \bar{Y}_k = \bar{c}_k = 0, A = a \quad (7.7)$$

Therefore, in addition to the observed exposure, at each follow-up time censoring due to loss to follow-up is independent of future counterfactual outcomes had everyone followed $A = a$ and censoring were eliminated. In reference to the DAG, Figure 7.4, we now have the possibility of additional lines from U to D_k and D_{k+1} , meaning that the presence of these relationships does not violate the conditional exchangeability assumptions needed for the total effect. Therefore, we only have two conditions which need to hold, unlike the four in the direct effect. We have (a) the condition that absence of any unblocked backdoor paths between A and

both Y_k and Y_{k+1} conditional on Z_0 , and we have (b) the absence of such paths between C_{k+1} and Y_{k+1} , conditional on Z_k, Y_k, D_k, A , and Z_0 . We no longer have to have the additional assumption of independence between outcome and competing event, as in the direct effect.

2. Positivity

For direct effect, we have

$$f_{A, \bar{Z}_k, D_k, C_k, Y_k}(a, \bar{z}_k, 0, 0, 0) \neq 0 \Rightarrow \\ \Pr[C_{k+1} = 0, D_{k+1} = 0 | \bar{Z}_k = \bar{z}_k, Y_k = C_k = D_k = 0, A = a] > 0$$

where $f_{A, \bar{Z}_k, D_k, C_k, Y_k}(a, \bar{z}_k, 0, 0, 0)$ is the joint density of $(A, \bar{Z}_k, D_k, C_k, Y_k)$ evaluated at $(a, \bar{z}_k, 0, 0, 0)$. Similarly, for the total effect

$$f_{A, \bar{Z}_k, \bar{D}_k, C_k, Y_k}(a, \bar{z}_k, \bar{d}_k, 0, 0) \neq 0 \Rightarrow \\ \Pr[C_{k+1} = 0 | \bar{Z}_k = \bar{z}_k, \bar{D}_k = \bar{d}_k, Y_k = C_k = 0, A = a] > 0$$

where $f_{A, \bar{Z}_k, \bar{D}_k, C_k, Y_k}(a, \bar{z}_k, \bar{d}_k, 0, 0)$ is the joint density of $(A, \bar{Z}_k, \bar{D}_k, C_k, Y_k)$ evaluated at $(a, \bar{z}_k, \bar{d}_k, 0, 0)$. For the direct effect this is the assumption that death or loss to follow-up is possible at each time-step, and for the total effect, this is the assumption that loss to follow-up is possible. The example given in the Young et al. paper is the study of two events for patients in hospital: prostate cancer death as the event of interest and other death as the competing event. In the direct effect, covariates could indicate that a patient is very healthy, implying the probability of other death is close to zero. However, in our study all patients have the possibility of loss to follow-up or death at every time point, therefore this assumption should hold. We also have sufficient sample size that all patients will have a non-zero probability of experiencing an event.

3. Consistency

As a general definition, consistency is that the values of treatment under observation correspond to well-defined interventions that, in turn, correspond to the treatment regimen in the data. If an individual has observed exposure $A = a$

then their observed outcome Y is equal to Y^a . An example for this would be if a study focused on the effect of ‘smoking’; some may interpret this as ‘cigarette smoking’ only, but others may include items such as pipes or e-cigarettes, and these have different effects on the outcome risk.

We discuss this for our study. For the direct effect, if $A = a$ and $\bar{C}_{k+1} = \bar{D}_{k+1} = 0$ then $\bar{Z}_{k+1} = \bar{Z}_{k+1}^{a, \bar{c}=\bar{d}}$ and $\bar{Y}_{k+1} = \bar{Y}_{k+1}^{a, \bar{c}=\bar{d}=0}$. For the total effect, if $A = a$ and $\bar{C}_{k+1} = 0$ then $\bar{Z}_{k+1} = \bar{Z}_{k+1}^{a, \bar{c}=0}$, $\bar{D}_{k+1} = \bar{D}_{k+1}^{a, \bar{c}=0}$ and $\bar{Y}_{k+1} = \bar{Y}_{k+1}^{a, \bar{c}=0}$. Often in studies, consistency concerns how exactly the treatment/outcomes have been defined, as demonstrated above with regard to the type of smoking. For us, the diseases, treatment, and covariates are well-defined, with yes or no variables, and therefore consistency holds for our dataset in terms of what we have observed and are observable outcomes. For our study, this imprecision is hopefully minimised, as there is a classification system in place for SSc diagnosis. However, we cannot eliminate the possibility of human errors in data collections due to a delay in SSc onset and date of diagnosis. There could be discrepancies for SSc, which is hard to diagnose due to wide ranging characteristics and rarity, however we feel this is a minor issue that should not affect the data.

However, issues arise with how we define the direct effect. As noted by Young et al for the direct effect, the consistency assumption requires well-defined interventions, but the elimination of death is an unspecified intervention. The occurrence of one event with the elimination of the other is an outcome which is not observable and is unlikely to ever be observable, yet we assume that it is possible for this hypothetical situation. Therefore, our risk results are not possible to observe in the real world, hence they are not well-defined. However, it can be argued that the risks under elimination of events is a better interpretation of the causal effect of SSc on cancer, and provides a better interpretation than including death, as risk of death will change over time, and is therefore necessary. Another method that does not rely on this untestable assumption may be more preferable. A very recent article by the same team as Young et al. (2020) uses the theory of separable effects, which does not have this issue (Stensrud et al.,

2021). This method also allows for a causal interpretation of competing events. We discuss this as a possibility for further work in Chapter 10.

7.3.3 Regression models to predict hazard

We wish to estimate $\Pr[Y_{k+1} = 1 | \bar{Z}_k = \bar{z}_k, \bar{Y}_k = \bar{C}_{k+1} = \bar{D}_{k+1} = 0, A = a]$ for each patient when estimating the direct effect, and both this and $\Pr[D_{k+1} = 0 | z_k, \bar{C}_{k+1} = \bar{D}_k = \bar{Y}_k = 0, A = a]$ for the total effect.

Let us look at a simpler case where time is not a covariate. Let Z be a vector of covariates Z_0, Z_1, \dots, Z_p , and A is treatment/exposure. Then $g\{E[Y|Z, A]\} = \sum_{i=1}^p \theta_i Z_i + \theta_{p+1} A$, where $g\{\cdot\}$ is a link function. Often the logit link is used, $\log\left\{\frac{E[Y|Z, A]}{1 - E[Y|Z, A]}\right\} = \sum_{i=1}^p \theta_i Z_i + \theta_{p+1} A$, which has the advantage that all predicted values will be greater than 0 and less than 1. An alternative method is to fit generalized additive models (GAM). This method replaces the linear combination of $\sum_{i=1}^p \theta_i Z_i + \theta_{p+1} A$ with the sum of the smooth function $\sum_{i=0}^p f_i(Z_i, A_i)$. The GAM allows more flexibility, however the simpler GLM may fit more efficiently. We shall test both of these for our data.

If we wished to include time, we would fit a logistic model to $\Pr[Y_{k+1} = 1 | \bar{Y}_k = 0, A, Z]$ such that, for example,

$$g\{\Pr[Y_{k+1} | \bar{Y}_k = 0, A, Z]\} = \theta_0 + \theta_1 k + \theta_2 A + \theta_3 Z$$

where $k = 1, \dots, K$ is linear time, and K is the maximum time under study. We could include more time variables, such as k^2 or k^3 . Also, if we had covariates or a treatment effect that is time dependent, these could be included in the model with terms such as $\theta_4 k \times A$, for example.

7.4 Application of g-formula to incident cohort

To calculate the g-formula for our data we need to model the cause-specific hazards for cancer and death prior to cancer. The conditional probability of the event cancer occurring at time $k + 1$ is $\Pr[Y_{k+1} = 1 | \bar{C}_{k+1} = \bar{D}_{k+1} = \bar{Y}_k = 0, A = a, Z_0]$. For the total effect where cancer is the event of interest we also require $\Pr[D_{k+1} = 1 | \bar{C}_{k+1} = \bar{D}_k = \bar{Y}_k = 0, A = a, Z_0]$ to be estimated. Note the use of Z_0 , as we do not have time-varying covariates.

We propose either a binomial GLM (generalized logistic model) or a GAM (generalized additive model), and perform selection via minimising AIC. There are two things to consider here: which model to use and which covariates to include. Time is required in the model to find the hazard at each time-step, as well as SSc type. As a reminder, time for us is from the point of SSc diagnosis onward. The decision was made to include age at SSc diagnosis, sex and smoking as covariates. This is as the other three covariates (BMI, alcohol intake and location) did not appear to be statistically significant when estimating cancer or death with the GLM. Location is based on GP practice and is available for all patients, but it is not a good indicator of cancer or death. This has been shown by proportional hazards models in Chapters 4, 5 and 6, and by minimum AIC while optimizing GLMs in this chapter (not shown here). The inclusion of alcohol and BMI status would require the removal of approximately 25% of original patients and would thus lose too many patients from our analysis.

We include time as possibly a linear, quadratic and cubic term in the GLM, i.e. $\theta_1 k, \theta_2 k^2, \theta_3 k^3$. Here k is time after SSc diagnosis in months, where the patient's outcome time is rounded up to the nearest month. The model for the GLM is presented later in the chapter. If a GAM were to be used, shown by Equation (7.8), it would allow flexibility of time and age at SSc diagnosis, as the prediction would be based on the model.

$$\begin{aligned}
&g\{Pr[Y_{k+1} = 1 | \bar{C}_{k+1} = \bar{D}_{k+1} = \bar{Y}_k = 0, A = a, Z_0]\} \\
&= \theta_0 + f_k(k) + \theta_1 \times A + \theta_2 \times Male + f_{Age}(Age) \\
&+ \theta_3 \times Smoker + \theta_4 \times Exsmoker
\end{aligned} \tag{7.8}$$

7.4.1 Results of model-fitting

We use AIC to compare different regression models. When using a model with only the above variables (i.e. k but no higher order power terms for time, and no interaction terms), the GAM, which allowed more flexibility for factors k and Age , showed no improvement on the GLM (AIC=8397 for GLM, AIC=8398 for GAM). Therefore, we shall use a GLM. For death, the AIC for the GAM was actually slightly smaller (GLM AIC=5548 to the GAM's AIC=5547), however it has been decided to keep the GLM for consistency between models in this thesis as there was only a small difference.

After establishing we will be using a GLM, we also consider interacting terms and higher order terms where minimum AIC determined their inclusion. A mixture of sex, age at SSc diagnosis, smoking and time were considered, as well as higher order terms. Each extra term was trialled one variable at a time, and was considered for inclusion if the AIC was minimised. Then all these terms were added to the model, and a backwards stepwise regression was then used to remove extra terms. However, even after this a few terms were removed. For cancer, $k * Age^2$ was removed due to it neither being statistically significant (or close to statistically significant) and a high order of magnitude (e^{-5}). Similarly for death, $Male * Age^2$ was neither significant and nor did it have a high magnitude (e^{-5}).

In summary we therefore use the following GLM models:

$$\begin{aligned}
g\{\Pr[Y_{k+1} = 1 | \bar{C}_{k+1} = \bar{D}_{k+1} = \bar{Y}_k = 0, A = a, Z_0]\} = \\
\theta_0 + \theta_1 \times A + \theta_2 \times k + \theta_3 \times Male + \theta_4 \times Age + \theta_5 \times Age^2 \\
+ \theta_6 \times Smoker + \theta_7 \times Exsmoker + \theta_8 \times Age \times k + \theta_9 \times Age \times Male
\end{aligned} \tag{7.9}$$

$$\begin{aligned}
g\{\Pr[D_{k+1} = 1 | \bar{C}_{k+1} = \bar{D}_k = \bar{Y}_k = 0, A = a, Z_0]\} = \\
= \phi_0 + \phi_1 \times A + \phi_2 \times k + \phi_3 \times Male + \phi_4 \times Age + \phi_5 \times Age^2 + \\
\phi_6 \times Smoker + \phi_7 \times Exsmoker + \phi_8 \times k \times Age
\end{aligned} \tag{7.10}$$

The inclusion of these extra interaction terms reduces the AIC to 8387 for cancer and to 5539 for death.

After the exclusion of patients who do not have smoking information, we have n=5,493 incident patients, 803 SSc and 4,690 non-SSc; the GLMs are fitted to these data. The coefficient values using the logistic regression of cancer are reported in Table 7.2. The hazard appears to increase with time. Having SSc ($A = 1$) increases the hazard too, especially for death. Being a smoker increases the hazard of both cancer and death, and being an ex-smoker increases the hazard of death. Age is difficult to discern due to interaction terms, but investigation shows both hazards increase with age, as expected (not shown here).

7.4.2 Selecting a pseudo-population

As previously stated, the g-formula method is most commonly used to estimate the average treatment effect (ATE). G-formula methods are often used when a cohort of patients is representative of the baseline population of interest or a randomly selected clinical trial whose covariates represent the population of interest. It is this property that allows marginalisation over the original population of interest,

	Cancer	[95% Confidence interval]	Death	[95% Confidence interval]
Intercept	-12.208	[-14.594, -10.039]	-10.534	[-13.682, -7.822]
k (month)	1.641e-2	[6.038e-3, 2.674e-2]	-1.092e-2	[-2.564e-2, 3.444e-2]
SSc (A=1)	0.257	[0.025, 0.478]	1.283	[1.040, 1.519]
Male	-1.710	[-3.097, -3.798]	0.210	[-0.124, 0.524]
Age	0.119	[0.046, 0.184]	-2.587e-2	[-1.044e-1, 6.379e-2]
Age ²	-4.445e-4	[-9.901e-4, 6.036e-5]	9.431e-4	[2.966e-4, 1.520e-3]
Smoker	0.293	[0.075, 0.505]	1.008	[0.751, 1.260]
Ex-Smoker	0.159	[-0.040, 0.355]	0.364	[0.119, 0.604]
$k \times$ Age	-2.017e-4	[-3.676e-4, 3.660e-5]	2.851e-4	[7.774e-5, 4.957e-4]
Male \times A	-	-	0.428	[0.132, 0.963]
Male \times Age	3.028e-e2	[9.729e-3, 5.127e-2]	-	-

Table 7.2: Coefficient values for logistic regression on the incident cohort. The baseline comparator for sex is female, for A is non-SSc (A=0), and the baseline for smoking is 'non-smoker'. The 95% confidence intervals are in brackets.

where a pseudo-population is created from the treated and untreated. The ATE is the difference between the pseudo-population hypothetically all being assigned the exposure/treatment or all not being assigned the exposure/treatment. This is particularly useful for clinical trials when the covariate distributions might differ between treated and non-treated. By combining the treated and non-treated groups the pseudo-population is a better representation of the population of interest.

While most studies investigate the ATE, we have the option of quantifying the ATT, the average treatment effects on the treated. The ATT is used to estimate the difference if the treated had not been treated. The ATU (the average treatment effects on the untreated) is also used to decide if intervention would have an impact on a group, as it estimates the average effect if the non-treated/non-exposed had been treated or exposed, so the non-exposed is the pseudo-population.

Some examples of studies which have used the g-formula are as follows:

- Taubman et al. (2009) performed a comprehensive study discussing the benefits of different interventions to reduce coronary heart disease. The

data is taken from the Nurses Health Study, a well-known and large study comprising 121,701 nurses (although only 78,746 were used in the study). The dataset contained time-varying covariates, and therefore a Monte Carlo sampling scheme was used². Prior to sampling time-varying covariates and outcomes, the pseudo-population was first derived by using baseline covariates from the whole, combined dataset, regardless of whether the patients in the original dataset had the exposure or not. This is an example of a cohort study that contains a large sample of patients of the specified population of interest (nurses) and is therefore representative of the larger population.

- Keil et al. (2014) investigated the implication of a time-varying exposure of a hypothetical drug that prevents graft-versus-host disease in bone marrow transplant patients. They were interested in modelling time-varying risk. Therefore, they estimated the probability of the status of the patient at each time-step (normal platelet levels, relapse, death, censored) by logistic regression. Again, Monte Carlo sampling was used, where $n=137,000$ patients were sampled with replacement from 137 patients in the original dataset, with several baseline covariates. Outcomes at each time-step were simulated with probabilities based on the earlier regression models to create time-varying covariates. This demonstrates the use of a dataset which is a good representation of the target population, those who had bone marrow transplants.
- In the aforementioned Young et al. paper, their application is focused on a clinical trial without time-varying covariates for a follow-up of 60 months, estimating the average causal effect of oestrogen therapy on prostate cancer risk (with other deaths being competing events). The lack of time-varying coefficients means that Monte-Carlo sampling is not required. They combine the treated and placebo group together to form one pseudo-population.

²Monte-Carlo sampling is often used for modelling the g-formula when time-varying covariates are present to create a pseudo-population, however as we do not have time-varying covariates we do not discuss this in more depth.

They take two copies of this dataset and then set the covariate to $A = 1$ in one, and $A = 0$ in the other, and they predict risk based on the assigned treatment. Therefore, the population of interest is the combined dataset, as they are estimating the average causal risk. Although not stated in their article, the reasoning for this is the assumption that treatment was given randomly, regardless of covariates. This allows for exchangeability, and therefore the combined pseudo-population with the treated and untreated combined covariate distributions is the best representation of the population of interest, namely those with prostate cancer.

The above are all examples of ATEs. The populations of interest (nurses, bone marrow transplant patients, prostate cancer patients) are well defined, and the pseudo-population is created out of those who have been treated and not treated to allow for exchangeability between groups. This allows for a quantification of risk between the two methods.

However, for our study we question what our population of interest is. An incorrect assumption would be to estimate the ATE by combining our SSc and non-SSc cohort together to estimate an ATE of patients in our dataset, as would be done with a clinical trials dataset. Our dataset of matched patients is not a ‘target population’, but a dataset formed such that SSc patients have 6 matches each based on sex, age, and location to reduce confounding. It is neither reflective of the UK population or those who are most at risk of getting SSc.

One hypothesis might be to use a pseudo-population which is reflective of the covariates found in the general population. Then the difference between the predicted risks, where the population had been assigned SSc or not, may be of interest if we wished to quantify what the difference would be had the entire population had SSc or not. However, we believe this would not be appropriate, as we know some patients are more likely to have SSc than others, therefore this target population is not representative of the risk we would observe in the tangible world. Not only that, but our dataset is a matched cohort-comparator study, and therefore we do not have a dataset to represent the general population.

The ATT is a more logical pseudo-population for this research. The ATT would provide the difference between the risk experienced by SSc patients versus their counterfactual risk of not having SSc. This is a more intuitive interpretation for the impact that SSc will have on cancer (and death without cancer), and it is quantifying the difference had the SSc group not been exposed to SSc. This is taking the SSc cohort to be the population of interest. The pseudo-population we will use is therefore the SSc group and the SSc group's baseline covariates. Then the expectation $E[Y^{a=1}|A = 1]$ is the risk of SSc estimated using the SSc cohort with their original exposure, and $E[Y^{a=0}|A = 1]$ is the hypothetical intervention where we set $A = 0$ to all patients to estimate the counterfactual risk. The SSc group is therefore the target population. For example, the ATE would be Equation (7.1), however for the ATT it would be Equation (7.11),

$$ATT = E(Y^{a=1} - Y^{a=0} | A = 1) = E(Y | A = 1) - \sum_z E(Y | A = 0, Z = z) P(Z = z | A = 1) \quad (7.11)$$

Although we are only using the SSc cohort's covariate distribution, we continue to follow the exact methodology as above, and as described in the Young at al paper. However, while we perform the regression model fitted on both the SSc and non-SSc patients, these models will only be applied to model the risk of the SSc pseudo-population such that:

- they had been assigned their original exposure, $a = 1$, or
- they had been assigned their counterfactual exposure, $a = 0$.

Therefore, the risk ratio is the risk we observe in the SSc patients over the hypothetical risk that would be expected if the SSc patients had never had SSc. We rewrite the risks of interest as follows:

The direct risk of cancer at $k + 1$

$$\Pr[Y_{k+1}^{a, \bar{d}=\bar{0}} = 1 | A = 1]$$

The direct risk of death at $k + 1$

$$\Pr[D_{k+1}^{a, \bar{y}=\bar{0}} = 1 | A = 1]$$

The total risk of cancer at $k + 1$

$$\Pr[Y_{k+1}^a = 1 | A = 1]$$

The total risk of death at $k + 1$

$$\Pr[D_{k+1}^a = 1 | A = 1]$$

One reason for this choice is that it allows a comparison for SSc patients, as we are now finding the treatment effect between an SSc population and what their risk would have been if they did not have SSc, allowing for a causal interpretation of SSc. Another reason is that, due to the way the dataset was matched based on SSc patients, it will also help in the methodology of later chapters when we use IPW methods to correct for calendar time. However, due to it being a matched study, we note that the difference between using the SSc set as the pseudo-population or the whole dataset (SSc and non-SSc) would be very similar, as of the three baseline covariates used it is only smoking habits that differ.

At the time of this study, we have found no mention of the g-formula being applied to a matched or semi-matched study, i.e., where selection into the study is not random. Therefore, we suspect the above is a novel idea. However, hypothetically it may be that if there was such a study perhaps propensity score methods may have been used to correct for differing baseline covariates.

We note that for these results to be applicable to SSc patients, we need to assume that our SSc cohort is representative of all SSc patients in the UK (for example, we would not want more females in the study than the proportion in the SSc UK population). However, because as many SSc patients in the UK were recruited as possible and there was limited bias in the selection of patients, we believe our

cohort is an accurate representation of SSc in the UK.

7.4.3 Results of g-formula application to the incident cohort.

The risk estimates are calculated for the incident SSc data for their original exposure ($a = 1$) and their counterfactual ($a = 0$), and the cumulative incidence curves are shown in Figure 7.5 for the direct effect based on Equation (7.3.2), with cancer risk on the left and death risk on the right. Note that in both of these, the opposing event is treated as eliminated (death is eliminated for cancer, and cancer is eliminated for death). Time is calculated at intervals of a month, with maximum follow-up time of $K = 240$. We also include the NPMLE from the previous chapter (one minus the naive Kaplan-Meier) for comparison.

Figure 7.5 illustrates that the 1-KM and g-formula curves are similar. To some extent this is to be expected, as our study has matched data, whereas other studies with larger differences in covariate distributions between the exposed and unexposed may observe larger differences. The g-formula direct risk is slightly higher than the NPMLE curve, which on further inspection (verification not shown here) is due to the inclusion of age at SSc diagnosis in the model. Once this covariate is removed from the model the two curves appear very similar (again, not shown here). As we have a matched dataset, this is not due to a correction of confounding for age at SSc diagnosis alone. It is likely due to the way that censoring is now accounted for. The g-formula treats informative censoring differently from the NPMLE, as in the g-formula patients are simulated beyond the times when they were, in fact, censored (from competing event or loss to follow-up). It may be that because those who are older are more likely to get cancer, they are also could be more likely to die (or be censored from loss to follow-up), creating dependence based on covariates which need to be accounted for. Alternatively, it could be that due to the low number of patients still alive in the sample at later survival times, who may carry more weight when fitting the GLM, we have not fitted the GLM optimally at these later times.

There is the possibility of extrapolating the estimation beyond the maximum

event time to later times in the dataset, however this is extrapolating beyond the data available and is not advisable.

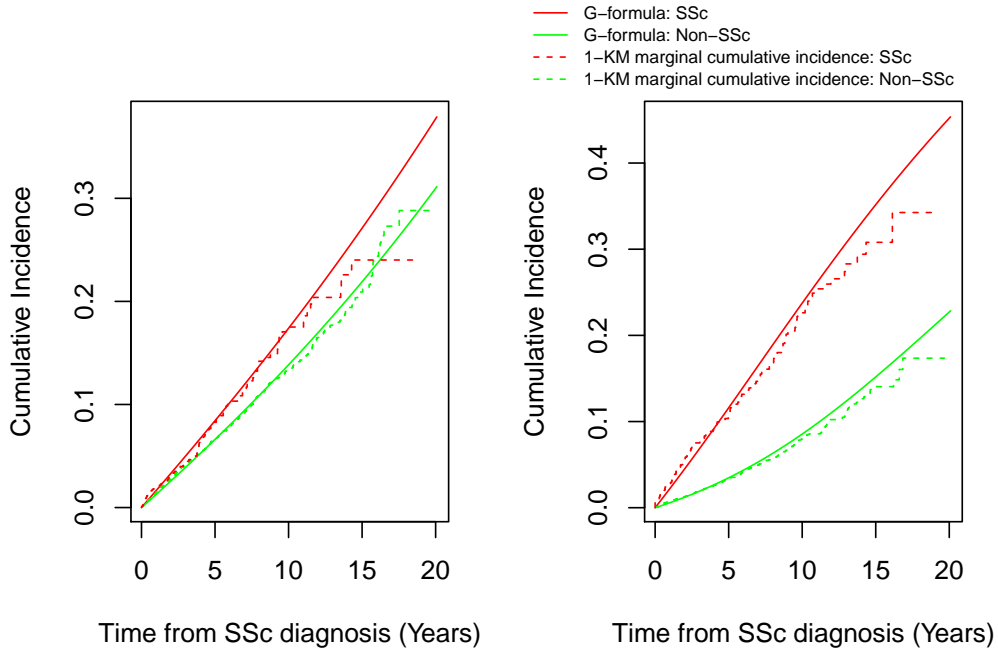


Figure 7.5: G-formula risk estimation for risk of cancer (left) and death (right), direct effect (elimination of competing events), with the regression and prediction based on incident cohort only. The dashed lines are the incident NPMLE 1-Kaplan-Meier estimation.

Figure 7.6 presents the risk without elimination of competing events based on Equation 7.2, along with the NPMLE cause-specific cumulative incidence curves of the previous chapter. While there is a small difference in the larger times for the SSc group for death without cancer, the NPMLE method and the g-formula estimator give very similar results. The fact that the gap between the two methods here is smaller than the gap for the direct effect again suggests that the difference in the direct effect is due to the method of accounting for informative censoring, where ‘censoring’ is due to the competing event. It is reassuring these are similar, as we neither expect to account for informative censoring nor greatly differing covariate distributions between exposure groups, which would be reasons for differences between the curves.

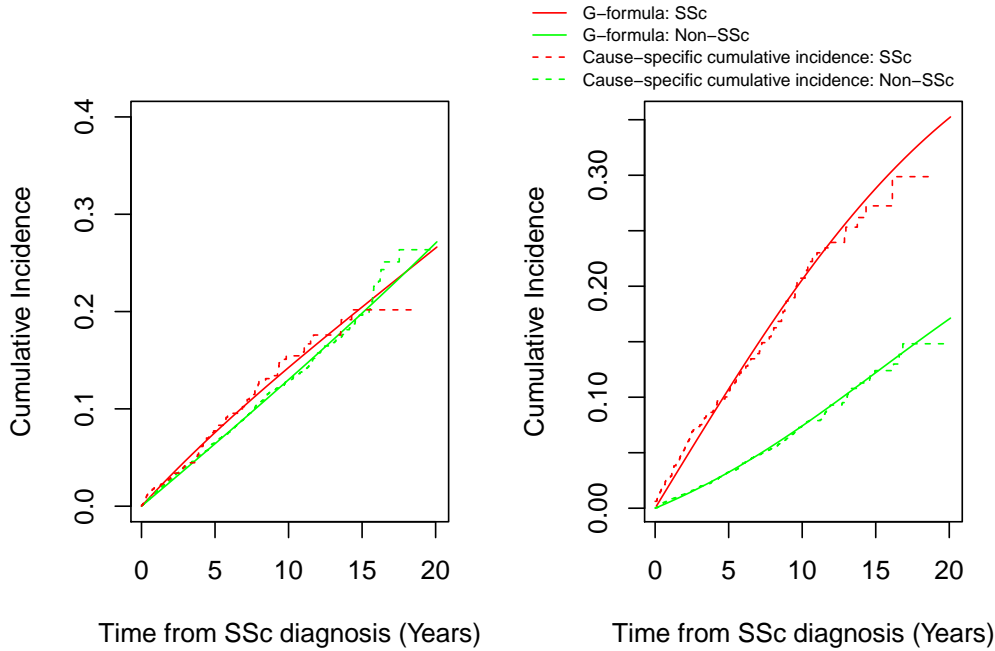


Figure 7.6: G-formula risk estimation for risk of cancer (left) and death (right), total effect (without elimination of competing events), with the regression and prediction based on incident cohort only. The dashed lines are the incident NPMLE Aalen-Johansen estimation.

The risk ratios and confidence intervals are shown in Table 7.3 for the direct effect on cancer and Table 7.4 for the total effect on cancer. Confidence intervals are estimated via bootstrapping, by selecting patients with replacement to form a bootstrapped dataset of the same size as the original dataset, and simulating 500 point estimates. We sample SSc patients and include their corresponding matches accordingly so that it continues to be a matched study. As above, the logistic regression, after being fitted to both SSc and non-SSc patients, is applied to the SSc pseudo-population.

For the confidence interval of the cumulative incidence curves, we do not wish to assume that $\hat{F}_k(t)$ is normally distributed (where t is time in years), as often this will result in a skew if $\hat{F}_k(t)$ is close to 0 or 1. Therefore, we use a log-log transform, which transforms $\hat{F}_k(t)$ onto a $(-\infty, \infty)$ scale. We take 500 bootstrap samples such that we have B point estimates of risk, $\bar{F}_{k,b}(t)$, where $b = 1, \dots, B$,

and k here represents the possible presence of competing risks. The $100(1 - \alpha)\%$ confidence intervals are as follows

$$\left([\hat{F}_k(t)]^{\exp(z_{(1-\alpha)/2} \sqrt{\text{Var}(\log(-\log(\bar{F}_{b,k}(t))))}), [\hat{F}_k(t)]^{\exp(-z_{(1-\alpha)/2} \sqrt{\text{Var}(\log(-\log(\bar{F}_{b,k}(t))))})} \right) \quad (7.12)$$

where $\hat{F}_k(t)$ is the cumulative incidence estimate of the original data.

Similarly, the risk ratio can take values between $(0, \infty)$, so a log transform is appropriate here. We take 500 bootstrapped risk ratios, $\bar{R}R_{k,b}(t)$ for $b = 1, \dots, B$. The confidence intervals for the risk ratio are also as follows:

$$\left([\hat{R}R_k(t)]^{\exp(-z_{(1-\alpha)/2} \sqrt{\text{Var}(\log(\bar{R}R_{b,k}(t))))}, [\hat{R}R_k(t)]^{\exp(z_{(1-\alpha)/2} \sqrt{\text{Var}(\log(\bar{R}R_{b,k}(t))))} \right) \quad (7.13)$$

where $\hat{R}R_k(t)$ is the cumulative incidence estimate of the original data.

Table 7.3 shows the risk ratios for the direct effect and Table 7.4 shows the risk ratios for the total effect, along with the corresponding NPMLEs from the previous chapter. The numbers are slightly different for NPMLE to the last chapter as, for consistency with the g-formula, we remove patients without smoking information here. The g-formula and NPMLE risk ratios are similar in that the confidence intervals are quite large and the risk ratios are within 0.1 of each other. For the total effect the confidence intervals contain 1, however the new confidence intervals for the direct effect no longer contain 1. Therefore, there is the suggestion of a causal impact on cancer. There is a large difference at time 10 between the NPMLE and g-formula, due to the NPMLE having a large jump in risk at this time compared to the smoothed parametric version of the g-formula.

Direct effect		
Time from SSc diagnosis (years)	NPMLE [95% CI]	G-formula [95% CI]
5	SSc 0.083 [0.061, 0.104] Non-SSc 0.066 [0.058, 0.074] Ratio 1.258 [0.940, 1.684]	SSc 0.084 [0.067, 0.104] Non-SSc 0.066 [0.059, 0.074] Ratio 1.273 [1.020, 1.590]
10	SSc 0.175 [0.136, 0.212] Non-SSc 0.135 [0.121, 0.148] Ratio 1.302 [1.025, 1.652]	SSc 0.174 [0.141, 0.209] Non-SSc 0.138 [0.126, 0.151] Ratio 1.261 [1.025, 1.539]
15	SSc 0.24 [0.179, 0.297] Non-SSc 0.209 [0.186, 0.232] Ratio 1.148 [0.878, 1.50]	SSc 0.271 [0.222, 0.323] Non-SSc 0.219 [0.199, 0.240] Ratio 1.237 [1.029, 1.488]

Table 7.3: Cumulative incidence and risk ratio for the outcome of cancer for the marginal NPMLE (1-KM) and direct effect g-formula methods using solely the incident data. The 95% confidence intervals are given in brackets.

Total effect		
Time from SSc diagnosis (years)	NPMLE [95% CI]	G-formula [95% CI]
5	SSc 0.077 [0.059, 0.101] Non-SSc 0.065 [0.057, 0.073] Ratio 1.196 [0.895, 1.599]	SSc 0.076 [0.060, 0.093] Non-SSc 0.064 [0.057, 0.072] Ratio 1.188 [0.940, 1.470]
10	SSc 0.155 [0.125, 0.191] Non-SSc 0.130 [0.118, 0.143] Ratio 1.191 [0.942, 1.505]	SSc 0.142 [0.116, 0.172] Non-SSc 0.130 [0.119, 0.142] Ratio 1.092 [0.887, 1.354]
15	SSc 0.202 [0.161, 0.253] Non-SSc 0.196 [0.176, 0.219] Ratio 1.028 [0.799, 1.322]	SSc 0.205 [0.167, 0.245] Non-SSc 0.198 [0.180, 0.217] Ratio 1.035 [0.840, 1.260]

Table 7.4: Cumulative incidence and risk ratio for the outcome of cancer for the cause-specific NPMLE and total effect g-formula methods using solely the incident data. The 95% confidence intervals are given in brackets.

7.5 Alternative to the g-formula: IPCW

This section discusses and applies the alternative method to the g-formula provided by Young et al. (2020), the inverse probability censoring weighting (IPCW) method. The IPCW is another method provided by Young et al. which is a weighted Kaplan-Meier, where patients are given a greater weight if they are more likely to be censored, hence accounting for dependent censoring. The IPCW was first proposed by Robins & Rotnitzky (1992), where the hazard of being censored was modelled using the Cox model. There was a subsequent article detailing IPCW log-rank tests (Robins & Finkelstein, 2000). Satten & Datta (2001) furthered this work, where the probability of censoring is estimated using an additive hazards model. Datta & Satten (2002) also placed IPCW in the context of a multistage system.

Note that the definition for ‘censoring’ here includes both loss to follow-up and the competing event when the direct effect is being considered.

The IPCW direct effect estimate is as follows (Young et al., 2020),

$$\sum_{k=0}^K \hat{h}_k(a; \hat{\alpha}, \hat{\eta}) \prod_{j=0}^{k-1} \left[1 - \hat{h}_j(a; \hat{\alpha}, \hat{\eta}) \right] \quad (7.14)$$

where

$$\hat{h}_k(a; \hat{\alpha}, \hat{\eta}) = \frac{\sum_{i=1}^n y_{k+1i} (1 - y_{ki}) w_{ki}(\hat{\alpha}, \hat{\eta}) I(a_i = a)}{\sum_{i=1}^n (1 - y_{ki}) w_{ki}(\hat{\alpha}, \hat{\eta}) I(a_i = a)} \quad (7.15)$$

and

$$w_{ki}(\hat{\alpha}, \hat{\eta}) = \frac{\prod_{j=0}^k (1 - c_{j+1i})(1 - d_{j+1i})}{\prod_{j=0}^k (1 - r(a, \bar{z}_{ji}, j; \hat{\alpha}))(1 - q(a, \bar{z}_{ji}, j; \hat{\eta}))} \quad (7.16)$$

where $r(a, \bar{z}_{ji}, j; \hat{\alpha})$ and $q(a, \bar{z}_{ji}, j; \hat{\eta})$ are models for the hazard of loss to follow-up and death, respectively, indexed by parameters. The hazard for the cause-specific loss to follow-up is $\Pr[C_{j+1} = 1 | \bar{Z}_j = \bar{z}_{ji}, \bar{C}_j = \bar{D}_j = \bar{Y}_j = 0, A = a]$

and competing event hazards, $\Pr[D_{j+1} = 1 | \bar{Z}_j = \bar{z}_{ji}, \bar{C}_{j+1} = \bar{D}_j = \bar{Y}_k = 0, A = a]$, with $\hat{\alpha}$ and $\hat{\eta}$ as consistent estimators of α and η , respectively, with notation $(a_i, \bar{z}_{Ki}, \bar{c}_{K+1i}, \bar{d}_{K+1i}, \bar{y}_{K+1i})$ as individual i 's values of $(A, \bar{Z}_K, \bar{C}_{K+1}, \bar{D}_{K+1}, \bar{Y}_{K+1})$.

Similarly, the following is the risk for the total effect, with estimates based on the observed cause-specific hazards of the event of interest and the competing event³,

$$\sum_{k=0}^K \hat{h}_k^1(a; \hat{\alpha})(1 - \hat{h}_k^2(a; \hat{\alpha})) \prod_{j=0}^{k-1} [(1 - \hat{h}_j^1(a; \hat{\alpha}))(1 - \hat{h}_j^2(a; \hat{\alpha}))] \quad (7.17)$$

$$\hat{h}_k^1(a; \hat{\alpha}) = \frac{\sum_{i=1}^n y_{k+1i}(1 - y_{ki})(1 - d_{ki})w_{ki}(\hat{\alpha})I(a_i = a)}{\sum_{i=1}^n (1 - y_{ki})(1 - d_{ki})w_{ki}(\hat{\alpha})I(a_i = a)} \quad (7.18)$$

$$\hat{h}_k^2(a; \hat{\alpha}) = \frac{\sum_{i=1}^n d_{k+1i}(1 - y_{ki})(1 - d_{ki})w_{ki}(\hat{\alpha})I(a_i = a)}{\sum_{i=1}^n (1 - y_{ki})(1 - d_{ki})w_{ki}(\hat{\alpha})I(a_i = a)} \quad (7.19)$$

and

$$w_{ki}(\hat{\alpha}) = \frac{\prod_{j=0}^k (1 - c_{j+1i})}{\prod_{j=0}^k (1 - r(a, \bar{z}_{ij}, j; \hat{\alpha}))} \quad (7.20)$$

If there was no informative censoring then these models would be equivalent to the NPMLE models of the last chapter. The IPCW method has a few advantages over the g-formula. It is a more intuitive comparison between the now censoring-weighted models and the unweighted NPMLE, which can demonstrate the differences in censoring (a difference would imply informative censoring). If we believe we do not have informative loss to follow-up, then we do not need to weight for $r(\cdot)$. Also, as in general with the IPCW method, we can stabilise the weights for more accurate results (Hernan et al., 2000). This will reduce the variability when a few patients are given too higher weightings. The stabilizer can be added to the numerator of the weight.

³An alternative version using the subdistribution hazard form is discussed in Young but not discussed here.

There are a few downsides. The first is the redundant inclusion of modelling loss-to-follow up in our study. We have two reasons for loss to follow-up, either leaving the dataset due to emigration, which we expect to be rare and mostly non-informative, or the end of the study, which is non-informative. We unfortunately have no way of detecting which of these types applies to a particular patient. It would still be of benefit to use the IPCW to weight for the competing event for the direct effect, therefore we shall do so in this section. The g-formula has allowed for the ATT analysis we have been using as it has allowed for the use of a pseudo-dataset, which might have been an advantage over the g-formula if we did not have a matched study. The g-formula adjusts for confounding when covariates differ between the groups, which the IPCW does not do without additional adjustment. Lastly, as we shall see in later chapters, using the g-formula has allowed more flexibility when considering differences between recent cases and historic ones (temporal trends).

We again consider whether to use GAM or GLM for our model in order to model censoring. If we were to not include interaction terms, the AIC for the GLM model is 49478.12 and the AIC for GAM is 49444.06. Therefore we shall use a GAM to model censoring, but continue to use GLM for modelling the outcome death. When considering interaction terms, the only interaction that provided smaller AIC was the interaction of time, k , and age at SSc diagnosis, and therefore this is included, reducing the AIC to 49430.01

Table 7.5 shows the results for the parametric coefficients from modelling loss to follow-up, C , with a GAM. Time is significant at the 5% level, and higher orders of time were tested but did not minimise the AIC. Most covariates show no significance when modelling loss to follow-up: non-SSc patients and males are more likely to be lost although neither at a statistically significant level. We also require the hazard of the competing events when estimating the direct effect, and these remain unchanged from Section 7.4.1.

Figure 7.7 shows how the smoothing function changes in relation to the variable. We observe that the hazard of censoring increases over calendar time. There is

	Loss to follow-up	[95% Confidence interval]
Intercept	-4.7481	[-4.7948, -4.7014]
SSc (A=1)	-0.0508	[-1.395, 0.0379]
Male	0.0588	[-0.0222, 0.1397]
Smoker	0.0352	[-0.0375, 0.1078]
Ex-Smoker	0.0121	[-0.0650, 0.08909]

Table 7.5: Coefficient values for censoring logistic regression, estimated using the incident cohort only. The baseline for A is non-SSc, the baseline for sex is male and the baseline for smoking and ex-smoker is non-smoker.

	edf	Ref.df	Chi.sq	p-value
s(k)	8.611	8.922	193.70	< 2e-16
s(Age)	7.070	8.129	21.84	0.00749
s(k*Age)	7.603	8.498	61.28	6.75e-10

Table 7.6: Approximate significance of smooth terms, estimated using the incident cohort only.

a drastic increase in the hazard of censoring at higher ages and small increase at lower ages. Arguably, the variability of the interaction term $k * Age$ may suggest it is not needed in the model, however after redoing the plots without this term the difference in minimal (not shown here), therefore we keep this term in.

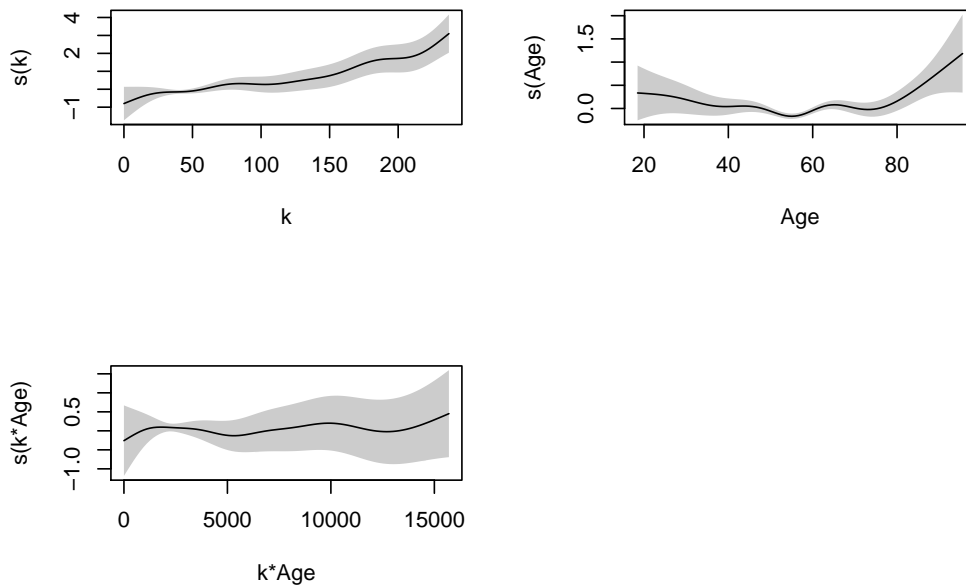


Figure 7.7: Plots of the smmothing functions used to model the heazard of censoring. The grey shading is the 95% confidence intervals.

Figure 7.8 shows the application of the IPCW to our incident data to estimate the direct risk, and Figure 7.9 shows the same for the total effect. The unweighted 1-KM and the Aalen-Johansen cause-specific cumulative incidence (solid lines) are included in the direct and total effect respectively for comparison. There is very little difference in the total effect between the weighted and unweighted versions, which is as expected as our censoring is mostly administrative at the end of the study and is not thought to be dependent on covariates. However, it is clear there is informative censoring for the direct effect, and therefore accounts for the competing events respectively for cancer and death without cancer. This difference is particularly noticeable for cancer in SSc patients, which is to be expected as SSc patients are at a much higher risk of death, and therefore many are weighted higher. This, as with the g-formula, implies that there is dependent censoring in the direct effect, and therefore it would be inappropriate to model this data without dependent censoring considerations. The IPCW gives very similar

results to those obtained from the g-formula estimator (Figure 7.5 and 7.6), and therefore we do not provide risk ratios or confidence intervals due to the preferred smoothed g-formula curves.

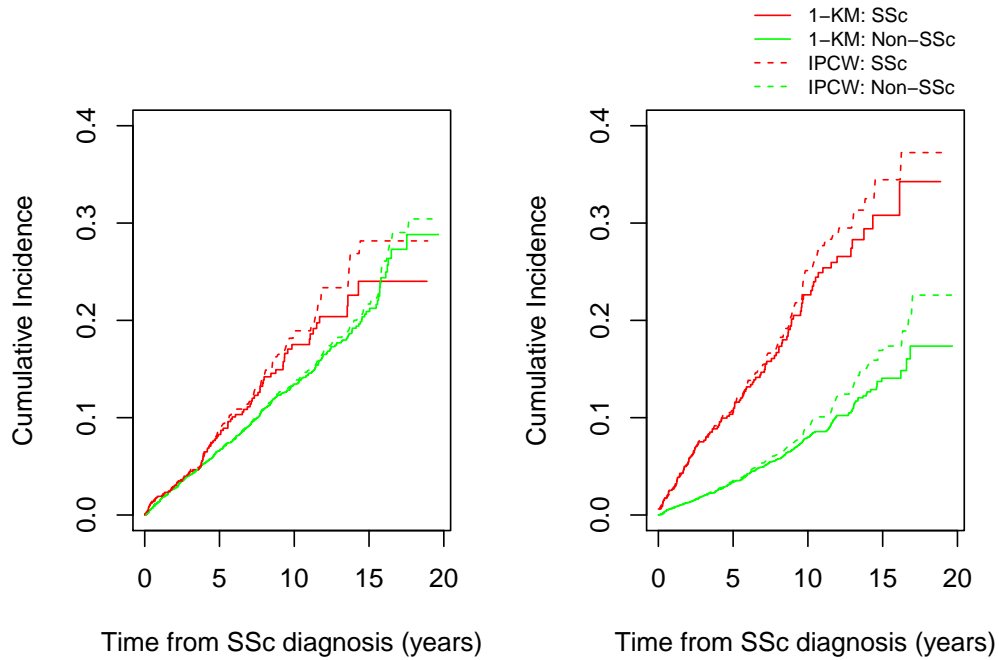


Figure 7.8: IPCW for risk of cancer (left) and death (right), direct effect, with data from the incident cohort. The unadjusted 1-Kaplan-Meier estimations are shown as the solid lines, and the IPCW are the dashed lines.

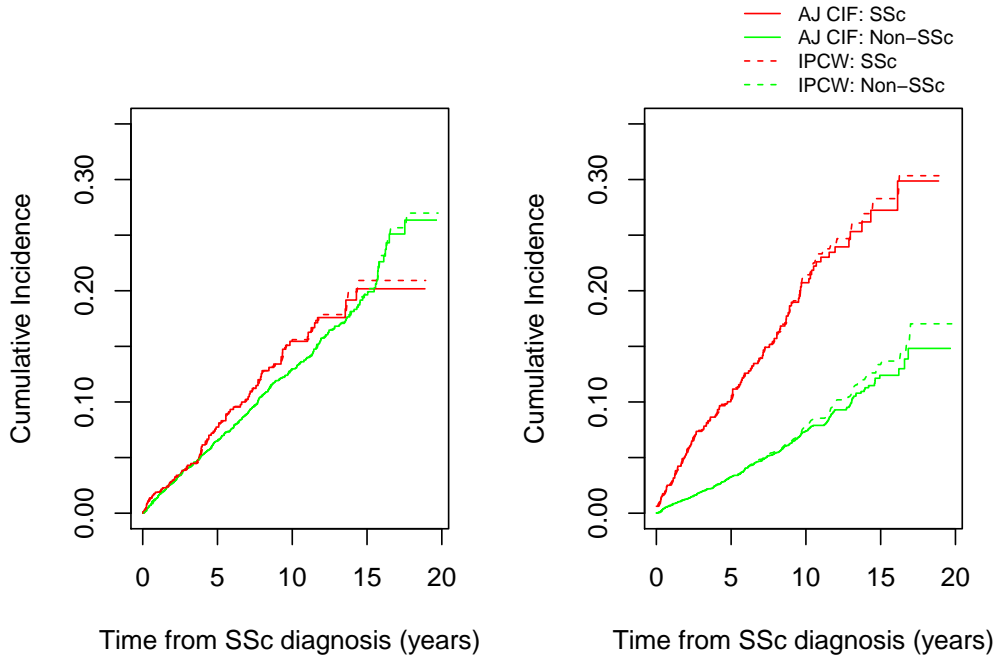


Figure 7.9: IPCW for risk of cancer (left) and death (right), total effect, with data from the incident cohort. The unadjusted Aalen-Johansen estimations are shown as the solid lines, and the IPCW are the dashed lines.

7.6 Summary

This chapter has demonstrated the use of g-formula estimation on our incident SSc dataset. The theory of the g-formula was explained along with the conditions which need to be assumed for successful implementation of the g-formula, which was then applied to the incident data. We have applied this methodology as a difference between SSc patients having their original exposure and then compared to them not having it (ATT), as opposed to the usual comparison between a whole population having the exposure and then not (ATE). We also included results for the IPCW method, however this is not our primary method for this thesis.

The g-formula has many advantages to the NPMLE discussed in the previous chapter, with a main one being the interpretation of an average causal effect as opposed to an association due to accounting for differing covariates. It also adjusts

for informative censoring, which is especially useful for the interpretation of the direct effect that the naive Kaplan-Meier lacks. However, parametric models require that the regression models used give unbiased/consistent estimates for the models to be correctly specified (Hernan & Robins, 2020).

For the incident dataset, there appears to be good agreement between the total effect methods, as shown in the Tables 7.3 and 7.4 above. The direct effect shows a minor increased risk for SSc and non-SSc patients, but with very similar risk ratios, compared to the previous NPML. The risk ratios of slightly above 1 suggest, when we consider the solely incident cohort, there is an indication of a small direct causal effect but not a total effect of SSc on cancer.

Due to the matched nature of this dataset (and smoking data not differing much between SSc and non-SSc patients), the full potential of this method as a causal estimator cannot be demonstrated here. It would be of interest to apply this methodology to other, unmatched, datasets.

One disadvantage of the methods presented in Young is the hypothetical interpretation of the elimination of death. This questions the interpretability of this way to quantify risk. An alternative method may be separable effects, as described in a very recent paper as presented by the same team as the Young paper, Stensrud et al. (2021). This may be an interesting place for further work.

In the next chapter, we recognize the difficulty of applying this theory to prevalent cohorts and discuss an adjustment which will allow for their inclusion.

Chapter 8

G-formula: Application to a prevalent cohort

8.1 Introduction

In the previous chapter, we introduced the g-formula, including the assumptions that are required for its use, and we then applied the estimator to the incident dataset.

In this chapter, we discuss how best to utilize our prevalent patients. The consideration of left truncated data when using the g-formula is rare, however Keil and Edwards consider the problems encountered with right censoring and late entry in a dataset, although not to the extent of solution or application. They highlight two areas of concern: exchangeability and differing covariate distributions.

To quote Keil & Edwards (2018) when briefly discussing exchangeability considerations encountered in data with late entry:

"...Roughly, however, for causal inference to be possible in survival analysis subject to left truncation, exchangeability must hold conditional on the measured past values of exposure and confounders,

among those who have entered into the study by time k and remain uncensored at time k . This strengthened exchangeability assumption implies that underlying health status should not depend on entry time, conditional on past values of exposure and confounders."

In this chapter, we shall assume exchangeability holds, and in the next chapter we relax this assumption. However, an issue with using the g-formula estimator with left truncated data is biased covariate distributions. To quote Keil & Edwards (2018) on covariate distributions:

"Under the definition of exchangeability extended to account for left truncation, Taubman et al's algorithm (Taubman et al., 2013) (*a paper demonstrating the usage of g-formula estimation in an epidemiological - rather than clinical - dataset*) cannot be used to estimate risk in target population data subject to left truncation. Frequently, the study population (re-imagined as a closed cohort) is the target population of interest, so the issue may arise in any data subject to left truncation. In order to account for these missing individuals, we would have to simulate extra members of the population, but it is not clear how these pseudo-individuals would be assigned baseline or time-fixed covariates, given that they were never observed. Unfortunately, there is no apparent solution given in the methodologic literature to this exact problem."

As above, using prevalent cohorts is not as simple as applying the previous chapter's methodology to the prevalent cohort, as we no longer have a pseudo-population that represents the true covariate distribution that we would expect in an SSc cohort. In prevalent cohorts, some types of patients are more likely to be left truncated than others. This is the selection bias we referred to in Section 5.6, simulation 1. The g-formula works by selecting a pseudo-population based on the patients in the dataset, then creating cloned

sets/pseudo-populations of the data and applying treatment types to each set. However, with left truncation, we may have a biased covariate distribution which will result in our distribution of prevalent covariates being significantly different from the covariate distribution of an incident cohort.

As there is no current solution, we propose 2 different methods to account for this difference depending on the dataset:

Method 1 : Use the prevalent cohort alongside the incident cohort to estimate the regression coefficients for the hazards. It is then possible to use only the incident SSc cohort for the pseudo-population, which has a cohort distribution representative of what we would expect in a population of current SSc patients. However, this is a particular benefit for our study, as we have both an incident and prevalent cohort, and can apply our regression coefficients to a dataset that is a recent and reliable representation of an SSc cohort. This allows the prevalent cohort to provide more information and increase the accuracy of the regression coefficients without application to a biased pseudo-population. However, we are aware that some studies do not benefit from an incident cohort. If the underlying population covariate distribution is known, then this method can perhaps also be used, where the hazard at each time step is estimated using the prevalent cohort, but only applied to an artificially constructed population.

Method 2 : Treat the prevalent cohort as incomplete, and use methods to approximate the missing data. We shall be using inverse probability weighting, where we give a greater weight to patients who are less likely to enter the study due to left truncation. To do this, we weight the cumulative incidence of each patient by the probability of a patient with their covariates surviving until their entry time given their covariates. This is then marginised over.

The chapter begins by justifying the use of regression models to model prevalent data. We then cover the first method, and explain the methodology of the second

method. To do this, we also estimate the covariate distribution we expect if the sampled prevalent dataset had been in its incident form. From here, we shall apply the weighted g-formula to the prevalent cohort in order to observe differences between this and the nonparametric method. Lastly, we shall utilize both cohorts for a final result for this chapter.

8.2 Estimating regression coefficients in a prevalent cohort

This section demonstrates that we can model prevalent data with a discrete time logistic model by including patient data from the time of entry into the study to their exit time. To do this, we return to similar notation we used in Chapters 4-6 and use maximum likelihood theory. This section's material can be found in other publications, but illustrates the use of modelling left truncated data with GLMs.

Assume temporarily that we do not have left truncation. We have information about n patients, $(x_1, \delta_1), \dots, (x_n, \delta_n)$, with time to endpoint x_i and censoring status δ_i . The endpoint, $X_i = \min(T_i, C_i)$, is either T_i , the minimum of event time, or C_i , censoring time, which depends on whether the patient is first observed to have the event of interest or is right censored. Censoring status, δ_i , is the indication of having the event of interest ($\delta_i = 1$) or not ($\delta_i = 0$).

Assume T and C are independent. Let $S(t) = S_T(t) = \Pr(T > t)$, $H(c) = \Pr(C > c)$, and let $f(t)$ and $h(c)$ be the corresponding density functions. The likelihood function on the basis of $(x_1, \delta_1), \dots, (x_n, \delta_n)$ is

$$\mathcal{L} = \prod_i^n \{ [f(x_i)^{\delta_i} S(x_i)^{1-\delta_i}] [h(x_i)^{1-\delta_i} H(x_i)^{\delta_i}] \}$$

Since our interest is only in survival times and not follow-up times, and our censoring time distributions do not depend on the same parameters of survival time, the second half of the above equation is treated as a constant (Klein &

Moeschberger, 2003). Under the assumption of independent censoring conditional on measured covariates the likelihood reduces to

$$\mathcal{L} \propto \prod_{i=1}^n \Pr[x_i, \delta_i] = \prod_{i=1}^n [f(x_i)^{\delta_i} S(x_i)^{1-\delta_i}]$$

The likelihood in the presence of left truncation, as given by Klein & Moeschberger (2003), page 74, is

$$\mathcal{L} \propto \prod_{i=1}^n \left(\frac{f_T(x_i)}{S_T(\ell_i)} \right)^{\delta_i} \left(\frac{S_T(x_i)}{S_T(\ell_i)} \right)^{1-\delta_i}$$

Let us now work in discrete time, where λ_{ij} is the interval hazard for patient i at time-step j such that $\lambda_{ij} = P(T_i = j | T_i \geq j, z_i)$, where z_i represents the covariate distribution of patient i . We drop the use of z_i notation from here, but the subscript i denotes patient i whose information differs from other patients because of their covariate distribution. The discrete time hazard can also be written as

$$\begin{aligned} \lambda_{ij} &= P(T_i = j | T_i \geq j) \\ &= \frac{\Pr(T_i = j \cap T_i \geq j)}{\Pr(T_i \geq j)} \\ &= \frac{\Pr(T_i = j)}{\Pr(T_i \geq j)} \\ &= \frac{\Pr(T_i = j)}{\Pr(T_i > j - 1)} \\ &= \frac{f_{ij}}{S_{i(j-1)}} \end{aligned}$$

where f_{ij} and S_{ij} are defined below. For us, we wish to estimate the hazard at each time-step for patient i , so $p(a, z_i, j; \hat{\theta})$, $q(a, z_i, j; \hat{\eta})$ and $r(a, z_i, j; \hat{\alpha})$ for cancer, death and loss to follow-up, respectively.

Let S_{ij} be the survival until time j of patient i . Then:

$$\begin{aligned}
S_{ij} &= \Pr(T_i > j) \\
&= \Pr(T_i > j | T_i \geq j) \Pr(T_i > j - 1 | T_i \geq j - 1) \dots \Pr(T_i > 1 | T_i \geq 1) \\
&= \prod_{k=1}^j (1 - \lambda_{ik})
\end{aligned}$$

where $P(T_i > 0) = 1$. We also define

$$\begin{aligned}
f_{ij} &= \Pr(T_i = j) = \Pr(T_i = j | T_i > j - 1) \Pr(T_i > j - 1) \\
&= \lambda_{ij} S_{j-1} \\
&= \lambda_{ij} \prod_{k=1}^{j-1} (1 - \lambda_{ik})
\end{aligned}$$

Let $\delta_i = 1$ if x_i is the event time and 0 otherwise, and x_i be the survival time of the i th subject. The log-likelihood, when there is no left truncation, for the grouped survival data is given by

$$\begin{aligned}
\mathcal{L} &\propto \prod_{i=1}^n [f(x_i)^{\delta_i} S(x_i)^{1-\delta_i}] \\
&\propto \prod_{i=1}^n \Pr(X_i = x_i)^{\delta_i} \Pr(X_i > x_i)^{1-\delta_i} \\
&\propto \prod_{i=1}^n \left[\lambda_{ix_i} \prod_{k=1}^{x_i-1} (1 - \lambda_{ik}) \right]^{\delta_i} \left[\prod_{k=1}^{x_i} (1 - \lambda_{ik}) \right]^{1-\delta_i} \\
&\propto \prod_{i=1}^n \left[\frac{\lambda_{ix_i}}{1 - \lambda_{ix_i}} \prod_{k=1}^{x_i} (1 - \lambda_{ik}) \right]^{\delta_i} \left[\prod_{k=1}^{x_i} (1 - \lambda_{ik}) \right]^{1-\delta_i} \\
&\propto \prod_{i=1}^n \left(\frac{\lambda_{ix_i}}{1 - \lambda_{ix_i}} \right)^{\delta_i} \prod_{k=1}^{x_i} (1 - \lambda_{ik}) \\
&\propto \prod_{i=1}^n \prod_{k=1}^{x_i} \left(\frac{\lambda_{ik}}{1 - \lambda_{ik}} \right)^{d_{ik}} (1 - \lambda_{ik})
\end{aligned}$$

As shown in the last line, δ_i can be instead written as d_{ik} , where $d_{ik} = 1$ if the i th subject experienced an event at time $x_i = k$ and 0 otherwise. This is then the same as the binary response model with hazard λ_{ik} for the probability. This is what allows the use of a regression response model. We repeat similar steps for

left truncated data,

$$\begin{aligned}
\mathcal{L} &\propto \prod_{i=1}^n \left[\frac{f(x_i)}{S(\ell_i)} \right]^{\delta_i} \left[\frac{S(x_i)}{S(\ell_i)} \right]^{1-\delta_i} \\
&\propto \prod_{i=1}^n \left[\frac{P(X_i = x_i)}{P(X_i > \ell_i)} \right]^{\delta_i} \left[\frac{P(X_i > x_i)}{P(X_i > \ell_i)} \right]^{1-\delta_i} \\
&\propto \prod_{i=1}^n \left[\frac{\lambda_{ix_i} \prod_{k=1}^{x_i-1} (1 - \lambda_{ik})}{\prod_{k=1}^{\ell_i} (1 - \lambda_{ik})} \right]^{\delta_i} \left[\frac{\prod_{k=1}^{x_i} (1 - \lambda_{ik})}{\prod_{k=1}^{\ell_i} (1 - \lambda_{ik})} \right]^{1-\delta_i} \\
&\propto \prod_{i=1}^n \left[\frac{\lambda_{ix_i}}{1 - \lambda_{ix_i}} \prod_{k=\ell_i+1}^{x_i} (1 - \lambda_{ik}) \right]^{\delta_i} \left[\prod_{k=\ell_i+1}^{x_i} (1 - \lambda_{ik}) \right]^{1-\delta_i} \\
&\propto \prod_{i=1}^n \left(\frac{\lambda_{ix_i}}{1 - \lambda_{ix_i}} \right)^{\delta_i} \prod_{k=\ell_i+1}^{x_i} (1 - \lambda_{ik}) \\
&\propto \prod_{i=1}^n \prod_{k=\ell_i+1}^{x_i} \left(\frac{\lambda_{ik}}{1 - \lambda_{ik}} \right)^{d_{ik}} (1 - \lambda_{ik})
\end{aligned}$$

The above likelihood is the same form as the previous likelihood function without truncation, but we include patients from the time of entry into the study. Again, δ_i in the last line can be d_{ik} which is 1 when the event for patient i occurs at time $x_i = j$ and 0 otherwise. This allows for the use of logistic regression to model prevalent patients in discrete time (on the assumptions of non-informative censoring and independent truncation). In our study, we model the occurrence of cancer and death as Y_k , where 0 indicates the event not happening and 1 the event happening. Therefore, we model the hazard of cancer and death using logistic regression from the time the patients enter to their exit. For example, if a patient enters the study at time 20 and is diagnosed with cancer at time 40 then $Y_{21} = 0, \dots, Y_{39} = 0$ and $Y_{40} = 1$ will contribute to the logistic regression model.

8.3 Using the prevalent cohort for regression coefficients, applied to the incident pseudo-population

Here we fit the logistic regression coefficients from a) the prevalent cohort alone, and b) the combined incident and prevalent cohorts, but we shall only apply the

fitted models to the pseudo-populations based on the incident SSc cohort. In the latter case, we use all 10,770 patients to fit the logistic model, but then only apply the model to the baseline characteristics of 803 incident SSc patients, with one pseudo-population assigned $a = 1$ and another assigned $a = 0$. We then predict the hazard for each patient at each time-step, and apply Equations (7.2) and (7.3). This is because the incident cohort has the distribution of covariates we expect in the baseline population. The fitting based on the combined dataset will most likely produce a better estimate than using just the prevalent fit due to the increased sample size with the more recent incident data, but we additionally include the result obtained when only the prevalent cohort is used to fit the GLM to compare the differences between the two applications. A difference between the combined and prevalent hazards could indicate a possible temporal trend where the hazard has changed over time, as prevalent patients are often diagnosed prior to 1998. This method is applicable in our study due to us having a solely incident cohort for the complete covariate distribution, however if the data consists of only the prevalent cohort then this approach would not be possible, unless prior knowledge of the population is known.

When fitting the model, the same process as the last chapter was used. The base model was k , A , $Male$, Age (age at SSc diagnosis), and smoking status. Without interaction terms, a GLM or GAM was chosen due to minimal AIC. For cancer, the AIC was smaller for the GLM (8397 for GLM compared to 8398), and for death the AIC were the same (5548), therefore the GLM shall be used for both. After this, interaction terms and higher order terms were tested one at a time via AIC, and were considered for inclusion if the AIC was smaller. These were all then included in the model and a backward stepwise algorithm was trialled and extraneous variables were removed. Also, covariates were considered for removal if the coefficient was too close to zero or not statistically significant (or not close to being statistically significant).

We report the coefficients for the variables of interest for cancer and death in Table 8.1 for when only the prevalent cohort is used and Table 8.2 when using both

incident and prevalent cohorts, with the time coefficient, k , in months, the same as for the incident cohort previously. Continuous variables, k and Age , were tested with higher order terms until these terms started to increase the AIC, and often age at SSc diagnosis was an important factor, and therefore interaction terms with SSc diagnosis often minimised the AIC. The GLM applied to the prevalent cohort to estimate the hazard of cancer shows that k^2 is included (although weakly) and $A \times Age$ is now included as a term, which differs to the incidence cohort in Table 7.2. For death, we no longer have the interaction term between sex and SSc ($Male \times A$) but do have $A \times Age$. The combined GLM does not add any additional covariates from the prevalent, but k^2 did not minimise the AIC for cancer but did so for death.

	Cancer	[95% Confidence interval]	Death	[95% Confidence interval]
Intercept	-12.892	[-15.239, -10.718]	12.501	[-14.947, -10.218]
k	1.217e-2	[0.0057, 0.0190]	3.260e-3	[-3.926e-4, 6.917e-3]
k^2	-5.210e-6	[-1.10e-5, 4.219e-7]	-	-
SSc (A=1)	1.054	[0.261, 1.826]	2.262	[1.435, 3.091]
Male	-0.858	[-1.918, 0.150]	0.145	[-0.111, 0.386]
Age	0.133	[0.068, 0.204]	3.404e-2	[-3.626e-2, 1.090e-1]
Age ²	-6.227e-4	[-1.165e-3, -1.108e-4]	5.198e-4	[-4.008e-5, 1.047e-3]
Smoker	0.110	[-0.0891, 0.304]	0.842	[0.638, 1.047]
Ex-Smoker	-0.106	[-0.301, 0.0847]	0.104	[-0.105, 0.308]
$k \times Age$	-1.180e-4	[-2.100e-4, -3.116e-5]	1.057e-4	[3.634e-5, 1.750e-4]
$A \times Age$	-1.403e-2	[-2.910e-2, 9.946e-4]	-1.816e-2	[-3.249e-2, -4.008e-3]
$Male \times Age$	2.125e-2	[2.860e-3, 3.999e-2]	-	-

Table 8.1: Coefficient values for logistic regression, estimated using the prevalent cohort only. The baseline for A is non-SSc, the baseline for sex is female and the baseline for smoking and ex-smoker is non-smoker. Age at SSc diagnosis is increase in hazard per year (as opposed to decade).

Interpretation of coefficients between the prevalent and combined is difficult due to models including different terms. However, there is the suggestion that the confidence intervals are smaller for the combined than the prevalent, for example smoking status and sex in death. This is what is expected by adding more data to the model.

	Cancer	[95% Confidence interval]	Death	[95% Confidence interval]
Intercept	-10.005	[-12.875, -10.292]	-10.911	[-12.297, -9.018]
k	5.128e-3	[4.89e-3, 9.06e-3]	2.119e-3	[-7.679e-3, 3.738e-3]
k^2	-	-	5.644e-6	[4.789e-7, 1.056e-5]
SSc (A=1)	0.302	[0.149, 0.500]	1.983	[1.328, 2.637]
Male	-0.977	[-1.771, -0.214]	0.263	[0.042, 0.403]
Age	0.101	[0.060, 0.144]	-1.413e-2	[-6.770e-2, 4.352e-2]
Age ²	-4.182e-4	[-7.600e-4, -8.951e-5]	8.741e-4	[4.664e-4, 1.257e-3]
Smoker	0.199	[0.053, 0.343]	0.905	[0.746, 1.063]
Ex-Smoker	2.911e-2	[-0.109, 0.165]	0.203	[0.045, 0.389]
$k \times Age$	-6.281e-5	[-1.04e-4, -2.13e-5]	1.617e-4	[8.794e-5, 2.318e-4]
$A \times Age$	-	-	-1.126e-2	[-2.163e-2, -9.534e-4]
$Male \times Age$	0.0208	[0.0081, 0.0338]	-	-

Table 8.2: Coefficient values for logistic regression, estimated using both the prevalent and incident cohorts. The baseline for A is non-SSc, the baseline for sex is female and the baseline for smoking and ex-smoker is non-smoker. Age at SSc diagnosis is increase in hazard per year (as opposed to decade).

We use these estimated hazards in the g-formula in accordance with Equation (7.2) and Equation (7.3) to estimate the cumulative risks. Figure 8.1 is the application for the direct effect, and Figure 8.2 is the application for the total effect. As with the last chapter, the pseudo-population that we have applied the GLMs to is the SSc cohort, so we are finding the average exposure effect in those who are exposed. In both sets of graphs, we have the three g-formula curves, with the incident, prevalent and combined sets used to fit the GLM and then applied to the incident SSc cohort's pseudo-population. The incident curves are an extrapolation of the data to later time points, which is not advised, however the curves are useful for comparison.

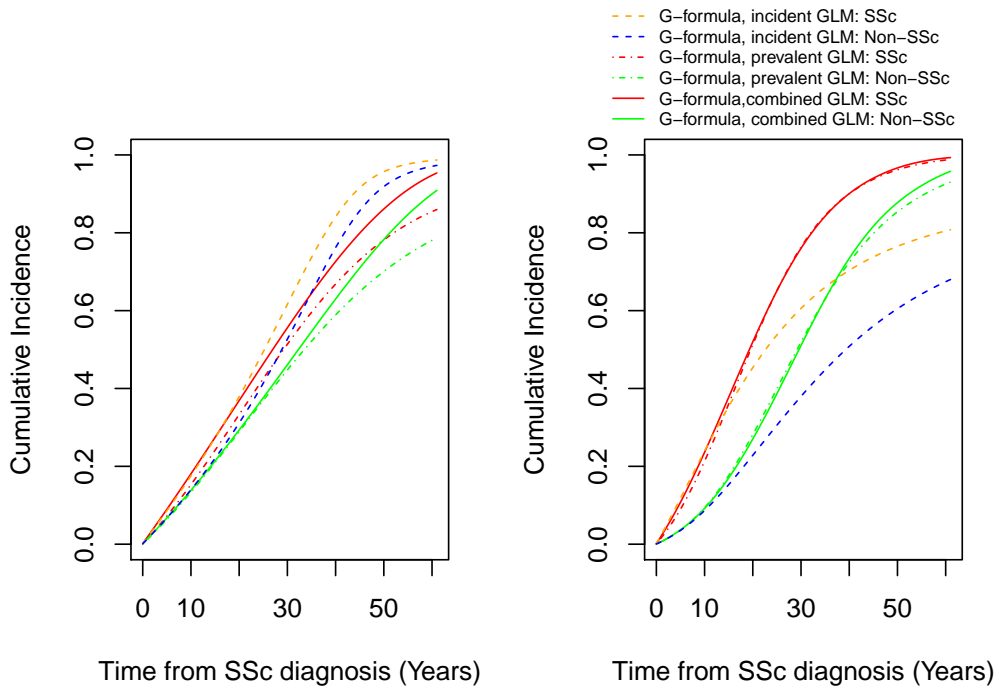


Figure 8.1: G-formula for risk of cancer (left) and death (right), direct effect, with the logistic regression model being fitted on either solely the incident cohort (blue/orange dashed lines), the prevalent cohort (red/green dot dashed line) or both the incident and prevalent cohort (red/green solid line), but with the pseudo-population based solely on the incident SSc cohort.

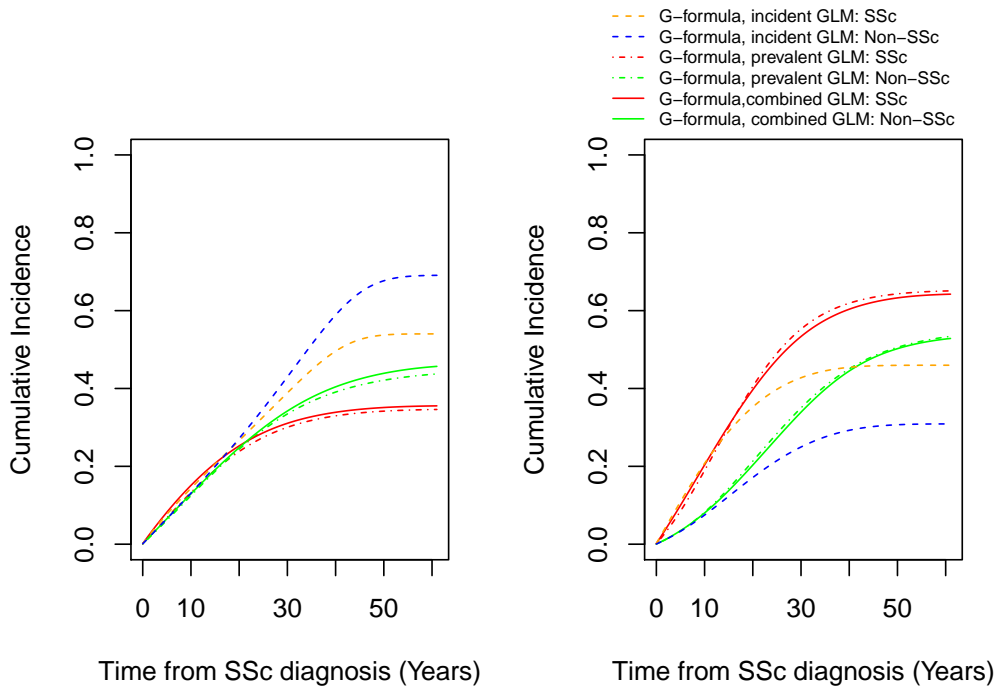


Figure 8.2: G-formula for risk of cancer (left) and death (right), total effect, with the logistic regression model being fitted on either solely the incident cohort (blue/orange dashed lines), the prevalent cohort (red/green dot-dashed line) or both the incident and prevalent cohort (red/green solid line), but with the pseudo-population based solely on the incident SSc cohort.

The combined and prevalent curves are very similar, with overlapping lines for the risk of death and a slight difference for the risk of cancer. The incident curves for the direct effect clearly have an increased direct risk of cancer compared to the prevalent and combined GLM applied to the incident pseudo-population. This noticeable difference is most likely due to the unreliability of extrapolating the data. Although not shown here, the removal of some terms, such as the interaction terms, from the incident GLM greatly changes the curves. There is the implication that prevalent patients have lower direct cancer risk than the combined data. This gives the impression that prevalent cases, who on average are diagnosed further back in calendar time, have a lower cancer risk. We explore this more in the next chapter which covers temporal trends.

The most noticeable consequence of using the g-formula compared to the

nonparametric methods of the previous chapter (Figure 6.3 and Figure 6.4) is the appearance of increased risk with increased time in the g-formula, in both the direct effect and total effect for both cancer and death. As we will discuss later in this chapter, this is most likely due to underlying differences between the incident and prevalent cohorts' covariate distribution. The nonparametric method does not take covariate information into consideration, but those in the risk set and having events at times greater than 20 will be prevalent, and therefore have characteristics of prevalent patients. However, in the parametric method we are predicting survival at each time-step based on the incident cohort's covariate distribution. If the prevalent cohort has more covariates of a type which has better cancer and death survival, possibly due to temporal trends, this may explain why the parametric method appears to give an overestimation of risk. Also, for the direct effect it could be due to better handling of informative censoring. There is a greater difference between the nonparametric and the g-formula in the direct effect, where death is treated as censoring in the nonparametric, than in the total effect. The g-formula pseudo-population is formed such that all patients are present over all time under study. So a patient from our dataset who was censored or died soon after SSc diagnosis will no longer appear in a nonparametric estimator after their removal from the risk set, but in the g-formula they will continue to contribute their covariate information over the time under study for the cancer risk as well.

While quadratic and cubic terms were tested and often the AIC showed no improvement, it is worth noting that piece-wise time was also investigated (results not shown) and showed very similar results.

Table 8.3 and Table 8.4 estimate the risk ratios and confidence intervals for the direct effect and total effect, respectively, at 10- 20- and 30- years, along with NPMLE estimated from Table 6.4. The bootstrapped confidence intervals are found using the same methods as in Section 7.4.1.

Even with large differences between the parametric and nonparametric curves, the ratios are close to each other. In the direct effect there is a slight increased risk

Time from SSc diagnosis (years)	Left truncation Kaplan-Meier [95% CI]	Direct effect with combined incident and prevalent GLM fitting [95% CI]
10	SSc 0.152 [0.126, 0.178] Non-SSc 0.122 [0.112, 0.132] Ratio 1.247 [1.031, 1.507]	SSc 0.179 [0.166, 0.192] Non-SSc 0.137 [0.130, 0.144] Ratio 1.307 [1.147, 1.489]
20	SSc 0.294 [0.252, 0.334] Non-SSc 0.253 [0.237, 0.269] Ratio 1.163 [0.998, 1.355]	SSc 0.367 [0.349, 0.386] Non-SSc 0.292 [0.281, 0.303] Ratio 1.258 [1.130, 1.401]
30	SSc 0.400 [0.335, 0.459] Non-SSc 0.362 [0.334, 0.389] Ratio 1.105 [0.930, 1.313]	SSc 0.554 [0.535, 0.573] Non-SSc 0.459 [0.448, 0.471] Ratio 1.207 [1.112, 1.309]

Table 8.3: Cumulative incidence risk ratio comparisons for the left truncation Kaplan-Meier and direct g-formula estimator methods using both the incident and prevalent data to model the GLM but then applying it to the incident pseudo-population. The NPMLE is estimated from both the incident and prevalent patients. The 95% confidence intervals are given in brackets.

Time from SSc diagnosis (years)	Left truncation cause-specific cumulative incidence [95% CI]	Total effect with combined incident and prevalent GLM fitting [95% CI]
10	SSc 0.136 [0.115, 0.162] Non-SSc 0.118 [0.109, 0.128] Ratio 1.153 [0.954, 1.393]	SSc 0.150 [0.138, 0.162] Non-SSc 0.128 [0.122, 0.135] Ratio 1.169 [1.021, 1.338]
20	SSc 0.235 [0.206, 0.270] Non-SSc 0.234 [0.219, 0.249] Ratio 1.008 [0.868, 1.171]	SSc 0.252 [0.235, 0.269] Non-SSc 0.247 [0.237, 0.257] Ratio 1.019 [0.899, 1.155]
30	SSc 0.290 [0.253, 0.332] Non-SSc 0.319 [0.297, 0.342] Ratio 0.909 [0.780, 1.058]	SSc 0.310 [0.290, 0.330] Non-SSc 0.341 [0.330, 0.353] Ratio 0.909 [0.804, 1.026]

Table 8.4: Cumulative incidence risk ratio comparisons for the left truncation cause-specific and total g-formula effect estimator methods using both the incident and prevalent data to model the GLM but then applying it to the incident pseudo-population. The NPMLE is estimated from both the incident and prevalent patients. The 95% confidence intervals are given in brackets.

ratio compared to the nonparametric estimator. There is no change in conclusion for the direct effect from the solely incident cohort however, with there being a small increased risk of cancer in those with SSc if death is eliminated, implying a small causal effect. However, when death is considered for the total effect, the narrower confidence intervals for the g-formula in Table 8.4 imply statistically significant results, with risk in the first 10 years being increased in those with SSc, however risk significantly decreased in later years such that there is a reduced risk of cancer in SSc patients compared to non-SSc patients. This is due to mortality being higher in non-SSc patients. Again, we note that while the confidence intervals are smaller here there is still the possibility of bias due to the g-formula's sensitivity to a violation of the three key assumptions (Taubman et al., 2009), therefore even if there is significance in the total effect at times 10 and 30 years, the closeness to 1 prevents us from conclusively saying there is a difference in risk between SSc and non-SSc patients. We note that the nonparametric estimator is using the incident and prevalent data, and is therefore estimating the risk of SSc and non-SSc separately, however we are looking for the average exposure in the exposed (ATE), therefore we would expect these results to be different if the SSc and non-SSc cohorts had drastically different covariate distributions (which ours do not have, due to matching).

The inclusion of prevalent patients here when fitting the GLM allows us to estimate the risk at longer durations from diagnosis more reliably than the solely incident cohort, as if we were only using the incident cohort we would be extrapolating to later time points. Also, using both sets increases the sample size.

8.4 Consequence of unadjusted prevalent inclusion

As previously stated, our study has the advantage of having an incident cohort, however many other studies will only have a prevalent cohort. Therefore, it is important to establish a method for just a prevalent cohort, as this will enable

the techniques described here to be applied beyond this thesis.

As we have highlighted, the g-formula is estimated by fitting a GLM to the data, creating pseudo-populations by assigning exposure and non-exposure to copies of the SSc cohort, and then marginalising over the hazards applied to these pseudo-populations. However, we expect the covariate distribution to be different in the prevalent cohort compared to the incident cohort. Patients with smaller survival times are under-represented, and patients with longer survival times are more likely to be included in a prevalent cohort. Therefore, without knowing or estimating the baseline covariate distribution (i.e. what we would expect in an incident cohort) the g-formula would average over a biased population, a population which will (most likely) produce an underestimation of risk. For example, if men have a higher mortality, then there will be a smaller proportion of males in the prevalent cohort, and males will therefore be underrepresented. This is a consequence of length-biased sampling, where we are more likely to recruit patients with longer survival times. Age at SSc diagnosis is of particular concern in our study, as a set of prevalent patients will be biased towards lower ages at SSc diagnosis, as these patients are more likely to have longer survival times and therefore long truncation times.

We demonstrate this possible but suspected underestimation, by fitting the GLM based on the prevalent cohort, and then applying it to pseudo-populations of the solely prevalent SSc cohort. Figure 8.3 shows the consequences of applying a logistic regression fitted to the prevalent SSc cohort for the direct effect, and Figure 8.4 shows the same for the total effect. We also include the previous risk curve where the GLM fitted from the solely prevalent cohort was applied to a pseudo-population of only the incident SSc cohort, for comparison.

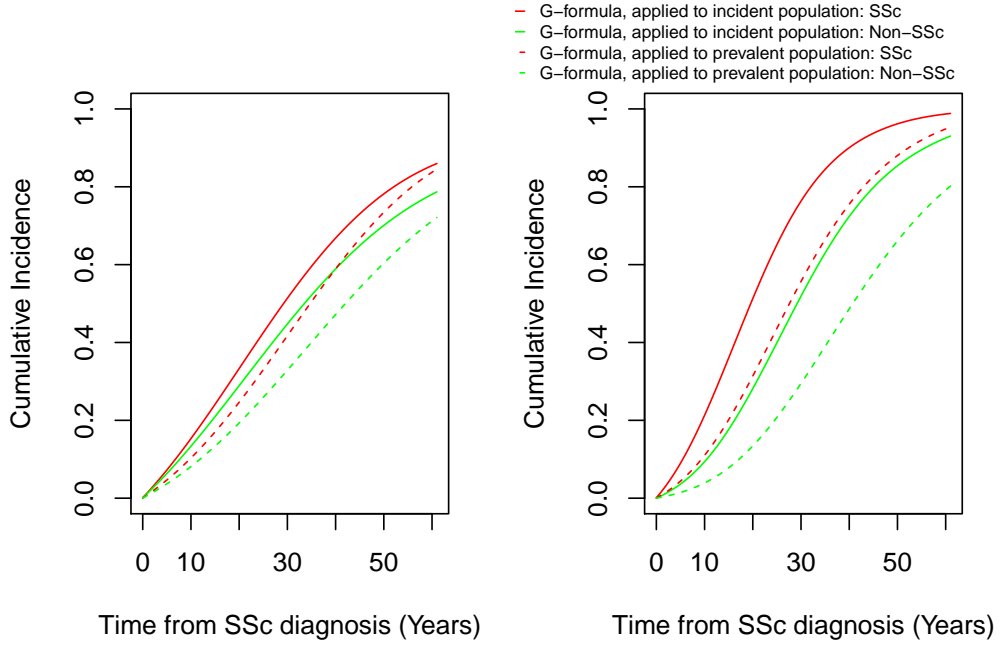


Figure 8.3: G-formula for risk of cancer (left) and death (right), direct effect, with the hazards being predicted from the prevalent cohort, and either applied to the incident SSc pseudo-population (solid line) or the prevalent SSc pseudo-population (dashed line).

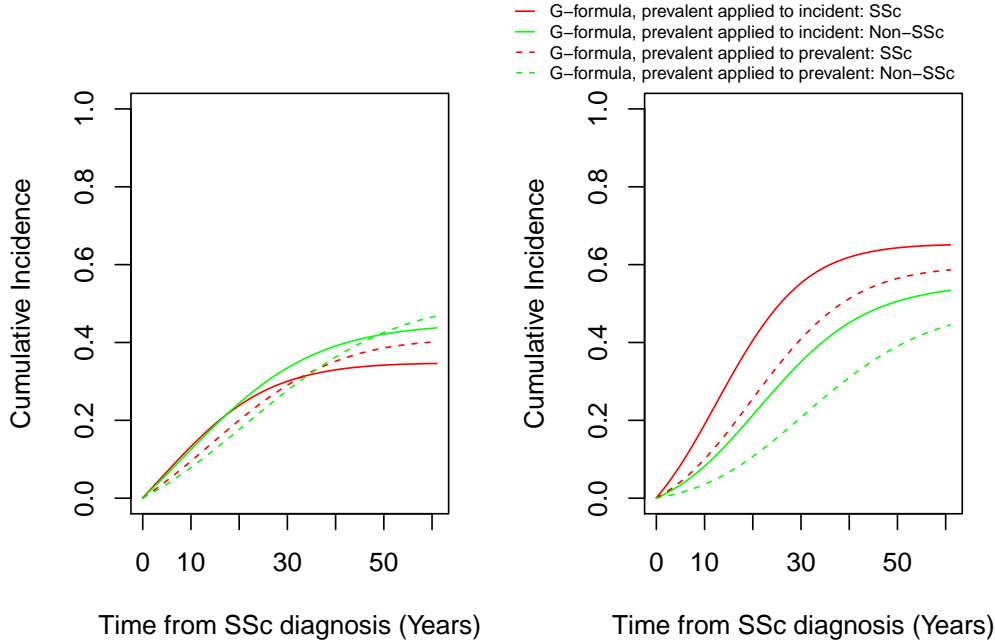


Figure 8.4: G-formula for risk of cancer (left) and death (right), total effect, with the hazards being predicted from the prevalent cohort, and either applied to the incident cohort (solid line) or the prevalent cohort (dashed line).

The prevalent regression applied to the prevalent cohort appears to indicate an underestimation of risk compared to the application to the incident SSc pseudo-population. This is the change expected as we suspect we have a biased covariate distribution where those with the worse survival are underrepresented in the prevalent cohort. However, we also note that it may be that the prevalent cohort naturally has a different covariate distribution from the incident cohort because of temporal trends.

We consider how the covariate distribution might differ between the incident and prevalent cohorts due to selection bias. There might be a small bias towards less males in the prevalent cohort, as we have observed that males have higher hazard of death and cancer in the Cox models as discussed in previous chapters, and also there are less males in the prevalent cohort (17.2% in the incident but only 12.2% in the prevalent cohort). There may be a lower number who smoke or who are ex-smokers, as we have seen smokers are at an increased risk of death and cancer.

We would expect the greatest bias to arise from age at SSc diagnosis. As we have seen in Table 5.1, the prevalent cohort has a mean age 10 years younger than the incident cohort (48.1 years compared to 58.1 years), which is due to the negative correlation between age at SSc diagnosis and truncation time (See Figure 3.10). Patients with a larger truncation time will disproportionately be younger at age of SSc diagnosis, as patients diagnosed in later life have smaller survival times. Therefore, the age profile of prevalent patients will be skewed to younger patients. Patients diagnosed with SSc at younger ages have a lower hazard of both cancer and death, as shown in the Cox models from the previous chapters, so averaging the risk over more of these patients will show an underestimation.

In the next section we propose an approach which addresses this bias, and will then use simulations to demonstrate both the concept of baseline covariate proportion estimation and the use of the novel weighted g-formula. Then we shall apply this to our data.

8.5 Methodology for weighting the g-formula

We wish to account for missing patients through the use of weighting. The goal is to weight patients higher if their survival probability due to their covariates makes them less likely to have entered the study. This aims to account for patients who are missing due to left truncation.

We begin by investigating how to estimate the underlying distribution of covariates, which is also useful for the wider study of SSc to see how covariate distributions of the prevalent cohort at baseline are different from that of the incident cohort. This may highlight possible temporal trends that are occurring. This may show, for example, if even after prevalent cohort correction, SSc is being diagnosed at later ages now than further back in calendar time, or if more males are being diagnosed with SSc now compared to past cases.

Following work by Chan & Wang (2012) an estimator of the baseline covariate distribution function (what we would expect the prevalent sample to look like if

it has been sampled as an incident cohort) is given by

$$\hat{F}_{Z_0}(z; \hat{S}) = \left(\sum_{i=1}^n \frac{1}{\hat{S}(L_i|Z_i)} \right)^{-1} \left(\sum_{i=1}^n \frac{\mathbb{I}(Z_i \leq z)}{\hat{S}(L_i|Z_i)} \right) \quad (8.1)$$

Here, $\hat{S}(L_i|Z_i)$ is each patient's estimated probability of surviving until their entry time given their covariate values. This formula gives higher weightings to patients whose characteristics lead them to be less likely to survive to the time of their study entry. In their paper, Chan and Wang provide both nonparametric and semiparametric methods. In the nonparametric form, survival probability can be estimated for categorical covariates, but cannot be estimated consistently with continuous data. In the nonparametric form, $\hat{S}(t|Z)$ can be estimated by the left truncated Kaplan-Meier estimator for each individual covariate. However, for entry into our dataset we require patients to be both alive and cancer-free. Therefore, if we were using a nonparametric method to estimate the survival of all our patients into the study we require the competing events estimator of survival. We use $\hat{S}(t) = \prod_{j:t_j < t} \left(1 - \frac{d(t_j)}{r(t_j)} \right)$ with $d(t_j) = \sum_{k=1}^K d_k(t_j)$, where d_k is the number who have an event of any event type, k . For us, this would be cancer and death without cancer.

From the logistic regression used to estimate the g-formula, we are able to estimate the survival of each patient given their individual covariate profile. We remind ourselves of the decision to use the ATT as opposed to the ATE, Section 7.4.2. This choice also has an advantage when we consider how to weight the population. Patients with SSc were first selected to be in this study, and then each of these was matched to non-SSc patients. As SSc patients are the ones under study and therefore form the basis of who entered the study, we would like to weight each SSc patient by the probability that a patient with those covariates entered the study. It would be nonsensical to weight non-SSc patients, as it is not their survival which dictates their entry but their SSc match.

We propose that each SSc patient has their own weighting given their covariates, where SSc patients retain their original exposure of $A = 1$ for SSc to estimate

the probability of their survival, where their hazard is estimated from the logistic regression models for cancer and death. Hence, the survival probability of patient i is the probability of survival up until their entry time L_i , given as

$$\hat{S}_i(L_i|z_{0i}, A = 1) = \prod_{j=0}^{L_i} [1 - p(A = 1, z_{0i}, j; \hat{\theta})][1 - q(A = 1, z_{0i}, j; \hat{\eta})], \quad (8.2)$$

where z_{0i} is the baseline covariates for patient i . This can then be used in Equation (8.1).

Using this to weight each patient's estimate of risk, the estimate of the direct effect (ATT) is therefore,

$$\hat{\text{Pr}}[Y_{K+1}^{a, \bar{d}=\bar{0}} = 1|A = 1] = \left(\sum_{i=1}^n \frac{1}{\prod_{j=0}^{L_i} [1 - p(A = 1, z_{0i}, j; \hat{\theta})][1 - q(A = 1, z_{0i}, j; \hat{\eta})]} \right)^{-1} \times \sum_{i=1}^n \frac{\sum_{k=0}^K p(a, z_{0i}, k; \hat{\theta}) \prod_{j=0}^{k-1} [1 - p(a, z_{0i}, j; \hat{\theta})]}{\prod_{j=0}^{L_i} [1 - p(A = 1, z_{0i}, j; \hat{\theta})][1 - q(A = 1, z_{0i}, j; \hat{\eta})]} \quad (8.3)$$

and the estimate of the total effect (ATT) is,

$$\hat{\text{Pr}}[Y_{K+1}^a = 1|A = 1] = \left(\sum_{i=1}^n \frac{1}{\prod_{j=0}^{L_i} [1 - p(A = 1, z_{0i}, j; \hat{\theta})][1 - q(A = 1, z_{0i}, j; \hat{\eta})]} \right)^{-1} \times \sum_{i=1}^n \frac{\sum_{k=0}^K p(a, z_{0i}, k; \hat{\theta}) [1 - q(a, z_{0i}, k; \hat{\eta})] \prod_{j=0}^{k-1} [1 - p(a, z_{0i}, j; \hat{\theta})][1 - q(a, z_{0i}, j; \hat{\eta})]}{\prod_{j=0}^{L_i} [1 - p(A = 1, z_{0i}, j; \hat{\theta})][1 - q(A = 1, z_{0i}, j; \hat{\eta})]} \quad (8.4)$$

We now demonstrate the empirical performance of these estimators through simulation.

8.6 Simulation of proportion estimation and weighted g-formula estimator

We perform a simulation to demonstrate the application of the Chan and Wang estimator, and also to highlight a danger of studying a timeframe which does not have representation of all covariates under study.

There is a concern that after a certain length of study time, a subset of patients is much less likely to be recruited into the study due to the time under study often being larger than the survival of these patients. Often these patients have lower survival probabilities, and the inclusion of these patients will provide a large weighting but their absence will lead to no weighting. As an extreme example, suppose a patient with covariate z has a 0.0001% chance of surviving until time 20 years. If a patient does indeed survive and is in the study, they could have a huge weighting to account for other patients who failed to make it into the study. However, if that patient is not recruited, then there shall be no weighting. This is comparable to the positivity condition/positivity assumption in missing data literature. To quote Molina et al. (2018),

"It [positivity condition] states, in the missing data setting, that the propensity score is bounded away from zero ... the violation of the strict positivity condition causes the estimates to be very unstable and to have a large variability."

This may demonstrate the possible need to set a maximum truncation time for the dataset, as opposed to our study which is naturally bounded by the patient with the largest truncation time.

We wish to demonstrate this positivity condition in the left truncation setting. To this end, we perform a simulation study as follows. For consistency with this study, we simulate patients born under a uniform distribution between calendar times 1800 and 2000. We simulate two covariates, sex and age at diagnosis. Each patient is assigned male or female randomly with equal probability, where $z_{Sex} = 0$

if female, and $z_{Sex} = 1$ if male. If they are male, then their age at SSc diagnosis follows a normal distribution $Age \sim N(40, 10^2)$ and if they are female they will have a later mean age of diagnosis normally distributed $Age \sim N(60, 10^2)$. Both sex and age at SSc diagnosis will affect survival, with event time distributed exponentially with $T \sim Exp(0.1 + 0.2z_{age} + 0.1z_{Sex})$, where $z_{age} = 1$ if SSc diagnosis occurred younger than 50, and $z_{age} = 0$ if the patient was above 50 when they were diagnosed. Therefore, patients are more at risk if they are older and they are male. We simulate this dataset so that 10,000 patients are recruited during $(\tau, 2000)$, where we vary τ . In an incident cohort we expect half to have covariate $z_{sex} = 1$, and half to have $z_{age} = 1$. Note that the prevalent cohort covariate distribution will become biased the longer the truncation time, with less males and less elderly patients.

We provide two different ways of modelling survival, either via NPMLE or logistic regression (LR). The NPMLE is the Kaplan-Meier estimator for survival, where survival is estimated for each covariate of interest separately ($\hat{S}(L_i|z_{sex})$ and $\hat{S}(L_i|z_{age})$ as opposed to $\hat{S}(L_i|z_{sex}, z_{age})$). If we use the logistic regression to model survival, then patient i 's survival is estimated

$$\hat{S}_i(L_i|z_{0i}) = \prod_{j=0}^{L_i} [1 - p(z_{0i}, j; \hat{\theta})] \quad (8.5)$$

The logistic model is taking into account covariates z_{Age} and z_{Sex} as well as their interaction, $z_{Age} \times z_{Sex}$. This is comparable to Equation (8.2), but with no competing risks ($q(\cdot) = 0$).

We perform the simulation 300 times. Figure 8.5 demonstrates the difference in the adjusted and the unadjusted covariate proportions for sex, with the survival probability used for the weightings estimated either by the NPMLE method or logistic regression. The black points show the bias towards females with increasing truncation time in the prevalent cohort, therefore the unadjusted proportion is not representative of the true incident baseline proportion. The red points are the nonparametric estimator from the Kaplan-Meier estimator, and the green is

the approximation where survival is estimated using logistic regression (Equation (8.5)). There appears to be good agreement between the two different survival methods, and both show good approximations of the expected proportion at baseline. We may expect to see differences between the two methods if any of the covariates were continuous, as the NPMLE would model this less well as they would need to be considered as categorical variables. Chan and Wang's semiparametric method could be used, or the logistic regression method we will be using. Confidence bars are also included at the 95% level. Let b_i be the vector of point estimates at time τ , where $i = 1, \dots, B$, with mean $\bar{b} = \frac{1}{B} \sum_{i=1}^n b_i$ and confidence intervals (95%) for the sampling are included $(\bar{b} \pm z_{(1-\alpha)/2} \frac{s}{\sqrt{B}})$, where $s = \sqrt{\frac{\sum_{i=1}^n (b_i - \bar{b})^2}{B-1}}$.

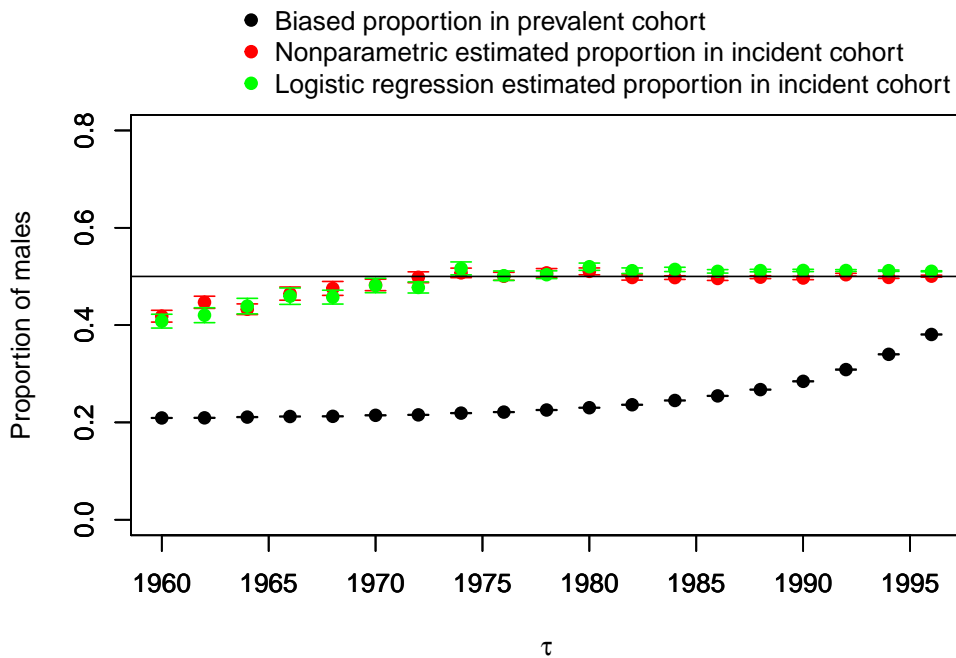


Figure 8.5: Simulation of biased prevalent covariate distribution and adjusted estimated baseline distribution for males. The points are the mean of 300 simulations and the bars are the 95% confidence intervals.

In keeping with the theory of inverse probability weightings, these weightings can become insufficient when the probability of survival is close to 0, resulting in the

possibility of high weightings or no patients to weight at all. This happens in our study for $\tau < 1975$, where males are unlikely to be recruited prior to this time. If the weightings are too unstable for a few patients, these few observations can greatly influence and disproportionately affect the results, depending on how biased these patients are. Unequal weights in a missing data setting can be accounted for by using a variety of methods:

- Trimming the tails - The removal of patients with large weighting. For us, the large weightings may be the result of the large survival/truncation time or the rare covariates under study. If we removed patients based on having a long truncation time, then we would effectively be restricting our study to observe less patients further back in calendar time. Removing patients based on rare covariates may bias the dataset further by removing these patients.
- Weight truncation - Truncate weights so that there is a maximum allowable weight, such that if a weight exceeds this value it is set to the maximum. Again, this may reduce variance but it will increase bias.
- There are other methods, such as matching weights (Greene & Li, 2013) and overlap weights (F. Li & Thomas, 2019), however consideration of these is beyond the scope of this study.

We have not set a calendar timeframe for when SSc patients may enter the study. There is an incentive to set one so that patients with large truncation times have less influence, and we can be confident that patients of all covariate types are represented. Not only will this help with extreme weightings, but due to possible temporal trends, where covariates change over time and past cases do not reflect the current distribution, the removal of these past cases may aid in estimating the current covariate proportions. The above estimator marginalises over all cases, so if historic cases have greatly different covariate distributions due to temporal trends, then this may bias the results towards a covariate distribution

which is no longer appropriate (see Chapter 9). Including less historic cases is more appropriate when we wish to study the current disease distribution. The drawback of this is a loss of patients, so we do not want to restrict the calendar time-frame of recruitment too much, creating a possible trade-off of sample size and bias reduction.

We also note that the above simulation is only one type, which is prone to many small survival times and limited long survival times. The exponential distribution is used, which results in most patients having small survival times with a few extremes, so having a few patients with large truncation times is expected to result in extreme weightings.

We now demonstrate the application to estimate the cumulative risk. We apply weightings to estimate the risk, based on Equation (8.3) (although in this case $q(a, k, z_{0i}; \eta) = 0$ due to the absence of competing risks). For this, we only recruit patients diagnosed with ‘SSc’ between 1990 and 2000. Figure 8.6 shows a demonstration of the application of weightings for risk from one simulation with 100,000 prevalent patients. The black line is the cumulative incidence line produced with the application of the left truncated Kaplan-Meier estimator¹ to the prevalent data based on the above simulation. This is the expected cumulative incidence curve despite the different covariate proportions, as shown in Simulation 1 of Section 5.6. The red line is the application of the g-formula estimator without weighting, which produces an underestimation of risk, due to a smaller proportion of $z_{sex} = 1$ and $z_{age} = 1$ in the pseudo-population. The green line is now the adjusted, weighted version, and is close to one minus the Kaplan-Meier curve. This demonstrates the similarity between the nonparametric and parametric methods we are using, however the g-formula will allow for the estimation of the direct risk which the simple Kaplan-Meier will not.

¹Estimated using the *survival* package command `survfit`.

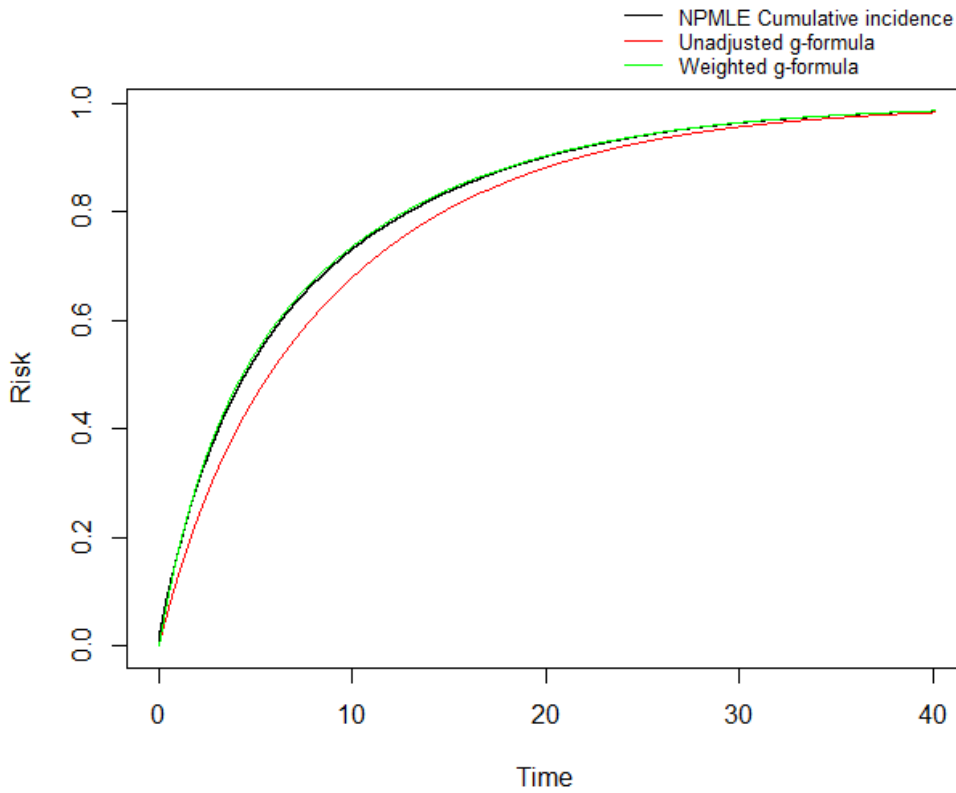


Figure 8.6: The cumulative risk curves produced by simulation. The black curve is the 1-KM estimated from an incident cohort, the red is the biased prevalent curve and the green is the g-formula bias adjusted curve.

8.7 Estimating the baseline covariate distribution

In this section we apply the methodology above to our dataset in the following order:

1. We apply the nonparametric survival and the logistic regression survival methods to the incident cohort to examine the use of this method on a cohort where the underlying (baseline) distribution is known.
2. We apply the nonparametric survival and the logistic regression survival methods to the prevalent cohort to investigate the estimated covariate distribution.

Only the SSc patients are used in this section, as the recruitment was based on SSc patients and therefore the SSc covariate distribution is what we wish to estimate.

8.7.1 Applying the NPML estimator of baseline covariate proportions to the incident data

Before we apply Equation (8.1) to the prevalent cohort we wish to use the estimator on our incident cohort. This provides an example of how the estimator works and is applied, and the differences we might observe. It may also serve to illuminate temporal trends over the 20 years of the incident cohort. To do this, we use our incident dataset by ‘moving’ the study start date such that our incident patients become prevalent patients. The hope is that we shall observe approximations close to the true incident cohort distributions.

To demonstrate how this is done, we consider four hypothetical incident patients, see Table 8.5. If we set the new study start date to 01/01/1999, Patient 1 and Patient 2 would be included, each with a left truncation time of one. Patients 3 and 4 would not be included, as we are only considering ‘prevalent’ patients.

If we set the new study start date to 01/01/2000, we would also now have Patient 3 included as a prevalent patient. Patients 1, 2 and 3 would then have left truncation times of 2, 2 and 0.5 years respectively. For the year 2002, we would have Patient 2 and Patient 3 included, however Patient 1 is now not included.

Patient id	Entry date	Outcome date
1	01/01/1998	03/10/2000
2	01/01/1998	01/01/2010
3	01/07/1999	05/11/2008
4	17/05/2005	05/11/2016

Table 8.5: Example incident patients.

These cohorts of ‘prevalent’ patients will have different, biased proportions, and it is the ability to estimate the baseline proportion that we wish to investigate.

Our survival probability is also estimated from the newly formed prevalent cohort. Therefore, we expect less accuracy in the estimate of the adjusted proportion when the study start date is earlier in time, as we will only have a few patients in the study. For example, if we recruited patients between 1998 and 2000 for a study start of 2000 there is only a 2 year window for patients to be diagnosed with SSc and therefore be eligible for recruitment. Equally, we expect less accuracy when the study start time is set to later years, as there will be less time for follow-up in the study. For example, if we set a new study start date of 2014, we have from 1998 to 2014 to recruit patients, but only 4 years of possible follow-up in which events can occur. There is also the possibility of left truncation issues as described in Section 5.4.1, as we have fewer patients in the times closest to time 0 (time from SSc diagnosis). This could cause a large ‘jump’ in survival when the first event occurs, which may lead to a bias that will affect larger survival times as well. As stated above, survival is the survival of both cancer and death until entry.

We are interested in the overall baseline covariate distribution changes under different artificial study start times of the following three covariates: sex, age at SSc diagnosis, and smoking. In Figures 8.7, 8.8 and 8.9 the x-axis is the new study start date, and the y-axis is the proportion of each covariate type in the study, so for example a study start year of 2000 will include those who were diagnosed with SSc between 1998 and 2000 who survived and were cancer-free until 2000. The black points are the proportion in the newly formed prevalent cohort without adjustment. The red dots are the bias adjusted estimate for the baseline covariates using the nonparametric estimate of survival, and the green dots are the logistic regression proportion estimates. The black horizontal line represents the true, known, incident covariate distribution. When the hypothetical study start time is set to the earlier years, we do not expect a difference between the unadjusted and the adjusted proportion estimate due to small truncation times, but also as it is a small sample it may be different to the whole incident proportion (the black lines). In the later years we would observe a difference between the unadjusted

and the adjusted if there was a bias occurring over time, but would hope that the adjusted was close to the black line.

Sex Figure 8.7 is the proportion of females depending on study start date. In the incident cohort, 82.8% are female, but there are more females in the prevalent cohort with 87.8% female. However, survival between males and females is very similar, and therefore there is little change between the prevalent cohort's covariate distribution and that of the bias adjusted proportion. There does not appear to be a large difference in the proportion of females over time, and the bias correction has minimal effect. When viewing the survival NPMLE for males and females (whether just incident or both incident and prevalent, not shown here), there is actually very little difference in survival between males and females, explaining why we may be seeing little difference between the adjusted and unadjusted proportions. The logistic regression appears to be a better approximation, except for the outlier 2009.

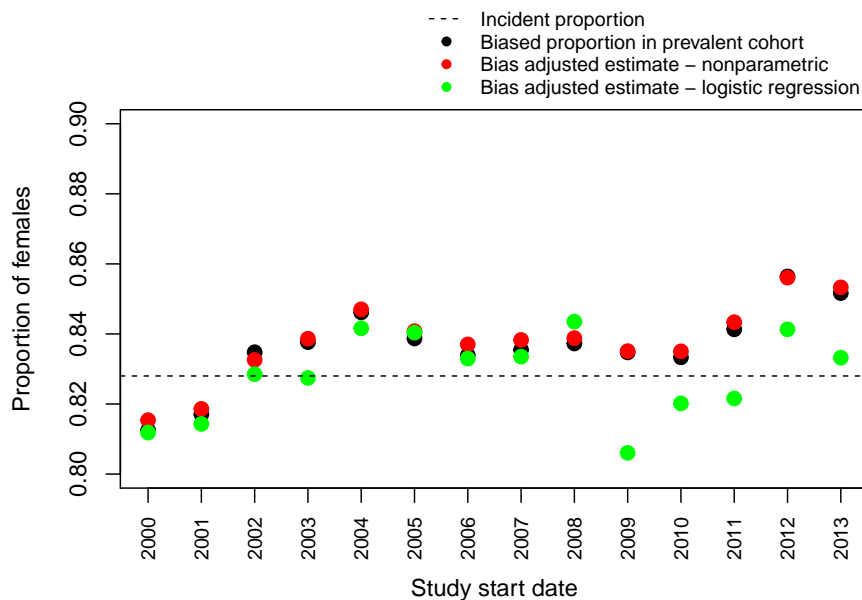


Figure 8.7: Estimation of the baseline covariate proportion for females in the incident cohort depending on the hypothetical study start.

Age at SSc diagnosis We expect the distribution of age at SSc diagnosis to be heavily biased in prevalent cohorts. This is due to the correlation between age at SSc diagnosis and truncation time, as those who were diagnosed with SSc at an advanced age will be less likely to survive to study entry. As the nonparametric estimator we are using here does not allow for continuous variables, we stratify the ages into 4 categories, <40 years, 40-50 years, 50-60 years, and 60+ years. The bias adjusted proportion estimates are shown in Figure 8.8. We see the black points deviate more from the true proportion in the study the later we move the recruitment time, where there are fewer patients in the oldest age set. The adjusted proportions are an improvement in the later age categories compared to the biased prevalent distributions, with the estimations being closer to the true overall covariate distribution than the unadjusted proportion in the artificially created prevalent cohorts. This is reassuring despite the small sample used. The bias discussed in Pan and Chappell can be observed here. For example, note that when the study start is set to 2010 there is a particularly large gap between the adjusted and unadjusted proportions, which is due to a patient in the 60+ category having an event soon after diagnosis while there were less than 6 patients in the study. An extension to this would be to apply the Lai and Ying NPMLE estimator, but as this section is simply an example this is not undertaken here. The logistic regression does not suffer from this problem, hence the better approximation at these times, although there is an outlier in the 60+ category with truncation year of 2008.

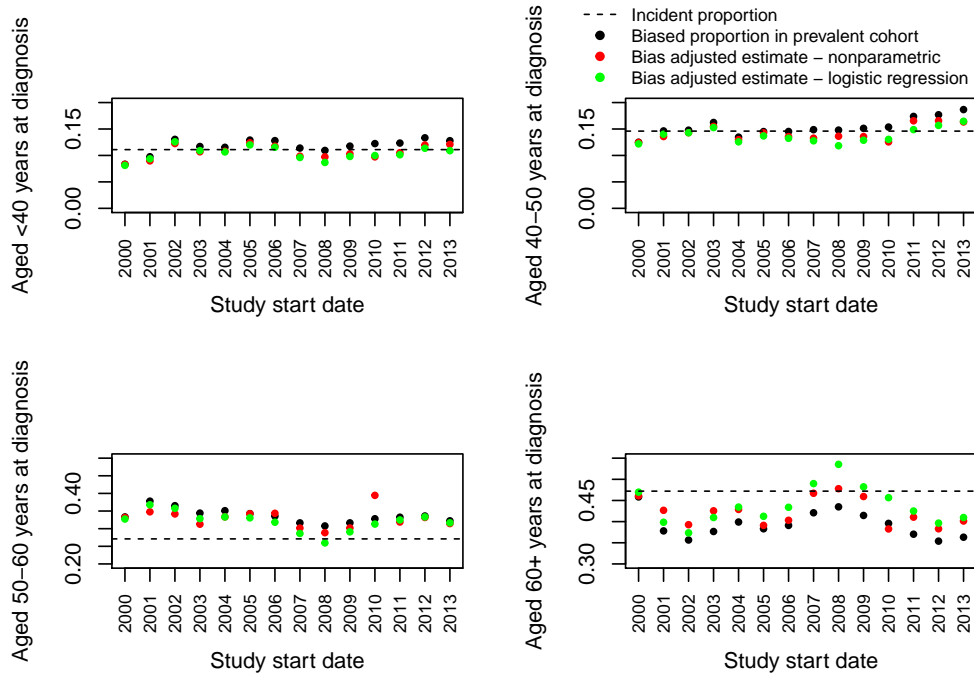


Figure 8.8: Estimation of the baseline covariate proportion for age at SSc diagnosis in the incident cohort depending on the hypothetical study start.

Smoking habits Figure 8.9 shows the biased proportion and the adjusted proportion for the three smoking categories. Both the adjusted and unadjusted vary greatly. In the early study start times there is a large deviation from the known incident cohort proportion, suggesting more smokers. In the ex-smoker category, the adjusted appears to produce a better estimate, however this is not the case in the other two categories. This again may be due to small numbers in the risk set at early survival times. Both the nonparametric and g-formula seem to be similar approximations.

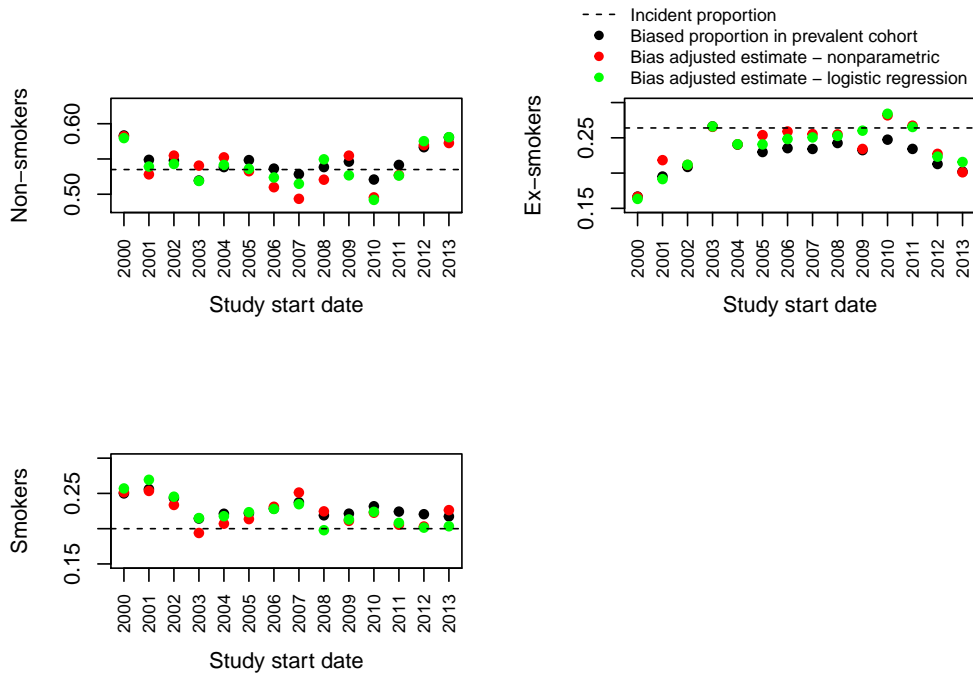


Figure 8.9: Estimation of the baseline covariate proportion for smoking types at SSc diagnosis in the incident cohort depending on the hypothetical study start.

8.7.2 Applying the nonparametric and logistic regression estimator of baseline covariate proportions to the prevalent data

We shall now apply the above principles to our prevalent cohort. We are interested in seeing if there is a difference between the observed incident proportions and the bias adjusted baseline estimated proportions using the prevalent data, which would imply that there is a difference in covariate distribution over time or that survival probability is perhaps not being modelled correctly. For this, we also use our incident cohort in addition to the prevalent cohort to estimate survival.

Referring back to our simulation on recruitment times, Figure 8.5, there will be a trade-off between setting an accurate recruitment time (the calendar times between which patient diagnosis date can occur) and having a robust number of

patients in the study. The patient with the longest truncation time is 47 years, with an SSc diagnosis in 1962. However, among all the covariates, it is age which would be problematic, as those with a large age at SSc diagnosis would only be in the study for small truncation times. The longest entry time for a patient in the age category 60+ years is 24.7 years, with a diagnosis in 1978. Therefore, using patients with truncation times greater than this may introduce bias in the weightings, as demonstrated in the above simulations. It may be best to only estimate the baseline covariate proportions using patients with left truncation time smaller than this.

Below we provide the estimation of baseline covariate distributions. We do this for prevalent patients only diagnosed after τ , with the x-axis being τ . Therefore, we will expect less patients when τ is later in calendar time. For example, a sample including prevalent patients from 1960 will contain more patients, but the weightings will be biased due to under-representation of patients who are unlikely to survive to the longer times (such as those 60+ years). A point corresponding to 1995 will represent those diagnosed from 1995 onward, so the cohort will have less patients but will not be biased by large truncation times.

We use all our patients to estimate survival, including both incident and prevalent cohorts. Therefore, the survival probability will not change when we move τ , we are just altering the patients used to estimate Equation (8.1). If we wished to investigate the change in survival over calendar time, we might want to also estimate survival from each reduced cohort, but for now we want to keep possible temporal trends and biased weights separate.

Figures 8.10, 8.12 and 8.14 plot each patient's probability of survival (given their covariates) to their entry time given their personal entry time, L_i , where survival is estimated using logistic regression, given their covariates. As will be shown, it is age at SSc diagnosis which has the most notable relationship with probability of surviving until study entry.

In Figures 8.11, 8.13 and 8.15, the horizontal black line represents the incident cohort proportion, and therefore what we might expect in the adjusted prevalent

proportion. The black points are the unadjusted (biased) prevalent proportions. The red points are the adjusted proportions based on nonparametric survival estimation, and the green points are the proportion found using the individual survival of patients based on logistic regression-based estimation of survival, Equation (8.2). This will be done for an array of truncation ranges, but we note that the survival curves used here were based on all 1,574 combined incident and prevalent SSc patients. This is because we obtain a better estimate of survival when all patients are utilised. While we are still using categorical covariates below, the logistic regression form does allow for continuous covariates when modelling the GLM and predicting the hazard, so age at SSc diagnosis is modelled as a continuous covariate.

Sex Figure 8.10 shows the probability of survival for each patient until their truncation time, colour-coded for gender. There seems to be no strong, observable relationship between sex and probability of survival to entry. Neither does there appear to be more men at smaller truncation times, nor does a particular sex appear to have higher weightings. From Figure 8.11, there is the implication that there is no strong difference between males' or females' survival, with the unadjusted and adjusted points close together. The logistic regression weightings imply slightly more males than the nonparametric estimator, although the difference is very small. The separation between the unadjusted and adjusted in earlier τ demonstrate that there was a bias towards females due to the lower survival times of males. The logistic regression appears to be closer to the incident proportion than the biased prevalent proportion, however the nonparametric curve is closer. For both adjustment methods, there continues to be a large difference between the incident and prevalent distributions. This could be due to estimates not being accurate, or that there has been a change in covariate distribution over calendar time, with more males now being diagnosed.



Figure 8.10: Truncation time and probability of surviving until that truncation time for each patient, depending on sex of the patient.

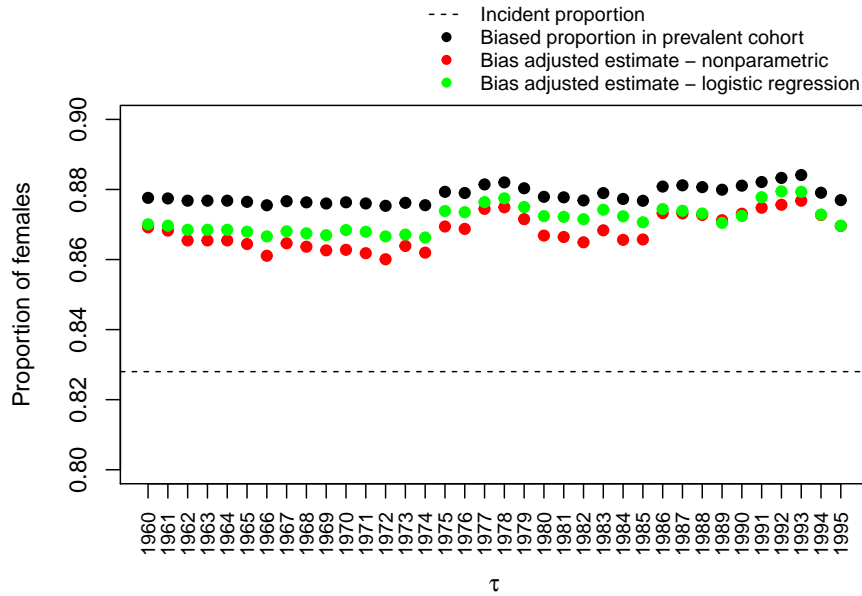


Figure 8.11: Estimation of the baseline covariate proportion for females in the prevalent cohort depending on the hypothetical recruitment start (from when SSC patients are diagnosed), including logistic regression based survival.

Age at diagnosis Figure 8.12 demonstrates the lower survival to entry time for older patients, with many having higher weightings despite short truncation times. Also, we see patients being weighted highly even if they are young, if they have long truncation times. This nicely shows the relationship between a weighting and age, where age is the biggest contributing factor to higher weightings. There are some small differences between the nonparametric and the logistic regression methods, as shown in Figure 8.13, in particular the logistic regression estimations are often closer to the unadjusted proportions than the nonparametric estimates. Note that we potentially stop observing differences with earlier $\tau < 1984$, as the gap between the adjusted and unadjusted stops widening in earlier calendar times. This may demonstrate the lack of patients representing all patients under study at these times. We still see a large difference between the incident cohort and the bias adjusted cohort. It could be that the estimators are insufficient to account for these differences. This could be due to temporal trends, where the incident cohort

has a later age at diagnosis due to improving healthcare and therefore longer survival times before SSc diagnosis.

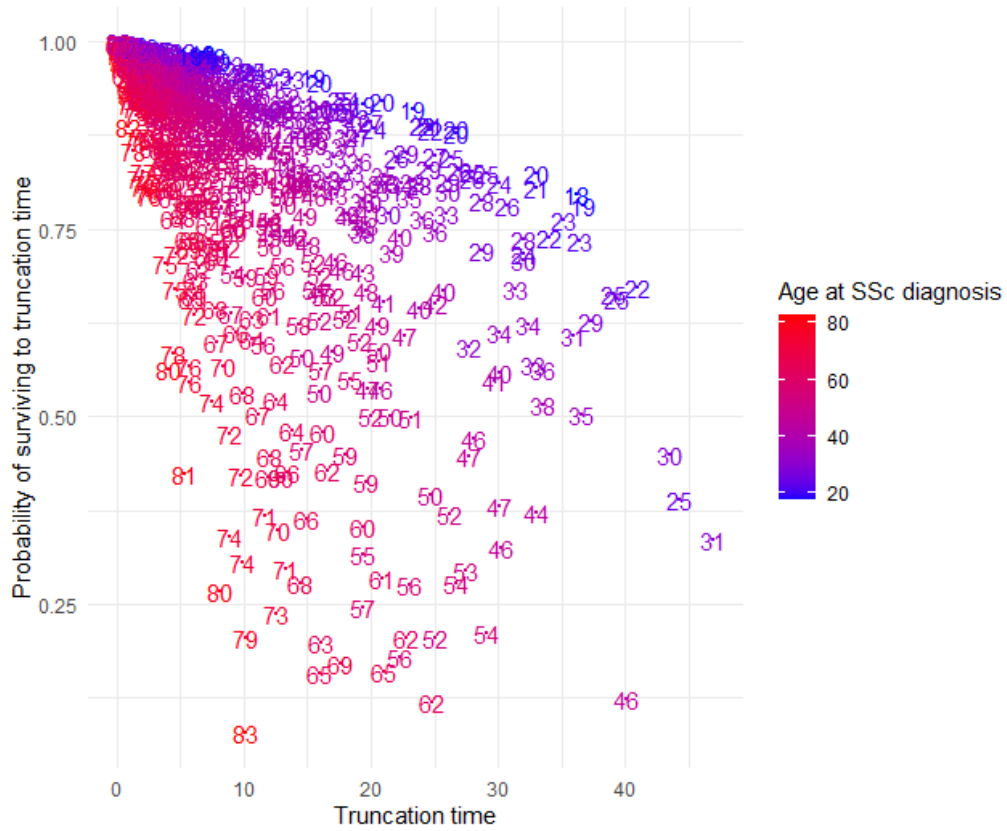


Figure 8.12: Truncation time and probability of surviving until that truncation time for each patient, depending on age at diagnosis. The number shows the exact age of the patient at diagnosis, and the colour of the number also depicts age at SSc diagnosis with blue being younger and red being older.

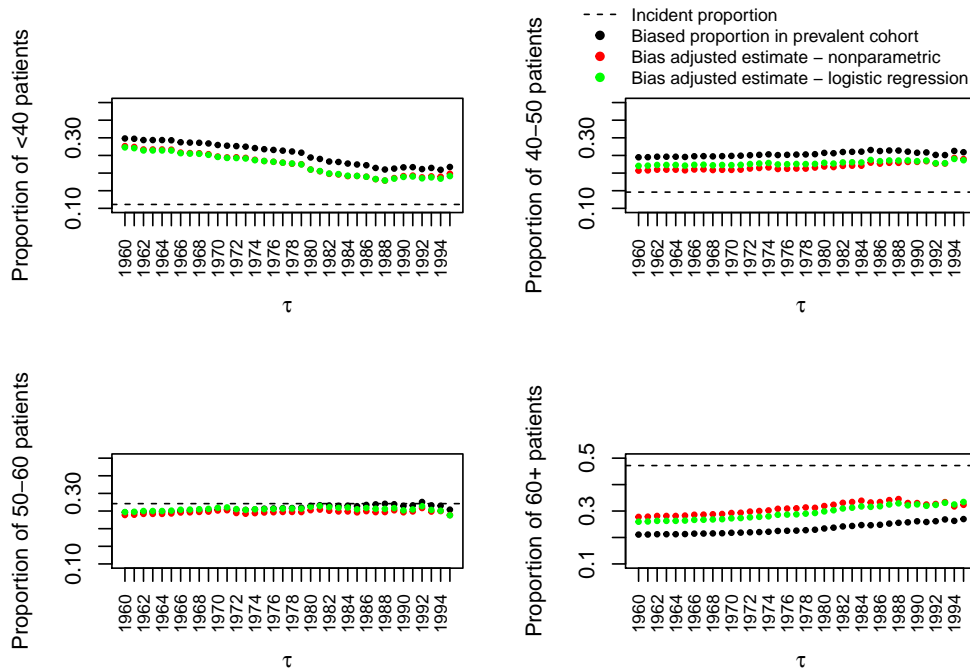


Figure 8.13: Estimation of the baseline covariate proportion for age at SSC diagnosis in the prevalent cohort depending on the hypothetical diagnosis start, including logistic regression based survival.

Smoking habits Figure 8.14 shows the survival until entry stratified by smoking status. Like sex, there appears to be no discernible relationship, as no category appears to only occur at small truncation times, nor does any category appear to have particularly large weightings. There may be the possible indication of more smokers with small truncation times, or ex-smokers with higher weightings, but it is not clear. As shown in Figure 8.15, the adjusted proportions appear to be poor estimates, if we expect the prevalent baseline proportion to be comparable to the true incident proportion. Of the three factors, it is smoking that may have temporal trends, and this may be why we are seeing large differences. We would expect less smokers with increasing calendar time, and the bias of the prevalent is towards more smokers. Similarly the incident cohort has more non-smokers and, as we know non-smokers have better survival, we would expect the bias correction to give less non-smokers, even if this is in

the direction away from the incident cohort. This highlights the need to investigate temporal trends.

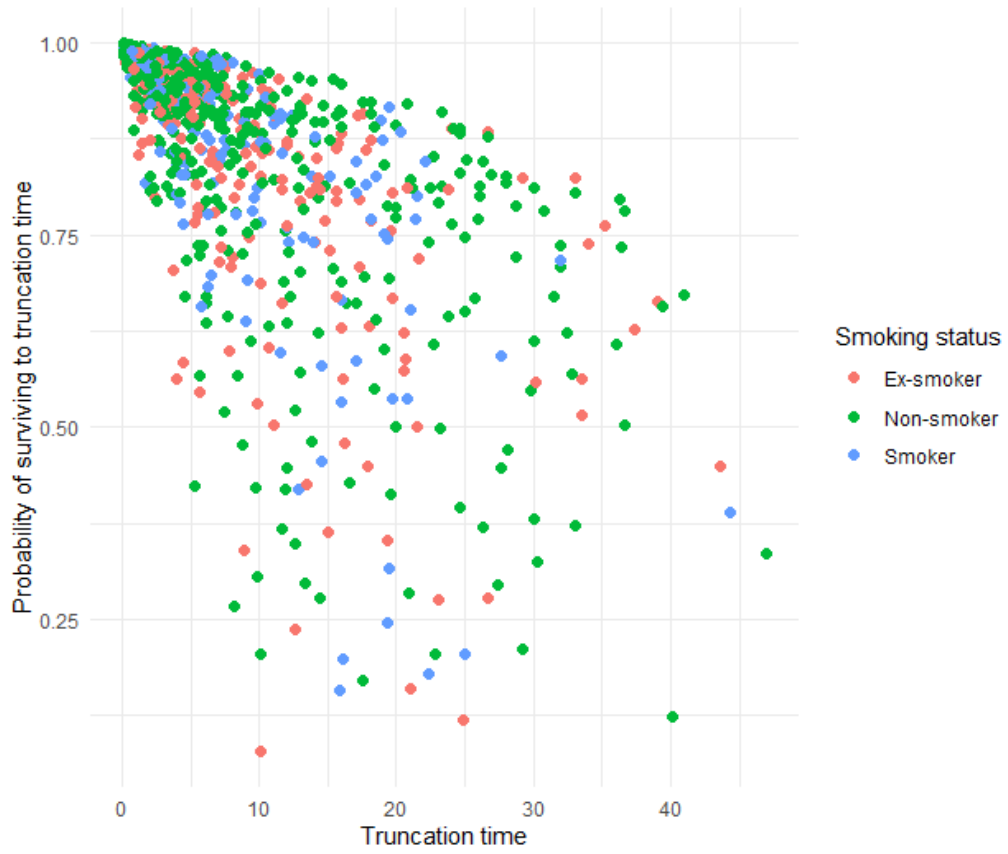


Figure 8.14: Truncation time and probability of surviving until that truncation time for each patient, depending on smoking status.

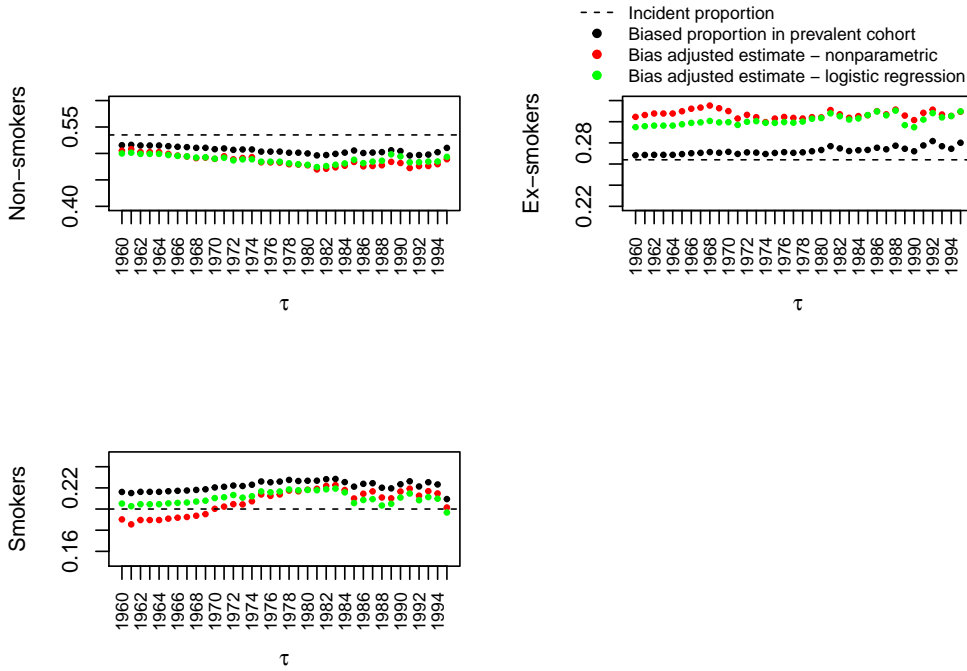


Figure 8.15: Estimation of the baseline covariate proportion for smoking status in the prevalent cohort depending on the hypothetical diagnosis start, including logistic regression based survival.

In reference to the previous trade-off in setting a diagnosis window to avoid extreme weightings but still having a large sample size, we hypothesise that setting a timeframe for diagnoses might result in estimates closer to the incident cohort than if we used all data. This is due to both a) needing all patients of all types under representation and b) less patients who, due to temporal trends, may no longer be representative of the incident cohort even after bias adjustment. Therefore, we provide two tables, one where all SSc patients are used, and one where only SSc patients diagnosed after 1980 are used. This year was chosen due to it being a midpoint between when the first patient was diagnosed and the study start, and leads to a reduction in the number of SSc patients from 1,574 to 1,502 (and to a 9.5% reduction of prevalent SSc patients).

Table 8.6 shows the proportion of each covariate in the incident cohort, the prevalent cohort, and the estimated baseline covariate distribution based on the whole prevalent cohort. Table 8.7 shows the proportion of each covariate in

the incident cohort, the prevalent cohort, and the estimated proportion of the baseline covariate distribution based on a reduced prevalent cohort of only those diagnosed with SSc after 1980. Survival continues to be estimated using only the SSc patients.

	Incident cohort	Unadjusted prevalent cohort	Adjusted prevalent proportion (NPMLE)	Adjusted prevalent proportion (LR)
Gender				
Female	82.8%	87.8%	86.9%	87.0%
Age at diagnosis				
<40 years	11.0%	29.8%	27.7%	27.3%
40-50 years	14.5%	24.5%	20.7%	22.0%
50-60 years	27.3%	24.5%	23.9%	24.7%
60+ years	47.1%	21.1%	27.8%	26.0%
Smoking				
Smoker	20.1%	21.6%	19.0%	20.5%
Ex-smoker	26.4%	26.8%	30.5%	29.5%
Non-smoker	53.5%	51.6%	50.5%	50.0%

Table 8.6: Percentage in each age group for the incident and prevalent cohorts, and then estimated proportions based on NPMLE and logistic regression methods. The prevalent cohort is defined as all SSc patients (i.e. no set window of diagnoses dates).

Despite this small removal of patients, the distribution of adjusted age at SSc diagnosis where only those diagnosed after 1980 are used is much closer to the distribution of the incident cohort than the adjusted full cohort distribution. This could be because survival is not accurate due to risks of cancer or death either increasing or decreasing over calendar time, or the covariate distribution has changed over calendar time. We shall look at this further in the next chapter.

	Incident cohort	Unadjusted prevalent cohort	Adjusted prevalent proportion (NPMLE)	Adjusted prevalent proportion (LR)
Gender				
Female	82.8%	87.8%	86.7%	87.2%
Age at diagnosis				
<40 years	11.0%	24.4%	21.0%	21.0%
40-50 years	14.5%	25.7%	21.9%	22.9%
50-60 years	27.3%	26.5%	25.2%	26.2%
60+ years	47.1%	23.4%	31.9%	29.9%
Smoking				
Smoker	20.1%	22.7%	21.8%	21.8%
Ex-smoker	26.4%	27.3%	30.4%	30.3%
Non-smoker	53.5%	50.0%	47.8%	47.8%

Table 8.7: Percentage in each age group for the incident and prevalent cohorts, and then estimated proportions based on NPMLE and logistic regression methods. Reduced model with patients diagnosed after 1980, with 1502 SSc patients.

8.8 Weighted g-formula estimation using the prevalent cohort

We now apply this theory of weightings to the risk estimates. This section applies the weightings to the prevalent cohort and compares the difference between this and the non-weighted. Figure 8.16 shows the direct effect estimates, and Figure 8.17 shows the total effect estimates. We include three curves in each plot, all using the g-formula method:

- The solid lines (red for SSc, green for non-SSc) are derived using the prevalent cohort only to fit the GLM, and then applied to the incident SSc cohort, so that the incident cohort is providing the covariate distribution (incident application). These curves are the same as those shown in Figure 8.1 and Figure 8.2, and are included for comparison. We might expect the weighted g-formula curves to be close to these.
- The dashed lines (red for SSc, green for non-SSc) are derived from the GLM

being estimated using the prevalent cohort only, and then applied to the prevalent SSc cohort (prevalent application). These curves are an expected underestimation as shown previously in Figures 8.3 and 8.4.

- The dashed orange and blue lines (orange for SSc, blue for non-SSc) are the application of GLM estimated using the prevalent cohort and then applied to a pseudo-population of prevalent SSc patients, with weightings used in accordance with Equation (8.3) and Equation (8.4) (weighted). It is worth remembering that the weighting for non-SSc patients is the same as the weighting for their matched SSc patient, as their entry is conditional on the SSc patient's entry.

For the direct effect, there appears to be a small increase in risk when weightings are applied for both cancer and death without cancer, however only by a small amount. As we have seen from the distribution estimates, Table 8.6, the estimated baseline proportions for the prevalent cohort do not change greatly from the biased prevalent proportion, so this small difference in the curves is expected. This is perhaps surprising. This lack of change could be for a couple of reasons. Firstly, we have an insufficient array of all types of people under study (e.g. male, old and smoker), hence there is an incentive to set a calendar time of recruitment. Repeating the above but only for those diagnosed after 1980 was considered, but as we do something similar in the next chapter it is not included here. Secondly, it could be due to temporal trends, and the survival weightings not sufficiently account for them. It could be that survival has changed over time and therefore survival curves are not accurate. We shall investigate this in the next chapter.

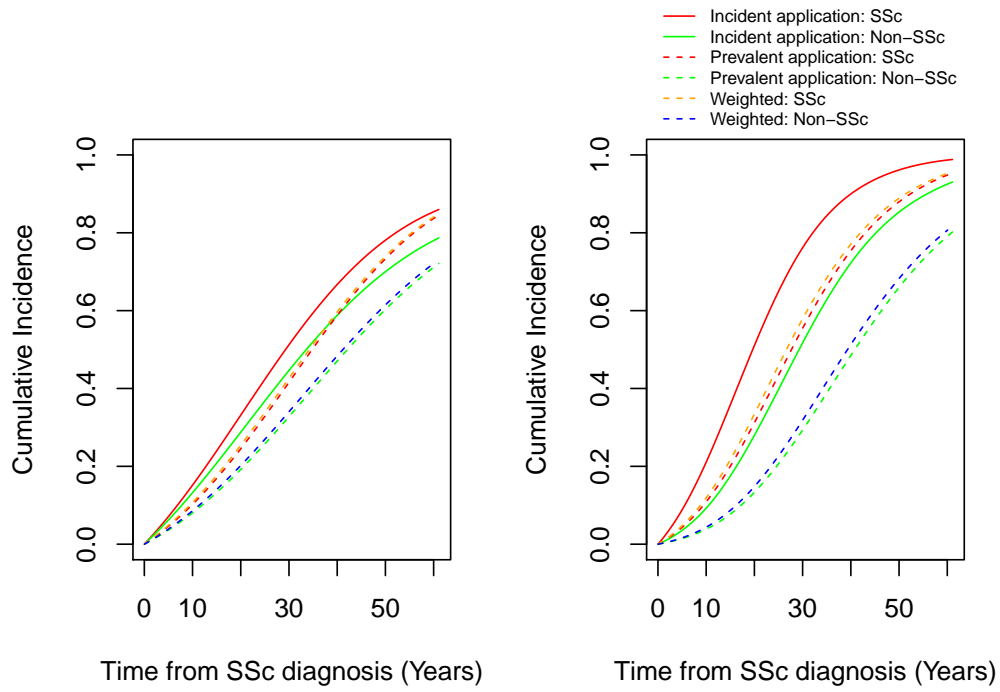


Figure 8.16: G-formula for risk of cancer (left) and death (right), direct effect, with hazard fitted from the prevalent GLM only and applied to the a) incident SSc cohort (solid lines), b) prevalent SSc cohort (dashed red/green lines) and then a weighted prevalent (dashed orange/blue lines).

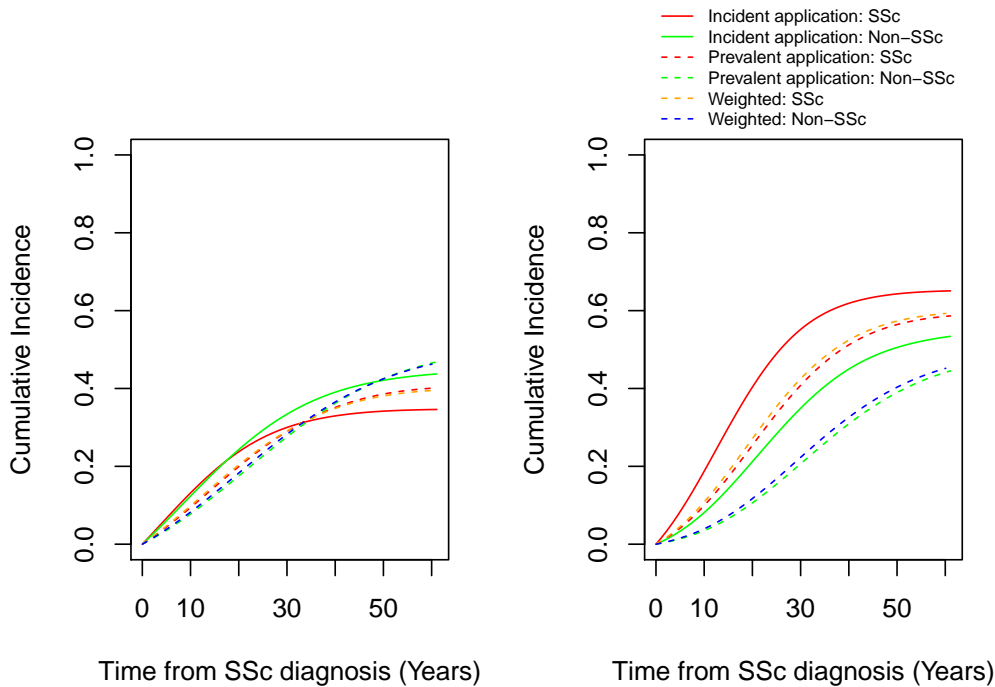


Figure 8.17: G-formula for risk of cancer (left) and death (right), total effect, with hazard fitted from the GLM and applied to the a) incident (solid lines), b) prevalent (dashed red/green lines) and then a weighted prevalent (dashed orange/blue lines), full dataset.

8.9 Weighted g-formula risk estimation using the combined dataset

We lastly produce the combined incident and prevalent curves using our weighted risk estimates, which will utilise all our data, so that both survival and the pseudo-population come from the combined dataset. We would hope, on the assumption that covariate distribution and survival has not changed over time, that by using all data we would increase the sample size and thereby obtain the best estimates.

As a reminder of the theory, each SSc patient will have a survival probability at

each time-step and therefore a weighting, which is the inverse of the probability that they survived until their truncation time given their covariates. Incident patients, who have a truncation time of zero, will have a weighting of 1. Prevalent patients will have a weighting greater than 1. Each patient's cumulative incidence curve will be weighted given the probability of themselves (if they are SSc) or their SSc match (if they are non-SSc) surviving until entry.

Figure 8.18 is for the direct effect and Figure 8.19 is the total effect. The solid lines are the weighted g-formula using the combined dataset to predict the risk, and the combined dataset for the covariate distribution, weighted for left truncation. The dashed lines are the nonparametric estimators using only the combined cohorts.

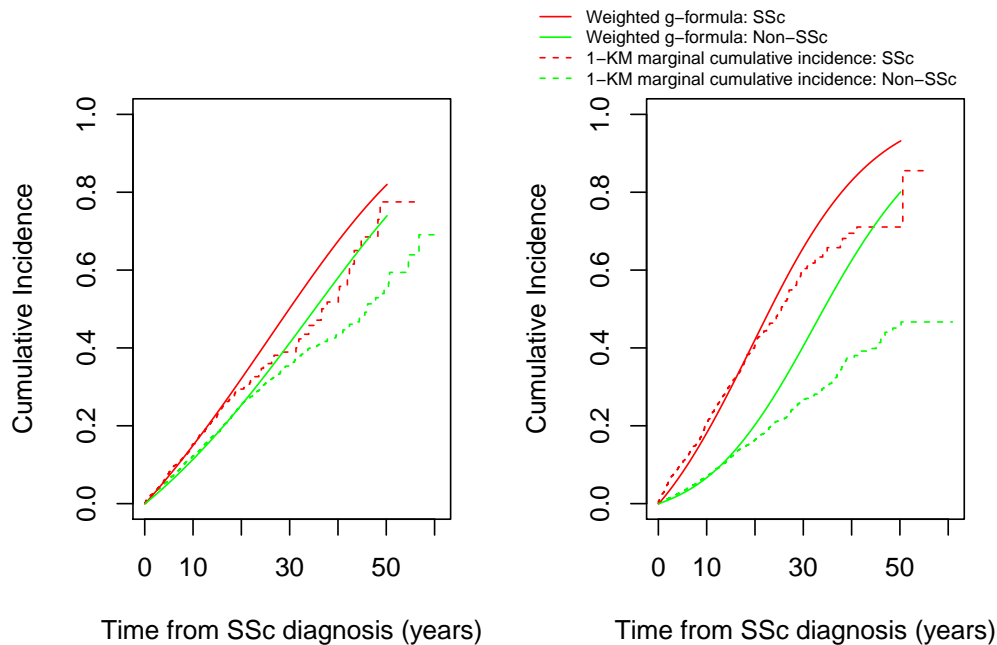


Figure 8.18: G-formula for risk of cancer (left) and death (right), direct effect, with the g-formula using both incident and prevalent cohorts weighted based on survival, and the nonparametric method which is also estimated from the combined dataset.

In Figure 8.18, the two different methods are close to each other at small values, but there is a difference at larger values. This could be due to two reasons. Firstly, this may be an impact of temporal trends where covariates have changed

over time. For example, prevalent patients could be associated with covariates that have lower risk, such as younger age at diagnosis. The g-formula averages these covariates over the whole time under study, however in the NPMLE at longer times from SSc diagnosis we expect prevalent patients to dominate the risk set, hence the appearance of the underestimation of risk at these times in the NPMLE. The second is that the g-formula estimator gives a better estimation in the presence of informative censoring, which has an impact here. This could also be why we see large differences in this figure but less so in Figure 8.19.

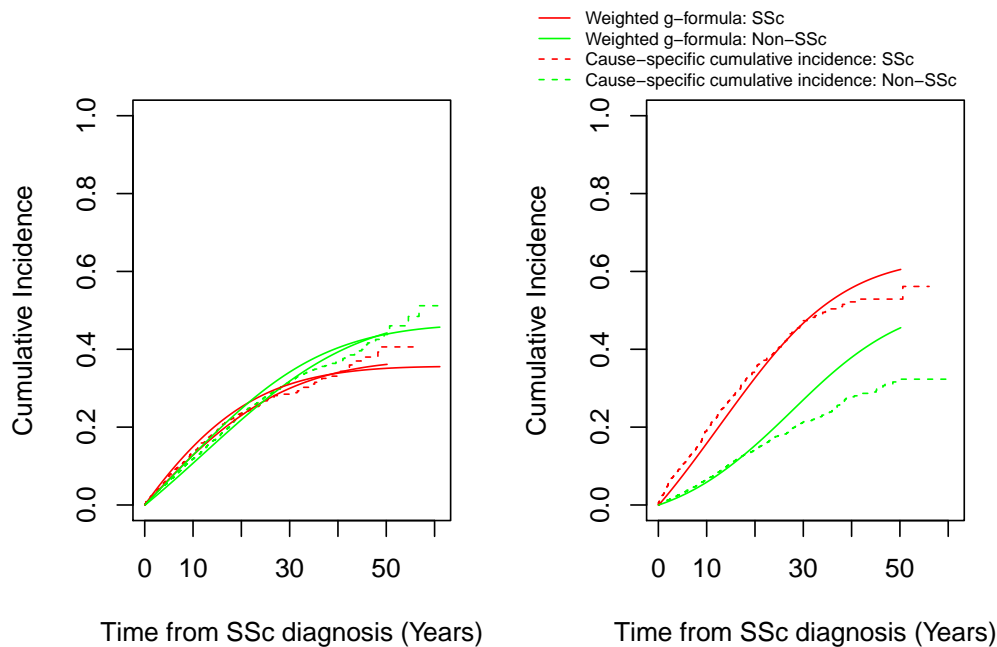


Figure 8.19: G-formula for risk of cancer (left) and death (right), total effect, with the g-formula effect estimator using both incident and prevalent cohorts weighted based on survival probability, and the nonparametric cause-specific cumulative curves which are also estimated from the combined dataset.

Figure 8.19 shows good agreement between the nonparametric and g-formula methods when estimating the risk of cancer. While death without cancer shows a good fit for the non-SSc group at smaller times after SSc diagnosis, the g-formula shows a higher risk at higher times. There are two reasons why we may be observing this:

- The fit may be dominated by events at smaller survival times due to the large number of patients and events at these times, and therefore the fit may be closer at the smaller survival times better than the times further from SSc diagnosis. This may indicate that the fit may not be accurate at these higher times.
- Again, there could be a change in covariates over time. We have seen that even after the adjustment of weightings, we still have a younger average age and more females being represented in the prevalent cohort. As we discuss in the next chapter, this changing covariate profile over calendar time might be a reason why the g-formula shows a higher risk at higher times. If prevalent/historic patients have baseline covariates which are associated with lower risk, and incident/recent patients have covariates which are associated with higher risk, we would expect the two curves to differ. The g-formula utilizes the set of both incident and prevalent covariates as a pseudo-population and then predicts the hazard for all patients over all time of interest. However, the NPMLE will account for the recent patients at early time-points and historic patients at later time-points. Therefore, we might expect the NPMLE curves to indicate higher risk than the g-formula at shorter survival times and the NPMLE curves to indicate lower risk at the longer survival times. This is what we observe for death for both the SSc and non-SSc curves.

8.9.1 Summary table

The following tables present the risk ratios for cancer for the combined weighted g-formula effect estimator compared to the NPMLE at selected time-points. Despite the large difference in curves for the SSc and the non-SSc for the first method of including prevalent patients (Tables 7.3 and 7.4) and the second method of weighting the g-formula (Tables 8.8 and 8.9), the risk ratios are very similar, for both the direct and total effect.

Again, from the direct effect it appears that there is a small causal impact of SSc

Time from SSc diagnosis (years)	Left truncation Kaplan-Meier [95% CI]	Weighted g-formula [95% CI]
10	SSc 0.152 [0.126, 0.178] Non-SSc 0.122 [0.112, 0.132] Ratio 1.247 [1.031, 1.507]	SSc 0.150 [0.137, 0.163] Non-SSc 0.114 [0.107, 0.120] Ratio 1.315 [1.148, 1.507]
20	SSc 0.294 [0.252, 0.334] Non-SSc 0.253 [0.237, 0.269] Ratio 1.163 [0.998, 1.355]	SSc 0.321 [0.3, 0.342] Non-SSc 0.253 [0.243, 0.264] Ratio 1.267 [1.130, 1.421]
30	SSc 0.400 [0.335, 0.459] Non-SSc 0.362 [0.334, 0.389] Ratio 1.105 [0.930, 1.313]	SSc 0.502 [0.479, 0.524] Non-SSc 0.413 [0.401, 0.424] Ratio 1.215 [1.111, 1.330]

Table 8.8: Cumulative incidence risk ratios for cancer, comparisons for the left truncated Kaplan-Meier and direct effect g-formula methods, using both the incident and prevalent data for both methods. The 95% confidence intervals are given in brackets.

on cancer, with none of the confidence intervals containing one. The total effect implies that those with SSc have a slightly higher risk of cancer in the first few years, however this equalizes over later years to the point where death shows a greater discrepancy. From this we achieve very similar conclusions to the previous chapter, that there is a small causal impact of SSc on cancer, however this may not always be observable due to death being at such an increased risk in SSc patients.

Time from SSc diagnosis (years)	Left truncation cause-specific cumulative incidence [95% CI]	Weighted g-formula [95% CI]
10	SSc 0.136 [0.115, 0.162] Non-SSc 0.118 [0.109, 0.128] Ratio 1.153 [0.954, 1.393]	SSc 0.128 [0.116, 0.140] Non-SSc 0.107 [0.101, 0.114] Ratio 1.192 [1.037, 1.371]
20	SSc 0.235 [0.206, 0.270] Non-SSc 0.234 [0.219, 0.249] Ratio 1.008 [0.868, 1.171]	SSc 0.230 [0.212, 0.249] Non-SSc 0.219 [0.209, 0.229] Ratio 1.052 [0.923, 1.198]
30	SSc 0.290 [0.253, 0.332] Non-SSc 0.319 [0.297, 0.342] Ratio 0.909 [0.780, 1.058]	SSc 0.300 [0.277, 0.322] Non-SSc 0.318 [0.307, 0.330] Ratio 0.942 [0.831, 1.068]

Table 8.9: Cumulative incidence risk ratios for cancer, comparisons for the Aalen-Johansen and total effect g-formula methods using both the incident and prevalent data for both methods. The 95% confidence intervals are given in brackets.

8.10 Alternative method: IPCW with left truncated data

This section discusses how the IPCW could be altered to include left truncated data, however in this thesis we have focused on the g-formula and there could be further work done on this aspect, which we shall highlight. The IPCW does not have the benefit of using a pseudo-population of our choice (having a pseudo-population will be beneficial when discussing temporal trends in the next chapter). It does, however, have the advantage of more clearly demonstrating the difference between the IPCW and the unweighted (for censoring) Kaplan-Meier.

A few works have mentioned left truncation in combination with informative censoring, however all use slightly different approaches. Shen (2003) discusses IPW methods when using left truncation in the discussion section of their paper. The discussion section cites both the above Robins & Rotnitzky (1992) and

Datta & Satten (2002) papers, and how dependent/informative censoring can be accounted for, however Shen does not suggest the method we use below, and says “it is not clear how to estimate $\Lambda_c[t|Z_i(t)]$ for left truncated and right-censored data”. This paper suggest modelling the hazard of censoring on the residual time scale², where the hazard can then be estimated using the Cox model or Aalen’s additive hazards model. They then propose simultaneously modelling the left truncation and event distributions, $G(\ell)$ and $F(t)$, possibly through an iterative process. Allignol et al. (2014) discuss dependent censoring and left truncation in an illness-death model format. Vakulenko-Lagun et al. (2021) mention Robins & Finkelstein (2000) and suggest using this form of IPCW to correct for dependent truncation (not dependent censoring), with higher weights for those who are more likely to be left truncated.

Despite the absence of using the IPCW with prevalent cohorts in prior literature, the most natural method to estimate the hazard, h_k , would be to continue to only include patients in the risk set from the time they enter the study. We also continue to model the hazard of loss to follow-up for the weighting as specified in Section 8.2, where patients continue to only contribute to the fitting of the hazard if they are in the risk set.

We propose the IPCW direct effect estimator for left truncated data is as follows,

$$\sum_{k=0}^K \hat{h}_k(a; \hat{\alpha}, \hat{\eta}) \prod_{j=0}^{k-1} \left[1 - \hat{h}_j(a; \hat{\alpha}, \hat{\eta}) \right] \quad (8.6)$$

where

$$\hat{h}_k(a; \hat{\alpha}, \hat{\eta}) = \frac{\sum_{i=1}^n y_{k+1i} (1 - y_{ki}) w_{ki}(\hat{\alpha}, \hat{\eta}) I(a_i = a) I(L_i > k)}{\sum_{i=1}^n (1 - y_{ki}) w_{ki}(\hat{\alpha}, \hat{\eta}) I(a_i = a) I(L_i > k)} \quad (8.7)$$

and

²Residual time scale is the time from when a patient enters the study to their event/censoring in the study, as opposed to the original time scale where time is measured from disease onset to event. This is covered more in Qian & Betensky (2014).

$$w_{ki}(\hat{\alpha}, \hat{\eta}) = \frac{\prod_{j=0}^k (1 - c_{j+1i})(1 - d_{j+1i})}{\prod_{j=0}^k (1 - r(a, \bar{z}_{ji}, j; \hat{\alpha}))(1 - q(a, \bar{z}_{ji}, j; \hat{\eta}))} \quad (8.8)$$

where $r(a, \bar{z}_{ji}, j; \hat{\alpha})$ and $q(a, \bar{z}_{ji}, j; \hat{\eta})$ are models for the hazard of loss to follow-up and death, respectively, indexed by parameters. The hazard for the cause-specific loss to follow-up is $\Pr[C_{j+1} = 1 | \bar{Z}_j = \bar{z}_{ji}, \bar{C}_j = \bar{D}_j = \bar{Y}_j = 0, A = a]$ and competing event hazards, $\Pr[D_{j+1} = 1 | \bar{Z}_j = \bar{z}_{ji}, \bar{C}_{j+1} = \bar{D}_j = \bar{Y}_k = 0, A = a]$, with $\hat{\alpha}$ and $\hat{\eta}$ as consistent estimators of α and η , respectively, with notation $(a_i, \bar{z}_{Ki}, \bar{c}_{K+1i}, \bar{d}_{K+1i}, \bar{y}_{K+1i})$ as individual i 's values of $(A, \bar{Z}_K, \bar{C}_{K+1}, \bar{D}_{K+1}, \bar{Y}_{K+1})$.

Similarly, the following is the risk for the total effect, with estimates based on the observed cause-specific hazards of the event of interest and the competing event,

$$\sum_{k=0}^K \hat{h}_k^1(a; \hat{\alpha})(1 - \hat{h}_k^2(a; \hat{\alpha})) \prod_{j=0}^{k-1} [(1 - \hat{h}_j^1(a; \hat{\alpha}))(1 - \hat{h}_j^2(a; \hat{\alpha}))] \quad (8.9)$$

$$\hat{h}_k^1(a; \hat{\alpha}) = \frac{\sum_{i=1}^n y_{k+1i}(1 - y_{ki})(1 - d_{ki})w_{ki}(\hat{\alpha})I(a_i = a)I(L_i > k)}{\sum_{i=1}^n (1 - y_{ki})(1 - d_{ki})w_{ki}(\hat{\alpha})I(a_i = a)I(L_i > k)} \quad (8.10)$$

$$\hat{h}_k^2(a; \hat{\alpha}) = \frac{\sum_{i=1}^n d_{k+1i}(1 - y_{ki})(1 - d_{ki})w_{ki}(\hat{\alpha})I(a_i = a)I(L_i > k)}{\sum_{i=1}^n (1 - y_{ki})(1 - d_{ki})w_{ki}(\hat{\alpha})I(a_i = a)I(L_i > k)} \quad (8.11)$$

and

$$w_{ki}(\hat{\alpha}) = \frac{\prod_{j=0}^k (1 - c_{j+1i})}{\prod_{j=0}^k (1 - r(a, \bar{z}_{ij}, j; \hat{\alpha}))} \quad (8.12)$$

Observe that there is no change in the formula for weight from that in Section 7.5 for the solely incident case. Patients only have an event or contribute to the risk set for the hazard if they were in the dataset at the previous time point. As we have already shown in Section 8.2, we can estimate the hazard of loss to follow-up, C , and the competing event, Y or D , with GLMs for the prevalent cohort.

However, we need to work under a set of assumptions that may not hold for

our study, particularly pertaining to independence assumptions between factors, which we discuss more in the next chapter. We have to assume that a prevalent patient’s event time is independent of their study entry time, $L \perp T|L \leq T$, so that their weighting is independent of their time of entry into the study. Also, that $C \perp L|L \leq C$, where C is time from diagnosis to loss to follow-up. Having both of these hold will allow for exchangeability, and are both required for both hazard prediction for w_{ik} and the MLE derivation for h_k (Section 8.2).

We suspect these do hold in our study due to either covariates changing over calendar time or the hazard conditional on covariates changing over calendar time. The majority of loss to follow-up occurs in the final year, or close to the final year when practices last updated their records to CPRD GOLD, however there are a small number of patients in the study who leave due to emigration or due to the practice failing to record the end event. In many studies it is necessary to model loss to follow-up as some patients may be more likely to leave a study (e.g. patients observed in a hospital environment may leave the hospital if they are ‘healthier’ patients). However, there is no connection between our covariates (age at diagnosis, sex or smoking) and any explanation for them leaving the study. Our loss to follow-up occurs mostly in the last few years of the study, so if prevalent patients are associated with certain characteristics and also experience more cancer or death events, then these characteristics may be associated less with loss to follow-up, failing $C \perp L|L \leq C$. Incident patients may inadvertently be weighted higher, but to what extent is not investigated here due to lack of time. As was said in Section 7.5, we do not feel the need to model loss to follow-up as other studies might, but if we observe a loss to follow-up effect it may be an effect of left-truncation and a violation of an above assumption.

We find that a GAM has a slightly smaller AIC for modelling censoring than a GLM, therefore we shall use a GAM (98643 GAM compared to 98576 GLM). Once more we also include an interaction term of time and age, reducing the AIC to 49430, however no other interaction minimised the AIC. Table 8.10 shows the results of modelling loss to follow-up via a GAM for the non-smoothed

terms for the combined prevalent and incident cohort. Table 8.11 shows the smoothing terms and the approximate significance of smooth terms. Figure 8.20 demonstrates the smoothing over each covariate.

All factors other than smoking are now significant predictors compared to the results of Section 7.5. This could be due to the inclusion of more patients, increasing the sample size and therefore the accuracy, and indeed younger people or males are more at risk of leaving the dataset. Also, hypothetically, it could also be due to changing covariates over calendar time. We have already observed above that the prevalent cohort differs to the incident in that it is on average younger at diagnosis and has more females, and weightings did not appear to overcome this gap between the incident and prevalent covariate proportions which imply these differences are intrinsic. We have also observed that prevalent patients have more events and less loss to follow-up, due to our loss to follow-up occurring mostly in late 2017. Therefore, we may be introducing bias by accounting for loss to follow-up, as we no longer have independence between left truncation time and censoring time.

	Loss to follow-up	[95% Confidence interval]
Intercept	-4.8300	[-4.8615, -4.7984]
SSc (A=1)	-0.0973	[-0.1629, -0.0318]
Male	0.1109	[0.0476, 0.1741]
Smoker	0.0284	[-0.0273, 0.0842]
Ex-Smoker	0.0151	[-0.0387, 0.0690]

Table 8.10: Coefficient values for censoring logistic regression, estimated using the incident cohort only. The baseline for A is non-SSc, the baseline for sex is male and the baseline for smoking and ex-smoker is non-smoker.

	edf	Ref.df	Chi.sq	p-value
s(k)	3.607	4.579	90.18	< 2e-16
s(Age)	6.556	7.684	109.06	<2e-16
s(k*Age)	7.760	8.605	192.50	<2e-16

Table 8.11: Approximate significance of smooth terms, estimated using the incident cohort only.

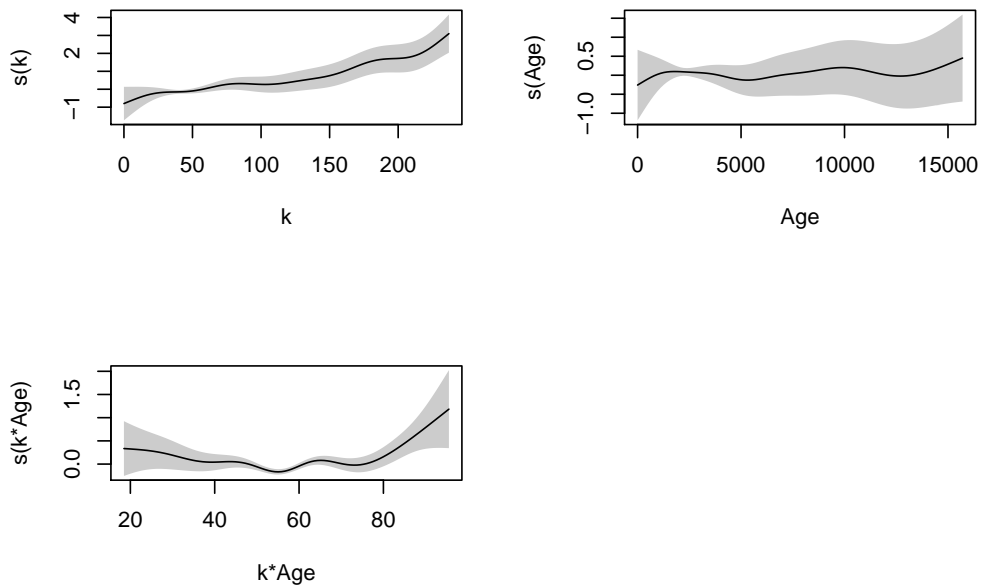


Figure 8.20: Plots of the smoothing functions used to model the hazard of censoring. The dark grey shading is the 95% confidence intervals.

Figure 8.21 and Figure 8.22 show the IPCW results compared to the NPMLE methods where there are no weightings for censoring. The full model for the direct effect is weighted for the hazard of loss to follow-up and the competing event, therefore in order to assess the impact of these weightings we also include the IPCW where loss to follow-up is not included as a weighting but the competing event is ($r(a, \bar{z}_{ij}, j; \alpha) = 0$ for all time j); these are shown in orange and blue. We show these separate curves with different censoring types to distinguish which type of censoring is having an impact.

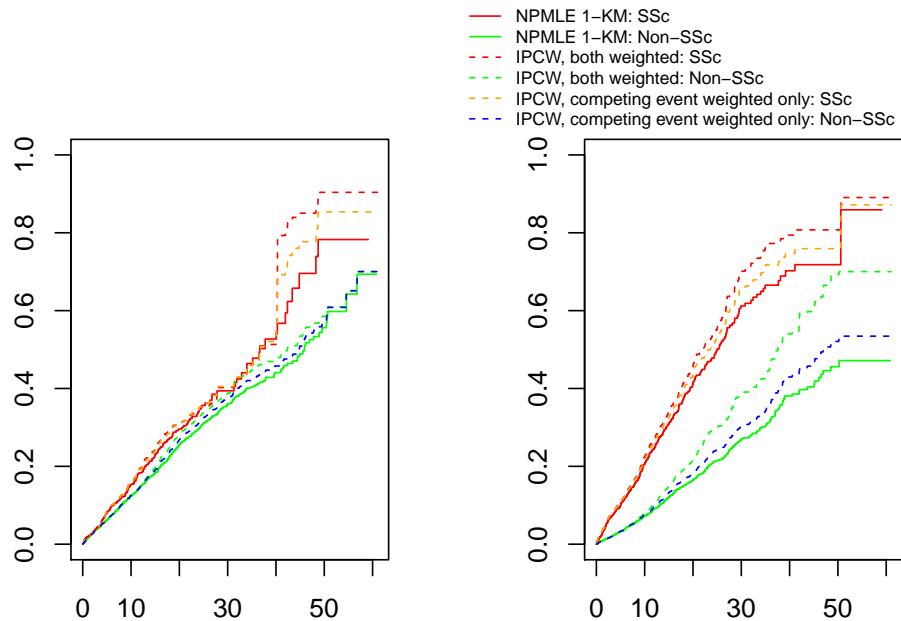


Figure 8.21: IPCW for risk of cancer (left) and death (right), direct effect. The hazard is fitted from the GLM of the combined incident and prevalent cohorts and we then show the a) unweighted/unadjusted cause-specific cumulative incidence (solid red/green), b) IPCW including a weighting for loss to follow-up and competing event (dashed red/green lines) and c) IPCW where loss to follow-up is not weighted but the competing event is (dashed orange/blue).

The direct effect, Figure 8.21, for cancer (left) in SSc patients shows hardly any difference between the curves except for a large jump at time 41 years after SSc diagnosis. With so few patients in the risk set at this time it may be a consequence of unstable weightings (Hernan et al., 2000). The most noticeable difference is the small increase in death with the addition of weightings for non-SSc patients. All the curves are similar (but less smooth) to the results of the weighted g-formula, Figure 8.18. We would expect these figures to be similar as they are both accounting for informative censoring.

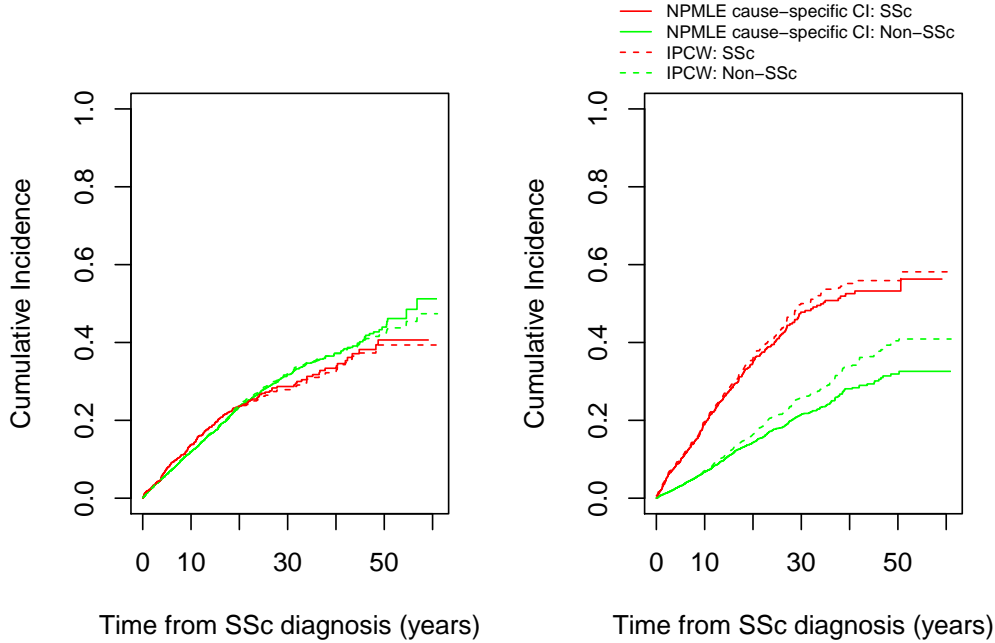


Figure 8.22: IPCW for risk of cancer (left) and death (right), total effect, with hazard fitted from the GLM and applied to the a) unweighted, b) IPCW (dashed red/green lines).

The total effect, Figure 8.22, demonstrates that weighting for loss to follow-up gives the appearance of increased risk for death. However, we might not expect this to be the case, as our loss to follow-up is purely administrative. We hypothesise that we are inaccurately accounting for something else, and therefore introducing bias. It may be that a change in covariates over time is contributing to this difference.

It may be that our method here is insufficient, as by modelling loss to follow-up, which we believe to be administrative only, we may be introducing bias. Although there may now be a relationship between covariates and censoring due to more patients improving the power of the study, it may be that there needs to be an extra weighting for temporal trends (or in other studies, dependent truncation). With more time, we would have liked to have proved an original formula for left truncation to verify this method, however we have specifically focused on the g-formula estimator for this thesis, as it has allowed for more flexibility. The

g-formula has the advantage of a different pseudo-population based on the incident SSc cohort, but with the hazard being predicted with prevalent data. Due to the way we are using the ATT method, the g-formula may be useful to other researchers if their dataset needs to adjust for differences in covariates to find an average effect or average effect of treatment.

In the next chapter we shall discuss how the g-formula can be altered to allow for temporal trends. The IPCW is harder to adjust for and this has been left to later research. Regardless of the drawbacks summarised above, the IPCW may be preferred in other studies for its simpler interpretation and its similarity to other methods. It may also be easier to adjust for time-varying covariates.

8.11 Summary

This chapter has demonstrated the use of g-formula estimation on our SSc prevalent dataset.

The first method presented, using the prevalent dataset to inform the GLM but then only applying the predictions to the incident cohort, allows both incident and prevalent cohorts to be used, increasing the sample size to fit the GLM, but based on a dataset whose covariates are more reflective of patients diagnosed in more recent times. This method was possible for our study due to the large number of incident patients, however other studies may not benefit from this. We note the possibility that knowledge of patient characteristics of an incident cohort could be used as a basis for a pseudo-dataset if the data for an incident cohort is not available but a prevalent cohort is. Also, this method would not suffer from the same problems put forward by Pan & Chappell (1999). A slightly greater risk was observed in the prevalent cohort compared to the incident cohort, however this did not change the conclusion of results from the previous chapter.

For the second method, an incident baseline covariate distribution estimator was used to weight the g-formula such that patients who are underrepresented in the study are given higher weightings. This is a new method, and the hope

is that this can provide incentive for the inclusion of prevalent patients in a causal framework. The estimated baseline covariate distributions for the adjusted prevalent proportions were not as significant as possibly expected when compared to the incident, which could be due to a limitation of the model, or temporal trends over time being present in our study. The results for the weighted prevalent curve when compared to the incident curve were quite different, however we suspect this is due to natural differences between the prevalent and incident datasets, possibly due to the aforementioned temporal trends, for example changes in survival probability or covariate distribution over time. The possibility of temporal trends incentivised the next chapter.

When the incident and prevalent cohorts are combined to both predict hazard and to be used as the pseudo-population, we see close similarities between the nonparametric curves and the g-formula curves, with a small deviation at later times. This effect at later times could be due to informative censoring consideration in the g-formula. As can be seen from Table 8.8 and Table 8.9, there is little change in the conclusion from previous chapters in that there is a direct effect of SSc on cancer, but limited total effect due to the higher mortality in those with SSc.

The use of the IPCW for a left truncated cohort was discussed with a method suggested, however as loss to follow-up is not something that is necessary to model in this thesis, the formal proof of this method is left to further work.

Further analysis could have been done to deduce if there was any statistical difference in the weighted g-formula to the non-weighted g-formula (Figure 8.16 and 8.17), however we have not done so as this report is particularly concerned with quantification of risk of cancer in SSc rather than methodology analysis.

In the next chapter, we look at temporal trends in nonparametric models and one application in the g-formula.

Chapter 9

Accounting for temporal trends

9.1 Introduction

With the inclusion of our prevalent cohort, we now have a long follow-up time between SSc diagnosis and an event. In this study, dates of SSc diagnoses range over a large number of years, from 1957 to 2018. There is a possibility that there could be intrinsic differences between those diagnosed in later years compared to those diagnosed in earlier years. Diagnosis in earlier years may not be representative of the current relationship between the exposure and an outcome. The possibility of these differences has been highlighted by the greatly differing covariate distributions in the incident and the prevalent cohorts after adjustment using the Chan & Wang (2012) method, as was seen in Chapter 8. If there are temporal trends, then the survival times observed will be associated with the patient's calendar time information (DOB, diagnosis date, entry year), whether this is due to characteristics of the cohorts varying over calendar time or caused by differences arising from the calendar period patients live through due to changes in lifestyle and/or medicine.

It has been shown that the risk-set adjusted Kaplan-Meier estimator for left truncation (and by extension the risk-set adjusted Aalen-Johansen cause-specific estimator) accounts for 'length-biased' data when we have independent truncation

(Woodroffe (1985), Tsai et al. (1987)), however, these methods may not be accurate if there is not independent left truncation. Due to the long duration of our dataset, there is a risk that including patients with large truncation times will not reflect the current risk of cancer or death without cancer. Adjusting for calendar time is common in studies (Cole et al. (2013), Murray et al. (2018)), however the concern is that the left truncated nature of the dataset means the relationship may not be as easily observable as if we had a solely incident cohort. How best to account for temporal trends is dependent on what investigators wish to learn from the data:

- If we wish to use our dataset for prognostic reasons for application in future healthcare, then the more recently diagnosed patients are likely to better represent the current risk, and patients born/diagnosed further back in calendar time may introduce bias. The use of historic cases (most likely to be our prevalent patients) strengthens our data by increasing sample size. If, due to temporal trends, cancer risk and mortality in the historic cases are significantly different to patients who were more recently diagnosed then we will be introducing a set of patients inconsistent with the incident SSc patients in the database, distorting our study. Therefore, if we decide to include these patients, we will need to make appropriate adjustments for temporal bias. This is often termed *baseline survival* in the literature, but we shall term this **recent risk** to avoid confusion with the use in previous chapters of baseline covariate distribution (M.-C. Wang et al., 1993). Even though the method originated in a paper on left truncated data, the method we will be using is not specific to left truncated data, and could be used on any longitudinal incident data.
- However, if we are interested in the risk of cancer over all the calendar years covered by the database, either from a retrospective perspective or to contrast with an estimate of current risk, then we would wish to quantify the marginal/average cumulative incidence over all patients, including the

source population represented by the prevalent cohort: for example, in our study we would want the average risk over 1957 to 2018. Due to potential patients being unobservable due to the selection bias (we only observe patients whose survival time exceeds their truncation time), incident cases will have a greater presence than the historic cases. Our theory is therefore weighting based, where we assign greater weight to historic patients to represent unobservable patients. The adjusted cumulative incidence curve has been termed *marginal cumulative incidence* in the literature (Stegherr et al., 2020), however we shall term this **historic risk** to avoid confusion with the term ‘marginal cumulative incidence’ used in this thesis in the setting of competing risks. The methods we shall use are specific to prevalent data, as the marginal estimator needs to account for unobserved patients which are not observed in the prevalent cohort.

- While not an original aim of the study, it would be of interest to observe how the risk of cancer (and death without cancer) have changed over calendar time. Prior research into SSc has highlighted a possible decrease in mortality compared to the general population (Mayes et al. (2003), Steen & Medsger (2007), Butt et al. (2018)), however one study found no significant difference (Elhai et al., 2012). There has not been a sufficiently large study of SSc and cancer to investigate a changing risk of cancer over time, however we note that even if there were to be a sufficiently large study we may observe an increase, decrease or no change at all. There may, for example, be a decrease in cancer incidence due to better healthcare in the general population, but there could be an increase if cancer detection has improved.

There are a handful of studies which investigate dependent truncation in prevalent cohorts, but consideration of temporal trends (and conditionally independent truncation) is rarer. The methods from two papers described later in this chapter appear to have been applied only occasionally. The papers themselves each apply their own methods but this has been done in the context of dependent truncation, not conditionally independent truncation. The analysis in this thesis is therefore

often overlooked in other studies, and is particularly novel in the application to a dataset as large as ours.

In this chapter, we describe why it is important to consider temporal trends and we then test for temporal trends in our dataset. We apply modifications to our dataset for these trends to enable prevalent cohort inclusion using nonparametric weightings. This modification, similar to Chan & Wang (2012) from the previous chapter, will weight an observation by the reciprocal of its inclusion probability. We then include temporal trends for our g-formula estimator, our estimator of choice in the previous chapters.

9.2 Dependent truncation and the quasi-stationarity assumption

Many studies that involve left truncated data may have independent truncation, where there is no association between entry time and event time. Some studies may have an association, implying that knowledge of entry time will provide us with information on their event time.

Let L be time between SSc diagnosis date (or match date) and entry date, let T be the time between SSc diagnosis date and event date, and let Z be patient covariates. Independent truncation is where the entry time does not affect event time, $L \perp T$. It could be that there is dependent truncation, where the entry time affects the event time. An example of this is given in M.-C. Wang et al. (1993), where patients with a prior diagnosis of AIDS were given a drug at study start and monitored for survival. With time 0 being infection with AIDS and time L being time from infection to treatment (and simultaneously entry to study), there was a possible negative correlation between entry time, when they were given treatment, and event time, as patients with later entry times may have been sicker/frailer when they entered the study. These authors hypothesise that this could be a declining effect of treatment in that treatment is most effective used closer to initial infection. However, another reason for a relationship between entry

time and event time would be conditionally independent truncation, $L \perp T|Z$, due to underlying confounders which may result in the appearance of dependent truncation. For us, each patient’s entry date is an arbitrary date at which they were first available to be recruited, so we do not expect the entry date to causally affect the event time and therefore do not have dependent truncation. However, if we have ‘temporal trends’ then we may expect to observe an association between entry time, L , and calendar date of diagnosis, Z_{Cal} .

Keiding & Moeschberger (1992) demonstrate that if we have independence between T and L conditional on Z the hazard of an event T given $T > L$ similarly reduces to the independent case,

$$\frac{\Pr[T = t|L = \ell, Z = z, T > L]}{\Pr[T \geq t|L = \ell, Z = z, T > L]} = \frac{\Pr[T = t|Z = z]}{\Pr[T \geq t|Z = z]}$$

i.e. the hazard of T does not depend on entry time if covariates are accounted for. They also give a good example of frailty, where an unmeasured confounder with a positive association with entry time and survival time has been introduced into the underlying model, leading to a decreasing hazard with larger truncation times. They note that if the covariates are measured, this can be accounted for in the Cox model. However, accounting for Z when using the Kaplan-Meier is not mentioned.

We recognize ‘temporal trends’ may be an umbrella term for possible confounders, and we use the covariate calendar time of SSc diagnosis as a proxy. These could include:

- Measured patient covariates - for example, either a change in proportion of sex, age at SSc diagnosis or smoking habits over calendar time, or a change in hazard relating to these covariates. For the former, it could be that more men are now diagnosed with SSc, or we have less smokers, or SSc is being diagnosed at an advanced age due to longer lifespans. An example of the latter would be if a new treatment for males was implemented.

- Unmeasured patient covariates - we acknowledge that we may not have information on all covariates which could impact temporal trends, such as a new national implementation of treatment of SSc, but the hope is that the method detailed below, by using calendar time, will account for changes in unmeasured confounders.
- External, environmental factors - there could be changes in cancer or mortality hazard that are not specific to our SSc dataset but to the UK as a whole, such as improved cancer screening or healthier lifestyles. This would mean historic cases no longer represent the current hazard of SSc patients or the general population.

We assume that calendar time, Z_{Cal} , is representative of these changes over time in the models for this chapter. Some reasons for observed changes in calendar time in our study may be:

- Changes in cancer risk in SSc patients. Reasons for increased cancer risk in those with SSc were noted by Weeding et al. (2020). If there has been a change in SSc treatment such as more use of cytotoxic therapies or immunosuppressive drugs we might observe changes in cancer risk. There may have also been an increase in ionizing radiation from SSc treatment. If an increase in cancer is present in SSc patients but not non-SSc patients, it may be something specific to SSc patients.
- Changes in cancer risk in the general population. There could be many different factors that may have an effect here. Some types of cancer are now picked up earlier because of screening, so possibly more cases at younger ages will be found, but as people now live to greater ages we may also observe more cases in older patients as well. As mortality in those with SSc has improved over time, due to improved treatment, we may observe more cancer events at older ages due to longer survival. This may interact with our competing events over time. Possible lifestyle changes may decrease the risk of cancer in the general population, such as better diet, less smoking and

better health regimes. It will be hard to detect which of these differences has the most impact.

- Changes in mortality in the general population. Patients born in 1930 will have had different healthcare than someone born in 1980, for example. We would expect mortality to decrease over calendar time, therefore inclusion of historic cases will lead an overestimation of risk if we wished to estimate the recent risk.
- Changes in mortality in those with SSc. As there have been improvements in recent years in the treatment of SSc, the SMR is believed to have decreased over calendar time, and the leading cause of death has changed from renal failure (the predominant cause of death) to pulmonary complications (Steen & Medsger (2007), Rubio-Rivas et al. (2014)). Therefore we may expect SSc patients further back in calendar time to have higher mortality, and their inclusion would lead to an overestimation of risk if we wished to estimate recent risk.
- Changes in known covariate distribution. If the covariate distribution of those with SSc has changed over time (for example, the average age at SSc diagnosis has increased over time due to longer survival times in the general population) then this will increase the risk of cancer and death in those with SSc. Therefore we might expect higher risks in the recent cases compared to the prevalent/historic cases. This would appear as a temporal trend if these covariates were not adjusted for. Also, we have seen that there has been a reclassification of SSc (see Section 2.1), which may now result in SSc being diagnosed at different ages, and with possibly less severe cases now being recorded. Depending on which covariate distributions have changed, the inclusion of historic patients could lead to an over- or under-representation of recent risk. Healthier smoking habits may lead to lower mortality today, therefore the application of only historic patients would lead to an overestimation of both cancer risk and mortality. A greater age at SSc diagnosis would lead to a higher mortality in the more recent cohort, so the inclusion of historic patients would be an

underestimation. It should be noted that this does not include covariate differences that we see between prevalent and incident cohorts naturally due to length biased left truncation. For example, we expect a lower average age at SSc diagnosis in the prevalent cohort due to those with lower ages having better survival, but this difference is accounted for in the weighting of the risk-adjusted Kaplan-Meier estimation, on the assumption that the prevalent and incident cohorts have the same baseline covariate distribution (See Section 5.6, Simulation 1, 2 and 3).

Often, prior literature in this area looks specifically into dependent left truncation. Although we are working with conditionally independent truncation, we review dependent truncation due to an overlap in the methodology we will be using.

9.2.1 Dependent truncation

Before we discuss working with calendar time of diagnosis as a covariate, we shall give an overview of dependent truncation, as it ties in heavily with prevalent cohorts and is used in our work.

During the formulation of Kaplan and Meier, the authors briefly discussed the validity of their method for left truncated data (Kaplan & Meier, 1958). The use of the Kaplan-Meier and the Aalen-Johansen cause-specific cumulative incidence estimators in the presence of independent (or random) left truncation can be found in fundamental statistical textbooks (for example Aalen et al. (2008)), as the estimator used can be altered by including the patients in the risk set once they enter the study, as was done in Chapter 5.

As in previous chapters, let L denote time to study entry (left truncation time), T be time to final event. Many of the studies do not include competing risks, therefore we assume one outcome. Let C be time from SSc diagnosis to censoring and $X = \min(T, C)$. In general, dependent truncation cannot be tested for, due to the unobserved truncated data. However, quasi-independence between T and L could be ascertained. Tsai (1990) defines quasi-independence as

$$\Pr(T \leq t, L \leq \ell | L \leq T) = \int \int_{\Delta(t, \ell)} d\Pr(T \leq u) d\Pr(L \leq v) / \alpha$$

where $\alpha = \Pr(L \leq T) = \int \int_{u \geq v} d\Pr(T \leq u) d\Pr(L \leq v)$, $\Delta(t, \ell) = \{(u, v) | u \leq t, v \leq \ell, u \geq v\}$.

It should be noted that many articles which involve dependent truncation make assumptions about how censoring interacts with truncation, and the independence between left truncation time, event time and censoring time, (Qian & Betensky, 2014). Distinctions are made in this article depending on whether the residual time scale or the original time scale is used. We have the original time scale as we are measuring risk from SSc onset to an event, not entry time to an event. When considering the original nonparametric form for left truncation (left truncated Kaplan-Meier), the two assumptions that were made were $\Pr(L \leq C) = 1$ and $T \perp (L, C) | L \leq T$. The first assumption is that patients cannot be censored until they enter the study. The second assumption is that event time is independent from both entry time and censoring time, i.e. entry time is not correlated with event time, and event time and censoring time are independent given covariates. However, we believe that we do not have $\Pr(L \leq C) = 1$ but $\Pr(L \leq C) < 1$ due to emigration from the dataset. We make the assumption that firstly this level is low and also that it is mostly random/non-informative. Regardless, as stated in Qian & Betensky (2014), the left truncated NPMLE survival estimator ($F(t)$) still holds for $\Pr(L \leq C) < 1$ but not the left truncation distribution ($G(t)$) where $\Pr(L \leq C) = 1$ is required. In regards to $T \perp (L, C) | L \leq T$, we assume that $T \perp C | L \leq T$ holds if we do not account for informative censoring/competing risks.

One of the more common methods to test for quasi-independence is the conditional Kendall's tau test and its generalisations (Tsai (1990), Martin & Betensky (2005)). However, recent research by Vakulenko-Lagun et al. (2019) suggests that this test may incorrectly suggest that the null hypothesis holds, and that there is no dependent truncation.

As a reminder of previous notation: $F(t) = \Pr(t \leq T)$ and $G(\ell) = \Pr(\ell \leq L)$, and $f(t)$ and $g(\ell)$ are the respective densities.

As stated in Vakulenko-Lagun et al. (2019) there is a factorisation condition,

$$\Pr(\ell \leq L < \ell + d\ell \in d\ell | T = t) = dA(\ell) = a(\ell)d\ell \quad (\ell < t), \quad (9.1)$$

where $a(\ell) \geq 0$ need not equal $g(\ell)$ and is defined on the support of L in the observable region $L < T$. This is a weaker assumption to the above quasi-independence which would allow for the nonparametric estimation of the distribution of T . However there is another factorisation condition

$$\Pr(t \leq T < t + dt | L = \ell) = dA^*(t) = a^*(t)dt \quad (\ell < t), \quad (9.2)$$

where $a^*(t) \geq 0$ need not equal $f(t)$. Condition (9.1) and Condition (9.2) combined make up quasi-independence, and quasi-independence implies both Condition (9.1) and Condition (9.2). There is a non-identifiability issue. If Condition (9.1) holds, then $f(t)$ is identifiable, but not necessarily $g(\ell)$. If Condition (9.2) holds, then $g(\ell)$ is identifiable but not necessarily $f(t)$. Kendall's tau can test the null hypothesis that either of the conditions hold, against the alternative hypothesis that neither hold. Therefore if there is insufficient evidence to reject the null hypothesis, we cannot be sure which condition holds, and if it is only Condition (9.2) then we cannot estimate survival. This is an evolving research area, and demonstrates the questioning of a method traditionally used since 1991.

Another recommended method, and possibly the most popular, is Keiding and Gill where truncation time is included as a covariate in the Cox model to test quasi-independence (Keiding & Gill, 1990). However, as we are applying the Cox model here we have to work under the proportionality assumption, that the relative hazard remains constant over time for the covariates used, including calendar time.

9.2.2 Quasi-stationarity

M.-C. Wang et al. (1993) discuss the assumption of quasi-stationarity. This is the assumption that is made for the hazard function $\lambda(t; \tau_1, \tau_2) = \lambda_0(t)$ where $\lambda_0(t)$ is the baseline hazard, therefore the hazard function $\lambda(t; \tau_1, \tau_2)$ is independent of the calendar time of exposure, τ_1 , and the calendar recruitment time, τ_2 , for $T \geq L$. If there are temporal trends, this quasi-stationarity assumption would not hold as the calendar time of exposure would be associated with the hazard.

M.-C. Wang et al. (1993) allows for a relaxation of this quasi-stationary assumption. The method is a follow-on from the previously published method of Keiding and Gill where truncation time is included as a covariate in the Cox model to test quasi-independence (Keiding & Gill, 1990). We can test the underlying assumption using a transformation of (τ_1, τ_2) as covariates in the proportional hazards model:

$$\lambda(t; \tau_1, \tau_2) = \lambda_0(t) \exp(\phi(\tau_1, \tau_2)\beta)$$

where β is the coefficient and ϕ is a specified function. It could be that we set $\phi(\tau_1, \tau_2) = \tau_1$ to investigate calendar time of diagnosis. If $\phi(\tau_1, \tau_2) = \tau_2 - \tau_1$ then we would be investigating truncation time, and if β is negative then as truncation time increases the hazard decreases. Quasi-stationarity holds if $\beta = 0$. Whether truncation time or calendar time effects are present can be tested by the hypothesis test $H : \beta = 0$, which can be tested by the significance of the p-value(s) by the partial score test statistic. This will be the method we use to test for temporal trends in our data.

9.3 Testing quasi-stationarity in the SSc dataset

We will use Wang's methodology to test for a violation of the quasi-stationarity assumption using calendar time of SSc diagnosis (M.-C. Wang et al., 1993). We have the option of treating the covariate of calendar time as continuous or

categorical. We would use continuous if we believed that changes over calendar time occur gradually. This would be the case if there was a gradual improvement in healthcare, for example. We could also include calendar time as a categorical variable, which would be preferable if there was a sudden change in survival. It is not clear which is more appropriate for us, as while we expect mortality to decrease over calendar time, we know from prior knowledge that there was a sudden change in survival due to the use of ACE inhibitors among those diagnosed with SSc, which reduced the risk of death from scleroderma renal crisis (SRC) (Steen & Medsger, 2000). Therefore we look at calendar time of diagnosis as both a continuous and categorical covariate. We also include additional covariates in the analysis to separate the effects of calendar time from other covariates.

The following analysis will include both incident and prevalent cohorts, increasing our sample size and accuracy with more recent data. Let W be the covariate ‘calendar time’, which will quantify the difference between a set calendar time (such as most recent or study start) and a patient’s calendar date of diagnosis. We set our baseline for W as the last date anyone in our study was diagnosed with SSc, which is 18/09/2017, so patient i will have calendar time $W_i = (18/09/2017 - \text{Date of diagnosis of patient } i)$. Being diagnosed further back in time will increase W , so a patient who was diagnosed on 18/09/2016 will have covariate $W = 1$, as they were diagnosed a year prior. Therefore, if we have a hazard ratio greater than one when using the Cox model, the higher calendar time implies a higher hazard, indicating worse hazard in historic cases. A hazard ratio smaller than 1 indicates smaller hazard of cancer (or death) in those who were diagnosed further back in calendar time.

9.3.1 Continuous calendar time and no additional covariates

Table 9.1 contains estimates of the effect of calendar time of diagnosis on our SSc and non-SSc groups, and then combined, using the Cox model with calendar time as a continuous covariate. This is done for our two outcomes: cancer and death without cancer. Calendar time is shown per ten years, so for example the

hazard ratio shown is the ratio attributable to being diagnosed 10 years earlier. The p-values on the right of the table are results from the Schoenfeld residual test to test for the proportionality assumption. When both SSc and non-SSc cohorts are used, SSc status is also included as a covariate (hazard ratio for SSc type not shown here).

	HR	95% CI	p-value
Cancer, Both SSc and non-SSc	0.74	[0.66, 0.84]	0.53
Cancer, SSc	0.77	[0.59, 1.02]	0.79
Cancer, Non-SSc	0.74	[0.65, 0.84]	0.59
Death without cancer, Both SSc and non-SSc	0.83	[0.72, 0.94]	0.29
Death without cancer, SSc	0.86	[0.69, 1.09]	0.59
Death without cancer, Non-SSc	0.79	[0.67, 0.92]	0.29

Table 9.1: Cox regression model for calendar time of diagnosis per 10 years, with coefficient and hazard ratio (HR) and 95% confidence intervals, left, along with the global Schoenfeld residual test with p-values, right. Grey highlights the statistically significant hazard ratios.

In non-SSc patients, the hazard ratios for both cancer and death without cancer are statistically significantly less than one, suggesting recent cases have a higher hazard than those diagnosed further back in calendar time. This is not necessarily what we would expect for death, as from prior knowledge we know survival in the general population has increased over calendar time. However, we note that these results may be due to differing covariate distributions over calendar time. We discuss this more later, after including other covariates. Also, the confidence interval for cancer hazard is almost significant for SSc patients, however this group has wider confidence intervals than the non-SSc patients, and it may be that the sample size is too small to detect a difference at the 5% level. Note that none of these models fail the proportional hazard assumption.

9.3.2 Categorical calendar time with no additional covariates

We repeat the above but instead use calendar time of SSc diagnosis as a categorical variable. We split patients into those diagnosed between 1957-1990 (13.0% of all patients), 1990-1995 (11.2% of all patients), 1995-2000 (16.4% of all patients) and 2000+ (59.4% of all patients). The last group will be our reference group/baseline factor. The reason for the large number of patients in the last category is that there are less events and more censoring in the later time periods, so to ensure a sufficient number of events in the baseline category we need more patients.

Table 9.2 shows the hazard ratios for cancer and death for SSc and non-SSc.

	HR	95% CI	p-value
Cancer, SSc			
Calendar category: 1995-2000	0.79	[0.52, 1.18]	0.56
Calendar category: 1990-1995	1.00	[0.61, 1.62]	
Calendar category: 1957-1990	0.77	[0.38, 1.56]	
Cancer, Non-SSc			
Calendar category: 1995-2000	0.70	[0.59, 0.84]	0.84
Calendar category: 1990-1995	0.60	[0.47, 0.76]	
Calendar category: 1957-1990	0.57	[0.42, 0.78]	
Death without cancer, SSc			
Calendar category: 1995-2000	1.21	[0.88, 1.66]	0.13
Calendar category: 1990-1995	1.05	[0.70, 1.59]	
Calendar category: 1957-1990	0.62	[0.34, 1.13]	
Death without cancer, Non-SSc			
Calendar category: 1995-2000	0.81	[0.65, 1.03]	0.08
Calendar category: 1990-1995	0.95	[0.72, 1.24]	
Calendar category: 1957-1990	0.66	[0.45, 0.97]	

Table 9.2: Cox PH regression model, calendar time of diagnosis as categorical variable, with coefficients and hazard ratio (HR) with 95% confidence intervals, left, along with Schoenfeld residual test with p-values for each category, right. Grey highlights the statistically significant hazard ratios.

The SSc groups continue to show little evidence of temporal trends. This is again perhaps surprising, considering there is evidence to suggest survival in those with SSc has been improving over calendar time (Steen & Medsger, 2007). Again, it

may be that we have too small a sample size, producing wide confidence intervals. In the non-SSc patients for cancer there is a declining hazard with increasing calendar time, implying patients further back in time had lower hazard of cancer diagnosis. It is only the most historic cases that are statistically significant for death without cancer in those without SSc, and the confidence intervals are wide for all categories, hence there is uncertainty in these results. Note that none of these failed the proportional hazards test (right of table).

9.3.3 Continuous calendar time with additional covariates

We now add previously used covariates alongside continuous calendar time, to observe any impact on calendar time once these other covariates are accounted for (Table 9.3). If changes over calendar time are due to a change in covariate distribution, then we may observe that the hazard ratio for calendar time of diagnosis is closer to one when other covariates are considered.

None of the covariates fail the Schoenfeld goodness-of-fit test, either individually or globally (global results not included here).

We first consider non-SSc patients. When covariates are not included, we previously observed a decreased hazard ratio for both cancer and death without cancer in the more historic cases of non-SSc patients. However, once covariates are included this continues to be the case for cancer (although to a lesser extent) but for death without cancer we observe a reversal in the results at a statistically significant level, as now the historic cases indicate a worse survival. The change in hazard ratio from below one to above one for death without cancer once covariates are included may mean that calendar time is associated with a change in other covariates. In fact, it may be changes in patient covariate distributions of SSc over calendar time that we are detecting, as opposed to changes in external factors (e.g better testing, better healthcare).

When considering SSc patients, the hazard for cancer has the same direction as non-SSc patients, where earlier calendar times have lower risk. However,

	HR	95% CI	p-value
Cancer, SSc			
Calendar time (10 years)	0.84	[0.64, 1.12]	0.85
Age at SSc diagnosis (10 years)	1.52	[1.34, 1.72]	0.22
Sex (Male)	1.42	[0.98, 2.06]	0.98
Ex-smoker	1.30	[0.94, 1.78]	0.69
Smoker	1.27	[0.88, 1.85]	
Cancer, Non-SSc			
Calendar time (10 years)	0.87	[0.77, 1.00]	0.90
Age at SSc diagnosis (10 years)	1.62	[1.53, 1.71]	0.05
Sex (Male)	1.27	[1.07, 1.50]	0.68
Ex-smoker	0.99	[0.85, 1.15]	0.81
Smoker	1.23	[1.05, 1.44]	
Death without cancer, SSc			
Calendar time (10 years)	1.05	[0.83, 1.32]	0.72
Age at SSc diagnosis (10 years)	2.32	[2.15, 2.70]	0.51
Sex (Male)	1.14	[0.88, 1.70]	0.07
Ex-smoker	1.11	[0.86, 1.48]	0.29
Smoker	1.65	[1.22, 2.24]	
Death without cancer, Non-SSc			
Calendar time (10 years)	1.24	[1.05, 1.47]	0.82
Age at SSc diagnosis (10 years)	3.42	[3.15, 3.73]	0.26
Sex (Male)	1.17	[0.94, 1.45]	0.35
Ex-smoker	1.26	[1.04, 1.53]	0.87
Smoker	2.79	[2.31, 3.36]	

Table 9.3: Cox proportional hazard regression model, calendar time of diagnosis as a continuous covariate, with hazard ratio and 95% confidence intervals on the left and Schoenfeld test p-values on the right, where the p-value assigned to ex-smoker is for the smoking category as a whole. The baseline for sex and smoking are female and non-smoker, respectively. Grey highlights the statistically significant hazard ratios.

the confidence intervals are wide and therefore this is not at a significant level. Mortality does not seem to have changed over calendar time with a hazard ratio close to 1.

That we may be seeing a different ‘temporal trend’ due to changes in measured covariates can possibly be seen in Section 8.3. On the right of Figure 8.1, the prevalent regression applied to the incident pseudo-population for cancer shows slightly lower risk than the combined regression applied to the incident pseudo-population, which agrees with what we observe in Table 9.3. Also, when the baseline covariate distributions were estimated using weightings, as shown in Tables 8.6 and 8.7, the average age was still younger in the prevalent cohort than was observed in the incident cohort, indicating patients are now diagnosed with SSc at later ages.

For death in non-SSc patients, the decreased risk when other covariates are not accounted for are most likely due to covariate changes over calendar time, where younger age at SSc diagnosis and more females will lead to patients with better survival in historic cases. The implication is males and older patients are now diagnosed proportionally more in recent cases. We have already observed from the Chan and Wang baseline estimation of the previous chapter that there is still a large difference in covariate distributions between the incident and prevalent cohorts even with weightings. Once covariates have been taken into account, the increased hazard in historic cases is consistent with what is expected, as we would hope healthcare and average lifespan improves over calendar time. This is certainly true for non-SSc patients, but less so for those with SSc. This may still be due to a small sample size or may suggest a lack of progress in improving SSc mortality.

For cancer, we continue to see a decreased hazard in historic cases, even when covariates are taken into account, however the differences are smaller. Therefore, again, we may be observing a decreased risk as historic patients are younger, and if we continue to see a decreased risk after other covariates have been adjusted for it may be due to better diagnoses practices (and therefore more frequent diagnoses)

of cancer in more recent times. SSc patients have a similar calendar time hazard ratio for cancer to the non-SSc patients.

Some limitations of the above models are noted here. Firstly, due to an expected high correlation between calendar time and age at diagnosis, the model may not separate out the effects for these two correctly (collinearity). Another concern is that the confidence intervals are quite wide, and even significant results are close to 1. In particular, these results are not significant for SSc patients. While we have covered three important additional covariates, there is also the possibility of important associated covariates, such as general lifestyle habits (such as exercise, food consumption) that were not collected for this study.

The results from Table 9.3 do not detract from the need to adjust for calendar time in the NPMLEs (naive Kaplan-Meier and cause-specific cumulative incidence) using the hazard ratios **without covariates**, however, as the NPMLE is biased because historic cases have better survival (even if this improved survival arises due to the differing covariate distributions) these differing covariates still need to be adjusted for, as we shall now do.

9.4 Accounting for conditionally independent truncation using weightings

If the quasi-stationarity assumption is violated, whether through dependent truncation or changes over calendar time, then we can make adjustments using weightings. To do this appropriately we need to consider whether our risk of interest based on target population is:

- The risk at baseline, or what we term **recent risk**, which would be the risk of the average SSc patient at either truncation time $L = 0$ (if we wished to account for dependent truncation time) or calendar time of SSc diagnosis $W = 0$ (for example, in our dataset we could set $W = 0$ to be the last date of diagnosis in our dataset). Here we will use $W = 0$ and will term this

‘recent cumulative incidence’. Adjusting for calendar time of SSc diagnosis is not unique to prevalent cohorts (or left truncated data) and is simply used to account for calendar time.

- What we term **historic risk** (but is termed the marginal distribution in other work) is the average cumulative incidence over the whole calendar time investigated. For our study, we have patients who were diagnosed with SSc between 1957 and 2000, so we require a marginal estimated risk that also accounts for unobserved patients in these early years whose survival time was too short to survive until study entry, i.e. the average risk over these years. Let us take an example: if we saw better survival at later years of calendar time, there would be fewer patients observed than expected with large truncation times. Due to the correlation between truncation time and survival time, we might expect longer times of survival to be predominantly prevalent patients, and shorter times of survival to be incident patients. Therefore we might observe an unadjusted left truncation Kaplan-Meier that would overestimate the average survival soon after SSc exposure compared to the average person over the calendar time of interest, and likewise an underestimate of survival at larger times after SSc exposure. Marginalizing over all patients allows us to find the average risk of all patients over the calendar timeframe of interest. However, we then have unobserved prevalent patients which would mean the average is biased towards more recent patients. Therefore we also need to include weightings to account for the unobserved patients. We term the corrected risk estimate the ‘historic cumulative incidence’.

The recent risk will be most useful from a prognostic perspective, as it will be the risk of the most recent set of patients, and is therefore the one we are most interested in. The historic risk would be useful for those who wish to look over a previous time period. It may also be useful to compare how the risk of cancer has changed over calendar time by comparing the recent and historic risks. We acknowledge that differences in the curve may not be due to a change in hazard

conditional on covariates, but the distribution of measured baseline covariates (e.g. age at SSc diagnosis) that might have changed, such as a younger average age at SSc diagnosis in the more historic cases. Therefore, the historic cumulative incidence should not be mistaken for the cumulative incidence that would be observed if our cohort of recent patients who were diagnosed later in calendar time had lived through an earlier time period.

We are using methodology developed by M.-C. Wang et al. (1993), Mackenzie (2012) and Stegherr et al. (2020). Wang et al. first suggested the Breslow estimator to estimate the recent risk. Mackenzie first estimated the historic risk in relation to dependent left truncation for one event outcome, and it was subsequently developed by Stegherr et al. for competing risks. These last two papers use similar methods to each other but different notation, and therefore for the purpose of consistency of methods we use notation similar to Stegherr et al. Both methods are originally formulated to account for dependent left truncation, however it can be easily seen that this can be reinterpreted to calendar time, such as birth year, calendar time of exposure or calendar time of entry. The methods in both papers work by modelling the conditional distribution of the event time given the truncation variable (or calendar time variable) using the Cox model and then averaging weighted covariate-specific survival functions. There are therefore similarities in method between this and the weighting of the g-formula from the previous chapter. The use of these methods for the adjustment of calendar time is novel.

For consistency of notation, we shall use $L = \ell$ for truncation time and $W = w$ for calendar time of diagnosis (and baseline calendar time of diagnosis being $w = 0$). When estimating survival weightings they shall be conditional on calendar time, $\hat{\Pr}(t < T|W = w)$, but the methodology is similarly applicable for left truncation, or indeed any other covariate which affects recruitment time. We describe the method when we have one outcome of interest, and then for multiple outcomes of interest based on Stegherr et al.

9.4.1 One outcome event

M.-C. Wang et al. (1993) suggests the recent risk estimator method under the ‘relaxation of the quasi-stationarity assumption’. For this, we use the Breslow estimator, which is used to estimate the baseline hazard/survival. If each subject i has a calendar time of SSc diagnosis w_i then, by fitting a Cox regression model, we assume that $\lambda_i(t) = e^{\beta \cdot w_i} \lambda_0(t)$. The cumulative baseline hazard for the baseline of calendar time $w = 0$, which term recent risk, can be estimated using this Breslow estimator,

$$\hat{\Lambda}_0(t) = \sum_{s \leq t} \frac{\Delta N(s)}{\sum_{i=1}^n R_i(s) \exp(\hat{\beta} w_i)} \quad (9.3)$$

where $N(t)$ is the total number of events in the interval $[0, t]$ and $\Delta N(t) = N(t) - N(t_-)$, where $N(t_-)$ is the number of events up until the previous time-step. The risk set at time t is defined as $R_i(t) = \mathbb{I}(L_i < t \leq T_i)$, and $\hat{\beta}$ is the coefficient estimated using the Cox model for the hazard depending on calendar time of SSc diagnosis. We have used w_i here for the calendar time of patient i , however it could also be left truncation time, or it could be used for categorical covariates of calendar time (or left truncation time) as well (i.e. $\exp(\hat{\beta}_i \phi_i)$, where ϕ_i is the category type for patient i).

The cumulative hazard of each patient with calendar time of diagnosis can be estimated as $\hat{\Lambda}(t|W = w_i) = \hat{\Lambda}_0(t) \exp(\beta w_i)$. Therefore survival given patient i 's calendar time is $\hat{S}(t|W = w_i) = \exp(-\hat{\Lambda}(t|w_i))$. In the one event setting the cumulative incidence is $\hat{F}(t|W = w) = \hat{\Pr}(T \leq t|W = w) = 1 - \hat{S}(t|W = w)$. The survival of a patient at baseline is therefore $S(t|W = 0)$, which we take to be our **recent risk estimator**.

For the historic risk estimation, we wish to marginalise over the cumulative incidences of all patients to find the average risk over the calendar time period of interest. However, due to left truncation we know that our sample has a biased population, with more patients with recent dates of SSc diagnosis.

Therefore marginalising over these patients will not account for those diagnosed further back in calendar time who are not observed. We therefore need to give larger weightings to patients who have longer truncation times who have lower probability of surviving until entry into the dataset. We estimate survival conditional on their calendar time of diagnosis. Suppose that we only observe n out of m patients over the calendar time period of interest, for us 1957 to 2018. The marginal inclusion probability is $Q = \Pr(T > L) \approx \frac{n}{m}$, where an individual represents $1/Q$ of the total patients. We can estimate $\hat{m} = \sum_{i=1}^n \frac{1}{\hat{P}(T > \ell_i | W = w_i)}$. Then the marginal probability of being included in the study is $\hat{Q} = \frac{n}{\hat{m}}$. The left truncation distribution (as first seen in Section 5.4.2) is estimated as

$$\hat{G}_L(\ell) = \left(\sum_{i=1}^n \frac{1}{\hat{P}(T > \ell_i | W = w_i)} \right)^{-1} \sum_{i=1}^n \frac{\mathbb{I}(\ell_i \leq \ell)}{\hat{P}(T > \ell_i | W = w_i)} \quad (9.4)$$

and the **historic cumulative incidence** estimator is

$$\hat{P}(T \leq t) = \left(\sum_{i=1}^n \frac{1}{\hat{P}(T > \ell_i | W = w_i)} \right)^{-1} \sum_{i=1}^n \frac{\hat{P}(T \leq t | W = w_i)}{\hat{P}(T > \ell_i | W = w_i)} \quad (9.5)$$

The form of Equation (9.5) is similar to the theory we have previously used from Chan & Wang (2012), where patients are weighted higher if they are less likely to survive until their entry time. However, we cannot use Equation (9.5) in our study to estimate the historic direct effect, as we have the possibility of ‘censoring’ due to death occurring prior to study entry. Therefore we shall justify Equation (9.9) instead, although without further corrections this still lacks the ability to adjust for informative censoring.

For the above to be valid, we assume that the hazard estimated in the Cox model follows a Cox proportional hazards model. This assumption is also only testable on the observable region $T > L$ and remains untestable on the region $T \leq L$. There may be a higher variability of results when there are patients with high weightings who represent many more patients who are missing, as highlighted in the last chapter. Depending on what the investigator is interested in, it may be

of merit to restrict the calendar time period under study for the historic risk.

9.4.2 Multiple outcome events

We can alter the above to include competing risks. Prevalent patients will only enter the study if they are both cancer free and alive, therefore we need to consider the probability of patients surviving both these events in order to be in the study. Let patients start in state 0 and then transition to a state 1 to j , such that their state at time t is $E_t \in \{1, \dots, j\}$.

The Cox model is used to fit the cause-specific hazard of transitioning from state 0 to j as $\exp(\hat{\beta}_{0j}w)$. This is done separately for each event, j . The cumulative baseline hazard for event j at time t can be estimated using the Breslow estimator $\hat{\Lambda}_{0j;0}(t) = \sum_{s \leq t} \frac{\Delta N_{0j}(s)}{\sum_{i=1}^n R_i(s) \exp(\hat{\beta}_{0j}w_i)}$, where $N_{0j}(t)$ is the total number of transitions out of state 0 to state j in the interval $[0, t]$.

Each patient's cumulative hazard is the estimate $\hat{\Lambda}_{0j}(t|W = w_i) = \hat{\Lambda}_{0j;0}(t) \exp(\hat{\beta}_{0j}w_i)$. The estimated cumulative all-cause hazard is the summation of the above cumulative baseline hazard, $\hat{\Lambda}_0(t|W = w_i) = \sum_j \hat{\Lambda}_{0j}(t|W = w_i)$. The estimated survival probability is the survival from all events

$$\hat{S}(t|W = w_i) = P(T > t|W = w_i) = \prod_{u \leq t} (1 - \Delta \hat{\Lambda}_0(u|W = w_i)) \quad (9.6)$$

The predicted conditional cumulative incidence functions for different event types are

$$\hat{\Pr}(T \leq t, E_T = j|W = w_i) = \sum_{u \leq t} \hat{S}(u_-|W = w_i) \cdot \Delta \hat{\Lambda}_{0j}(u|W = w_i) \quad (9.7)$$

where $\hat{S}(u_-|W = w_i)$ is the survival probability at the previous timepoint. Note how each patient has their own individual conditional cumulative incidence, and therefore we can take the 'baseline' patient (the patient with $W = 0$) to represent

the baseline distribution, the **recent cumulative incidence** under competing events.

From here the steps are similar to the one state case. We define $Q = P(T > L) \approx \frac{n}{m}$, where we observe n out of m events due to left truncation. We estimate $\hat{m} = \sum_{i=1}^n \frac{1}{\hat{P}(T > \ell_i | W = w_i)}$, and define the left truncation distribution as

$$\hat{G}_L(\ell) = \frac{\hat{Q}}{n} \sum_{i=1}^n \frac{\mathbb{I}(\ell_i \leq \ell)}{\hat{P}_r(T > \ell_i | W = w_i)} \quad (9.8)$$

Finally, we have what Stegherr et al. term the marginal cumulative incidence function, for us the **historic cumulative incidence** in a competing risks setting given left truncation:

$$\hat{P}_r(T \leq t, E_T = j) = \frac{\hat{Q}}{n} \sum_{i=1}^n \frac{\hat{P}_r(T \leq t, E_T = j | W = w_i)}{\hat{P}_r(T > \ell_i | W = w_i)} \quad (9.9)$$

This estimates the historic risk, comparable to Figure 6.4, now weighted for changes over calendar time of SSc diagnosis.

As noted previously, when we wish to estimate the historic risk as a direct effect we still require the weighting to take into account the possibility of death prior to entry to the study as if it were a competing event. So there is a need to weight the cumulative incidence of each patient for their survival of both events. Therefore, we estimate the marginal cumulative incidence for our study as Equation (9.6), but use the survival from the above multiple event estimator. Let $T = \min(Y, D)$, where Y is the time until event of interest and D is the time until the competing event. The direct historic risk is therefore:

$$\hat{P}(Y \leq t) = \left(\sum_{i=1}^n \frac{1}{\hat{P}(T > \ell_i | W = w_i)} \right)^{-1} \sum_{i=1}^n \frac{\hat{P}(Y \leq t | W = w_i)}{\hat{P}(T > \ell_i | W = w_i)} \quad (9.10)$$

Although we now weight for patients being unobserved due to calendar time, this continues to be a weighted naive Kaplan-Meier, and therefore does not account for informative censoring. There is a need to adjust these methods, which can

be done on a continuous time scale with Cox proportional hazards models, or we can use our alternative method from Chapters 7 and 8 provided by Young et al. (2020) with discrete time. We will use the latter in Section 9.6.

9.5 Application of temporal trend allowances to SSc dataset

The appendix contains simulations to demonstrate the use of the above estimators for better understanding (Appendix A.2). However, this is not novel material and is therefore separated from the main thesis.

We apply the above methods to our data. We will be using continuous calendar time for our weightings. We apply the weightings to both the SSc and non-SSc patients. The left of Table 9.4 shows the hazard coefficient, hazard ratio and 95% confidence intervals, and the right shows the proportionality test for the proportional hazards assumption. This table is similar to Table 9.1, however we now have 1 year of calendar time as opposed to 10 years. However the above findings still hold, where we see a decreased risk of both cancer and death prior to cancer in the more historic cases. As shown on the right, none of the four hazard ratios failed the proportional hazards test.

	coef	HR = exp(coef)	95% CI	p-value
Cancer, SSc	-0.0258	0.975	[0.948, 1.002]	0.79
Cancer, Non-SSc	-0.0303	0.970	[0.958, 0.983]	0.59
Death without cancer, SSc	-0.0146	0.985	[0.963, 1.008]	0.53
Death without cancer, Non-SSc	-0.0241	0.976	[0.961, 0.992]	0.29

Table 9.4: Cox regression models with calendar time of diagnosis as covariate (left), along with Cox proportional hazards test (right).

We have continued to remove patients who did not have smoking information for consistency with the previous two chapters. Also we have removed one SSc patient whose truncation time L was greater than the length of the last event time

of other SSc patients. We have also removed this patient's non-SSc comparators. Also, for simplicity of computing, we are working on discrete time as opposed to continuous scale, similar to that of the g-formula. We work to the nearest month with the timestep being each month, as opposed to the incremental increases in hazard we have whenever an event happens. This holds for time until event, T , time from diagnosis to entry to study, L , and calendar time from SSc diagnosis, W . We are working with discrete time and have rounded each event to the nearest month.

9.5.1 Direct effect

Figure 9.1 shows the adaption to demonstrate the 'recent' risk, where we wish to consider one event and treat the other as censoring (direct effect). The solid lines are the left truncated adjusted Kaplan-Meier from previous chapters under the assumption there is independent truncation. The dashed lines are the Breslow baseline adjusted estimations (i.e. the recent cumulative incidence) using Equation (9.3).

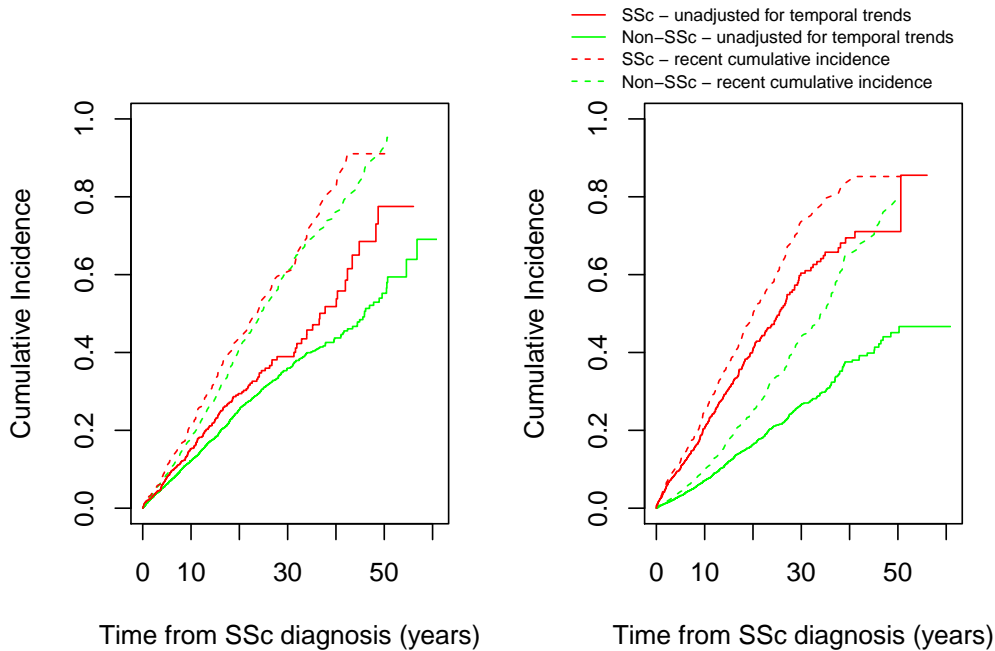


Figure 9.1: Cumulative incidence using the independent Kaplan-Meier (solid lines) and the temporally adjusted recent cumulative incidence direct effect (dashed line). Cancer (left) and death (right).

We have cancer on the left and death on the right of Figure 9.1. Both demonstrate a significantly increased risk in the recent cumulative incidence, which is as expected due to the observed hazard ratios above, where those further back in calendar time have a lower hazard. While these new curves look very different to the independent naive Kaplan-Meier, the new curves are close to that of the g-formula results where the combined fitted GLM was applied to the incident SSc pseudo-population (Figure 8.1). This again may highlight that it is the differences in measured covariate distributions over calendar time rather than an increase of hazard conditional on covariates over calendar time that is leading to the appearance of temporal trends, as the g-formula accounted for covariates by using the incident cohort as the pseudo-population. The difference in the SSc and non-SSc curves for cancer incidence is now reduced, implying that there is little to no direct effect on cancer in those with SSc compared to those without. The gap between the SSc and non-SSc curves for death is also narrower, but still

remains substantially different. However, these curves treat the competing event as independent censoring, when the competing event is not independent from the event of interest.

Figure 9.2, shows estimates of the historic risk, i.e. the marginal risk over the calendar time period of interest, using Equation (9.10). While not as high as the recent risk, the historic risk is again greater than that obtained from the unadjusted Kaplan-Meier at longer times from SSc diagnosis. The gap between the unadjusted and the adjusted for death in SSc patients is smallest, which is expected due to the hazard ratio for calendar time being closer to 1.

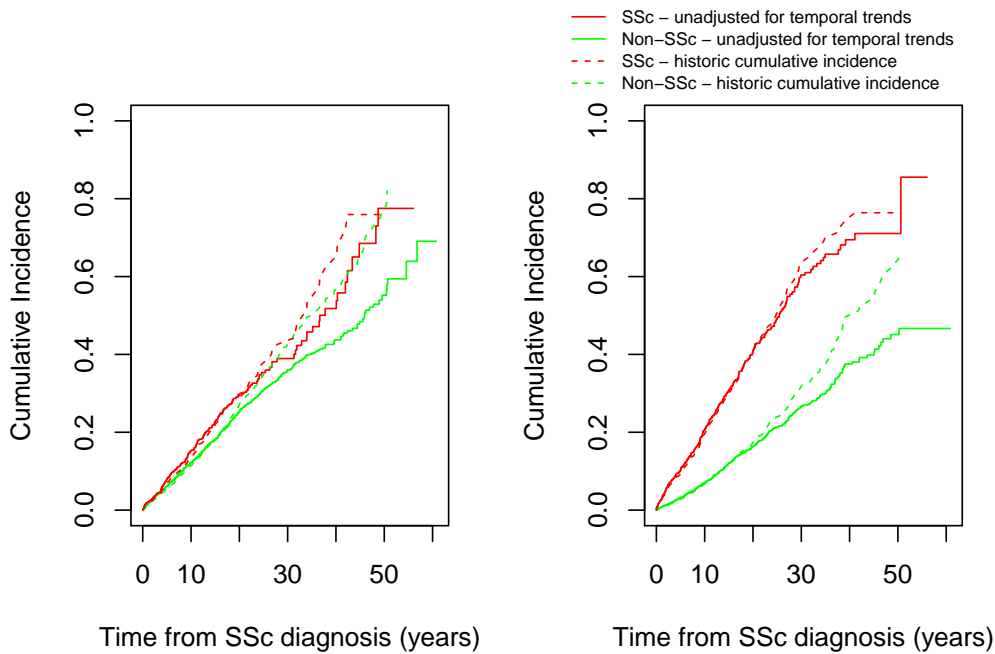


Figure 9.2: Cumulative incidence using the independent Kaplan-Meier (solid lines) and the temporally adjusted historic cumulative incidence direct effect (dashed line). Cancer (left) and death (right).

9.5.2 Total effect

We now assume that the competing event is not treated as censoring (total effect), and estimate the cause-specific cumulative incidences. Figure 9.3 shows

the estimations of cause-specific cumulative incidence for the recent effect. The solid red and green lines are the risk-set adjusted Aalen-Johansen curves under the assumption that we have no temporal trends. The dashed green lines are the estimation of cumulative risk if all patients had been diagnosed in 18-09-2017 (i.e. the recent risk) which is estimated using the conditional cumulative incidence, Equation (9.7).

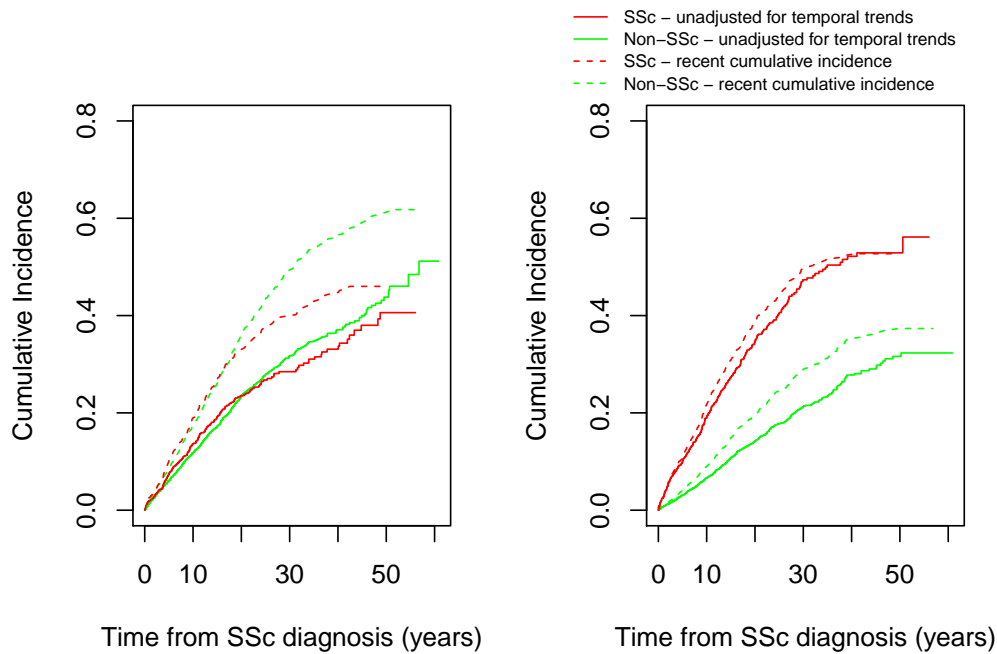


Figure 9.3: Comparing the non-adjusted cause-specific cumulative incidence with the recent cumulative incidence curves. The non-adjusted cause-specific cumulative incidence are the solid lines, and the temporal adjusted recent cumulative incidence curves are the dashed lines. Cancer (left) and death without cancer (right).

Again, we see that the recent cases show an increased risk of both death and cancer compared to the unadjusted cause-specific cumulative incidence, with a much greater difference between the SSc and non-SSc at large survival times. In the cancer plot (left) we see that the non-SSc patients have a large difference between the adjusted and unadjusted curve. For death (right) we see a much smaller difference between the adjusted and unadjusted curves. This is due to the

smaller hazard ratios for risk of death compared to the risk of cancer, but may also be due to the interaction of cancer and death events in the competing risk setting. If we expected many more cancer outcomes in more recent patients, we may expect to observe less death events. However as recent patients have an increased hazard for both death and cancer, an increased cumulative incidence is observed in both events. All cumulative incidences converge to a probability between 0 and 1 as $t \rightarrow \infty$ (which differs to the direct effect, where the probability will always tend to 1), and we observe that approximately 45% of SSc patients will have cancer prior to death and 55% will die prior to a cancer diagnosis. For non-SSc patients we estimate more observable cancer events, with approximately 60% having received a cancer diagnosis prior to death. This is a slightly larger number than we might expect, as several reports in the literature place the current risk as approximately 50% of people (Ahmad et al. (2015), Smittenaar et al. (2016)). However, we do not have a full UK general population study but a matched study based on SSc matching, and risk is measured from SSc onset, and this could be the cause of the discrepancy. Also, the confidence intervals at longer times from SSc diagnosis are possibly quite large.

Again, these curves are similar to the g-formula curves in Figure 8.2, possibly implying that there is more of an adjustment for covariates over calendar time rather than a change in hazard over calendar time even after adjusting for measured covariates.

Figure 9.4 demonstrates the difference for the unadjusted cause-specific incidence (solid line) and the historic cumulative incidence accounting for temporal trends (dashed line), estimated from Equation (9.9). Fifteen years after SSc diagnosis, there is clearly an increase in risk of both cancer and death prior to cancer for both SSc and non-SSc patients. However, there is a small decreased risk in the early years after SSc diagnosis/match for the adjusted curve compared to the unadjusted Aalen-Johansen curve, but this equalizes 15 years after SSc diagnosis. This demonstrates that patients diagnosed earlier in calendar time have a slightly decreased risk of both cancer and death without cancer compared to the average

over the time period studied, so their risk is less than the unadjusted risk. In death for non-SSc patients, we see that the historic curves are similar to the recent curves. Even though death risk decreases for both SSc and non-SSc patients with increasing W , i.e. further back in calendar time, it is less than for cancer with calendar time. Cancer risk continues to be higher for non-SSc patients at later times from SSc diagnosis, due to the increased mortality of SSc patients.

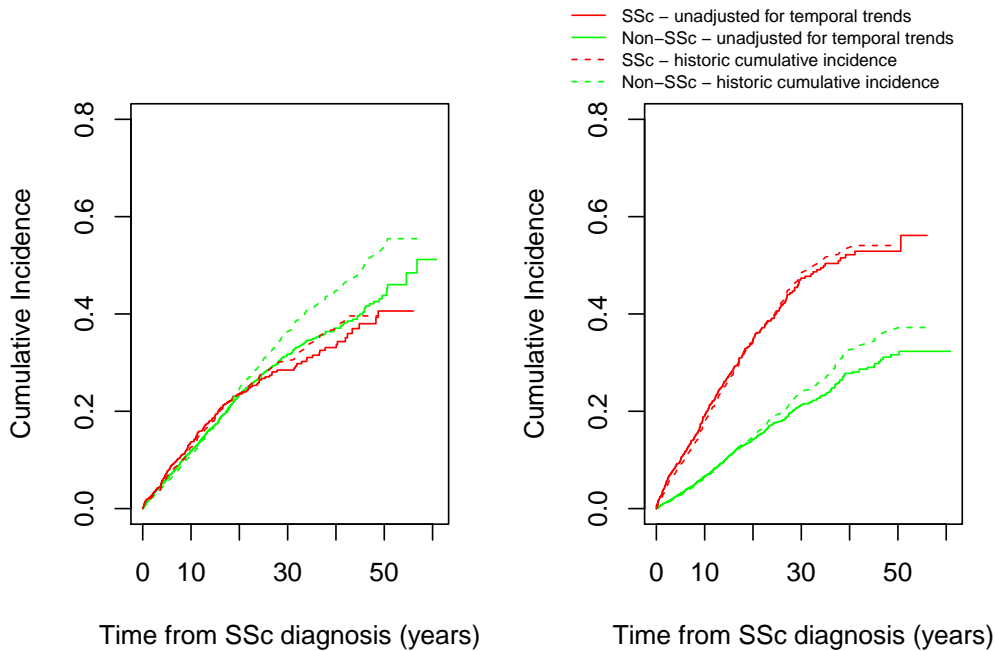


Figure 9.4: Comparing the non-adjusted cause-specific cumulative incidence with the historic cumulative incidence. The non-adjusted cause-specific cumulative incidence are the solid lines, and the temporal adjusted historic cumulative incidence curves are the dashed lines. Cancer (left) and death without cancer (right).

While all these curves may indicate that there are substantial changes when we consider calendar time, there is more uncertainty in later years due to the limited number of patients in the risk set, and these patients who have longer truncation times are therefore representing a large number of unobservable patients who did not survive to study start. As we shall see from the confidence intervals, there is a concern about the stability of this method, as patients with large truncation

times may have more effect in the Cox model, and therefore may have a large impact on the curves.

9.5.3 Summary table

Tables 9.5 and 9.6 provide a summary of the new temporally adjusted risk ratios alongside bootstrapped confidence intervals for the direct and total effect respectively. As a reminder from Chapters 5 and 6, the non-temporally adjusted left truncated Kaplan-Meier 95% confidence intervals are estimated from the Greenwood formula for the $1 - KM$ cumulative incidence curve, and the non-temporally adjusted cause-specific (Aalen-Johansen) 95% confidence intervals are estimated using jackknifed confidence intervals as specified in Section 6.3.2 which is a product of *cuminc* in the R package *cmprsk*. Therefore the non-temporally adjusted risk ratio is $RR(t) = F_{SSc}(t)/F_{nSSc}(t)$. The confidence intervals are calculated as (Stegherr et al., 2021)

$$CI(t) (95\%) = RR(t) \exp \left(\pm 1.96 \sqrt{\left(\frac{se(F_{SSc}(t))}{F_{SSc}(t)}\right)^2 + \left(\frac{se(F_{nSSc}(t))}{F_{nSSc}(t)}\right)^2} \right) \quad (9.11)$$

For our new estimators in this chapter, we use bootstrapping to estimate the confidence intervals. We sample our SSc dataset with replacement to be the same size as the original dataset for SSc, and create an SSc cohort from these as well as a complementary non-SSc of the SSc set's matches so that we continue to have a matched study. We then record the point estimates at time points $t=10, 20$ and 30 and the risk ratio. We do this for 1,000 samples. We note that not every sample will contain the patient whose diagnosis was the above baseline date (18/09/2017), and we shall take the patient who was diagnosed at the latest calendar time as the recent patient risk. The confidence interval estimate for the cumulative incidence (both recent and historic, direct and total effect) is as follows

$$\left([\hat{F}_k(t)]^{\exp(z_{(1-\alpha)/2} \sqrt{\text{Var}(\log(-\log(\bar{F}_k(t))))}), [\hat{F}_k(t)]^{\exp(-z_{(1-\alpha)/2} \sqrt{\text{Var}(\log(-\log(\bar{F}_k(t))))})} \right) \quad (9.12)$$

where $\hat{F}_k(t)$ is the recent/historic cumulative incidence estimates of the original data, and $\log(\bar{F}_k(t)) = \sum_{b=1}^B \log(\bar{F}_{k,b}(t))/B$, and where $\bar{F}_{k,b}(t)$ is the point estimate of bootstrap sample b .

The confidence intervals for the risk ratio are as follows:

$$\left([\hat{R}R_k(t)]^{\exp(-z_{(1-\alpha)/2} \sqrt{\text{Var}(\log(\bar{R}R_{b,k}(t))))}, [\hat{R}R_k(t)]^{\exp(z_{(1-\alpha)/2} \sqrt{\text{Var}(\log(\bar{R}R_{b,k}(t))))} \right) \quad (9.13)$$

where $\hat{R}R_k(t)$ is the cumulative incidence estimate of the original data, and $\log(\bar{R}R_k(t)) = \sum_{b=1}^B \log(\bar{R}R_{b,k}(t))/B$, and where $\bar{R}R_{b,k}(t)$ is the point estimate of bootstrap sample b ¹.

The confidence intervals are wide here, particularly for the recent risk. The use of weightings may be widening the confidence intervals, as there is now the extra influence of the weightings. Patients with large truncation times, and whether they are included or not in the bootstrapped sample, greatly influence the Cox model. It may be that outliers at large truncation times will cause more variability in the simulation. It could be that this is due to the greater influence of a few patients with large truncation times. With more time, it may have been interesting to restrict this analysis to those diagnosed after a later time period, such as from 1980.

Table 9.5 is for the direct effect. In the recent risk, which is what we wish to use to infer current risk and is best from a prognosis perspective, we see the risk ratio becomes closer to 1 for all timepoints and the wide confidence intervals

¹Equation (9.11) could also have been used here instead and would produce similar CI to Equation (9.13).

Time from SSc diagnosis (years)	Non-adjusted Kaplan-Meier	Recent adjusted direct effect	Historic adjusted direct effect
10	SSc 0.15 [0.13, 0.18] Non-SSc 0.12 [0.11, 0.13] Ratio 1.25 [1.03, 1.51]	SSc 0.21 [0.14, 0.30] Non-SSc 0.18 [0.15, 0.22] Ratio 1.18 [0.79, 1.77]	SSc 0.14 [0.11, 0.17] Non-SSc 0.11 [0.10, 0.12] Ratio 1.23 [1.00, 1.51]
20	SSc 0.29 [0.25, 0.33] Non-SSc 0.25 [0.24, 0.27] Ratio 1.16 [1.00, 1.36]	SSc 0.44 [0.26, 0.60] Non-SSc 0.41 [0.32, 0.49] Ratio 1.07 [0.69, 1.65]	SSc 0.30 [0.25, 0.34] Non-SSc 0.27 [0.25, 0.29] Ratio 1.11 [0.93, 1.32]
30	SSc 0.39 [0.33, 0.44] Non-SSc 0.36 [0.34, 0.38] Ratio 1.08 [0.92, 1.27]	SSc 0.61 [0.29, 0.82] Non-SSc 0.61 [0.46, 0.73] Ratio 0.99 [0.63, 1.57]	SSc 0.44 [0.34, 0.52] Non-SSc 0.43 [0.38, 0.48] Ratio 1.02 [0.81, 1.29]

Table 9.5: Direct effect on cancer for incident and combined datasets at time points 10, 20 and 30 years, also the ratio between them (SSc over non-SSc). The 95% confidence intervals are provided in brackets.

Time from SSc diagnosis (years)	Non-adjusted Aalen-Johansen	Recent adjusted total effect	Historic adjusted total effect
10	SSc 0.14 [0.11, 0.16] Non-SSc 0.12 [0.11, 0.13] Ratio 1.15 [0.95, 1.39]	SSc 0.19 [0.16, 0.22] Non-SSc 0.17 [0.14, 0.21] Ratio 1.09 [0.72, 1.66]	SSc 0.13 [0.10, 0.15] Non-SSc 0.11 [0.10, 0.12] Ratio 1.14 [0.92, 1.42]
20	SSc 0.24 [0.21, 0.27] Non-SSc 0.23 [0.22, 0.25] Ratio 1.01 [0.87, 1.17]	SSc 0.33 [0.26, 0.40] Non-SSc 0.36 [0.29, 0.43] Ratio 0.92 [0.59, 1.42]	SSc 0.24 [0.20, 0.27] Non-SSc 0.24 [0.23, 0.26] Ratio 0.96 [0.82, 1.13]
30	SSc 0.28 [0.25, 0.32] Non-SSc 0.32 [0.30, 0.34] Ratio 0.90 [0.78, 1.04]	SSc 0.40 [0.29, 0.51] Non-SSc 0.49 [0.39, 0.59] Ratio 0.81 [0.51, 1.29]	SSc 0.30 [0.26, 0.35] Non-SSc 0.36 [0.33, 0.40] Ratio 0.84 [0.70, 1.00]

Table 9.6: Total effect on cancer for incident and combined datasets at time points 10, 20 and 30 years, also the ratio between them (SSc over non-SSc). The 95% confidence intervals are provided in brackets.

mean that at times further from SSc diagnosis we no longer have a significant difference between SSc and non-SSc patient risks. The historic risk is similar to the non-temporally adjusted risk, but also has large confidence intervals. From the historic risk, there is the implication that in the past there might have been a small increase in the direct effect of SSc on cancer, however this is no longer the case at larger times from diagnosis, judging from the recent risk ratios. This difference might imply that historically SSc patients were more at risk of cancer than their non-SSc counterparts but this is no longer the case. However, it could also be due to an under-reporting of cancer in non-SSc patients further back in calendar time, or under-reporting of certain SSc patients, such as males.

In Table 9.6 we have the cause-specific risk ratios and total effect. The confidence intervals are particularly wide here, so despite the risk ratios being large at times from diagnosis greater than 15 years, there are no statistically significant risk ratios. The recent risk appears to show a decreasing risk ratio with survival time, due to death being a competing event. This is similar for the historic risk but the historic risk ratios are slightly higher than the recent risks.

9.6 Temporal trends in the g-formula

In the prior section, we acknowledge that the direct effect method does not account for informative censoring. While we could alter the above methods to account for this as well as for temporal trends, as continuous methods can also be weighted to account for informative censoring, we wish to continue with the g-formula from the previous chapter. The reason for this is that the g-formula method also accounts for differing covariates between the exposed and non-exposed, which is one of its key advantages, even if our dataset is matched on two of the three covariates used.

We give an overview of how the g-formula could be altered to include temporal trends. We can adjust for calendar time by treating it as a covariate. This is done in other g-formula studies which account for differences over long study times, however they do not feature prevalent cohorts (Cole et al. (2013), Murray et al.

(2018)).

We firstly review the three conditions for the g-formula. As stated in Section 8.1, Keil & Edwards (2018) discussed that exchangeability² would need to hold where the “underlying health status should not depend on entry time, conditional on past values of exposure and confounders”. In our study, we have independence between entry time and event time given calendar time, therefore by including calendar time as a covariate we will remove dependence between truncation time and event time. How the g-formula could be used if dependent truncation is present would be more complicated, and a possible avenue for further work. Robins, in their original paper on the g-formula (Robins, 1986), discusses prevalent cohorts under the subsection ‘Selection bias caused by cohort definition’, stating that while they were currently investigating this selection bias, in the meantime “the only safe option is to match on exposure and work history until time of selection”.

The other two criteria are positivity and consistency. Positivity holds, as the probability of being censored (both competing event and loss to follow-up) is greater than zero under all time-points under study. Consistency by definition is not affected by left truncation, it only concerns the time under study, $k + 1$ (see Section 7.3.2).

As covered in Chapter 8, we need to weight the g-formula to represent unobserved prevalent patients. It is not immediately obvious how to account for prevalent cohorts when the baseline covariate distribution changes over calendar time, which we suspect is occurring in our study. If we account for prevalent cohorts with weightings, we would produce the marginal covariate distribution over the calendar time period of SSc diagnoses (1957-2018), not the covariate distribution most applicable to recent time periods. How best to account for a change in covariate distribution is left for future work. A simple approach to investigate the presence of changing covariates would be to perform a similar analysis as in Section 8.7.2, where we estimated the adjusted covariate distributions by restricting the recruitment of patients to those diagnosed after a certain calendar

²See Section 7.3.2

time, τ . A proportion increasing/decreasing due to more historic (and therefore biased) cases may be an indication of changing covariate distribution over time, as this implies that the more historic cases there are, the more different from the more recent distribution the estimate of baseline covariate distribution becomes. Also, applying the Cox model as was done above, both without covariates and with covariates, would test whether the covariate distribution impacts the temporal trend.

Due to our large dataset, we believe the best way to estimate the recent risk using the g-formula estimator will be to fit the GLM for all patients to get the most precise estimates of the coefficients, but then only use the most recent set of patients diagnosed with SSc as our pseudo-population, therefore representing the most recent cohort. This includes both incident and prevalent patients diagnosed after a certain calendar time. We can then adjust for temporal trends by treating calendar time as a covariate.

9.6.1 Including calendar time as a covariate

We choose to add calendar time to the model. A GAM was tested compared to a GLM, where k , *Age* and *Calendartime* were smoothed and *A*, *Sex* and *Smoking* were not, however the AIC was not minimised for- To allow for the possibility that calendar time effects may differ between the SSc and non-SSc groups, we include an interaction between calendar time and SSc status. As with previous chapters we test for interaction terms with other covariates as well, where they are included if they minimise the AIC. We show the results for the model fitting, where we have the following models:

$$\begin{aligned}
 g\{Pr[Y_{k+1} = 1 | \bar{C}_{k+1} = \bar{D}_{k+1} = \bar{Y}_k = 0, A = a, Z_0]\} \dots \\
 = \theta_0 + \theta_1 A + \theta_2 k + \theta_3 k^2 + \theta_4 Male + \theta_5 Age + \theta_6 Age^2 + \dots \\
 \theta_7 Smoker + \theta_8 Exsmoker + \theta_9 W + \theta_{10} A \times W + \dots \\
 \theta_{11} k \times Age + \theta_{12} Male \times Age
 \end{aligned}$$

$$\begin{aligned}
&g\{Pr[D_{k+1} = 1 | \bar{C}_{k+1} = \bar{D}_k = \bar{Y}_k = 0, A = a, Z_0]\}... \\
&= \phi_0 + \phi_1 A + \phi_2 k + \phi_3 k^2 + \phi_4 Male + \phi_5 Age + ... \\
&\phi_6 Age^2 + \phi_7 Smoker + \phi_8 Exsmoker + \phi_9 W + ... \\
&\phi_{10} A \times W + \phi_{11} k \times Age + \phi_{12} A \times Age
\end{aligned}$$

where W is calendar time of diagnosis going backwards from the latest date of diagnosis.

Table 9.7 shows the summary of the fit of the GLM. Again, with time from SSc diagnosis, k , was not significant for terms higher than quadratic for the hazard of cancer and death. While an interaction of sex and age at SSc diagnosis minimised the AIC for cancer hazard, it did not do the same for death, and similarly the interaction of SSc and age at diagnosis is included for death but not for cancer.

	Cancer	[95% Confidence interval]	Death	[95% Confidence interval]
Intercept	-12.257	[-13.887, -10.673]	-11.791	[-13.977, -9.776]
k	0.0125	[0.0076, 0.0177]	-2.568e-3	[-8.404e-3, 3.563e-3]
k^2	-4.06e-6	[-8.740e-6, 4.221e-7]	5.126e-6	[-2.707e-7, 1.030e-5]
SSc (A=1)	0.156	[-0.171, 0.478]	3.131	[2.097, 4.173]
Male	-0.955	[-1.746, -0.193]	0.219	[0.0356, 0.396]
Age	0.123	[0.076, 0.173]	-3.404e-3	[-0.0586, 0.0560]
Age ²	-5.617e-4	[-9.37e-4, -2.04e-4]	8.400e-4	[4.253e-4, 1.229e-3]
Smoker	0.195	[0.049, 0.338]	0.900	[0.741, 1.058]
Ex-Smoker	0.0252	[-0.113, 0.161]	0.206	[0.0483, 0.362]
W (Calendar time)	-0.0178	[-0.0299, -0.006]	0.0248	[0.010, 0.040]
$k \times Age$	-1.162e-4	[-1.852e-4, -5.001e-5]	1.519e-4	[7.644e-5, 2.236e-4]
$A \times Age$	-	-	-0.0220	[-0.0348, -0.0093]
$A \times W$	7.333e-3	[-0.0071, 0.0213]	-0.0238	[-0.0404, -0.007]
Male \times Age	0.0205	[0.0078, 0.0334]	-	-

Table 9.7: Coefficient values for logistic regression, estimated using both the prevalent and incident cohorts. The baseline for A is non-SSc, the baseline for sex is female and the baseline for smoking and ex-smoker is non-smoker. Age is the age at which SSc patients were diagnosed. The baseline for calendar time is $W=0$ in 18/09/2017, and each yearly increment is further back in calendar time.

For cancer, calendar time, W , is statistically significant at the 5% level, implying

that historic cases have better survival of cancer which agrees with the Cox model previously (Table 9.3). The interaction of calendar time and SSc type (WA) is not statistically significant at the 5% level, but the interaction term implies there is less of a difference in hazard over calendar time for SSc patients. For death without cancer, calendar time is significant at the 5% level, with hazard increasing with calendar time such that recent cases have better survival (again, in agreement with Table 9.3). This increase is much less for SSc patients due to the interaction term, AW , which is at a statistically significant level.

We produce g-formula estimates similar to Section 8.9, but now include calendar time in the modelling of the g-formula. We fit the GLM with incident and prevalent patients. To minimise the impact of changing covariates over time, when applying the fitted GLM to a pseudo-population, the pseudo-population will only consist of patients who were diagnosed after the year 2000, which reduces the number of SSc patients used in the pseudo-population to 60% of the original. Prevalent patients make up 20% of this group, and we therefore continue with the weighted g-formula for the probability of their inclusion to adjust for the unobserved patients due to left truncation, Equations (8.3) and (8.4).

The results for these g-formula estimates are shown in Figure 9.5 for the direct effect and Figure 9.6 for the total effect. We also include the semi-parametric ‘recent risk’ estimates (Breslow baseline estimates) from the prior section for comparison. Note that the Breslow estimator does not account for informative censoring, and is a comparison for the recent risk based on one calendar date as opposed to time since the year 2000.

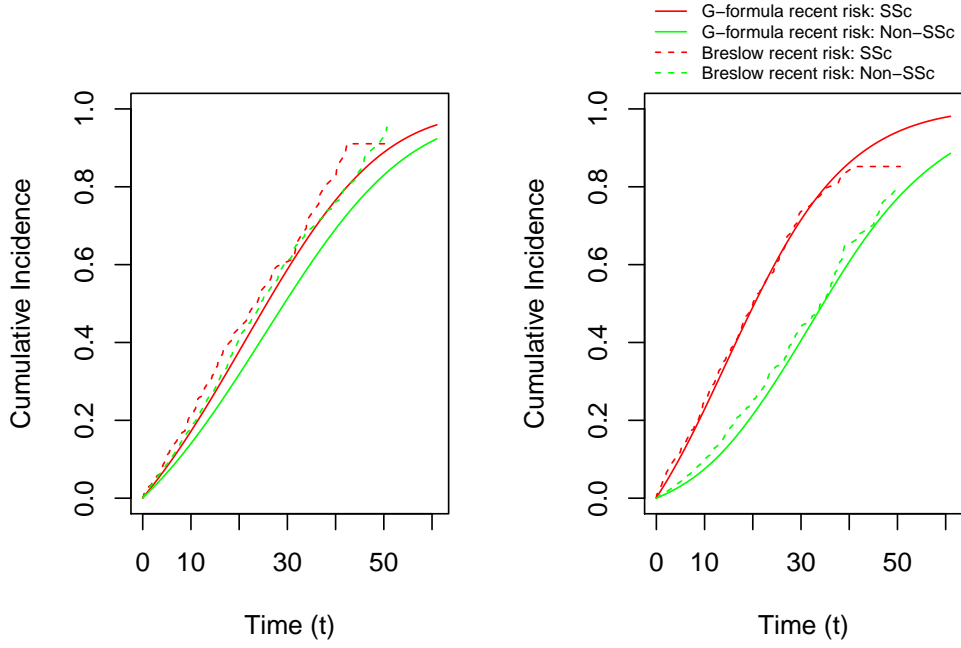


Figure 9.5: G-formula recent direct effect on cancer (left) and death (right), with the logistic regression model being fitted on either the incident and prevalent cohort with temporal trends (solid line, red and green) or the Breslow estimated recent risk with both the incident and prevalent cohort (dot-dashed line, red and green).

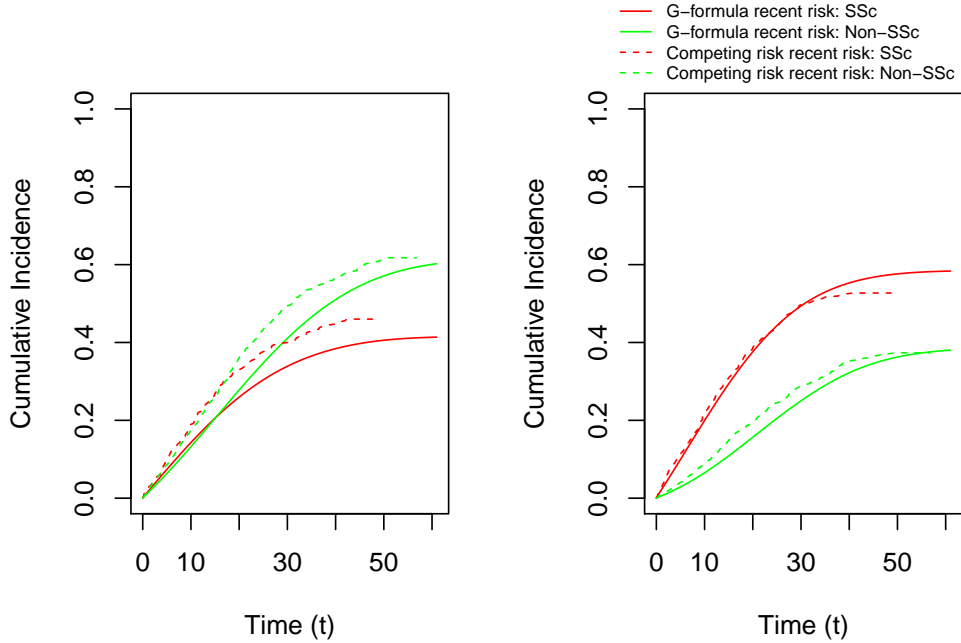


Figure 9.6: G-formula recent total effect on cancer (left) and death (right), total effect, with the logistic regression model being fitted on either the incident and prevalent cohort with temporal trends (solid line, red and green) or the competing risk estimated recent risk with both the incident and prevalent cohort (dot-dashed line, red and green).

Reviewing Figure 9.5, we observe that the recent risk naive Kaplan-Meier curve is very similar to the g-formula curve. While this is possibly expected, we might expect the g-formula estimator to be higher as it is adjusting for informative censoring which the Breslow estimator is not. Although, we expect there to be differences as the naive Kaplan-Meier adjusted for baseline risk is assuming every patient entered at time $w = 0$ whereas the g-formula pseudo-population has diagnosis dates distributed between 2000-2018 ($w = 0$ to $w = 18$). Comparisons between this graph and Figure 8.1 show the largest change in risk is for non-SSc patients. We observe that cancer risk increases slightly for both SSc and non-SSc patients with the temporal trends adjustment, and there is a change in death without cancer where risk has decreased slightly for SSc patients and more noticeably for non-SSc patients. For death, the difference between the SSc

and non-SSc curves is now larger, implying the mortality between SSc and non-SSc-patients is now greater.

Figure 9.6 demonstrates a difference in the two presented methods. We see that there is a small increased risk of cancer in the g-formula compared to the temporally adjusted Aalen-Johansen method. Again, this could be due to differences between the methods due to which time-point defines the ‘recent’ risk. We note the lack of time-varying covariates in this study, despite this being one of the main advantages of the g-formula as a method. This is an area that could be researched further for a study that has time-varying covariates. In such an investigation, firstly, exchangeability would need to hold. Another consideration is that the weightings we have used for the baseline proportion estimation would not work when we need to simulate patients with changing covariates over time. We direct anyone with an interest in this to a very recent article by Vakulenko-Lagun et al. (2021), who use inverse probability weightings at each time step (allowing for time-varying covariates) to adjust for dependent truncation induced by covariates to model survival distribution nonparametrically. This may be a good place to start a study of this aspect.

9.6.2 Summary table

The following tables, Table 9.8 (direct effect) and Table 9.9 (total effect), summarise the risk ratios for the above graphs, as well as the Breslow estimate (‘recent risk’) and the original unweighted and unadjusted NPMLEs (naive Kaplan-Meier and Aalen-Johansen cause-specific cumulative incidence). The 95% confidence intervals for the g-formula are again found via bootstrapping.

Table 9.8 shows the direct effect. There are close similarities between the two temporal adjusting methods of this chapter. The risk is closer to 1 for both the recent adjusted Kaplan-Meier and the recent adjusted g-formula, compared to the original, unadjusted Kaplan-Meier. This gives the impression that SSc has little to no causal impact on cancer. However, all the risk ratios are higher than 1. We

Time from SSc diagnosis (years)	Non-adjusted Kaplan-Meier	Recent adjusted direct effect	Weighted direct g-formula
10	SSc 0.15 [0.13, 0.18] Non-SSc 0.12 [0.11, 0.13] Ratio 1.25 [1.03, 1.51]	SSc 0.21 [0.14, 0.30] Non-SSc 0.18 [0.15, 0.22] Ratio 1.18 [0.79, 1.77]	SSc 0.17 [0.13, 0.21] Non-SSc 0.14 [0.12, 0.16] Ratio 1.22 [0.92, 1.63]
20	SSc 0.29 [0.25, 0.33] Non-SSc 0.25 [0.24, 0.27] Ratio 1.16 [1.00, 1.36]	SSc 0.44 [0.26, 0.60] Non-SSc 0.41 [0.32, 0.49] Ratio 1.07 [0.69, 1.65]	SSc 0.38 [0.29, 0.46] Non-SSc 0.32 [0.26, 0.37] Ratio 1.19 [0.90, 1.57]
30	SSc 0.39 [0.33, 0.44] Non-SSc 0.36 [0.34, 0.38] Ratio 1.08 [0.92, 1.27]	SSc 0.61 [0.29, 0.82] Non-SSc 0.61 [0.46, 0.73] Ratio 0.99 [0.63, 1.57]	SSc 0.59 [0.45, 0.70] Non-SSc 0.51 [0.41, 0.60] Ratio 1.15 [0.89, 1.48]

Table 9.8: Cumulative incidence of cancer for incident and combined datasets at time points 10-, 20- and 30-years, also the ratio between them (SSc over non-SSc). The 95% confidence intervals are provided in brackets.

Time from SSc diagnosis (years)	Non-adjusted Aalen-Johansen	Recent adjusted total effect	Weighted total g-formula
10	SSc 0.14 [0.11, 0.16] Non-SSc 0.12 [0.11, 0.13] Ratio 1.15 [0.95, 1.39]	SSc 0.19 [0.16, 0.22] Non-SSc 0.17 [0.14, 0.21] Ratio 1.09 [0.72, 1.66]	SSc 0.14 [0.11, 0.18] Non-SSc 0.131 [0.11, 0.15] Ratio 1.08 [0.81, 1.45]
20	SSc 0.24 [0.21, 0.27] Non-SSc 0.23 [0.22, 0.25] Ratio 1.01 [0.87, 1.17]	SSc 0.33 [0.26, 0.40] Non-SSc 0.36 [0.29, 0.43] Ratio 0.92 [0.59, 1.42]	SSc 0.26 [0.20, 0.32] Non-SSc 0.28 [0.23, 0.32] Ratio 0.94 [0.69, 1.27]
30	SSc 0.28 [0.25, 0.32] Non-SSc 0.32 [0.30, 0.34] Ratio 0.90 [0.78, 1.04]	SSc 0.40 [0.29, 0.51] Non-SSc 0.49 [0.39, 0.59] Ratio 0.81 [0.51, 1.29]	SSc 0.34 [0.25, 0.43] Non-SSc 0.41 [0.33, 0.49] Ratio 0.83 [0.60, 1.14]

Table 9.9: Cumulative incidence of cancer for incident and combined datasets at time points 10-, 20- and 30-years, also the ratio between them (SSc over non-SSc). The 95% confidence intervals are provided in brackets.

have not investigated subsets of cancer and it may be that a particular cancer or cancers have a higher risk.

Table 9.9 shows the total effect. In the first few years after SSc diagnosis, there does not appear to be a difference in risk ratio for SSc and non-SSc. However, now that we have adjusted for differing covariate values and temporal trends, the total risk of cancer for SSc patients is lower than that of non-SSc patients as death is such a significant competing risk. This implies that SSc patients are less likely to be diagnosed with cancer than their non-SSc counterparts. However, the confidence intervals continue to be large. From the plot, there is a clear separation between SSc and non-SSc for cancer 20 years after diagnosis.

The g-formula has smaller confidence intervals than the recent Breslow estimator, possibly suggesting this method might be more robust against more influential highly truncated patients, however we have observed that the g-formula has smaller confidence intervals in other sections of this work too.

9.7 Conclusions

There is a small subsection of research which focuses on adjusting for dependent truncation. While we do not have dependent truncation, there is an overlap in methodology with conditionally independent truncation, where we utilised research in this area to both investigate temporal trends in a prevalent cohort, and adjust for both a ‘recent risk’ and a ‘historic risk’.

Chapter 8 highlighted that the inclusion of our prevalent cohort in the modelling of cumulative incidence curves gave an underestimation of the risk at later survival times compared to the risk we believe to be the most ‘recent’ (i.e. comparable to 2017-2018). We suspect this is due to both differing hazards conditional on measured covariates and differing covariates. Once covariates have been taken into account we noted an increased hazard of mortality and decreased hazard of cancer in more recent cases, especially for non-SSc patients. For SSc patients this was not at a significant level, however this may be due to an insufficient number

of SSc patients in our study.

From the cumulative incidence curves for the direct effect, we see the risk ratios become closer to 1 after temporal adjustment. The confidence intervals are larger for the methods in this chapter, especially our Breslow estimator for recent risk. Looking at the recent risk when death is treated as censoring in Table 9.5, the results are no longer significant, implying that there is no difference in risk of cancer between SSc and non-SSc patients, however we note the much wider confidence intervals in the risk ratios. The historic risk does still have significant results at smaller times, implying that it may be that there was once an increased risk of cancer in SSc patients compared to non-SSc patients, but this is no longer the case. Looking at the cause-specific risk in Table 9.6, our conclusion is the same as in previous chapters, which is that when death is treated as a competing event then the risk of cancer is no greater in SSc patients than non-SSc patients.

The final method we present is a culmination of what we wish to achieve: a g -formula estimator which can account for left truncation, dependent censoring and confounders. We again found no significant direct impact of SSc on cancer, and the results for the total effect of SSc on cancer highlights that the probability of observing cancer is less in SSc patients due to their increased risk of death.

We note some limitations of the above methods. The first is that we can only observe patients who have survived until their entry time, and without observation of the truncated region we do not know anything about the actual distribution of T and L , i.e. our methods work when they are based on estimating information in the observable region. The Breslow recent risk estimator has produced wide confidence intervals, suggesting this method may not be accurate. We consider that it is still preferable to adjust for temporal trends if we feel they are present rather than neglect them in favour of seeking smaller confidence intervals. We note that we have not taken the IPCW further in this study (it could be adjusted to allow for temporal trends), and neither have we taken continuous time methods further (which could potentially be altered to account for informative censoring).

Chapter 10

Discussion

10.1 Summary of work undertaken

In this thesis we have investigated whether there is an increased risk of cancer (the outcome of interest) in patients diagnosed with SSc (the exposure of interest) and to do so we have utilised a large, matched dataset. Our motivation was to assess the risk of cancer in SSc, which is an underdeveloped area and where current studies suffer from small sample sizes. An increased risk may highlight a need for a different allocation of healthcare resources to reduce the risk of cancer in SSc patients, or it may indicate a need for further study on the link between SSc and cancer.

An allocation of healthcare resources could, for example, fund research into possible biological mechanisms or research into whether there may be a reaction to SSc treatment which might cause this increased risk, such as has been suggested by Weeding et al. (2020) and discussed in Section 2.3.

A particular focus of the study was the justification and possible benefits of including a prevalent cohort, and considerations for the inclusion of such cohorts. In our study, the prevalent cohort provided a valuable additional resource for analysis, greatly increasing sample size and follow-up times.

A competing risk framework was required because of the high level of mortality,

leading to a necessary distinction being made between direct and total effects and also the consideration of informative censoring. Research was undertaken on alternative methods for such analyses. As the more common nonparametric methods fail to account for informative censoring and, in many studies, also fail to account for confounding, we suggest the application of an increasingly popular causal inference method, the g-formula. We needed to adjust this g-formula methodology to allow for prevalent cohort inclusion, so we investigated weightings to account for those not under observation due to left truncation. Lastly, temporal trends were considered due to the large span of calendar time of diagnosis under study.

In this chapter we discuss specific findings of this thesis. In Section 10.4 we highlight the strengths and limitations, and in Section 10.5 we highlight the originality of this thesis. A summary of what could be considered the key takeaway points is found in Section 10.6.

10.2 Specific conclusions

10.2.1 Mortality in SSc

Although this study has primarily considered the risk of cancer in those diagnosed with SSc, we have also investigated all-cause mortality. In Chapters 4 and 5, this was done with nonparametric methods. However, we note that in Chapters 4 and 5 we lacked patients who were diagnosed with cancer prior to the SSc diagnosis date (or the match date in non-SSc patients), and the results in Chapter 5 were additionally missing patients whose cancer diagnosis preceded study entry. Chapter 6 and later chapters look at the risk of death *without cancer* (which is the competing risk of cancer), so this part of the work excludes the risk of death after a diagnosis of cancer. Therefore, none of these results should be taken to be a complete analysis of mortality of SSc patients.

In Chapters 4 and 5, we demonstrate that the research presented here agrees with

previous research, namely that all-cause mortality is greater in those with SSc than without, with the relative risk being close to 2 in the 20 years after SSc diagnosis, and the hazard ratio being approximately 3. Later, in Chapters 6, 7, 8 and 9, the increased risk of death without cancer in those with SSc is clear throughout, with the mortality risk ratio prior to cancer for SSc patients being more than double that of non-SSc patients. Zeineddine et al. (2016) highlighted the increasing need to study cancer related to SSc following the reduction in deaths due to renal crisis and lung fibrosis. However, in this thesis we have seen that other types of mortality in those with SSc appears at this time to remain a concern, especially looking at the total effect of SSc on cancer and death without cancer.

The presence of a high number of death events is what led to the need to place our work in a competing risk framework. Most previous studies have compared the risk of cancer in SSc patients using SIR metrics, and while this allows for a cause-specific interpretation of risk, this metric is dependent on death occurrences and therefore does not allow for a causal interpretation. Acknowledging this higher mortality led to the application of the g-formula explored by Young et al. (2020) to provide a direct effect of SSc on cancer risk, an interpretation that can be used regardless of the change in mortality.

10.2.2 The inclusion of a prevalent cohort

The inclusion of prevalent patients is often done by treating them as left truncated data. Including them increases the sample size and therefore the power of the study. It also allows for longer follow-up times than we would have had without their inclusion, which would have been a maximum follow-up time of 20 years. However, the traditional NPMLE methods to adjust for left truncation did not sufficiently allow for the direct effect method we wished to utilize due to informative censoring. We provided two ways to use the g-formula when we have a prevalent cohort. One way was to include the prevalent cohort alongside the incident cohort when fitting the logistic models but then to only use the

incident cohort as the pseudo-population when applying the hazard estimates. This was a benefit of our study as we have an incident cohort, and this allowed for a more accurate representation of current covariate distributions. It is also a more intuitive method than weighting patients. The second method is to weight the pseudo-population of SSc patients by their probability of entering the study to account for the unobserved patients. However, we found that this was not effective in our study due to the possibility of temporal trends, which could affect both the hazard over calendar time (dependent on covariates) and the covariate distribution over time. Other studies may observe a greater benefit of the weighted g-formula method than presented here, if the covariate distribution does not change over calendar time. We also note that weightings run the risk of being unstable due to the uncertainty of certain patients being included if probability of left truncation is close to zero, and it is required that patients of all types be in the study over the timeframe of recruitment to avoid this issue. Therefore setting a timeframe for SSc diagnosis may be preferable.

10.2.3 Competing risks and causality

For an accurate interpretation of cancer, death should be treated as a competing event depending on what type of effect we want to estimate: a direct effect or a total effect. We proposed the g-formula to account for this, and also discuss the alternative approach of the IPCW method which was also utilised by Young et al. (2020).

Using a causal framework, as we have done in this study, allowed for the consideration of competing events. The g-formula has also allowed more flexibility in the model. The use of the ATT allowed for a comparison of the hypothetical intervention of SSc patients not being ‘exposed’ to SSc, and if we did not have a matched study this would be a large advantage due to this approach being able to account for confounders. We believe this is quite a novel approach and more epidemiological studies may benefit from using the ATT g-formula estimator, especially as matched datasets can be more costly or time

consuming to compile.

A weighted g-formula estimator was suggested here to account for informative censoring, and this is a novel method. The weighted method did not necessarily produce the results we were expecting, possibly due to temporal trends or differences between the incident and prevalent cohorts. It would be of interest to apply this methodology to other epidemiological studies.

10.2.4 Temporal trends

When considering inclusion of the prevalent cohort, it is important to consider the presence of either independent, dependent or conditionally dependent left truncation. While we do not have dependent truncation we believe we have conditionally independent truncation, and therefore we use calendar time as a covariate to adjust for temporal trends. If we wished to estimate the recent risk then whether we have left truncated data or not is not important to the estimation, and we apply the Breslow estimator. However, if we wanted to estimate the historic risk then we would need to account for missing patients with a weighted marginal cumulative incidence. For the consideration of competing risks we then revert to using the g-formula, although we recognise that we may have continued with a semi-parametric method to account for informative censoring.

We investigated two possibilities for temporal effects: changing hazard over time and changes in measured covariate distributions over time. We conclude that due to changes in measured covariate distributions over time, there appears to be an increasing risk of cancer (and death without cancer) in the more recent cases. Once these covariates have been adjusted for, there are still changes in hazard over calendar time, in that mortality has decreased over calendar time and the hazard of cancer has increased over calendar time. Once calendar time is adjusted for, the direct effect risk of cancer is larger, especially for non-SSc patients, and this results in reduced risk ratios of cancer for SSc patients compared to non-SSc patients. Calendar time is often not accounted for, certainly not in SSc studies

when considering cancer risk. Also, the impact of calendar time in prevalent studies is a topic where there are very limited publications in the literature.

We note that the way we model temporal trends may not be optimal, as with the Cox model we are over-reliant on the proportional hazard model, and patients who were diagnosed further back in calendar time may have more influence due to there being so few patients with these earlier calendar times of SSc diagnosis. It may be of benefit to reduce the calendar time period covered.

10.3 Summary risk ratios of cancer in SSc patients depending on model used

Throughout this study we have used different methods in order to assess the effect of SSc on cancer. These improvements were a) the inclusion of prevalent patients, b) the use of causal models to identify the direct and total effect as separate methods to analyse risks, and c) accounting for calendar time of SSc diagnosis. The plots below summarise the main risk ratios throughout this study, where Figure 10.1 shows the risk ratios of the direct effect, and Figure 10.2 shows the total effect. The risk ratios shown are:

- Incident NPMLE. This is the simplest method and possibly the one most commonly used by other studies, however without prevalent inclusion there are fewer patients under study and there are wider confidence intervals. The time from which SSc is under study is also smaller, hence why only 10-years is shown in the figure. Also, in the direct effect case we may be suffering from informative censoring, and therefore an underestimation of risk. These results can be found in Chapter 6.
- Combined NPMLE. This method benefits from the inclusion of prevalent patients, however we have not yet adjusted for temporal trends. Also, the nonparametric form does not account for informative censoring, making it

possibly the weakest model in our study for the direct effect. These results can be found in Chapter 6.

- Incident g-formula. The incident cohort is used to fit the logistic regression and is then applied to the incident SSc pseudo-population for the g-formula estimation. This allows for a causal interpretation, informative censoring, and a more reliable source of covariate distributions, but only uses half of the patients we have available, along with a shorter follow-up. These results can be found in Chapter 7.
- Combined GLM applied to incident SSc pseudo-population g-formula (denoted as ‘combined applied to incident g-formula’ in Figures 10.1 and 10.2). The combined GLM is applied to the incident SSc pseudo-population g-formula, where both the prevalent and incident cohorts were used to fit the regression model but then the incident SSc group were used as the pseudo-population. This allows for a causal interpretation and a more reliable source of covariate distributions. However, we have not yet adjusted for temporal trends. Also, while we have the advantage of an incident cohort, this method is not possible for a solely prevalent cohort. These results can be found in Section 8.3.
- Combined GLM applied to combined SSc pseudo-population g-formula (denoted as ‘combined applied to combined g-formula’ in Figures 10.1 and 10.2), a novel weighted method where both the prevalent and incident cohorts were used to fit the regression model and were used as the pseudo-population, and where prevalent patients were weighted higher depending on their probability of surviving until entry to the study. This method is preferable if we wished to use the covariate distributions of all patients, where past cases are believed to provide additional information about the true covariate distribution of SSc patients. However, we believe the difference between this method and when the combined GLM is applied to the solely incident pseudo-population is due to changing covariates over time, therefore this method is not ideal for this study, although it could be

applied in other studies. However, this method is preferred if no incident cohort is available. These results can be found in Chapter 8.

- Temporal NPMLE. the Breslow estimator is weighted to represent the most recent risk. This accounts for both covariate distribution changes and hazard changes over calendar time. This method produced larger confidence intervals than other methods. This method does not account for informative censoring. These results can be found in Chapter 9.
- Temporal g-formula. We apply both the prevalent and incident data to fit the regression model, including terms for calendar time and the interaction of calendar time and SSc status, and apply this to a pseudo-population of any patient who was diagnosed after 2000 (so that these patients represent the most recent cohort of patients). This could be argued to be the most accurate model, as we are using the g-formula to account for informative censoring and confounding, and consideration of temporal trends means we are using historic cases to strengthen the hazard estimates but only applying this to the more recent pseudo-population. These results can be found in the final section of Chapter 9.

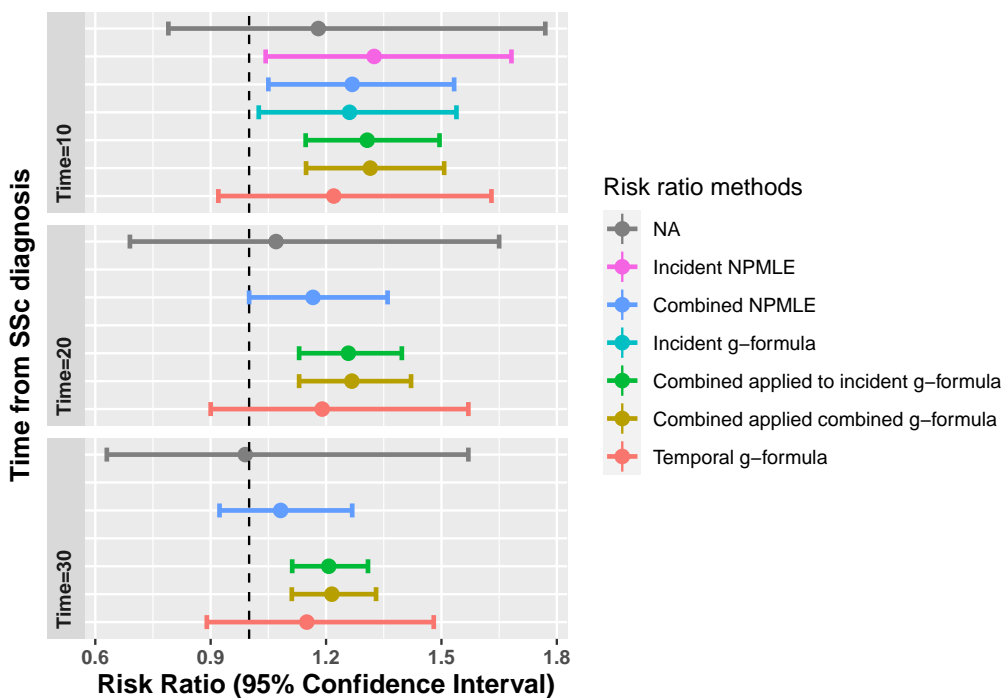


Figure 10.1: Risk ratios for the direct effect of SSc on cancer depending on methods used throughout the thesis. The times taken are at 10, 20, and 30 years after SSc diagnosis.

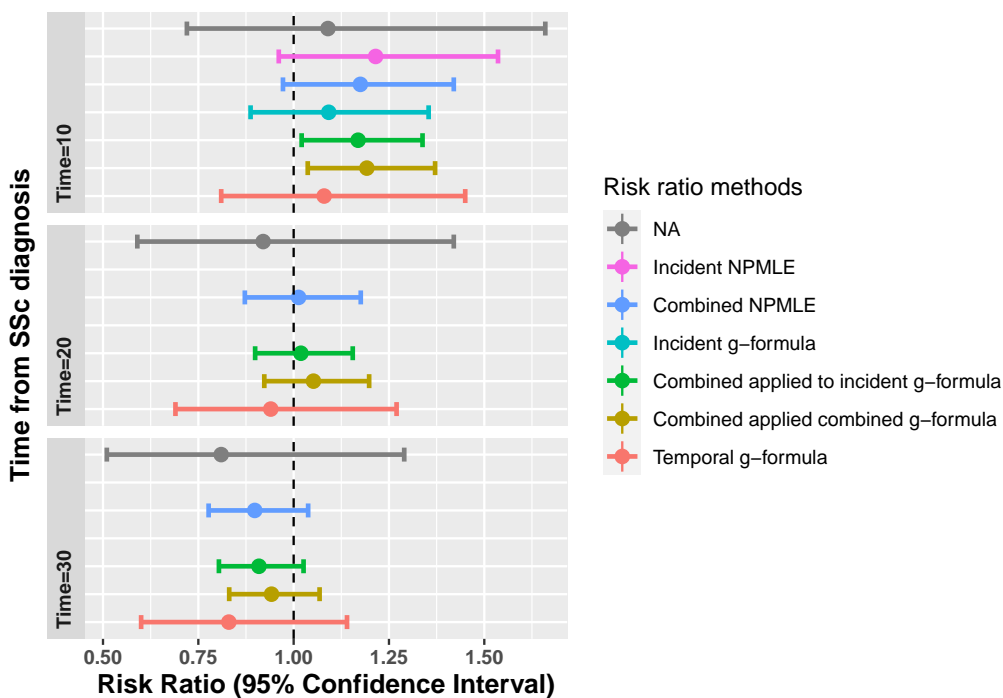


Figure 10.2: Risk ratios for the total effect of SSc on cancer depending on methods used throughout the thesis. The times taken are at 10, 20, and 30 years after SSc diagnosis.

Reviewing Figure 10.1, the risk ratios do not differ greatly between methods except for the recent risk semiparametric method (Temporal semiparametric). The temporally adjusted have wider confidence intervals, and if we believe this method to be the best, then we cannot sufficiently conclude that there is an effect of SSc on cancer. Figure 10.2 shows that the risk ratios have a noticeable decrease over time from SSc diagnosis, however most stay close to 1. For most methods the risk ratio confidence intervals demonstrate that the interval contains 1 for the total effect, implying no total effect of having SSc on the risk of cancer compared to not having SSc. Both figures show that the temporally adjusted curves have a smaller risk ratio, possibly implying a decreasing difference in risk over time.

10.4 Strengths and limitations

10.4.1 Strengths

The major strength of this thesis is that it is a robust analysis of a large dataset. As we have seen from Section 2.3, many studies have previously been done with small, secondary healthcare datasets. Here, we have a large, primary healthcare dataset so that our number of patients increases the power of the study and we were able to match patients.

The advantage of having a matched dataset as a way to account for confounding has the major benefit of making the non-parametric methods applied here more reliable. The matched dataset also allows for direct comparison between SSc and non-SSc patients, without the need for either a general dataset from an outside source, such as by SIR or SMR calculations, or a non-matched study which would need adjustment for confounders. Throughout the study, when considering different methodologies, we have considered the impact a non-matched study would have, and hence what the benefit of the g-formula estimator is.

A key component of this study was using g-formula estimators for direct and total effects as described by Young et al. (2020), which we feel brings to light a very important distinction between direct and total effect which may previously not have been sufficiently defined in competing risk theory. Prior studies of SSc have not made the distinction, often using SIRs which are based on real-world (total) effects as opposed to causal (direct) effects.

When using the g-formula or IPCW, we used regression models to predict the hazard of cancer, death and censoring. We made sure we were using an optimal model by choosing a GAM or GLM. We also investigated higher order terms and interaction terms in order to minimise AIC, and stopped additional terms if the gain for adding the term was minimal. We therefore are confident that our regression models fit the data based on our measured covariates.

Our study is broad, and the aim is that it covers all major considerations of a prevalent dataset. The direction of the study was intentionally flexible to allow for the best analysis of the data, and this permitted development and application

of various methods which then enabled investigation of different aspects as the study progressed. This allowed for the consideration of a competing risk setting as well as the need to consider temporal trends after the findings of Chapter 8.

A variety of methods are described, demonstrating the range available to study a dataset such as this, together with the associated advantages and disadvantages of each. We provided an application of the g-formula in an epidemiological study and suggested an adaptation to allow for prevalent cohorts, which is novel. We found that the different methods produced very different cumulative risk curves, if not different risk ratios.

We reviewed temporal trend analysis and considered how our previous methods could be adjusted to allow for changes over time to reflect risks for patients diagnosed more recently. Temporal trends are often not considered. By introducing temporal trends we found a distinct change in covariate distributions over time for SSc patients, and possible changes in hazards for cancer and death given covariates. The fact that males and older SSc patients are more common in recent times may be of interest in the broader study of SSc.

We hope that this study gives incentive for the consideration of prevalent cohorts, in both how to quantify intrinsic differences that might be observed between incident and prevalent cohorts and how to utilize prevalent cohorts accordingly, whether this is alongside the incident or as a weighted g-formula.

10.4.2 Limitations

Firstly, we should note limitations with the dataset. The dataset does not subset SSc type (either limited or diffuse) which is often done in SSc studies, and which has the advantage of considering the different effect these types have on the body. Also, we have used a limited number of covariates, and have not included other covariates such as individual medical history, environmental exposures or wider lifestyle choices. This is mostly due to restrictions in information available from CPRD GOLD.

There is the concern about CPRD GOLD and its cancer records, due to either missing data or misdiagnosis. There was the potential to use an outside source, such as the hospital Episodes Statistics (HES) or Cancer Registry (CR). As stated in Arhi et al. (2018), “the 10% of cancer cases identified in CPRD or HES not confirmed by the CR in this study are likely to represent false negative cases, rather than true missing cases from the CR. Such false negatives may be a result of a presumed diagnosis by the GP.” Also, the CPRD will underestimate cancer incidence in patients who are older and those diagnosed through an emergency route, giving the possible impression of better survival. Lastly, CPRD has a delay of cancer recordings compared to the CR. There is an incentive for linkage to verify CPRD cases and add possible cancer non-recorded cases. This was not done in this study due to the extra cost associated with data-linkage, but it may have been a large benefit.

Within this study we assume that the diagnosis date of SSc is the start of SSc, time zero, as inevitably there is no information on the actual onset time of SSc. However, there will clearly be a delay between onset of SSc and diagnosis, as patients will be diagnosed once they start to present symptoms. In addition, SSc diagnosis dates may not be precise. Earlier diagnosis dates of SSc may be recorded in the database based on when the patient first thinks they were diagnosed with SSc, and this is therefore prone to human error. Also, many patients prior to 1998 have their diagnosis set to the 1st of January of the year they were diagnosed, which would extend the time between SSc and end event by up to twelve months. Furthermore, we have previously mentioned that there are four patients whose diagnosis date coincided with their death date, and these patients were included in our study. However, as it is almost certain that these four SSc patients developed SSc before the date of their death, this makes the date of diagnosis unreliable and delayed. It may have been preferable to remove them.

In consideration of the g-formula, the exchangeability assumption of the g-formula requires that there are no unmeasured confounders, however there may be some of potential significance we have not been able to consider, such as silica exposure

or prior medical treatment. In particular, there is no information from secondary care datasets, which may have more accurate and detailed account of SSc treatment, or even cancer treatment (as we believe there may be a possible relationship between SSc and cancer treatment).

Smoking, alcohol habits and BMI information for prevalent patients were taken at entry to the study as opposed to prior to SSc diagnosis, which is not the assumption made when using the g-formula. This may be a particular issue for BMI, as BMI often rises over time and therefore prevalent patients may be viewed as less healthy on entry to the study, hence introducing misclassification error. The only one of these covariates used for the g-formula estimator was smoking, and in particular smoking may cease after an SSc diagnosis, which would therefore differ if smoking had been recorded at baseline. However we believe this difference in habits is small, and we believe that including smoking is informative despite this weakness in the data.

In Section 4.2.1, matching was discussed. We decided for the analysis to treat the matching as an unmatched study, however it is possible (depending on if the unmatched covariates are conditionally independent of the exposure) that GP practice should have been accounted for as a random effect. This feature of the dataset could have been investigated more thoroughly.

We have included sections deriving and estimating the IPCW, but the g-formula estimator was the method that was the focus of this thesis. The IPCW section for how prevalent patients are included could be expanded. For example, we could test on simulations or on a dataset with known informative loss to follow-up. In particular, there remains some concern regarding which conditions need to hold for this method to be viable, such as independence between entry time and event time, or entry time and censoring. This could be explored in further work.

We have separated what we qualify as risk between the ‘direct’ and ‘total’ effect. However, the direct effect is attempting to quantify the risk of cancer under the assumption of the elimination of death. Death cannot be eliminated in practice, hence this is a hypothetical quantity which makes interpretation difficult, but

we believe this is the best way to assess the causal impact of SSc on cancer. A different approach may be via separable effects, which could provide a subject for further work (Stensrud et al., 2021).

10.5 Original contribution

This section summarises the novel contributions of the thesis, and where the originality lies. We separate the original contribution into ‘epidemiological’ and ‘statistical’.

1. Epidemiological considerations. We have investigated both the direct and total effect of cancer risk in patients with SSc. In an area of research that tends to use SIRs to detect an increased risk of cancer in SSc, our large and comprehensive dataset has allowed for a more thorough investigation. Due to competing events, we have defined the difference between the direct effect and total effect, which has never been used in SSc research before. We have investigated the use of prevalent patients to strengthen an incident dataset, and have determined how to assess the impact and usefulness of their inclusion. The inclusion of our prevalent patients allowed for longer follow-up times, which has often not been available in other studies. Including prevalent patients may result in less expensive studies, as new and costly studies requiring fresh data do not need to be done if retrospective data collection can be utilised. We have concluded that there may be a small statistically significant increase in the direct effect of SSc on cancer risk, however, after temporal trends are included this is no longer at a statistically significant level. But in contrast, the total effect implies no significant difference in risk between SSc and non-SSc patients at shorter times, while conversely suggesting a more significant difference at longer times. These findings differ to many other studies, which have produced a wide array of different SIRs, with a predominance of results indicating a raised risk higher than our results suggest.

2. Statistical considerations. We built upon prior research to allow for prevalent data to be used in both a competing risk and causal framework. The use of the g-formula in a matched study has rarely been done before, and its use in a competing risk setting is also rare. The causal framework allowed for a more theoretical interpretation of direct and total effect. The inclusion of prevalent cohorts has not been done before, and we hope that this exploration of left truncated data might be used and tested in the future. Temporal trends, and the concept of a recent and historic risk, have previously been investigated only rarely.

10.6 Possibilities for future work

We note some key areas for future work that would extend the results obtained in this thesis:

- Testing on other datasets - The methods proposed here have only been performed on one dataset. It would be of interest to apply these methods to other datasets. In particular, it would be interesting to apply the suggested weighted g-formula to a dataset with a long follow-up time, which can then be transformed into a prevalent cohort (as was done to our incident dataset in Section 8.7.1), and compared to the true risk, to demonstrate accuracy. It may be of interest to apply the weighted g-formula to a dataset that does not have temporal trends.
- Time-varying covariates – An advantage of the g-formula is that it permits the use of time-varying covariates. Methods to approximate time-varying covariates, especially in prevalent cohorts, could be beneficial. Young et al. (2020) mention this as a possibility in their work, however for simplicity they too do not use them.
- Average treatment effect in an unmatched study - We have a matched study, and as can be seen in Chapter 7 there is limited difference in the NPMLE

and g-formula curves for the solely incident dataset, and the risk ratios are very similar. We hypothesise that this is due to the matching, meaning the pseudo-populations had a similar covariate distribution to the SSc and non-SSc dataset. It would be very interesting to apply the same theory to a dataset where there was more substantial confounding, and observe if the g-formula is correcting for confounding. This could suggest an improvement compared to the non-parametric method.

- One of the key challenges of using the g-formula is the underlying assumptions. Being able to test for these and/or undertaking more work into the consequences of the assumptions not holding would enable more confidence to be placed in this method.
- Young et al. (2020) provided the IPCW method alongside the g-formula but as stated above this is a less developed area for left truncation. Further work in this area would be beneficial.
- We focused on Young et al.'s methodology, however an alternative method using separable effects has been recently published, which does not need the assumption of the elimination of the competing event which is required for the direct effect (Stensrud et al., 2021), and this is an approach worthy of further investigation.

10.7 Overall conclusion

This thesis presents different statistical approaches to answer the question of ‘does SSc increase the risk of cancer?’. We hope that we have answered that question, the conclusion being that there is not a statistically significant causal impact of SSc on the overall risk of cancer (RR approximately 1.1-1.3 between 0 to 30 years from diagnosis), but with a risk ratio slightly above one we cannot rule out a much larger causal effect for an individual cancer type combined with no effect in other cancer types. We have also demonstrated that all-cause mortality is significantly

greater in those with SSc than without, indicating that causes of death other than cancer in SSc patients remain a cause for concern. We have also highlighted that the increased mortality in SSc patients will obscure the observation of diseases which could occur prior to death. The consideration of prevalent cohorts has been an overarching theme, and we hope this thesis provides encouragement for their inclusion alongside the typical incident cohort and serves as a useful overview of this area. We hope the use of causal methods adds to this growing and increasingly popular area, and especially now, with the additional ability to include prevalent cohorts. A review of recent research in temporal trends provides a further way to accommodate prevalent cohorts and data with large follow-up times. Finally, we present possible future areas of research suggested by this study.

Chapter 11

References

Aalen, O., Borgan, O., & Gjessing, H. (2008). *Survival and event history analysis: A process point of view*. Springer Science & Business Media.

Abbot, S., Bossingham, D., Proudman, S., Costa, C. de, & Ho-Huynh, A. (2018). Risk factors for the development of systemic sclerosis: A systematic review of the literature. *Rheumatology Advances in Practice*, 2(2), rky041.

Ahmad, A. S., Ormiston-Smith, N., & Sasieni, P. D. (2015). Trends in the lifetime risk of developing cancer in Great Britain: Comparison of risk for those born from 1930 to 1960. *British Journal of Cancer*, 112(5), 943–947.

Alba, M. A., Velasco, C., Simeón, C. P., Fonollosa, V., Trapiella, L., Egurbide, M. V., Sáez, L., Castillo, M. J., Callejas, J. L., Camps, M. T., & others. (2014). Early-versus late-onset systemic sclerosis: Differences in clinical presentation and outcome in 1037 patients. *Medicine*, 93(2).

Allcock, R., Forrest, I., Corris, P., Crook, P., & Griffiths, I. (2004). A study of the prevalence of systemic sclerosis in northeast England. *Rheumatology*, 43(5), 596–602.

Allignol, A., Beyersmann, J., Gerds, T., & Latouche, A. (2014). A competing risks approach for nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis*, 20(4), 495–513.

- Allignol, A., Schumacher, M., & Beyersmann, J. (2010). A note on variance estimation of the Aalen-Johansen estimator of the cumulative incidence function in competing risks, with a view towards left-truncated data. *Biometrical Journal*, *52*(1), 126–137.
- Allison, P. (2018). *For causal analysis of competing risks, don't use Fine and Gray's subdistribution method*. Statistical Horizons. <https://statisticalhorizons.com/for-causal-analysis-of-competing-risks/>
- Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (2012a). *Statistical models based on counting processes*. Springer Science & Business Media.
- Andersen, P. K., Geskus, R. B., Witte, T. de, & Putter, H. (2012b). Competing risks in epidemiology: Possibilities and pitfalls. *International Journal of Epidemiology*, *41*(3), 861–870.
- Arhi, C. S., Bottle, A., Burns, E. M., Clarke, J. M., Aylin, P., Ziprin, P., & Darzi, A. (2018). Comparison of cancer diagnosis recording between the clinical practice research datalink, cancer registry and hospital episodes statistics. *Cancer Epidemiology*, *57*, 148–157.
- Austin, P. C., Lee, D. S., & Fine, J. P. (2016). Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, *133*(6), 601–609.
- Bonifazi, M., Tramacere, I., Pomponio, G., Gabrielli, B., Avvedimento, E., La Vecchia, C., Negri, E., & Gabrielli, A. (2013). Systemic sclerosis (scleroderma) and cancer risk: Systematic review and meta-analysis of observational studies. *Rheumatology*, *52*(1), 143–154.
- Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 89–99.
- Breslow, N., & Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 437–453.
- Butt, S. A., Jeppesen, J. L., Fuchs, C., Mogensen, M., Engelhart, M.,

- Torp-Pedersen, C., Gislason, G. H., Jacobsen, S., & Andersson, C. (2018). Trends in incidence, mortality, and causes of death associated with systemic sclerosis in Denmark between 1995 and 2015: A nationwide cohort study. *BMC Rheumatology*, *2*(1), 36.
- Calvert, G., Rice, F., Boiano, J., Sheehy, J., & Sanderson, W. (2003). Occupational silica exposure and risk of various diseases: An analysis using death certificates from 27 states of the United States. *Occupational and Environmental Medicine*, *60*(2), 122–129.
- CancerResearchUK. (2020). Cancer mortality statistics (CRUK). In *Cancer Research UK*. Cancer Research UK. <https://www.cancerresearchuk.org/health-professional/cancer-statistics/mortality/age>
- Chan, K. C. G., & Wang, M.-C. (2012). Estimating incident population distribution from prevalent data. *Biometrics*, *68*(2), 521–531.
- Chatterjee, S., Dombi, G. W., Severson, R. K., & Mayes, M. D. (2005). Risk of malignancy in scleroderma: A population-based cohort study. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, *52*(8), 2415–2424.
- Chisholm, J. (1990). The read clinical classification. *BMJ: British Medical Journal*, *300*(6732), 1092.
- Coi, A., Barsotti, S., Santoro, M., Almerigogna, F., Bargagli, E., Caproni, M., Emmi, G., Frediani, B., Guiducci, S., Cerinic, M. M., & others. (2021). Epidemiology of systemic sclerosis: A multi-database population-based study in Tuscany (Italy). *Orphanet Journal of Rare Diseases*, *16*(1), 1–13.
- Cole, S. R., Richardson, D. B., Chu, H., & Naimi, A. I. (2013). Analysis of occupational asbestos exposure and lung cancer mortality using the g-formula. *American Journal of Epidemiology*, *177*(9), 989–996.
- Datta, S., & Satten, G. A. (2002). Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under dependent censoring. *Biometrics*, *58*(4), 792–802.

- Derk, C. T., Rasheed, M., Artlett, C. M., & Jimenez, S. A. (2006). A cohort study of cancer incidence in systemic sclerosis. *The Journal of Rheumatology*, *33*(6), 1113–1116.
- Elhai, M., Meune, C., Avouac, J., Kahan, A., & Allanore, Y. (2012). Trends in mortality in patients with systemic sclerosis over 40 years: A systematic review and meta-analysis of cohort studies. *Rheumatology*, *51*(6), 1017–1026.
- Englert, H., Small-McMahon, J., Chambers, P., O'Connor, H., Davis, K., Manolios, N., White, R., Dracos, G., & Brooks, P. (1999). Familial risk estimation in systemic sclerosis. *Australian and New Zealand Journal of Medicine*, *29*(1), 36–41.
- Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, *94*(446), 496–509.
- Gabrielli, A., Avvedimento, E., & Krieg, T. (2009). Scleroderma. *New England Journal of Medicine*, *360*(19), 1989–2003.
- Geskus, R. B. (2011). Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring. *Biometrics*, *67*(1), 39–49.
- Gray, R. J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 1141–1154.
- Greene, T., & Li, L. (2013). A weighting analogue to pair matching in propensity score analysis. *The International Journal of Biostatistics*, *9*(2), 215–234.
- Greenwood, M. (1926). A report on the natural duration of cancer. *A Report on the Natural Duration of Cancer*, *33*.
- Hernan, M. A., & Robins, J. M. (2020). *Causal Inference: What If*. CRC Boca Raton, FL.
- Hernan, M. A., Brumback, B., & Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive

men. *Epidemiology*, 561–570.

Herrett, E., Gallagher, A. M., Bhaskaran, K., Forbes, H., Mathur, R., Staa, T. van, & Smeeth, L. (2015). Data resource profile: Clinical practice research datalink (CPRD). *International Journal of Epidemiology*, 44(3), 827–836.

Hill, C., Nguyen, A., Roder, D., & Roberts-Thomson, P. (2003). Risk of cancer in patients with scleroderma: A population based cohort study. *Annals of the Rheumatic Diseases*, 62(8), 728–731.

Ioannidis, J. P., Vlachoyiannopoulos, P. G., Haidich, A.-B., Medsger Jr, T. A., Lucas, M., Michet, C. J., Kuwana, M., Yasuoka, H., Van Den Hoogen, F., Te Boome, L., & others. (2005). Mortality in systemic sclerosis: An international meta-analysis of individual patient data. *The American Journal of Medicine*, 118(1), 2–10.

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5), 402.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481.

Keiding, N. (1991). Age-specific incidence and prevalence: A statistical perspective. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 154(3), 371–396.

Keiding, N., & Gill, R. (1990). Random truncation models and Markov processes. *The Annals of Statistics*, 582–602.

Keiding, N., & Moeschberger, M. (1992). Independent delayed entry. In *Survival analysis: State of the art* (pp. 309–326). Springer.

Keil, A. P., & Edwards, J. K. (2018). A review of time scale fundamentals in the g-formula and insidious selection bias. *Current Epidemiology Reports*, 5(3), 205–213.

Keil, A. P., Edwards, J. K., Richardson, D. R., Naimi, A. I., & Cole, S. R. (2014). The parametric G-formula for time-to-event data: Towards intuition with

- a worked example. *Epidemiology (Cambridge, Mass.)*, 25(6), 889.
- Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data* (Vol. 1230). Springer.
- Lai, T. L., & Ying, Z. (1991). Estimating a distribution function with truncated and censored data. *The Annals of Statistics*, 417–442.
- Li, F., & Thomas, L. (2019). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, 188(1), 250–257.
- Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices of the Royal Astronomical Society*, 155(1), 95–118.
- Mackenzie, T. (2012). Survival curve estimation with dependent left truncated data using Cox’s model. *The International Journal of Biostatistics*, 8(1).
- Manno, R. L., Wigley, F. M., Gelber, A. C., & Hummers, L. K. (2011). Late-age onset systemic sclerosis. *The Journal of Rheumatology*, 38(7), 1317–1325.
- Mansfield, K., Crellin, E., Denholm, R., Quint, J. K., Smeeth, L., Cook, S., & Herrett, E. (2019). Completeness and validity of alcohol recording in general practice within the uk: A cross-sectional study. *BMJ Open*, 9(11), e031537.
- Mansournia, M. A., Hernán, M. A., & Greenland, S. (2013). Matched designs and causal diagrams. *International Journal of Epidemiology*, 42(3), 860–869.
- Mansournia, M. A., Jewell, N. P., & Greenland, S. (2018). Case-control matching: Effects, misconceptions, and recommendations. *European Journal of Epidemiology*, 33(1), 5–14.
- Martin, E. C., & Betensky, R. A. (2005). Testing quasi-independence of failure and truncation times via conditional Kendall’s tau. *Journal of the American Statistical Association*, 100(470), 484–492.
- Masi, A. T., American Rheumatism Association Diagnostic, S. F. S. C. of the, & Committee, T. C. (1980). Preliminary criteria for the classification of systemic sclerosis (scleroderma). *Arthritis & Rheumatism*, 23(5), 581–590.

- Mayes, M. D., Lacey Jr, J. V., Beebe-Dimmer, J., Gillespie, B. W., Cooper, B., Laing, T. J., & Schottenfeld, D. (2003). Prevalence, incidence, survival, and disease characteristics of systemic sclerosis in a large US population. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, *48*(8), 2246–2255.
- Molina, J., Sued, M., & Valdora, M. (2018). Models for the propensity score that contemplate the positivity assumption and their application to missing data and causality. *Statistics in Medicine*, *37*(24), 3503–3518.
- Morrisroe, K., Hansen, D., Huq, M., Stevens, W., Sahhar, J., Ngian, G.-S., Ferdowski, N., Hill, C., Roddy, J., Walker, J., & others. (2020). Incidence, risk factors, and outcomes of cancer in systemic sclerosis. *Arthritis Care & Research*, *72*(11), 1625–1635.
- Murray, E. J., Robins, J. M., Seage III, G. R., Lodi, S., Hyle, E. P., Reddy, K. P., Freedberg, K. A., & Hernán, M. A. (2018). Using observational data to calibrate simulation models. *Medical Decision Making*, *38*(2), 212–224.
- NHS. (2020). Health Survey for England, 2019: Data tables. In *NHS digital*. NHS digital. <https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2019/health-survey-for-england-2019-data-tables>
- Nietert, P., & Silver, R. (2000). Systemic sclerosis: Environmental and occupational risk factors. *Current Opinion in Rheumatology*, *12*(6), 520–526.
- Olesen, A. B., Svaerke, C., Farkas, D., & Sørensen, H. T. (2010). Systemic sclerosis and the risk of cancer: A nationwide population-based cohort study. *British Journal of Dermatology*, *163*(4), 800–806.
- ONS. (2020). Births, Deaths and Marriages, ONS. In *Overview of the UK population - Office for National Statistics*. Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/overviewoftheukpopulation/mar2017>
- Pan, W., & Chappell, R. (1999). A note on inconsistency of NPMLE of the distribution function from left truncated and case I interval censored data.

Lifetime Data Analysis, 5(3), 281–291.

Pauling, J., McGrogan, A., Snowball, J., & McHugh, N. J. (2021). Epidemiology of systemic sclerosis in the UK: An analysis of the Clinical Practice Research Datalink. *Rheumatology*, 60(6), 2688–2696.

Pearl, J. (2009). *Causality*. Cambridge university press.

Peng, H., Wu, X., Wen, Y., Li, C., Lin, J., Li, J., Xiong, S., Zhong, R., Liang, H., Cheng, B., & others. (2020). Association between systemic sclerosis and risk of lung cancer: Results from a pool of cohort studies and Mendelian randomization analysis. *Autoimmunity Reviews*, 102633.

Pokeerbux, M., Giovannelli, J., Dauchet, L., Mouthon, L., Agard, C., Lega, J.-C., Allanore, Y., Jegou, P., Bienvenu, B., Berthier, S., & others. (2019). Survival and prognosis factors in systemic sclerosis: Data of a French multicenter cohort, systematic review, and meta-analysis of the literature. *Arthritis Research & Therapy*, 21(1), 1–12.

Prentice, R. L., & Kalbfleisch, J. D. (1979). Hazard rate models with covariates. *Biometrics*, 25–39.

Qian, J., & Betensky, R. A. (2014). Assumptions regarding right censoring in the presence of left truncation. *Statistics & Probability Letters*, 87, 12–17.

Randall, M. (2017). Overview of the UK population: March 2017. In *Overview of the UK population - Office for National Statistics*. Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/overviewoftheukpopulation/mar2017>

Rennert, L., & Xie, S. X. (2021). Cox regression model under dependent truncation. *Biometrics*.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12), 1393–1512.

- Robins, J. M., & Finkelstein, D. M. (2000). Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*, *56*(3), 779–788.
- Robins, J. M., & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology* (pp. 297–331). Springer.
- Rodríguez, L. A. G., González-Pérez, A., Michel, A., & Sáez, M. E. (2019). Contemporary epidemiology of systemic sclerosis: A population-based cohort study in the united kingdom. *Seminars in Arthritis and Rheumatism*, *49*, 105–111.
- Rosenthal, A. K., McLaughlin, J. K., Gridley, G., & Nyrén, O. (1995). Incidence of cancer among patients with systemic sclerosis. *Cancer*, *76*(5), 910–914.
- Rothman, K. J. (2012). *Epidemiology: An introduction*. Oxford University Press.
- Royle, J. G., Lanyon, P. C., Grainge, M. J., Abhishek, A., & Pearce, F. A. (2018). The incidence, prevalence, and survival of systemic sclerosis in the UK Clinical Practice Research Datalink. *Clinical Rheumatology*, *37*(8), 2103–2111.
- Rubio-Rivas, M., Royo, C., Simeón, C. P., Corbella, X., & Fonollosa, V. (2014). Mortality and survival in systemic sclerosis: Systematic review and meta-analysis. *Seminars in Arthritis and Rheumatism*, *44*, 208–219.
- Sackett, D. L. (1979). Bias in analytic research. In *The case-control study consensus and controversy* (pp. 51–63). Elsevier.
- Satten, G. A., & Datta, S. (2001). The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, *55*(3), 207–210.
- Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, *69*(1), 239–241.
- Shen, P.-s. (2003). The product-limit estimate as an inverse-probability-weighted average. *Communications in Statistics-Theory and Methods*, *32*(6), 1119–1133.

- Sjölander, A., & Greenland, S. (2013). Ignoring the matching variables in cohort studies—when is it valid and why? *Statistics in Medicine*, *32*(27), 4696–4708.
- Smittenaar, C., Petersen, K., Stewart, K., & Moitt, N. (2016). Cancer incidence and mortality projections in the UK until 2035. *British Journal of Cancer*, *115*(9), 1147–1155.
- Steen, V. D., & Medsger, T. A. (2000). Long-term outcomes of scleroderma renal crisis. *Annals of Internal Medicine*, *133*(8), 600–603.
- Steen, V. D., & Medsger, T. A. (2007). Changes in causes of death in systemic sclerosis, 1972–2002. *Annals of the Rheumatic Diseases*, *66*(7), 940–944.
- Steenland, K. (2005). One agent, many diseases: Exposure-response data and comparative risks of different outcomes following silica exposure. *American Journal of Industrial Medicine*, *48*(1), 16–23.
- Stegherr, R., Allignol, A., Meister, R., Schaefer, C., & Beyersmann, J. (2020). Estimating cumulative incidence functions in competing risks data with dependent left-truncation. *Statistics in Medicine*, *39*(4), 481–493.
- Stegherr, R., Beyersmann, J., Jehl, V., Rufibach, K., Leverkus, F., Schmoor, C., & Friede, T. (2021). Survival analysis for AdVerse events with VarYing follow-up times (SAVVY): Rationale and statistical concept of a meta-analytic study. *Biometrical Journal*, *63*(3), 650–670.
- Stensrud, M. J., Hernán, M. A., Tchetgen Tchetgen, E. J., Robins, J. M., Didelez, V., & Young, J. G. (2021). A generalized theory of separable effects in competing event settings. *Lifetime Data Analysis*, *27*(4), 588–631.
- Strickland, G., Pauling, J., Cavill, C., Shaddick, G., & McHugh, N. (2013). Mortality in systemic sclerosis—a single centre study from the UK. *Clinical Rheumatology*, *32*(10), 1533–1539.
- Taubman, S. L., Robins, J. M., Mittleman, M. A., & Hernán, M. A. (2009). Intervening on risk factors for coronary heart disease: An application of the parametric g-formula. *International Journal of Epidemiology*, *38*(6), 1599–1611.

- Therneau, T., Crowson, C., & Atkinson, E. (2020). Multi-state models and competing risks. In *CRAN-R*. CRAN-R. <https://cran.r-project.org/web/packages/survival/vignettes/compete.pdf>
- Thomas, E., Brewster, D., Black, R., & Macfarlane, G. (2000). Risk of malignancy among patients with rheumatic conditions. *International Journal of Cancer*, *88*(3), 497–502.
- Tsai, W.-Y. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika*, *77*(1), 169–177.
- Tsai, W.-Y., Jewell, N. P., & Wang, M.-C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, *74*(4), 883–886.
- Vakulenko-Lagun, B., Qian, J., Chiou, S. H., Wang, N., & Betensky, R. A. (2021). Nonparametric estimation of the survival distribution under covariate-induced dependent truncation. *Biometrics*.
- Vakulenko-Lagun, B., Qian, J., Chiou, S., & Betensky, R. (2019). Nonidentifiability in the presence of factorization for truncated data. *Biometrika*, *106*(3), 724–731.
- Van Den Hoogen, F., Khanna, D., Fransen, J., Johnson, S. R., Baron, M., Tyndall, A., Matucci-Cerinic, M., Naden, R. P., Medsger Jr, T. A., Carreira, P. E., & others. (2013). 2013 classification criteria for systemic sclerosis: An American College of Rheumatology/European League against Rheumatism collaborative initiative. *Arthritis & Rheumatism*, *65*(11), 2737–2747.
- Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika*, *76*(4), 751–761.
- Wang, M.-C. (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association*, *84*(407), 742–748.
- Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, *86*(413), 130–143.
- Wang, M.-C., Brookmeyer, R., & Jewell, N. P. (1993). Statistical models for prevalent cohort data. *Biometrics*, 1–11.

- Weeding, E., Casciola-Rosen, L., & Shah, A. A. (2020). Cancer and scleroderma. *Rheumatic Disease Clinics*, *46*(3), 551–564.
- Wolf, A., Dedman, D., Campbell, J., Booth, H., Lunn, D., Chapman, J., & Myles, P. (2019). Data resource profile: Clinical practice research datalink (cprd) aurum. *International Journal of Epidemiology*, *48*(6), 1740–1740g.
- Wolfson, D., Best, A., Addona, V., Wolfson, J., & Gadalla, S. (2019). Benefits of combining prevalent and incident cohorts: An application to myotonic dystrophy. *Statistical Methods in Medical Research*, *28*(10-11), 3333–3345.
- Woodroffe, M. (1985). Estimating a distribution function with truncated data. *The Annals of Statistics*, *13*(1), 163–177.
- Young, J. G., Stensrud, M. J., Tchetgen Tchetgen, E. J., & Hernán, M. A. (2020). A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in Medicine*, *39*(8), 1199–1236.
- Zeineddine, N., El Khoury, L., & Mosak, J. (2016). Systemic sclerosis and malignancy: A review of current data. *Journal of Clinical Medicine Research*, *8*(9), 625.

Appendix A

Appendix

This chapter contains simulations which are not directly required within the thesis but which may be of interest for understanding.

A.1 Relating to Chapter 5: Simulation of truncation times

We perform simulations to demonstrate reasons as to why we might observe a non-uniform truncation distribution. We set a study we we time $\tau = 2000$ calendar year and patient size $n_{obs} = 50,000$, however some of these patients will be left truncated. Let I_i be initial disease onset simulation, L_i time from initial disease to study entry and T_i be the time from initial disease to event time, in this case death.

- Simulation 1, S1 has a static entry date, 2000, where onset of initial disease is distributed $I_{S1} \sim U(1800, 2100)$, and from this time distribution of event time $T_{S1} \sim Exp(0.05)$. This is to replicate the stationarity assumption from Wang.
- Simulation 2, S2, will have a non-uniform initial event time such that $I_2 = 2000 - Exp(0.05)$, and one survival distribution $T_{S1} \sim Exp(0.05)$. This

is to replicate a non-stationarity assumption, as the distribution leads to more patients with short truncation times. This could be due to more patients entering the study at later times, or it could be due to an increase in incidence of SSc.

- Simulation 3, S3, will have a uniform distribution of onset $I_{S3} \sim Unif(1800, 2100)$, but shall have two different survival distributions dependent on whether the patient's diagnosis was prior to 1990, $T_{S3, I_3 < 1990} \sim Exp(0.1)$ or after 1990, $T_{S3, I_3 > 1990} \sim Exp(0.05)$. This is to represent the possibility of a new treatment which increases survival. While the stationarity assumption is satisfied here, we have temporal trends for survival, implying a dependence between survival and onset time.

In all three studies, if a patient does not survive until 2000 they are removed. For simplicity, there is no censoring. The results are shown in Figure A.1. The non-weighted CDF is $\mathbb{I}(L_i \leq t)$, where we count the number who have entered by time t , not weighted by the probability of surviving to time t . The weighted distribution is estimated using Equation (5.1).

- Simulation 1 demonstrates the straight uniform distribution expected from the stationarity assumption. In this circumstance, the weighted left truncation distribution shows a uniform distribution.
- Simulation 2 shows a non-uniform CDF. This curve is due to the distribution of initial exposure, where we have more patients with small truncation times than large ones. The curve follows the CDF of $Exp(0.05)$, which is the same as the distribution from the study entry.
- Simulation 3 again shows a curved estimation, even for the IPW estimator. This simulation is lower than the comparison curve for long truncation times, as the survival prediction, which uses both the pre-1990 and post-1990 patients to predict survival, gives an overestimated survival probability,

hence there are fewer patients surviving to this truncation time than would otherwise be expected.

Both Simulation 2 and Simulation 3 have truncation distributions comparable to our curves. Simulation 1 and Simulation 2 have a hazard that is not dependent on calendar time. These both reflect the true truncation distribution. Simulation 3 does not demonstrate the most recent risk of mortality, due to the inclusion of pre-1990 prevalent cases, however this can be corrected using methods to account for dependent truncation. We examined this further in Chapter 9, with a complementary simulation in Appendix A2.

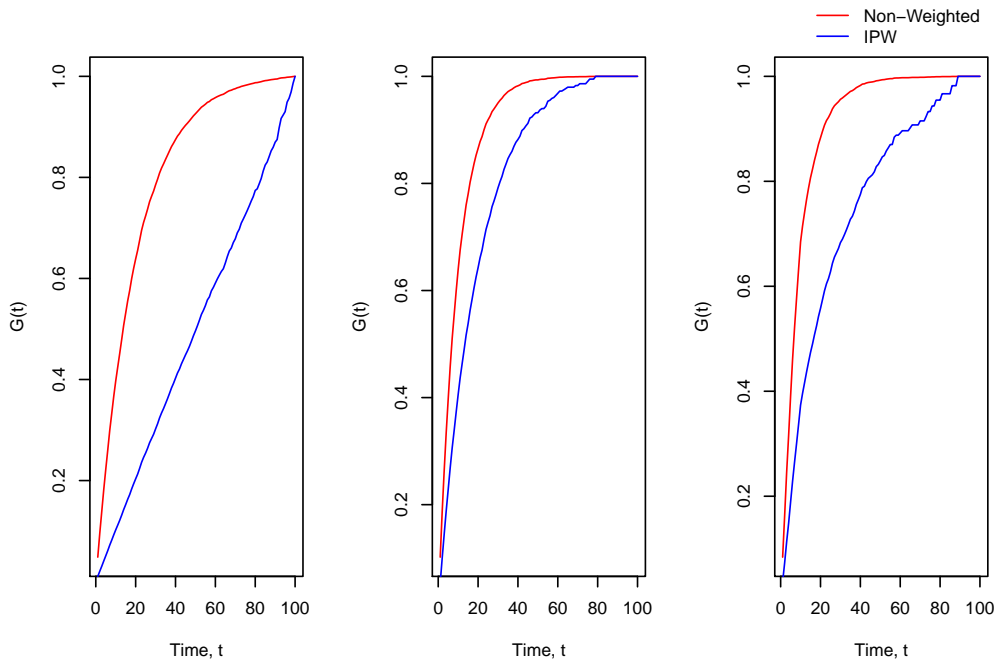


Figure A.1: Simulations of cumulative distribution functions for left truncation distribution. Left: S1, demonstrating a uniform distribution. Middle: S2, demonstrating changes in SSc incidence over calendar time. Right S3, demonstrating changes in survival probability over calendar time

A.2 Relating to Chapter 9: Simulation of conditionally independent truncation

This section contains simulations which help with the understanding of Chapter 9. As the novel material here is limited, the content has been excluded from the main thesis.

We demonstrate the application of the calendar time estimators for simulations, one with one event of interest and the other with competing risks.

A.2.1 Simulation 1: One outcome

We will have patients born with a uniform distribution between 1800 and 2100. In order to be similar to the patients in our study, we allow patients to enter in different years, between 2000 and 2010[^][This creates a distinction between dependent truncation and conditionally independent truncation, as if we have a static recruitment the date truncation time and calendar time would be equal. This was already established by Stegherr et al. (2020). Therefore, if a patient is exposed prior to their entry time and are event free by that time, then they will be prevalent patients, and if SSc diagnosis happens after their entry time then they are incident patients. The initial event, SSc, is simulated from the time of birth in years $N(50, 10^2)$. After SSc, their risk of death is distributed $T_{SSc < 1995} \sim Exp(0.2)$ if the patient was diagnosed with SSc before 1995 and $T_{SSc > 1995} \sim Exp(0.05)$ if they were diagnosed after 1995. This is to add a calendar effect. We will not have censoring.

A simulation to estimate the hazard ratio of calendar time was performed with 500,000 simulated patients and averaged over 100 simulations. The categories are under the year 1990, 1990-1995, 1995-1998, 1998-2000 and 2000+, where the baseline variable is 2000+. While this is a large sample, many of these patients will either not survive until entry to the study or are incident patients. The Cox model shown in Table A.1 estimates the average hazard ratio of patients

depending on their truncation time, with truncation as a categorical variable and the baseline covariate as the year 2000+. Both incident and prevalent patients were included.

As is expected, the hazard ratio is close to one in those whose calendar time of SSc diagnosis was after 1995, but we then observe four times when the hazard in those diagnosed with SSc was seen prior to the year 1995. The confidence intervals are percentile intervals based on 100 simulations and show a higher uncertainty at larger calendar times of SSc diagnosis.

Calendar year of SSc diagnosis	Mean HR	95% CI
<1990	4.01	[3.70, 4.36]
1990-1995	4.01	[3.77, 4.27]
1995-1998	1.00	[0.97, 1.03]
1998-2000	1.00	[0.97, 1.04]

Table A.1: Cox model for simulation, with calendar time of diagnosis compared to those diagnosed with SSc after 2000. The hazard ratio is the mean of the 500 samples.

Figure A.2 demonstrates the differences in truncation distribution depending on whether dependent truncation is accounted for. This is performed using one simulation of originally 500,000 patients, although we only use the simulated prevalent patients here. The black curve shows the application of the independent estimator of the truncation distribution from Section 5.4.2, and the dependent truncation corrected truncation distribution based on Equation (9.4), shown in green. The black curve has a distinct change in gradient at 5 years, where the change in hazard distribution occurs. Once dependency is considered, we see a more uniform distribution. It is not a perfectly uniform distribution, but this is due to there being fewer patients (and therefore less power) at higher calendar times of SSc diagnosis.

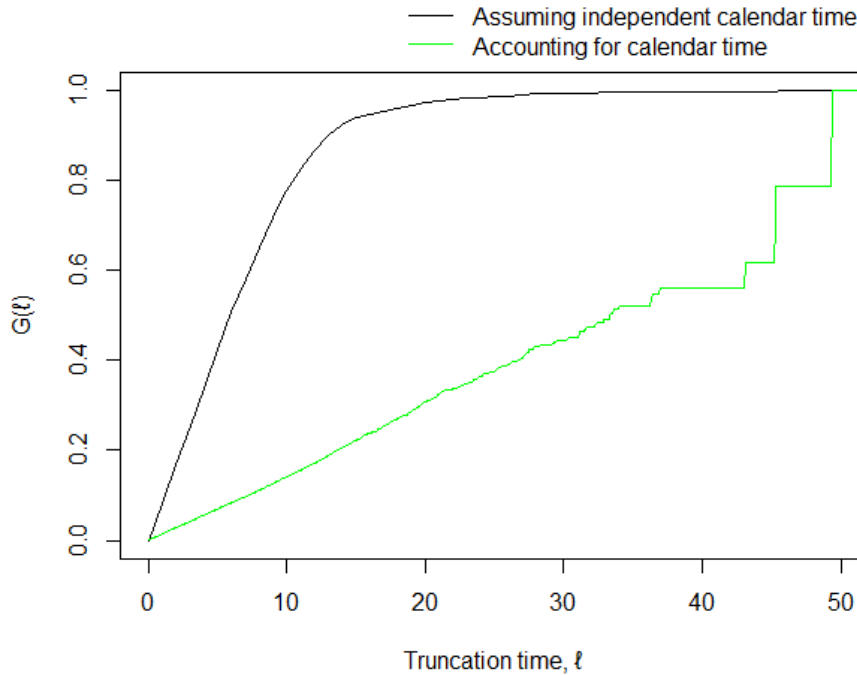


Figure A.2: Simulation of truncation distribution in the one outcome event setting.

Figure A.3 demonstrates the different cumulative incidence curves that could be estimated if we assume dependent or independent truncation. The black line is the 1-KM (1-Kaplan-Meier) using the incident cohort, which is close to the expected $1 - \exp(-0.05t)$. The red line is the 1-KM under the assumption of independent truncation in the prevalent set, which are patients who were diagnosed with SSc prior to 2000 but were still alive at 2000; and hence some were diagnosed after 1995 and some were diagnosed before 1995. This is representative of the prevalent simulated cohort available to us, but is neither a reflection of current survival as with the incident cohort, or the historic survival (for example those diagnosed between 1940 and 1995), as we are missing some historic patients with worse survival (i.e those where $T < L$).

We can correct for these temporal trends. If we wish to approximate the recent risk based on the prevalent data, we can apply the coefficients from the Cox model

in the Breslow estimator to find the baseline cumulative incidence using Equation (9.3). We apply this to our prevalent cohort only, and the estimation is shown in green in the plot, and is comparable to the incident dataset, the more recent cumulative incidence.

We also have the option of finding what we term the historic risk. Patients diagnosed with SSc prior to 1995 are weighted higher, due to more of these cases being left truncated. Due to weighting we observe an increased risk due to the inclusion of historic missing patients who have worse survival. The estimated historic risk is shown as the blue curve. If we were to only include patients prior to 1995, we would expect the cumulative incidence to be equivalent to $1 - \exp(-0.2t)$, therefore we might expect what we term the historic risk to be close to this. They will not be exact, as the historic risk is finding the average risk of all patients diagnosed between 1950 and 2000, and therefore includes patients diagnosed after 1995 who will have lower risk. The hypothetical risk of those diagnosed prior to 1995 is shown as the grey curve.

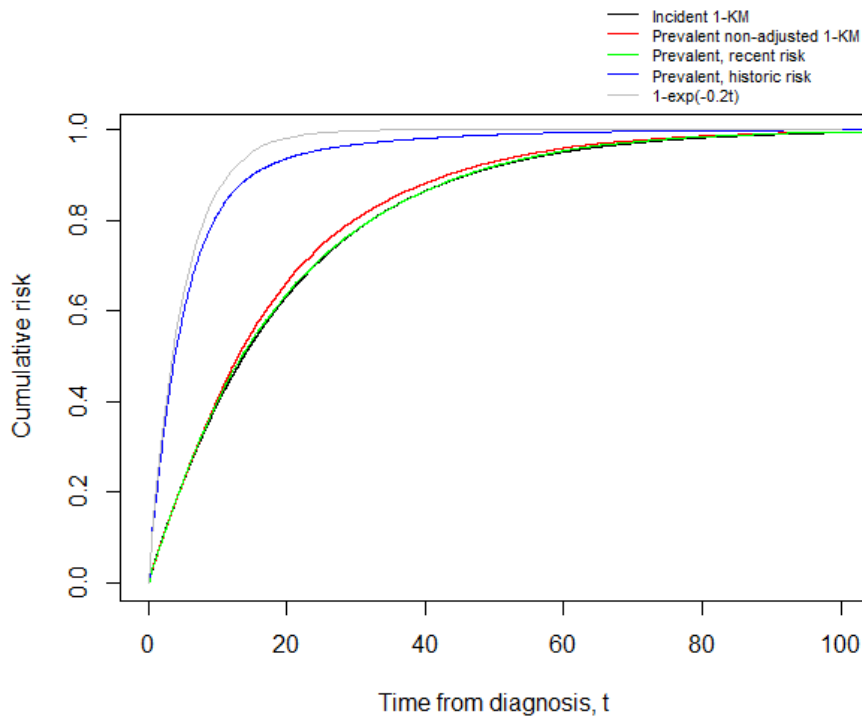


Figure A.3: Simulation of survival given dependent left truncation in the one event setting. The black is the simulated 1-KM curve of the incident data. The red is the same for the prevalent data using the left truncated Kaplan-Meier estimator. The green is the dependent-adjusted recent risk estimator. The blue is the dependent-adjusted historic risk. The grey is the theoretical cumulative incidence of the patients with the longest truncation.

This simulation is only one version of possible simulations and can vary depending on the constraints and rates chosen. We have chosen to use truncation time as a categorical variable, however as it is unlikely that survival changes so suddenly at one point in time, we would expect the change to be more gradual. In this study, we have the benefit of being able to simulate a large number of patients, however we are unsure of the success of this method when there is a smaller sample size, and this factor could potentially lead to bias (especially in large truncation times). We have not included loss to follow-up, which may make the Cox model estimates less precise due to less frequent events.

A.2.2 Simulation 2: Competing risks

We repeat a similar simulation to above, but we now include another event type, which to be consistent with our study we shall term ‘cancer’, with the previous risk being ‘death without cancer’. This is to demonstrate the differences in the estimators depending on whether the direct effect or total effect is being estimated. Here, we decrease the risk of cancer if the patient is diagnosed with SSc after 1998. Therefore, with Y as the event of interest and D as the competing risk of death without cancer, the risks are distributed $Y_{SSc < 1998} \sim Exp(0.1)$ and $Y_{SSc > 1998} \sim Exp(0.025)$, and death is distributed $D_{SSc < 1995} \sim Exp(0.2)$ and $D_{SSc > 1995} \sim Exp(0.05)$.

The two following Cox tables, Table A.2 and Table A.3, show the average hazard ratio of 500 simulations of 500,000 patients with pointwise confidence intervals.

Calendar year of SSc diagnosis	coef	HR = exp(coef)	95% CI
<1990	4.02	[2.73, 5.64]	
1990-1995	4.02	[3.37, 4.79]	
1995-1998	4.00	[3.78, 4.21]	
1998-2000	1.00	[0.94, 1.08]	

Table A.2: Cox model for simulation of the hazard ratios for cancer, Y , with calendar year of SSc diagnosis compared to baseline for year 2000+ .

Calendar year of SSc diagnosis	coef	HR = exp(coef)	95% CI
<1990	4.03	[3.18, 5.01]	
1990-1995	4.01	[3.50, 4.50]	
1995-1998	1.00	[0.92, 1.08]	
1998-2000	1.00	[0.95, 1.05]	

Table A.3: Cox model for simulation of the hazard ratios for death without cancer, D , with calendar year of SSc diagnosis compared to baseline for year 2000+ .

Despite having the same number of simulated patients (500,000) and the same number of simulations (500) as the one outcome simulation above, the possibility of an extra event results in weak confidence at the larger calendar times due to the fewer number having an event of interest. However the averages are close to

the hazard ratios we expect.

We shall now apply the different cumulative incidence estimators to one simulation of an original 500,000 patients. There are two ways to define risk: direct and total, therefore we first show the naive Kaplan-Meier (direct risk) estimation under the elimination of competing events and then the competing risk setting (total effect).

Direct effect

We demonstrate the application to both event Y (left) and D (right) in Figure A.4, where the left is the cancer risk and the right is death without cancer risk, both treating the competing risk as censoring. The black line is the non-adjusted one minus Kaplan-Meier of the incident cohort, and the red line the non-adjusted one minus Kaplan-Meier for the prevalent cohort, and both demonstrate that the inclusion of historic cases with worse survival will show an increased risk compared to those patients who have been diagnosed more recently. We restrict our dataset to prevalent patients with maximum recruitment time of 40 years to reduce noise. The green line is the application of the weighted Breslow estimator to the prevalent cohort, so for event Y this is $1 - \exp\left(-\sum_{s \leq t} \frac{\Delta N_{0Y}(s)}{\sum_{k=1}^n Y_k(s) \exp(\beta_{0Y} \ell_k)}\right)$. Both the cancer and death estimations show an improvement with the black and green line overlapping, however uncertainty may be present at later times, as shown by the simulation in the Cox model. The blue line is the dependent weighted marginal cumulative incidence, estimated using Equation (9.10), with inverse probability weighting based on the probability of surviving both cancer and death. This will estimate a cohort of SSc diagnosed at times ranging from to 1950 to 2000, therefore we expect the more severe historical cases to be weighted higher and be close to the historic hazard. The grey lines plot $1 - \exp(-0.1t)$ for cancer and $1 - \exp(-0.2t)$ for death, the cumulative incidence expected in the historic cases. We see the marginal curves (historic risk) close to these exponential curves for historic cases, but we would expect them to show some distortion due to the inclusion of recent, healthier cases under study as well.

Note that we have not included informative censoring, which is the disadvantage of the methods shown here compared to the proposed g-formula estimator or other

estimators that adjust for informative censoring.

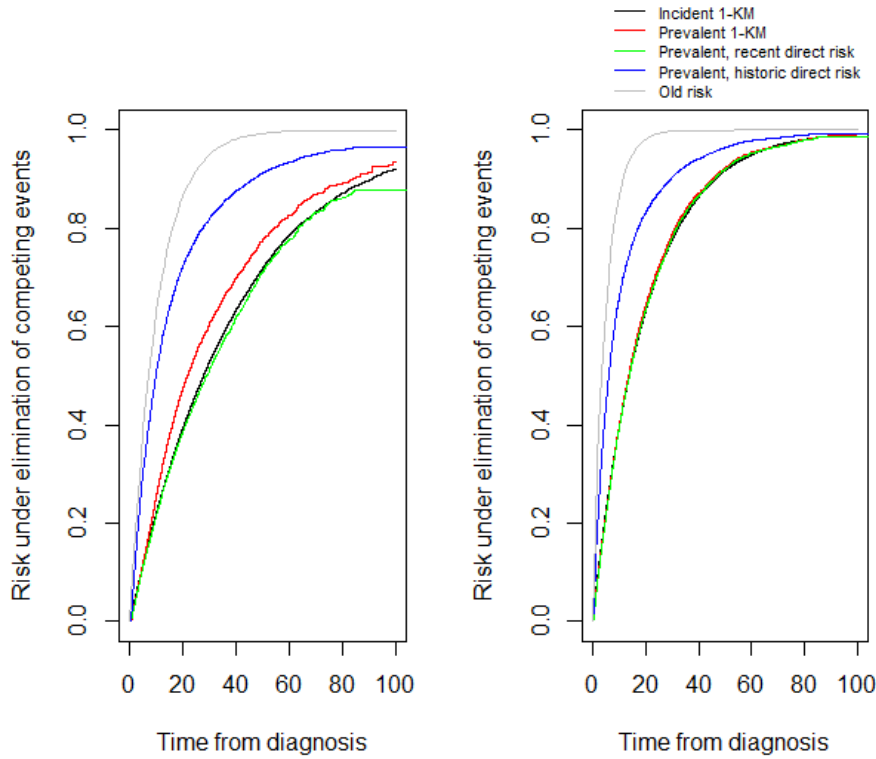


Figure A.4: Simulation of survival given dependent left truncation in the two event setting under the assumption of the elimination of the competing event (*i.e* naive Kaplan-Meier). Left: Cancer, Right: Death without cancer. The black is the simulated 1-KM curve of the incident data. The red is the same for the prevalent data using the left truncated Kaplan-Meier estimator. The green is the dependent-adjusted recent risk estimator. The blue is the dependent-adjusted historic risk. The grey is the theoretical cumulative incidence of the patients with the longer truncation.

Total effect

Figure A.5 demonstrates the application of the estimation of the truncation distribution in the multiple event setting, with survival weightings based on both surviving cancer and death (Equation (9.8)). In our analysis we can observe that, similar to the one event case, there are changes in gradient in the non-adjusted truncation estimations at times 2 and 5, and the results also show that we do not have a uniform cumulative incidence. The estimator for conditionally independent truncation is more uniform, however again it less uniform at larger

ℓ times due to there being few patients with large truncation times.

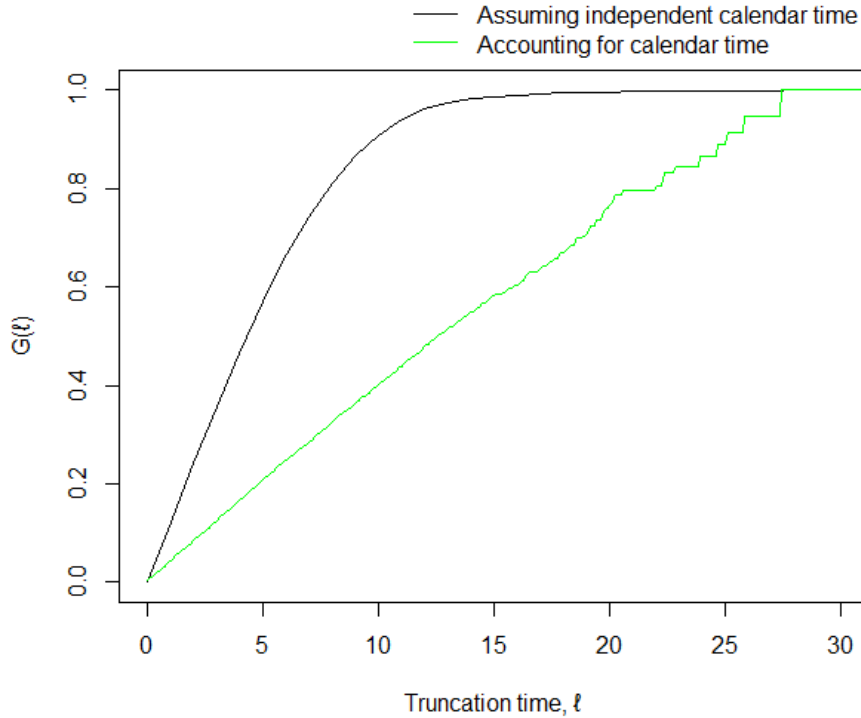


Figure A.5: Simulation of truncation distribution. The black line is the independent distribution assuming left truncation distribution (as shown in Chapter 5), and the green is the dependent-adjusted left truncation distribution.

We remind ourselves of the theory of competing risks. Let the hazard of event Y be $\lambda_Y(t)$ and the hazard of the competing event $\lambda_D(t)$, then the cumulative incidence function for event Y is

$$F_Y(t) = \int_0^t \lambda_Y(u) \exp\left(-\int_0^u (\lambda_Y(s) + \lambda_D(s)) ds\right) du$$

Therefore we expect the incident/baseline cumulative incidence to follow $F_Y(t) = \frac{1}{3} \exp(-0.75t)$ and $F_Y(t) = \frac{2}{3} \exp(-0.75t)$. The cumulative incidence for historic cases (pre-1995) would be $F_Y(t) = \frac{1}{3} \exp(-0.3t)$ and $F_Y(t) = \frac{2}{3} \exp(-0.3t)$.

Figure A.6 shows the estimations in a competing risk setting, with cancer on the left and death on the right. The black lines are the incident cohort using

the cause-specific Aalen-Johansen estimator under the assumption of independent truncation, and is close to the expected $F_Y(t) = \frac{1}{3} \exp(-0.75t)$ and $F_D(t) = \frac{2}{3} \exp(-0.75t)$, respectively. The red lines are the equivalent for the prevalent set. There are very large differences here between the plots for cancer and death without cancer when the prevalent is compared to the incident, with prevalent cancer being a large overestimation of risk and prevalent death being a large underestimation of risk compared to the expected recent risk. Using the Breslow estimator (green lines) for estimating the baseline risk from the prevalent data we observe a close fit to both the cancer and death incident cohort estimations, which is what we expect. The blue lines are the estimator using the historic risk, Equation (9.9). We additionally add the expected distribution for the historic cases, $F_Y(t) = \frac{1}{3} \exp(-0.3t)$ and $F_D(t) = \frac{2}{3} \exp(-0.3t)$ for cancer and death respectively (the grey lines). As in the one case system, we do not expect these to be the same due to the inclusion of recent cases, however there are similarities.

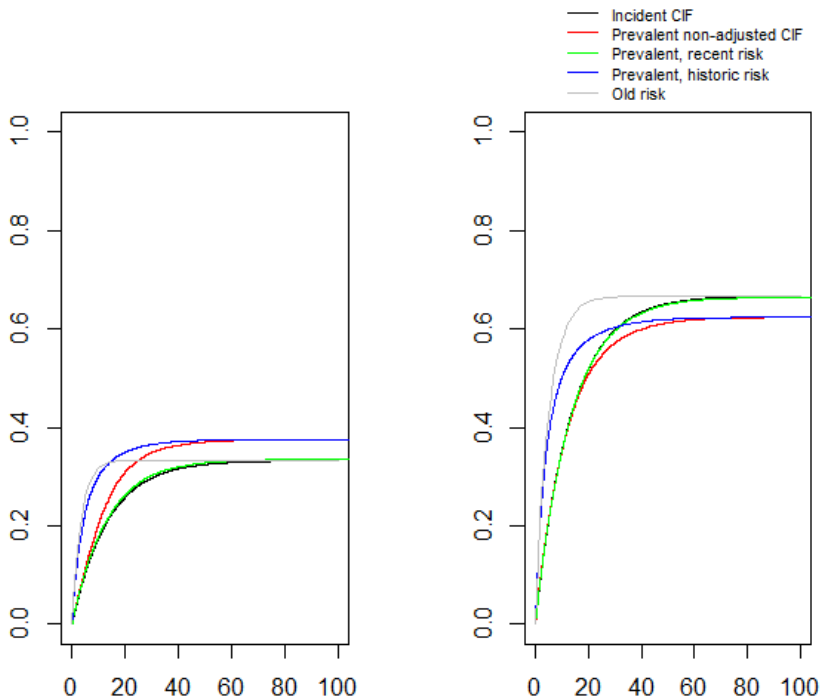


Figure A.6: Simulation of survival given dependent truncation

A.3 ISAC Application form

ISAC APPLICATION FORM PROTOCOLS FOR RESEARCH USING THE CLINICAL PRACTICE RESEARCH DATALINK (CPRD)

For ISAC use only																						
Protocol No.	IMPORTANT Please refer to the guidance for 'Completing the ISAC application form' found on the CPRD website (www.cprd.com/isac). If you have any queries, please contact the ISAC Secretariat at isac@cpdr.com .																				
Submission date (DD/MM/YYYY)																					
SECTION A: GENERAL INFORMATION ABOUT THE PROPOSED RESEARCH STUDY																						
1. Study Title[§] (Please state the study title below) Epidemiology and outcomes in patients with systemic sclerosis in the UK <small>[§]Please note: This information will be published on the CPRD's website as part of its transparency policy.</small>																						
2. Has any part of this research proposal or a related proposal been previously submitted to ISAC? Yes* <input type="checkbox"/> No <input checked="" type="checkbox"/> X <small>*If yes, please provide the previous protocol number/s below. Please also state in your current submission how this/these are related or relevant to this study.</small>																						
3. Has this protocol been peer reviewed by another Committee? (e.g. grant award or ethics committee) Yes* <input type="checkbox"/> No <input checked="" type="checkbox"/> X <small>*If Yes, please state the name of the reviewing Committee(s) below and provide an outline of the review process and outcome as an Appendix to this protocol :</small>																						
4. Type of Study (please tick all the relevant boxes which apply) <table border="0"> <tr> <td>Adverse Drug Reaction/Drug Safety</td> <td><input type="checkbox"/></td> <td>Drug Effectiveness</td> <td><input type="checkbox"/></td> </tr> <tr> <td>Drug Utilisation</td> <td><input checked="" type="checkbox"/></td> <td>Pharmacoeconomics</td> <td><input type="checkbox"/></td> </tr> <tr> <td>Disease Epidemiology</td> <td><input checked="" type="checkbox"/></td> <td>Post-authorisation Safety</td> <td><input type="checkbox"/></td> </tr> <tr> <td>Health care resource utilisation</td> <td><input type="checkbox"/></td> <td>Methodological Research</td> <td><input type="checkbox"/></td> </tr> <tr> <td>Health/Public Health Services Research</td> <td><input type="checkbox"/></td> <td>Other*</td> <td><input type="checkbox"/></td> </tr> </table> <small>*If Other, please specify the type of study in the lay summary</small>			Adverse Drug Reaction/Drug Safety	<input type="checkbox"/>	Drug Effectiveness	<input type="checkbox"/>	Drug Utilisation	<input checked="" type="checkbox"/>	Pharmacoeconomics	<input type="checkbox"/>	Disease Epidemiology	<input checked="" type="checkbox"/>	Post-authorisation Safety	<input type="checkbox"/>	Health care resource utilisation	<input type="checkbox"/>	Methodological Research	<input type="checkbox"/>	Health/Public Health Services Research	<input type="checkbox"/>	Other*	<input type="checkbox"/>
Adverse Drug Reaction/Drug Safety	<input type="checkbox"/>	Drug Effectiveness	<input type="checkbox"/>																			
Drug Utilisation	<input checked="" type="checkbox"/>	Pharmacoeconomics	<input type="checkbox"/>																			
Disease Epidemiology	<input checked="" type="checkbox"/>	Post-authorisation Safety	<input type="checkbox"/>																			
Health care resource utilisation	<input type="checkbox"/>	Methodological Research	<input type="checkbox"/>																			
Health/Public Health Services Research	<input type="checkbox"/>	Other*	<input type="checkbox"/>																			
5. Health Outcomes to be Measured[§] <small>[§]Please note: This information will be published on CPRD's website as part of its transparency policy.</small> <u>Please summarise below the primary/secondary health outcomes to be measured in this research protocol:</u> <table border="0"> <tr> <td>• Cancer</td> <td>• Peripheral vascular disease</td> <td>•</td> </tr> <tr> <td>• Myocardial infarction</td> <td>• Cerebrovascular disease</td> <td>•</td> </tr> <tr> <td>• Pulmonary hypertension</td> <td>• All-cause mortality</td> <td>•</td> </tr> </table> <small>[Please add more bullet points as necessary]</small>			• Cancer	• Peripheral vascular disease	•	• Myocardial infarction	• Cerebrovascular disease	•	• Pulmonary hypertension	• All-cause mortality	•											
• Cancer	• Peripheral vascular disease	•																				
• Myocardial infarction	• Cerebrovascular disease	•																				
• Pulmonary hypertension	• All-cause mortality	•																				

08 August 2016 Version 1.0

6. Publication: This study is intended for (please tick all the relevant boxes which apply):	
Publication in peer-reviewed journals <input checked="" type="checkbox"/>	Presentation at scientific conference <input checked="" type="checkbox"/>
Presentation at company/institutional meetings <input checked="" type="checkbox"/>	Regulatory purposes <input type="checkbox"/>
Other* <input type="checkbox"/>	
<i>*If Other, please provide further information:</i>	
SECTION B: INFORMATION ON INVESTIGATORS AND COLLABORATORS	
7. Chief Investigator[§]	
Please state the full name, job title, organisation name & e-mail address for correspondence - see guidance notes for eligibility. Please note that there can only be one Chief Investigator per protocol.	
Anita McGrogan Senior lecturer in pharmacoepidemiology and statistics Department of Pharmacy and Pharmacology University of Bath Bath BA2 7AY a.mcrogan@bath.ac.uk	
[§] Please note: The name and organisation of the Chief Investigator and will be published on CPRD's website as part of its transparency policy	
CV has been previously submitted to ISAC <input type="checkbox"/>	CV number:
A new CV is being submitted with this protocol <input checked="" type="checkbox"/>	
An updated CV is being submitted with this protocol <input type="checkbox"/>	
8. Affiliation of Chief Investigator (full address)	
Department of Pharmacy and Pharmacology, University of Bath, Bath, BA2 7AY, UK	
9. Corresponding Applicant[§]	
Please state the full name, affiliation(s) and e-mail address below: Anita McGrogan, Department of Pharmacy and Pharmacology, University of Bath, Bath, BA2 7AY, UK; a.mcrogan@bath.ac.uk	
[§] Please note: The name and organisation of the corresponding applicant and their organisation name will be published on CPRD's website as part of its transparency policy	
Same as chief investigator <input checked="" type="checkbox"/>	CV number:
CV has been previously submitted to ISAC <input type="checkbox"/>	
A new CV is being submitted with this protocol <input checked="" type="checkbox"/>	
An updated CV is being submitted with this protocol <input type="checkbox"/>	
10. List of all investigators/collaborators[§]	
Please list the full name, affiliation(s) and e-mail address* of all collaborators, other than the Chief Investigator below:	
[§] Please note: The name of all investigators and their organisations/institutions will be published on CPRD's website as part of its transparency policy	
Other investigator: Professor Neil McHugh Department of Pharmacy and Pharmacology, University of Bath, Bath, BA2 7AY, UK Royal National Hospital for Rheumatic Diseases, Upper Borough Walls, Bath, BA1 1RL, UK. n.j.mchugh@bath.ac.uk	
CV has been previously submitted to ISAC <input type="checkbox"/>	CV number: 035_15CE
A new CV is being submitted with this protocol <input type="checkbox"/>	

An updated CV is being submitted with this protocol

Other investigator: Dr John Pauling
 Department of Pharmacy and Pharmacology, University of Bath, Bath, BA2 7AY, UK
 Royal National Hospital for Rheumatic Diseases, Upper Borough Walls, Bath, BA1 1RL, UK.
j.d.pauling@bath.ac.uk

CV has been previously submitted to ISAC **CV number:**
 A new CV is being submitted with this protocol
 An updated CV is being submitted with this protocol

Other investigator: Dr Alison Nightingale
 Department of Pharmacy and Pharmacology, University of Bath, Bath, BA2 7AY, UK
a.nightingale@bath.ac.uk

CV has been previously submitted to ISAC **CV number:** 033_15CES
 A new CV is being submitted with this protocol
 An updated CV is being submitted with this protocol

Other investigator: Mrs Julia Snowball
 Department of Pharmacy and Pharmacology, University of Bath, Bath, BA2 7AY, UK
j.snowball@bath.ac.uk

CV has been previously submitted to ISAC **CV number:**
 A new CV is being submitted with this protocol
 An updated CV is being submitted with this protocol

[Please add more investigators as necessary]

Please note that your ISAC application form and protocol **must be copied to all e-mail addresses listed above at the time of submission of your application to the ISAC mailbox. Failure to do so will result in delays in the processing of your application.*

11. Conflict of interest statement*
 Please provide a draft of the conflict (or competing) of interest (COI) statement that you intend to include in any publication which might result from this work

All investigators have confirmed that they have no conflicts of interest to declare.

**Please refer to the International Committee of Medical Journal Editors (ICMJE) for guidance on what constitutes a COI.*

12. Experience/expertise available
 Please complete the following questions to indicate the experience/ expertise available within the team of investigators/collaborators actively involved in the proposed research, including the analysis of data and interpretation of results.

	Previous GPRD/CPRD Studies	Publications using GPRD/CPRD data
None	<input type="checkbox"/>	<input type="checkbox"/>
1-3	<input type="checkbox"/>	<input type="checkbox"/>
> 3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Experience/Expertise available	Yes	No
Is statistical expertise available within the research team? <i>If yes, please indicate the name(s) of the relevant investigator(s)</i> Dr Anita McGrogan	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Is experience of handling large data sets (>1 million records) available within the research team? <i>If yes, please indicate the name(s) of the relevant investigator(s)</i> Mrs Julia Snowball, Dr Alison Nightingale, Dr Anita McGrogan	<input checked="" type="checkbox"/>	<input type="checkbox"/>

<p>Is experience of practising in UK primary care available to or within the research team? <i>If yes, please indicate the name(s) of the relevant investigator(s)</i> Dr Linda McHugh and links also in place with a local practice (Oldfield Park) managed by Helen Harris. Two the investigators (Professor Neil McHugh and Dr John Pauling) have undertaken medical training thus gaining a detailed knowledge of primary care practice in the UK.</p>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<p>13. References relating to your study Please list up to 3 references (most relevant) relating to your proposed study:</p> <p>Mayes MD, Lacey JV, Jr., Beebe-Dimmer J, et al. Prevalence, incidence, survival, and disease characteristics of systemic sclerosis in a large US population. <i>Arthritis Rheum</i> 2003; 48:2246-55</p> <p>Au K, Singh MK, Bodukam V, et al. Atherosclerosis in systemic sclerosis: a systematic review and meta-analysis. <i>Arthritis Rheum</i> 2011; 63:2078-90</p> <p>Chaisson NF, Hassoun PM. Systemic sclerosis-associated pulmonary arterial hypertension. <i>Chest</i> 2013; 144:1346-56</p>		
<p>SECTION C: ACCESS TO THE DATA</p>		
<p>14. Financial Sponsor of study[§] [§]Please note: The name of the source of funding will be published on CPRD's website as part of its transparency policy</p> <p>Pharmaceutical Industry <input type="checkbox"/> <i>Please specify name and country:</i> Academia <input type="checkbox"/> <i>Please specify name and country:</i> Government / NHS <input type="checkbox"/> <i>Please specify name and country:</i> Charity <input checked="" type="checkbox"/> <i>Please specify name and country:</i> Scleroderma and Raynaud's UK; Bath Institute of Rheumatic Diseases; both UK charities. Other <input type="checkbox"/> <i>Please specify name and country:</i> None <input type="checkbox"/></p>		
<p>15. Type of Institution conducting the research</p> <p>Pharmaceutical Industry <input type="checkbox"/> <i>Please specify name and country:</i> Academia <input checked="" type="checkbox"/> <i>Please specify name and country:</i> University of Bath, UK Government Department <input type="checkbox"/> <i>Please specify name and country:</i> Research Service Provider <input type="checkbox"/> <i>Please specify name and country:</i> NHS <input type="checkbox"/> <i>Please specify name and country:</i> Other <input type="checkbox"/> <i>Please specify name and country:</i></p>		
<p>16. Data access arrangements</p> <p>The financial sponsor/ collaborator* has a licence for CPRD GOLD and will extract the data <input type="checkbox"/> The institution carrying out the analysis has a licence for CPRD GOLD and will extract the data** <input type="checkbox"/> A data set will be provided by the CPRD*[€] - <input checked="" type="checkbox"/> CPRD has been commissioned to extract the data <u>and</u> perform the analyses[€] <input type="checkbox"/> Other: <input type="checkbox"/> <i>If Other, please specify:</i></p> <p><small>*Collaborators supplying data for this study must be named on the protocol as co-applicants. **If data sources other than CPRD GOLD are required, these will be supplied by CPRD [€]Please note that datasets provided by CPRD are limited in size; applicants should contact CPRD (kc@cprd.com) if a dataset of >300,000 patients is required. [€]Investigators must discuss their request with a member of the CPRD Research team before submitting an ISAC application. Please contact the CPRD Research Team on +44 (20) 3080 6383 or email (kc@cprd.com) to discuss your requirements. Please also state the name of CPRD Research team with whom you have discussed this request (provide the date of discussion and any relevant reference information).</small></p>		

Name of CPRD Researcher	Reference number (where available)	Date of contact
17. Primary care data		
Please specify which primary care data set(s) are required		
Vision only (Default for CPRD studies)	<input type="checkbox"/>	Both Vision and EMIS®* <input checked="" type="checkbox"/>
EMIS® only*	<input type="checkbox"/>	
<p><i>Note: Vision and EMIS are different practice management systems. CPRD has traditionally collected data from Vision practice. Data collected from EMIS is currently under evaluation prior to wider release.</i></p> <p><i>*Investigators requiring the use of EMIS data must discuss the study with a member of the CPRD Research team before submitting an ISAC application</i></p>		
Please state the name of the CPRD Researcher with whom you have discussed your request for EMIS data:		
Name of CPRD Researcher	Daniel Dedman	Reference number (where available) OCR8454
Date of contact	1/11/2016	
SECTION D: INFORMATION ON DATA LINKAGES		
18. Does this protocol seek access to linked data		
Yes* <input checked="" type="checkbox"/>	No <input type="checkbox"/>	If No, please move to section E.
<p><i>*Research groups which have not previously accessed CPRD linked data resources must discuss access to these resources with a member of the CPRD Research team, before submitting an ISAC application. Investigators requiring access to HES Accident and Emergency data, HES Diagnostic Imaging Dataset and PROMS data must also discuss this with a member of the CPRD Research team before submitting an ISAC application. Please contact the CPRD Research Team on +44 (20) 3080 6383 or email kc@cprd.com to discuss your requirements before submitting your application.</i></p>		
Please state the name of the CPRD Researcher with whom you have discussed your linkage request.		
Name of CPRD Researcher	Dipesh Patel and Jessie Oyinlola	Reference number (where available)
NCQ9158	Date of contact	9/2/2017
<p><i>Please note that as part of the ISAC review of linkages, your protocol may be shared - in confidence - with a representative of the requested linked data set(s) and summary details may be shared - in confidence - with the Confidentiality Advisory Group of the Health Research Authority.</i></p>		
19. Please select the source(s) of linked data being requested[§]		
[§] Please note: This information will be published on the CPRD's website as part of its transparency policy.		
<input type="checkbox"/> ONS Death Registration Data	<input type="checkbox"/> MINAP (Myocardial Ischaemia National Audit Project)	
<input checked="" type="checkbox"/> HES Admitted Patient Care	<input checked="" type="checkbox"/> Cancer Registration Data*	
<input checked="" type="checkbox"/> HES Outpatient	<input type="checkbox"/> PROMS (Patient Reported Outcomes Measure)**	
<input type="checkbox"/> HES Accident and Emergency	<input type="checkbox"/> CPRD Mother Baby Link	
<input type="checkbox"/> HES Diagnostic Imaging Dataset		
<input type="checkbox"/> Practice Level Index of Multiple Deprivation (Standard)		
<input type="checkbox"/> Practice Level Index of Multiple Deprivation (Bespoke)		
<input type="checkbox"/> Patient Level Index of Multiple Deprivation***		
<input type="checkbox"/> Patient Level Townsend Score ***		
<input type="checkbox"/> Other**** Please specify:		
<p><i>*Applicants seeking access to cancer registration data must complete a Cancer Dataset Agreement form (available from CPRD). This should be submitted to the ISAC as an appendix to your protocol. Please also note that applicants seeking access to cancer EMIS must provide consent for publication of their study title and study institution on the UK Cancer Registry website.</i></p> <p><i>**Assessment of the quality of care delivered to NHS patients in England undergoing four procedures: hip replacement, knee replacement, groin hernia and varicose veins. Please note that patient level PROMS data are only accessible by academics</i></p> <p><i>*** Patient level IMD and Townsend scores will not be supplied for the same study</i></p> <p><i>****If "Other" is specified, please provide the name of the individual in the CPRD Research team with whom this linkage has been discussed.</i></p>		
Name of CPRD Researcher	Dipesh Patel and Jessie Oyinlola	Reference number (where available)
OCR9251	Date of contact	16/2/2017

<p>20. Total number of linked datasets requested <u>including</u> CPRD GOLD</p> <p>Number of linked datasets requested (<i>practice/ 'patient' level Index of Multiple Deprivation, Townsend Score or the CPRD Mother Baby Link should <u>not</u> be included in this count</i>) 4</p> <p><i>Please note: Where ≥5 linked datasets are requested, approval may be required from the Confidentiality Advisory Group (CAG) to access these data</i></p>												
<p>21. Is linkage to a <u>local</u>[¶] dataset with <1 million patients being requested?</p> <p>Yes* <input type="checkbox"/> No <input checked="" type="checkbox"/></p> <p><i>*If yes, please provide further details:</i> [¶] Data from defined geographical areas i.e. non-national datasets.</p>												
<p>22. If you have requested one or more linked data sets, please indicate whether the Chief Investigator or any of the collaborators listed in question 5 above, have access to these data in a patient identifiable form (e.g. full date of birth, NHS number, patient post code), or associated with an identifiable patient index.</p> <p>Yes* <input type="checkbox"/> No <input checked="" type="checkbox"/></p> <p><i>* If yes, please provide further details:</i></p>												
<p>23. Does this study involve linking to patient <i>identifiable</i> data (e.g. hold date of birth, NHS number, patient post code) from other sources?</p> <p>Yes <input type="checkbox"/> No <input checked="" type="checkbox"/></p>												
SECTION E: VALIDATION/VERIFICATION												
<p>24. Does this protocol describe a purely observational study using CPRD data?</p> <p>Yes* <input checked="" type="checkbox"/> No** <input type="checkbox"/></p> <p><i>* Yes: If you will be using data obtained from the CPRD Group, this study does not require separate ethics approval from an NHS Research Ethics Committee. ** No: You may need to seek separate ethics approval from an NHS Research Ethics Committee for this study. The ISAC will provide advice on whether this may be needed.</i></p>												
<p>25. Does this protocol involve requesting any additional information from GPs?</p> <p>Yes* <input type="checkbox"/> No <input checked="" type="checkbox"/></p> <p><i>* If yes, please indicate what will be required:</i></p> <table style="width: 100%; border: none;"> <tr> <td style="width: 60%;">Completion of questionnaires by the GP^v</td> <td style="width: 10%;">Yes <input type="checkbox"/></td> <td style="width: 10%;">No <input type="checkbox"/></td> <td style="width: 10%;"></td> </tr> <tr> <td> Is the questionnaire a validated instrument?</td> <td>Yes <input type="checkbox"/></td> <td>No <input type="checkbox"/></td> <td></td> </tr> <tr> <td> If yes, has permission been obtained to use the instrument?</td> <td>Yes <input type="checkbox"/></td> <td>No <input type="checkbox"/></td> <td></td> </tr> </table> <p>Please provide further information:</p> <p>Other (please describe)</p>	Completion of questionnaires by the GP ^v	Yes <input type="checkbox"/>	No <input type="checkbox"/>		Is the questionnaire a validated instrument?	Yes <input type="checkbox"/>	No <input type="checkbox"/>		If yes, has permission been obtained to use the instrument?	Yes <input type="checkbox"/>	No <input type="checkbox"/>	
Completion of questionnaires by the GP ^v	Yes <input type="checkbox"/>	No <input type="checkbox"/>										
Is the questionnaire a validated instrument?	Yes <input type="checkbox"/>	No <input type="checkbox"/>										
If yes, has permission been obtained to use the instrument?	Yes <input type="checkbox"/>	No <input type="checkbox"/>										

Any questionnaire for completion by GPs or other health care professional must be approved by ISAC before circulation for completion.

26. Does this study require contact with patients in order for them to complete a questionnaire?

Yes* No

**Please note that any questionnaire for completion by patients must be approved by ISAC before circulation for completion.*

27. Does this study require contact with patients in order to collect a sample?

Yes* No

** Please state what will be collected:*

SECTION F: DECLARATION

28. Signature from the Chief Investigator

- I have read the guidance on '**Completion of the ISAC application form**' and '**Contents of CPRD ISAC Research Protocols**' and have understood these;
- I have read the submitted version of this research protocol, including all supporting documents, and confirm that these are accurate.
- I am suitably qualified and experienced to perform and/or supervise the research study proposed.
- I agree to conduct or supervise the study described in accordance with the relevant, current protocol
- I agree to abide by all ethical, legal and scientific guidelines that relate to access and use of CPRD data for research
- I understand that the details provided in sections marked with (8) in the application form and protocol will be published on the CPRD website in line with CPRD's transparency policy.
- I agree to inform the CPRD of the final outcome of the research study: publication, prolonged delay, completion or termination of the study.

Name: Anita McGrogan Date: 14/2/2017 e-Signature (type name): Anita McGrogan

PROTOCOL INFORMATION REQUIRED

The following sections below **must** be included in the CPRD ISAC research protocol. Please refer to the guidance on '**Contents of CPRD ISAC Research Protocols**' (www.cprd.com/isac) for more information on how to complete the sections below. Pages should be numbered. All abbreviations must be defined on first use.

Applicants must complete all sections listed below Sections which do not apply should be completed as ' <i>Not Applicable</i> '
<p>A. Study Title[§] [§]<i>Please note: This information will be published on CPRD's website as part of its transparency policy</i> Epidemiology and outcomes in patients with systemic sclerosis in the UK</p>
<p>B. Lay Summary (Max. 200 words)[§] [§]<i>Please note: This information will be published on CPRD's website as part of its transparency policy</i></p> <p>Systemic sclerosis (SSc) is a disease which in its more serious form, affects the skin and internal organs through the immune system and the connective tissue. While some work has been done to estimate the rate of new cases and number of existing cases of this disease in different areas of the UK, these estimates are lower than those found in Europe and the US. This may suggest underreporting or a lower disease burden in the UK.</p> <p>A diagnosis of SSc may predispose patients to developing other diseases: greater knowledge about this is needed to inform patient care. For example, the burden of cancer may be greater in patients with SSc, but few studies have fully investigated this link. It has been found that some cancers (particularly breast) occur before a diagnosis of SSc whereas other cancers (oropharyngeal, oesophageal, lung) occur after SSc diagnosis: this relationship needs to be looked at in more detail. Increased rates of other diseases including pulmonary hypertension, cerebrovascular disease, myocardial infarction and vascular disease must be considered when evaluating mortality. This study will use national healthcare databases to describe the disease burden of SSc and investigate potential associations between SSc and other serious diagnoses.</p>
<p>C. Technical Summary (Max. 200 words)[§] [§]<i>Please note: This information will be published on CPRD's website as part of its transparency policy</i></p> <p>This study will investigate the incidence and prevalence of systemic sclerosis (SSc) in the UK. Cohort studies will be used to investigate serious outcomes: people with SSc will be matched to those who do not have SSc by age, sex, GP practice and calendar time will be accounted for. The outcomes investigated in separate studies will be cancer, myocardial infarction, cerebrovascular disease, peripheral vascular disease and pulmonary hypertension. Time to diagnosis of these outcomes and death will be compared using survival analysis. Covariates including smoking, alcohol, BMI, comorbidities and co-prescribing will be adjusted for.</p> <p>The temporal relationship between cancer and SSc is less well understood than in other diseases: in some patients cancer pre-dates the diagnosis of SSc and this could be a factor in the development of SSc. This will be investigated using a case control study where the cases have an incident diagnosis of SSc and the controls do not, cases and controls will be matched as before. The analysis will use logistic regression and backdating of the index date to investigate this further. Where patients have a diagnosis of cancer and SSc at similar time points, inclusion in the analytical studies may not be possible.</p>
<p>D. Objectives, Specific Aims and Rationale</p> <p>Objectives:</p> <ul style="list-style-type: none"> • To evaluate the epidemiology of SSc in the UK • To investigate if there is an association between SSc and cancer and if this exists, the

1

Applicants must complete all sections listed below
Sections which do not apply should be completed as 'Not Applicable'

temporal relationship between diagnoses

- To investigate if there is an association between SSc and other serious outcomes including pulmonary hypertension, cerebrovascular disease, peripheral vascular disease, myocardial infarction

Aims:

- Calculate the age- and sex- specific and standardised incidence rates, annual prevalence and standardised mortality rates for SSc in the UK population.
- Describe the patients who have a diagnosis of cancer and SSc within a very short time interval
- Determine the crude and standardised incidence rates and gender specific incidence rates for cancer, myocardial infarction, pulmonary hypertension, cerebrovascular disease and peripheral vascular disease in people with SSc and compare these to people who do not have SSc
- Identify any differences in incidence of different types of cancer in people with a diagnosis of SSc compared to people who do not have SSc
- Determine if there are differences in time to diagnosis of serious outcomes including cancer, myocardial infarction, cerebrovascular disease, pulmonary hypertension and peripheral vascular disease and time to death in people with SSc compared to people who do not have SSc
- Determine if there are any additional risk factors associated with the occurrence of these serious outcomes in SSc patients compared to people who do not have SSc
- Explore the temporal relationship between cancer followed by a diagnosis of SSc

Rationale:

There is limited information about the incidence and prevalence of SSc in the UK and of the information that has been published, rates appear lower than in other parts of the world. Determining these rates in the CPRD will add to the body of knowledge and through comparison with rates from Europe and the US, will indicate whether this disease is being under-diagnosed / underreported in the UK or if the UK does have lower rates. It is vital to have more accurate data than that which may be available in specialist centres: for example by recognising pulmonary hypertension earlier and treating it more effectively, its impact on mortality could be delayed.

It is thought that there is a link between SSc and cancer: for some patients the diagnosis of cancer occurs before the diagnosis of SSc but for others it is after this diagnosis. Using the CPRD it will be determined if there is an increased risk of cancer in patients with a diagnosis of SSc. Other risk factors for cancer in these patients will be evaluated and this information will be used to inform the management of patients.

Where cancer occurs before the diagnosis of SSc any differences in time to diagnosis of SSc and the type of cancer diagnosed will be identified. Previous work has found differences in the types of cancer diagnosed and it has been hypothesised that there are differences in terms of disease process and the treatments used for cancer.

Some work investigating atherosclerosis in patients with SSc has been undertaken by looking at level of atherosclerosis in patients using measures including carotid intima-medial thickness and flow mediated dilation but very few studies have looked at longer term outcomes related to atherosclerosis. This will also be investigated to give a more complete picture of potential serious outcomes in patients with SSc.

By answering these specific aims, useful information about the impact of SSc in terms of serious

2

Applicants must complete all sections listed below
Sections which do not apply should be completed as 'Not Applicable'

outcomes and the possible risk factors for this disease will be described. This will enable further work in improving the management of patients and earlier diagnoses of comorbidities.

E. Study Background

Systemic sclerosis (SSc) is an autoimmune disease that occurs as either localised disease which mainly affects the skin, usually the hands, arms and face, or diffuse disease that affects a large area of the skin and at least one internal organ.¹ It is the diffuse form of the disease that this study will investigate. Very little is known about the epidemiology of SSc and outcomes that may be associated with this: this study will investigate and quantify these.

Epidemiology of SSc in the UK

A review of the published epidemiological studies of SSc suggests a marked temporal increase in the occurrence of SSc and significant geographical variation worldwide. Reported incidence rates of SSc in the US have risen from 0.6/million/year (1947-52) to 21/million/year (1989-1991).^{2,3} There has been a similar changes in estimates of prevalence rising from 138/million (1950-79) to 276/million (1989-1991). It has been suggested that there is a higher prevalence of SSc in the US compared with Europe, as low estimated prevalence rates have been reported in the UK. A UK study in the northeast identified cases using hospital admission data, local membership records from major UK scleroderma charities, data from local primary care centres and an approach to all relevant clinicians to report all cases of SSc seen over a 12-month period found an estimated prevalence of SSc of 88/million population which remains significantly lower than contemporary reports from the US and Europe.^{2,4,5} No attempts have yet been made to estimate the incidence and prevalence of SSc in the UK using national healthcare databases.

Using the CPRD to evaluate the incidence and prevalence of SSc in the UK will give much needed information about these rates and how the disease is managed in general practice. Investigating the incidence of SSc over time could identify aetiological drivers while having a more accurate assessment of the prevalence of SSc in the UK will have useful implications for service planning and accurate modelling of high-cost drug use.

The relationship between SSc and serious outcomes

An association between SSc and malignancy was first proposed in the 1950s and in recent years has attracted renewed attention. A systematic review of cohort studies reporting SSc mortality identified cancer as a contributing factor to mortality in up to 30% of patients.⁶ A large retrospective study of 2,177 patients with SSc from the Royal Free Hospital identified a history of cancer in 7.1% of patients.⁷ The most commonly occurring cancers were breast (42.2%) followed by haematological (12.3%), gastrointestinal (11.0%) and gynaecological (11.0%) cancers.⁷ Breast carcinoma has repeatedly emerged as the most commonly occurring cancer in SSc.⁷⁻⁹ A number of studies⁸⁻¹¹ have investigated this potential association but with mixed results that could be due to different methods of case ascertainment and comparators from different populations used. No longitudinal population-based databases have simultaneously assessed the burden of cancer-related mortality in SSc, the temporal relationship between cancer and SSc and additional risk factors influencing specific cancer occurrence in SSc, using the same data source.

It is important to consider other serious outcomes that are likely to have a higher prevalence in people with SSc and that may contribute to higher mortality rates. Endothelial damage which occurs early in patients with SSc can trigger vasculopathy the effect of which was thought, until recently, to be mainly on the capillaries and small arteries. Some research has suggested that

Applicants must complete all sections listed below
Sections which do not apply should be completed as 'Not Applicable'

patients with SSc have increased atherosclerosis compared to people who do not have SSc which can lead to serious outcomes including cerebrovascular disease, myocardial infarction and peripheral vascular disease. Evidence indicating involvement of macrovascular vessels includes work done to review flow mediated dilation and carotid intima-medial thickness in patients¹² with limited population based studies evaluating outcomes including cardiovascular disease and mortality¹³. Given the paucity of data available combined with the likely beneficial impact to patients of identifying cardiovascular disease early, outcomes of myocardial infarction, cerebrovascular disease and peripheral vascular disease will also be investigated in this study.

While pulmonary hypertension is thought to be a leading cause of death in people with SSc, it can be challenging to identify early signs of this and instigate treatment.¹⁴ Given the relationship with mortality, early treatment is recommended therefore determining rates of diagnosis of pulmonary hypertension in people with SSc, when this is diagnosed in relation to SSc and whether this has changed over time as well as whether certain patient groups are more likely to be diagnosed will provide information about clinical management and patient susceptibility.

F. Study Type

The first part of the study will be descriptive and will describe the population of people with a diagnosis of SSc; this will be followed by hypothesis testing studies to investigate if there is a link between SSc and serious outcomes: cancer, myocardial infarction, cerebrovascular disease, peripheral vascular disease, pulmonary hypertension. ~~In order to overcome the potential difficulties of a small sample size of people with SSc diagnosed, permission to use EMIS data has been agreed therefore this ISAC includes both CPRD and EMIS data.~~

G. Study Design

Descriptive study:

The descriptive study will identify all patients diagnosed with SSc and characteristics of these patients will be described. Incidence, prevalence and mortality rates of patients with SSc will be determined. Where patients are diagnosed with SSc and cancer with the diagnoses occurring very closely in time, these patients will be described: it has been observed in clinical practice that sometimes patients are diagnosed with both diseases simultaneously therefore it is not possible to include these patients in either of the hypothesis testing studies evaluating SSc and cancer given the likely uncertainty of which disease occurred first.

Hypothesis testing studies:

Cohort studies

The hypothesis testing studies will be used to investigate the diagnosis of cancer, myocardial infarction, cerebrovascular disease, peripheral vascular disease and pulmonary hypertension in patients who have had a diagnosis of SSc. As the population with SSc is relatively small and well defined, a cohort study of patients with a diagnosis of SSc matched to patients who do not have a diagnosis of SSc will be used to investigate whether there are any associations with SSc and these outcomes.

A first cohort of people with a diagnosis of SSc and no prior diagnosis of cancer will be identified and matched on age, sex and GP practice to up to six randomly sampled patients who have not had SSc diagnosed or any diagnosis of cancer at the date that they are enrolled into the study. The index date will be defined as the date of entry into the cohort.

A second cohort study of people with a diagnosis of SSc and no prior diagnosis of myocardial infarction, peripheral vascular disease or cerebrovascular disease will be identified. This group will be matched on age, sex and GP practice to up to six randomly sampled patients who have not had SSc diagnosed or any diagnosis of the outcomes being investigated at the date that

Applicants must complete all sections listed below
Sections which do not apply should be completed as 'Not Applicable'

they are enrolled into the study. The index date will be defined as the date of entry into the cohort.

To identify the burden associated with pulmonary hypertension in people with SSc, a third cohort will be used and defined as people with a diagnosis of SSc and no prior diagnosis of pulmonary hypertension; this group will be matched as before on age, sex and GP practice to up to six randomly sampled patients who have not had SSc diagnosed or a diagnosis of pulmonary hypertension at the date that they are enrolled into the study. The index date will be defined as the date of entry into the cohort.

For these cohort studies, patients will leave the study when they are diagnosed with one of the outcomes being studied, or when they have a date of death recorded or when their practice stop contributing research standard data or when the study period ends (31/12/2016).

Case control study

The odds of SSc being diagnosed after a diagnosis of cancer compared with the odds of SSc occurring where there has not been a previous occurrence of cancer will also be investigated. For this part of the study it is not feasible to use a cohort study of all people with a diagnosis of cancer hence a case control study will be used instead: the cases will have an incident diagnosis of SSc and exposure will be determined by identifying those with a previous diagnosis of cancer.

Cases will be patients with an incident diagnosis of SSc and their date of first diagnosis of SSc will be the index date. Controls will not have had any diagnosis of SSc by their index date and will be matched to cases on age, sex, GP practice and they must be contributing up to standard data in the same calendar time as their matched case. The index date for the controls will be the same as their matched case's index date. Diagnoses for cancer before index date will be identified for all patients included in the study and the researcher will be blinded to each individual's case or control status. To understand the temporal relationship between the diagnosis of cancer and any subsequent diagnosis of SSc, backdating the cases' index dates and Cox proportional hazards methods will be used to analyse this data.

H. Feasibility counts

Feasibility counts have been determined using CPRD data until the end of 2014 however for this study we will have a new load of data that runs until the end of 2016; ~~we will also be using EMIS data so these numbers will increase.~~

Descriptive study determining prevalence of SSc and cohort study

There should be at least 5400 cases of SSc in the CPRD including both incident and prevalent cases. Potential cases of cancer: 420

Descriptive study determining incidence of SSc and case control study

Number of incident cases of SSc: 2200 patients.

Descriptive study: mortality rates

For the determination of mortality rates, there are at least 280 deaths in people with SSc.

I. Sample size considerations

Sample sizes were calculated assuming significance level of 0.05, power = 0.8.

Cohort study

To look at time to diagnosis of cancer in this cohort, the proposed 1:6 matching should allow a hazard ratio of 2.1 to be detected. For this, 77 events would be needed which would require a sample size of 26338 patients altogether (3762 with SSc) assuming incidence of cancer in the general population as 396 / 100 000/year¹⁵ and incidence of cancer in those with scleroderma: 819/100 000/year¹¹. For peripheral vascular disease incidence is around 19/1000¹⁶, cerebrovascular disease with incidence of 4.9/1000¹⁷ and myocardial infarction with incidence of

Applicants must complete all sections listed below
Sections which do not apply should be completed as 'Not Applicable'

2.42/1000¹⁸ these sample sizes will also be sufficient and for the more common outcomes will allow smaller hazard ratios to be detected. For pulmonary hypertension the incidence is much lower in the general population (prevalence of 10-52 / 1000 000¹⁹) however occurrence in SSc patients is thought to be higher and therefore this is still useful to investigate.

Case control study

There should be at least 2200 patients with incident SSc hence matching the case control study with one case and six controls should allow an odds ratio of at least 2.5 to be detected.

J. Data Linkage Required (if applicable):[§]

§Please note that the data linkage/s requested in research protocols will be published by the CPRD as part of its transparency policy

The linkage to full and out patient Hospital Episode Statistics will be used to identify any further cases of SSc that have not been recorded in the CPRD. The number of patients who will be identified in the CPRD with SSc is likely to be low therefore being able to identify further patients in HES data with SSc that may not have been recorded in the CPRD will be beneficial to the study and will also allow us to see how well SSc is recorded in general practice.

The cancer registry will be used to identify cases of cancer in patients included in the study in linked practices. Information requested from cancer registry data will give definitive information on cancer diagnoses for the sample of the population that has linked data and will also provide information that will help in determining if cancers are detected earlier or later in people with SSc compared to people who do not have SSc. This will allow comparisons about diagnosis (behaviour, histology, lesion size, laterality, multifocal, nodes excised and involved, detected by screening, HER2 status, NPI score, Clark's level, Breslow thickness, Gleason primary pattern, secondary grade, tertiary, combined and tumour count), prognosis (basis of diagnosis, morphology, grade, tumour size, staging, receptor status, screening) and about treatment and any operations that have taken place (number of tumours, imaging code and site). This information will be very important and informative for clinical practice to aid understanding of whether patients with SSc have more aggressive cancers with potentially poorer outcomes. Also, knowing whether patients with SSc have cancers picked up by national surveillance programmes (screen detected) will have implications for how patients are managed. Data on ethnic origin will supplement the data available in the CPRD and data on sex, age at diagnosis and vital status will provide a useful check with that held on the CPRD. By cross checking with CPRD data it will be determined if there is a lower rate of recording of cancer in the population whose data is not linked.

The time periods available for all linked data mostly overlap the study period therefore lack of coverage is not expected to be problematic.

K. Study population

a) *Source population:* Patients included in this study will be selected from those who are permanently registered acceptable patients with at least twelve months of data between their Icen and rcens dates. Only patients with data up to research standard recorded between 1998 and 2016 will be included. ~~Practices contributing to EMIS that have previously contributed to CPRD will be excluded.~~ The start of follow up for patients will be the date that they join the GP practice, the date that the GP practice starts contributing up to standard data to the CPRD or 1/1/1998, whichever is later. The end of follow up for patients is the date that the GP practice stops contributing up to standard data to the CPRD, the date that the patient is diagnosed with the outcome under study, the date that the patient dies or 31/12/2016, whichever is earlier. The CPRD date of death will be used in defining exit from the cohort, where applicable.

b) *Study population*

Applicants must complete all sections listed below
Sections which do not apply should be completed as 'Not Applicable'

Identification of patients with SSc in the CPRD: Patients with a diagnosis of SSc will be identified using medical diagnosis codes (given in Appendix 1) with supportive evidence such as the presence of other ACR/EULAR classification criteria (e.g. Raynaud's phenomenon, digital ulcers, pulmonary hypertension, interstitial lung disease etc.), positive autoimmune serology and the use of drugs commonly prescribed in SSc such as vasodilator medication, proton pump inhibitors and immunosuppressive agents (including the regular assessment of specific blood tests that might indicate immunosuppressive medication usage). Identified patients will be described in terms of their characteristics including comorbidities and prescribing. Efforts will be made to exclude localised forms of scleroderma (e.g. morphoea). Relevant medication including, PPIs, calcium channel blockers especially nifedipine and ACE inhibitors will also be identified as supporting a diagnosis of SSc. The CPRD and EMIS data used to identify these patients will include clinical, therapy, test and referral data.

Identification of SSc in the HES: In order to ensure that we have identified all potential cases of SSc from the CPRD, we will search the linked HES full and inpatient admission data for all diagnoses of SSc. We will then compare the date of the first record of SSc in the CPRD and HES records and, where no diagnosis of SSc is recorded in the CPRD, establish the presence of supporting evidence to confirm SSc diagnosis, such as those outlined above.

c) *Index date definitions:* For the cohort studies, patients diagnosed with SSc will be matched to patients who do not have a diagnosis of SSc. The index date will be defined as the date of diagnosis of SSc or the start date of the study (1/1/1998), whichever is later for those with SSc and the matching date for those who do not have SSc. Given the low number of people diagnosed with SSc, both incident and prevalent cases of SSc will be included in this study and adjustments made to the results to account for potential biases that arise from including prevalent cases.

For the case control study: Cases will be defined as those with an incident diagnosis of SSc that is made at least one year after their Icens date and before 31/12/2016. At least one year of data after Icens and before SSc diagnosis (index date) is needed to ascertain relevant exposures including any diagnoses of cancer and to ensure that the cases of SSc are incident. Controls will not have a diagnosis of SSc and will be matched to the cases in the ratio of one case to six controls. The index date will be the date of diagnosis of SSc for the cases and the matching date for the controls.

d) *Incidence, prevalence:* The incidence of SSc will be calculated by year of diagnosis (2000-2016) and age at diagnosis using the rest of the population present on the CPRD or EMIS who have at least one year of up to standard data after their Icens date, as the denominator. An incident diagnosis of SSc will be defined as the patient having been present in the CPRD or EMIS for at least one year after their Icens date without a diagnosis of SSc occurring until at least one year after their Icens date. Incidence rates will be standardised to the 2012 Eurostat population. Point prevalence rates will be calculated by counting the number of cases with SSc diagnosed at the point that the prevalence rate is being determined for; the denominator data will be found using the population count for the CPRD or EMIS for all people with at least one year of up to standard data and that their data contribution includes the year that the prevalence is being calculated for. These will be broken down by age and sex. The incidence of the outcomes cancer, myocardial infarction, cerebrovascular disease, peripheral vascular disease and pulmonary hypertension will be found in a similar way.

e) *Sampling from a base population:* This will not be used in this study.

f) *Definition of exposure window:* For the cohort studies, exposure will be defined as whether or not a patient has had SSc diagnosed; once a diagnosis of SSc has been made, the patient will remain exposed for the entire study. For the case control study, exposure will be a

7

Applicants must complete all sections listed below
Sections which do not apply should be completed as 'Not Applicable'

diagnosis of cancer before the index date.

~~g) **Data linkage information:** The coverage periods mostly overlap the study period with HES outpatient having the most limited coverage. It is hoped that HES data will help to identify patients with SSc not identified in CPRD records so this may have a small impact. Cancer registry data does not include the final few years of the study period however this information may be identified in HES records: this will provide useful information for future work.~~

L. Selection of comparison group(s) or controls

a) *How comparison groups differ from the main study population:* The comparison groups for the cohort and case control studies will be sampled from the main study population but will be matched by age, sex, GP practice and calendar time to the patients with SSc.

b) *The inclusion, exclusion and data used for each comparison group:* To be included in the comparison group, all patients will need to be in the source population and be selected as being matched to a patient with SSc for one of the cohort studies or the case control study. The data used to make this selection will include clinical, therapy, test and referral data.

c) *Matching:* Incident density sampling will be used to find the matched comparator groups. The data used in identifying these comparators will include clinical, referral and test data in order to ensure that the comparator patients do not have SSc and, for the cohort studies, a diagnosis of one of the outcomes before the index date.

Cohort study:

Patients with a diagnosis of SSc will each be matched to six controls who do not have a diagnosis of SSc; matching will be done on age, sex and GP practice. Matched sets will need to be contributing data to the CPRD ~~or EMIS~~ during the same time period to avoid any changes in rates of diagnosis of outcomes or patient management that might occur between the different years of the study. All patients included in each cohort study must not have had any diagnoses of the outcomes being investigated before the index date. It is recognised that including prevalent and incident cases of SSc in the cohort studies will result in bias through some patients not living long enough to be included in the cohort: this will be accounted for in the analysis of this data.

Case control study

Cases will be defined as those who have an incident diagnosis of SSc and these patients will be matched to a random sample of up to six controls who do not have any diagnosis of SSc, by age, sex, GP practice and index date.

M. Exposures, Health Outcomes[§] and Covariates

[§]*Please note: Summary information on health outcomes (as included on the ISAC application form above) will be published on CPRD's website as part of its transparency policy*

Descriptive study:

Identification of deaths in the study population:

Records will be identified from the information held in the CPRD, ~~EMIS, HES and cancer registry data~~ regarding the date of death and this will be confirmed using follow up information such as records recording an autopsy or issue of death certificate. Where Read codes for cause of death are given, these will be reported on.

Cohort study:

Exposures: The exposure being investigated in the cohort study is the diagnosis of SSc which will be identified using Read codes for SSc along with supporting evidence as described in section K. A sample of patients who are identified as being diagnosed with SSc and any patients where the diagnosis is uncertain will be reviewed in the medical records browser by Dr John Pauling who is a consultant rheumatologist. The medical records browser that we have developed in-house allows us to view records in chronological order.

Applicants must complete all sections listed below
Sections which do not apply should be completed as 'Not Applicable'

Outcomes:

Cancer: For practices in the CPRD that are linked to the cancer registry data, date of diagnosis, type of cancer diagnosed and severity will be determined from the registry information held using information on site of neoplasm, morphology, behaviour of the cancer, lesion / tumour (size, number, multifocal, grade), nodes excised / involved, laterality, excision margin, cancer stage and imaging (site and code) used to confirm diagnosis. Where CPRD and EMIS practices are not linked to the cancer registry, information about the date and type of diagnosis of cancer will be found by identifying Read codes, and referral, treatment information including surgery, chemotherapy and radiotherapy will be used to verify diagnosis.

Myocardial infarction, cerebrovascular disease, peripheral vascular disease and pulmonary hypertension: where CPRD practices are linked to HES data, information from HES will be used to supplement the data available in the CPRD. For all patients, Read codes will be identified in the CPRD and EMIS for these outcomes or related symptoms (e.g. intermittent claudication for peripheral vascular disease²⁰) and supporting information including relevant prescribing for each of these outcomes (ACE inhibitor, statin, aspirin, beta blocker, anticoagulant, angiotensin receptor blockers, diuretics), follow up referrals and other tests will be used to identify the outcome.

Lists of Read codes for all of these outcomes will be informed by work that we have done previously; these lists will be reviewed by Dr John Pauling to ensure their accuracy and any outcomes that are uncertain will be reviewed in the medical records browser. Product information will be found using BNF codes (ACE inhibitor [2.5.5.1], statin [2.12], aspirin [2.9], beta blocker [2.4], anticoagulant [2.8.2], angiotensin receptor blockers [2.5.5.2], diuretics [2.2]).

Case control study

Outcome: The outcome is the incident diagnosis of SSc; these will be determined in the same way as in the cohort studies. An incident case will need to have their first diagnosis of SSc during the study period with at least one year of data after the patient's left censor date before this first record of SSc.

Exposures:

The exposure that will be investigated is the diagnosis of cancer before the index date for cases and controls. Records prior to the index date will be checked to identify those patients who have a diagnosis of cancer or treatment for cancer before their diagnosis of SSc using cancer registry or CPRD or EMIS data, as described above (under cohort study outcome). The identification of cancer related records will be undertaken blinded to case or control status of the patients.

Covariates:

Information on the following covariates at index date will be obtained from the CPRD and EMIS: age, sex, BMI, smoking status, alcohol intake, prescribing of immunosuppressants, hormone replacement therapy, oral contraceptives, vasodilators, calcium channel blockers, proton pump inhibitors. Comorbidities including diabetes, cardiovascular disease, previous myocardial infarction, previous cerebrovascular disease and SSc related comorbidities including pulmonary hypertension, inflammation in the lungs, kidney and liver problems will also be identified. Prescribing for SSc including ACE inhibitors (2.5.5.1), calcium channel blockers (2.6.2) especially nifedipine and PPIs (1.3.5) will also be identified using the BNF codes in the product table, in the absence of alternative explanatory diagnoses for the prescriptions.

N. Data/ Statistical Analysis

All results will be presented with 95% confidence intervals. The analyses will include both data from the CPRD and from EMIS although comparisons will be made between results obtained

Applicants must complete all sections listed below
Sections which do not apply should be completed as 'Not Applicable'

from both databases to determine if there are any differences in recording of data.

Descriptive study:

- Calculate the age- and sex- specific and standardised incidence rates, annual prevalence and standardised mortality rates for SSc in the UK population.

Incidence and prevalence of SSc will be determined as described in section K (d). All patients with an incident diagnosis of SSc will be described by sex, age, smoking status, BMI, alcohol intake and other related comorbidities/co-prescribing at diagnosis as described in section M. Patients with a prevalent diagnosis of SSc will be described separately in terms of these patient characteristics at the date that they join the study population and the time since the diagnosis of SSc will be included. Descriptions will include counts for each category and the proportion of patients in each category.

~~We will investigate the impact of potential under recording of SSc diagnoses in the CPRD and EMIS by adjusting the incidence and prevalence estimates for any additional cases of SSc identified using the CPRD data that is linked to HES. The initial analysis will only include data from linked practices. Cases of SSc identified from both HES and the CPRD and those identified from the CPRD and EMIS alone will be used to calculate two sets of age and sex stratified rates. The ratio of these two sets of rates will be modelled using inverse polynomial regression to create a set of correction factors for each age group by sex. The correction factors will then be applied to the entire CPRD and EMIS data to obtain age and sex specific directly standardised age adjusted incidence and prevalence estimates.~~

- Describe the patients who have a diagnosis of cancer and SSc within a very short time interval

All patients with a diagnosis of cancer and SSc close in time will be described in terms of the patient's age, sex, smoking status, alcohol intake and other covariates (given in section M). The type of cancer diagnosed will be identified and described for the patient group and by time.

Cohort study:

- Determine the crude and standardised incidence rates and gender specific incidence rates for cancer, myocardial infarction, pulmonary hypertension, cerebrovascular disease and peripheral vascular disease in people with SSc and compare these to people who do not have SSc
- Identify any differences in incidence of different types of cancer in people with a diagnosis of SSc compared to people who do not have SSc

The characteristics of the patients in each cohort study will be described at index date in terms of age, sex and the covariates described above (section M). Crude and standardised incidence rates for each of the outcomes diagnosed after the index date will be determined (as described in section K). Comparisons will be made between the standardised rates for different types of cancer. ~~For patients with records in the cancer registry, comparisons of severity of cancer will be made between those with SSc and those who do not have SSc (as outlined in section J) using relevant information about diagnosis and histopathology (morphology, histology, behaviour, grade, lesion size, multifocal, nodes excised and involved, whether cancer detected through screening, stage of cancer from Duke's stage / FIGO stage/ Clark's level / Gleason pattern or grade / T, N or M stage as relevant, number of tumours and tumour size, prognosis / NPI score) and how the diagnosis was made (basis of the diagnosis, whether the patient underwent screening). This will enable us to determine if there is potentially later diagnosis of cancer occurring in groups of patients or if cancer is more serious in different patient groups. Any further details about the type of cancer diagnosed (HER2 status, progesterone receptor status and score, oestrogen receptor status and score) and pathology information (T/N/M stage) will be described if sufficient patients are included as this could be useful in thinking about the~~

Applicants must complete all sections listed below
Sections which do not apply should be completed as 'Not Applicable'

immunogenicity of cancers. Information from the cancer registry includes ethnic origin, age, sex and date of birth of the patient; this data will be compared with that held in the CPRD to determine if there are any differences. Where differences exist, these will be noted and the CPRD data will be used preferentially to maintain consistency as this data is available for all patients in the study.

- Determine if there are differences in time to diagnosis of serious outcomes including cancer, myocardial infarction, cerebrovascular disease, pulmonary hypertension and peripheral vascular disease and time to death in people with SSc compared to people who do not have SSc
- Determine if there are any additional risk factors associated with the occurrence of these serious outcomes in SSc patients compared to people who do not have SSc

Time to event analysis will be used to compare the time from SSc diagnosis to diagnosis of each of the outcomes and the time to death with adjustment for confounders for those with SSc compared with those who do not have SSc. Covariates will be included in the model where initial analyses have indicated that these could confound the relationship between SSc and cancer occurrence. Where there is potentially effect modification, stratified analyses will be used to investigate this further.

Survival analysis with left truncation will be used to account for the bias included by not counting those people who have one of the outcomes being studied but do not survive to be included in the cohort and the corresponding prevalent cases of SSc diagnosed before the time period of interest who do survive to be included. A number of methods for dealing with this situation have been proposed including non-parametric models²¹⁻²³, semi-parametric based on Cox models²⁴ and accelerated failure time models.²⁵ Work will be undertaken to ensure that the best model for this data is selected by using simulation and sensitivity analyses. For example an analysis of just incident cases of SSc and the occurrence of the outcomes will provide an estimate of the relationship but this will suffer due to the small sample size. By using the data from this to power simulation studies this will help to inform overall model selection. Sensitivity analyses will include: analysis of just those with an incident diagnosis of SSc matched to those who do not have a diagnosis of SSc; analysis of those with an SSc diagnosis at five year intervals (five years, ten years, fifteen years etc.) before index date. Any substantial differences in results will be investigated further.

Case-control study:

- Explore the temporal relationship between cancer followed by a diagnosis of SSc

The characteristics of those included in the case control study will be described and compared. A sub group comparison of those cases who have and do not have a history of cancer will be made. Conditional logistic regression will be used to determine the odds of getting a diagnosis of SSc following a diagnosis of cancer; this will be adjusted for the covariates identified. Cox proportional hazards models, weighted for use with case-control data will also be used, if the model assumptions are met. To explore any temporal relationship between cancer diagnosis and SSc and determine if there is a difference in odds of SSc related to the time since cancer diagnosis, the index date of the cases will be backdated and new risk estimates calculated.²⁶

O. Plan for addressing confounding

Matching the patients included in the cohort studies and matching the cases and controls in the case control study will help to overcome confounding by age, sex and GP practice. Matching by GP practice can overcome differences in socioeconomic status across different patient groups attending different surgeries as well as potential differences in treatment and referral practice

Applicants must complete all sections listed below
 Sections which do not apply should be completed as 'Not Applicable'

between different general practices in different parts of the UK which could have an impact on time to cancer diagnosis. Patient characteristics will be described for the cohort and case control studies with comparisons made between the comparator groups to identify any potential confounders that need to be taken into account. These potential confounders will each be included in the regression analyses and assessed to determine their impact on the model and whether they should be kept in the final model for adjustments to be made.

P. Plans for addressing missing data

~~The diagnosis of SSc may only be recorded in secondary care and not primary care. To gauge the extent of this, HES data will be checked for those practices that are linked and any extra diagnoses will be identified. The difference in ascertainment between primary and secondary care records will be determined and analysed as described above.~~
 It is expected that the recording of cancer diagnoses will be complete for those whose records are linked to cancer registry data. For patients whose records are not linked, we will use all relevant medical, referral and test codes to identify records of cancer diagnosis and verify that these are accurate. ~~Similarly for the atherosclerosis related outcomes, HES data will be used for linked practices to identify cases.~~ Any differences between numbers of outcomes identified in linked records compared with CPRD records alone will be used to inform the project in terms of the likely number of missing cases. It is likely that there will be missing data on BMI, smoking status and alcohol consumption: these will be imputed by multiple imputation.

Q. Patient or user group involvement (if applicable)

The study has been part funded by Scleroderma UK who will help with dissemination of findings from this through their annual conference and publications. Dr. Pauling has a Patient Public Involvement group who have an interest in research work in SSc and he will take this study to this group for input on patients' views and experiences.

R. Plans for disseminating and communicating study results, including the presence or absence of any restrictions on the extent and timing of publication

The results of this study will be presented at internal research meetings, national and international conferences and submitted for publication in peer reviewed journals. It is expected that papers on the epidemiology of SSc and any association or not between SSc and cancer will be published. ~~This work may also result in methods papers assessing the impact of EMIS data on the study as well as the extent of missing data coming from the work with linked CPRD data.~~

S. Limitations of the study design, data sources, and analytic methods

~~As described above, there are limitations for the practices that are not linked to HES or cancer registry data with respect to some diagnoses of SSc or cancer with the potential of being missed.~~ Members of this research group have studied colorectal cancer in previous work without using linkage data therefore input will be gained from them on the methods used to ensure as complete a set of outcomes as possible is identified.
 There is the potential for including cases of localised rather than systemic scleroderma: medical records and prescribing records along with patient record review where the diagnosis is uncertain will be used to verify the disease diagnosed.
 The inclusion of prevalent cases of SSc has been done in order to increase the sample size of the study and be able to take into account the occurrence of cancer that has a longer latency period or occur later in the disease process. This has the potential to bias the results and underestimate the risk of cancer in people with SSc due to an overinflated denominator population. This will be overcome using left truncation methods in survival analysis and through the use of sensitivity analysis to estimate the effect this bias may have.

Applicants must complete all sections listed below
Sections which do not apply should be completed as 'Not Applicable'

References

1. Gabrielli A, Avvedimento EV, Krieg T. Scleroderma. *N Engl J Med* 2009; 360:1989-2003
2. Mayes MD, Lacey JV, Jr., Beebe-Dimmer J, et al. Prevalence, incidence, survival, and disease characteristics of systemic sclerosis in a large US population. *Arthritis Rheum* 2003; 48:2246-55
3. Medsger TA, Jr., Masi AT. Epidemiology of systemic sclerosis (scleroderma). *Ann Intern Med* 1971; 74:714-21
4. Allcock RJ, Forrest I, Corris PA, Crook PR, Griffiths ID. A study of the prevalence of systemic sclerosis in northeast England. *Rheumatology (Oxford)* 2004; 43:596-602
5. Arias-Nunez MC, Llorca J, Vazquez-Rodriguez TR, et al. Systemic sclerosis in northwestern Spain: a 19-year epidemiologic study. *Medicine (Baltimore)* 2008; 87:272-80
6. Elhai M, Meune C, Avouac J, Kahan A, Allanore Y. Trends in mortality in patients with systemic sclerosis over 40 years: a systematic review and meta-analysis of cohort studies. *Rheumatology (Oxford)* 2012; 51:1017-26
7. Moinzadeh P, Fonseca C, Hellmich M, et al. Association of anti-RNA polymerase III autoantibodies and cancer in scleroderma. *Arthritis Res Ther* 2014; 16:R53
8. Roumm AD, Medsger TA, Jr. Cancer and systemic sclerosis. An epidemiologic study. *Arthritis Rheum* 1985; 28:1336-40
9. Abu-Shakra M, Guillemin F, Lee P. Cancer in systemic sclerosis. *Arthritis Rheum* 1993; 36:460-4
10. Derk CT, Rasheed M, Artlett CM, Jimenez SA. A cohort study of cancer incidence in systemic sclerosis. *J Rheumatol* 2006; 33:1113-6
11. Chatterjee S, Dombi GW, Severson RK, Mayes MD. Risk of malignancy in scleroderma: a population-based cohort study. *Arthritis Rheum* 2005; 52:2415-24
12. Au K, Singh MK, Bodukam V, et al. Atherosclerosis in systemic sclerosis: a systematic review and meta-analysis. *Arthritis Rheum* 2011; 63:2078-90
13. Cannarile F, Valentini V, Mirabelli G, et al. Cardiovascular disease in systemic sclerosis. *Ann Transl Med* 2015; 3:8
14. Chaisson NF, Hassoun PM. Systemic sclerosis-associated pulmonary arterial hypertension. *Chest* 2013; 144:1346-56
15. (2017) Cancer Research UK: Cancer statistics for the UK. Available from: <http://www.cancerresearchuk.org/health-professional/cancer-statistics>, accessed: 20/2/2017
16. Fowkes G. Peripheral Vascular Disease. Available from: www.birmingham.ac.uk/Documents/college-mds/haps/projects/.../09HCNA3D2.pdf, accessed: 20/2/2017
17. Sozio SM, Armstrong PA, Coresh J, et al. Cerebrovascular Disease Incidence, Characteristics, and Outcomes in Patients Initiating Dialysis: The CHOICE (Choices for Healthy Outcomes in Caring for ESRD) Study. *American journal of kidney diseases : the official journal of the National Kidney Foundation* 2009; 54:468-77
18. Scottish Heart Disease Statistics: National Statistics; UK Statistics Authority 2016.
19. Hoepfer MM, Simon R, Gibbs J. The changing landscape of pulmonary arterial hypertension and implications for patient care. *European Respiratory Review* 2014; 23:450
20. Clarson LE, Hider SL, Belcher J, et al. Increased risk of vascular disease associated with gout: a retrospective, matched cohort study in the UK Clinical Practice Research Datalink. *Annals of the Rheumatic Diseases* 2014;
21. Asgharian M, Cyr Emile ML, Wolfson DB. Length-Biased Sampling with Right Censoring: An Unconditional Approach. *Journal of the American Statistical Association* 2002; 97:201-9
22. Wang M-C. A Semiparametric Model for Randomly Truncated Data. *Journal of the American Statistical Association* 1989; 84:742-8
23. Wang M-C. Nonparametric Estimation from Cross-Sectional Survival Data. *Journal of the American Statistical Association* 1991; 86:130-43

Applicants must complete all sections listed below
Sections which do not apply should be completed as 'Not Applicable'

24. Qin J, Shen Y. Statistical Methods for Analyzing Right-censored Length-biased Data under Cox Model. *Biometrics* 2010; 66:382-92
25. Shen Y, Ning J, Qin J. Analyzing Length-biased Data with Semiparametric Transformation and Accelerated Failure Time Models. *Journal of the American Statistical Association* 2009; 104:1192-202
26. Charlton RA, Snowball JM, Bloomfield K, de Vries CS. Colorectal Cancer Risk Reduction following Macrogol Exposure: A Cohort and Nested Case Control Study in the UK. *PLoS ONE* 2013; 8:e83203

List of Appendices (Submit all appendices as separate documents to this application)

- Appendix 1: Code list for SSc
- Appendix 2: Sample size calculations